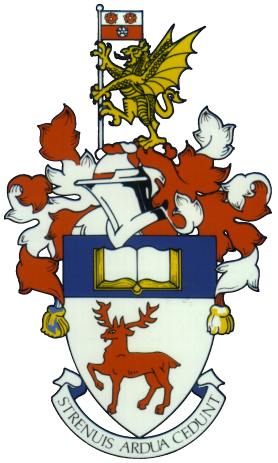
UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

Mathematical Sciences



Risk Analysis of User Satisfaction in Online Communities

by

Philippa Alice Hiscock

Thesis submitted for the degree of Doctor of Philosophy
October 2014

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES
Mathematical Sciences

Doctor of Philosophy

RISK ANALYSIS OF USER SATISFACTION IN ONLINE COMMUNITIES

by Philippa Alice Hiscock

Question-and-answer online communities require people to actively participate, and people can be motivated by their desire for recognition. Recognition may be in the form of an answer to a question, or the award of points by a peer. Businesses exploit these facts to provide cost-effective customer service platforms in the form of online communities.

As setting-up and maintaining an online community can be expensive, community managers are incentivised to understand the dynamics of their community. The satisfaction of *users* of the community is a predominant factor in understanding community dynamics. Any event which negatively impacts upon the satisfaction of individual users can, in-turn, negatively alter community dynamics. Such an event poses a *risk* to community "health".

We show that events which are well defined and well formulated can be predicted to provide real-time, cost-effective, risk management. To demonstrate this, we formulate two novel interpretations of risks pertaining to user satisfaction: low questioner satisfaction, and individual user "churn". In formulating these, we show that improvements in risk analysis can be achieved by focusing on the process of knowledge creation, rather than on isolated actions.

Risks related to user satisfaction are traditionally formulated as binary events, and binary events are usually modelled by classifiers. We consider a series of classifiers to model each risk formulation. In contrast to previous works, we use two-dimensional performance measures which allow us to analyse the many aspects of classifier performance. Consequently, the models we propose are suitably verified for their inclusion in an automated risk management platform.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Philippa Alice Hiscock, declare that this thesis entitled 'Risk Analysis of User Satisfaction in Online Communities' and the work presented in it are my own. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed: .....

Date: .....

Acknowledgements

First of all, Michale, děkuji ti z celého srdce za všechnu tvoji trpělivost. Most importantly, I would especially like to thank my father for being a constant source of inspiration and my mother for her never ending support (and infallible proof-reading abilities). I am eternally grateful for the weeks, days and hours spent in assisting me with my writing over the years. Although only a recent addition to the family, Hunter has given me a great excuse to escape the confines of the office over the past few years.

My thanks go out to my supervisors, Jörg Fliege and Athanassios Avramidis for their support and encouragement, particularly during times of deadline induced panic. In addition, I would like to thank Jon Forster for his statistical guidance and agreeing to be an addition to the team. Also, I would not have stayed dry whilst writing this thesis if not for the many porters, electricians, and other general maintenance people.

The EU FP7 project ROBUST (EC Project Number 257859) was a truly fantastic experience, and I am eternally grateful to have had so many opportunities to further my development. This project provided such a unique and interesting basis for my research. One of the highlights has to be walking up Mount Carmel accompanied by my fellow colleagues who took it in turns to sing me happy birthday in their native languages. My sincere gratitude goes out to Dr. Adrian Mocan who acted as the representative for SAP, his inside knowledge of the studied online community was invaluable. In addition, I am eternally grateful to Toby Mostyn from Polecat for assisting me in developing my programming skills.

Edwin Tye has been my constant companion during my time at Southampton. His memory is unbelievable, storing an incredible amount of information which he is always happy to share. I would finally like to acknowledge my other office mates and fellow PhD students across the entire maths department for their company over the years.

Chapter 1

Introduction

The origin of online communities can be traced back to the 1970's when the first computer mediated communication systems were developed (Hiltz, 1985). The International Telecommunication Union (2014) was formed in 1965 as a specialized agency of the United Nations (concerned with information and communication technologies). This agency has observed the use of the internet proliferate throughout the world, monotonically non-decreasing between 1997 and 2013 in both the developed and developing worlds. By the end of 2013, approximately 77% of inhabitants in the developed world (39% globally) were using the internet, having increased monotonically from 40% (10% globally) a decade earlier (International Telecommunication Union, 2013). The omnipresence of the Web has enabled millions of individuals to participate in online communities. As the number of participants in online communities continues to grow, so too does the breadth and depth of the knowledge base available within these communities. The knowledge shared in online communities may be amassed by the service provider in large repositories. Anderson et al. (2012) and others recognise the long term value of such repositories for Question and Answer (Q&A) sites.

Online communities are not limited to social domains, being widespread in various business, scientific and public service domains. Likewise, substantial economic value is no longer only generated by high profile public communities, e.g. Twitter and Facebook, but also by business communities, such as the SAP Community Network (SCN) (`http://scn.sap.com/`) and IBM's Connections (`http://www-03.ibm.com/software/products/us/en/conn`). Both SAP and IBM are multinational corporations specialising in software. Prior to the close of the 20$^{\text{th}}$ century, the lack of business exploitation of the vast and rapidly growing online communities was beginning to be noticed (Armstrong and Hagel, 1999). Around

this time, marketing professionals were exploring ways to utilize online communities to strengthen brands (McWilliam, 2000). Online communities are now pivotal elements in corporate management and marketing, product support, customer relations management, product innovation and targeted advertising (Franke and Shah, 2003; Franke et al., 2006; Lin and Lee, 2006). Members of such communities are connected in a way that opinion, knowledge and ideas may be shared to facilitate collaboration.

Each online community is a valuable ecosystem, full of information. It is obvious that risks and overlooked emerging opportunities present threats to the health of such an ecosystem. Consider for example Yahoo Answers, which was one of the first Q&A online communities. In 2011, traffic on the service was reported to have dropped, with user growth stalling (Wang et al., 2013). In addition, Google's own Google Answers community was shut down before the end of 2006 after less than five years of service. Nonetheless, the Stack Overflow community has reached its sixth year and is still well regarded, whilst SAP's Community Network (originally launched as the SAP Developer Network in 2003) and IBM's Connections (launched in 2007) remain integral to the respective companies. Iriberri and Leroy (2009) empirically find the causes of online community death to be a lack of contribution, participation and quality content generation. Techniques that enable the health in online communities to be measured, managed, analysed, protected and optimized are therefore invaluable.

1.1 Addressing the Problem/Background

The need to model online communities is just one aspect of Information Communication Technology (ICT) in need of development. Therefore, it is not surprising that ICT was the largest research theme in the Seventh Framework Programme (FP7) of the European Commission that funded European research and technological development between 2007 and 2013. The desire of the European Commission to invest so heavily in this research theme was to place Europe at the forefront of shaping the future of ICT and to ensure that the benefits were widely disseminated throughout infrastructure and to all citizens. Full details of the FP7-ICT research theme are available at `http://cordis.europa.eu/fp7/ict/home_en.html`.

One of the successful proposals to the FP7-ICT call was Risk and Opportunity management of huge-scale BUSiness communiTy cooperation (ROBUST), a consortium of ten partners that commenced work in November 2010 and completed

in October 2013. The ROBUST project addressed the call by promising "methods to understand and manage the business, social and economic objectives of the users, providers and hosts and to meet the challenges of scale and growth in large [online] communities." (Gottron, 2010, p. 2). One of the most attractive prospects of the ROBUST project was researching on the industrial platforms of the three use-case partners SAP, IBM and Polecat. We performed initial testing on data and scenarios from all three use-case partners but only SAP provided data rich enough for detailed analysis. Therefore, we only discuss the SAP community and drivers within this thesis. All three use-case partners are multinational corporations but each is in a different stage of development, with IBM being the most established and Polecat the least. SAP was formed by workers from IBM and has headquarters in Germany. The company specialises in enterprise software for managing customer relations and business operations. IBM itself specialises in both hardware and software technology in addition to providing consultancy. Finally, Polecat is a fast growing firm that specialises in providing solutions which output meaningful visualisations of information.

The University of Southampton partner was contracted to develop and deliver a framework and a fully functional Java web applet to enable proactive risk and opportunity management for online communities in real time. The process of risk management as defined by ISO31000:2009 (2009) is illustrated in Figure 1.1. Hence the initial motivation for this thesis was to discover from the use-case partners what sort of user behaviours were of concern (i.e. viewed as a risk event to user satisfaction) and to build a framework to assess the likelihood of these (i.e. risk assessment). With regard to the European Commission's expectation of ROBUST deliverables, the risk assessment framework was to be coded in Java to provide a validated tool which was fully integrated within the final risk management tool. All methods developed for ROBUST were evaluated throughout development on live industrial test beds provided by the use-case partners SAP, IBM and Polecat to determine success. Finally, the methods were incorporated to produce software solutions and integrated into partner services. An open source demonstrator was released to the public. In addition to this, the agreement with the European commission required periodic documentation of all ROBUST work and software components as well as yearly review progress meetings. Details of our significant direct collaborations with the ROBUST partners is given in Table 1.1.

Table 1.1: ROBUST partners by work package and details of our direct collaboration.

Work Package	Lead Partner	Collaboration/Interaction directly related to this thesis
1	University of Southampton (CORMSIS & IT Innovations)	**We provided:** proof of risk analysis tool concept in R; validated and compatible risk analysis tool in Java; and co-design of graphical user interface in Java.
2	Technische Universität Berlin (Germany)	**They provided:** computational solutions for huge data; calculation of page rank.
3	The Open University	N/A
4	Universität Koblenz-Landau (Germany)	N/A.
5	National University of Ireland, Galway (Ireland)	Discussion on churn analysis for community level metrics.
6	Software Mind SA (Poland)	**We provided:** Java project for extracting features from community database. **They provided:** streaming integration capability and managed the overall ROBUST platform integration.
7	IBM Israel – Science and Technology LTD (Israel)	**They provided:** employee use-case, platform "Connections". **We provided:** little analysis due to lack of data.
8	SAP AG (Germany)	**They provided:** business partners use-case, platform "SAP Community Network". **We provided:** detailed analysis on fora of interest.
9	Polecat (Ireland)	**They provided:** web community use-case, cleaned data from open platforms (e.g. boards.ie). **We provided:** preliminary analysis on fora of interest due to non-business connection; validated risk analysis tool in R and Java.
10	TEMIS S.A. (France)	**They provided:** overall dissemination and exploitation. **We provided:** presentations of our research as papers, posters and talks.
11	Universität Koblenz-Landau (Germany)	**They provided:** overall project management. **We provided:** progress reports.

To allow for the demands of ROBUST on this work, the primary original aspect of this thesis is in the novel interpretation and development of formal risk event definitions. As by-product of the ROBUST demands, we produced a fully incorporated risk analysis tool which is used by the SAP and Polecat use-case partners. Following the completion of the ROBUST project and related demands in late 2013, we explored a secondary original aspect: how to robustly pick the most suitable classifier of risk occurrence given the event defined. To take account of the abstract research nature, this thesis is not structured typically (see Section 1.5).

1.2 Definition of an Online Community

The term *community* was originally coined by sociologists in a physical sense (such as size) to quantify a group of individuals. As the migration of people increased, becoming more fluid with the ease of travel, defining a community with respect to its size appears flawed and the word is now interpreted in terms of peoples' relationships (Preece, 2001). The working definition given by Preece (2000) states that an online community consists of a group of people with a shared purpose who are guided by a set of policies and whose interactions are supported and facilitated by computer systems. Where the interactions between people are not driven by the need for "life-supporting resources", the purpose of the interaction is primarily the exchange of information. For example, businesses supporting an online community platform enable their employees and/or customers to engage with other like-minded individuals.

There are many alternatives in the literature for the term 'online community' (Preece, 2001), which itself has many different interpretations (Preece, 2000). Two such alternatives are 'community of practice' (Wenger, 1998) and 'virtual community' (Rheingold, 1994). However, Preece (2001) is of the opinion that 'online community' is the most widely used and this is found to be true to this day within the literature (Lin and Lee, 2006; Maxwell Harper et al., 2008; Nasser et al., 2013a; Rowe et al., 2013; Tausczik and Pennebaker, 2011). There are still those who take care to avoid the use of this term, preferring to use only 'community' (Anderson et al., 2012).

Online communities offering a platform to support the asking and answering of questions is known as a question-and-answer (Q&A) online community. Such communities are knowledge-intensive (Schall and Skopik, 2011). By 2008, online Q&A communities were one of the most popular means of seeking information

on the web due to their effectiveness (Liu et al., 2008; Nam et al., 2009). Yahoo Answers was one of the first Q&A online communities and was recorded to still be the largest by Wang et al. (2013). Such a community hosted by a business can serve to provide customers with a moderated platform where they can ask questions regarding company products (e.g. SAP's Community Network), or to provide employees with an outlet to share ideas for product development or future research (e.g. IBM's Connections).

1.3 Definition of Risk Analysis

Let *risk* be defined as the "effect of uncertainty on objectives" (ISO31000:2009, 2009, p.1); *event* as the "occurrence or change of a particular set of circumstances" (ISO31000:2009, 2009, p.4); and *level of risk* as the "magnitude of a risk or combination of risks, expressed in terms of the combination of consequences and their likelihood" (ISO31000:2009, 2009, p.6). Given these definitions, the British Standards Institution define risk analysis as the "process to comprehend the nature of risk and to determine the level of risk" (ISO31000:2009, 2009, p.5) which inherently includes risk estimation. We see from Figure 1.1 that risk analysis is the basis for risk evaluation and subsequent decisions about risk treatment.

1.4 Objectives and Research Questions

There are three global objectives, or research questions, that we aim to address within this thesis:

1. Real-time automated risk analysis of user satisfaction in question and answer online communities.

2. Whether a simpler classification method would have been more suitable than the generalised linear model with probit link function under Bayesian inference that we implemented for the ROBUST project.

3. How to best analyse and compare classifiers considering both graphical and scalar metrics.

The first objective is a more specific version of the task assigned to us by the ROBUST consortium and in meeting this, we fulfil our agreement with the European Research Council. Given that the classifier we delivered to the ROBUST

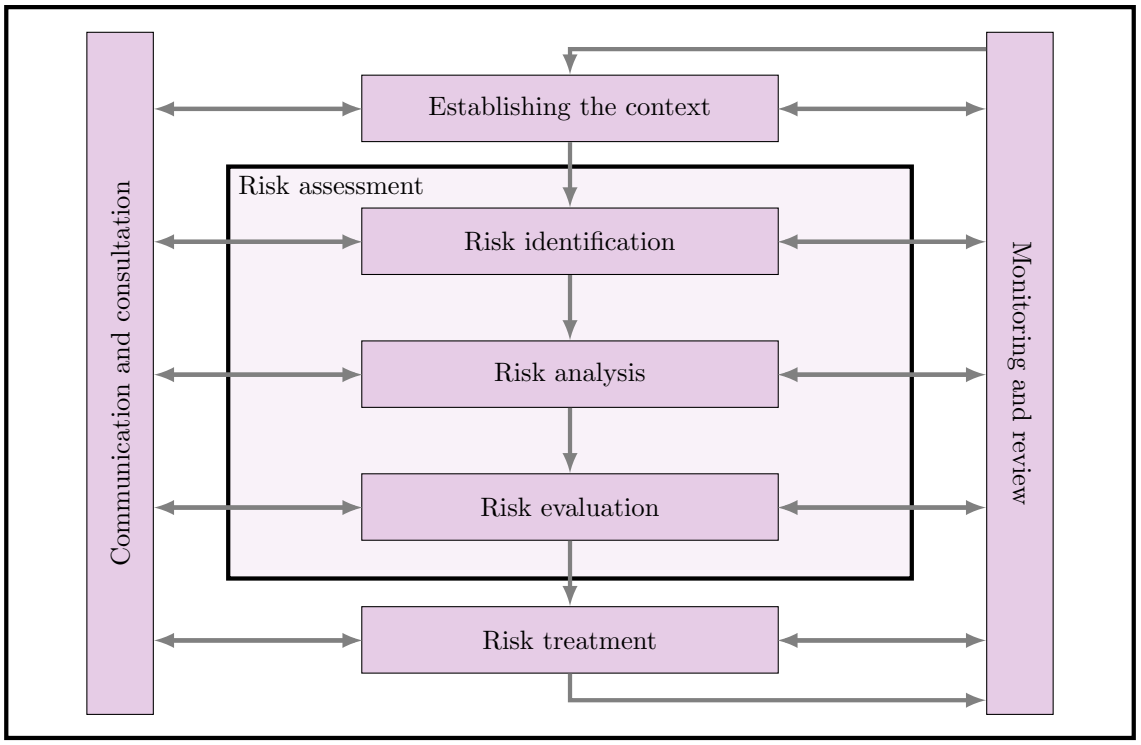


Figure 1.1: Risk management process as defined by ISO31000:2009 (2009).

consortium was an implementation of the generalised linear model with probit link function under Bayesian inference, the second objective is necessary to determine whether a simpler classification method would have given us comparable performance and thus been preferable. The third objective is driven by our understanding that the one-dimensional performance measures typically used are unreliable and misleading when comparing classifiers over all misclassification distributions. Therefore the third objective improves the quality of the analysis required by the second objective.

1.5 Outline of Thesis

The following six chapters may be briefly described as follows. Chapter 2 consists of three parts: firstly, it exposes the SAP Community Network (SCN) as the context of the analysis within this thesis; secondly, it introduces the concept of online community "health" with a focus on user satisfaction; thirdly, it summarises past research on modelling user satisfaction and extends to provide our novel formulations of the problem. This last part of Chapter 2 contains our main contributions within this thesis. Chapters 3 and 4 provide the theoretical under-

pinnings and fully describe the classifiers considered, from discriminant analysis to generalised linear models. The latter of these chapters exposes material for the classifier with Bayesian inference and includes a validation of our implementation of the classifier in R based on previously modelled data. Following the provision of information about the classifiers used, Chapter 5 offers novel insights into how to "best" analyse and compare classifier performance. Chapter 6 illustrates the results from the application of the given theoretical material to the risk events formulated in Chapter 2 and demonstrates how the different measures of classifier performance can lead to conflicting representations and unreliable conclusions. A closing discussion and exploration of possible extensions is found in Chapter 7.

Chapter 2

Problem Definition and Context

The objective of this chapter is three-fold. Firstly, to describe the SAP Community Network (SCN) and the data provided to us. Secondly, to communicate in general terms what is meant by online community health, highlighting that user satisfaction is a key aspect which should be monitored. Thirdly, to expose the gaps within the literature in accordance with SAP's concerns for user satisfaction (i.e. our problem space) and hence to highlight the importance of our research in providing novel event formulations to address the observed deficiencies.

2.1 Introduction

The health of question-and-answer (Q&A) online communities is largely dependant on the process of knowledge generation. For each question asked, there should be an efficiency and efficacy in the responses provided. Q&A communities are typically moderated by a person employed by the platform provider. It is assumed that this moderator is able to ensure than no duplicate questions exist and that all questions raised are topically relevant to the forum within which they are contained. Therefore there is a potential for the Q&A community knowledge repository to be expanded with every question asked, duplicate questions being either merged or pointed at a pre-existing question. We will show that this process of knowledge generation requires the community users to be satisfied. Hence the user posing the question must believe that their question post will be responded to and the user supplying the response must believe that they will be acknowledged by their peers. Events which may have a negative impact on user satisfaction therefore actually threaten community health.

In this chapter we first provide an outline of the SCN structure, following which

we summarise the data provided by the community's characteristics. We further explore the more general concept of online community health, before focusing on user satisfaction for community health. We then provide details of the three main approaches to modelling user satisfaction reported in the literature. These approaches are role analysis, questioner satisfaction and churn analysis. To avoid duplication of work within the ROBUST consortium, we only expose and do not develop further role analysis, that being the specific responsibility of our Open University partner within ROBUST. However, we observe a very clear gap in the literature concerning questioner satisfaction, which exposes the need for us to identify a new meaningful (questioner satisfaction) event to enable problem formulation to become straightforward. Regarding churn analysis, we find that the literature does not clearly define the concept of churn in online communities. Therefore we examine the literature addressing churn analysis with respect to various industrial applications (e.g. telecommunications, marketing and, where available, online communities). We expose the deficiencies found and address these as we formulate our proposed novel clear definition of the churn event specifically for online communities. Finally, we close this chapter by drawing conclusions concerning problem definition, highlighting our novel event formulations developed to address the observed gaps and deficiencies (i.e. to provide solutions to solve our problem space).

2.2 SCN: The SAP Community Network

The SCN is a business community platform where any uniquely registered person, referred to as a *user*, may discuss and share their ideas and issues regarding SAP products. This community mainly consists of a number of *fora*, each relating to a unique product or topic. A user may post a *message* in any forum, and a collection of messages forms a *thread*. Figure 2.1 provides an outline of containment in the SCN. With the heterogeneity of forum topics, the time stamp of a user's first message within a forum becomes that user's forum specific *registration date*. The first message in a thread is the *question* with subsequent messages being the *responses*. All messages are linked via a tree-like structure as demonstrated in Figure 2.2. In linking messages, they are given a *time rank* and *wall clock*, that is an arrival order and minutes since thread creation. The user who posts the 'parent' message, is known as the *original poster* (OP); user A is the OP in Figure 2.2. A user who makes a post in response to the parent message is called

a *respondent*. In Figure 2.2, users B, C, D and E are respondents. Within the tree-like message structure of a thread, the *most responded to message* (MRTM) is that with greatest number of messages posted in direct response. Similarly, the *most responded to user* (MRTU) (including the OP) is the user to whom the greatest number of direct responses is made of all users to post in the thread. We see in Figure 2.2 that the MRTM is the original post (3 responses) and the MRTU is both users A and B (4 responses).

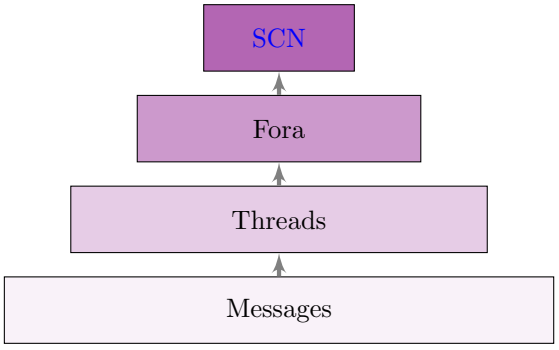


Figure 2.1: Outline of containment for elements in the SAP Community Network (SCN).

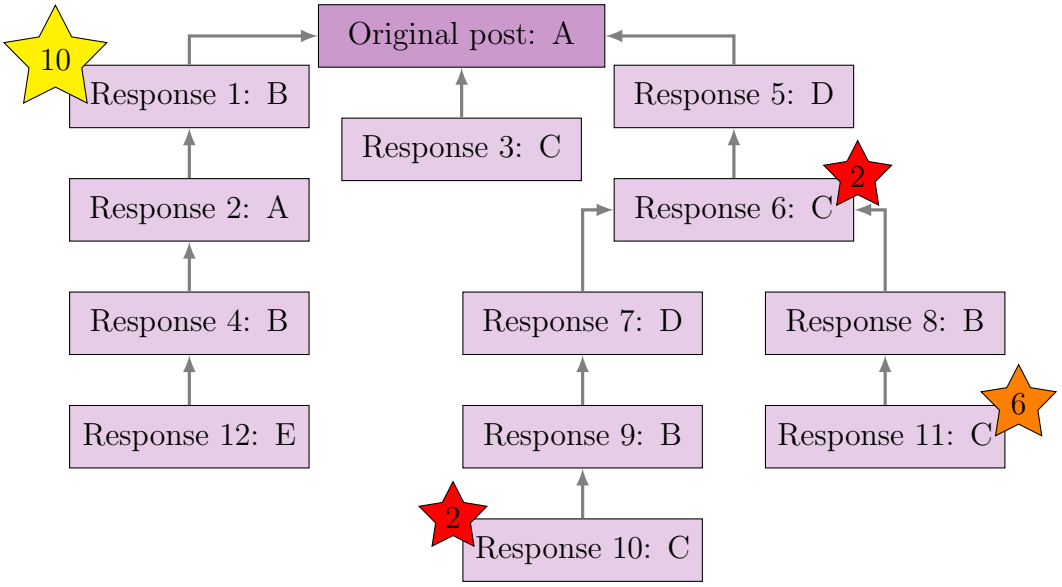


Figure 2.2: Example tree-like message structure between messages in a thread. Points awarded are represented by coloured stars and users by upper-case characters. An arrow from response index y to response index x, where the original post takes response index 0, indicates response y to have directly responded to response x.

The OP is the only user capable of making certain actions with respect to

their thread. Each respondent may be awarded *points* by the OP based on the quality of their response (see Table 2.1). The SCN places restrictions on the way an OP awards points in a thread such that only one 10 and two 6 point scores may be awarded. Consequently, we define a thread to be *solved* if, and only if, the OP has awarded a 10 point score to a response; the associated respondent is known as the *thread solver* (TS). A more relaxed version of the TS is the *highest point scorer* (HPS). Where the HPS is not the TS, there may be more than one HPS. Within Figure 2.2 user B is the TS, and also the HPS. In the SCN, points awarded are connected to the corresponding respondent's message. Therefore users can accumulate a peer-awarded *reputation* over time. The SCN "Contributor Reputation Program" tracks user reputation over a 12 month rolling window, providing users with a peer-awarded score. Being visible to others, the program's purpose is to motivate users to provide knowledge to the community in the form of answers to questions. Similar programs exist within the Q&A communities of Stack Overflow and Yahoo Answers. We view respondent reputation as being forum-specific due to forum topic inhomogeneity. Assuming a thread has at least one respondent, the *most reputable respondent* (MRR) is that with greatest lifetime reputation.

Table 2.1: SCN point awarding system via the original poster.

Original poster's view of respondent's post	Points awarded
Respondent 'solved' the issue of the parent post	10
Respondent was 'very helpful' on the issue of the parent post	6
Respondent was 'helpful' on the issue of the parent post	2

In addition, the OP can change the *status* of a thread from the default 'Unanswered' to 'Answered', the latter status implying a lack of point availability to potential respondents. However, there are no restrictions on when an OP may change the status of a thread. For example, a thread does not have to be solved to have status 'Answered' — the converse also holds true.

2.3 SCN Data Summary

SAP made available to us a complete trace of actions for 95 fora (a third of the total byte size of the SCN at time of extraction) from February 2004 to July 2011. Given that users of the SCN are spread across different time-zones, all timestamps in the data provided by SAP are standardised to Central European Time. Figure 2.3

shows how the number of fora posted in (i.e. active) per day increases as the SCN becomes a more established community. However, there is no obvious connection between the number of fora which are active and the number of messages posted as some fora have more user interest (and hence presence) than others. For example, over one half of the responses made and questions asked (threads created) are contained in the ten most active fora. Upon considering Figure 2.4, however, we hypothesise that there is a direct relationship between the number of questions asked (threads created), and the number of responses made. Table 2.2 shows almost 90% of all threads to have at least one response. Figure 2.5 implies the time series of these daily counts to not be normally distributed. Consequently we can use Kendall's tau rank to determine the level of co-monotonicity between the number of questions per day and the number of responses per day. An extremely high coefficient of 0.9374 (to 4 decimal places) supports our hypothesis; there is a monotonic relationship between the number of questions posted and the number of responses made per day (across the entire partition of the SCN).

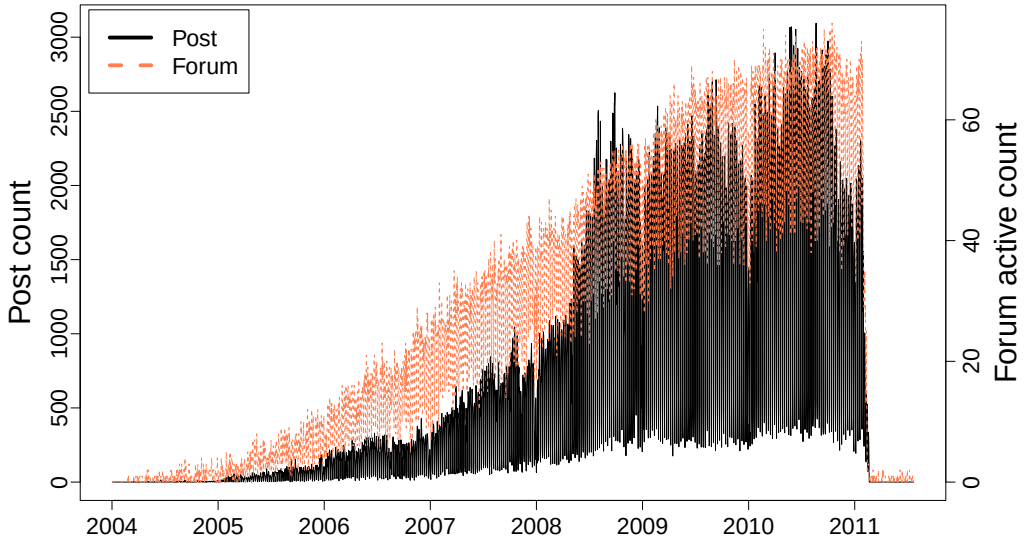


Figure 2.3: Number of posts, and fora posted in, per day in the SCN.

Reconsidering Figures 2.3 and 2.4, we see very little activity across the SCN prior to the year 2008. This indicates that the SCN became an established community at some time in the year 2008. That these graphics indicate the activity within the SCN to die in 2011 is a result of all fora being migrated to the updated SCN platform at this time. We deem it necessary to consider only those actions made during the years 2008 to 2010. During this window, 85.10% of messages were posted, 84.16% of threads were created and 74.40% of users made their first

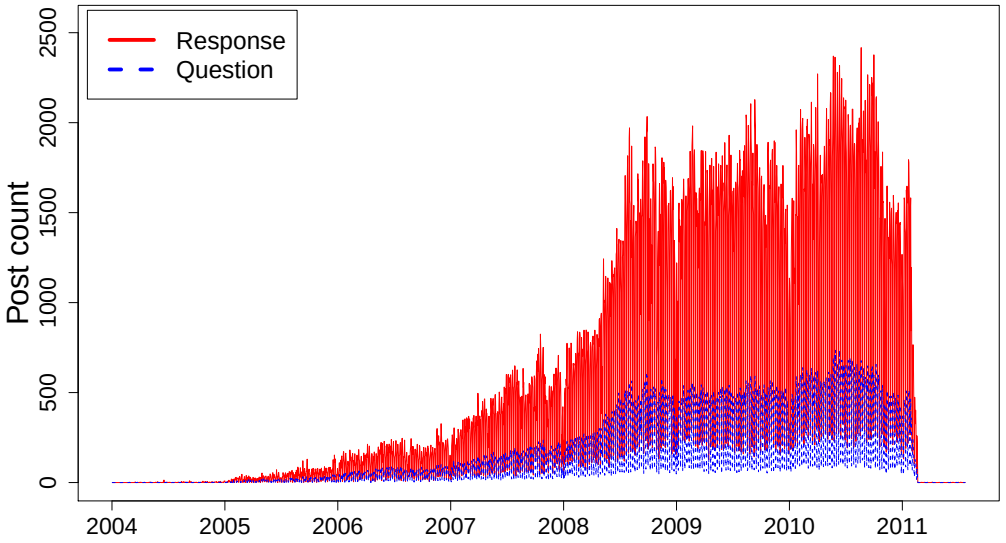


Figure 2.4: Number of question and response posts made per day in the SCN.

post, that is, *registered*.

Table 2.2: SCN annual summary statistics.

Year	Fora Active	Users Active	Threads with Response	Threads Solved	Solvers	Point Earners
2005	23	2840	89.40%	41.53%	7.78%	14.58%
2006	46	10449	86.24%	24.97%	7.02%	15.10%
2007	60	20666	89.74%	24.21%	7.47%	15.64%
2008	73	35351	91.25%	26.99%	9.86%	18.62%
2009	86	40407	92.63%	27.66%	10.74%	19.25%
2010	94	46044	89.33%	23.56%	7.56%	17.21%

For the sake of brevity we now consider a subset of the 95 fora to which we have access, the titles of the fora selected are given in Table 2.3. With regard to these fora, we detail various interesting characteristics in Table 2.4. Forum activity can be quantified with reference to users, posts or threads as totalled in the first 3 columns in Table 2.4 during the years 2008 to 2010. Figure 2.6 serves to show that a user's reputation gains are less noisy when considered on a weekly, rather than daily, basis.

From the 95 fora available, we select five to represent active fora with: very high activity; high activity; bursty activity; low activity; and very low activity. These fora are highlighted in grey in Tables 2.3 and 2.4 and carry the respective numerical identifiers: 142, 141, 50, 156 and 418.

The word "huge" is incorporated within the ROBUST project title. Within

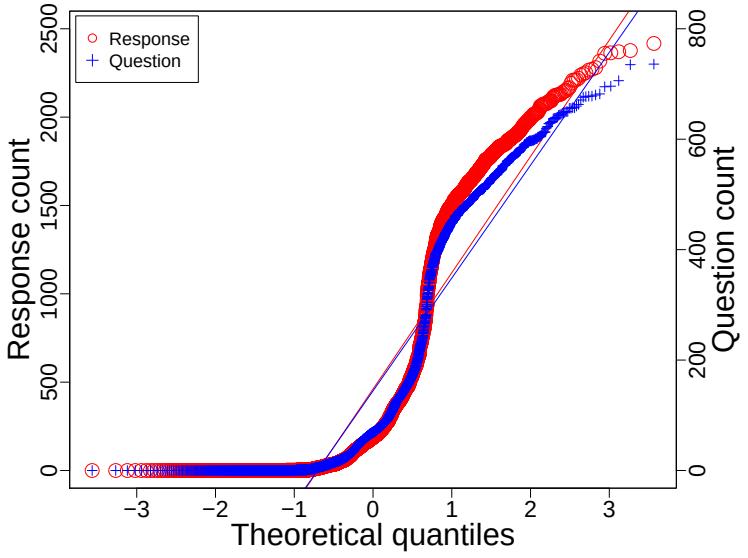


Figure 2.5: Quantile plot of number of questions and number of responses per day assuming underlying normal distribution.

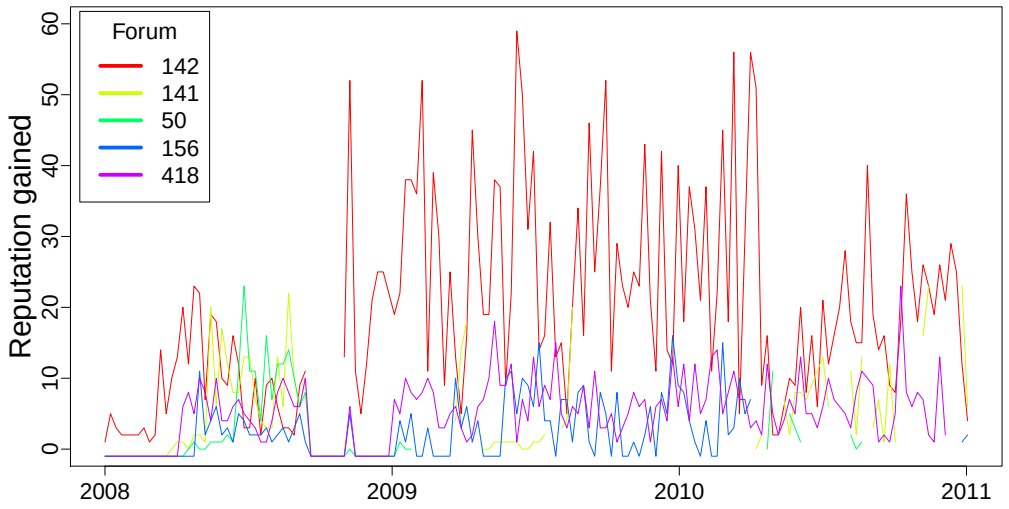
the initial deliverables, the use-case partners envisaged having *huge* data, however, the graphics and tables included within this section clearly indicate that we do not have *huge* data. For example, consider the heavy skew across fora implied by the mean number of threads per fora during 2008 and 2010 being more than four times the median (see Table 2.4). This indicates that, of the 95 available fora in the SCN, the majority have few questions asked over the three year period considered. Therefore, if we wanted to model individual threads it would be more realistic to have small rather than huge sample sizes. A similar observation holds for the number of users. The implication of this is that rather than attempting to model huge scale data we eventually realised that we would be more likely to be modelling a small sample. Thus, we are more likely to experience a lack of evidence rather than hit computational limitations.

Table 2.3: SCN forum community names.

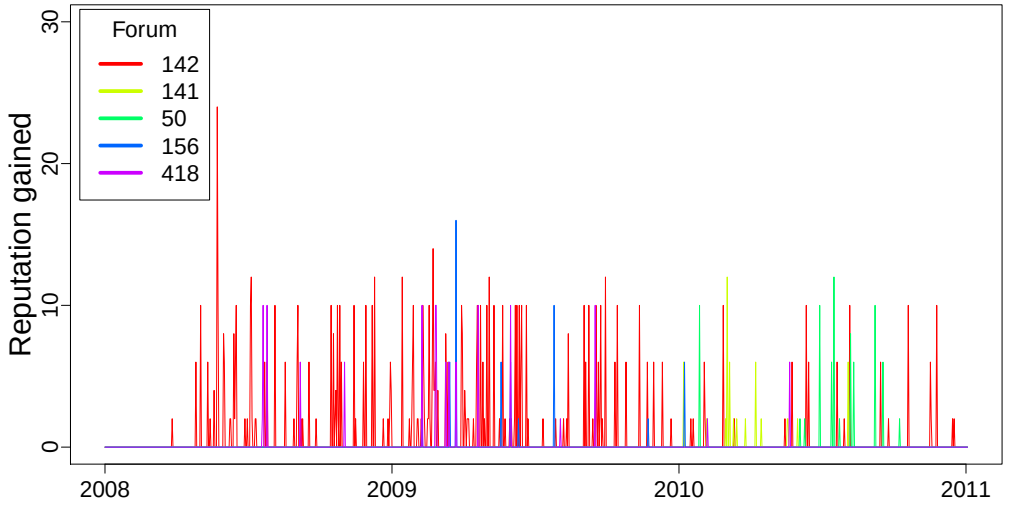
Forum ID	Forum title
142	ERP HCM (HR)
246	ERP - Sales and Distribution (SD) General
144	ERP Manufacturing (PP)
245	ERP - Logistics Materials Management (MM)
264	SAP Business One Core
323	ERP Financials - Controlling
141	ERP Financials
50	ABAP, General
143	SRM - General
156	SAP Solution Manager
56	SAP Business One SDK
159	Enterprise Asset Management (EAM)
284	Project System (PS)
324	ERP Financials - Asset Accounting
239	SCM APO Master Data and General
140	Enterprise Resource Planning (ERP)
145	Product Lifecycle Management (PLM)
405	ERP Operations - Quality Management (QM)
413	Business Planning and Consolidations (Microsoft Platform)
256	Governance, Risk and Compliance
418	SAP Business One - SAP Add-ons
244	CRM - Mobile Applications
399	ERP HCM Payroll North America

Table 2.4: SCN basic statistics by forum community. End rows summarise over all 95 SCN fora available.

Forum ID	Users	Posts	Threads	% Threads w/o response	% Threads w response solved	% Posts are replies	% Respondents with reputation	% Users post once
142	10041	153809	34842	11.32	19.61	77.70	19.58	31.40
246	10143	146702	33211	7.07	24.20	77.64	24.31	33.04
144	6596	144288	30095	2.91	30.53	79.40	26.94	28.35
245	10219	141437	32624	7.23	27.57	77.11	23.87	31.70
264	3604	78324	16418	1.25	45.32	79.17	26.64	27.61
323	6272	58376	15486	9.23	23.89	73.48	26.00	33.08
141	7045	54495	15088	14.50	17.32	72.73	19.05	38.55
50	8961	52286	13223	15.97	16.32	75.60	16.88	39.66
143	3933	50126	13085	11.27	20.34	75.13	23.27	34.40
156	5581	42973	11243	14.69	20.05	74.69	20.22	39.29
56	2492	42888	9815	3.61	40.18	77.98	23.66	25.52
159	2240	40467	7930	3.08	28.94	80.61	24.39	31.43
284	2842	39561	9141	6.47	21.38	77.03	24.39	32.09
324	3780	31017	8075	6.80	23.61	74.00	23.07	31.35
239	2762	30081	7669	6.48	37.34	75.30	31.35	33.31
140	5903	22587	7727	24.51	14.02	67.61	18.89	53.40
145	2324	20491	4927	9.58	24.83	76.75	24.77	37.35
405	1431	18886	3993	4.81	34.28	78.88	25.43	30.54
256	1510	17439	3854	8.30	22.95	78.20	23.55	34.90
418	1819	17204	3497	2.92	39.41	79.85	23.87	27.54
244	410	3667	923	7.91	25.06	75.67	26.59	37.07
399	484	3414	816	5.02	33.81	76.13	38.82	35.33
Median	676	3792	972	11.22	21.44	75.09	23.59	38.37
Mean	1721	17720	4195	14.14	23.27	72.80	23.29	42.19



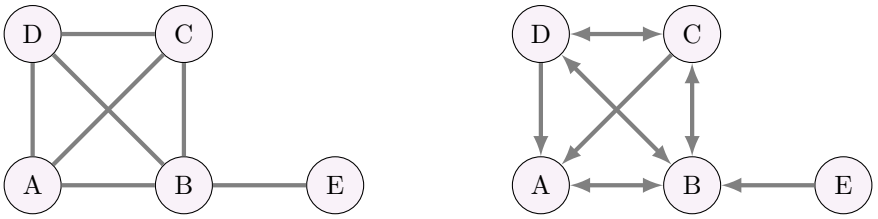
(a) Weekly activity



(b) Daily activity

Figure 2.6: Time series of user reputation gained for the most reputable user within each of five fora with differing activity levels between 2008 and 2010 inclusive.

Later, when formulating problem definitions for the SCN community, we shall make use of the fact that Q&A communities can be considered as social networks. This allows us to apply network theory to extract non-obvious features. For example, if we apply network theory to Figure 2.2, we can alternatively depict the social connections as shown in Figure 2.7. When performed forum-wide, we may derive features of a user's social connectedness. Such features are used later in the novel definition of user churn in online communities in Section 2.8.



(a) Undirected graph. (b) Directed graph.

Figure 2.7: Social network representation of Figure 2.2 with unweighted edges.

2.4 Online Community Health

Wasted resources can be costly and this is the motivation to ensure that, once an online community is established, it is successful, i.e. that it is healthy and remains healthy. Community managers or moderators are responsible for continually assessing community health in a cost-effective manner. Online communities can evolve extremely rapidly. Twitter, for example, generates hundreds of millions of posts per day with almost 300 million users active each month (https://about.twitter.com/company). As a result, real-time evaluation of community health metrics is necessary. Once some threat (risk) or potential benefit (opportunity) is predicted, it can be capitalised on at the earliest opportunity, giving access to the best outcome, in the most cost effective manner.

Within the literature on community health, one of the most popular models is the "Information System Success" (ISS) model, first published by DeLone and McLean (1992) and later revised (Delone and McLean, 2003). The ISS model was the first to make the many aspects of success in information systems clear, rather than adopting a narrow view. They identified six categories of information system success: system quality; information quality; user satisfaction; service quality; usage; and net benefits. As far as we are aware, all alternatives within the literature incorporate these six categories in some form (for example (Grover and Segars, 1996; Preece, 2000; Smithson and Hirschheim, 1998)).

Due to the widespread adoption of the ISS model by the research community, Petter and McLean (2009) were able to empirically assess the validity of the revised model. The majority of model aspects were found to be supported in a variety of information systems. Delone and McLean acknowledge that their model is a guiding framework and advise researchers to judge what aspects of it are in compliance with their own research objective(s). Our objective is to analyse the health of Q&A online communities, and hence the process of knowledge generation.

For users to ask questions, they must feel reasonably confident that knowledgeable responses will be provided; and for knowledgeable users to provide answers to questions, there must be some incentive (Koh et al., 2007). Essentially, for Q&A community health, it is important for users both seeking knowledge and providing knowledge to feel satisfied (Nam et al., 2009). Therefore, we focus our efforts on modelling the user satisfaction category of the ISS model.

2.5 User Satisfaction

Petter and McLean (2009) discovered, from various meta-analyses in the literature, that the relationship between user satisfaction and intention to use is one of the strongest in the updated ISS model framework. Additionally, many researchers observe that higher usage levels correlate to greater user performance (Iriberri and Leroy, 2009; Nam et al., 2009). Another reasonably strong relationship exists between information quality and user satisfaction, whilst service quality has a relatively insignificant relationship to user satisfaction.

The updated ISS model by Delone and McLean (2003) acknowledges the application of surveys as a way to measure user satisfaction. However, it is widely known that surveys are costly, potentially unreliable and commonly unrepresentative of the target population. Nonnecke and Preece (2000) find that one of the most common roles which a user takes in an online community is a *lurker*, i.e. someone who follows but does not contribute to content generation. Such users commonly revisit the same content area of the community and it can be argued that this is because they achieve some satisfaction from the community but, these users do not directly affect community health or success. Therefore, where such users exist and are included in a survey, their presence has the capacity to distort survey findings. Distortion may also occur because it is typical for people to either not complete surveys, or to complete them, but not in an entirely truthful manner. Additionally, surveys, by their nature, result in either textual answers (which are difficult to analyse) or answers on some discrete scale. However, the satisfaction of human beings is rarely one-dimensional. Consequently, online communities need an alternative way to monitor user satisfaction. This approach should be both representative and unobtrusive, where unobtrusive refers to modelling users from the meta-data their actions create.

Previous attempts to model user satisfaction from a user's trace of meta-data include role analysis (Section 2.6), questioner satisfaction (Section 2.7) and churn

analysis (Section 2.8). As stated in Section 2.1, the ROBUST consortium members from The Open University specialised in role analysis, but we briefly expose the corresponding literature which is most relevant to modelling user satisfaction for completeness in Section 2.6. For questioner satisfaction and churn analysis, we discuss the related literature and identified deficiencies within (that correspond to SAPs concerns), prior to defining our novel interpretations of user satisfaction aimed at addressing these in Sections 2.7 and 2.8 respectively.

2.6 Role Analysis For Online Community Health

Modelling user satisfaction via role composition is motivated by suggestions by Preece (2000) that, in communities where there is a dominant behavioural type, or role, users are more prone to churn. Work by Rowe et al. (2013) concludes that changing community role composition implies potential change in the type or focus of the community rather than ill health. This is supported by the earlier work of Fisher et al. (2006), who observed, through the examination of social community network features, that Q&A communities differ greatly in role composition to those of social support, discussion or flame (brainstorming) communities. A large amount of literature exists about the role composition of communities which is not limited to online communities. Some of these studies focus on all roles available (Angeletou et al., 2011; McWilliam, 2000; Nonnecke et al., 2006; Rowe et al., 2013); whilst others consider the impact of individual roles such as "newbies" (Joyce and Kraut, 2006), experts (Pal et al., 2011), or lurkers (Nonnecke and Preece, 2000; Preece et al., 2004; Tagarelli and Interdonato, 2013). However, so far these studies are limited to either forum-level or community-level events.

Within a community, roles are determined by a set of observed behaviours that are described by a set of features with size p. Typically, each of the features are binned into k categories (e.g. small, medium and large) (Angeletou et al., 2011; Hautz et al., 2010; Karnstedt et al., 2011; Rowe et al., 2013), leading to a loss of information. Each role is then comprised of one of the k^p unique combinations of feature binnings (for fixed k). As p becomes large for $k > 1$, the number of possible roles tends to infinity. In practice, not all of the unique combinations are considered as roles. A natural result of this is that not all users can be assigned a role, an issue observed in Angeletou et al. (2011) and Rowe et al. (2013) where 21% and 7% of users respectively are consequently dropped from further analysis.

Whilst role composition has been used to analyse communities at the forum-

level, there is little in the literature thus far to suggest that this approach is suitable for user-level or thread-level analysis. Both Zhu et al. (2011) and Angeletou et al. (2011) model the behaviour of individuals in an attempt to assign each a role. However, both use individual role composition to make inferences at the forum-level. Interestingly, both see the future of role analysis to lie in modelling individual user-level churn but, the issue remains that some individuals are not assigned a role. Rowe et al. (2013) improves on Angeletou et al. (2011), leaving fewer individuals without an assigned role. The goal of modelling individual churn remains as future work, although the aims of role analysis are extended to include the inference of churn likelihood from the trace of a user's role development. Role analysis does not currently seem to be suited to modelling user satisfaction, but may be an avenue for future developments.

2.7 Questioner Satisfaction

We now consider modelling user satisfaction via metrics motivated by questioner satisfaction. Initially, we give an overview of the related work in the literature, we then expand this to give our own unique formulation and finally we describe the features extracted from the community meta-data of the SCN.

2.7.1 Related work

Since the widespread recognition of the value of Q&A online communities, many researchers have attempted to provide suitable response to the fundamental task of predicting whether a questioner will be satisfied by the set of responses they receive. Researchers hail from a diverse range of backgrounds including social science, computer science, statistics and machine learning.

The vast participation in various topic areas in Q&A online communities has led to the generation of large knowledge repositories. Given the format of community question-and-answering, there is an inevitable time lag between the posting of the question and any responses made, which may or may not solve the original question. If an answer to a question is found to be similar to a previously solved question in the community repositories, this time lag can be averted. As a consequence some researchers in the literature focus more on whether an answer for the questioner's query already exists either in the community or the wider web (Hao and Agichtein, 2012; Jeon et al., 2005; Liu et al., 2011; Xue et al., 2008).

This research is closely connected to the work of moderators in ensuring that no duplicate questions exist.

The first large-scale study of questioner satisfaction was published by Liu et al. (2008) with later significant development by Agichtein et al. (2009b); Anderson et al. (2012); Keegan and Gergle (2010); Maxwell Harper et al. (2008); Tausczik and Pennebaker (2011). This first work modelled questioner satisfaction as a binary response event: either the questioner was satisfied by the set of responses or he was not. Classification algorithms applied include Decision trees, Support Vector Machines (SVM), Boosting and Naive Bayes as provided by the Weka framework Witten et al. (2011). Ideas in (Liu et al., 2008) are extended in (Agichtein et al., 2009b). This latter work demonstrates the feasibility of predicting questioner satisfaction as a binary classification task, using features of question content and structure as well as more macro-level community features. Agichtein et al. (2009a) further discuss the process of extracting features for the analysis of questioner satisfaction. Within this discussion, PageRank (see Appendix A.2) was identified as being useful for predicting questioner satisfaction. Subsequently it became increasingly common for network features to be incorporated in modelling user satisfaction, for example see Anderson et al. (2012).

Others consider more practically how user reputation alone contributes to answer quality and, implicitly, questioner satisfaction (Keegan and Gergle, 2010; Maxwell Harper et al., 2008; Tausczik and Pennebaker, 2011). Opinions are split in the literature as to whether it is best to maintain a large set of expert users or whether to encourage a wide range of users for the health of a Q&A community. Tausczik and Pennebaker (2011) attempt to address this split, asking whether users with high reputation should be farmed to create a set of experts or whether the importance of user reputation should be minimized so that users who are not well-established within the OC or question area are not discouraged from contributing. A prior study (Maxwell Harper et al., 2008), which compared a selection of Q&A online communities, indicates that the latter opinion holds more merit. This study found the communities whose users exhibited more varying levels of experience produced higher quality answers. Even so, others in the literature (for example (Keegan and Gergle, 2010)) have observed online communities where this is not the case. Tausczik and Pennebaker (2011) support (Maxwell Harper et al., 2008) by observing that users with a lower median reputation score would typically submit questions, whilst those with a higher median reputation score would submit answers.

Anderson et al. (2012) were one of the first in the literature to consider user satisfaction by viewing the whole user experience of the Q&A platform as being of a knowledge creation process. Rather than predicting questioner satisfaction via information extracted from question-response pairs (as done by Jeon et al. (2006)), they extract features from questions and their entire set of corresponding responses. Two events are considered: long-lasting value of the message tree and sufficient generation of knowledge within the message tree (Anderson et al., 2012). Our first novel formulation of user satisfaction extends the latter of these events to predict whether sufficient knowledge is generated *within some specified time* of the questioner seeking an answer. This risk event addresses both the issue of an acceptable level of knowledge being provided and the issue of knowledge provided prior to some unacceptable time lag.

2.7.2 Novel questioner satisfaction problem formulation

We observe from Table 2.2 that approximately 25% of threads created annually within the dataset were solved. Whilst we do not know if 25% is low or high with respect to other Q&A online communities, the SAP representative on the ROBUST project stated that this was a concerning level and stated a desire to model questioner satisfaction. We chose to formulate the problem of questioner satisfaction as a classification event which may be viewed either as a risk or an opportunity. That is, after a time threshold, t_s minutes, of creation, the thread is *solved* (opportunity) or *unsolved* (risk). Given the literature of which we are aware, this risk event is most similar to the Anderson et al. (2012) event which can be expressed as: sufficient knowledge is generated such that the thread is solved. The predominant difference, that is the novelty, in our event is that we consider an additional dimension by enforcing a time threshold on the thread being solved.

We now introduce some typical classification notation in the process of providing a more rigorous definition of our risk event. Let the categorical response G be in the set $\mathcal{G} = \{unsolved, solved\}$ where $\mathcal{G}$ is the complete set of possible classifications for an observation. Take K to be the cardinality of $\mathcal{G}$ such that the number of classes for our event is $K = 2$. By taking Y to be the discrete-coded response of G and given that $K = 2$, Y is binary and each response can be recorded as either 0 or 1. We choose $Y = 1$ ($Y = 0$) to correspond to the category *solved* (*unsolved*). Assuming the i^{th} thread to be eventually solved, we note the wall clock time (minutes since thread creation) of this event as w_i. The default value

of w_i is ∞. Thus the observed binary response, y_i, for the i^{th} thread, is

$$y_i = \begin{cases} 1 & \text{if } w_i \leq t_s, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

where $i = 1, \ldots, N$ and N is the number of threads observed within the sample population. What follows in the remainder of this section is entirely based upon our understanding of the SCN.

For the threads with at least one response in the fora of the SCN, approximately 60% (on average across these fora) have points awarded, and almost 25% (on average) are marked as "answered" by the original poster. We acknowledge the awarding of points by the questioner (to respondents' messages) to be a subjective act that requires the questioner to respect the underlying process in Q&A communities (where the answerer is awarded points for his knowledge as encouragement to continue). However, given that a Q&A community such as the SCN works on a peer-to-peer level, we find this to be an appropriate measure of the value of knowledge given.

Considering only those threads created at least one year prior to our last observation and having at least one respondent, 13.76% have not been responded to within the first 24 hours and 0.56% are only responded to after more than year since thread creation. This highlights that the vast majority of threads receive greatest attention within the first 24 hours after creation. Of the threads responded to within the first 24 hours, 28.56% are subsequently solved and of these 74.57% are solved within the first 24 hours. Within the threads solved in the first 24 hours, 62.64% are solved by the first response. By comparison, 15.18% of the threads responded to only after the initial 24 hour period are eventually solved. However, within these latter solved threads, 64.77% are solved by the first response. Thus, although a thread seems less likely to be solved if not responded to within the first 24 hours of creation, it is still most likely to be solved by the first respondent.

Figure 2.8 illustrates the percentage of threads solved across all fora, grouped by year, over hours since thread creation. In all cases, the curve incline begins to notably reduce six hours after thread creation and starts to level off 24 hours after thread creation. Therefore, we see a suitable choice for t_s in (2.1) to be 1440 minutes (24 hours).

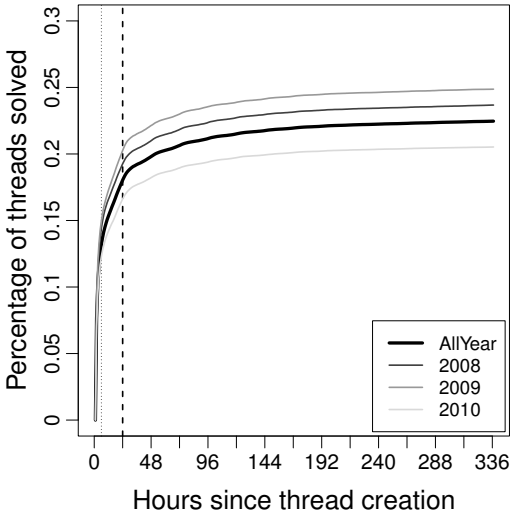


Figure 2.8: Percentage of threads solved, by hours since thread creation, for all threads within dataset created between 2008 and 2010.

2.7.3 Novel feature set to predict questioner satisfaction

The full set of features available for prediction following t minutes since thread creation is given below. The choice of t affects feature inclusion. Were the time of prediction (t) to be close to the time of event (t_s), one could argue that prediction is made too closely to the occurrence of the event to be of interest. In addition, for t close to zero (thread creation) there exist uninformative features where all observations hold the same value. We trialled $t \in \{20, 30, 60, 180, 360\}$ minutes; here we report only for $t = 20$ minutes due to lack of improvement in classification performance measures for larger t. There are a total of 44 features, each belonging to one *type*, which we use to model y_i in (2.1) given t.

Given the similarity between our risk event and that of Anderson et al. (2012), we analysed the corresponding features in the SCN where they existed, and expanded upon these. Those features which are present in (Anderson et al., 2012) are underlined. We cannot be certain of the level of similarity between our interpretation of the underlined features and that of Anderson et al. (2012) given that Anderson et al. (2012) did not publish full descriptions of the features used. We are not aware of any of the 30 features which are not underlined having been used previously towards modelling user satisfaction in online communities. Indeed, we have not come across the concept of structuring features about the most responded to message or user anywhere in the literature.

Let the vertices of a directed graph represent the users of an online community,

and take the directed edge from user v_i to user v_j to be indicative that user v_i posted at least one message in direct response to user v_j during the time frame considered. For illustrative example see Figure 2.7b. The directed edge between two users may be weighted by the number of unique direct responses.

In what follows, let p_x represent the post identified as x. Given that a user is attributed a title (e.g. OP) in accordance with a single post, we let $\mathcal{P}_{title}$ denote the set of posts made by any user marked with the designated title within the relevant forum. Likewise, we let $\mathcal{P}^*_{title}$ denote the set of reputable posts made by any user marked by the designated title within the relevant forum. In addition, we let $\mathcal{H}$ represent the set of posts within the thread up to t minutes after thread creation; and take $\mathcal{H}_{title}$ as the subset of posts made by any user who is marked by the designated title. We also take $r(p_x, p_y, t_{xy})$ to represent the post p_x being directly responded to by p_y where the lag between posts was t_{xy} decimal minutes.

OP features

- **OP reputation:** is a measure of how active the OP has been in terms of providing peer-perceived valuable knowledge throughout their lifetime in the thread relevant forum, prior to creating the thread. Take $|p_x|$ to denote the reputation awarded to the post p_x, where $p_x \in \mathcal{P}^*_{op}$ the reputation earned by the OP is
$$\sum_{\mathcal{P}^*_{op}} |p_x|.$$

- **OP reputation in past year:** is a measure of how active the OP has been in terms of providing peer-perceived valuable knowledge in the thread relevant forum, during the year prior to creating the thread. Where $\mathcal{P}^*_{op}(-1)$ is the subset of $\mathcal{P}^*_{op}$ containing posts made within a year of the thread being created and $p_x \in \mathcal{P}^*_{op}(-1)$, the reputation earned by the OP is
$$\sum_{\mathcal{P}^*_{op}(-1)} |p_x|.$$

- **# thread OP participated:** the count of distinct threads within the relevant forum in which the OP has made a post, excluding the currently referenced thread.

- **# thread OP created:** the count of distinct threads within the relevant

forum which the OP has made the creating post, excluding the currently referenced thread. This is consequently the number of times the OP has been the OP of a thread in the past within the relevant forum.

- **# thread OP solved:** the count of distinct threads within the relevant forum in which the OP has made the post which was awarded 10 points by that threads OP.
- **underline# messages OP posted:** the count of distinct posts made by the OP within the relevant forum prior to creating the current thread.
- **# messages OP posted in thread:** the count of distinct posts made by the OP within the current thread in the t minutes following creation, this includes the original post and as such has a minimum value of 1.
- **# days since OP registration (first appearance in relevant forum):** the number of full twenty-four hour periods which have elapsed since the OP first posted in the relevant forum.

MRR features

- **MRR reputation:** is a measure of how active the MRR has been in terms of providing peer-perceived valuable knowledge throughout their lifetime in the thread relevant forum, prior to the creation of the considered thread. Where $p_x \in \mathcal{P}^*_{mrr}$, the reputation earned by the MRR is

$$\sum_{\mathcal{P}^*_{mrr}} |p_x|.$$

- **MRR reputation in past year:** is a measure of how active the MRR has been in terms of providing peer-perceived valuable knowledge in the thread relevant forum, during the year prior to the creation of the considered thread. Where $\mathcal{P}^*_{mrr}(-1)$ is the subset of $\mathcal{P}^*_{mrr}$ containing posts made within a year of the thread being created and $p_x \in \mathcal{P}^*_{mrr}(-1)$, the reputation earned by the MRR is

$$\sum_{\mathcal{P}^*_{mrr}(-1)} |p_x|.$$

- **# messages MRR posted in thread:** the count of distinct posts made by the MRR within the considered thread in the t minutes following creation.

HPS features

- **# HPSs:** the count of the number of unique posts within the considered thread which have the highest point award. Let this be denoted by n. Due to the restrictions SAP place on how points are awarded within the SCN (Table 2.1), where 10 points have been awarded, there is one HPS and they are the TS; where 6 points is the highest awarded, there are either one or two HPS users; where 2 points is the highest individual award, there may be many HPS users; and where no points have been awarded, the HPS concept is void and this feature takes a default value of -1.

- **HPS reputation:** the forum consistent life-time reputation up to posting the message which was awarded the highest point score. This sum is averaged where there are multiple users marked as the HPS in the considered thread in the following manner

$$\frac{1}{n} \sum_{p_x \in \mathcal{P}^*_{hps}} |p_x|.$$

- **# responses to HPS:** the count of the number of messages posted in direct response to the user(s) whose post(s) was(were) awarded the highest point score by the OP. This count is averaged where there are multiple users marked as the HPS in the considered thread in the following manner

$$\frac{1}{n} \left| \{p_y | p_x \in \mathcal{H}_{hps}\} \right|.$$

MRTM features

- **# MRTMs:** the count of the number of unique messages within the considered thread which have the highest number of posts made in direct response. Let this be denoted by n. Where there are only two posts in the considered thread, the first is the original post and the second must be a direct response to the original post. In this scenario the original post is the MRTM. If no posts have been made other than the original post, the MRTM concept is void and this feature takes a default value of -1.

- **MRTM reputation:** the forum consistent life-time reputation of the user(s) up to posting the message which became the most responded to message. This sum is averaged where there are multiple messages with the highest

number of direct responses in the following manner

$$\frac{1}{n} \sum_{p_x \in \mathcal{P}^*_{mrtm}} |p_x|.$$

- **# responses to MRTM:** the count of the most direct responses to a single post.

- **MRTM points earned:** this conveys the worth of the most responded to message to the original poster and can indicate the message type. When $n > 1$ we let $|p_{mrtm}|$ denote the reputation awarded to a message with the greatest number of direct responses and average as follows

$$\frac{1}{n} \sum |p_{mrtm}|.$$

MRTU features

- **# responses to MRTU:** the count of the number of direct responses to the user(s) who are the most responded to within the thread considered. Let this value be assigned to s.

- **# MRTUs:** the count of the number of unique users who have received s direct responses in total to their posts in the thread. Let this be denoted by n. Where there are only two posts in the thread, the first is made by the OP and as the second post must be a direct response to the original post, the OP must be the MRTU and $s = 1$. If no posts have been made other than the original post, the MRTU concept is void and this feature takes default value $s = -1$.

- **MRTU reputation:** the forum consistent life-time reputation of the most responded to user(s) up to making their first post in the thread. This sum is averaged where there are multiple users who received s direct responses in total within the thread in the following manner

$$\frac{1}{n} \sum_{p_x \in \mathcal{P}^*_{mrtu}} |p_x|.$$

- **MRTU points earned:** this conveys the level of knowledge generated by the most responded to user. When $n > 1$ we let $|p_{mrtu}|$ denote the reputation

awarded to a post made by a user who has s total direct responses and average as follows

$$\frac{1}{n}\sum_{mrtu}|p_{mrtu}|.$$

- **# thread MRTU created:** the count of the number of threads which a user marked as having the most responses has created. Let $\mathcal{I}_{mrtu}$ be the set of threads initialised by a most responded to user in the relevant forum prior to their first post in the considered thread. Then the number of threads initialised by such a user is

$$\frac{1}{n}\sum_{mrtu}|\mathcal{I}_{mrtu}|.$$

- **# messages MRTU posted:** the count of the number of messages which the user marked as having the most responses has posted. This can be portrayed as

$$\frac{1}{n}\sum_{mrtu}|\mathcal{P}_{mrtu}|.$$

Temporal features

- **minutes until first reply:** the difference in minutes between the original post and the first subsequent post.

- **mean minutes until respondents first message:** the mean average difference in minutes between the original post and a respondent's first post in the considered thread across all respondents who post in the considered thread prior to t minutes after the original post.

- **mean minutes between messages:** the mean average difference in decimal minutes between posts in the considered thread. Where the original post is the only message in the considered thread, this takes a default value of -1. Let $r(p_x, p_y)$ signify that post p_y was consecutive to post p_x so that the mean time between postings within the thread is

$$\frac{1}{|\mathcal{H}|}\sum_{r(p_x,p_y)\in\mathcal{H}} t_{xy}.$$

- **median minutes between messages:** the median average difference in

minutes between posts in the considered thread. Where the original post is the only message in the considered thread this takes a default value of -1.

- **minimum minutes between messages:** the minimum difference in minutes between posts in the considered thread. Where the original post is the only message in the considered thread this takes a default value of -1.
- **MRR time rank:** the rank of the post made by the most reputable respondent in the thread considered.
- **MRR wall clock:** the decimal minutes passed from thread creation to the first post of the most reputable respondent.
- **mean HPS time rank:** the mean rank of the post made by the highest point scorer(s).
- **mean HPS wall clock:** the mean decimal minutes passed from thread creation to the post(s) awarded the highest points by the original poster.
- **minimum MRTM time rank:** the posting rank of the most responded to message in the thread considered. Where there are multiple messages with the maximum number of direct responses the rank of the first is taken.
- **minimum MRTM wall clock:** the decimal minutes elapsed from thread creation to the posting of the most responded to message. Where there are multiple messages with the maximum number of direct responses the minimum is taken.
- **minimum MRTU time rank:** the posting rank of the first message of the most responded to user in the thread considered. Where there are multiple such users the measure of the earliest to participate is taken.
- **minimum MRTU wall clock:** the decimal minutes elapsed from thread creation to the first post of the most responded to user within the considered thread. Where there are multiple such users the measure of the earliest to participate is taken.

Thread summary features

- **indicator of thread status:** a binary value indicating whether the status of the considered thread has been changed from the default of "unanswered" to "answered".

- **# users to participate:** the count of unique users who post messages in the considered thread within the first t minutes since thread creation. Where the original post is the only message in the thread this count takes value 1 to represent the participation of the OP.

- **sum points awarded:** the sum of the point scores awarded by the OP in the considered thread within the first t minutes since thread creation. Where the original post is the only message in the thread, this feature takes the default value of -1; whereas if an additional post is made but no points are awarded, the value 0 is taken.

- **# messages posted:** the count of the number of messages posted in the considered thread within t minutes of the original post. This count includes the original post.

- **mean respondent reputation:** the mean forum-specific lifetime reputation of any respondent up to the time of their first participation in the considered thread. Where there are no respondents in the considered thread the value -1 is taken.

- **median respondent reputation:** the median forum-specific lifetime reputation of any respondent up to the time of their first participation in the considered thread. Where there are no respondents in the considered thread the value -1 is taken.

- **mean respondent reputation in past year:** the mean forum-specific reputation of any respondent within the year period leading up to the timestamp of their first participation in the considered thread. Where there are no respondents in the considered thread the value -1 is taken.

All the features listed above are used to model the novel questioner satisfaction risk event outlined earlier in this section. The features are used in *sets* according to *type* both individually and additively. We conduct an analysis on whether the features of one type provide greater classification accuracy of questioner satisfaction than another through comparing classifier performance across individual feature sets. In addition we consider the degree to which there is an additive effect across feature sets. An example of the individual feature case is: if those features about the OP are more informative than those about the MRR on whether a questioner is satisfied within 20 minutes. This may be demonstrated by comparing classifier

performance measures. When investigating the additive effect of feature sets, this may demonstrate whether there is information contained in the features about the MRR which is not available in those features focused on the OP. The corresponding results of this classification analysis are discussed in Section 6.4.

2.8 Churn Analysis

We now consider modelling user satisfaction as churn analysis. Prior to giving our unique problem formulation of churn in Q&A online communities, we review the related literature. There is very little which addresses the concept of churn for online communities and this highlights the vital need to consider new alternative definitions of the churn event for the users of online communities. Subsequent to this, we describe those features extracted from the meta-data of the SCN.

2.8.1 Related work

We consider two disparate aspects of churn analysis in the literature: the concept of churn and the methods used to perform churn analysis. The first of these outlines the literature that conceptualises the churn event at a customer, or user, level across different industries. Given that churn is typically modelled as a binary event, the second aspect compiles analysis from the literature on those classifiers previously used to perform churn analysis.

The concept of churn

The majority of literature on churn analysis is associated with the telecommunication industry (Ngonmang et al., 2012), for example: (Fawcett and Provost, 1997; Jadhav and Pawar, 2011; Lemmens and Croux, 2006; Provost and Fawcett, 2001; Richter et al., 2010; Zhu et al., 2011). To this day, churn is most prominently discussed and analysed with respect to some monetary service, where it is typically understood as total customer defect (i.e. a customer "defecting" to another service provider). Churn analysis is critical in this industry, given the comparatively cheap cost of customer retention versus customer acquisition (Hadden et al., 2007). Nonetheless, customer retention efforts are still costly, meaning that the accuracy of churn predictions is very important.

The users of online communities are generally not charged a subscription fee by the platform provider and as such there is no financial motivation for users to

notify the provider that they no longer wish to be registered as a user. Therefore, the concept of churn as total customer defect does not apply to the users of online communities. As a consequence, we require an alternative definition of the churn event with respect to online communities that will account for the required difference in the behaviours of the respective industries.

With the recognition of the value of Q&A online communities, telecommunication companies are moving away from providing telephone helplines and towards incorporating dedicated support communities. Consequently, some researchers who previously focused on analysing churn in the telecommunication industry have moved to perform similar analysis for Q&A online communities (Rowe et al., 2013). The telecommunication industry has benefited from the growth in online community research, harvesting ideas from social network analysis to model churn (Dasgupta et al., 2008; Phadke et al., 2013; Richter et al., 2010). Both Dasgupta et al. (2008) and Richter et al. (2010) are widely accredited with pioneering the use of social network analysis in predicting churn in the telecommunications industry. A by-product of the work by Richter et al. (2010), which analyses group rather than individual behaviour, is that customers who have the greatest influence on their peers can be pinpointed.

One of the first Q&A online communities, Yahoo Answers, was launched in 2004. Karnstedt et al. (2010a) remark that no research previously existed on the meaning and consequence of churn in online communities. Nonetheless, we find evidence of churn consideration in online communities in the observations of Jones et al. (2004), who observed that the less active users in online newsgroups were more likely to *decrease in activity* than those who were more active. Similar observations were made by Stutzbach and Rejaie (2006) when investigating why churn was so poorly understood in Peer-to-Peer (P2P) systems at the time. One of their most interesting findings is the stability of active peers whilst those less active appear less stable. If we conceive churn as a concerning level of decrease in user activity then this finding can be interpreted as: less active users in P2P systems appear more likely to churn. These research articles are published in social network, management science and internet structure research fields, which suggests a historical duplication of work. Churn has been studied in such a variety of fields that it is common to find duplicated material within the literature.

At the time when social network analysis began to be used as a solution to marketing problems, Doyle (2007) marked churn to be a key output. Prior to this explicit mention of churn in a marketing scenario, Baesens et al. (2004) used

classifiers to model the sign of the slope of customer spending in a market place. Again, if we think of churn as a decline in activity (but not necessarily to the point of leaving) then the work of Baesens et al. (2004) can be referred to as churn analysis.

The first formal proposal of an alternative definition of the churn event for the users of online communities was given and assessed by Karnstedt et al. (2010a). As far as we are aware, this is the earliest work in the literature that explicitly conceptualises churn with respect to a measure of decreased user activity over time. We give this definition in Definition 2.1 for completeness.

Definition 2.1.

"*The previous activity (PA) window is a time window consisting of time steps* t_1 *to* $t_1 + n - 1$ *inclusive,* $n \in \mathbb{N}, n \geq 0$. *Let* $\mu_{PA}(v_i)$ *denote the average activity of a user* v_i *over the previous activity window. The churn (C) window is a time window* $t_2 = t_1 + n$ *to* $t_2 + m - 1$ *inclusive,* $m \in \mathbb{N}, m \geq 0$. *Let* $\mu_C(v_i)$ *denote the average activity of a user* v_i *over the churn window. A user* v_i *is considered to have churned during the churn window if:*

$$\mu_C(v_i) \leq T(S) \cdot \mu_{PA}(v_i)$$

$0 \leq T(S) < 1$ *is a threshold factor dependent on the relevant system parameters* S" (Karnstedt et al., 2010a, p. 191).

Based upon this definition of churn specific to online communities, Karnstedt et al. (2010a) derive the corresponding probability estimate of the individual churn event as follows

$$\mathbb{P}(churn|v_i) = \begin{cases} 0 & \text{if } \mu_C(v_i) \geq \mu_{PA}(v_i) \\ 1 - \left(\frac{\mu_C(v_i)}{\mu_{PA}(v_i)}\right) & \text{otherwise} \end{cases}$$

The authors use two differently structured Q&A online communities to demonstrate that Definition 2.1 is meaningful when user activity is measured by the number of posts made within the time windows specified. From these case studies, Karnstedt et al. (2010a) show the existence of a relationship between network features, influence and churn (given Definition 2.1) at the user-level. This suggests the need to focus on structural relations and features for the prediction of churn in online communities (Karnstedt et al., 2010b). The main result of (Karn-

stedt et al., 2010a) was the discovery of correlation between individual churn (see Definition 2.1) predictions and forum-level churn events. Not all researchers of churn analysis in online communities use the updated definition of churn given in Definition 2.1. For example, Ngonmang et al. (2012) choose to use the original definition from the telecommunications industry: a user churns only if he becomes completely inactive. Given the previously observed duplication of churn analysis material within the literature across different research fields, it is possible that this is due to researchers not being aware of research from other areas. However, it is also possible that churn as given in Definition 2.1 is not widely used given that it contains inconsistencies. These inconsistencies are discussed in Section 2.8.2 in accordance with our suggested alterations, and it is these that led us to propose Definition 2.2 as an alternative.

Classifiers for churn analysis

Classification methods used to model churn in the literature include (but are not limited to): logistic regression (Rowe et al., 2013); decision trees (Richter et al., 2010); bagging and boosting (Burez and Van den Poel, 2009; Lemmens and Croux, 2006); random forests (Burez and Van den Poel, 2009). To analyse classifiers inherently requires some classification quality characteristic measure and the Area Under (ROC) Curve (AUC) is the most widely used (Hand, 2009). However, as with the classifiers themselves, these can be susceptible to class imbalance.

A fundamental characteristic of churn is that it is a rare event. In all datasets, whether taken from the telecommunications industry or some online community, class imbalance is present. Some researchers choose to balance the classes, for example (Burez and Van den Poel, 2009; Lemmens and Croux, 2006), whilst others accept the imbalance (Karnstedt et al., 2011). It is suggested by Burez and Van den Poel (2009) that simply choosing an appropriate performance measure can reduce the significance of class imbalance.

The principle author of (Baesens et al., 2004), in which churn was considered in a manner suitable for the users of online communities, explicitly modelled churn for the marketing industry in (Glady et al., 2009). In this later work, standard accuracy-based classification performance measures are declared to be unsuitable for modelling churn in a business scenario and alternative profit-focused measures are deemed to be desirable. Even with this new perspective, the methods of logistic regression, decision trees, neural networks, AdaCost and cost-sensitive tree are not found to be significantly different.

Lemmens and Croux (2006) use churn in a US telecommunications company as an application in the marketing sector to compare whether bagging and stochastic gradient boosting classification trees outperform logistic regression (one of the most popular classification models). They advocate balanced sampling followed by bias correction, given that churn is a rare event. However, they do not provide evidence to support this. The Gini coefficient and lift metric are the given measures of classifier performance. These are both one-dimensional performance measures and as such are relatively uninformative compared to other measures available (see Chapter 5). Some small difference in performance is observed between the logit model and the classification tree model but this is inconclusive given the measures used. Certainly the logit is the preferable model when performing real-time analysis, considering the relative computation involved.

Hadden et al. (2007) provides a review of the popular advancements in creating a platform for managing customer churn. Features that reflect frequency, recency or monetary behaviours of the users are found to be informative for churn prediction. With respect to the best classification model for predicting churn, the findings of Hadden et al. (2007) imply that there is no significant advantage between any of the following methods regarding accuracy-based performance: binomial generalised linear models; neural networks; decision trees; k-nearest neighbour; Bayesian network classifiers; genetic algorithms. To the best of our knowledge of the current literature, Zhao et al. (2005) were the first to apply support vector machines to churn analysis for telecommunications - but this article is not referenced by Hadden et al. (2007). To our knowledge, the concept of fuzzy logic has not yet been used to model churn. Given the fuzziness of the problem at hand it seems a natural development and such thoughts are echoed by Hadden et al. (2007).

All who compare classification models for churn analysis within the literature, find little evidence to suggest that decision trees or neural networks are significantly superior to binomial generalised linear models - with or without boosting being applied (for example (Hadden et al., 2007; Mozer et al., 2000)). In essence, everyone is trying to solve the same problem - how to predict some loosely defined behaviour for some interesting and reasonably large dataset. Given that our aim is to integrate a real-time risk analysis service into the risk management platform of Nasser et al. (2013a), we do not consider the more computationally expensive decisions trees, neural networks and so forth.

2.8.2 Novel user churn problem formulation

We now propose our formulation of user churn, with specific regard to online communities, following our analysis of the SCN community and observations from the related work. Firstly, we explicitly detail the inconsistencies within Definition 2.1 and address these to provide a more usable definition of the user churn event in our novel Definition 2.2. Secondly, we detail our corresponding experimental set-up for using this novel interpretation of churn in practice and extended this to consider which users we model with churn analysis. Within this we highlight the shortfalls of the previous applications of Definition 2.1. Thirdly, we discuss the relationship between this and our previous novel formulation of user satisfaction (see Section 2.7.2). The next section (Section 2.8.3) provides the set of features which we use as the independent variables when predicting the user churn event for a population of users via the classifiers given in Chapters 3 and 4.

Our response to our perceived inconsistencies in Definition 2.1

As argued during our exposition of the related work, churn has not yet been appropriately defined with regard to online communities as far as we are aware. Karnstedt et al. (2010a) has come closest to providing a suitable definition of churn for this field but we are aware of inconsistencies within their formal definition (Definition 2.1). The first and most obvious inconsistency concerns the definitions of the time windows: for $n = 0$ the previous activity window includes the time windows t_1 *to* $t_1 - 1$; and with $n, m = 0$ the churn activity window includes time windows $t_2 = t_1$ *to* $t_2 - 1 = t_1 - 1$.

A second inconsistency is that churn, as given in Definition 2.1, is unable to represent the disparity between a regularly contributing user whose activity decreases (relative to their previous activity) and the user who contributes irregularly. This can be seen as an inconsistency between the formulation of the churn event and the reality of the problem. In comparison with the inconsistency of defining the time windows, this inconsistency is significantly more difficult to address in a meaningful way. We include our initial attempt to address this inconsistency in Appendix F.1 which directly targets the obvious issue: that two users v_k and v_ℓ may have the same probability of churning regardless of v_k being significantly more active than v_ℓ in the current time window. However, due to time constraints, we only include some preliminary results in Appendix F and raise this as being an issue in the current understanding and corresponding defi-

nition of churn in online communities which requires more detailed consideration in the future. Nonetheless, the initial inconsistency regarding the specification of the activity time windows is more easily addressed and consequently we propose Definition 2.2 as an alternative to Definition 2.1.

Definition 2.2.

Let $\mathbb{N}$ denote the set of natural numbers (excluding zero). The previous activity (PA) window is the time window comprised inclusively of the time steps t_1 to $t_1 + n - 1$ $n \in \mathbb{N}$. Denote the average activity of user v_i over the previous activity window by $\mu_{PA}(v_i)$. Take the churn (C) window to be the time window including the time steps $t_2 = t_1 + n$ to $t_2 + m - 1$ inclusive, $m \in N$. Denote the average activity of a user v_i over the churn window as $\mu_C(v_i)$. The i^{th} user therefore churns during the churn window if the following condition holds:

$$\mu_C(v_i) \leq T(S) \cdot \mu_{PA}(v_i) \tag{2.2}$$

$0 \leq T(S) < 1$ is a threshold factor dependent on the relevant system parameters S.

From Definition 2.2, let the relative percentage decrease in average activity from the previous activity, to the churn, time window be

$$q_i = 1 - \left(\frac{\mu_C(v_i)}{\mu_{PA}(v_i)} \right). \tag{2.3}$$

Given that users are incapable of negative activity, it naturally follows that $\mu_{PA}(v_i) > 0$. Using (2.3), (2.4) can be written instead as

$$\mathbb{P}(churn|v_i) = \begin{cases} 0 & \text{if } \mu_C(v_i) \geq \mu_{PA}(v_i) \\ q_i & \text{otherwise.} \end{cases} \tag{2.4}$$

We can infer from (2.4) that for the probability of churn to be non-zero $\frac{\mu_C(v_i)}{\mu_{PA}(v_i)} \in (0, 1)$. Given $T(S) \in [0, 1)$, we can use this to to rewrite (2.2) as

$$\frac{\mu_C(v_i)}{\mu_{PA}(v_i)} \leq T(S).$$

As a result we can ascertain that the i^{th} user churns if

$$\mathbb{P}(churn|v_i) \geq T(S)$$

Subsequently, the observed binary response of churn for the i^{th} user is

$$y_i = \begin{cases} 1 & \text{if } \mathbb{P}(churn|v_i) \geq T(S), \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Experimental set-up for our novel interpretation of churn

The adopted general definition of churn requires choosing the *churn threshold* $T(S)$, which is something that lacks attention in the literature as far as we are aware. As a result, we conduct an analysis with respect to the SCN community towards the sensitivity of churn prediction on this threshold. Considered values for this threshold are in the set $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$.

For n and m in Definition 2.2, we take both equal to one: we predict one time window into the "future" upon having observed user behaviour over one time window. This better enables the automation of the risk analysis for the risk management platform proposed by Nasser et al. (2013a).

All previous applications of Definition 2.1 have measured user activity by the number of posts made per stated time period (Karnstedt et al., 2011). Therefore, all prior churn analyses, on individual users, assumes the behaviours of users who predominantly post questions to be homogeneous with users who predominantly post responses. In addition, as a user's posting behaviour is commonly erratic (Karnstedt et al., 2011), this reduces the significance of churn analysis on posting activity. We intend to measure user activity by the peer-awarded reputation score for two reasons. Firstly, this has more meaning, given that knowledge is the currency of these communities. Secondly, we may limit our attention to the subset of users who are awarded reputation, resulting in a more homogeneous sample population. Consequently, we argue for the study of churn analysis on reputable users.

The Lorenz curves in Figure 2.9 show that only a minority of reputable users earn over 90% of the reputation in the SCN. Additionally, of those users who have solved threads, the top 20% solve more than 80% of the threads created. As in the telecommunication industry, not all users are important to maintaining the health of the online community (Hadden et al., 2007). For example, Nonnecke

and Preece (2000) showed users who lurk have little impact on the health of an online community. In discussion with our contact at SAP, user contentment was identified as being important; more particularly, the contentment of users who actively earn reputation (referred to as *reputable respondents*) was important. It has been shown by Richter et al. (2010) in the telecommunication industry that, if a discontented user is in the vicinity of another user, i, within the social network, the user i is more likely to churn. (We do not explore this avenue within this thesis as this was actively considered by our ROBUST partners from The National University of Ireland Galway as part of Work Package (WP) 5.) Therefore, when performing churn analysis, we only consider users who are *reputable respondents* as determined in Definition 2.3 with $\alpha_r = 0$.

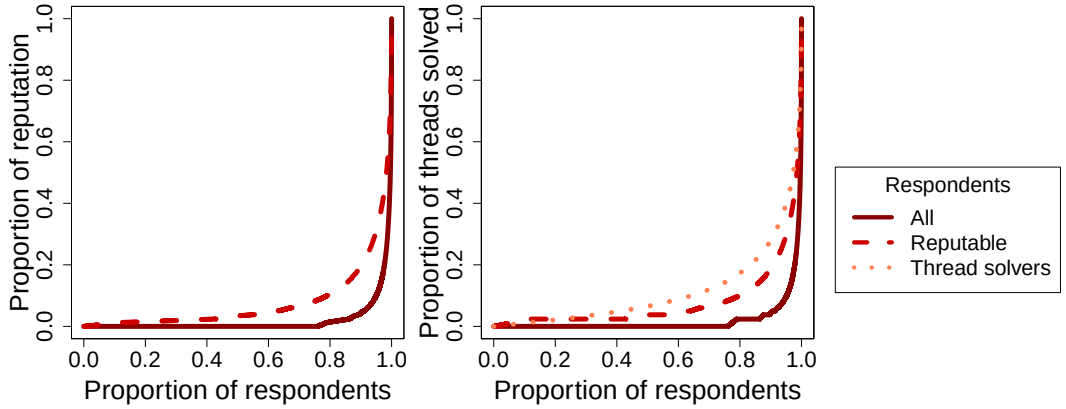


Figure 2.9: Lorenz curves for reputation wealth and thread solving capacity across all fora.

Definition 2.3.

Let $\mu_{(S)}(v_i)$ be the average reputation earned per point award of the i^{th} user between registration and analysis (feature extraction), given some system (S) specific time window for analysis (where $i = 1, \ldots, N$). With respect to some threshold α_r, if for the i^{th} user the inequality

$$\mu_{(S)}(v_i) > \alpha_r$$

holds, this user is a *reputable respondent* of the community.

Each training sample, $\mathcal{T}$, contains all reputable respondents, as defined by Definition 2.3, who post at least once (are active) during the previous activity window. The sample sizes for $\alpha_r = 0$ are illustrated in Figure 2.10 for all five fora

considered. We see that, even for the most active forum in the SCN, the sample sizes appear small. Consequently, we do not place any further restrictions. Although we use forum 50 to demonstrate the capability of our problem formulation on a forum with unstable activity, we are uncertain of the cause of the burst in the number of users participating in the forum during 2010. We suspect that the sudden increase was caused by a rise in the use of Advanced Business Application Programming (ABAP) which may have resulted from SAP releasing an update to the ABAP debugger.

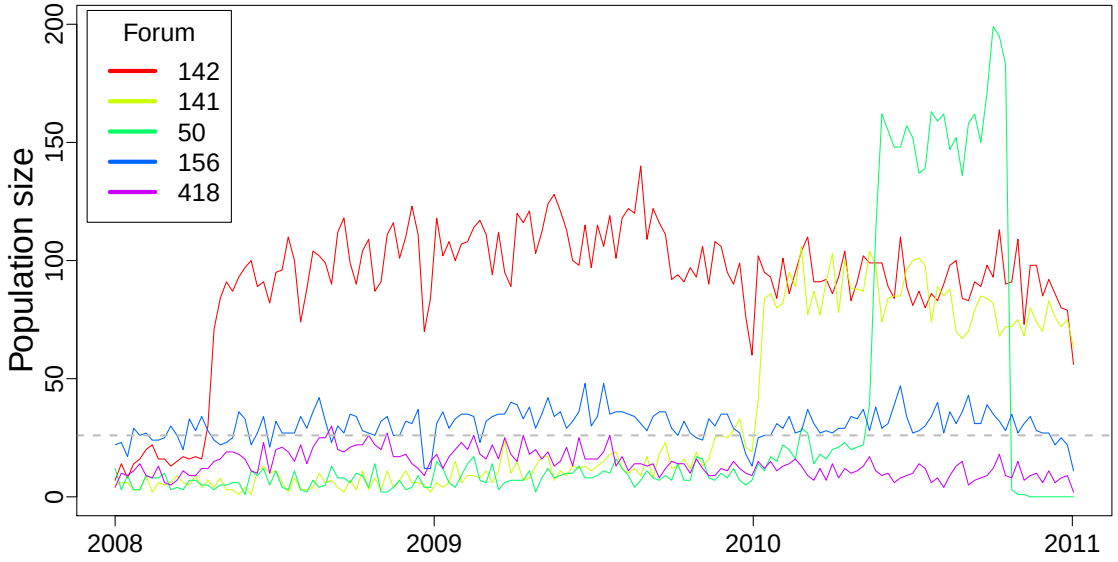
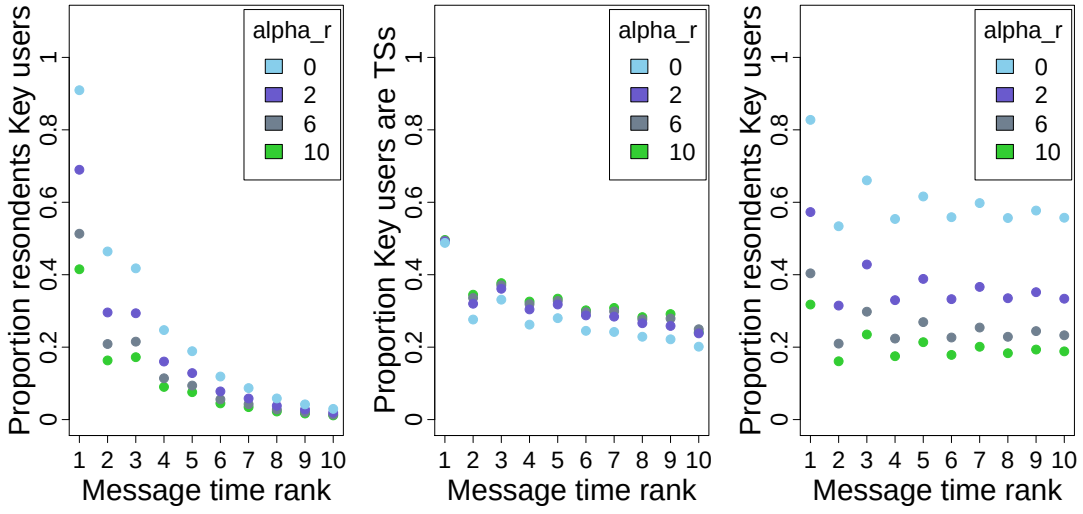


Figure 2.10: Number of reputable respondents making posts per week in the SCN by forum. The size of the novel feature set to predict user churn is indicated by the horizontal grey line.

Relation between our novel user satisfaction formulations

Using the SCN, we can show churn analysis to directly relate to the previously discussed questioner satisfaction problem. Assume the length of time period is a week and, for some solved thread, take n as the number of weeks (time periods) during which the thread solver has been posting in the forum of the considered thread. Let the term *key user* refer to any user whose average weekly reputation gains (since registration) are above α_r. We see that of the threads solved, 92.06% are solved by a key user for $\alpha_r = 0$. Where $\alpha_r = 2$, this percentage decreases to 70.09%; and for $\alpha_r = 6$, this falls to 52.48%. This indicates that any given thread is most likely to be solved by a key user with $\alpha_r \leq 6$.

We now consider the percentage of thread respondents who are key users for $\alpha_r \in \{0, 2, 6, 10\}$. Upon comparing Figure 2.11a to Figure 2.11c we see little evidence of a pattern for those threads which are unsolved in comparison to those which are observed to be solved. For the latter, it is more strongly implied that the first response to a thread is to be made by a key user. However, the later responses in a solved thread, where they exist, are made by respondents who are much less likely to meet the key user criteria. In addition, Figure 2.11b clearly shows the proportion of key users (who are also previous thread solvers), within a solved thread to be much higher for message time rank 1 than for any other message time rank. We also see the proportion of key users at message time rank 1 who are thread solvers, is approximately the same across all thresholds considered (including here $\alpha_r = 10$). This directly links to our previous observation (Section 2.7.2) that of the threads solved, approximately 60% are solved by the first response.



(a) In solved threads. (b) Thread solved at rank. (c) In unsolved threads.

Figure 2.11: Percentage of thread respondents who are key users given some α_r across all fora. Users are included individually for all messages with rank between one and ten.

2.8.3 Novel feature set to predict user churn

We now describe the set of features used when modelling our previously defined user churn event (see Definition 2.2). There are a total of 26 features, and we have

organised these by *type* in this section. Given the close connection between our churn event and that of Karnstedt et al. (2010a), we incorporate, where possible, the features given later (Karnstedt et al., 2011), which are entirely network based. These features are underlined in the list that follows to enable clear identification. Although the majority of our network based features are taken from Karnstedt et al. (2011), we consider a much less costly (in terms of computer memory, computational effort, time for feature extraction) previous activity window of one week as opposed to six months.

We are unaware if the features which are not underlined have been explicitly used previously with regard to churn events in online communities. These features are mostly derived from our previous work on modelling the questioner satisfaction event (see Section 2.7).

Time window user network features

Let the vertices of a directed graph represent the users of an online community, and take the directed edge from user v_i to user v_j to be indicative that user v_i posted at least one message in direct response to user v_j during the time frame considered (see Figure 2.7b). The directed edge between two users may be weighted by the number of unique direct responses. With this set-up, we can consider the graph of users as a social network (see Appendix A.1) and apply algorithms such as *PageRank* to measure the influence of a user.

In what follows, let p_x denote the post identified as x; $\mathcal{P}(t_1, t_2)$ to be the set of all posts written during the period $[t_1, t_2]$; and $\mathcal{P}_i(t_1, t_2)$ to be the set of posts written by the user v_i during the period $[t_1, t_2]$. In addition, let t_{xy} denote the time lag in decimal minutes between two consecutive posts p_x and p_y, where p_x precedes p_y, and take $r(p_x, p_y, t_{xy})$ to represent that the post p_x was directly responded to by the post p_y following a lag of t_{xy} minutes.

- **PageRank:** see Appendix A.2. This measure ranks users by their influence on all others, within the considered forum, during the specified week.

- **indegree:** a count of the number of unique users to have posted in direct response to the user v_i during the specified week. This is a measure of the number of unique incoming connections to some user v_i.

- **outdegree:** a count of the number of unique users to whom the user v_i has posted a direct response during the specified week. This is a measure of the

number of unique outgoing connections to some user v_i.

- **mean reciprocity:** let *reciprocity* be a measure of the time taken for a post made by user v_i to be responded to. Then, assuming that the characteristics about a post can be inferred from the time it takes for a response to be made, reciprocity can convey the importance of user v_i and the type of post that was made. Where $R = |r(p_x, p_y, t_{xy})|$, $p_x \in \mathcal{P}_i(t_1, t_2)$ and $p_y \in \mathcal{P}(t_1, t_2)$, the mean reciprocity for a user v_i over all posts in $\mathcal{P}_i(t_1, t_2)$ is

$$\frac{1}{R} \sum_{r(p_x, p_y, t_{xy})} t_{xy}.$$

- **minimum reciprocity:** using the same concept of reciprocity, this is the minimum time taken for the post of a user v_i to be responded to within the specified time frame. Where $p_x \in \mathcal{P}_i(t_1, t_2)$ and $p_y \in \mathcal{P}(t_1, t_2)$, the minimum reciprocity for a user v_i is

$$\min t_{x,y}$$

- **maximum reciprocity:** using the same concept of reciprocity, this is the maximum time taken for the post of a user v_i to be responded to within the specified time frame. Where $p_x \in \mathcal{P}_i(t_1, t_2)$ and $p_y \in \mathcal{P}(t_1, t_2)$, the maximum reciprocity for a user v_i is

$$\max t_{x,y}$$

- **popularity:** a percentage measure of a user's posts to which there is a direct response. Where $R = |r(p_x, p_y, t_{xy})|$, $p_x \in \mathcal{P}_i(t_1, t_2)$ and $p_y \in \mathcal{P}(t_1, t_2)$ the popularity measure for user v_i is

$$\frac{R}{|\mathcal{P}_i(t_1, t_2)|}.$$

- **initialisation:** a popularity measure for the threads initialised by a user. By implication, the more popular a user, the more likely it is that any thread for which they are the OP will have at least one response. Let the thread l be represented as h_l, and take $|h_l|$ to be the number of posts made in that thread. Where $\mathcal{I}_i(t_1, t_2)$ is the set of all threads initialised by user v_i during

the period $[t_1, t_2]$, we define the initialisation of user v_i to be

$$\frac{|\{h_l | h_l \in \mathcal{I}_i(t_1, t_2) \bigwedge |h_l| > 1\}|}{|\mathcal{I}_i(t_1, t_2)|}$$

Time window summary features

- **# forum participated in:** the count of unique forums in which the user has posted in within the specified week. This measure reflects the variety of topics in which a user has interest and/or knowledge.
- **# threads participated in:** the count of distinct threads within the considered forum in which the user has made within the specified week.
- **# threads created (# times OP):** the count of distinct threads within the considered forum which the user has made the creating post within the specified week. This is consequently the number of distinct threads in which the user has taken the role of OP within the specified week.
- **# threads solved (# times TS):** the count of distinct threads within the considered forum which the user has made the post which solved the thread within the specified week. This is consequently the number of distinct threads in which the user has taken the role of TS within the specified week.
- **# messages:** the count of unique posts in the considered forum which the user has made within the specified week. This gives an indication of how active a user has been terms of sheer mass.
- **reputation earned:** a measure of how active a user has been in terms of providing peer-perceived valuable knowledge during the specified period. Let $|p_x|$ denote the reputation which is awarded to the post p_x, where $p_x \in \mathcal{P}_i(t_1, t_2)$, then the reputation earned by user v_i is

$$\sum_{\mathcal{P}_i(t_1, t_2)} |p_x|.$$

- **mean reputation gained per point awarding:** a measure of how knowledgeable a user is during the specified period. For user v_i, it is defined as

$$\frac{1}{|\{p_x | p_x \in \mathcal{P}_i(t_1, t_2) \bigwedge |p_x| > 0\}|} \sum_{\{p_x | p_x \in \mathcal{P}_i(t_1, t_2) \bigwedge |p_x| > 0\}} |p_x|.$$

- **mean posts in initialisations:** an indicator of the length of discussion had by those users who post in a users thread. The mean length of threads where user v_i initialised the thread as the OP is

$$\frac{1}{|\mathcal{I}_i(t_1,t_2)|}\sum_{h_l\in\mathcal{I}_i(t_1,t_2)}|h_l|.$$

- **mean posts in thread participations:** this indicates the length of a discussion occurring in threads where a user participates during the specified period. Let $\mathcal{H}(t_1,t_2)$ be the set of all active threads (posted in) during the period $[t_1,t_2]$; and $\mathcal{H}_i(t_1,t_2)$ be the set of threads posted in by user v_i during the period $[t_1,t_2]$ such that $\mathcal{H}_i(t_1,t_2)\subset\mathcal{H}(t_1,t_2)$. The mean length of threads in which user v_i participates is

$$\frac{1}{|\mathcal{H}_i(t_1,t_2)|}\sum_{h_l\in\mathcal{H}_i(t_1,t_2)}|h_l|.$$

- **proportion of posts as TS:** an informal expression of the likelihood that some post made by a user provides sufficient knowledge to solve the thread in which it was posted. Given that an award of 10 points to a post within a thread indicates that the thread is solved, the proportion of posts as TS is

$$\frac{|\{p_x|p_x\in\mathcal{P}_i(t_1,t_2)\bigwedge|p_x|=10\}|}{|\mathcal{P}_i(t_1,t_2)|}.$$

Lifetime summary features

- **total # threads participated in:** the count of distinct threads within the considered forum in which the user has posted since registration.

- **total # messages:** the count of unique posts in the considered forum which the user has made since registration.

- **total # replies to user:** the count of all posts made in direct response to any of the unique posts made by the user in the considered forum since the users registration.

- **total reputation earned:** a measure of how active a user has been in terms of providing peer-perceived valuable knowledge since their registration in the considered forum. Let $|p_x|$ be the reputation which is awarded to the post

p_x and let $\mathcal{P}_i(t_2)$ be the set of posts made by user v_i from the beginning of his/her registration time period until t_2. With $p_x \in \mathcal{P}_i(t_2)$, the total reputation earned by user v_i is

$$\sum_{\mathcal{P}_i(t_2)} |p_x|.$$

- **total mean reputation gained per point awarding:** a measure of how knowledgeable a user has been throughout their lifetime in the considered forum. For simplicity, we denote the set of posts made by user v_i that are awarded points between user registration and the end of t_2, $\{p_x \in \mathcal{P}_i(t_2) \bigwedge |p_x| > 0\}$, as $\mathcal{P}_i^*(t_2)$. The average reputation gained per point awarding by user v_i is thus

$$\frac{1}{|\mathcal{P}_i^*(t_2)|} \sum_{p_x \in \mathcal{P}_i^*(t_2)} |p_x|.$$

Temporal features

- **days since registration:** the number of full twenty-four hour periods which have elapsed since the user first posted in the considered forum.

- **mean minutes between messages:** the mean average difference in decimal minutes between posts in the considered thread. Where $r(p_x, p_y)$ signifies that post p_y is a direct response to post p_x, the mean time between postings for user v_i in the considered forum is

$$\frac{1}{|\mathcal{P}_i(t_1, t_2)|} \sum_{r(p_x, p_y) \in \mathcal{P}_i(t_1, t_2)} t_{xy}.$$

- **mean minutes between reputation gains:** an indicator of a users activity level. Let $\mathcal{T}_i(t_1, t_2)$ denote the set of time lags, t_{xy}, between consecutive posts, p_x and p_y, made by user v_i, that have points awarded. The mean time between posts which are recognised for user v_i is

$$\frac{1}{|\mathcal{T}_i(t_1, t_2)|} \sum_{t_{xy} \in \mathcal{T}_i(t_1, t_2)} t_{xy}.$$

All the features listed above are used to perform churn analysis as outlined

earlier in this section. We do not perform feature analysis or selection due to time constraints and that, with respect to the ROBUST project, SAP were greatly concerned about whether the choice of churn threshold significantly affects classifier performance. That is, given that this event was to be used by SAP community managers, the SAP representative wanted to know whether some churn thresholds led to a lower level of accuracy when predicting the churn event than others. (We recommend that anyone wishing to use our churn event, including the features, should perform feature selection.) The corresponding results of this classification analysis are discussed in Section 6.5.

2.9 Conclusion

This chapter began with addressing its first objective: to describe the SCN and the available data. Within this, we saw how the SCN is structured; that it is comprised of messages which are posted in threads, which are contained by topic in fora, which are hosted on the SCN platform that is owned by SAP. We introduced different labels for users who meet certain criteria in the context of a thread, and described the peer-awarded user reputation program which is similar to those in other question and answer online communities. Following this, we noted that the initial four years of the community were relatively unstable compared to the following three year period, where the community seemed to be well established. The fora identified by the IDs 142,141,50,156 and 418 were selected (from the 95 to which we had access to) and these were shown to have different activity patterns.

We then addressed the second objective of this chapter: to relate to the reader what is meant by online community health, without reference to any specific community, highlighting that user satisfaction is a key aspect. This was done through summarising one of the most popular models for community health and finding user satisfaction to be one of the six categories identified for the success of the community. An investigation into previous attempts to model user satisfaction found the approaches given by the majority of the literature fitted into three main categories: role analysis, questioner satisfaction, and churn analysis.

This naturally leads us to the third, final, and most significant objective of this chapter: to review the relevant literature given the objectives of the ROBUST project and to develop novel event formulations of user satisfaction where necessary. In accordance, the latter part of this chapter our preliminary analysis of the SCN identified the event that a questioner is satisfied within some relevant

time frame to be of interest to the SAP representative. A study of the literature related to questioner satisfaction lead us to believe that the closest event definition in the literature was that the questioner is satisfied. This existing event does not incorporate time as a dimension of the response and, as such, was deemed unsuitable. Therefore, we determine that our formulation of the questioner satisfaction event, which considered the time element, does provide a necessary aspect of modelling user satisfaction in online communities that has not been previously investigated.

That users may leave a community (i.e. churn) is an integral concern to any platform provider. We have shown that prior to this work no rigorous definition of churn in online communities existed and have therefore provided such a definition. This definition took care to expand on the many facets of churn analysis including setting up the experiments from which empirical analysis is made (Chapter 6) and identifying the users for whom churn analysis provides valuable insights.

For both of our proposed problem formulations, we have provided a comprehensive set of features, some of which are driven by social network analysis on the users of the SCN. All our features, and the features taken from elsewhere, have been fully described to enable our formulations of user satisfaction to be applied to different communities, or for others to analyse our work. The code which was written to extract the features from the SQL data provided by SAP was written in Java and, with the assistance of ROBUST consortium members from IT Innovations, was parallelised across all available cores of a standard computer. This code is owned by the University of Southampton. We also demonstrated the presence of a connection between churn analysis and our given formulation of questioner satisfaction.

Chapter 3

Classification Methods

The objective of this chapter is to convey the theory from the literature that is required to understand and appropriately use the *lda*, *qda* and *glm* functions of the MASS package within R written by Venables and Ripley. These functions are verified and validated implementations of the linear discriminant analysis, quadratic discriminant analysis and generalised linear model type classifiers respectively.

3.1 Introduction

In what follows, we give the theoretical details of the classification methods used to model our novel risk event formulations of user satisfaction in online communities. In the context of the ROBUST project, we had previously provided ROBUST with an implementation of the Bayesian classification method described in Chapter 4. The classifiers considered in this chapter do not incorporate Bayesian inference and are thus used to validate whether the model given to ROBUST is really suitable; or whether there exists a more simplistic, less computationally expensive, classification method that performs as well (or better). As a result, we did not aim to code the classification methods given in this chapter, but used the corresponding standard functions of the "MASS" package within R written by Venables and Ripley. (Classifiers were compared against our implementation of the Bayesian classification method in R, see Section 4.3.1 for a discussion of this implementation.)

The content of this chapter therefore aims to give the reader a sufficient overview of the theory so that the corresponding standard functions in the "MASS" package within R can be used appropriately. A by product of this is that it also

highlights any advantages of one method over another which relates to our comparison of the characteristics of classifier performance in Chapter 6.

3.1.1 Outline of chapter

As the nature of our problem is to assign a value to $G(x) \in \mathcal{G}$, the input space defined by the features $\mathbf{X}$ is divided into labelled regions according to some classification. The boundaries of such regions are the *decision boundaries*. Where these are linear in $\mathbf{X}$, the classification method is said to be linear. One such method is linear regression, given in Subsection 3.1.3. This method is an example of *discriminant analysis* given that it models *discriminant functions* denoted $\delta_k(x)$ for each class $(k = 1, \ldots, K)$, and classifies an observation to that class for which the function has greatest value. Additional types of discriminant methods are *linear discriminant analysis* (Section 3.3.1) and *quadratic discriminant analysis* (Section 3.3.2). Included in the class of linear classification methods are those which model the posterior probabilities of class membership written as $\mathbb{P}(G = \mathcal{G}_k | X = x)$. Typical examples of such methods are those which can be formulated as generalized linear models (see Section 3.4). Inference about generalized linear models is most commonly made via the *maximum likelihood* approach, but Bayesian approaches can also be made to include prior information about beliefs held as described in Chapter 4. We take our baseline model as the random assignment of class membership — this is laid out in Section 3.2. A brief theoretical comparison of classification methods is given in Section 3.5. For completeness, the reader is referred to Chapter 6 where we provide an empirical analysis using the quality characteristics of Chapter 5 with respect to the problems in Chapter 2.

3.1.2 Notation

The risk and/or opportunity events which we consider have many common characteristics. In both cases, there are inputs, or *features*, which are observations of the past or present. These inputs are connected in some way to one or more outputs, or *responses*. Our aim is to perform supervised learning, using the features to predict the corresponding responses. More traditionally, the inputs (or features) are called the independent variables and the outputs (or responses) are called the dependent variables.

In the events formulated, the responses are categorical (qualitative), belonging to the finite set $\mathcal{G}$ with cardinality of two. Given a set of features for an observation,

we want to predict its class label (response). Therefore our prediction task is one of classification. Our features are mainly quantitative; those which are qualitative are categorical, represented by a binary digit (i.e. either 0 or 1). As is convention, we will denote all input variables by X. Where X is a vector, we reference its components by the subscript j as X_j. Our qualitative outputs are noted as G where $G \in \mathcal{G}$. When referring to values which are observed, these shall be noted in the lower case; as such, the i^{th} observed value of X is x_i, where x_i is either a scalar or vector. Matrices are indicated in bold, for example, the $N \times p$ dimensional feature matrix $\mathbf{X}$. Vectors on the other hand, are not indicated in bold unless of length N. This enables clear distinction between the p length feature vector x_i of the i^{th} observation from the N length vector $\mathbf{x}_j$ consisting of all observations of the j^{th} feature, that is $\mathbf{x}_j = X_j$. Given the assumption that all vectors are column vectors, x_i^T represents the i^{th} row of $\mathbf{X}$.

Due to the qualitative response G being finite categorical, it may be coded by an indicator variable Y. Let K be the cardinality of $\mathcal{G}$, such that there exist K possible indicators Y_k where $k = 1, \ldots, K$ and

$$Y_k = \begin{cases} 1 & \text{if } G = \mathcal{G}_k, \\ 0 & \text{otherwise,} \end{cases}$$

thus $Y = [Y_1, \ldots, Y_K]$. For the training data $\mathcal{T}$ of N observations we have the $N \times K$ dimensional matrix $\mathbf{Y}$ representing the *indicator response matrix* where all rows sum to unity. In the case $K = 2$ it is sufficient to take $\mathbf{Y}$ as the $N \times 1$ dimensional matrix with binary entries.

The supervised learning task can be stated as: given a vector of feature observations X, predict the qualitative response G, denoted $\hat{G}$, where $\hat{G}$ take values in the set $\mathcal{G}$ corresponding to G. As we consider only G where $\mathcal{G}$ has cardinality of two, we may zero-one code the quantitative response Y such that Y is a binary response variable. The predicted value of this quantitative response is denoted $\hat{Y}$, and thus, $\hat{y} \in \{0, 1\}$. Where probabilistic classifiers are used, the model provides an estimate of the posterior probability of class one membership denoted as $\hat{P}(Y = 1|X = x)$, which for brevity may be written as $\hat{p}(x)$. A *discriminating threshold*, d, on $\hat{p}(x)$ gives $\hat{y} = 1$ where $\hat{p}(x) \geq d$ and $\hat{y} = 0$ otherwise. The estimated qualitative response $\hat{g}$ takes value $G_{\hat{y}}$, assuming some discrimination threshold $d \in (0, 1]$. To construct any prediction rule, we require *training data*, noted by $\mathcal{T}$, comprised of the sets (x_i, g_i) for $i = 1, \ldots, N$.

We let $\mathbf{X}$ be the normalized column matrix with rows $x_i^\top = [x_{i,1}, \ldots, x_{i,p}]$ where $x_{i,j}$ is the i^{th} observation of the j^{th} feature; and $i = 1, \ldots, N$, for N the number of observations within the sample population. To avoid identifiability or non-integrability issues later on, we assume that $\mathbf{X}^\top\mathbf{X}$ is non-singular. Given that $\mathbf{X}$ has full column-rank, this assumption is always satisfied. In addition let β be the corresponding p length vector of unknown (elasticity) coefficients.

3.1.3 Linear model (fit by least squares)

The linear model is, and has been for several decades, the backbone of statistics (Hastie et al., 2009). The response Y, is assumed to be an independent normally distributed random variable, with mean μ and constant variance σ^2. Note that whilst this assumption is in direct contradiction with our definition of Y, there exist classification tasks where the observed response is continuous, denoted as Z, over some interval but later dichotomised to provide class membership G.

Taking the feature vector $X = [X_1, X_2, \ldots, X_p]$, Z is predicted as a vector of coefficient estimators $\tilde{\beta}$,

$$\tilde{Z} = \mathbb{E}(Z) = \mu = \tilde{\beta}_0 + \sum_{j=1}^{p} X_j \tilde{\beta}_j, \tag{3.1}$$

where $\tilde{\beta}_0$ is the intercept coefficient. Assuming the constant variable 1 to be included in X, such that $\tilde{\beta}_0$ is added to the p length vector of coefficients $\tilde{\beta}$, (3.1) can be rewritten (as an inner product) in vector form,

$$\tilde{Z} = X\tilde{\beta}. \tag{3.2}$$

We see that without the addition of any restrictions, $\tilde{Z}$ may take any real value in the interval $(-\infty, \infty)$. From here on, we assume that the intercept coefficient is included within $\tilde{\beta}$.

The simplest and most popular way to determine the coefficients $\tilde{\beta}$ is via least squares. For least squares we take β such that the residual sum of squares (RRS) is minimized, so

$$\tilde{\beta} = \underset{\beta \in \Re^p}{\arg\min}\ \text{RSS}\,(\beta) = \underset{\beta \in \Re^p}{\arg\min} \sum_{i=1}^{n_o} \left(z_i - x_i^\top \beta\right)^2. \tag{3.3}$$

This is a straightforward convex optimization problem. Given $\mathbf{X}$ is the $N \times p$

matrix, each row relates to a feature vector $(x_i^\top)$, and $\mathbf{z}$ the corresponding N length vector of responses, we can write $\mathrm{RSS}(\beta)$ in matrix notation

$$\mathrm{RSS}(\beta) = (\mathbf{z} - \mathbf{X}\beta)^\top (\mathbf{z} - \mathbf{X}\beta). \tag{3.4}$$

Differentiating (3.4) with respect to β provides the standard equations

$$\mathbf{X}^\top \left(\mathbf{z} - \mathbf{X}\tilde{\beta}\right) = 0. \tag{3.5}$$

Assuming that $\mathbf{X}^\top\mathbf{X}$ is non-singular, the unique solution of the standard equations (3.5) is

$$\tilde{\beta} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{z}. \tag{3.6}$$

With this in mind, the fitted value for the feature vector x_i is $\tilde{z}_i = \tilde{z}(x_i) = x_i^\top\tilde{\beta}$. Hence for some x_0, the predicted response is $\tilde{z}(x_0) = x_0^\top\tilde{\beta}$. Without placing restrictions on x_0 the response $\tilde{z}(x_0)$ may take values in the interval $(-\infty, \infty)$. This method is consequently inappropriate for classification, although may be used as described in the following examples regarding our specified novel problem formulations given in the previous chapter.

Example: questioner satisfaction

In the questioner satisfaction classification problem outlined in Subsection 2.7.2, there is no underlying continuous response. The categorical response G is observed directly as either *unsolved* or *solved*. Consequently the only response available is the indicator response variable Y having value 1 (0) if the thread is *solved* (*unsolved*). Thus the least squares linear model is most unsuited to this problem. If one were to apply the method to the task, a discrimination threshold d in the interval $(-\infty, \infty)$ would be required to convert the value fitted by the linear model to one of the possible class responses using the rule

$$\tilde{G} = \begin{cases} \textit{solved} & \text{if } \tilde{Y} \geq d, \\ \textit{unsolved} & \text{otherwise.} \end{cases} \tag{3.7}$$

Example: individual user churn

For the problem definition of individual user churn, the set of possible values for the categorical response G is $\mathcal{G} = \{no - churn, churn\}$. Correspondingly, the indicator variable Y takes the value 1 if the user activity *churn*s and 0 otherwise.

However, given the problem formulation in Section 2.8.2, the occurrence of "churn" is dependant on the subjective choice of the churn threshold, $T(S)$. The response G is a dichotomised interpretation of the response Z which is continuous in the range $[0, 1]$. Given a prediction of Z denoted $\tilde{Z}$, one may use the α_q churn threshold to produce a prediction of class membership:

$$\tilde{G} = \begin{cases} churn & \text{if } \tilde{Z} \geq \alpha_q, \\ no - churn & \text{otherwise.} \end{cases}$$

3.1.4 Statistical modelling for classification

In the general case, where we assume quantitative response, let $X \in \Re^p$ and $Y \in \Re$ with joint probability distribution $\mathbb{P}(X, Y)$. We seek the function $h(X)$ capable of predicting Y. To penalize errors in the prediction of Y requires a *loss function*, the most common being the one used in (3.3), the squared error loss or L_2-norm $(Y - h(X))^2$. McCullagh and Nelder (1989) show that for the L_2-norm to be a suitable measure of deviation we require stochastic independence of observations, in addition to the variance being independent of the mean for each observation. Assuming these conditions to be met the expected prediction error (EPE) for some $h(X)$ is

$$\text{EPE}(h(X)) = \mathbb{E}(Y - h(X))^2. \tag{3.8}$$

Conditioning the EPE on X gives

$$\text{EPE}(h(X)) = \mathbb{E}_X \mathbb{E}_{Y|X} \left((Y - h(X))^2 | X \right).$$

If we minimize the conditional EPE pointwise we see that the best function $h(X)$ is the conditional expectation

$$h(x) = \mathbb{E}(Y|X = x).$$

Therefore, when using the squared error loss function, the best prediction of Y at some $X = x$ is the conditional mean.

Consider for a moment $h(X)$ as defined for linear regression (3.2), and replace this for $h(X)$ in (3.8):

$$\text{EPE}(h(X)) = \mathbb{E}\left(Y - X^\top \beta\right)^2. \tag{3.9}$$

Differentiating (3.9) and solving analytically for β gives

$$\beta = \left(\mathbb{E}\left(XX^\top\right)\right)^{-1}\mathbb{E}\left(XY\right); \tag{3.10}$$

which is not conditioned on X. If we replace the expectation in (3.10) by averages on the training data $\mathcal{T}$, we get the solution found by solving least squares (3.6).

Recall that the predicted value $\hat{G}$ should take values in the set $\mathcal{G}$. The loss function may be represented by the $K \times K$ matrix L, where K is the cardinality (size) of the set $\mathcal{G}$, whose elements take value zero on the diagonal and non-negative reals elsewhere. The value of the element in position (k, ℓ) is the cost for incorrectly classifying an observation of class $\mathcal{G}_k$ as one of $\mathcal{G}_\ell$. Most commonly, we take the off-diagonal elements of L equal to 1, giving the zero-one loss function. In general, the expected prediction error for categorical response is $\text{EPE} = \mathbb{E}\left(L\left(G, \hat{G}(X)\right)\right)$. Conditioning on X gives

$$\text{EPE} = \mathbb{E}_X \sum_{k=1}^{K} L\left(\mathcal{G}_k, \hat{G}(X)\right)\mathbb{P}\left(\mathcal{G}_k, X\right),$$

minimizing this pointwise and assuming the zero-one loss function leads to

$$\hat{G}(X) = \mathcal{G}_k \quad \text{if} \quad \mathbb{P}\left(\mathcal{G}_k | X = x\right) = \max_{g\in\mathcal{G}}\mathbb{P}\left(g | X = x\right). \tag{3.11}$$

Given that we use the 'dummy-variable' approach as laid out in Section 1.4, coding our categorical responses G by binary Y, $\hat{h}(X) = \mathbb{E}(Y|X) = \mathbb{P}\left(G = \mathcal{G}_1 | X\right)$ where $\mathcal{G}_1$ corresponds to class 1, as in the example of (3.7). This approach has drawbacks where the values of the probability $\hat{h}(X)$ may be outside the interval $[0, 1]$. Consequently we do not consider $h(X)$ as in (3.2).

The aim is thus to find some approximation $\hat{h}(X)$ of $h(X)$ such that

$$Y = \hat{h}(X) + \epsilon,$$

where $\epsilon \sim N(\mu, \sigma)$, is maintained within the interval $[0, 1]$. (Where $h(x)$ is the basic linear function, $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$, our statistical model is linear regression.) Typically we assume some cumulative probability distribution for Y giving

$$Y = \int_{-\infty}^{t} f(s)ds,$$

assuming $f(s) \geq 0$ and $\int_{-\infty}^{\infty} f(s)ds = 1$. We term the probability density function $f(s)$ the *tolerance distribution*. The choice of tolerance distribution dictates the *link function* between the response Y and *linear predictor* $\eta = \beta x$. For example where the tolerance distribution assumed is

$$f(s) = \frac{\beta \exp(\beta s)}{[1 + \exp(\beta s)]^2},$$

we have

$$Y = \int_{-\infty}^{x} f(s)ds = \frac{\exp(\beta x)}{1 + \exp(\beta x)}$$

which leads to the logit link function

$$\log\left(\frac{Y}{1-Y}\right) = \beta x.$$

Recall the training set denoted as $\mathcal{T}$ to be comprised of pairs of observations (x_i, g_i) where g_i is the classification of the i^{th} observation, for $i = 1, \ldots, N$. To ensure the L_2-norm is a valid criterion, we assume that all pairs $(x_i, g_i) \in \mathcal{T}$ are independent random samples from the underlying population.

The decision theory culminating in (3.11) implies that optimal classification requires the class posteriors $\mathbb{P}(G = \mathcal{G}_k | X = x)$ to be known. Given $K = \text{card}(\mathcal{G})$, let the prior probability of an observation being of the k^{th} class be $\pi_{(k)}$, with $\sum_{k=1}^{K} \pi_{(k)} = 1$. Each $\pi_{(k)}$ may be estimated empirically from the observations within $\mathcal{T}$

$$\hat{\pi}_{(k)} = \frac{\sum_{i=1}^{N} 1(g_i = \mathcal{G}_k)}{N}, \tag{3.12}$$

where $1()$ is the indicator function. Take $f_k(x)$ to represent the class-conditional density of X when in class $G = \mathcal{G}_k$. Applying Bayes' theorem provides

$$\mathbb{P}(G = \mathcal{G}_k | X = x) = \frac{f_k(x)\pi_{(k)}}{\sum_{\ell=1}^{K} f_\ell(x)\pi_{(\ell)}};$$

which implies $f_k(x)$ to be approximately equivalent to $\mathbb{P}(G = \mathcal{G}_k | X = x)$. Some modelling techniques model the class-conditional densities, for example discriminant analysis (Section 3.3); whilst others model the posterior probabilities of the classes ($\mathbb{P}(G = \mathcal{G}_k | X = x)$ for $k = 1, \ldots, K$), for example generalised linear models (Section 3.4) (McCullagh and Nelder, 1989).

3.2 The (Random) Baseline Model (RAND)

We assume the data are from independent binary random trials

$$Y_i \overset{\text{i.d.}}{\sim} \text{Ber}\left(\pi_{(1)}\right),$$

where $\pi_{(1)}$ is the prior probability of an observation being in class 1 (positive), as opposed to class 0 (negative). The prior probability for the k^{th} class is taken as the empirical estimate given in (3.12), for $K = 2$. Where Y is coded as the zero-one version of the categorical response G, $\mathbb{E}(Y|X = x) = \pi_{(1)}$ and $\text{Var}(Y|X = x) = \pi_{(1)}(1 - \pi_{(1)}) = \pi_{(1)}\pi_{(0)}$.

3.3 Discriminant Analysis

Discriminant analysis began with Fisher (1936) using linear discriminant functions to model the famous Iris flower dataset. This was the first well defined statement on the problem of discrimination — with Bliss (1934) concerned more with the effect of a toxic agent on survival — as well as the first proposed solution. Fisher's linear discriminant functions do not require any distributional assumptions, except when justifying the likelihood ratio (Goldstein and Dillon, 1978). This is key for Goldstein and Dillon (1978), who promote modelling discrete (or qualitative) variables rather than treating them as continuous variables. However, the justification of the likelihood ratio is eased for both linear and quadratic discriminant analysis if we assume multivariate normality of the vector $\mathbf{X}$ (Fisher, 1936). Here on in, we will be considering the likelihood and hence will make the assumption of multivariate normality.

From Subsection 3.1.4, we know discriminant analysis takes the class-conditional density $f_k(x)$ as an estimate of the class posteriors $\mathbb{P}(G = \mathcal{G}_k|X = x)$. Both linear and quadratic discriminant analysis assume these densities to be multivariate Gaussian (Press and Wilson, 1978)

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x - \mu_k)^\top \boldsymbol{\Sigma}_k^{-1}(x - \mu_k)\right] \text{ for } x \in \Re^p \tag{3.13}$$

where

$$\delta(x, \mu_k) = \left((x - \mu_k)^\top \boldsymbol{\Sigma}_k^{-1}(x - \mu_k)\right)^{\frac{1}{2}}$$

is the Mahalanobis distance (Mahalanobis, 1936) between the features x and the

centre of the k^{th} class.

We compare two classes k and ℓ via the log ratio

$$\begin{aligned}\log \frac{\mathbb{P}\left(G=\mathcal{G}_k|X=x\right)}{\mathbb{P}\left(G=\mathcal{G}_\ell|X=x\right)} &= \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_{(k)}}{\pi_{(\ell)}} \\ &= \log \frac{\pi_{(k)}}{\pi_{(\ell)}} + \log \left|\frac{\boldsymbol{\Sigma}_\ell}{\boldsymbol{\Sigma}_k}\right| - \frac{1}{2}\left(\mu_k^\top \boldsymbol{\Sigma}_k^{-1}\mu_k - \mu_\ell^\top \boldsymbol{\Sigma}_\ell^{-1}\mu_\ell\right) \\ &\quad + x^\top\left(\boldsymbol{\Sigma}_k^{-1}\mu_k - \boldsymbol{\Sigma}_\ell^{-1}\mu_\ell\right) - \frac{1}{2}x^\top\left(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_\ell^{-1}\right)x.\end{aligned} \tag{3.14}$$

The decision boundaries for discriminant analysis may be found by setting the log ratio to zero. Where the covariance matrices $\boldsymbol{\Sigma}_k$ (for $k = 1, \ldots, K$) are assumed equal across all K classes, (3.14) becomes linear in x, and we have linear discriminant analysis (Subsection 3.3.1). Alternatively, the decision boundary is quadratic in x and we have quadratic discriminant analysis (Subsection 3.3.2).

3.3.1 Linear Discriminant Analysis (LDA)

For features X_j (where $X = [X_1, X_2, \ldots, X_p]$) which are normally distributed, the class-conditional density is as given in (3.13). To ensure linearity in the features x, we assume all classes have a common covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_k \ \forall k \in \{1, \ldots, K\}$. The necessity of this assumption becomes evident when comparing two classes, k and ℓ, via the log ratio without this assumption (3.14) where the last term is quadratic in x. In the case of linear discriminant analysis, (3.14) simplifies to

$$\begin{aligned}\log \frac{\mathbb{P}\left(G=\mathcal{G}_k|X=x\right)}{\mathbb{P}\left(G=\mathcal{G}_\ell|X=x\right)} = \ & \log\frac{\pi_{(k)}}{\pi_{(\ell)}} - \frac{1}{2}(\mu_k+\mu_\ell)^\top \boldsymbol{\Sigma}^{-1}\left(\mu_k-\mu_\ell\right) \\ & + x^\top \boldsymbol{\Sigma}^{-1}\left(\mu_k-\mu_\ell\right),\end{aligned} \tag{3.15}$$

now linear in x and the normalization factors cancel entirely. By assuming $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_k \ \forall k$, the class-conditional densities are merely shifted versions of one another. The linear nature of (3.15) implies the decision boundary between any two classes k and ℓ, for $k \neq \ell$, to be linear in the features x; a hyperplane in p dimensional space. Therefore, $\Re^p$ is partitioned for K classes by at most $(K-1)!$ hyperplanes, where $()!$ represents the factorial of $()$. Note that where the common covariance $\boldsymbol{\Sigma}$ is spherical, that is $\sigma^2\mathbf{I}$, and the class priors $\pi_{(k)}$ are equal, the hyperplanes are the perpendicular bisectors of those joining the centroids of each class. For optimal

classification:

$$\begin{aligned}\hat{G}(x) &= \arg\max_k \mathbb{P}\left(G = \mathcal{G}_k | X = x\right), \\ &= \arg\max_k \; f_k(x)\pi_{(k)} = \arg\max_k \; \log\left(f_k(x)\pi(x)\right), \\ &= \arg\max_k \left(x^\top \mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^\top \mathbf{\Sigma}^{-1}\mu_k + \log \pi_{(k)}\right). \end{aligned} \tag{3.16}$$

We see from (3.15) and (3.16) that the *linear discriminant functions*

$$\delta_k(x) = x^\top \mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^\top \mathbf{\Sigma}^{-1}\mu_k + \log \pi_{(k)} \tag{3.17}$$

describe the decision rule $\hat{G}(x) = \arg\max_k \mathbb{P}\left(G = \mathcal{G}_k | X = x\right)$ equivalently as $\hat{G}(x) = \arg\max_k \delta_k(x)$. The decision boundary between the classes k and ℓ may thus be expressed as $\{x : \delta_k(x) = \delta_\ell(x)\}$, which is equivalent to setting the log-ratio (3.15) to zero.

In practice, the parameters of the multivariate Gaussian distribution are estimated from the observations (x_i, g_i) within the training set $\mathcal{T}$:

$$\begin{aligned}\hat{\pi}_{(k)} &= N_k/N; \\ \hat{\mu}_k &= \sum_{g_i=k} x_i/N_k; \\ \hat{\mathbf{\Sigma}} &= \sum_{k=1}^{K}\sum_{g_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top/(N-K), \end{aligned}$$

where $N_k = \sum_{i=1}^{N} 1\left(g_i = \mathcal{G}_k\right)$ is the number of observations in class k. Given that we only need the differences between discriminant functions, $\delta_k(x) - \delta_K(x)$ where K is a pre-chosen class (we choose the last), each requiring $(p+1)$ parameters, there are $(K-1)\times(p+1)$ parameters in fitting linear discriminant analysis.

From (3.17) we see that we are required to compute the inverse of the common covariance matrix. For a unique inverse of $\mathbf{\Sigma}$ to exist we require the number of observations N to outnumber the dimensionality p. In this case, $\mathbf{\Sigma}$ is a $p \times p$ square matrix of full rank and we can use singular value decomposition on $\mathbf{\Sigma}$ to find the inverse. However, in practice, where $N \leq p$, we can apply singular value decomposition on $\mathbf{X}$ to compute an estimate of $\mathbf{\Sigma}^{-1}$.

3.3.2 Quadratic Discriminant Analysis (QDA)

We now drop the assumption that a common covariance matrix, $\mathbf{\Sigma}$, exists for all K classes, and instead allow each to have its own specific covariance matrix $\mathbf{\Sigma}_k$, for $k \in \{1, \ldots, L\}$. The log ratio for the comparison of the classes k and ℓ is consequently as given in (3.14), with the second term being a remnant of the normalisation factors in (3.13) and the last term being quadratic in x. As a result, the decision boundary between any two classes k and ℓ, where $k \neq \ell$, is quadratic in the features x. Again we have $\Re^p$ being partitioned by maximum of $(K-1)!$ hyperplanes. Optimal classification is found by:

$$\begin{aligned}\hat{G}(x) &= \arg\max_k \mathbb{P}\left(G = \mathcal{G}_k | X = x\right), \\ &= \arg\max_k \; f_k(x)\pi_{(k)} = \arg\max_k \; \log\left(f_k(x)\pi(x)\right), \\ &= \arg\max_k \left(x^\top \mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^\top \mathbf{\Sigma}_k^{-1}\mu_k + \log \pi_{(k)} \right. \\ &\qquad \left. + \log\left|\frac{1}{\mathbf{\Sigma}_k}\right| - \frac{1}{2}x^\top \mathbf{\Sigma}_k^{-1} x \right).\end{aligned} \tag{3.18}$$

From (3.14) and (3.18) the *quadratic discriminant functions* are

$$\delta_k(x) = x^\top \mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^\top \mathbf{\Sigma}_k^{-1}\mu_k + \log \pi_{(k)} + \log\left|\frac{1}{\mathbf{\Sigma}_k}\right| - \frac{1}{2}x^\top \mathbf{\Sigma}_k^{-1} x. \tag{3.19}$$

As for linear discriminant analysis, the decision boundary between the classes k and ℓ may be written as $\{x : \delta_k(x) = \delta_\ell(x)\}$, which is equivalent to setting the log-ratio (3.14) to zero. For quadratic discriminant analysis, where the discriminant functions are complete, there exist $(p+1)(p+2)/2$ terms which leads to $K(p+1)(p+2)/2$ parameters requiring estimation (Webb, 2002, p. 24).

We see from (3.19) that finding the quadratic decision boundaries requires the computation of the inverse of the within group/class covariance matrix. For the QR-decomposition to be uniquely solvable we require the matrix $\mathbf{\Sigma}_k$ to be of full rank, and thus we require $N_k > (p+1)$.

3.4 Generalised Linear Model (GLM)

Generalised linear models give a unified approach which can be shown to encompass many important models, for both continuous and discrete responses (and

features) (Agresti, 2013). For our purposes, the generalised linear model ensures the fitted value of the discrete response Y stays in the desired $[0, 1]$ interval by choosing function $g()$ of the mean which effectively maps the response variable $Y \in [0, 1]$ to a latent variable $Z \in (-\infty, \infty)$. Thus, rather than the model being of form $\mathbb{E}(Y) = \mu = X\beta$ (with constant variance σ^2), as in the case of a linear model, the generalised linear model has form

$$g\left(\mathbb{E}(Y)\right) = X\beta, \tag{3.20}$$

where the right-hand side remains linear in the unknown coefficients β.

Generalised linear models relax the assumption of linear models that the distribution of the response Y is, or is well-approximated by, a normal distribution. Instead, generalised linear models assume some probability density, or mass function, from the exponential family (see (Casella and Berger, 2002, Chapter 3)) for the N independent observations $(y_1, \ldots, y_N)$ of Y. The typical form is

$$f_Y(y_i; \theta_i, \phi) = \exp\left\{\frac{(y_i\theta_i - b(\theta_i))}{a(\phi)} + c(y_i, \phi)\right\} \tag{3.21}$$

where $a()$ (usually of form ϕ/ω_i for prior weight ω_i known and varying over independent observations $i = 1, \ldots, N$), $b()$ and $c()$ are specific functions; θ_i is the canonical (or natural) parameter; and ϕ is the dispersion parameter (Jørgensen, 1987).

The general form of the log-likelihood is $l(\theta, \phi; y) = \sum_i l(\theta_i, \phi; y_i)$ due to independence between the observations, where $l(\theta_i, \phi; y_i) = \log f_Y(y_i; \theta_i, \phi))$. Therefore, using (3.21), the log-likelihood can be expressed as

$$l(\theta, \phi; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi). \tag{3.22}$$

The familiar general likelihood results below hold under regularity conditions satisfied by the exponential family (Cox and Hinkley, 1974, Section 4.8)

$$\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = 0,$$
$$\mathbb{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right) + \mathrm{Var}\left(\frac{\partial l}{\partial \theta}\right) = 0,$$

(Fisher, 1922). Applying these results here, it can be easily shown from (3.22)

that

$$\mathbb{E}(Y) = b'(\theta) =: \mu, \tag{3.23}$$

and

$$\mathrm{V}(Y) = a(\phi)b''(\theta) =: a(\phi)\mathrm{V}(\mu), \tag{3.24}$$

where $\mathrm{V}(\mu)$ represents the variance function corresponding to the assumed family. In addition, generalised linear models do not require the assumption of constant variance (unlike in linear discriminant analysis). Nevertheless, the relationship between the variance and the mean is assumed to be known. In fact, from (3.23) and (3.24), we see

$$\frac{\partial \mu}{\partial \theta} = \mathrm{V}(\mu). \tag{3.25}$$

Let $\eta = X\beta \in \Re$ be called the *linear predictor* associated to Y. Therefore we have

$$\frac{\partial \eta}{\partial \beta_j} = x_j.$$

With respect to the linear model of Section 3.1.3 we have $\mu = \eta$. A generalization of this involves the *link function* $g()$ which is any differentiable, monotonic, one-to-one function with

$$g(\mu) = \eta, \tag{3.26}$$

(for the linear model $g()$ is the identity function). Given $g()$ is on-to-one, there exists an inverse such that the expected response $\mathbb{E}(Y) = \mu$ may be expressed via the linear predictor as $\mathbb{E}(Y) = g^{-1}(\eta) = \mu$.

Given an assumed exponential family, if for the considered link function the canonical parameter θ is equivalent to the linear predictor η, the link function is called the natural, or canonical link function. However, as previously stated, we may choose any monotonic function for the link so long as it maps the linear predictor to the range for the mean of the assumed exponential family.

3.4.1 Estimation algorithms

Traditional theory supposes that there exists a single iteratively reweighted least squares (IRLS) algorithm capable of fitting all generalised linear models. However, as discussed in Hardin and Hilbe (2001), there are extensions to this viewpoint which allow for a greater range of possible methods.

Specific instances of generalised linear models may be estimated by Newton-Raphson methods. These estimates are more tedious to produce and present

theoretically than IRLS estimates; the latter based on the traditional Fisher scoring approach. It can be shown that generalised linear models can be unified and estimated by a single IRLS algorithm, hence the attraction and use by Venables and Ripley in the R function "glm" contained in the MASS package. There exist conditions under which the methodologies of the Newton-Raphson and IRLS techniques are equivalent. Nonetheless, there are still discrepancies, even where a unique solution exists, due to variation in starting values and convergence paths. Even when not equivalent, the methodologies are equal in the limit. For further discussion see (Hardin and Hilbe, 2001).

To estimate β in (3.20), we wish to maximize the log-likelihood (3.22), which is equivalent, for $l' = \partial l/\partial\beta$, to solving

$$l'(\beta) = 0, \tag{3.27}$$

where

$$\begin{aligned}
\frac{\partial l}{\partial \beta_j} &= \left(\frac{\partial l}{\partial \theta}\right)\left(\frac{\partial \theta}{\partial \mu}\right)\left(\frac{\partial \mu}{\partial \eta}\right)\left(\frac{\partial \eta}{\partial \beta_j}\right), \\
&= \sum_{i=1}^{N}\left\{\frac{y_i - b'(\theta_i)}{a(\phi)}\right\}\left\{\frac{1}{V(\mu_i)}\right\}\left(\frac{\partial \mu}{\partial \eta}\right)_i (x_{ji}), \\
&= \sum_{i=1}^{N}\frac{y_i - \mu_i}{a(\phi)V(\mu_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i x_{ji}.
\end{aligned} \tag{3.28}$$

The Hessian matrix has components

$$\begin{aligned}
\frac{\partial^2 l}{\partial\beta_j\partial\beta_k} &= \sum_{i=1}^{N}\frac{1}{a(\phi)}\left(\frac{\partial}{\partial\beta_k}\right)\left\{\frac{y_i-\mu_i}{V(\mu_i)}\left(\frac{\partial\mu}{\partial\eta}\right)_i x_{ji}\right\},\\
&= \sum_{i=1}^{N}\frac{1}{a(\phi)}\left[\left(\frac{\partial\mu}{\partial\eta}\right)_i\left\{\left(\frac{\partial}{\partial\mu}\right)_i\left(\frac{\partial\mu}{\partial\eta}\right)_i\left(\frac{\partial\eta}{\partial\beta_k}\right)_i\right\}\frac{y_i-\mu_i}{V(\mu_i)}\right.\\
&\quad\left.+\frac{y_i-\mu_i}{V(\mu_i)}\left\{\left(\frac{\partial}{\partial\eta}\right)_i\left(\frac{\partial\eta}{\partial\beta_k}\right)_i\right\}\left(\frac{\partial\mu}{\partial\eta}\right)_i\right]x_{ji},\\
&= \sum_{i=1}^{N}\frac{1}{a(\phi)}\left[\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\left\{\left(\frac{\partial}{\partial\mu}\right)_i\frac{y_i-\mu_i}{V(\mu_i)}\right\}x_{ki}\right.\\
&\quad\left.+\frac{y_i-\mu_i}{V(\mu_i)}\left\{\left(\frac{\partial}{\partial\eta}\right)_i\left(\frac{\partial\mu}{\partial\eta}\right)_i\right\}x_{ki}\right]x_{ji},\\
&= -\sum_{i=1}^{N}\frac{1}{a(\phi)}\left[\frac{1}{\mathrm{V}(\mu_i)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\right.\\
&\quad\left.-(\mu_i-y_i)\left\{\frac{1}{\mathrm{V}(\mu_i)^2}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\frac{\partial\mathrm{V}(\mu_i)}{\partial\mu}-\frac{1}{\mathrm{V}(\mu_i)}\left(\frac{\partial^2\mu}{\partial\eta^2}\right)_i\right\}\right]x_{ji}x_{ki}.
\end{aligned}\tag{3.29}$$

Let $l'' = \partial^2 l/(\partial\beta\partial\beta^T) \in \Re^{p\times p}$. Expanding the estimating equation in (3.27) as a Taylor series leads to

$$l'\left(\beta^{(0)}\right)+\left(\beta-\beta^{(0)}\right)l''(\beta^{(0)})+\frac{\left(\beta-\beta^{(0)}\right)^2}{2!}l'''\left(\beta^{(0)}\right)+\ldots=0.$$

If we assume that β is sufficiently close to $\beta^{(0)}$ then we have

$$l'\left(\beta^{(0)}\right)+\left(\beta-\beta^{(0)}\right)l''(\beta^{(0)})\approx 0,$$

which solving for β gives

$$\beta\approx\beta^{(0)}-\frac{l'\left(\beta^{(0)}\right)}{l''\left(\beta^{(0)}\right)},$$

viewing this iteratively yields

$$\beta^{(r)}=\beta^{(r-1)}-\frac{l'\left(\beta^{(r-1)}\right)}{l''\left(\beta^{(r-1)}\right)},\tag{3.30}$$

for some starting value $\beta^{(0)}$ and subsequent iterations $r=1,2,\ldots$. With respect to the linear regression model (generalised linear model with identity link and Gaussian variance), the linearised Taylor series approximation above is exact, requiring

only $r = 1$ for convergence.

To attain the maximum likelihood (ML) estimates for β via Newton-Raphson methodology, (3.27) is solved through iterative application of (3.30), which incorporates the first (3.28) and second (3.29) derivatives of the log-likelihood. Note that the variance matrix, for the estimates $\hat{\beta}$ of β, is taken to be the inverse of the negative of the observed Hessian matrix.

Alternatively, to get ML estimates of β by IRLS, we begin by rewriting (3.30)

$$\delta\beta^{(r-1)} = -\left\{\frac{\partial^2 l}{\partial\left(\beta^{(r-1)}\right)^T \partial\beta^{(r-1)}}\right\}^{-1} \frac{\partial l}{\partial\beta^{(r-1)}}, \tag{3.31}$$

in which rather than calculate the (negative) Hessian, we use its expectation (also known as the Fisher information for β)

$$-\mathbb{E}\left(\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right) = \sum_{i=1}^{N}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2 \frac{1}{\mathrm{V}(\mu_i)a(\phi)} x_{ji}x_{ki},$$

as we assume $\mathbb{E}\left\{(y_i - \mu_i)\right\} = 0$ (we correctly specify the conditional mean), giving (3.31) as

$$\left\{\sum_{i=1}^{N}\frac{1}{\mathrm{V}(\mu_i)a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2 x_{ji}x_{ki}\right\}\delta\beta^{(r-1)} = \sum_{i=1}^{N}\frac{y_i-\mu_i}{\mathrm{V}(\mu_i)a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i x_{ji}. \tag{3.32}$$

Assuming the linear predictor to be

$$\eta_i^{(r-1)} = \sum_{k=1}^{p} x_{ki}\beta_k^{(r-1)},$$

by pre-multiplying we have

$$\begin{aligned}\left\{\sum_{i=1}^{N}\frac{1}{\mathrm{V}(\mu_i)a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2 x_{ji}x_{ki}\right\}\beta^{(r-1)} \\ = \sum_{i=1}^{N}\frac{1}{\mathrm{V}(\mu_i)a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\left(\eta_i^{(r-1)}\right)x_{ji}.\end{aligned} \tag{3.33}$$

If we sum (3.32) and (3.33) and substitute (3.31) for $\delta\beta^{(r-1)}$ followed by observing

the relationship in (3.30), we see

$$\begin{aligned}\left\{\sum_{i=1}^{N}\frac{1}{\mathrm{V}(\mu_i)a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2 x_{ji}x_{ki}\right\}\beta^{(r)}\\ =\sum_{i=1}^{N}\frac{1}{\mathrm{V}(\mu_i)a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\left\{(y_i-\mu_i)\left(\frac{\partial\mu}{\partial\eta}\right)_i+\left(\eta_i^{(r-1)}\right)\right\}x_{ji}.\end{aligned} \tag{3.34}$$

Let

$$\mathbf{W}^{(r-1)}=\mathrm{diag}\left\{\frac{1}{\mathrm{V}(\mu_i)a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\right\},$$

an $N\times N$ dimensional diagonal matrix, and

$$\mathbf{z}^{(r-1)}=\left\{(y-\mu)\left(\frac{\partial\mu}{\partial\eta}\right)_i+\left(\eta_i^{(r-1)}\right)\right\},$$

a vector of length N; accordingly, (3.34) may be written as

$$\hat{\beta}^{(r)}=\left(\mathbf{X}^T\mathbf{W}^{(r-1)}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{W}^{(r-1)}\mathbf{z}^{(r-1)}, \tag{3.35}$$

assuming that the inverse exists. Suppose that our responses are correctly modelled by the binomial family, now assume the log-likelihood to be log concave (as for both logit and probit links under the binomial family) and the response y_i, for the i^{th} observation, to be within the open interval $(0,1)$. The results of Wedderburn (1976) and Haberman (1977) show the unique maximum to occur at the estimate $\hat{\beta}$, where $\hat{\beta}$ is known to be well defined.

We remark on the similarity of (3.6) and (3.35), where the estimate of β in the latter is obtained via weighted ordinary least squares. As when using the Newton-Raphson methodology, a variance matrix for the estimate of β is required. It is natural under IRLS to take the estimated expected Hessian matrix to be the variance matrix. Note that in the case that the canonical link function of the chosen family is used, the standard errors (square root of the diagonal elements of the variance matrix) constructed from the expected Hessian (IRLS) are equal to those from the observed Hessian (Newton-Raphson). This is as $\theta=\eta$, and hence, (3.25) becomes $\mathrm{V}(\mu)=\partial\mu/\partial\eta$ from which the second term of (3.29) reduces to be zero.

IRLS is said to have converged to the true estimates when the difference in the

deviance

$$D = 2\sum_{i=1}^{N} \left[y \left\{ \theta(y_i) - \theta(\mu_i) \right\} - b\left\{ \theta(y_i) \right\} + b\left\{ \theta(\mu_i) \right\} \right],$$

where $\theta()$ is the canonical parameter and $b\{\}$ the cumulant, from one iteration to another is below some specified tolerance (Hardin and Hilbe, 2001). Note that when the dispersion parameter ϕ is one, the deviance may also be written in terms of log-likelihoods

$$D = 2\phi \left\{ l(y; y) - l(y; \mu) \right\} \tag{3.36}$$

where $l(y; y)$ is the log-likelihood of the full model, and $l(y; \mu)$ that for the fitted model. We remark that those estimates which minimize the deviance also maximize the likelihood.

The benefit of IRLS versus Newton-Raphson is predominantly that IRLS does not require the likelihood to be known, instead a quasi-likelihood implied by the assumed moments may be incorporated. In the instance that the quasi-likelihood is the true likelihood, the estimates found by IRLS are the ML estimates.

3.4.2 Generalised linear models for binary data

The binomial family is the most commonly used other than the Gaussian family with canonical link (linear regression model). This family is suitable for both of our two class categorical events (Sections 2.7.2 and 2.8.2), where the categorical response G is zero-one coded as the binary (negative-positive) response Y. Each observation Y_i can take one of two possible values: 0 or 1. We treat each Y_i as the response of a single Bernoulli trial with $\mathbb{E}(Y_i) = \mathbb{P}(Y_i = 1|X = x_i)$, and denote $\mathbb{P}(Y_i = 1|X = x_i) = \pi(x_i)$ to indicate the dependence on the features $x_i = [x_{i,1}, \ldots, x_{i,p}]^\top$. Consequently, for the i^{th} observation, the variance in the response Y_i is

$$\text{Var}(Y_i) = \pi(x_i)\left(1 - \pi(x_i)\right).$$

Given $\pi(x_i)$, the linear probability model is

$$\pi(x_i) = x_i^\top \beta \tag{3.37}$$

(Agresti, 2013, Chapter 4). This becomes a generalised linear model, with binomial family and identity link function, following the assumption that all observations $(i = 1, \ldots, N)$ are independent. (Note that this assumption is frequently broken when applied to real data.) However, whilst the interpretation of (3.37) is trivial,

the identity link is not viable as it allows the probability measure $\pi(x_i)$ to fall outside of the $[0, 1]$ interval (unless β is restricted). Thus emphasizing what we know from (3.26): that the link function must be chosen such that the feasible interval for $\pi(x_i) = \mathbb{E}(\mathbf{Y}) = \mu$, $[0, 1]$, is mapped to the complete real line. Three of the most commonly used link functions which satisfy this requirement are given here.

- The (canonical) logit (logistic function):

$$g_1(\pi) = \log\left(\frac{\pi}{1-\pi}\right). \tag{3.38}$$

- The probit (inverse normal function):

$$g_2(\pi) = \Phi^{-1}(\pi), \tag{3.39}$$

  where Φ is the normal cumulative distribution function.

- The complementary log-log:

$$g_3(\pi) = \log\left(-\log(1-\pi)\right). \tag{3.40}$$

Derivation of the Bernoulli model

The probability mass function (PMF) for the model with Bernoulli response is

$$f(y;\pi) = \pi^y(1-\pi)^{(1-y)}.$$

Rewriting this in the exponential form of (3.21) we have

$$f(y;\pi) = \exp\left\{y \ . \ \log\left(\frac{\pi}{1-\pi}\right) + \log\left(1-\pi\right)\right\}, \tag{3.41}$$

from which, we can easily match the form against (3.21) to find

$$\theta = \log\left(\frac{\pi}{1-\pi}\right),$$

$$b(\theta) = -\log\left(1-\pi\right), \tag{3.42}$$

and $a(\phi) = 1$ with the dispersion parameter $\phi = 1$. The first and second derivatives of (3.42) are easily found to be:

$$b'(\theta) = \pi$$

$$b''(\theta) = \pi(1 - \pi).$$

Using (3.41), the log-likelihood (3.22) and deviance (3.36) for the Bernoulli response model are

$$l(\pi; y) = \sum_{i=1}^{N} \left\{ y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \log\left(1 - \pi_i\right) \right\},$$

$$D = 2 \sum_{i=1}^{N} \left\{ y_i \log\left(\frac{y_i}{\pi_i}\right) + (1 - y_i)\left(\frac{1 - y_i}{1 - \pi_i}\right) \right\}.$$

In this manner, we can use the general results for fitting the generalised linear model with exponential family to the case where we assume Bernoulli distribution.

Generalised linear model with probit (GLMP) link

Bliss (1934) was the first develop a probit model for the analysis of bioassay data by application of maximum likelihood methodology (Fisher, 1922). Beginning with the definition of the probit link function in (3.39) we have

$$\theta = \log\left(\frac{\pi}{1 - \pi}\right); \ \eta = g(\mu) = \Phi^{-1}\left(\mu\right),$$

from which we can derive the inverse link function

$$g^{-1}(\eta) = \Phi\left(\eta\right),$$

and the corresponding first derivative

$$g'(\mu) = \frac{1}{\phi\{\Phi^{-1}(\mu)\}}.$$

Above, Φ is the standard normal cumulative distribution function, and ϕ is the standard normal density function.

Generalised linear model with logit (GLML) (canonical)

The logit link function originated when Berkson (1944) formulated an alternative model for the bioassay data previously modelled by Bliss (1934) with the probit link. Given that the logit link (3.38) is the canonical link function for the assumed Bernoulli family (and more generally the binomial family), we know $\theta = \eta$. Using this identity in conjunction with (3.41) it is straightforward to see

$$\theta = \log\left(\frac{\pi}{1-\pi}\right) = \eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right).$$

The inverse of the canonical link is then

$$g^{-1}(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)} = \frac{1}{1+\exp(-\eta)},$$

and the first derivative of the link function is

$$g'(\mu) = \frac{1}{\mu(1-\mu)}.$$

Generalised linear model with complementary log-log (GLMCLL) link

Use of the complementary log-log link function can be traced back to Fisher (1922) analysing dilution assay data. However, unlike the probit and logit link functions, the complementary log-log link function has not yet found popularity. Assuming the complementary log-log link function as given in (3.40), we have

$$\theta = \log\left(\frac{\pi}{1-\pi}\right);\ \eta = g(\mu) = \log\left(-\log(1-\mu)\right),$$

with the corresponding inverse link function

$$g^{-1}(\eta) = 1 - \exp(-\exp(\eta)),$$

and first derivative easily found to be

$$g'(\mu) = \frac{1}{(\mu-1)\log(1-\mu)}.$$

Model output

Given that each binary response Y_i is treated as the outcome of a single Bernoulli trial, such that $\mathbb{E}(Y_i) = \mathbb{P}(Y_i = 1|X = x_i) = \pi(x_i)$, the generalised linear model for binary response data is a probabilistic classifier (independent of link function). That is, letting $\hat{\beta}$ be the parameter estimates of the fitted model, predictions of $\mathbb{P}(Y_i = 1|X = x_i)$ are provided as

$$\hat{p}(x_i) = \hat{\mathbb{P}}(Y_i = 1|X = x_i) = g^{-1}(x_i^\top \hat{\beta}). \tag{3.43}$$

A classification for the zero-one coded dummy response is then determined according to some discrimination threshold $d \in (0, 1)$ on (3.43)

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{p}(x_i) \geq d, \\ 0 & \text{if } \hat{p}(x_i) < d. \end{cases} \tag{3.44}$$

3.5 Model Comparison

Within applied research, we are concerned with the notion of performance. Here, this relates to the compatibility of the studied data with the underlying assumptions for the optimality of the considered classifier(s). In practice, it is rare for models to be perfectly in-line with the theoretical construction of the method used. In this section, we attempt to provide an overview of theoretical differences between the classification methods outlined earlier in this chapter.

We begin by analysing the choice of link function for generalised linear models in Section 3.5.1. In Section 3.5.2, we compare the generalised linear model with logit link to linear discriminant analysis (given that the two differ only in the estimation procedure for the parameters) and finally (in Section 3.5.3), we describe the main difference between the only quadratic classification method and all others. These observations aid the later analysis of classifier performance (Chapter 6) via the methods of Chapter 5, in those circumstances laid out in Chapter 2.

3.5.1 Generalised linear model: choice of link function

In choosing to fit a model, we implicitly assume that the structure of the classifier is capable of expressing the relationship between the chosen features and the binary response for each observation. With respect to generalised linear models,

we assume that the link function for the chosen family is specified appropriately.

Recall that we assume Bernoulli distribution on the modelled response. In Section 3.4.2 we described the corresponding three most commonly used link functions, which are summarised here in Table 3.1. The most immediate difference between these three functions is that, whilst the first two are symmetric, the third is skewed to the left (see Figure 3.1). However, Figure 3.1 illustrates that for very small values of μ, the logit and complementary log-log functions are similar; whilst for values of μ near 0.5, the probit and logit links are similar.

Table 3.1: Comparison of the the link functions considered for the binomial generalised linear model.

Link	Distribution	Shape	Link function $\eta = g(\mu)$
probit	Normal	symmetric	$\Phi^{-1}(\mu)$
logit	logistic	symmetric	$\log\left(\frac{\mu}{1-\mu}\right)$
complementary log-log	Weibull (min)	left skew	$\log(-\log(1-\mu))$

The probit and logit functions are both symmetric about $\mu = 0.5$ and produce extremely similar results over a sample of μ (see Figure 3.1), unless very many of the observations lie in the tails (i.e. μ is frequently close to zero or one within the sample). Given the close similarity, we expect the generalised linear models with probit and logit link to fit approximately equally well. There is often no discernible difference in performance except if the size of the sample, N, is very large. Note that with respect to our risk event formulations in Chapter 2, the sample sizes are small.

Due to the relative indifference in classifier performance between the generalised linear models with probit and logit link functions, preference is typically based on the ease of interpretation. The logit link function provides a direct interpretation as the log-odds of event occurrence (success). In addition, as the logit link is the canonical link under the Bernoulli assumption, it is mathematically convenient when computing model parameter estimates. The probit link function (see Table 3.1) conversely involves Φ, which does not have a closed form and the corresponding model requires numerical integration when computing the parameter estimates by maximum likelihood. Nonetheless, as the probit function scales μ to some normal cumulative distribution function, this link is more tractable when we want to apply Bayesian inference on the model. The relationship to a normal cumulative distribution function means that a special case of Markov Chain Monte Carlo may be used that is simpler to compute.

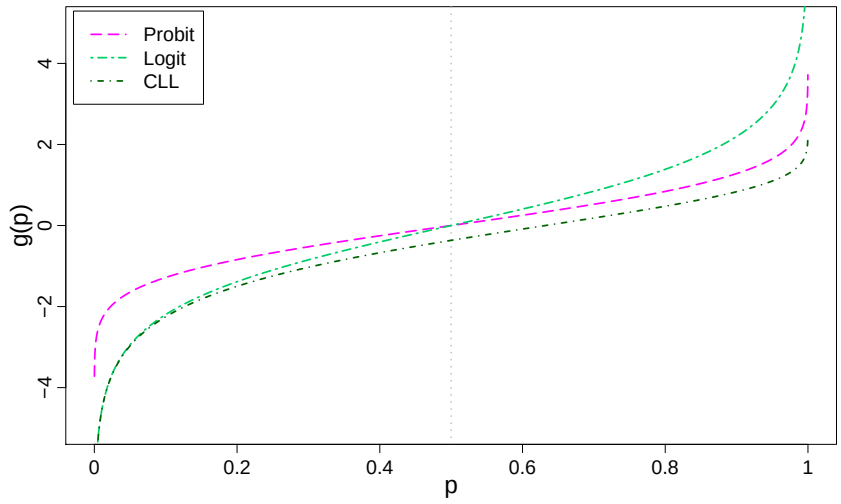


Figure 3.1: Comparison of the link functions considered for binomial generalised linear model, where CLL is used to denote the complementary log-log function.

3.5.2 Linear discriminant analysis versus generalised linear model with logit link

Linear discriminant analysis takes the same class posterior as the generalised linear model with logit link, but differs in the estimation procedure (Venables and Ripley, 2002, Chapter 12). The linear discriminant classifier is unsuited where homoscedasticity is an inappropriate assumption; the generalised linear model allows this, among other relaxations. Nonetheless, where the normality assumption of linear discriminant analysis is satisfied (i.e. the features follow a normal distribution conditional on the response y), this classifier is generally preferred to the generalised linear model with logit link (Agresti, 2013). This is primarily due to its superior efficiency arising from the utilisation of the distributional information about the design matrix $\mathbf{X}$. This superiority becomes considerable for more widely separated groups (Efron, 1975). However, as it is frequent for the features to break this normality assumption, Venables and Ripley (2002) typically prefer the generalised linear model.

To avert issues with the features not being from a multivariate normal, we have standardised our design matrix $\mathbf{X}$. In addition, extreme values in the design matrix $\mathbf{X}$ can have a considerable effect on the model when using discriminant analysis, but less so for the generalised linear model with logit link. Consequently, the generalised linear model with logit link has a broader scope by making fewer assumptions on the features included in the design matrix and being more robust to their eccentricities. A further benefit of using the generalised linear model with

logit link over linear discriminant analysis is the direct interpretation of features via the log-odds ratio.

3.5.3 Quadratic versus linear classifiers

Quadratic discriminant analysis is the only classifier considered which gives a non-linear decision boundary. The use of any other classifier given within this chapter assumes a linear decision boundary to be suitable, that is, that the data are linearly separable by class. Whilst linear decision boundaries have low variance, they typically exhibit high bias (Hastie et al., 2009, Chapter 2). Part of the appeal of linear classifiers is that they do not require vast amounts of data to provide a fitted model. In practice, the choice of model can be decided by the trade-off between bias and variance.

3.6 Conclusion

We are not aware of any standard set of descriptors used to characterise the different theoretical details of the classification methods that we have considered in this chapter. Therefore, it is not easy to directly compare them in a tabular fashion. However, we have created Table 3.2 from the material presented in Section 3.5 to provide a summary understanding of the relative strengths and weaknesses of the classification methods in the context of our research. We have outlined the theoretical details of those classification methods used which do not incorporate Bayesian inference within this chapter. Consequently, the reader should be sufficiently informed and able to interpret the implementation of all relevant methods within the R library MASS.

One issue that became apparent as the ROBUST project progressed was that the volume of data available from the ROBUST use-case partners would be much less than expected and highly variable over time (see Section 2.3). Hence we became aware of a requirement to support flexibility. Furthermore, computational efficiency is required to allow for real-time analysis across a range of partner host platforms. By inspecting Table 3.2, the linear discriminant analysis method appears to be most likely to meet both these requirements but, the table also shows that different classification methods may prove superior in communities or fora with different characteristics.

Comparison of these classification methods enhanced our empirical analysis in

Table 3.2: Summary comparison of methods.

Capability	Linear Discriminant Analysis	Quadratic Discriminant Analysis	Generalised Linear Model
Model Acronym	LDA	QDA	GLM
Discrete Responses Supported	Yes	Yes	Yes (with binomial family)
Linear decision boundaries	Yes	No	Yes
Similar to	Binonmial GLM with logit link	—	Binomial GLM with logit link similar to that with probit link
Relative data requirements	Low	High	Low
Relative ease of interpretation	High	Low	High for Binomial with logit link

Chapter refch:results and also, demonstrated that applying Bayesian inference to generalised linear models is significantly easier for the probit link function (assuming Binomial distribution for the event response). This naturally leads to the following chapter, where we provide the associated theoretical knowledge for applying Bayesian inference to the generalised linear model with probit link, as well as the implementation made for ROBUST.

Chapter 4

Bayesian Approach to Classification

The primary objective of this chapter is to provide the theoretical knowledge from the literature that was required to code the generalised linear model with probit link function using Bayesian inference as a classifier. The secondary objective of this chapter is to verify and validate our implementation that was coded for ROBUST.

4.1 Introduction

To our knowledge, no one has previously performed classification for online communities by generalised linear models with Bayesian inference before (Hiscock et al., 2013).

The foundation of Bayesian statistics is in its presentation of uncertainty as probability (Iversen, 1984). Tanner (1996) provides the earliest comprehensive treatment of Bayesian computation. Let $\mathbf{y}$ be the vector of observations of the random variable $\mathbf{Y}$ whose distribution is parametrised by the vector of parameters θ. With $f(\theta)$ as the prior distribution of the parameter vector θ, define $f(\theta|\mathbf{y})$ to represent the posterior distribution of the parameter vector θ given the observations $\mathbf{y}$; and the likelihood of the response $\mathbf{y}$ given the parameter vector θ to be $f(\mathbf{y}|\theta)$. For inference about the parameters θ Bayes' Theorem states:

$$f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)f(\theta)}{\int f(\mathbf{y}|\theta)f(\theta)d\theta}. \tag{4.1}$$

We see the posterior distribution is therefore formed from initial, or prior, beliefs about the values of the considered parameter(s) that are updated by the data. With the denominator in (4.1) acting as the normalising constant on the posterior distribution, Bayes' Theorem may be interpreted as the unnormalised joint posterior being proportional to the likelihood times the prior (4.2).

$$f(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)f(\theta) \tag{4.2}$$

The posterior is then the prior updated by the likelihood. We note that the term joint posterior comes from there being multiple parameters within the vector θ. The posterior of each individual parameter is called the marginal posterior distribution.

Whilst the prior distribution may be chosen to reflect the beliefs about the behaviour of the parameter vector θ, in practice it is common for a uniform (flat and uninformative) prior to be used (Box and Tiao, 1992). If we take the prior distribution to be a constant, the posterior is proportional to the likelihood. The use of a uniform prior on the vector of parameters θ, implies the belief that the value of the parameter θ_j is equally likely to take any value on the real line between two specified bounds a_j and b_j. That is, we are confident that the parameter θ_j lies somewhere in the region $[a_j, b_j]$, but believe no value within this range to be any more probable than any other. Such a prior does not represent complete ignorance, given its rectangular shape, but is uninformative having the lowest possible level of prior discrimination before the parameter values. There are instances when the assumption of a "flat", that is uninformative, prior for the parameters θ can lead to the posterior distribution being improper (Hobert and Casella, 1996; Natarajan and McCulloch, 1995). An improper prior is one which does not integrate to a finite number (Natarajan and McCulloch, 1998, p. 267). However, suitable model choice in addition to the use of an appropriate uninformative prior can allow the data to highlight dubious aspects about the model especially for small sample sizes (e.g. the extent of sensitivity to departures from normality on inferences about location parameters).

4.2 Bayesian Computation

Given Bayes' formula for the joint posterior of the parameter vector $\theta \in \mathcal{R} \subseteq \Re^p$ given the observations $\mathbf{y}$ (4.1), we see we must be able to calculate the normalising

constant, often referred to as the marginal likelihood,

$$\int f(\mathbf{y}|\theta)f(\theta)d\theta. \tag{4.3}$$

A common issue is (4.3) not being explicitly available due to complexity in high dimensionality. Where θ is p-dimensional we must integrate over the p-dimensional parameter space. Markov chain Monte Carlo methods are a numerical methods capable of computing the complex integrals and making inference about the parameters θ. Bayesian computation is used here as an alternative to the maximum likelihood methods used earlier; there are similarities but the Bayesian methods are typically much more computationally expensive given their iterative nature.

Bayesian computation requires the estimation of posterior distribution, in addition to various summary measures (e.g. the mean of the posterior distribution); given an assumed prior distribution. To do so requires the computation of the integral of some function of the posterior distribution . The mean of the posterior distribution for instance may be found by computing

$$\mathbb{E}(\theta) = \int \theta f(\theta|\mathbf{y})d\theta.$$

Such a calculation can be difficult, and is usually impossible to determine analytically. There are many methods developed to overcome this issue, the one used in this thesis is a particular instance of a special case of the Markov Chain Monte Carlo (MCMC) Metropolis algorithm (Barker, 1965; Chib and Greenberg, 1995) called the Gibbs sampler (Geman and Geman, 1984). Other methods include asymptotic approximations and numerical integration.

4.2.1 Markov chain Monte Carlo (MCMC)

Metropolis et al. (1953) were the first to publish on Markov Chain Monte Carlo (MCMC) sampling. MCMC integration is based on the elementary statistical theory that given sufficient samples generated from some unknown distribution, features of the distribution may be estimated. The *target distribution* is the joint posterior distribution $f(\theta|\mathbf{y})$, which itself, is known only up to some multiplicative constant as shown in (4.2). Let $f(\theta|\mathbf{y})$ be sufficiently complex such that it cannot be sampled from directly. Smith and Roberts (1993) show that we may indirectly sample from $f(\theta|\mathbf{y})$ by constructing an irreducible and aperiodic Markov chain in state space $\mathcal{R}$ with stationary distribution $f(\theta|\mathbf{y})$. Upon completion of an initial

transient phase, those subsequent values of an "appropriately" behaved Markov chain, can be viewed as dependent samples from the target distribution and used to summarise useful features of the joint posterior. See Roberts and Smith (1994) for a discussion on "appropriate" behaviour.

Take $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}$ to be an identically distributed sample from the posterior distribution $f(\theta|\mathbf{y})$. The estimated expectation for some function $g(\theta)$ may be found as:

$$\mathbb{E}\left(g(\theta)\right) \approx \frac{1}{M} \sum_{m=1}^{M} g\left(\theta^{(m)}\right). \tag{4.4}$$

Where $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)}, \ldots$ are samples from a Markov chain under suitable regularity conditions there exist asymptotic results which include:

$$\theta^{(M)} \xrightarrow[M\to\infty]{d} \theta \sim f(\theta|\mathbf{y}); \quad \frac{1}{M} \sum_{m=1}^{M} g\left(\theta^{(m)}\right) \xrightarrow[M\to\infty]{} \mathbb{E}_{f(\theta|\mathbf{y})}\left\{g(\theta)\right\} \text{ almost surely}$$

(Smith and Roberts, 1993).

Independently sampling from $f(\theta|\mathbf{y})$ is in practice troublesome. Generally, the sample is either dependent and/or is from a distribution different to $f(\theta|\mathbf{y})$. Whereas importance-sampling (Geweke, 1989; Stewart, 1979; Zellner and Rossi, 1984) uses weighted independent samples from a distribution similar to the posterior, the Gibbs sampler (Gelfand and Smith, 1990) uses dependent samples with equilibrium distribution the same as that of the posterior (Tierney, 1994).

The iterative Markov process, where the parameters at the m^{th} stage $\theta^{(m)}$ depend on those of the previous stage $\theta^{(m-1)}$, may be run until equilibrium is approximately reached at iteration M_0 — M_0 is referred to as the end of *burn-in*. Taking an additional M samples gives the dependent sample of parameter vectors $\theta^{(M_0)}, \theta^{(M_0+1)}, \ldots, \theta^{(M_0+M)}$ from the unknown posterior distribution $f(\theta|\mathbf{y})$. The respective sizes of M_0 and M are chosen such that the final M samples are representative of $f(\theta|\mathbf{y})$. To estimate various summaries of the posterior, (4.4) can be applied over the sample of size M.

In implementing an MCMC procedure, an ever present issue is the convergence of the Markov chain to its stationary distribution that is the same as the posterior (Banerjee et al., 2004). Firstly, convergence diagnostics are required to decide when the burn-in is complete so that we summarise only that part of the Markov chain with desired distribution. Secondly, we must estimate the Monte Carlo variances of the posterior estimates to determine the quality of those estimates

produced from the MCMC procedure, that is, whether the estimates of the chain cover all of the space.

4.2.2 The Gibbs sampler

The Gibbs sampler is a special case of Barker (1965)'s variation of the Metropolis-Hastings algorithm. The sampler was introduced by Geman and Geman (1984), without apparent knowledge of previous work on the Metropolis-Hastings algorithm, when utilising optimisation to find the posterior mode of a Gibbs random field. However, a similar methodology described as data augmentation was published in (Tanner and Wong, 1987), also without apparent knowledge of previous similar work. The potential of the Gibbs sampler to compute a particular numerical characteristic was not widely realised until Gelfand and Smith (1990) compared the three proposed alternatives: Gibbs sampler introduced by Geman and Geman (1984); data-augmentation prescribed by Tanner and Wong (1987); and importance-sampling described by Rubin (1987).

Given the vast literature on MCMC and the Gibbs sampler, we give a brief overview of the Gibbs method here and refer the interested reader to (Casella and George, 1992) for an introduction; and to (Robert and Casella, 2004, Chapters 9 and 10) for a thorough treatment.

Let the joint density $f(\theta|y)$ of the p-dimensional random variable θ be unknown. To make estimates of some summary inferences of the joint posterior distribution expectations we can simulate observations from the univariate full conditional densities $f_j(\theta_j|\theta_{-j}, \mathbf{y})$, where $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)$. The troublesome normalising constant of the joint density (4.3) does not appear in the full conditional densities. Where the joint posterior density is proper, that is it integrates to a finite number (Natarajan and McCulloch, 1998, p. 267), the Hammersley-Clifford theorem under light conditions gives us that the joint posterior density is perfectly summarised by the set of fully conditional densities (Davison, 2003, Section 11.3.3). (However it does not hold that any set of full conditional densities implies a proper joint posterior density to exist.) A Gibbs sampler thus takes successive samples from $f_j(\theta_j|\theta_{-j}, \mathbf{y})$ as prescribed in Algorithm 1. Notice that the algorithm only requires simulation from univariate densities and consequently does not suffer in high dimension.

To formulate the Gibbs sampler as a special case of the Metropolis-Hastings algorithm, take the proposed density at the j^{th} step of the iteration of the Metropolis-

Algorithm 1: Gibbs sampler

Result: Dependant sample $\theta^{(M_0)}, \theta^{(M_0+1)}, \ldots, \theta^{(M_0+M)}$ with equilibrium distribution the same as that of the desired posterior.

begin

Choose initial estimates $\theta^{(0)}$ of the parameter vector θ.

for $m \leftarrow 1$ **to** $(M_0 + M)$ **do**

Sample $\theta_1^{(m)}$ from $f_1(\theta_1|\theta_2^{(m-1)}, \theta_3^{(m-1)}, \ldots, \theta_p^{(m-1)}, \mathbf{y})$.

Sample $\theta_2^{(m)}$ from $f_2(\theta_2|\theta_1^{(m)}, \theta_3^{(m-1)}, \theta_4^{(m-1)}, \ldots, \theta_p^{(m-1)}, \mathbf{y})$.

$\vdots$

Sample $\theta_j^{(m)}$ from $f_j(\theta_j|\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_{(j-1)}^{(m)}, \theta_{(j+1)}^{(m-1)}, \ldots, \theta_p^{(m-1)}, \mathbf{y})$.

$\vdots$

Sample $\theta_p^{(m)}$ from $f_p(\theta_p|\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_{p-1}^{(m)}, \mathbf{y})$.

Hastings algorithm to be

$$q(\theta^*|\theta) = \begin{cases} f_j(\theta_j^*|\theta_{-j}) & \text{where } \theta_{-j}^* = \theta_{-j}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

The acceptance probabilities for the proposed move from θ to θ^* at the j^{th} step of the iteration for the Gibbs sampler can be shown to be

$$\begin{aligned} \alpha(\theta^*, \theta) &= \min\left\{\frac{f(\theta^*)}{f(\theta)}\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}, 1\right\} \\ &= \min\left\{\frac{f(\theta^*)/f(\theta_j^*|\theta_{-j})}{f(\theta)/f(\theta_j|\theta_{-j}^*)}, 1\right\} \\ &= \min\left\{\frac{f(\theta^*)/f(\theta_j^*|\theta_{-j}^*)}{f(\theta)/f(\theta_j|\theta_{-j})}, 1\right\} \\ &= \min\left\{\frac{f(\theta_{-j}^*)}{f(\theta_{-j})}, 1\right\} \\ &= 1 \end{aligned}$$

where $\theta_{-j}^* = \theta_{-j}$. Thus, given that all proposals have $\theta_{-j}^* = \theta_{-j}$, the proposed moves are always accepted. As a result, the convergence assessment of the Gibbs sampler can differ to that of the Metropolis-Hastings algorithm.

Where the majority of the full conditionals for the unknown parameters θ are from standard distributions (e.g. normal or gamma), use of the Gibbs sampler is appropriate (Brooks, 1998). This may involve the adoption of a hybrid approach,

including a Metropolis step in the sampling algorithm for those parameters whose conditional posterior distributions are of non-standard form. It can be shown that, under mild conditions, the complete set of full conditional distributions can determine the joint posterior distribution $f(\theta|\mathbf{y})$ uniquely and consequently also all marginal posterior distributions $f(\theta_j|\mathbf{y})$ for $j = 1, \ldots, p$ (Banerjee et al., 2004).

4.2.3 Convergence diagnostics

Formal convergence diagnostics exist, but in practice, most users of MCMC use informal graphical methods to check that their MCMC samplers are performing as required. The most typically used graphic is the trace plot which indicates how well the MCMC is mixing; i.e. exploring the parameter space.

Given the not inconsiderable number of models we fit, in addition to our aim to provide an automated model for each risk formulation, we deem it unrealistic to use graphical methods. We choose to allow the sampler a maximum burn-in period of 100,000 iterations and retain 10,000 iterations. Every 1,000 iterations in the transient phase, we take the mean and standard deviation of the simulated Markov chains for the individual parameters and calculate the absolute difference with that of the proceeding 1,000 iterations. If the average of these differences is not greater than a specified ϵ, of $0.01/p$ (for p the number of parameters) for the mean, and 0.1 for the standard deviation, we assume the chains have reached their steady states and end the transient phase. Empirical results (see Chapter 6) show the retained chains to provide estimates similar to the maximum likelihood estimates for that classifier defined in Section 4.3.

4.3 Bayesian Probit Regression Model (BP) for Binary Response

The Bayesian probit model is similar to the Generalised linear model (Section 3.4) in how it ensures the response Y stays in the interval $[0, 1]$. As in Section 3.4.2, assume the response variable Y to have Bernoulli distribution. Now, as for the GLMP model, take the link function between the linear predictor η and the binary Y as the probit link. Thus we may express the assumed distribution for the binary response Y as

$$Y \overset{\text{i.d.}}{\sim} \text{Ber}(\Phi(\eta)). \tag{4.6}$$

Let Z_i be the latent continuous variable corresponding to the i^{th} zero-one coded response Y_i, for $i = 1, \ldots, N$ where N is the number of observations. Now let this latent variable be independently normally distributed

$$Z_i \overset{\text{i.d.}}{\sim} \mathcal{N}(\eta_i, \sigma), \tag{4.7}$$

from which we may rewrite (4.6) as

$$Y_i = \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{if } Z_i \leq 0. \end{cases} \tag{4.8}$$

For the distribution on the error terms of Z_i to be standard normal, we must take σ in (4.7) equal to one.

Given (4.8) with Z_i as in (4.7), and having previously defined $\eta = \mathbf{X}\beta$, an inference problem about the coefficient vector β arises. Frequentist treatment of the model defined thus far, would result in the previously defined GLM with probit link. However, here we take a Bayesian approach to inference following (Albert and Chib, 1993), simulating from the exact posterior distribution of the unknown β given some prior probability density function $\pi_0()$ on β.

Let $(\beta, \mathbf{Z}) = [\beta_1, \ldots, \beta_p, Z_1, \ldots, Z_N]^\top$ be the $(p + N)$ length random vector. The posterior of $(\beta, \mathbf{Z})$ can be written

$$\pi(\beta, \mathbf{Z}|\mathbf{y}) \propto \pi_0(\beta) \prod_{i=1}^{N} \left[y_i \cdot 1(Z_i > 0) + (1 - y_i) \cdot 1(Z_i \leq 0) \right] \times \phi\left(Z_i; \eta_i\right) \tag{4.9}$$

assuming this is integrable; where

$$\phi\left(Z_i; \eta_i\right) \propto \exp\left(-\frac{(Z_i - \eta_i)^2}{2} \right)$$

is the Gaussian density with mean η_i, variance 1; and $1()$ is the indicator function. More concretely, letting $x = (\beta, \mathbf{Z})$ and taking μ as the $(p + N)$-dimensional Lebesgue measure, (4.9) is (a version of) the density of a probability measure on $\mathbb{R}^{(p+N)}$ with respect to μ only where $C \overset{\text{def}}{=} \int_{\mathbb{R}^{p+N}} \pi(x)\mu(\mathrm{d}x)$ is finite.

We follow the most common practice taking a flat uniform prior for π_0. Therefore, all points in $\mathbb{R}^p$ are, essentially, "equally likely". As the support is unbounded and the intended "density" is a positive constant, this π_0 does not define a probability measure on $\mathbb{R}^p$ and is hence *improper*. This is not a problem where (4.9)

defines a probability measure. Thus the target of inference, π_β, is the resulting β-marginal of (4.9).

The Bayesian probit method for binary response data described by Albert and Chib (1993) utilizes Gibbs sampling. Whilst the conditional distributions of the target (4.9) are easy to sample from (Albert and Chib, 1993), the level of ease depends on the choice of π_0. Under uniform prior, Albert and Chib (1993, Section 3.1) show the fully conditional densities of β and Z_i to be:

$$\beta \mid y, \mathbf{Z} \sim N_K\left(\hat{\beta}_{\mathbf{Z}}, \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right),$$

where $N_K(\mu, \Sigma)$ is the multivariate normal distribution with location (mean) μ and covariance Σ;

$$\hat{\beta}_{\mathbf{Z}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Z};$$

and

$$Z_i \mid y, \beta \overset{\text{i.d.}}{\sim} \begin{cases} N\left(x_i^\top \beta, 1\right) \text{ truncated at the left by } 0, & \text{if } y_i = 1, \\ N\left(x_i^\top \beta, 1\right) \text{ truncated at the right by } 0, & \text{if } y_i = 0. \end{cases}$$

These conditionals can be easily simulated. Simulation algorithms are given in (Devroye, 1986). In addition, Albert and Chib (1993) gives integrable, that is *proper*, possibilities for the prior π_0 which directly enable Gibbs sampling. Given that we assume a uniform prior distribution for the regression coefficients, related issues of the propriety of the posterior distribution are studied by Chen and Shao (1999). We take the initial state of the Markov chain for β to be the least squares estimate as given in (3.6) with $z = -1$ if $y = 0$ and $z = 1$ if $y = 1$.

For the i^{th} observation, with $i = 1, \ldots, N$, given the feature vector $x_i^\top$, the posterior mean of Y_i is

$$\mathbb{P}_{\pi_\beta}(Y_i = 1) = \mathbb{P}_{\pi_\beta}(Z_i > 0) = \mathbb{E}_{\pi_\beta}[\Phi(x_i^\top \beta)]$$

where $\mathbb{P}_{\pi_\beta}$ and $\mathbb{E}_{\pi_\beta}$ denote the probability and expected value with respect to π_β. Assuming certain conditions, a consistent estimator of this mean is the corresponding sample average of the Gibbs sample; where by consistency we mean convergence in probability to the correct value as the sample size tends to infinity (Robert and Casella, 2004, Theorem 6.63); (Cappé et al., 2005, Theorem 14.2.53)). Assume M_0 iterations are required for convergence to the true posterior and M subsequent iterations are made. Hence, given the retained sample

$\{\beta^{(M_0)}, \beta^{(M_0+1)}, \ldots, \beta^{(M)}\}$, with M sufficiently large,

$$\frac{1}{M}\sum_{m=M_0}^{M}\Phi(x_i^\top \beta^{(m)})$$

is an appropriate estimator of $\mathbb{P}_{\pi_\beta}(Y_i = 1)$.

4.3.1 Implementation for ROBUST

The Bayesian probit regression model previously detailed is the only classifier presented within this thesis to have been coded by the author in both the R and Java languages. It is this classifier which was provided as a Java project to be used in the risk management framework to provide estimates of future risk. Therefore, we use the vaso-constriction dataset first modelled by Finney (1947) to demonstrate the correctness of our implementation of this model in comparison with those results given in (Albert and Chib, 1993). The response Y is a record of whether or not vasorestriction on the skin occurs with the features being the volume of inspired air and the corresponding rate of inspiration. Figure 4.1 shows the shape of the density estimates achieved for all model parameters. We see the locations of all three parameter density estimates to be extremely similar to the corresponding results in (Albert and Chib, 1993), in addition to noticing that our densities appear skewed as was also found by Albert and Chib (1993). The trace of each parameters Markov chain post transient period is given in Figure 4.2. From this we conclude that all chains appear to be mixing well (i.e. covering the state-space well) and that our implementation is verified.

4.4 Comparison of Bayesian and other methods

We remark that, whilst it is possible for Bayesian models with flat or noninformative priors to mimic their maximum likelihood counterparts, that is not always the case. Where the prior is proper, the posterior will also be proper (McCulloch et al., 2001, p. 56). When the chosen noninformative prior leads to an improper posterior it is possible to retrieve the maximum likelihood estimates (Natarajan and McCulloch, 1995). Natarajan and McCulloch show that with a proper prior which is diffuse, the posterior modal estimates can be different to the corresponding maximum likelihood estimators — even for large sample sizes.

Given the use in Section 4.3 of a flat prior, it is possible to get the maximum

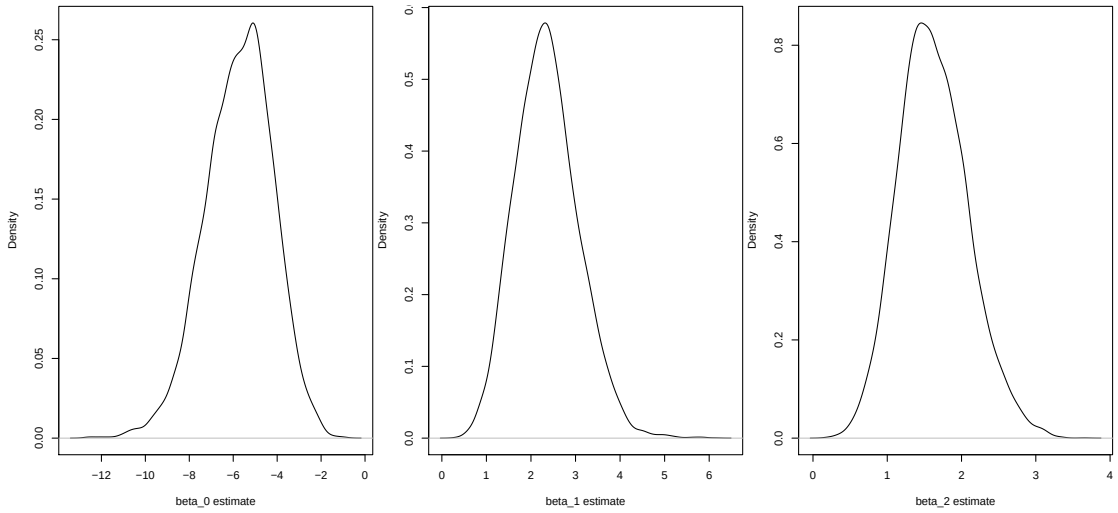


Figure 4.1: Density plot of the Markov chains for the parameters of Finney's Vaso-constriction dataset under our implementation of the Bayesian probit model in R.

likelihood estimators from the mean of the simulated posterior. However, the Bayesian approach gives us more, it gives us a simulation of the entire posterior distribution meaning that credible intervals may be found. We see the similarity between the estimates from the generalised linear model with probit link function (Section 3.4.2) and the Bayesian probit model (Section 4.3) as confirmation that the Gibbs sampler has converged to the true posterior, which gives importance to the credible intervals available through the Bayesian probit model.

4.5 Conclusion

We address our first objective for this chapter by introducing the essential details underlying inference via Bayesian computation and exposing the Bayesian probit regression model for binary response data, given the knowledge of generalised linear models provided by Chapter 3. The second objective of this chapter describing the implementation made for ROBUST, was addressed by comparing the output of our implementation in R against well known results from the literature. Consequently, we are assured that, where the assumptions of the model hold, our implementation in R, and inherently Java, is satisfactory. More details about our ROBUST related effort, and how our Java implementation of the Bayesian probit model fits into the final project deliverable, are given in Chapter 6.

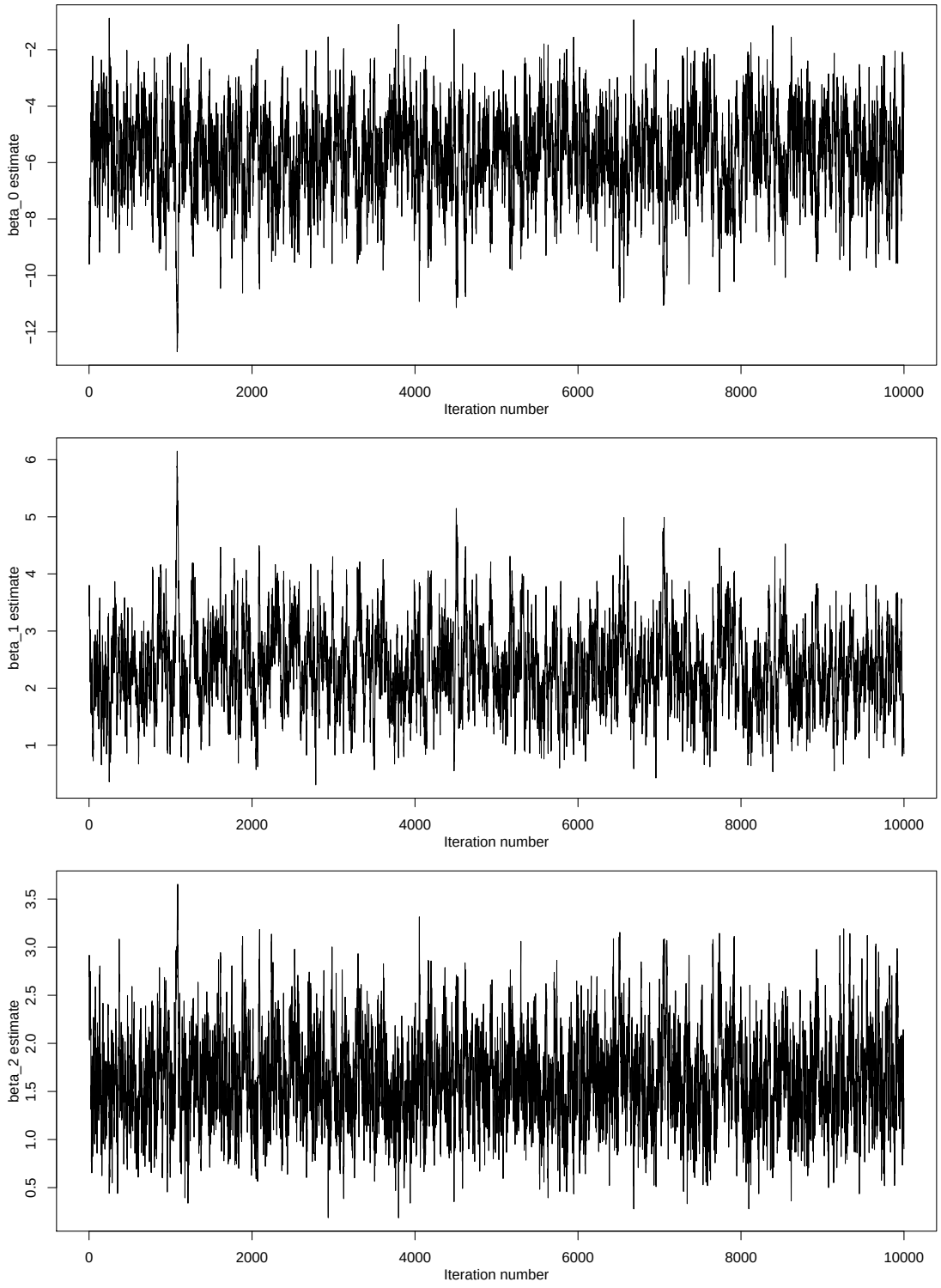


Figure 4.2: Trace of the Markov chains for the parameters of Finney's Vaso-constriction dataset under our implementation of the Bayesian probit model in R.

Chapter 5

Classification Quality Characteristics

The main objective of this chapter is to present the theory of the most popular classification quality characteristic measures and thus demonstrate why some are more suitable for analysing and comparing classifier performance than others. This chapter therefore directly addresses the third objective of this thesis: "how to best analyse and compare classifiers considering both graphical and scalar metrics".

5.1 Introduction

Given the categorical response set $\mathcal{G}$ has cardinality of two in all events considered (see Chapter 2), the discrete valued response Y is always a zero-one coded binary response. We therefore use classification methods to model the response Y. Here the state $Y = 0$ ($Y = 1$) corresponds to the state negative (positive).

The majority of classifiers outlined in Chapters 3 and 4 are probabilistic classifiers; that is, they output a probabilistic estimate, denoted as $\hat{\mathbb{P}}(Y_i = 1|X = x_i)$, for the expectation of Y. The estimate $\hat{\mathbb{P}}(Y_i = 1|X = x_i)$ is an estimate of the posterior probability that the i^{th} observation is positive ($Y_i = 1$), it hence lies in the interval $(0, 1)$. A subset of classification performance measures requires observations to be assigned to a state/class. Therefore, let $d \in (0, 1)$ be some discrimination threshold on $\hat{\mathbb{P}}(Y_i = 1|X = x_i)$ such that our prediction of the binary Y_i is

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{\mathbb{P}}(Y_i = 1|X = x_i) \geq d, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

In the past, it has been typical to take d as 0.5, this being the midpoint of the possible values. However, in the more recent literature (such as (Hernández-Orallo et al., 2012)), research is aimed at adapting the choice of the threshold d to performance measures. This raises the question: how do we best analyse and compare classifiers?

There are broadly speaking two approaches to analysing the performance of a classifier: numerical and graphical. In what follows, we consider some of the most popular measures in the literature and show how (the less used) graphical measures are significantly more understandable and interpretable than (the more common) scalar metrics.

The structure of this chapter begins with explicitly defining loss and afterwards gives an outline of the confusion matrix and associated scalar performance measures. After explaining the significant pitfalls of using the confusion matrix in practice, we define the receiver operating characteristic space and describe how both one- and two-dimensional performance measures can be derived. Following this, we consider the one- and two-dimensional measures in cost space prior to giving a definition of the Brier score (or mean squared error). We close this chapter with a conclusion which summarises our findings and specifies the measures that we use later in Chapter 6 to analyse classifier performance.

5.2 Loss Function

Let $L(\mathcal{G}_k, \mathcal{G}_\ell)$ represent the loss incurred by classifying an observation into class ℓ when it is of class k. The zero-one loss function is most commonly used

$$L(\mathcal{G}_k, \mathcal{G}_\ell) = \begin{cases} 0 & \text{if } \mathcal{G}_k = \mathcal{G}_\ell, \\ 1 & \text{if } \mathcal{G}_k \neq \mathcal{G}_\ell, \end{cases} \tag{5.2}$$

for $k, \ell \in \{1, \ldots, K\}$, where K is the cardinality of $\mathcal{G}$. For those risk events formulated in Chapter 2 as classification tasks, the cardinality of $\mathcal{G}$ is two, and observations are interpreted as being either negative or positive. Within this thesis there then are two types of possible misclassification (errors): either we falsely classify a positive observation as a negative observation (*false negative*), or we falsely classify a negative observation as a positive observation (*false positive*).

The loss function of (5.2) assumes the cost of a false negative to be equivalent to that of a false positive. In practice, this may be unrealistic. For example,

when predicting churn in the telecommunication industry, we know from Glady et al. (2009) that the consequence of a false negative error significantly outweighs that of a false positive error. Therefore, we may wish to use an alternative loss function defined via some cost function denoted as $c(\text{classification}, \text{class})$. The cost of a false negative is given by $c(\hat{y} = 0, y = 1)$ and that of a false positive is $c(\hat{y} = 1, y = 0)$. We may (for brevity) denote these costs by $c_0 = c(\hat{y} = 0, y = 1)$ and $c_1 = c(\hat{y} = 1, y = 0)$ respectively.

5.3 Confusion Matrix and Associated Performance Measures

Recall that probabilistic classifiers provide an estimate of the posterior probability that the i^{th} observation is positive ($Y_i = 1$), and that this is denoted as $\hat{\mathbb{P}}(Y_i = 1|X = x_i)$. Now let $\hat{p}_i(x_i)$ represent $\hat{\mathbb{P}}(Y_i = 1|X = x_i)$ for the sake of brevity. Given N observations, the probabilistic classifier gives us the set of predictions $\hat{\mathbf{p}} = \{\hat{p}_1(x_1), \ldots, \hat{p}_N(x_N)\}$. We can assign the predicted class labels $\hat{y}_i$ to each observation in this set using (5.1) for some discrimination threshold $d \in (0, 1)$. (Where the classifier is not probabilistic, the classifier naturally gives us the class labels $\hat{y}_i$.)

Let $\hat{\mathbf{y}}$ be the vector of all "predicted" class labels produced by a classifier and $\mathbf{y}$ that of observed class membership. The confusion of the classifier may be represented in a *confusion matrix* as presented in Figure 5.1. The confusion matrix typically forms the basis for any analysis of the performance of a 2-class classifier.

Figure 5.1 illustrates all four available outcomes of a classifiers prediction for a single observation, including the two possible misclassification types: false negative and false positive. From Figure 5.1 we can see the *true positive rate* of a classifier to be

$$\text{TPR} = \mathbb{P}(\hat{\mathbf{y}} = 1|\mathbf{y} = 1) \approx \frac{\sum_{i=1}^{n} 1(\hat{y}_i = 1, y_i = 1)}{\sum_{i=1}^{N} 1(y_i = 1)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.3}$$

with associated *false positive rate* given as

$$\text{FPR} = \mathbb{P}(\hat{\mathbf{y}} = 1|\mathbf{y} = 0) \approx \frac{\sum_{i=1}^{n} 1(\hat{y}_i = 1, y_i = 0)}{\sum_{i=1}^{N} 1(y_i = 0)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{5.4}$$

Both these measures are commonly used as scalar values in assessing classifier

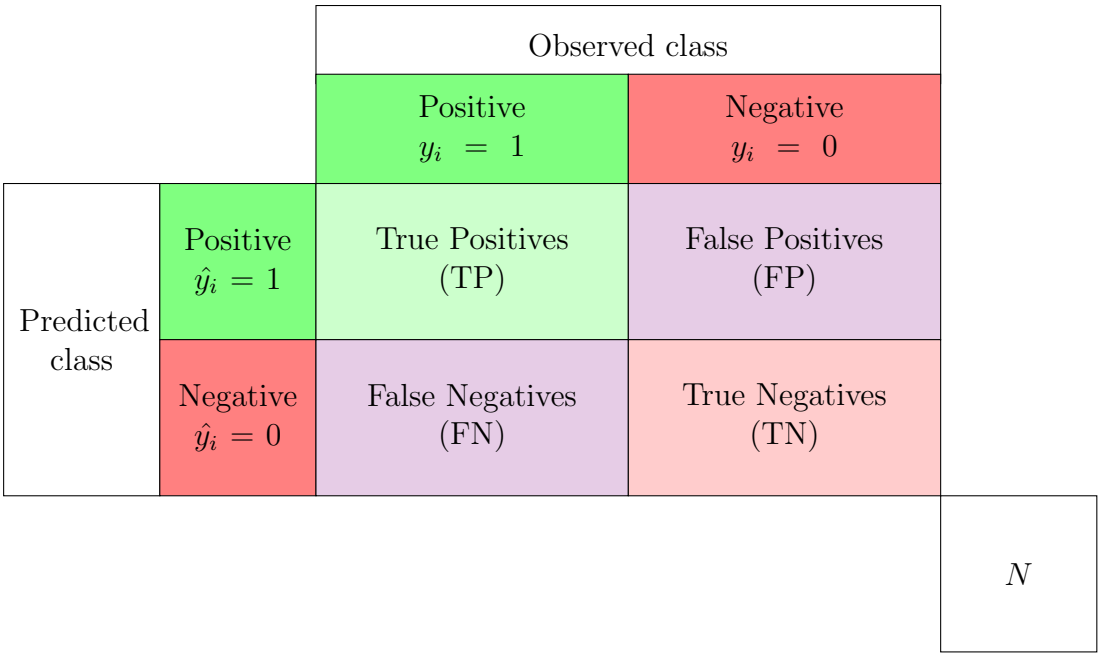


Figure 5.1: Confusion matrix for assessing classifier performance.

performance. Other popular scalar measures derived from Figure 5.1 are given in (5.3)—(5.9).

$$\text{Recall} = \text{Sensitivity} = \frac{\sum_{i=1}^{N} 1(\hat{y}_i = 1, y_i = 1)}{\sum_{i=1}^{N} 1(y_i = 1)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.5}$$

$$\text{Specificity} = \frac{\sum_{i=1}^{N} 1(\hat{y}_i = 0, y_i = 0)}{\sum_{i=1}^{N} 1(y_i = 0)} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{5.6}$$

$$\text{Precision} = \frac{\sum_{i=1}^{N} 1(\hat{y}_i = 1, y_i = 1)}{\sum_{i=1}^{N} 1(\hat{y}_i = 1)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.7}$$

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} 1(\hat{y}_i = y_i)}{N} = \frac{\text{TP} + \text{TN}}{N} \tag{5.8}$$

$$\text{F}_1 \text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = 2 \cdot \frac{\text{TP}}{2 \cdot \text{TP} + \text{FN} + \text{FP}} \tag{5.9}$$

Thus far, we have only considered single scalar performance measures, all derived from the confusion matrix. We showed earlier that the confusion matrix of a probabilistic classifier depends on a discrimination threshold d. Therefore, all the above measures can be criticised for judging classifier performance based on a single value of the discriminant threshold d and we implicitly assume the cost of a false negative to be equivalent to that of a false positive.

A tacit assumption of the accuracy statistic of (5.8) is for the distribution of observations between classes to be balanced and relatively constant (Hand, 1997; Provost and Fawcett, 1997) . However, in practice, it is common for one or both of these assumptions to be broken. For example, consider the case where 99% of the observations belong to one class (the *majority class*) and only 1% of observations to the other (the *minority class*). Classifying all observations into the maximum likelihood (majority) class results in an accuracy value of 99%, achieving 100% recognition rate for the majority class but 0% for the minority class. Such a classifier is trivial, being biased by the majority class, and should not be accepted although it has a high accuracy score. Thus in cases where the class distribution is skewed, conclusions drawn via use of the accuracy statistic become unreliable. As an alternative, Provost and Fawcett (1997) advocate ROC analysis (see Section 5.4).

5.4 Receiver Operating Characteristic (ROC)

Receiver Operating Characteristic (ROC) graphs are time-honoured and widely used tools for visualising classifier performance. Since the origin of ROC graphs in signal detection theory (Green and Swets, 1966), they have been generalised for use in diagnostic systems (Swets, 1988) with one of the earliest adoptions in machine learning made by Montana (1990). Classifier analysis by ROC graph only increased following the discrediting of accuracy (5.8) as a performance metric by Provost and Fawcett (1997). Unlike the performance measures of Section 5.3, this visual approach separates classifier performance from assumptions about both class and cost distributions. The two-dimensional ROC space is defined by true positive rate (y-axis) versus false positive rate (x-axis) where both exist continuously between zero and one. Curves through ROC space depict the trade-off between correctly classifying positive observations and incorrectly classifying negative observations as the discriminating threshold d is varied.

5.4.1 ROC curves

Each point in ROC space corresponds to a value of the discriminating threshold d. A discrete classifier assigns a class label to each observation, resulting in one single confusion matrix and corresponding point in ROC space. Whereas, we know probabilistic classifiers assign each observation a posterior probability, giving a

sample of N probabilities such that $\hat{\mathbf{p}} = [\hat{p}_1(x_1), \ldots, \hat{p}_N(x_N)]^\top$. From this a *ranking* or *scoring* classifier can be created by assigning classes according to (5.1) for some discriminating threshold d. By methodically considering all continuous values of d in the interval $(0, 1)$ we can trace a curve in ROC space. When $d = 1$, the true positive and false positive rates are zero, whilst when $d = 0$ both rates are one. As we decrease d, we trace the ROC curve step-wise from (0,0) to (1,1) and the resulting curve is piecewise-linear. Note that as the value of the discriminating threshold decreases, we move from more conservative to more liberal decisions. This is a computationally expensive way to plot the ROC curve, requiring the confusion matrix to be computed for each discriminating threshold possible.

Given finite N, the ROC curve is actually a step function, as N approaches infinity this curve approaches the true curve. Therefore an alternative method of creating the ROC curve begins with ordering observations by their posterior probabilities in a non-increasing manner. Represent this ordered vector by $\hat{p}_{(i)}(x_i)$ and let $y_{(i)}$ be the resulting ordered vector of observed classes. Take also N_+ (N_-) be the number of positive (negative) instances observed. The ROC curve can be traced, starting at (0,0), moving $1/N_+$ up if $y_{(i)} = 1$ and $1/N_-$ to the right if $y_{(i)} = 0$, until (1,1) is reached.

For example, take a sample of size 20 with balanced class distribution (equal number of positive and negative observations). The output of three probabilistic test classifiers, ordered by their "scores", is given in Table 5.1. Figure 5.2a shows the corresponding ROC curves as rainbow coloured lines, where the shade represents the discrimination threshold d. For all three classifiers, we see that for $d > 0.98$, the point (0,0) is produced. Reducing this to 0.76 leads all curves up to (0,0.1) and reducing d further to 0.75 extends the solid and dashed curves right to (0.1,0.1). Continuing to reduce d takes all three curves up and to the right until they reach the point (1,1). The data corresponding to those ROC curves in Figure 5.2b are given in Table 5.2.

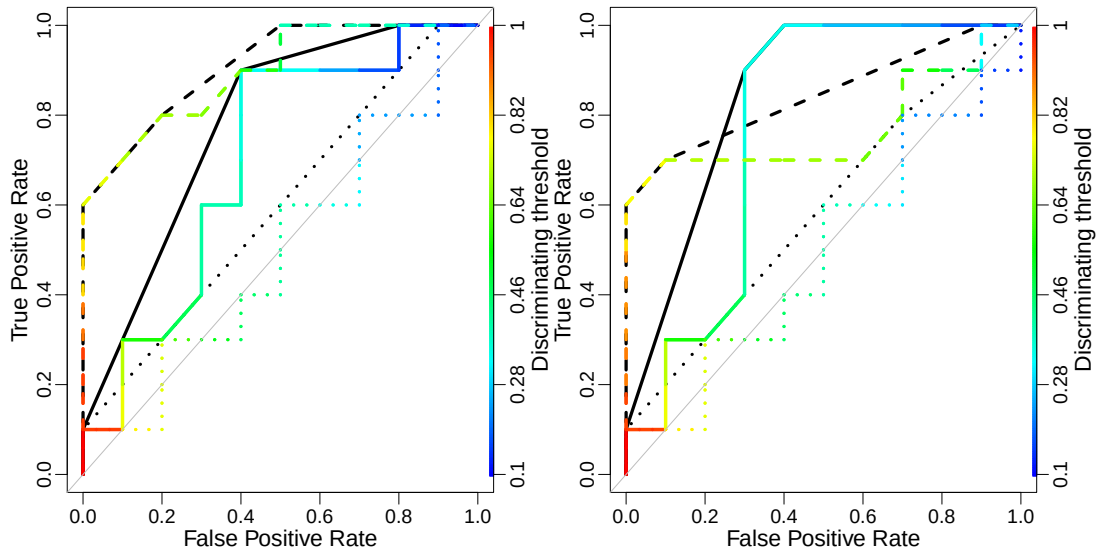
Any guess which is completely random, ignorant of the prior probabilities of class membership, $\pi_{(0)}$ and $\pi_{(1)}$, lies along the line $y = x$ in ROC space (given in grey in Figure 5.2). This line is referred to as the line of no-discrimination. Those classifiers which lie "north-west" of this line perform better than the random, whereas those which lie "south-east" perform less well. The "perfect" classifier will have a ROC curve that traces a direct path starting at (0,0), turning at (0,1), and ending at (1,1).

An iso-performance line is a straight line whose gradient is determined via

Table 5.1: Test data of three probabilistic classifiers with balanced class prior probability corresponding to the receiver operating characteristic curves in Figure 5.2a.

(a) Solid			(b) Dashed			(c) Dotted		
Rank	$\hat{p}_{(i)}(x_i)$	$y_{(i)}$	Rank	$\hat{p}_{(i)}(x_i)$	$y_{(i)}$	Rank	$\hat{p}_{(i)}(x_i)$	$y_{(i)}$
1	0.9800000	1	1	0.8951216	1	1	0.98000000	1
2	0.7598144	0	2	0.8947509	1	2	0.75981442	0
3	0.7144198	1	3	0.8438877	1	3	0.72309665	0
4	0.5693930	1	4	0.7528497	1	4	0.71441985	1
5	0.4808133	0	5	0.7389044	1	5	0.56939302	1
6	0.4136979	0	6	0.6926219	1	6	0.48081334	0
7	0.4136979	1	7	0.6400000	0	7	0.41369788	0
8	0.3970384	1	8	0.6400000	1	8	0.39703841	1
9	0.3970384	1	9	0.6300000	0	9	0.35522550	0
10	0.3867518	0	10	0.6300000	1	10	0.35061007	1
11	0.3552255	1	11	0.6029616	0	11	0.33440561	1
12	0.3506101	1	12	0.5191867	0	12	0.27357174	0
13	0.3344056	1	13	0.5191867	1	13	0.24715028	0
14	0.2471503	0	14	0.4306070	0	14	0.17143407	1
15	0.2471503	0	15	0.3248888	1	15	0.15611228	1
16	0.1714341	0	16	0.3247021	0	16	0.15298093	0
17	0.1561123	0	17	0.3240000	0	17	0.10487845	0
18	0.1338040	1	18	0.2855802	0	18	0.08470593	1
19	0.1048784	0	19	0.1125752	0	19	0.02660914	1
20	0.1000603	0	20	0.0200000	0	20	0.00000000	0

some specified cost or class distribution, see Definition 5.1. Given that (5.10) is dependent on both misclassification costs and class distributions, the classifiers corresponding to the points (FPR_1, TPR_1) and (FPR_2, TPR_2) incur the same expected cost i.e. have the same performance. A family of iso-performance lines exists for each set of cost and class distributions. For each family of lines, those which are more "north-west" and hence have greater intercept (true positive rate) provide classifiers with better performance (lower expected cost) for the given familial cost and class distributions. Therefore, the classifier with which the iso-performance line is tangential whilst having greatest intercept, is the optimal classifier for the corresponding cost or class distribution. For example, assuming the zero-one loss function ($c_0 = c_1$), the gradient of the respective iso-performance line is one. In Figure 5.2a, this iso-performance line chooses the dashed classifier with $d \in (0.6926, 0.6030)$. Whereas, for the classifiers of Figure 5.2b, the iso-performance line with gradient one identifies the dashed classifier



(a) Data given in Table 5.1 (b) Data given in Table 5.2

Figure 5.2: Example Receiver Operating Characteristic (ROC) space of probabilistic classifier performance. The coloured lines are the ROC curves whilst the black lines are the respective ROC convex hulls, each line type represents a different probabilistic classifier.

with $d \in (0.7264, 0.6494)$ or solid classifier with $d \in (0.3506, 0.3344)$. We clearly see from comparing the coloured lines in the two figures, that the global superiority of the dashed classifier in Figure 5.2a is not observed in Figure 5.2b, here we observe only local superiority.

Definition 5.1.

Provost and Fawcett (1997, 2001) convert some set of operating conditions into what they term an *iso-performance line*. Recall that $\pi_{(k)}$ is the prior probability of an observation belonging to class k and $c(\hat{g} = \mathcal{G}_\ell | g = \mathcal{G}_k)$ is the cost of misclassifying an observation of class k into class ℓ. The gradient of the iso-performance line intersecting any two points $(\text{FPR}_1, \text{TPR}_1)$ and $(\text{FPR}_2, \text{TPR}_2)$ in ROC space is given by:

$$\nabla(I_{a,b}) = \frac{\text{TPR}_1 - \text{TPR}_2}{\text{FPR}_1 - \text{FPR}_2} = \frac{\pi_{(0)} \cdot c(\hat{y} = 1 | y = 0)}{\pi_{(1)} \cdot c(\hat{y} = 0 | y = 1)}. \tag{5.10}$$

One property of ROC curves is their insensitivity to changes in misclassification costs or skew of class distribution; each curve conveys all information about possible misclassification costs or class skews. This is an attractive property, because we can alter the gradient of the iso-performance line to reflect different cost distri-

Table 5.2: Test data of three probabilistic classifiers with balanced class prior probability corresponding to the receiver operating characteristic curves in Figure 5.2b.

(a) Solid			(b) Dashed			(c) Dotted		
Rank	$\hat{p}_{(i)}(x_i)$	$y_{(i)}$	Rank	$\hat{p}_{(i)}(x_i)$	$y_{(i)}$	Rank	$\hat{p}_{(i)}(x_i)$	$y_{(i)}$
1	0.9800000	1	1	0.8951216	1	1	0.98000000	1
2	0.7598144	0	2	0.8438877	1	2	0.75981442	0
3	0.7144198	1	3	0.8285659	1	3	0.72309665	0
4	0.5693930	1	4	0.7528497	1	4	0.71441985	1
5	0.4808133	0	5	0.7528497	1	5	0.56939302	1
6	0.4136979	0	6	0.7264283	1	6	0.48081334	0
7	0.4136979	1	7	0.6655944	0	7	0.41369788	0
8	0.3970384	1	8	0.6655944	1	8	0.39703841	1
9	0.3970384	1	9	0.6493899	0	9	0.35522550	0
10	0.3552255	1	10	0.6493899	0	10	0.35061007	1
11	0.3506101	1	11	0.6447745	0	11	0.33440561	1
12	0.3506101	1	12	0.6029616	0	12	0.27357174	0
13	0.3344056	0	13	0.6029616	0	13	0.24715028	0
14	0.3344056	1	14	0.5863021	0	14	0.17143407	1
15	0.2735717	0	15	0.5863021	1	15	0.15611228	1
16	0.2471503	0	16	0.5191867	1	16	0.15298093	0
17	0.2471503	0	17	0.4306070	0	17	0.10487845	0
18	0.1714341	0	18	0.2855802	0	18	0.08470593	1
19	0.1561123	0	19	0.2401856	1	19	0.02660914	1
20	0.1048784	0	20	0.0200000	0	20	0.00000000	0

butions or class skews without altering the ROC curve itself. For the intersecting classifiers of Figure 5.2b, the dashed classifier outperforms the solid classifier for iso-performance lines with gradient greater than one, which corresponds to $c_0 < c_1$. Similarly, the solid classifier outperforms the dashed classifier for iso-performance lines with gradient less than one, which arises from $c_0 > c_1$. In practice, it is common for large class skews (such as class ratios of $1{:}10^2$) to be present in all domains concerned with modelling a rare event (Kubat et al., 1998). Even skews in the order of 10^6 have been observed in some domains (Fawcett, 2006). Changes in class skews are not unrealistic, for example, the occurrences of fraud temporally and spatially vary greatly (Fawcett and Provost, 1997).

Because the ROC curve is a step function, it is suitable for those "probabilistic" classifiers which output an uncalibrated score rather than a proper probability. For such classifiers, the only property of probability which holds is that higher scores imply higher probability. Therefore the ROC curve measures a classifier's abil-

ity to produce *relative* instance scores (or ranks) that reflect the ordering of the classes. The classifier does not need to produce accurate and calibrated probability estimates; it is necessary only to produce relative scores (or ranks) which discriminate well between the positive and negative instances. That is, to rank positive observations above negative observations. Let a *true* probabilistic classifier be one which produces calibrated probability estimates. If there exists a single d value for which a truly probabilistic classifier estimates all values $\hat{p}_i(x_i) \geq d$ ($\hat{p}_i(x_i) < d$) for $y_i = 1$ ($y_i = 0$), it will have the "perfect" ROC curve. The importance lies in all estimated probabilities corresponding to positive observations not being less than those of any negative observation.

ROC curves provide rich information about the trade-off between increased true positive rate and decreased false positive rate of a classifier under different assumptions. Unfortunately this information is lost when attempting to summarise these two-dimensional spaces by a single scalar metric. Figure 5.2b illustrates that it is possible for one classifier not to be globally superior, but only locally superior, to all others considered.

5.4.2 Receiver Operating Characteristic convex Hull (ROCH)

The convex hull is entirely comprised of iso-performance (see Definition 5.1) line segments, each assuming a different combination of cost and class distributions such that a convex upper boundary is formed on all ROC points considered. A convex hull may be given over individual ROC curves (as represented by the black lines in Figure 5.2), or a combination of ROC curves (as given by the black line in Figure 5.3). Note that for Figure 5.2a, the convex hull of the dashed classifier is the convex hull over all classifiers given that this classifier is globally superior.

Consider instances where the ROC curves of two classifiers contributing to the convex hull intersect. The point of intersection occurs between where one classifier leaves and another joins the convex hull, in Figure 5.3 this occurs for FPR $\in (0.1, 0.3)$. For the region where no classifier sits directly on the convex hull, Provost and Fawcett (1997, 1998) and Scott et al. (1998) present an elegant approach to combine individual ROC curves to produce what Scott et al. (1998) call "the maximum realisable ROC".

The ROC convex hull is not dependent on any operating conditions (such as cost or class distributions assumptions) and identifies where classifier is optimal. Those points along the ROC convex hull dominate all others. Any classifier which

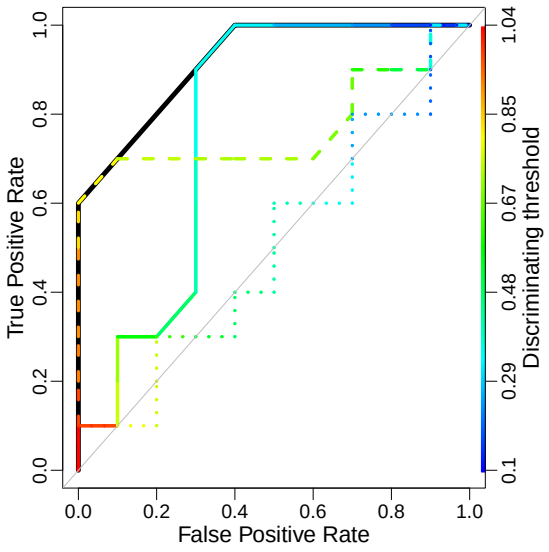


Figure 5.3: Global Receiver Operating Characteristic convex Hull (ROCH) across all classifiers of Figure 5.2b.

makes up part of the convex hull should be considered for any enforced operating conditions. Consequently, those classifiers which do not contribute to the convex hull need not be retained.

5.4.3 Area under curve (AUC)

When comparing classifier performance, it is convenient to use one-dimensional scalar measures as opposed to two-dimensional graphical spaces. Whereas the former can be assessed by a computer, the latter requires visual consideration. The Area Under (ROC) Curve (AUC) has been, and continues to be, widely used throughout various research and industrial fields such as applied statistics, credit scoring, psychology, medicine and bioinformatics (Bradley, 1997; Fawcett, 2006; Hanley and McNeil, 1982; Zweig and Campbell, 1993).

After the discrediting of accuracy (and other scalar measures of Section 5.3) as a performance measure, the AUC became vastly popular. The AUC, by definition, aggregates classifier performance over all possible discriminant thresholds, in addition to all distributions of class and cost. Therefore this measure provides a "global measure of the separability between the distribution of scores for positive and negative observations" (Krzanowski and Hand, 2009, p. 11).

Take s_i to represent the score for the i^{th} observation from a scoring classifier. Where the classifier is a true probabilistic classifier, $s_i = \hat{p}_i(x_i)$. Let $x(d) = \mathbb{P}(s \geq d|-)$ be the false positive rate and $y(d) = \mathbb{P}(s \geq d|+)$ be the true positive rate

(determined by the discriminating threshold d) on the ROC convex hull, so that the ROC convex hull is defined by the relation $y = h(x)$. This curve is a monotonic increasing function residing in the triangle above the line $y = x$ starting, at (0,0) and reaching (1,1). As the ROC space occupies the unit square, the AUC is bound within the interval $[0, 1]$. Given the ROC curve of the random classifier is the diagonal joining the coordinates (0,0) and (1,1) (illustrated by the grey line in Figure 5.2), this has AUC of 0.5. Any non-trivial classifier should therefore have AUC greater than 0.5. Note that whilst this is an indication, not all classifiers with AUC of 0.5 are trivial. In general, the area under such a curve is

$$\text{AUC} = \int_0^1 y(x)\, dx. \tag{5.11}$$

It follows directly that, given two classifiers where the ROC curve of one is globally superior to that of the other (such as the dashed and solid ROC curves in Figure 5.2a) the AUC measure is superior. For example, with respect to the classifiers of Figure 5.2a, we observe the dashed classifier to have AUC of 0.895 whereas the solid classifier has inferior AUC of 0.695. However, the reverse does not hold true given that the AUC is incapable of indicating whether the ROC curves of two classifiers intersect. For example, the solid and dashed classifiers of Figure 5.2b share an AUC value of 0.77, but this is because the curves are reflections in the line $y(x) = -x$. Without visually analysing Figure 5.2b, we would short-sightedly conclude that these classifiers perform equally well.

There are many interpretations of AUC, some more apparent than others. Given (5.11) and applying results from elementary calculus and probability theory, the AUC is the uniform average of the true positive rate taken over all feasible false positive rates. Hand (2005) showed the AUC to be a linear transformation of the weighted misclassification rate, where the weighting is determined by the mixture distribution of the positive and negative instances. Another interpretation of the AUC is its equivalence to the probability of the classifier ranking a randomly selected positive observation above a randomly selected negative observation (Fawcett, 2006). The proof of this can be found in (Krzanowski and Hand, 2009, p. 27). This, in turn, is equivalent to the Wilcox-Mann-Whitney statistic for ranks (Hanley and McNeil, 1982). In addition, Hand and Till (2001) show the AUC to be linearly related to the Gini coefficient and the ROC being effectively the Lorenz curve defined by Lorenz (1905).

As with all theory, the AUC is known to have disadvantages, such as being

misleading if two ROC curves cross but have similar AUC. However Hand (2009) recognised a much more serious deficiency in this performance metric: that different metrics are used to evaluate different classification rules. Consequently, the severity of a false negative for one classifier may be p times that of a false positive; whereas, for an alternative classifier this magnitude is allowed to be P, where $p \neq P$. Without care, practitioners may incidentally compare classifiers which have inherently different class or cost distributions. Thus, following the publication of Hand (2009), those researching how best to analyse probabilistic classifiers have looked to either better justify the use of AUC (Berrar and Flach, 2012) or into alternative, but related, metrics in cost space (Hernández-Orallo et al., 2012).

5.5 Cost Curves

Cost curves were first proposed by Drummond and Holte (2000) as an alternative to ROC analysis in graphically representing classifier performance. They argue increased interpretability is the primary benefit of visualising classifier performance in cost over receiver operating characteristic space. In addition, the authors demonstrate the point-line duality: how a point in ROC space corresponds to a line in cost space covering all cost and class distributions. Therefore, most performance measures for ROC space are translatable to cost space. Further, the authors conclude that, due to such transferability, it is unnecessary to choose between ROC representation and cost representation. However, some measures are more intuitively presented in one space over the other; for example, statistical significance when analysing the difference in classifier performance is visually clearer in cost than ROC space.

5.5.1 Expected cost

To measure the improvement of one classifier over another in ROC space, it is tempting use the Euclidean distance from the superior curve in direction normal to the inferior curve. Let the expected cost of a classifier be defined as

$$\mathbb{E}(C) = (1 - \text{TPR}) \cdot \pi_{(1)} \cdot c(\hat{y} = 0|y = 1) + \text{FPR} \cdot \pi_{(0)} \cdot c(\hat{y} = 1|y = 0)$$

The difference between two classifiers with respect to the expected cost is the weighted Manhattan distance given in (5.12).

$$\begin{aligned}\mathbb{E}(C_1) - \mathbb{E}(C_2) = (\text{TPR}_1 - \text{TPR}_2) \cdot \pi_{(1)} \cdot c(\hat{y} = 0|y = 1) \\ + (\text{FPR}_1 - \text{FPR}_2) \cdot \pi_{(0)} \cdot c(\hat{y} = 1|y = 0)\end{aligned} \tag{5.12}$$

In addition, when measuring the difference in performance, the distance should be taken between appropriate points of each ROC curve. These points are determined by

$$w_+ = \pi_{(1)} \cdot c(\hat{y} = 0|y = 1) \quad \text{and} \quad w_- = \pi_{(0)} \cdot c(\hat{y} = 1|y = 0).$$

From (5.10) of Section 5.4.2, we find the ratio of w_+ to w_- to be the slope of the iso-performance lines in ROC space that are used to determine points along ROC curves that have the same expected cost.

5.5.2 Cost space

The cost space is defined by the expected cost normalised by the maximum cost incurred (5.13) (y-axis) versus the probability-cost function (PCF) for positive observations (5.14) (x-axis). Curves in cost space thus depict the expected cost over all possible values of the probability cost function.

$$N\mathbb{E}(C) = \frac{(1 - \text{TPR}) \cdot w_+ + \text{FPR} \cdot w_-}{w_+ + w_-} \tag{5.13}$$

$$PCF(+) = \frac{w_+}{w_+ + w_-} \tag{5.14}$$

Note that where the misclassification costs are equal, $PCF(+)$ is equivalent to $\pi_{(1)}$. Define $PCF(-)$ as

$$PCF(-) = \frac{w_-}{w_+ + w_-} \tag{5.15}$$

and subsequently reinterpret the normalised expected cost in terms of the probability-cost function to give

$$N\mathbb{E}(C) = (1 - \text{TPR}) \cdot PCF(+) + \text{FPR} \cdot PCF(-). \tag{5.16}$$

From (5.14) and (5.15) we can infer

$$PCF(+) + PCF(-) = 1,$$

which we can use to further simplify the normalised expected cost of (5.16) to form (5.17).

$$N\mathbb{E}(C) = (1 - \text{TPR} - \text{FPR}) \cdot PCF(+) + \text{FPR} \tag{5.17}$$

Therefore, by definition, the cost space explicitly presents the difference in performance (with respect to expected cost) between two classifiers.

To move from ROC to cost space, we note that (5.17) converts the point (TPR,FPR) in ROC space to a line in cost space for some given operating conditions (class and cost distributions). Drawing this line may be achieved by simply connecting the normalised expected cost in the extreme $PCF(+)$ values where (5.17) simplifies to

$$N\mathbb{E}(C) = \begin{cases} \text{FPR} & \text{for } PCF(+) = 0, \\ 1 - \text{TPR} & \text{for } PCF(+) = 1. \end{cases}$$

An iso-performance line, with gradient $\nabla(I)$ and intercept TPR_I in ROC space, may be converted to a point in cost space via the relations given in (5.18).

$$\begin{aligned} PCF(+) &= \frac{1}{1 + \nabla(I)} \\ N\mathbb{E}(C) &= (1 - \text{TPR}_I) PCF(+) \end{aligned} \tag{5.18}$$

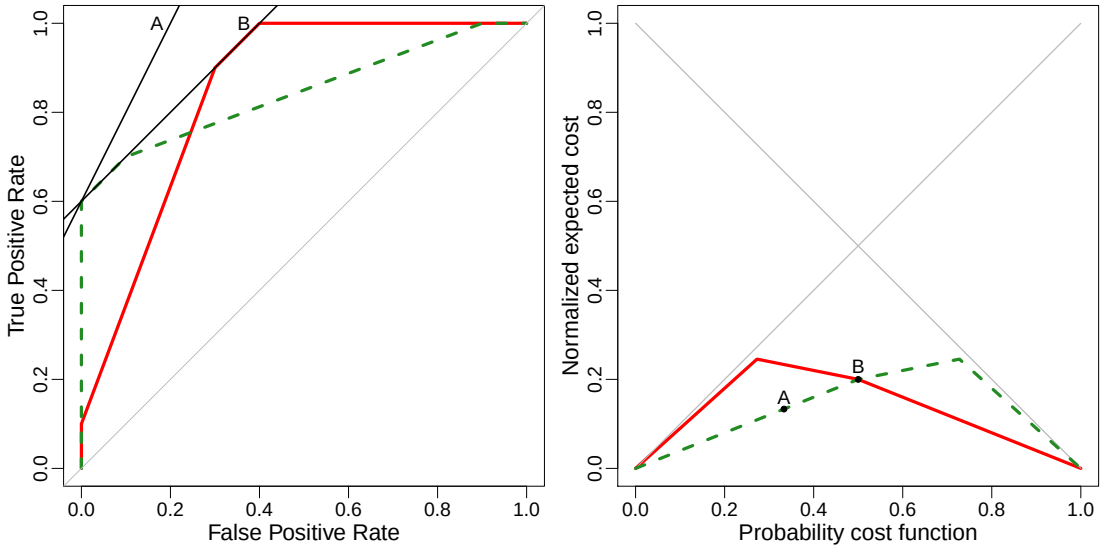
Given the obvious reversibility of (5.17) and (5.18), we have the previously mentioned point-line duality between the two considered spaces. Figure 5.4 indicates how two iso-performance lines A and B correspond to two points in cost space.

5.5.3 Lower envelope

For a given classifier, each instance of the set of possible (TPR,FPR) which generates the curve in ROC space produces a straight line in cost space. For the example classifiers presented in Figure 5.2 of Section 5.4, we have the cost curves given in Figure 5.5 where the red, green and blue colouring of classifiers in cost space represents the solid, dashed and dotted classifiers in ROC space respectively.

The *lower envelope* for a classifier is formed by those cost lines that have minimum normalised expected cost over the range of $PCF(+)$ values. This concept is demonstrated by the solid lines of Figure 5.5; here the solid blue/red/green curve is the lower envelope of the dotted light blue/red/green cost lines. The dashed black line in each cost space marks the lower envelope over all possible classifiers. From

Figure 5.5a, we see the green classifier is superior over all $PCF(+)$ values with respect to normalised expected cost. In Figure 5.5b, the classifier shown in green is only locally superior to the alternative classifiers for $PCF(+)$ values between 0 and 0.5, after which $PCF(+)$ values, the red classifier is superior. Considering this second set of classifiers, we say the minimum cost for the operating range (0,0.5) is achieved using the green classifier and achieved using the red classifier for the range (0.5,1).



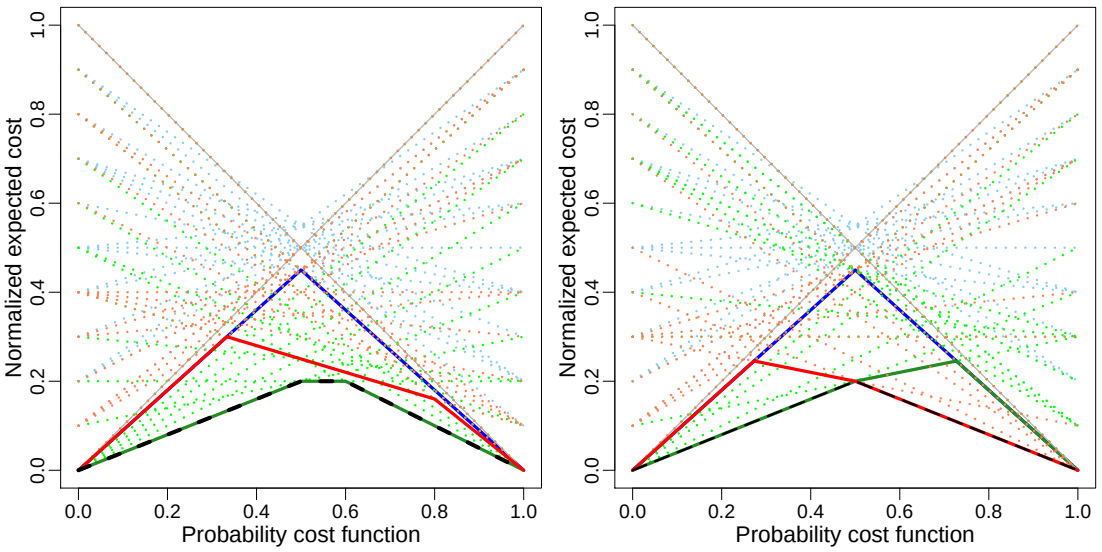
(a) ROC space (b) Cost space

Figure 5.4: Example illustration of point-line duality between ROC and cost spaces.

Unlike when illustrating classifier performance in ROC space, the representations in cost space are much easier to interpret. For example, from Figure 5.5a we instantly see that, for $PCF(+) < 0.5$, the green (dashed) classifier performs best and for $PCF(+) > 0.5$ the red (solid) classifier is superior to all others considered. This information was not readily observable from the equivalent figure in ROC space (Figure 5.2a). However, where a classifier is globally superior, there are no comparative apparent gains to the cost space representation.

5.5.4 Area under cost curve

No direct mapping exists between the area under the ROC curve and the area under the cost curve. We assume no knowledge of the $PCF(+)$ value used, however, in the instances this is known, we require the classifier corresponding to that part



(a) Table 5.1 (b) Table 5.2

Figure 5.5: Example cost space of probabilistic classifier performance.

of the lower envelope to be used. If in addition we assume the probability distribution of $\pi(PCF(+))$ to be uniform, then the area under the lower envelope is the expected cost. Where $\pi(PCF(+))$ is known, the expected cost can be expressed as

$$\int_0^1 N\mathbb{E}(C(PCF(+)))\pi(PCF(+))dPCF(+).$$

The difference between the area under cost curve for two classifiers is thus the difference in expected cost between the classifiers. We feel area under curve is more intuitive in the cost space. Those criticisms which Hand (2009) makes of the AUC and inappropriate use of the ROC space for probabilistic classifiers have no substance with respect to cost curves as proposed by Drummond and Holte (2000, 2006).

5.6 Brier Score

The Brier score is a measure of classifier performance designed to assess truly probabilistic classifiers (Brier, 1950). Recall that the output of a probabilistic classifier (for binary classification) is an estimate of the posterior probability that the i^{th} observation is in the positive class (represented by $\hat{p}_i(x_i)$). Given the vector of observed class labels $\mathbf{y}$, the Brier score can be formulated as in (5.19). A more general version of (5.19), for where $\mathcal{G}$ has cardinality greater than two, can be

found in (Brier, 1950).

$$BS = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i(x_i) - y_i)^2 \tag{5.19}$$

There are many decompositions of (5.19) which aim to provide a deeper insight into various aspects of classifier performance. However, these decompositions take a one-dimensional scalar measure and split it up into multiple one-dimensional scalar measures. This is no different to using a combination of the previously given scalar measures. Later, in Chapter 6, we empirically show that when interpreting multiple performance measures there can be issues with the consistency of findings.

In Section 5.4 we showed that ROC curves are designed for ranking classifiers, not probabilistic classifiers. The Brier score, being designed for probabilistic classifiers, is useful when representing or interpreting probabilistic classifiers.

5.7 Conclusion

In this chapter we have demonstrated that graphical performance measures are more suitable for analysing and comparing classifier performance than scalar performance metrics. With respect to graphical performance measures, we have identified cost curves to be more easily to interpret than ROC curves and demonstrated the point-line duality that exists between the two spaces.

Within the literature, claims about how one classifier out performs another are typically based on numeric measures such as AUC. There are two types of errors possible (false positive and false negative) where $\mathcal{G}$ has cardinality of two, and where the classification method is probabilistic, these errors depend on the choice of the discriminating threshold. The reduction of the two-dimensional error space to one-dimension results in a loss of information and thus decisions based on one-dimensional measures are open to criticism. Where one classifier is consistently superior to another over all discriminating thresholds, the one-dimensional scalar metrics provide an easy to comprehend measure of some aspect of superiority. However, what these measures cannot tell you is globally whether one classifier is superior to another, and locally under what costs (or class ratios) a classifier is superior. We conclude that any two-dimensional performance metric, capable of identifying the circumstances under which each considered classifier is superior, is preferable to any combination of scalar performance metrics.

ROC curves, are by definition, designed for ranking or scoring classifiers, and

as such, analysis made in ROC space is theoretically incompatible with true probabilistic classifiers. Cost curves maintain the benefits of ROC curves and in addition make the optimal operating conditions of each classifier considered extremely clear. A consequence of this is that cost curves are easier to interpret.

Much research is still being done into how best to analyse the performance of supervised learning models such as classification. Hand and Flach, among others, are keenly focused into how to best represent and analyse probabilistic classifiers as is evident from exchanges within the literature. This literature has directly impacted our work as it made us aware of a significant flaw so commonly present in the analysis of classifiers, basing judgement upon one-dimensional measures.

In Chapter 6 we will use cost curves, and compare these against ROC curves, to demonstrate the superior interpretability of the cost curves which we have shown to be an alternative representation of ROC curves. We also consider two scalar measures (AUC and Brier score) to empirically validate our claim that scalar measures are unreliable and are commonly used poorly in practice to incorrectly suggest global dominance of a classifier.

Chapter 6

Results

The primary objective of this chapter is to present an empirical analysis of the predictive classification of our novel user satisfaction risk event formulations. It also aims to demonstrate why the cost curve is the best classification quality characteristic. The secondary objective is to describe our integration with the ROBUST software platform to support the final deliverable of the project to the European Research Commission.

6.1 Introduction

In this chapter, we present our main results culminating from the work described in the previous chapters and give a description of our ROBUST related effort. The presentation of results includes details of our implementation of those classification methods from Chapters 3 and 4, as well as of the chosen classification quality characteristics from Chapter 5. Graphically, we use cost curves and ROC curves to analyse classifier performance for both of our novel formulations of risks on user satisfaction and demonstrate why cost-curves are preferred on our real-world case-study of the SCN. We also consider the popular AUC and Brier score measures to evidence the findings in Chapter 5 that scalar measures of classifier performance are unreliable because, when used incorrectly, they suggest the global dominance of a locally dominating classifier.

A by-product of using the four chosen classification quality characteristics in our analysis is that we shall be able to rigorously expose all aspects of classifier performance. Therefore, when comparing classifiers, we compare every aspect of the classifier for each of our user satisfaction risk event formulations. Our analysis, which forms the base of our discussions, is thus well evidenced and repeatable. In

practice, a community manager/moderator would be involved in the analysis and would provide appropriate relative misclassification costs so that the probability cost function can be determined. This value could then be used in cost space for model and feature selection, minimising the normalised expected cost.

Our analysis is performed per forum over five fora (with numerical identifiers 142, 141, 50, 156 and 418) that represent different activity patterns in the SCN. For each model considered, we implement 10-fold cross-validation as laid out in Section 6.2.2.

In detail, the contents of this chapter firstly provide details of our implementation work with respect to this thesis outlining the programming languages and packages used (Section 6.2.1) and why and how we used cross-validation (Section 6.2.2). Secondly, it provides a full description of our additional work required by the ROBUST project in Section 6.3 and this section includes details of the integration of our software within the ROBUST software platform that formed part of the final deliverable to the European Commission. Thirdly, we present our analysis on modelling questioner satisfaction (formulated in Section 2.7.2) in Section 6.4 and fourthly, we present our analysis of the individual user churn event (formulated in Section 2.8.2) in Section 6.5. Finally, we close this chapter with some concluding remarks in Section 6.6.

6.2 Implementation details

This section conveys the software aspects and efforts of our analysis regarding the implementation of our two novel risk event problem formulations described earlier in Chapter 2. The programming languages and associated packages that were used with respect to this thesis are outlined here but details which pertain specifically to the ROBUST project are given later in Section 6.3. The user of cross-validation in our implementation is also detailed in this section, including a high-level algorithmic overview of our code post-data-extraction.

6.2.1 Programming languages and packages used

Our statistical analysis was performed in the language and environment R (version 3.0.1) (R Core Team, 2013) on a stand-alone computer with a 64-bit operating system and 16 gigabytes of memory. We explored and extracted features from the SCN SQL database using PostgreSQL (known as "Postgres"). This was done

directly via the Linux terminal for simple queries; via R using the RJDBC package for statistical analysis and graphical exploration; and via Java using the Postgres library jar (specifically postgresql-9.3-1100.jdbc41.jar) for extracting all questioner satisfaction and user churn features. The features for both our event formulations were extracted using parallelised Java code and were written out to a table in the same database to be later queried by R when performing classification.

Two additional R packages were used: MASS and ROCR. We used the functions *lda*, *qda* and *glm* within the R package MASS for the classification methods of Chapter 3. However, as stated in Section 4.3.1, we fully implemented the Bayesian Probit classification method (described in Chapter 4) in both the R and Java languages. (The reason for implementing this classifier in Java was to meet a requirement of the ROBUST project (see Section 6.3).) We utilised the R package ROCR to plot the ROC curves and calculate the AUC, but wrote functions to produce the cost curves and calculate the Brier score.

6.2.2 Cross-Validation

All supervised learning methods produce models that are entirely dependant on the training data to which they are fitted. Therefore, it is not possible to calculate the true prediction error of such a model even when the training data is a fair representation of the underlying population. Cross-validation is one approach which can provide a relatively unbiased estimate of the true prediction error, or in the case of classification, a relatively unbiased representation of classifier performance.

Let the training set $\mathcal{T}$ contain the feature and response pairs (x_i, y_i) for N independent observations $(i = 1, \ldots, N)$, and let the underlying relationship between the feature, x_i, and response, y_i, be represented as $h(X, Y)$. Now let the pair (x^*, y^*) represent an unseen observation drawn from the same population as $\mathcal{T}$. Where $\hat{h}(X)$ is $h(X, Y)$ fitted on the observations in $\mathcal{T}$, the expected prediction error (EPE) for loss function $L(\cdot)$ is

$$EPE(x*) = \mathbb{E}\left(L\left(y^* - \hat{h}(X)\right) \mid X = x^*\right).$$

Given that $\hat{h}(X)$ is fitted on those observations in $\mathcal{T}$, the EPE will always be higher for $(x^*, y^*) \notin \mathcal{T}$ (out-of-sample), as opposed to $(x^*, y^*) \in \mathcal{T}$ (in-sample). Where $(x^*, y^*) \in \mathcal{T}$, the true prediction error is underestimated by some training optimism.

In practice, the true prediction error must be estimated. Cross-validation is

one of the most popular approaches (even though it is computationally intensive) given that it requires minimal assumptions, such as the data $((x^*, y^*))$ being from the same underlying population. Additionally, we know from Breiman (1992) that cross-validation can produce estimates of true prediction error that are relatively unbiased.

The concept of cross-validation is to partition a training set without overlap and train on a subset of the whole and then test the fitted model on the remaining observations which were not used for training. This can be generally formed as V-fold cross-validation where the training set is initially split into V non-overlapping approximately equally sized subsets. The model is then trained on all but one of the subsets and tested on that which was excluded. In this manner model performance can be assessed over all observations without the model being dependant on the observations upon which it is dependant. A simulation study by Breiman and Spector (1992) found $V = 5$ to be more true than complete cross-validation where $V = N$ given a training set with N observations, however left the question of "how big should V be?" as an open question which has yet to be answered. Within the literature it is very common to take $V = 10$ and therefore this is what we use. In Algorithm 2 we structure the concept of V-fold cross-validation to give a high-level overview of our code implementation assuming the data to have already been extracted.

In Section 2.8.1 we identified that churn is a rare event in the telecommunications industry and later in Section 5.4.1 we highlighted that it is common in real data to observe large skews between classes. Cross-validation has been used in the literature to address the issue of class imbalance by Stone (1974a) for general statistical predictions, and more specifically for multinomial predictions in (Stone, 1974b). A detailed comparison of cross-validation and other methods for estimating true prediction error is given by Breiman and Spector (1992).

6.3 ROBUST related effort

Figure 6.1 depicts the structure of the ROBUST project. The ROBUST consortium organised research, development and implementation about three aspects of community analysis: analysis of risks and opportunities; modelling individuals and behaviours; mining communities, behaviours and topics. Each aspect was assigned to one Work Package (WP), the first to WP1, the second to WP3 and the third to WP5. These were supported by the infrastructure of WPs 2 and 4 which provided

Algorithm 2: V-fold Cross-Validation on C classification methods

Result: Relatively unbiased measures of true classifier performance given the set $\mathcal{T}$.

begin

Choose the number of folds V.
Partition the observations $(x_i, y_i) \in \mathcal{T}$, for $i = 1, \ldots, N$, as close as possible into V equally sized non-overlapping subsets.
Denote the subsets by $T_1, \ldots, T_V$ and let $T^{(v)} = \mathcal{T} - T_v$ for $v = 1, \ldots, V$.
Let C be the number of classification methods considered.
for $v \leftarrow 1$ **to** V **do**
Let $N_v = |T^{(v)}| \approx (V-1)N/V$ be the number of observations in all subsets excluding the v^{th} subset.
for $c \leftarrow 1$ **to** C **do**
Train c^{th} classification method on $(x_i, y_i) \in T^{(v)}$ for $i = 1, \ldots, N_v$.
Apply fitted classifier to $(x_i^*, y_i^*) \in T_v$ giving predictions $\hat{p}_i(x_i^*)$.
for $c \leftarrow 1$ **to** C **do**
Compile the full set of predictions $\hat{\mathbf{p}} = \{\hat{p}_1(x_1^*), \ldots, \hat{p}_N(x_N^*)\}$.
Apply classification quality characteristics to $\hat{\mathbf{p}}$.
Analyse performance across all C classifiers.

a parallel data processing platform and simulation and mathematical models of online communities respectively.

The global objective of WP1 was "to develop a framework for risk management that can automate real-time risk monitoring, assessment and treatment by using rules and policies, whilst supporting human intervention and understanding where necessary through live-view dashboard type interfaces" (ROBUST Consortium, 2009, p. 36). To achieve this there were three specific objectives which we quote from the ROBUST Consortium (2009, p. 36):

1. *Apply formalised risk management processes to large-scale online communities.*

2. *Develop a high-performance framework for automated and policy based risk management driven by event streams from online communities.*

3. *Provide a community health decision support dashboard for visualising the condition of online communities and their participants.*

There were five tasks to meet these more specific objectives and *real-time risk-identification and forecasting* was the task designated to the Centre of Operational Research, Management Sciences and Information Systems (CORMSIS). This task

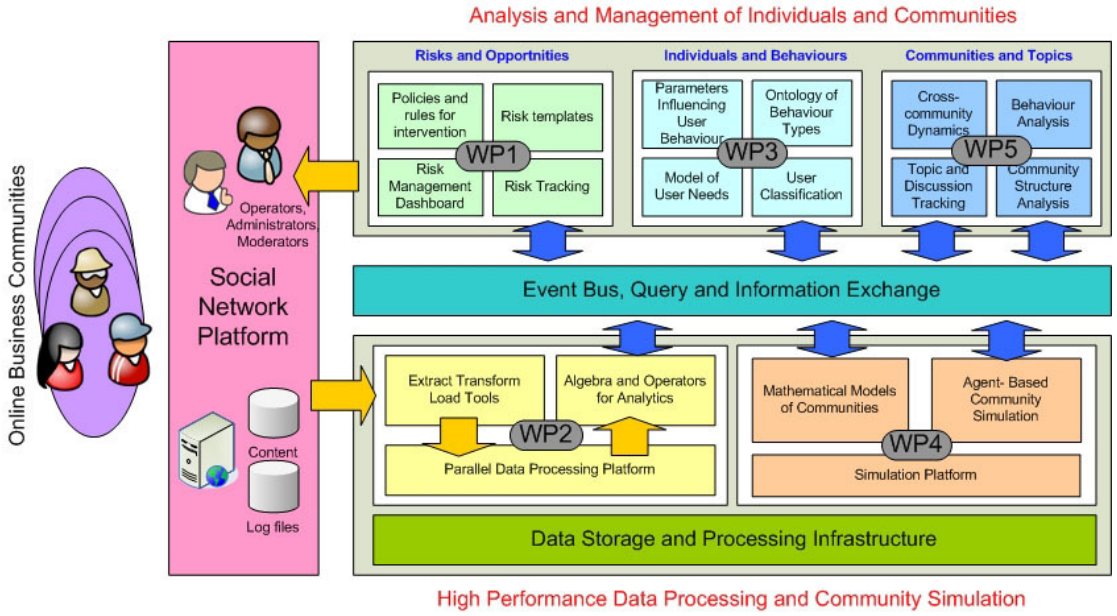


Figure 6.1: ROBUST Project Structure taken from (ROBUST Consortium, 2009) where WP stands for Work Package.

was to run from project start to completion, culminating in a final deliverable of "software for real-time risk-identification and forecasting" (ROBUST Consortium, 2009, p.37) that was to be part of deliverable number D1.3 (see Table 6.1). In agreement with the European Commission, the final deliverable was to include a fully integrated software platform that reflected the structure represented in Figure 6.1. Upon considering the similarities between Figure 1.1 and the top left of Figure 6.1, the task of real-time risk identification and forecasting can be seen to support all quadrants of the WP1 box.

Given the emphasis placed upon *real-time* community analysis and to avoid issues with software incompatibility, the ROBUST consortium chose to deliver the final platform as a web applet coded in Java. However, most consortium partners found it easier to develop and test in an array language such as Matlab or R before converting to Java for deliverables and software integration.

An overview of the final ROBUST platform implementation is given in (Janik et al., 2013). For comprehensive details of the Beta implementation of the real-time risk management framework that fulfils the third specific objective of WP1, see (Nasser et al., 2013b). A thorough description of the software framework for our implementation of the Bayesian Probit classification method can be found in (Fliege et al., 2012, Section 4.1).

Within Figure 4 of (Janik et al., 2013, p. 14) our implementation of the

Table 6.1: List of Work Package 1 ROBUST deliverables extracted from (ROBUST Consortium, 2009, Section 1.3.13.2).

Deliverable Number	Deliverable Name	Delivery Month	Deliverable Citation
D1.1	Representation of Risks in Online Communities	Nov 2011	(Nasser et al., 2011)
D1.2	Risk Monitoring and Tracking in Online Communities	Nov 2012	(Fliege et al., 2012)
D1.3	Real-time Risk Management Framework	Nov 2013	(Nasser et al., 2013b)

Bayesian Probit classification method is represented as a green rectangle labelled "WP1 GIBBS Sampler". This final implementation is fully integrated into the ROBUST platform such that it acts on features which are either extracted from live streaming data (via the *buffer* implemented by Software Mind) or from databases. Due to the questioner satisfaction formulation being developed during the final year of the ROBUST project, it was not possible to integrate the associated feature extraction procedures prior to November 2013. As a result, only the user churn event is supported in the final ROBUST platform. With this in mind, our risk assessment implementation is triggered by a call from the risk management framework (designed by Nasser et al.) represented as a green rectangle labelled "WP1 Evaluation Engine Service" which provides the following arguments:

- Event (user churn);
- Activity drop threshold - churn threshold;
- Activity filter threshold - α_q;
- Analysis period - previous activity window.

In return our classifier returns a Java object containing a probabilistic prediction for each observation (user) within the submitted sample. All source code was released by the end of 2013 and is publicly available from `https://robust.softwaremind.pl/svn/public/trunk/WP1/M30-release/`.

Given the strong statistical theme in the classifiers used, our complete inexperience in programming and fast approaching initial ROBUST deliverable (November 2011), we decided to develop our models in R.

Throughout our work, we used a version of the SCN data provided by SAP which was cleaned by Dr. Edwin Tye. During the early stages, we extracted our features directly from this Structured Query Language (SQL) database and later via a Java project which was parallelised with the help of Dr. Vegard Engen from IT Innovations (making use of all available cores of a standard desktop computer with 16GB of RAM). It was not until towards the end of the project that we integrated the parallelised code

However, to extract our novel feature sets from the Structured Query Language (SQL) data provided to us by SAP via calling SQL from R is a slow process. We were left with two options: code our models and data extraction in Java or code our models and data extraction in R and call these from Java. The latter of these caused concern within the ROBUST consortium as this would require users of the final web applet to have installed R, or to ensure that R was available on some reachable server. This, catalysed by the reputed inefficiency of R in performing looping procedures required by the iterative nature of the classification methods, lead to the ROBUST consortium requiring us to code the chosen Bayesian Probit classification model in Java. A long-term benefit of this was simplified integration with the front-end risk management aspect developed by Nasser et al. (2013a) as IT Innovations.

The work discussed in this thesis is substantially developed from its status when the final deliverables were made to the ROBUST project in late 2013. Therefore the software contained within the final WP1 deliverables is not directly comparable to the material within this thesis. Fro example, when the final deliverable was made we did not think of churn as *churn*, we had not developed the questioner satisfaction event, we had not developed the features we use now for churn and we did not have a well thought-out concept of a reputable respondent. The work in the deliverables was comparatively immature, being in the early stages of research and development.

6.4 Modelling Questioner Satisfaction

The questioner satisfaction event is modelled at the thread level given that, when a user asks a question, a corresponding thread is created. That is, the act of question asking is equivalent to the act of thread creation. This event is judged to be independent of any other such event. Whilst this (in all likelihood) is an unrealistic expectation, it is a necessary assumption for non-trivial classifiers.

For a given forum, our sample contains all threads created within that forum during the period of interest between 2008-2010 (inclusive). Sample sizes may be seen in the "Threads" column of Table 2.4. We only consider information about each thread 20 minutes after its creation; i.e. feature extraction occurs 20 minutes after question asking. Other time lags (t values) considered include 30, 60, 180 and 360 minutes to date. Results for these t values are not given within this thesis as performance does not differ sufficiently to justify later classification. Note that, unlike all other feature types (Section 2.7.3), the features about the Original Poster (OP) do not depend on t as these are extracted at the time of thread creation ($t = 0$).

We previously described our complete set of 44 features for our questioner satisfaction risk event (see Section 2.7.3) and arranged these into the seven non-overlapping feature types listed in Table 6.2. In partitioning the features by type, we can analyse the effect of the different feature types on classifier performance. For the feature types that correspond to a user type (e.g. OP), when one type produces globally superior classifiers to another, we can infer the relative level of influence that type of user has within a random thread. From such an analysis, we may gauge a relative level of informativeness across the feature types. This analysis of the individual feature types appears in Section 6.4.1.

Table 6.2: Acronyms for questioner satisfaction feature types (subsets).

Acronym	Feature type	Informativeness Order
op	Original poster	1
hps	Highest point scorer	2
mrr	Most reputable respondent	3
mrtm	Most responded to message	4
mrtu	Most responded to user	5
tmp	Temporal	6
smy	Summary	7

With a comparative level of informativeness across the feature types, we may compile a series of additive feature subsets; starting with the least informative individual subset of features and culminating with the full set of features. An analysis of these additive feature subsets (starting with the least informative feature type), can reveal the impact of the addition of more directly informative features on the ability to predict whether a questioner will be satisfied within twenty-four hours of asking a question. We analyse the feature types additively in such a way in Section 6.4.2 before discussing the cumulative observations in Section 6.4.3.

6.4.1 Features individually by type

We first consider forum 50, which displays bursty activity during 2010 and is otherwise relatively inactive (see Figure 2.10). Figures 6.2 to 6.4 illustrate the performance measures: AUC, Brier score and cost lower envelope above ROC convex hull. Recall from Chapter 5 that the main issue with one-dimensional measures is their oversimplification of classifier performance. For example, a two-dimensional ROC space cannot be comprehensively summarised by a one-dimensional measure. Here we empirically highlight this predominant weakness of one-dimensional classifier performance measures in the process of presenting our results.

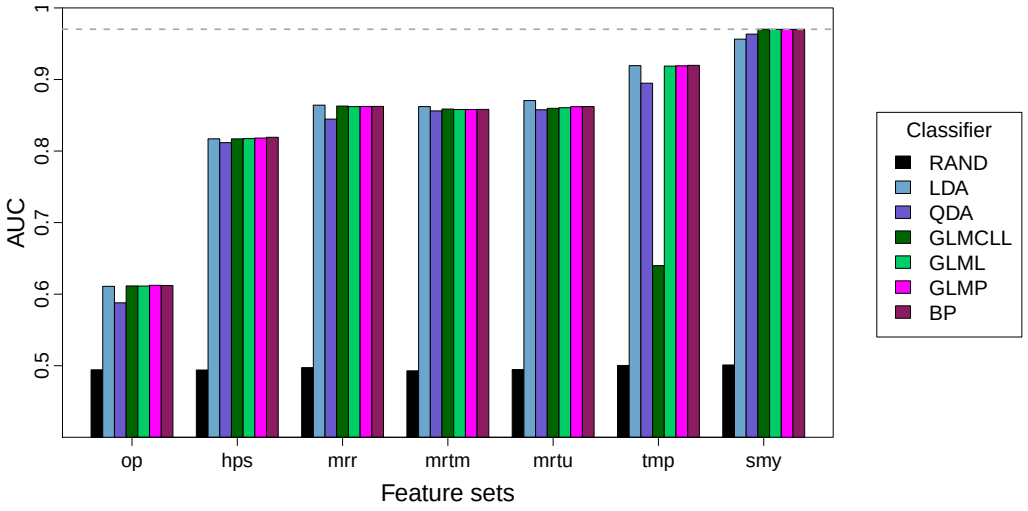


Figure 6.2: Area under ROC curve (AUC) by individual feature set for forum with identifier 50. The horizontal dashed grey line marks the highest (best) AUC measure observed.

We can infer from Figure 6.2 that the original poster features have reasonable predictive power across all classifiers. However, the remaining feature types are seen to provide significantly greater performance across all models — except for the generalised linear model with complementary log-log link function. As expected, the most reputable respondent features are highly informative — performing similarly for all measures, across all classifiers, to the most responded to feature types. Whilst the AUC measure does not imply features of highest point scorer type to be comparatively more informative, the Brier score (see Figure 6.3) presents disagreement for all classifiers but linear discriminant analysis. A significant AUC increase is seen for the temporal feature type across all classifiers (excluding the generalised linear model with complementary log-log link). However, the Brier score again presents conflicting information. All classifiers are recorded by both

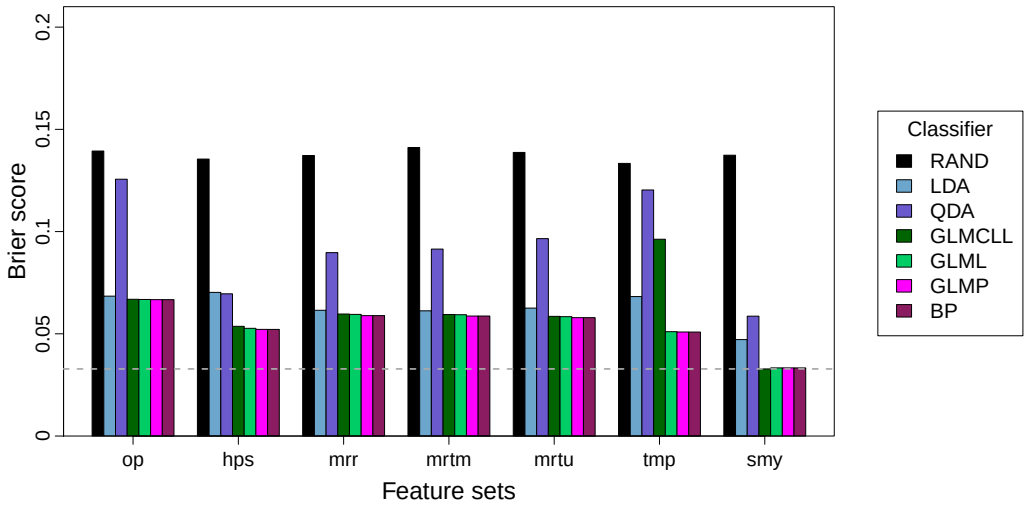


Figure 6.3: Brier score by individual feature set for forum with identifier 50. The horizontal dashed grey line marks the lowest (best) Brier score observed.

one-dimensional measures to perform best for the summary feature type.

For the other four fora considered, the corresponding figures are given in Appendix B. Similar observations to those for forum 50 stand for all fora. Again there is evidence of disagreement between the AUC and Brier score performance measures over all models and individual feature type subsets. Clearly there are inconsistencies between the two one-dimensional measures considered. Previous authors have used multiple one-dimensional measures in an attempt to capture all aspects of classifier performance. We have shown empirically that conclusions from such measures have the potential to be contradictory.

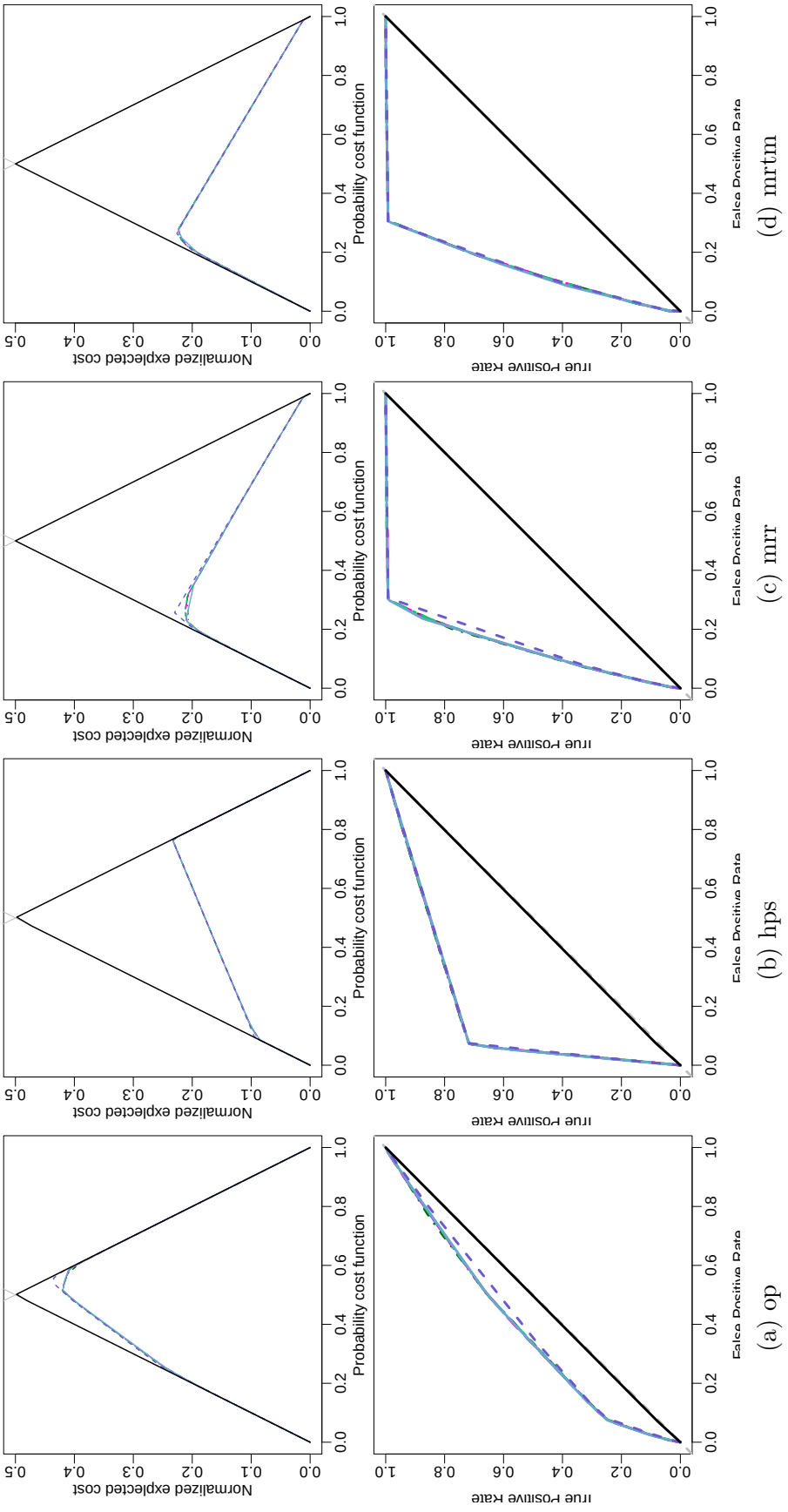
The two-dimensional spaces of Figure 6.4 correspond to the one-dimensional performance measures of Figures 6.2 and 6.3. Where the measures of AUC and Brier score have moments of incoherence, the cost and ROC spaces share a point-line duality. Supporting the one-dimensional measures, we instantly see Figure 6.4a to correspond to the poorest classifier performance, excluding the generalised linear model with complementary log-log link function in Figure 6.4f. Interestingly, although the most reputable respondent feature type generally gives AUC and Brier scores preferential to those by the highest point scorer features, in both cost and ROC space we observe instances which disagree. If the corresponding cost spaces were merged, we would see the lower envelopes to intersect approximately where the probability cost function is 0.4. For values below this, the highest point scorer features are preferential to those of the most reputable respondent. This poignant observation highlights the strength of analysing classi-

fier performance in cost space. All feature types give very skewed lower envelopes and convex hulls except for the more informative temporal and summary feature types (Figures 6.4f and 6.4g). The summary features are shown to provide a more globally superior performance than the temporal features across all models with the exception of linear discriminant analysis. Where the probability cost function is greater than 0.8, linear discriminant analysis using summary features does not appear superior to using the temporal features.

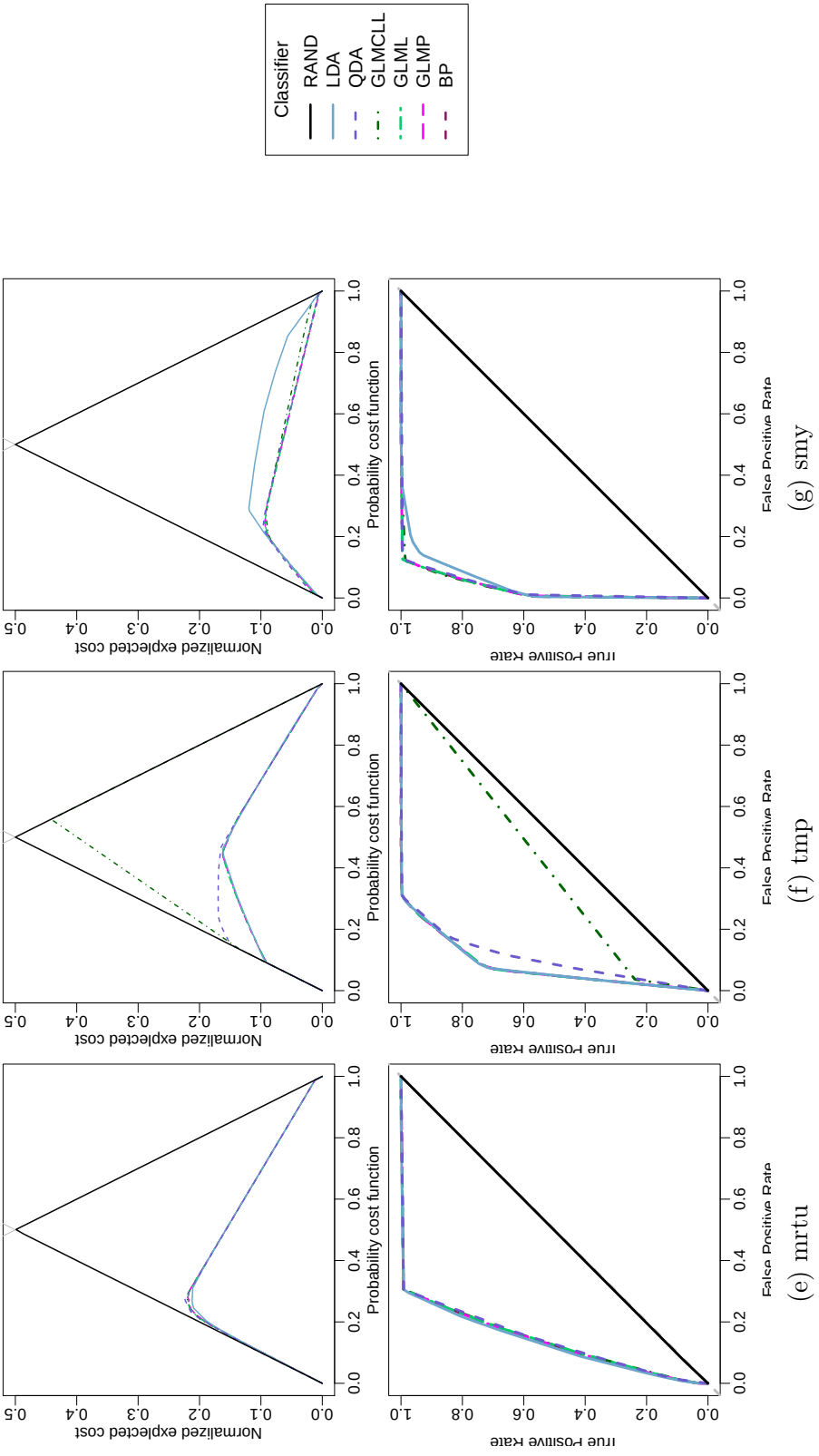
The cost and ROC spaces corresponding to the remaining four fora considered may be found in Appendix B. Those characteristics with respect to the reciprocal spaces of forum 50 are observed also for the alternate fora. Due to the findings of this section thus far, we declare the informativeness of the individual features types to rank in the order given in Table 6.2. Regarding the skew of the lower envelopes, we notice similar levels of skew across all fora. Interestingly, no matter the activity level of the forum, the original poster and highest point scorer feature types result in extreme left skew whilst the others exhibit some level of right skew. The models based upon the temporal and summary feature types have the lowest degree of skew. With the aforementioned line-point duality between cost and ROC space, we see that the lower envelope and convex hulls offer ease of access to different performance aspects. Thus we empirically show the cost space to be generally preferable for decision processes.

Across all fora of differing activity levels, the generalised linear model with complementary log-log link is unpredictably inconsistent against the other classifiers over the individual feature types. Due to rank deficiency in some group/class, we find quadratic discriminant analysis does not provide a fitted model where the feature subset is individually comprised of temporal or summary features in the lower activity fora 156 and 418. Most commonly, this occurs for class 1, which corresponds to the event that a thread is solved within 1440 minutes of creation. We hypothesise that this is because those users who ask questions which are solved, all have temporal and summary features that are sufficiently similar.

From this analysis, we have learnt that all models perform similarly for all feature types. Additionally, the summary features are the most informative in the questioner satisfaction problem formulated. We have found the two-dimensional spaces to be highly informative, with cost space appearing more tractable. In practice, curves in cost and ROC space do cross and, depending on the misclassification cost distribution, one should pick the classifier with minimal normalised expected cost corresponding to the inferred probability cost function value.



(a) op (b) hps (c) mrr (d) mrtm



(e) mrtu (f) tmp (g) smy

Figure 6.4: Lower envelope and ROC convex hulls for individual feature sets corresponding to forum with identifier 50.

6.4.2 Features additively by type

We previously analysed model performance regarding each feature type individually. We now perform a similar analysis which considers the feature types additively with respect to the order given in Table 6.2. Beginning with the original poster feature type, we add feature types one by one to observe whether any significant additive effect exists. For brevity we use the colon character between the acronym of one feature type and another to denote the subsets to which the classification methods were applied, given the ordering in Table 6.2. For example, when the original poster, highest point scorer and most reputable respondent feature types are used, this is denoted by "op:mrr". In this way, the complete set of features is denoted by "op:smy".

We initially consider forum 50 as before. Figures 6.5 to 6.7 illustrate the area under ROC curve, Brier score and lower envelope and convex hull performance measures for the additive feature sets. Obviously, performance corresponding to the original poster feature type remains unchanged. These are given as a benchmark.

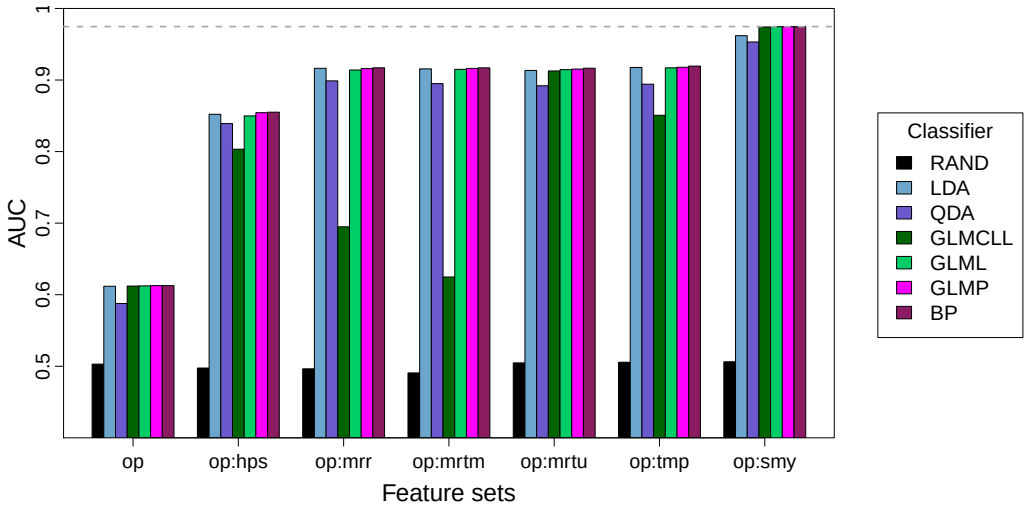


Figure 6.5: Area Under (ROC) Curve (AUC) by additive feature set for forum with identifier 50. The horizontal dashed grey line marks the highest (best) AUC measure observed.

With regard to the one-dimensional performance measures of Figures 6.5 and 6.6, we see some level of improved classifier performance with the addition of feature types. The combination of original poster and highest point scorer feature types provides comparable AUC values to the individual most reputable respondent feature type (see Figure 6.2) for most of the classifiers. By adding the most rep-

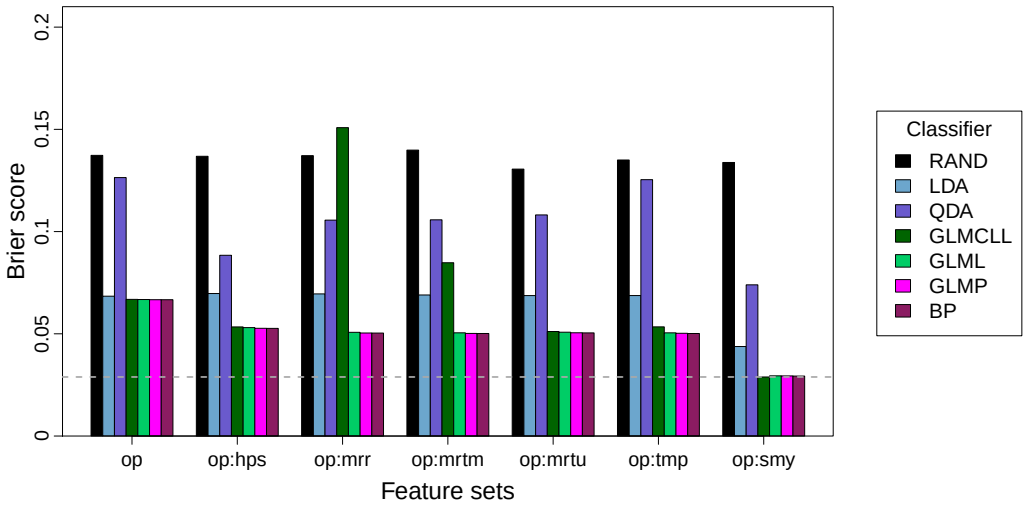
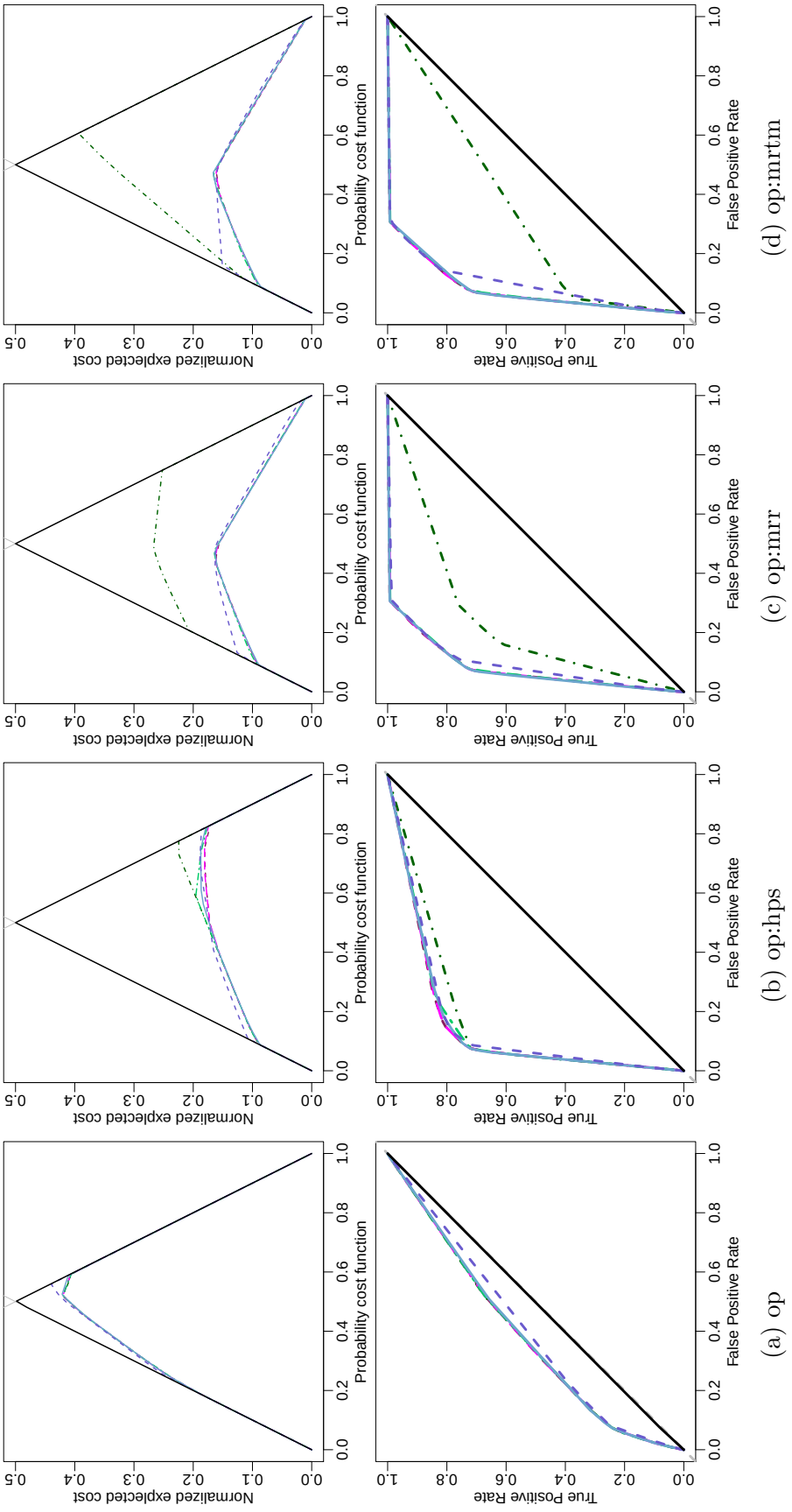


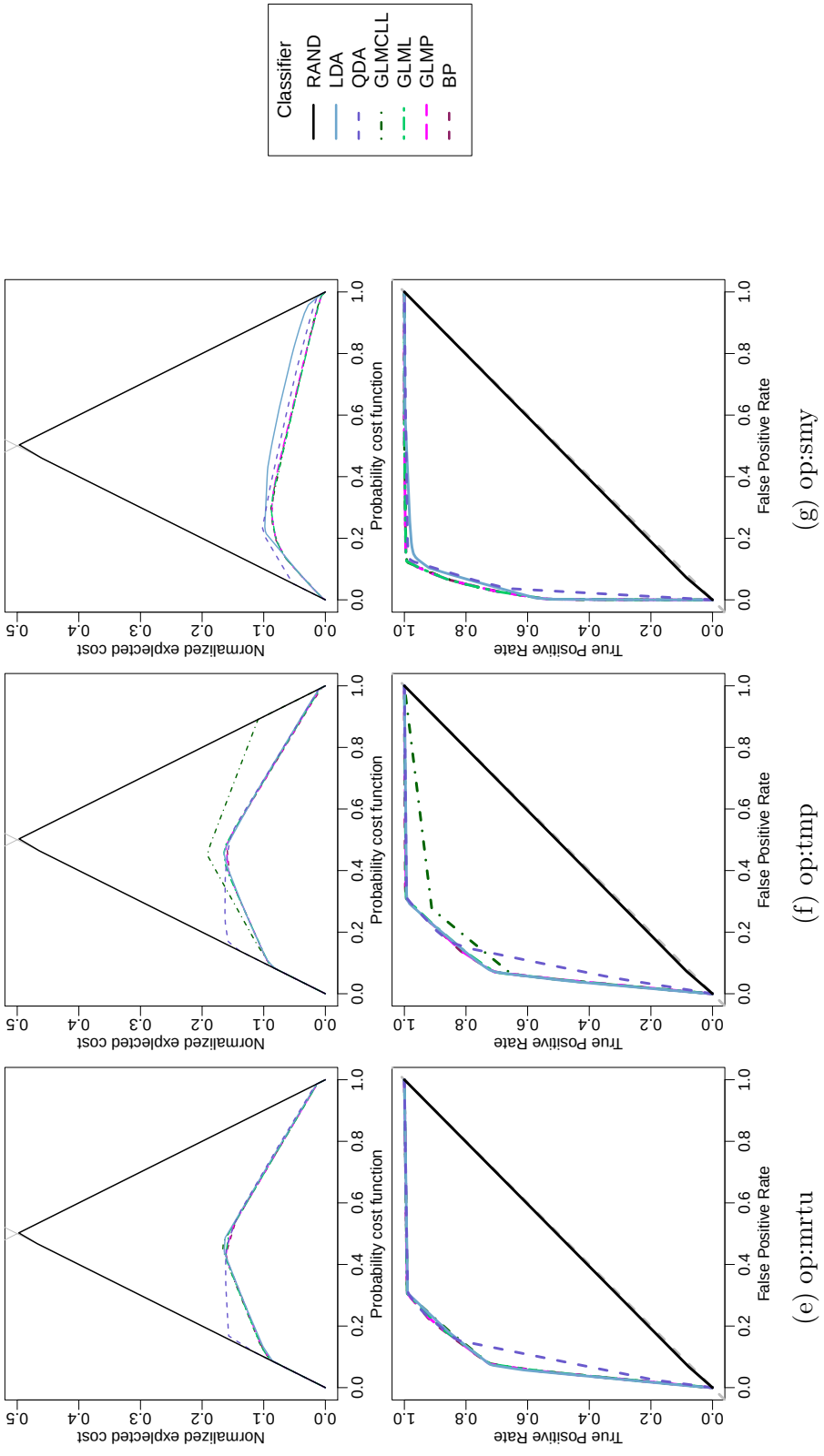
Figure 6.6: Brier score by additive feature set for forum with identifier 50. The horizontal dashed grey line marks the lowest (best) Brier score observed.

utable respondent features, we attain AUC values for most classifiers similar to the individual temporal feature type. We notice the addition of most responded to feature types does not provide any significant improvement in performance, except for the generalised linear model with complementary log-log link. However, with the inclusion of the temporal features, the AUC measure does not increase for any classifier. In fact, it appears that the combination of feature types from original poster to temporal does not provide greater AUC than when using the temporal features in isolation. A similar statement can be made about the summary feature type, although some small additive effect is observed in comparing Figure 6.5 to Figure 6.2.

Upon comparing the Brier scores between individual and additive feature sets (Figures 6.3 and 6.6 respectively), we generally observe similar findings as for the AUC. However, there are again inconsistencies between the AUC and Brier score measures. For example, Figure 6.5 shows the quadratic discriminant classifier to improve for the combination of the original poster and highest point scorer feature types, when compared with the highest point scorer features individually. However, Figure 6.6 indicates the additive model to perform worse.



(a) op (b) op:hps (c) op:mrr (d) op:mrtm



(e) op:mrtu (f) op:tmp (g) op:smy

Figure 6.7: Lower envelope and ROC convex hulls for additive feature sets corresponding to forum with identifier 50.

Figure 6.7 presents the two-dimensional performance measures corresponding to Figures 6.5 and 6.6. Those models using the feature types individually were shown in Figure 6.4 to display skewed lower envelope curves, except for the most informative feature types. In the analogous Figure 6.7, the lower envelopes of all models in cost space are less skewed. The combination of highest point scorer and most reputable respondent feature types, which were individually left and right skewed respectively, appear to take strengths from both feature types, producing lower envelope curves which are comparatively unskewed. By additively using the feature types, all classifiers (excluding the generalised linear model with complementary log-log link function) appear to have globally improved performance in both cost and ROC space. Of all the models, the best performance still corresponds to the inclusion of the summary feature type.

As for the previous individual feature subsets, those comments made regarding forum 50 generally apply to other fora with varying degrees of activity. This can be seen upon comparing the figures in Appendix C with those given above. However, with respect to the lower activity fora (156 and 418), the quadratic discriminant classifier fails to model the final two additive feature subsets (op:tmp and op:smy) because there is rank deficiency within some group/class. The same situation exist where the temporal and summary feature types are modelled individually and we hypothesise that the singularity among features is, as before, the most probably cause. In all fora, the generalised linear model with complementary log-log link again behaves inconsistently for some feature subsets. Therefore, we suggest against using the the generalised linear model with complementary log-log link function. Also, the additional assumptions and computational demand of the quadratic discriminant classifier appear unnecessary with the linear discriminant classifier performing so well. However, we still caution that business compatible misclassification costs need to be determined before settling on any one classifier, given that none are globally superior for the most informative feature subset: "op:smy".

6.4.3 Discussion

We have seen that model performance is generally more dependant on the feature types included than the classifier used. The results of Section 6.4.1 clearly show features of summary type to appear the most informative. The main value is therefore not from features which are direct evaluations of the questioner (thread

creator) but of the thread summary. We acknowledge the summary features to include an indicator of thread status. Preliminary analysis of classifier parameter values with respect to the summary feature types finds the generalised linear and Bayesian models to commonly rank this feature most highly. However, the averages of respondent reputation are closely ranked, being similar in magnitude. Linear discriminant analysis, meanwhile, ranks the feature of total points awarded highest, with some gap to the next ranking averages of respondent reputation. Of the temporal feature type, all classifiers (other than quadratic discriminant analysis and generalised linear model with complementary log-log link function) rank the first message time rank of the most responded to user most highly across the cross validation sets.

Thus we see that incorporating the rich, low-level community dynamics surrounding an original post significantly aids in determining whether a satisfactory response will be made in good time — no matter the method. In addition, we stress that these features are extracted only 20 minutes after the original post was made and are predicting whether a satisfactory response will arrive during the subsequent 1420 minutes. We find it promising towards real-time application that there is sufficient information after a mere 20 minutes to predict whether a thread will be solved (in comparatively long-term).

Regarding the differing metrics, we illustrate how the lower envelope in cost space is closely related to the convex hull in ROC space. However, we find the cost space to be more directly accessible to immediate performance assessment. We find little advantage of the ROC space representation.

For all measures, both one- and two-dimensional and both individual and additive feature set instances, we observe the generalised linear model with complementary log-log link function to behave differently to the other classifiers. The quadratic discriminant classifier is frequently inferior in cost space to the linear version, implying that, when modelling questioner satisfaction, we preferentially accept bias over variance.

6.5 Modelling Individual User Churn

Individual user churn is clearly a user level event. Those features outlined in Section 2.8.3 exploit characteristics about user behaviour individually and when interacting with others. Again, we assume all observations (here reputable respondents with $\alpha_r = 0$ in Definition 2.3) to be independent in order to satisfy

the assumptions of the non-trivial classifiers. In the formulation of the individual churn problem, the time window was determined to be one week. Predictions are therefore made for each week during the years 2008-2010 where possible.

The number of reputable respondents determines the sample size and is illustrated in Figure 2.10 for the five SCN fora considered. Within this graphic, the dashed grey line represents the number of features used in modelling the event (see Section 2.8.3). With respect to reputable respondents who make posts, forum 142 is clearly the most active, with stable population during the three year period. Both fora 141 and 50 are relatively inactive up to some point in the year 2010, when the number of reputable respondents posting per week significantly increases. Whereas forum 141 maintains the subsequent higher activity, forum 50 experiences a sharp decrease to become entirely inactive. Fora 156 and 418 are included to represent stable low activity fora within the SCN. However, we observe from Figure 2.10 that the population is rarely above the number of features, likewise for the majority of predictions for fora 141 and 50. We considered further limiting the population to only those reputable respondents who increased in reputation during the previous activity week period and this led to those sample sizes illustrated in Figure 6.8. However, we see forum 142 to be the only one for which the population (sample) size outweighs (is above) the number of features (the grey line) for the majority of the period analysed. Thus we reject this consideration for the current time.

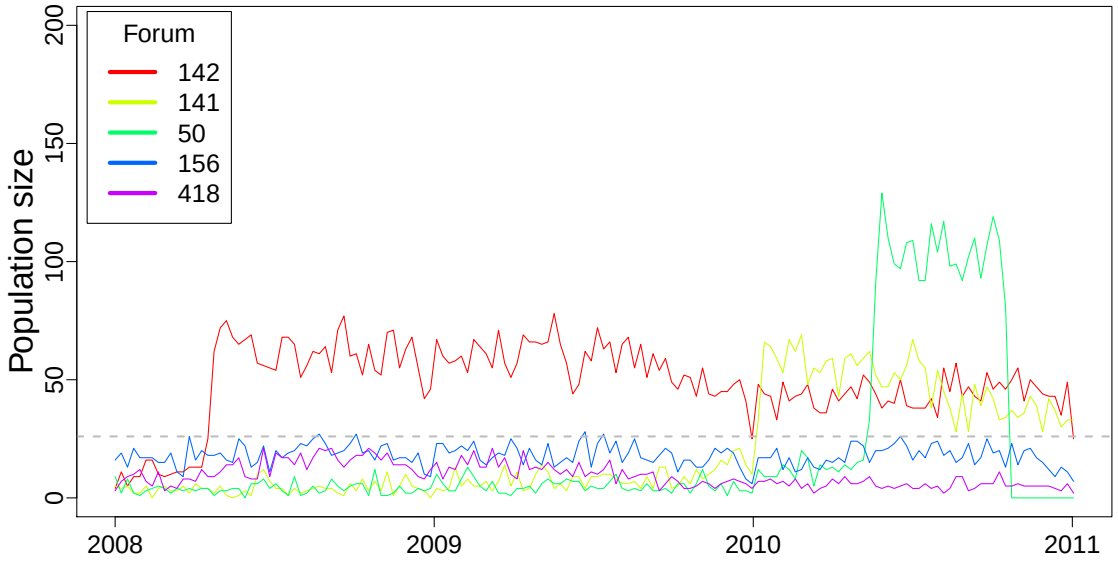


Figure 6.8: Number of reputable respondents awarded points (increasing reputation) per week in the SCN by forum. The size of the novel feature set to predict user churn is indicated by the horizontal grey line.

We find that it is not possible to meet the assumptions of all models for all fora and prediction dates. This is due either to sample sizes being too small or the churn event threshold ($T(S)$) leading one class to be too small. Small sample sizes are an issue as this leads the matrix $\mathbf{X}$ to be rank deficient. Periods for which this requirement is not met (for each of the considered fora) can be identified from Figure 2.10, where the number of reputable respondents falls beneath the horizontal dashed grey line that represents the total number of features (26).

Given that churn may be analysed as a continuous event (unlike questioner satisfaction), we additionally apply the generalised linear model with normal family and identity link function, otherwise known as linear regression. Classification is attained subsequently by empirically calculating the probability of individual user churn. As well as this additional "classifier", we also apply the same classifiers as in the previous section.

Recall that we take the average point score per time period as the measure of user activity in the definition of churn. Given the observed rate at which reputation is earned within even the most active fora of the SCN, we assume one week to be a reasonable time period over which to model user behaviour. The time series plots in Figure 2.6 illustrate that, even for the most reputable respondents in the most active forum, user activity is far too noisy when considered daily. Weekly user activity is much more smooth, being significantly less noisy. Whilst extracting features during the previous activity window of one week, we experiment by taking the subsequent churn window to be either one week (Section 6.5.1) or four weeks (Section 6.5.2).

6.5.1 Churn window is one week

Upon comparing Figures D.1 — D.3 , D.4 — D.6 and 6.9 — 6.11 for the more active fora (142,141 and 50), we immediately observe linear discriminant analysis to be far superior to the other classifiers considered. In the case of the low activity forum (156), all classifiers perform considerably less well (see Figure D.9). Those classification measures pertaining to the very low activity forum (418) (see Figures 6.12 — 6.14) appear to not follow any discernible pattern unlike those for the other fora considered.

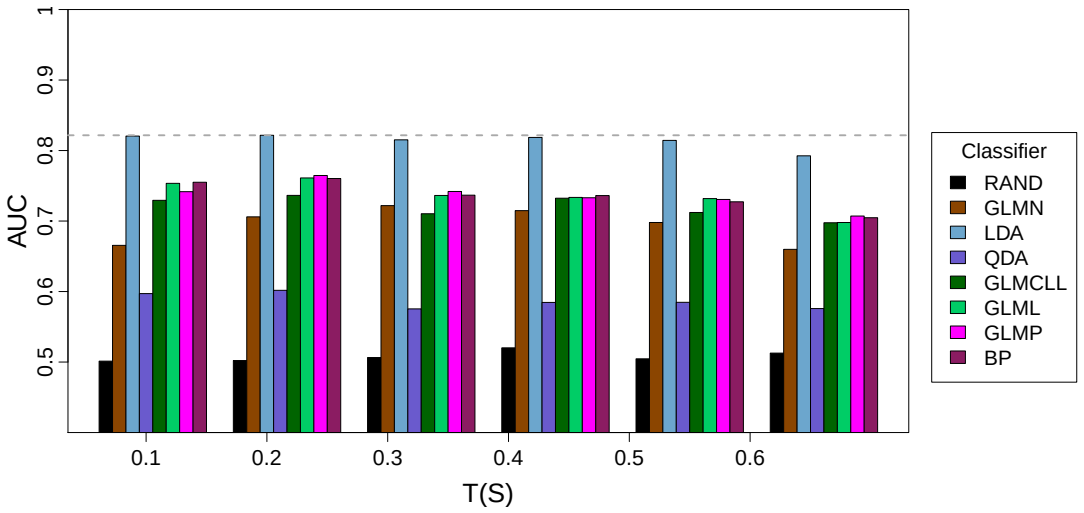


Figure 6.9: Area Under (ROC) Curve (AUC) by churn threshold $T(S)$ for forum with identifier 50 when churn window is one week. The horizontal dashed grey line marks the highest (best) AUC measure observed.

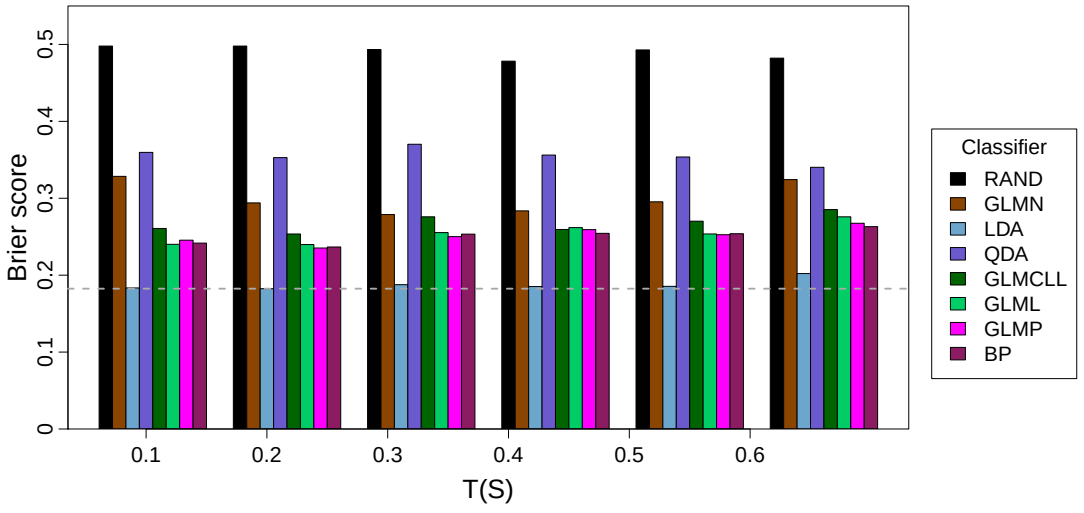
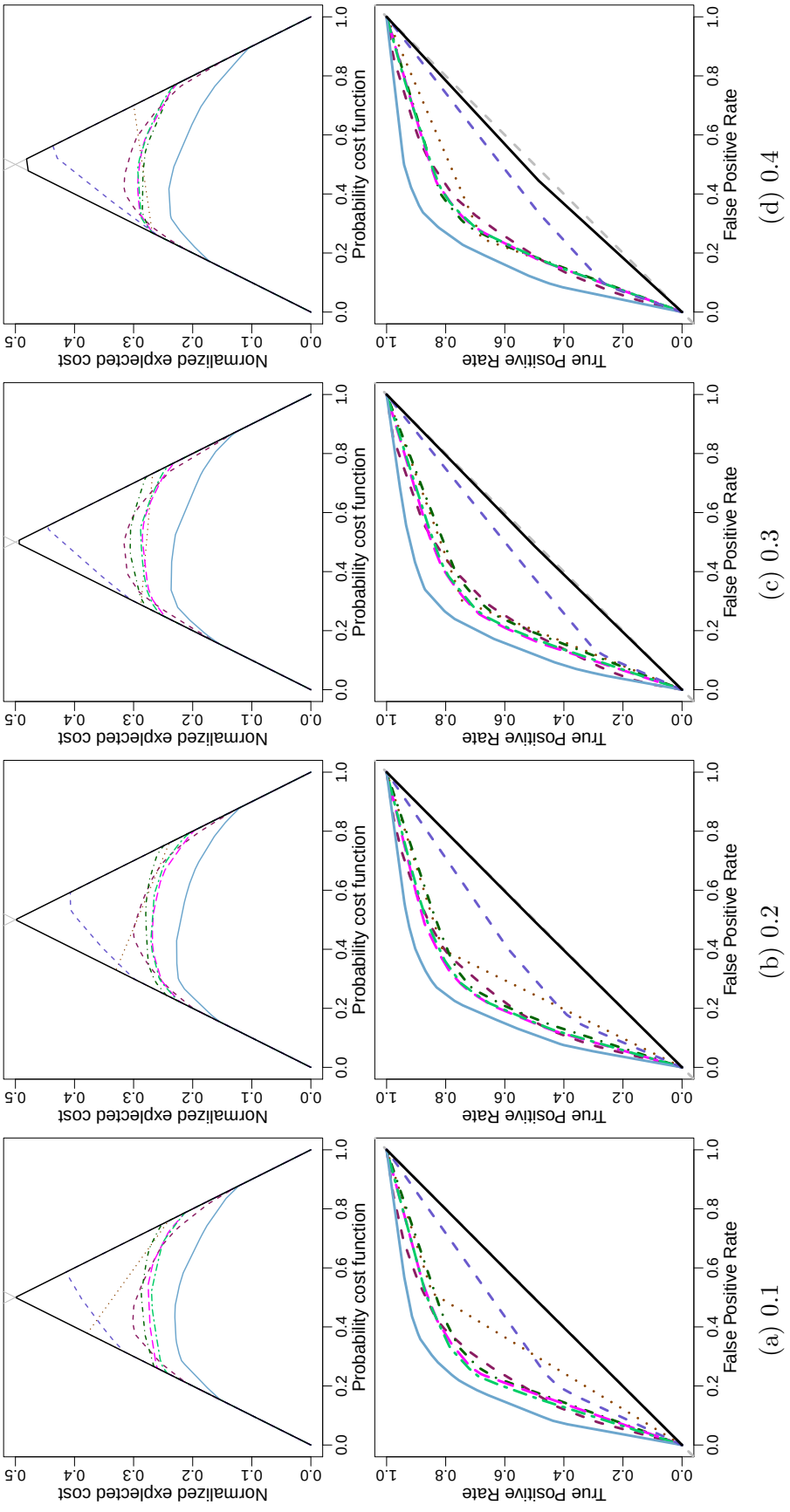
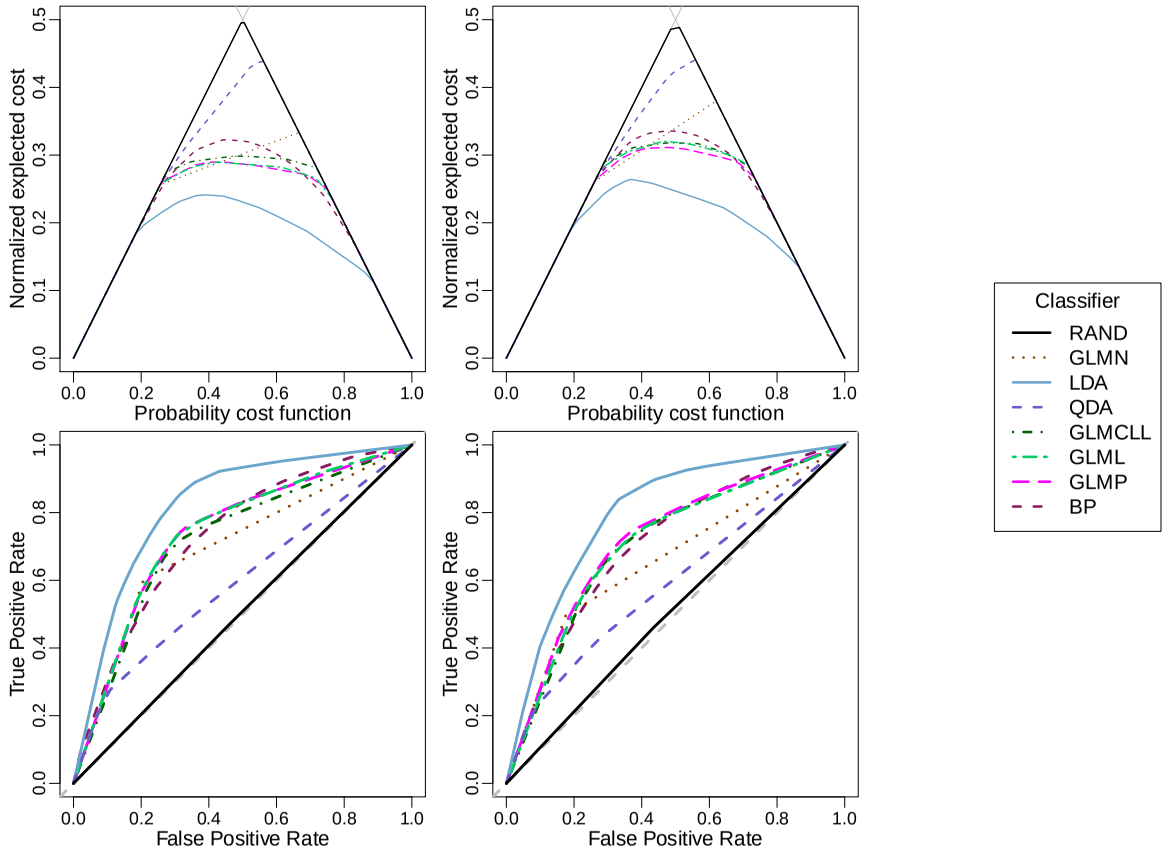


Figure 6.10: Brier score by churn threshold $T(S)$ for forum with identifier 50 when churn window is one week. The horizontal dashed grey line marks the lowest (best) Brier score measure observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure 6.11: Lower envelope and ROC convex hulls for varying churn threshold $T(S)$ corresponding to forum with identifier 50 when churn window is one week.

Recall that the performance measures presented here are taken globally over all predictions and hence all prediction dates. There are many dates for the more active fora where the models can infer a prediction but, for forum 418, there are a maximum of six dates for which some model is capable of providing a result. Those classifiers which model all six dates are linear discriminant analysis and generalised linear models without Bayesian inference. The Bayesian probit classifier models only the first of these dates (2008-09-14) which Figure 2.10 shows to correspond to the period of maximum activity for forum 418. Quadratic discriminant analysis is not possible in forum 418 for any period as the sample size of at least one class is consistently too small.

Therefore, although the relevant performance measures for forum 418 imply the Bayesian probit model to be superior globally where the churn threshold is 0.2 (Figure 6.14b) and locally where this is 0.1 (Figure 6.14a), we disregard this model, judging its inclusion to be unfair because there is only a result for one date. Following this, there is still not one model which outperforms all others across all churn thresholds. The AUC statistic (Figure 6.12) implies best classification to be using linear discriminant analysis on a churn threshold of 0.3, or generalised linear model with logit or probit link on a churn threshold of 0.4. Whilst the Brier score (Figure 6.13) agrees with the latter of these implications, it disagrees that linear discriminant analysis at any time outperforms the generalised linear model. However, the lower envelope and convex hull plots of Figure 6.14 clearly show linear discriminant analysis to be globally superior for churn thresholds of 0.3, 0.5 and 0.6. Additionally, Figure 6.14a shows this model is locally superior for probability cost function values above 0.45.

As expected, we find predicting the churn risk to be easier in those fora with greater activity. This is reflected by higher AUC values, lower Brier scores, more x-axis reaching lower envelopes in cost space and more top-right reaching higher convex hulls in ROC space for all models across all churn threshold levels.

Excluding forum 418, the churn threshold of 0.1 over all fora is generally mildly preferred by all classifiers. As we increase the value of the churn threshold $T(S)$, the proportion of users in the positive class decreases. This increases the sample size ratio between negative and positive classes: that is, the positive class representing churn occurrence becomes more of a minority class. A direct result of this is that we observe the classifiers to perform less well for much larger churn threshold values. These effects of increasing the value of the churn threshold are more difficult to see for those lower activity fora such as 156 and 418 because classifier

performance is stunted by small sample sizes.

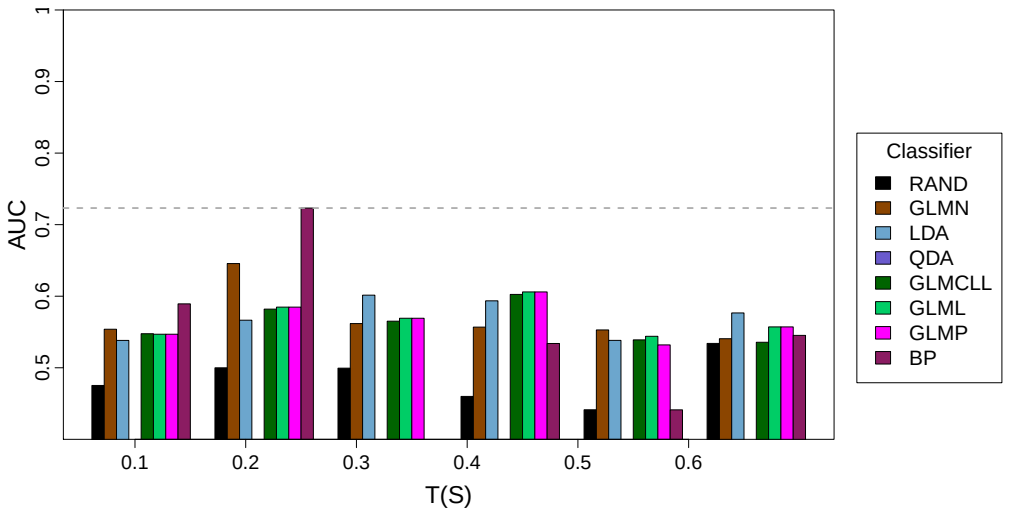


Figure 6.12: Area Under (ROC) Curve (AUC) by churn threshold $T(S)$ for forum with identifier 418 when churn window is one week. The horizontal dashed grey line marks the highest (best) AUC measure observed.

Ignoring fora 156 and 418, where the within-class-sample-size is inadequate for quadratic discriminant analysis, this classifier is consistently shown by the cost and ROC space projections to be globally inferior. As for the previously analysed questioner satisfaction risk event, the generalised linear models assuming binomial family perform similarly in all instances. The generalised linear model with normal family and identity link with subsequent classification generally performs less well than the generalised linear models with binomial family but better than quadratic discriminant analysis. The classifier with globally superior performance in most cases is linear discriminant analysis.

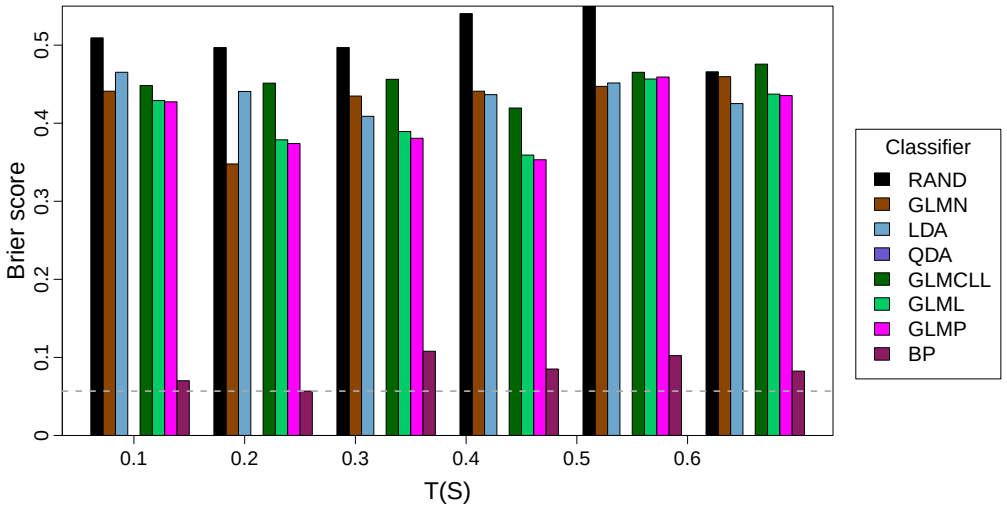
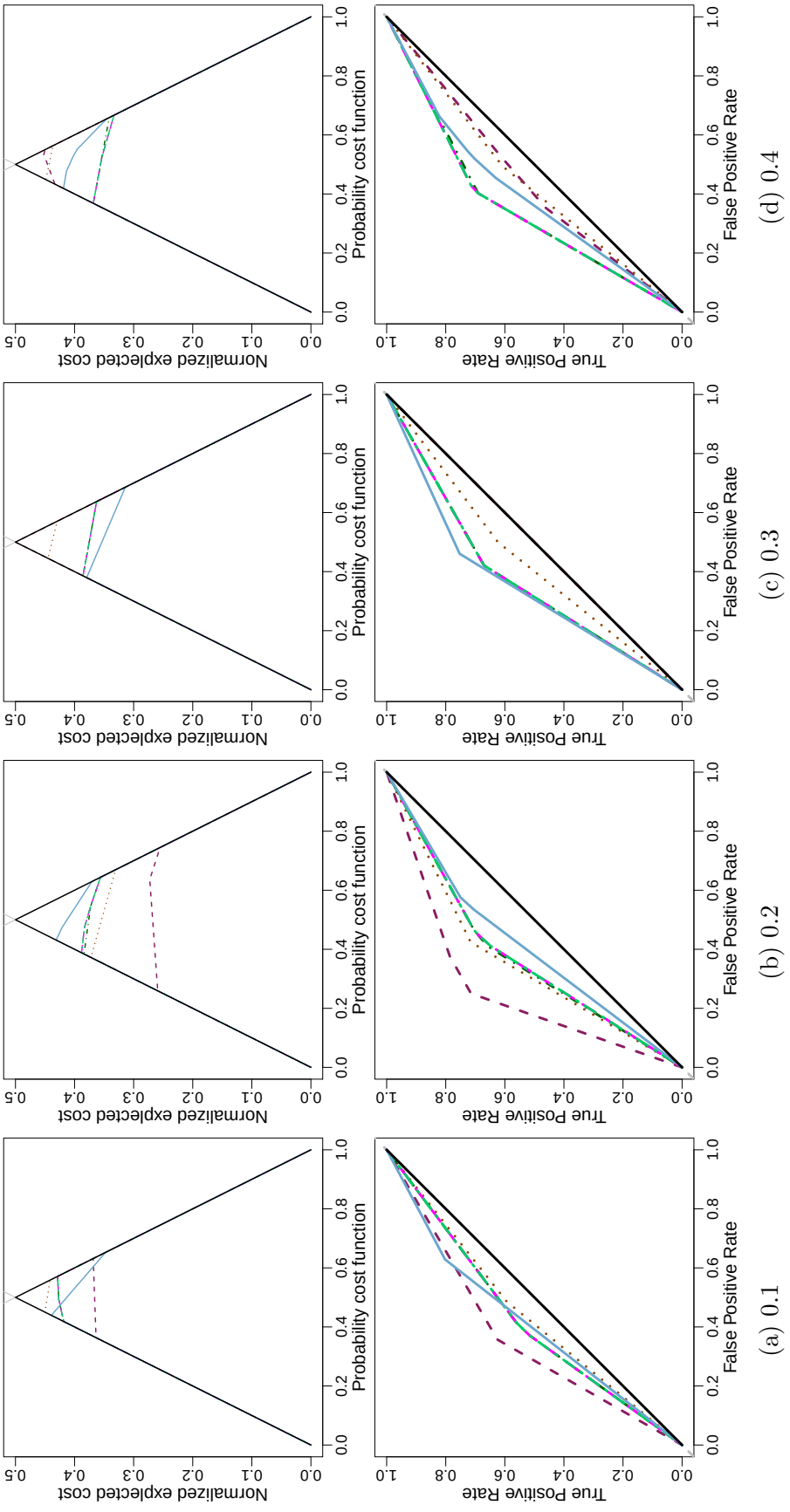
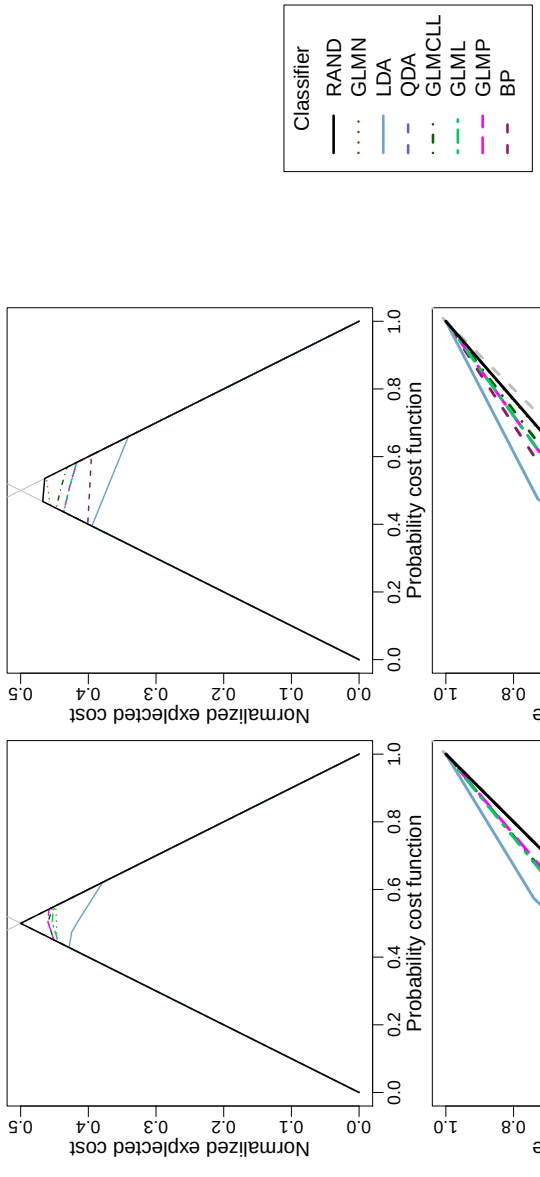


Figure 6.13: Brier score by churn threshold $T(S)$ for forum with identifier 418 when churn window is one week. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure 6.14: Lower envelope and ROC convex hulls for varying churn threshold $T(S)$ corresponding to forum with identifier 418 when churn window is one week.

6.5.2 Churn window is four weeks

We now consider the same problem but lengthen the churn window to four weeks — approximately one month. The previous activity window is maintained as one week meaning the features used are unchanged.

Comparing those figures illustrating the performance measures considered for the more active fora (142,141 and 50) in the same way as Section 6.5.1, we find linear discriminant analysis to again be globally superior to the other classifiers by a significant margin (see Figures E.4 - E.6 , E.7 - E.9 and E.1 - E.3). Both low activity fora (156 and 418) demonstrate poor classifier performance by all measures, barely improving over the random model (see Figures E.10 - E.12 and E.13 - E.15).

6.5.3 Discussion

In reality, the churn event for the SCN community is a small scale problem. We have seen that, even for the most active forum, the sample sizes are small because there are so few reputable respondents. By performing 10-fold cross-validation on this problem in an attempt to gauge more unbiased out-of-sample performance measures, the issue of scale is worsened. There are various ways to attempt to improve the scale-related issues but traditional approaches include feature selection and oversampling. Feature selection is a computationally costly process, particularly with respect to the Bayesian probit model, but it would reduce the likelihood of the feature matrix X being rank deficient. Oversampling is not suitable as, from our observations, we would argue that the small sample size is representative of the problem in hand. Additionally, in Section 2.3, we found only 15-20% of users to be be awarded points each year between 2005 and 2010. We marked these top few percent of users to be of interest to us in Section 2.8.2 and as being different in behaviour to others. Consequently we must conclude that the nature of the problem is defined by the minority of knowledgeable users — indeed this is the original motivation for modelling this problem.

We argue that the choice of the churn threshold is relatively unimportant for classifier performance. As a result, we believe this should be chosen by the community manager to reflect their view of what is a concerning level of churn for that community. For example, it may be typical for reputable respondents of high activity fora to have a relatively stable supply of threads in which to earn reputation compared to fora where fewer questions are posted. Therefore user activity in lower activity fora may appear very noisy in comparison to that

in higher activity fora. Figure 2.6 shows the activity of the user with highest reputation in lower activity fora to be less stable than for forum 142 (higher activity forum). Having observed all classifiers to struggle for low activity fora, we identify the main issue to be the development of mechanisms for performing classification for small sample sizes. However, the problem may not be one of small sample size but rather a problem of unstable user activity which changes the perspective completely. The population was found to be reasonably stable at the weekly level from the start of 2008 to the start of 2011 but this does not imply the stability of individual user activity and hence features. Unstable user features may be a sign of poor community health, or it may be a descriptive feature of very low activity fora.

As when classifying the thread-level event (results given in Section 6.4) we see that the incorporation of low-level features which describe some micro-level aspect of the community enlightens the determination as to whether user-level behaviour will change over time. In addition, where the sample sizes are smaller, significant value appears to be gained by the inclusion of features which are not directly related to the response classified. This is irrespective of the method considered (excluding the random model).

6.6 Conclusion

With respect to the questioner satisfaction risk event, we observed accurate predictions to be more dependent on the features included than the classifier used. For the most informative feature types, there is little need for additional features to improve classifier performance. Therefore, we can conclude that risk analysis corresponding to the event formulated in Section 2.7.2 is highly dependent on those underlying features which describe the behaviours and interactions of users.

We have observed the main restriction in modelling individual user churn to be the sample size. In the early stages, we anticipated having large data but there are few reputable respondents in even the most active forum of the SAP Community Network. As a result, the classifier performance in the lower activity fora suffers drastically, with no classifier able to significantly outperform the trivial random model. In contrast to the modelling of questioner satisfaction, linear discriminant analysis stood out as a superior classifier, given a reasonably sized sample.

Predicting the risk of individual user churn was found to be harder than predicting questioner satisfaction (as formulated in Sections 2.8.2 and 2.7.2 respectively).

Risk analysis of these risk formulations, especially with regard to questioner satisfaction, may be used to drive decisions about the structure of an online community such as the information to be logged and processed, and (observable) activities around which a community should be built.

Chapter 7

Discussion and Extensions

7.1 Discussion

We have shown that there is already a vast research population focused on studying online communities. Those in industry have become aware that a well-structured question and answer online community, with an integrated recognition system, is a cost-effective customer support service. SAP do not pay users of the SAP Community Network to participate or provide answers to other users. The established community is moderated by SAP employees, but knowledge is generated by users, who are motivated by earning peer-awarded reputations. Assuming that the community remains healthy (that is, knowledgeable users remain motivated to respond) the company need only employ community moderators to manage the service.

7.1.1 Objective One

Our first objective for this thesis was to explore real-time automated risk analysis of user satisfaction in question and answer online communities. We showed how our task in the ROBUST project (to perform *real-time risk-identification and forecasting*) fitted in with the wider ROBUST project and the final deliverable in Section 6.3. Therefore, meeting our first objective was crucial to the success of the ROBUST project.

In Chapter 2, we showed that user satisfaction is a vital element in maintaining the health of online communities. Within the literature, we found there to be three main approaches to modelling user satisfaction, one of which (role analysis) we did not address beyond providing an overview to avoid duplicating effort

within the ROBUST consortium. Reviewing the literature in accordance with the objectives of the SAP ROBUST partners revealed a lack of appropriate risk event formulations regarding the questioner satisfaction and churn analysis approaches (see Sections 2.7.1 and 2.8.1 respectively). In response, we formulated two new risk event formulations; the first to analyse the risk of a questioner not being satisfied within a relevant time (Section 2.7.2) and the second to analyse the risk of user churn in online communities with a rigorous concept of churn specifically for online communities (Section 2.8.2). For each risk event, our analysis was informed by a novel set of features that we developed based upon our knowledge of the SAP Community Network (SCN) (see Sections 2.7.3 and 2.8.3 respectively).

To ensure that our risk events can be analysed in real-time requires models that are stable and computationally inexpensive. In Chapter 6, we showed our risk event formulations of user satisfaction to be modelled well by simplistic and robust classifiers such as linear discriminant analysis. Both of our novel feature sets (Sections 2.7.3 and 2.8.3) are extracted from short time windows; a maximum of 20 minutes for questioner satisfaction, and one week for churn analysis. Those proven classifiers can be applied before the meta data is passed to the corresponding data warehouse, saving both processing time and effort and ensuring the real-time capability. In agreement with the ROBUST consortium, we have made available the source code for modelling the individual user churn event on stream data (see Section 6.3), therefore demonstrating that our concept is mature. This code is fully integrated within the novel risk management platform developed by Nasser et al. (2013a). The service enables SAP community managers to set up model parameters such as the previous activity and churn time windows, churn threshold and frequency of risk analysis (e.g. daily or weekly). The risk management platform then instructs risk analysis to be performed at the set frequency and resulting predictions are passed back (when available) to the risk management platform to be displayed to the community manager in a risk dashboard.

7.1.2 Objective Two

The second objective of this thesis was to investigate whether a simpler classification method would have been more suitable than the generalised linear model with probit link function under Bayesian inference that we implemented for ROBUST. (This directly relates to our third objective that is discussed in the succeeding paragraphs.) It is significant that we identified that no-one has previously

performed classification for online communities using generalised linear models under Bayesian inference (Chapter 4). Given that we integrated this Bayesian classification method into the final ROBUST software platform (Section 6.3), it is important for us to compare the performance of this method with the other classification methods that we have considered (i.e. linear discriminant analysis, quadratic discriminant analysis, generalised linear models under frequentist inference). In Chapter 4 we verified and validated that our implementation of the Bayesian probit model is capable of converging to the true underlying posterior distribution. Therefore, the Bayesian approach that we adopted for the ROBUST consortium was expected to provide benefits of which frequentist approaches are fundamentally not capable. The results of our analysis in modelling questioner satisfaction (Section 6.4) show that the Bayesian probit model performs at least as well as the other classifiers considered and that model performance is more reliant upon the features included than the classifier used. However, regarding user churn (Section 6.5), we found the unexpectedly small sample size and unstable user activity to be the predominant issues; indeed no classifier was able to significantly outperform the trivial random classifier. Where the sample size was reasonable, the linear discriminant classification method stood out as being the globally superior classifier. Therefore, in the context of the data available, the linear discriminant classification method would arguably have been more suitable than the Bayesian classification method that we incorporated into the ROBUST software platform.

7.1.3 Objective Three

The third, and final, objective of this thesis was how best to analyse and compare classifiers, considering both graphical and scalar metrics. Within our main results section (Chapter 6), we analysed and compared how the classification methods performed on our two novel risk event formulations defined in Chapter 2. Some preliminary results for one of our alternative definitions of churn in online communities is given along with the definition in Appendix F. To ensure that the study of our risk event formulations was appropriately verified and validated, we needed classification performance characteristics that enabled a traceable and fully evidenced analysis. Our summary of the most popular classification performance characteristics in Chapter 5 showed the graphical two-dimensional measures (specifically cost curves) to be most capable, whilst showing the widely used one-dimensional

scalar measures to be deceptive. This conclusion was further emphasised when we used the two-dimensional cost and ROC curves alongside one-dimensional Area Under (ROC) Curve (AUC) and Brier score measures in our empirical analysis of two of our novel risk event formulations on user satisfaction (Chapter 6).

The measures in two-dimensional space account for classifier performance over all possible misclassification cost distributions. We have taken care in this thesis not to use classifiers that assume prior knowledge of the misclassification cost distributions, which means that the cost and/or class distributional assumptions can be varied without altering the classifier. As a result, our findings in Chapter 6 are reliable and irrefutable. Our analysis in Chapter 6 has shown the features used to be more influential than the classifier selected in respect of the questioner satisfaction risk event. However, when modelling individual user churn, the linear discriminant classifier was found to be globally superior across fora of differing activity levels. As we analysed classifier performance with the two-dimensional measures, we were able to empirically prove that the choice of churn threshold does not significantly impact upon classifier performance for any possible misclassification cost distributions.

7.1.4 Conclusion

A prominent characteristic of our analysis for all our formulations of user satisfaction is that the simplistic classification methods perform well given suitably chosen features. Therefore, one should focus more on appropriately defining the problem rather than using complex classifiers on a poorly defined problem. A key aspect in developing our problem formulations was to recognise the value of perceiving each 'question' relating to a set of messages to be a series of connected information. This is a relatively new perspective to feature extraction in online communities which is becoming increasingly popular.

Just as no two people are the same, no two online communities are the same. We have shown that there are differences even between the fora of a community. Those features derived within this thesis are necessarily for the SAP Community Network and so cannot be guaranteed with respect to other communities. The most obvious example of an unsuitable application would be to a community where there is no comparable reputation system in place. Additionally, one cannot assume the structure of the data, or even the data format, to be the same from one community platform to another. To date, the applet we developed for the

ROBUST project has been used by SAP to monitor the users of the SCN in real-time, by Software Mind on users of the boards.ie online community and by Polecat on streams of Twitter data with a view to integrate this into their *fish tank* visualisation.

We therefore conclude, with regard to our stated research objectives (Section 1.4), that our research described in this thesis meets the requirements of the ROBUST consortium. In particular we:

1. Provide the means for ROBUST partners to conduct real-time risk analysis of user satisfaction noting that our tangible deliverable (i.e. our Java applet) has been successfully demonstrated within the SAP, Software Mind and Polecat user communities. We show that events which are well defined and well formulated can be predicted to provide real-time, cost-effective, risk management. We formulate two novel interpretations of risks pertaining to user satisfaction (i.e. low questioner satisfaction and individual user churn) showing that improvements in risk analysis can be achieved by focusing on the process of knowledge creation, rather than isolated actions.

2. We identified that the classification method with Bayesian inference is not more suitable than the others considered for the purpose of the ROBUST project. During our analysis of the individual churn event, we unexpectedly experienced sparseness in our data that impacted the performance of all classifiers. Our more important finding is that, regarding the questioner satisfaction event, model performance is more dependant on the features included than the classifier used.

3. In identifying and empirically proving how to best analyse and compare classifiers, we applied a series of classification methods to each of our risk event formulations. We used two-dimensional performance measures that allow us to comprehensively analyse all aspects of classifier performance which is a novel approach that is not adopted in the literature. This enables us to objectively verify our proposed models for their inclusion in an automated risk management platform. Furthermore, we find that the previously used scalar measures are unreliable and can incorrectly suggest global dominance of an only locally dominating classifier.

7.2 Extensions

The extensions suggested below address the main issues identified within the main body of this thesis. We regard alternative quantitative formulations of churn analysis for online communities to be one of the most important areas for extension. This is because there is relatively little published work on formulating a definition of churn with specific consideration to online communities. Therefore, we highlighted that there is not yet a conceptualisation of churn which can reflect the disparity between a regularly contributing user churning and an irregularly contributing user churning. Appendix F describes two additional novel formulations of the individual user churn event. The first of these formulations (Section F.1) attempts to directly address the issue and we include our preliminary results in Section F.1.1. To date, we have not had the opportunity to perform any tests on the second formulation (Section F.2) and this therefore remains an untested proposal.

Besides the above, we observed, in the process of analysis, that the individual user churn event is characterised by small sample sizes. In question and answer online communities, the reputable users are the significant minority. Thus, even for fora with greater user participation, small sample sizes can be an issue and consequently classifiers such as quadratic discriminant analysis frequently cannot be fitted. Therefore, any further consideration of events defined at the user level should attempt to address this issue.

When modelling questioner satisfaction, we found those features summarising user activity in a thread provided almost perfect classifier performance. In future, deeper analysis of the features should be performed and we would particularly wish to know whether the "indicator of thread status" feature is the most significant feature in modelling questioner satisfaction. If this were the case, there may be a relationship between an Original Poster (OP) marking his/her thread as answered and the thread being solved (and hence the questioner satisfied).

The questioner satisfaction risk event models whether a thread is solved within twenty-four hours of creation as outlined in Chapter 2. Given that the classifiers perform so well with the features extracted on this task, we consider further extensions of this event to be interesting. One extension of interest is whether users can be identified as likely to contribute knowledge to a particular thread. This event is clearly connected to research on recommender systems. We anticipate that analysing this event would require textual analysis of the original question,

and textual analysis of all questions previously responded to. Such an analysis would hope to provide some connection between each reputable user and the textual characteristics of the questions (threads) to which they provide knowledgeable responses.

The specification of a churn threshold is a fundamental aspect of the definition of individual user churn at this point in time. Although we have demonstrated that the performance of the classifiers does not vary significantly across churn thresholds, we see that there may be instances where a population-wide threshold is inappropriate. For example, it may be typical for some reputable users to fluctuate more greatly in activity than others and it may be that such users should not be held to the same churn threshold as those with more stable activity. In a similar vein, even though we limited our attention to model only the reputable users, we may not have sufficiently satisfied the assumption (of the classifiers used) that the sample population is homogeneous. Performing clustering on the population of reputable users could be an interesting exercise to overcome this potential problem.

Appendix A

Social Network Analysis

A.1 Social Network Structure

Social network analysis uses network theory (a subset of graph theory) to analyse social networks. Standard graph theory notation marks a graph G as having vertices V and edges E. In social network analysis, the graph $G = (V, E)$ represents users in an online community, with the vertices symbolising the users and the edges representing user interaction. In a directed social network, an edge from user A to user B infers that user A posts in response to user B but user B does not post in response to user A. An undirected social network with an edge between users A and B signifies that one or both users have previously posted in response to the other. Either graph type may be weighted. In a weighted graph, the edge weight is determined by the number of interactions between users — directed or otherwise.

A.2 PageRank Measure

The PageRank measure was developed by Page et al. (1999) to rank webpages by their *importance* according to the transition matrix for an "easily bored surfer". The transition matrix in (Page et al., 1999) is derived from the graph which has webpages as vertices and hyperlinks as edges. PageRank conceptualises the importance of a vertex in the graph as being proportional to the sum of the importance of all vertices which point to it. There are two extremes of importance: an *authority* which only has incoming links; and a *hub* which has only outgoing links (Kleinberg, 1999).

The PageRank algorithm is well-defined on any directed graph. Let N be the total number of vertices in the directed graph G and let $d^+(v_i)$ represent the outdegree of the vertex v_i of G. Define the transition matrix of the graph G to be $P = \{p_{ij}\}_{i,j=1}^{N}$ where

$$p_{ij} = \begin{cases} \frac{1}{d^+(v_i)} & \text{if } (v_i, v_j) \text{ is an edge in the graph } G \\ \frac{1}{N} & \text{if } d^+(v_i) = 0 \text{ } (v_i \text{ is a dangling vertex}) \\ 0 & \text{otherwise.} \end{cases}$$

Let E be some $N \times N$ matrix and c be a dampening parameter. With transition matrix P, the *perturbed Google matrix* is defined by

$$\tilde{P} = cP + (1-c)\frac{1}{N}E.$$

The largest eigenvalue of $\tilde{P}$ is $\lambda = 1$; for all other eigenvalues λ we have $|\lambda| \leq c$. Let R denote an eigenvector for the largest eigenvalue with $||R||_1 = 1$ and $R \geq 0$. Then R_i can be interpreted as the probability that a random surfer is found at vertex v_i within the directed graph G (Ermann et al., 2012). The vertices can consequently be sorted by decreasing probabilities to determine the vertex rank $K(v_i)$ that reflects the importance of v_i.

Definition A.1.

The *PageRank probability* of vertex v_i is R_i, where R is the eigenvector as above.

Whereas Page et al. (1999) define PageRank for a network of webpages linked by hyperlinks, the vertices in the case of social networks represent users whilst the edges represent user v_i posting to user v_j. For a question and answer network, the more unique reputable users who directly respond (in-links) to user v_i whilst user v_i maintains minimal activity (out-links), the higher the user v_i is ranked within the network and thus the more influential user v_i is measured to be (Karnstedt et al., 2010a). The measures resulting from PageRank represent the global influence well but do not fairly represent users on a relative scale.

Appendix B

Modelling Questioner Satisfaction: individual features

B.1 Forum 142: very high activity

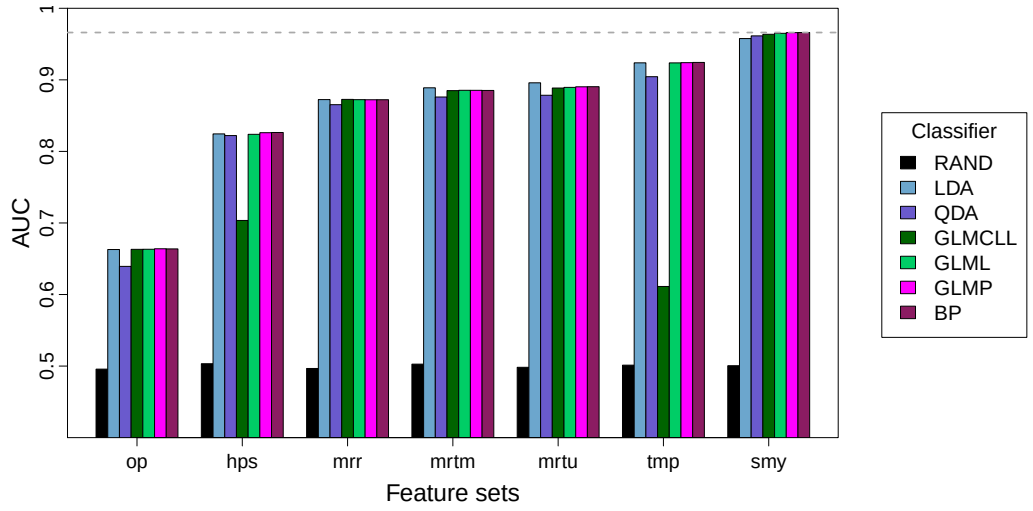


Figure B.1: Area Under (ROC) Curve (AUC) by individual feature set for forum with identifier 142. The horizontal dashed grey line marks the highest (best) AUC measure observed.

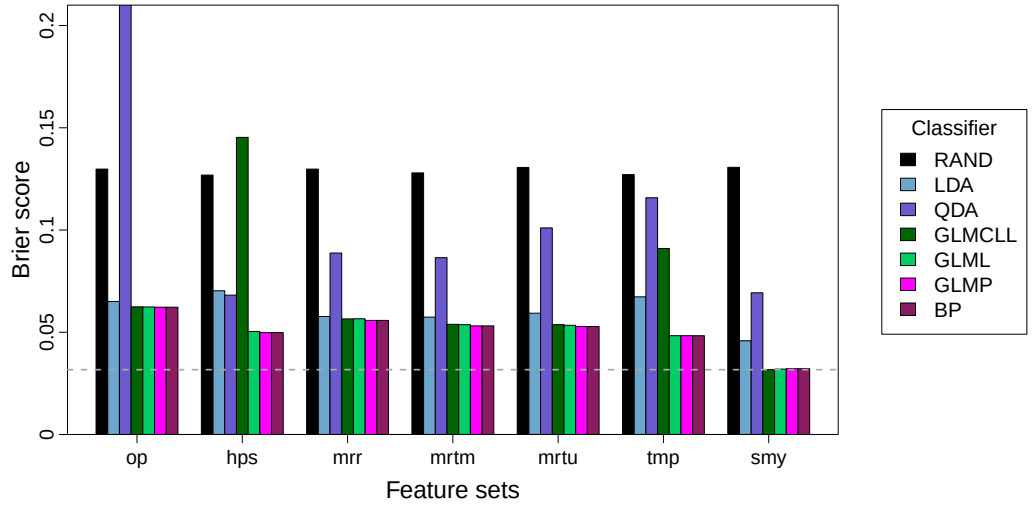
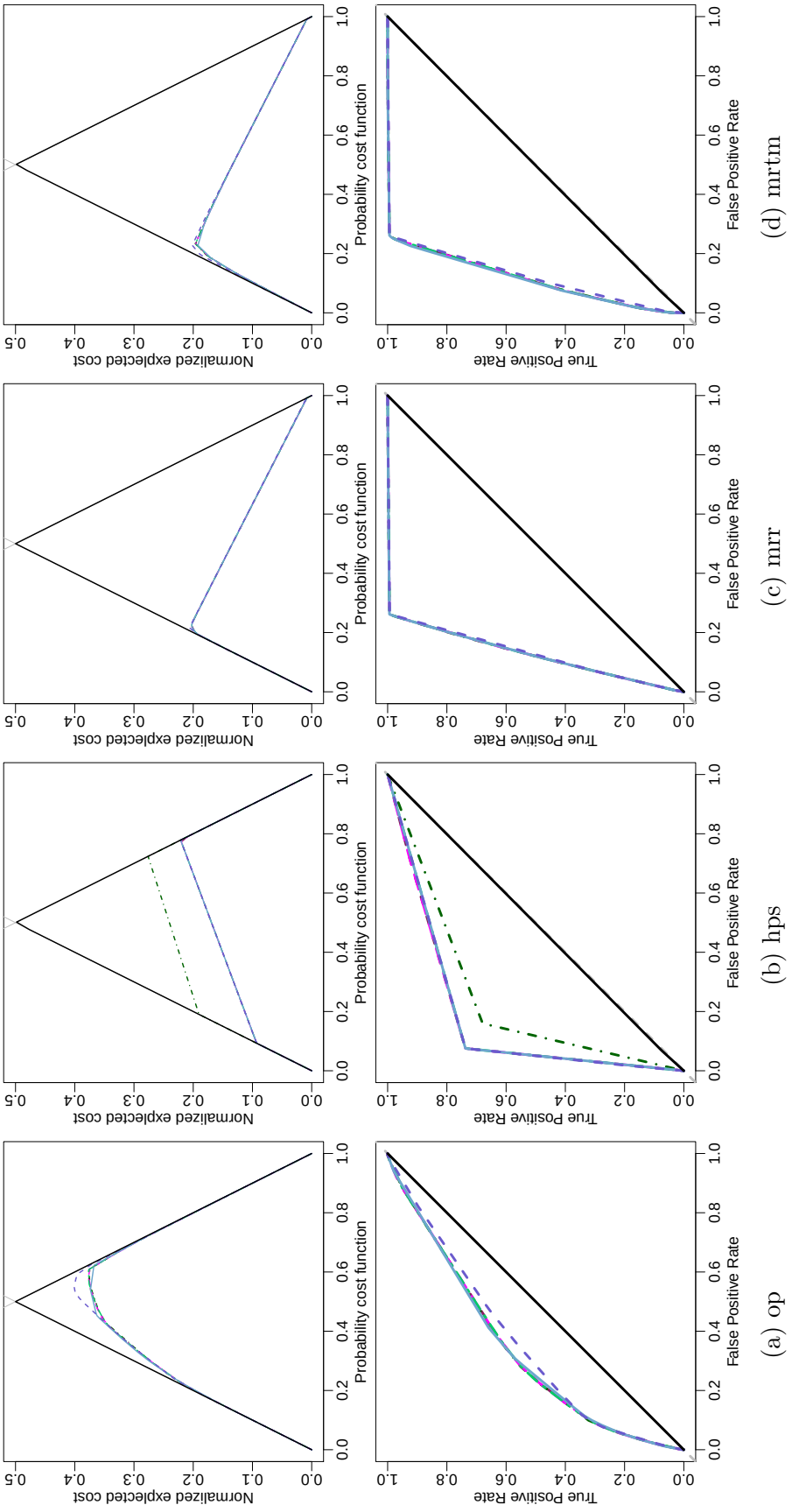
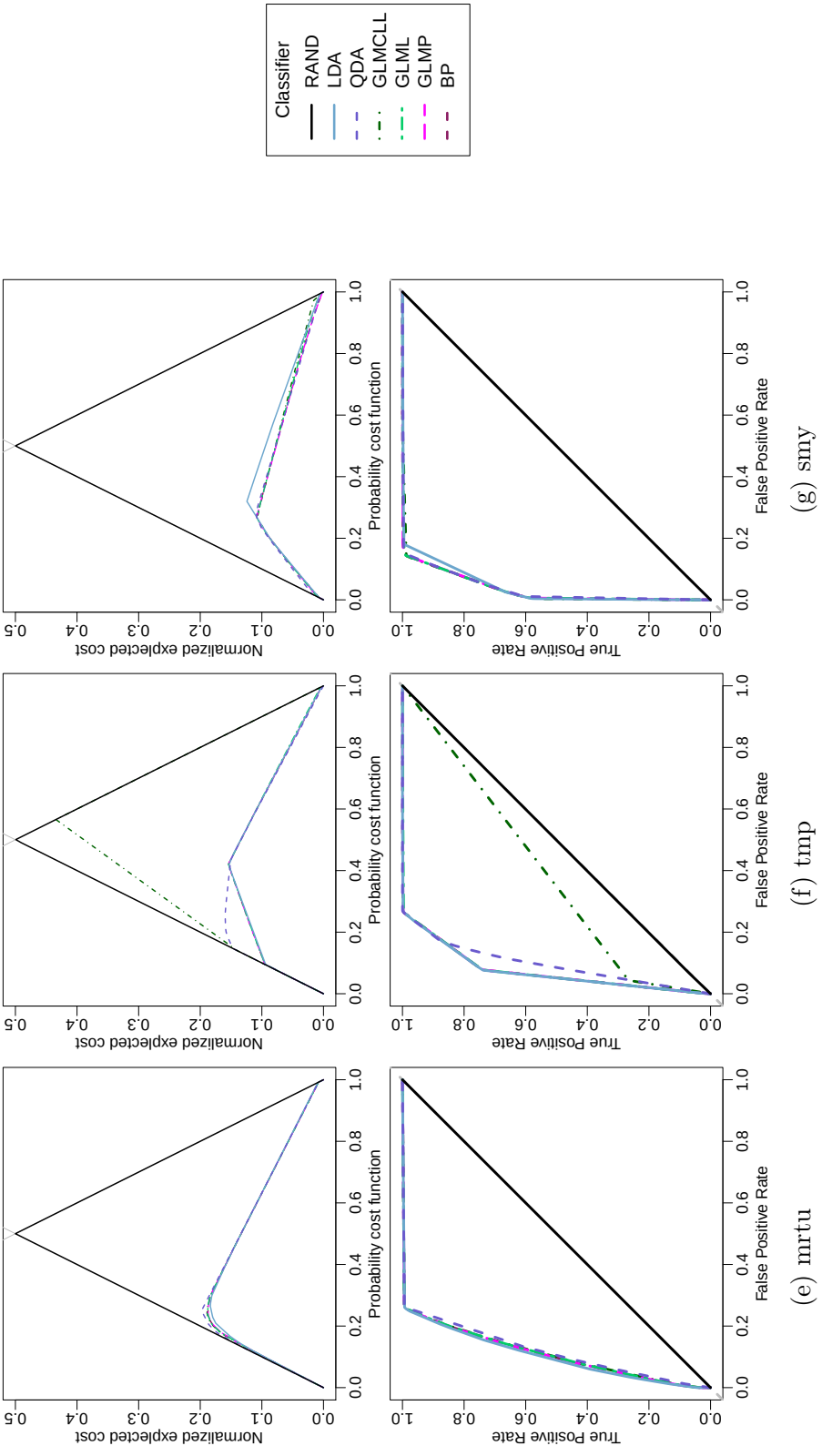


Figure B.2: Brier score by individual feature set for forum with identifier 142. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) op (b) hps (c) mrr (d) mrtm



(e) mrtu (f) tmp (g) smy

Figure B.3: Lower envelope and ROC convex hulls for individual feature sets corresponding to forum with identifier 142.

B.2 Forum 141: high activity

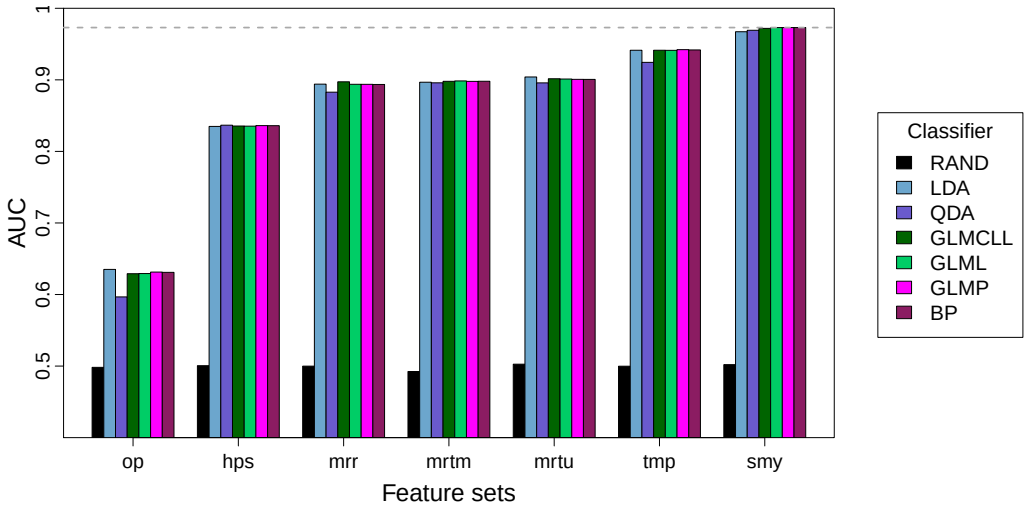


Figure B.4: Area Under (ROC) Curve (AUC) by individual feature set for forum with identifier 141. The horizontal dashed grey line marks the highest (best) AUC measure observed.

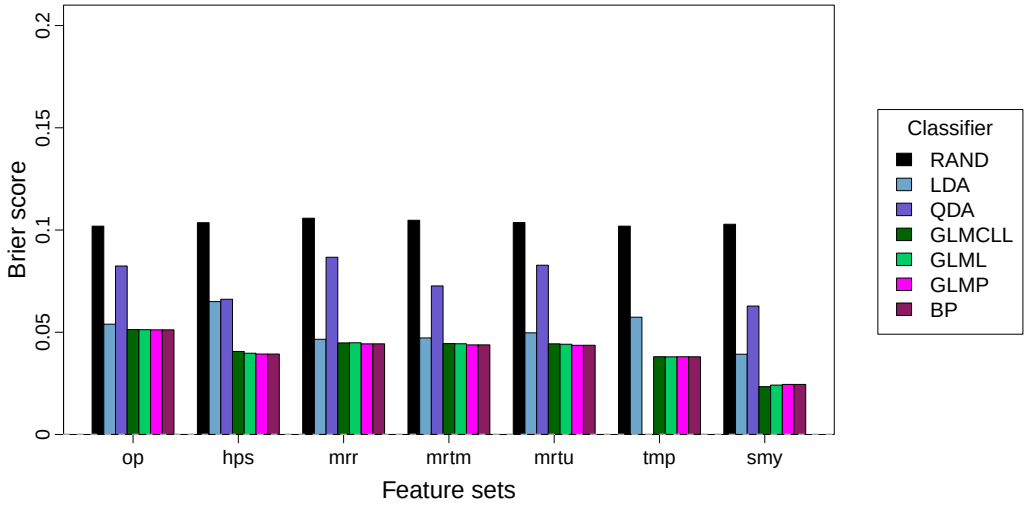
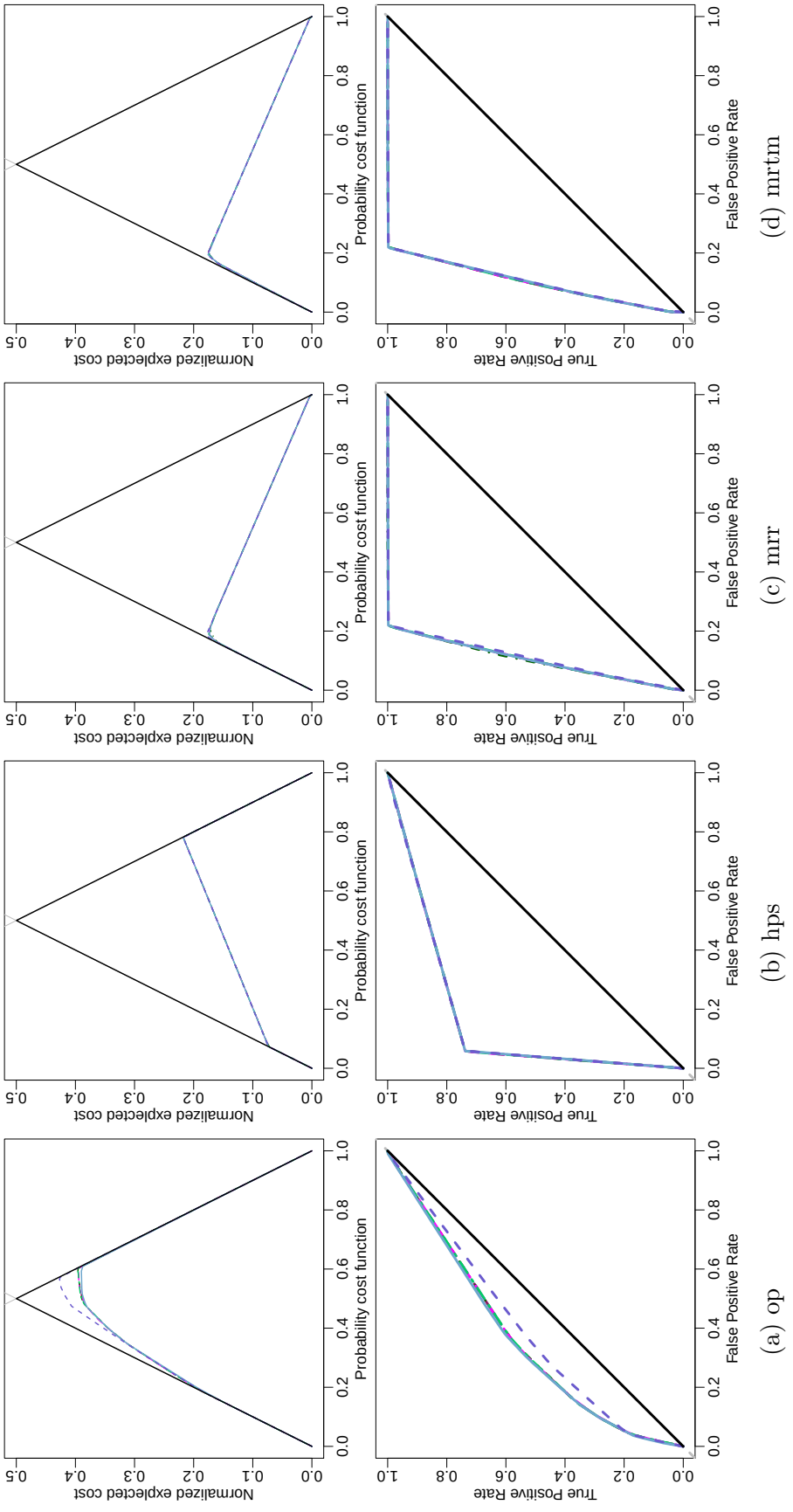
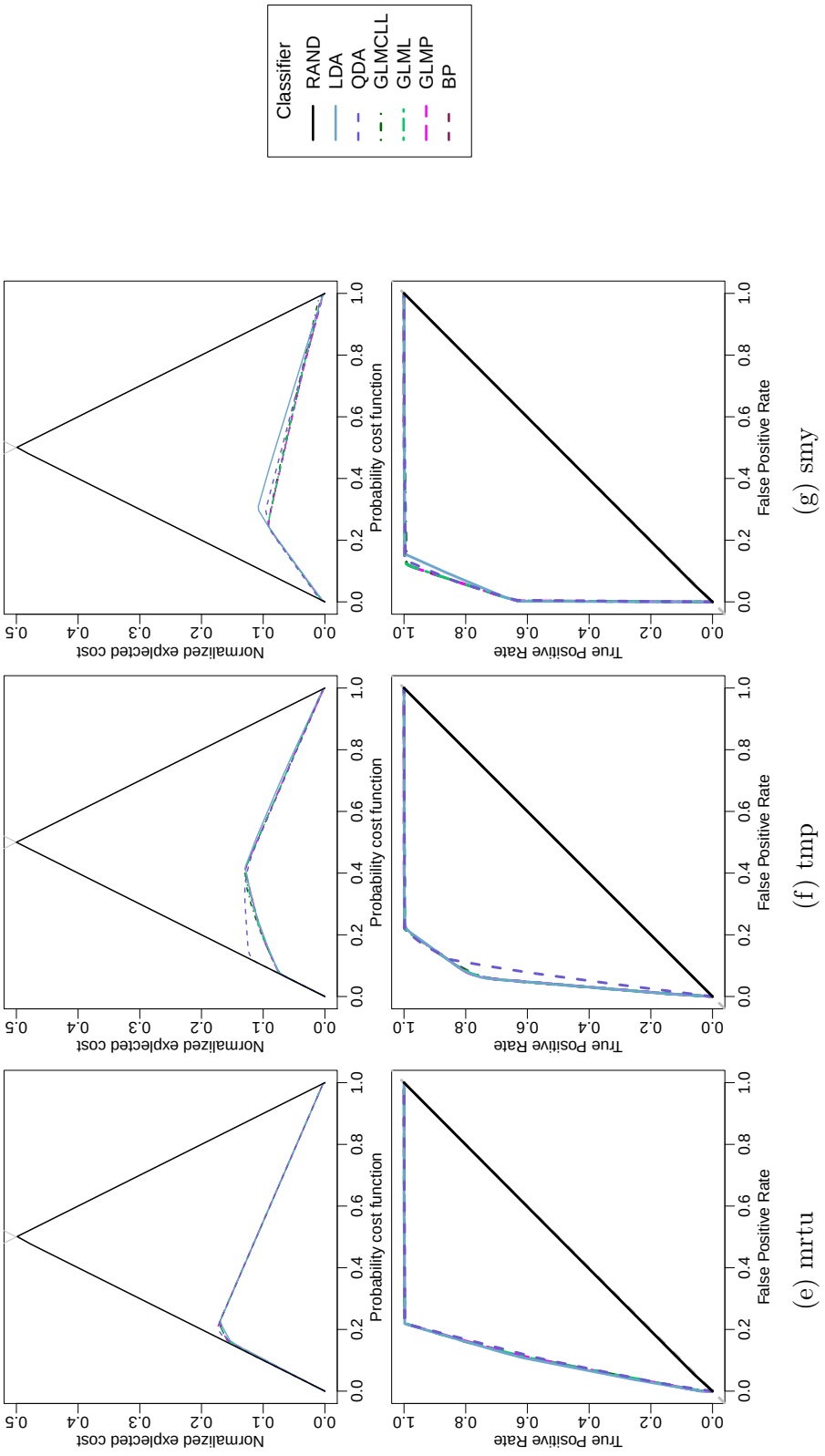


Figure B.5: Brier score by individual feature set for forum with identifier 141. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) op (b) hps (c) mrr (d) mrtm



(e) mrtu (f) tmp (g) smy

Figure B.6: Lower envelope and ROC convex hulls for individual feature sets corresponding to forum with identifier 141.

B.3 Forum 156: low activity

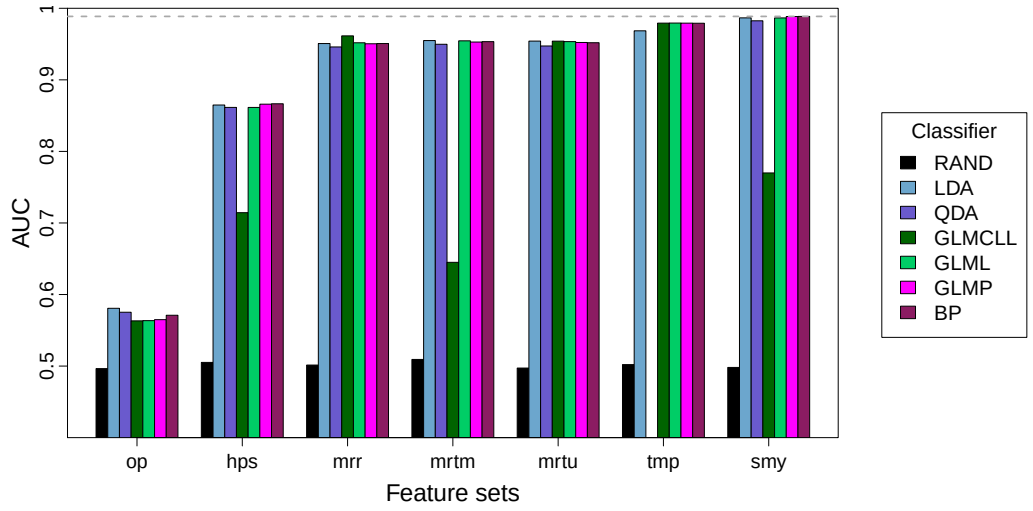


Figure B.7: Area Under (ROC) Curve (AUC) by individual feature set for forum with identifier 156. The horizontal dashed grey line marks the highest (best) AUC measure observed.

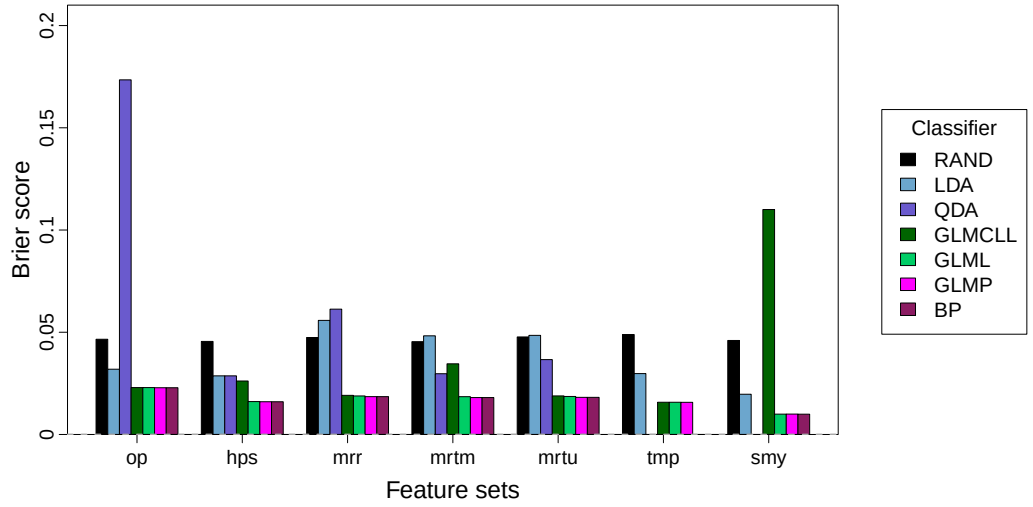
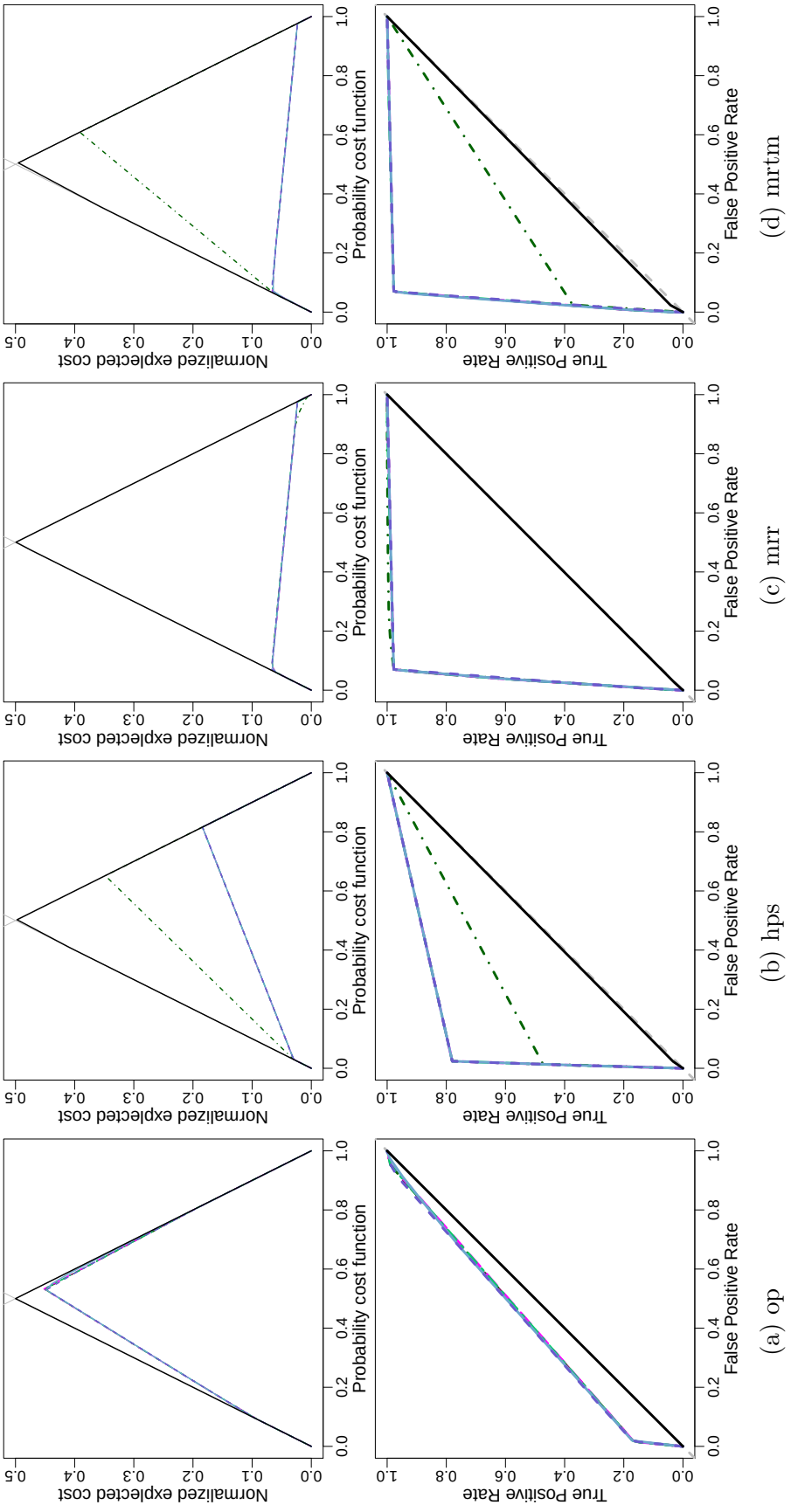
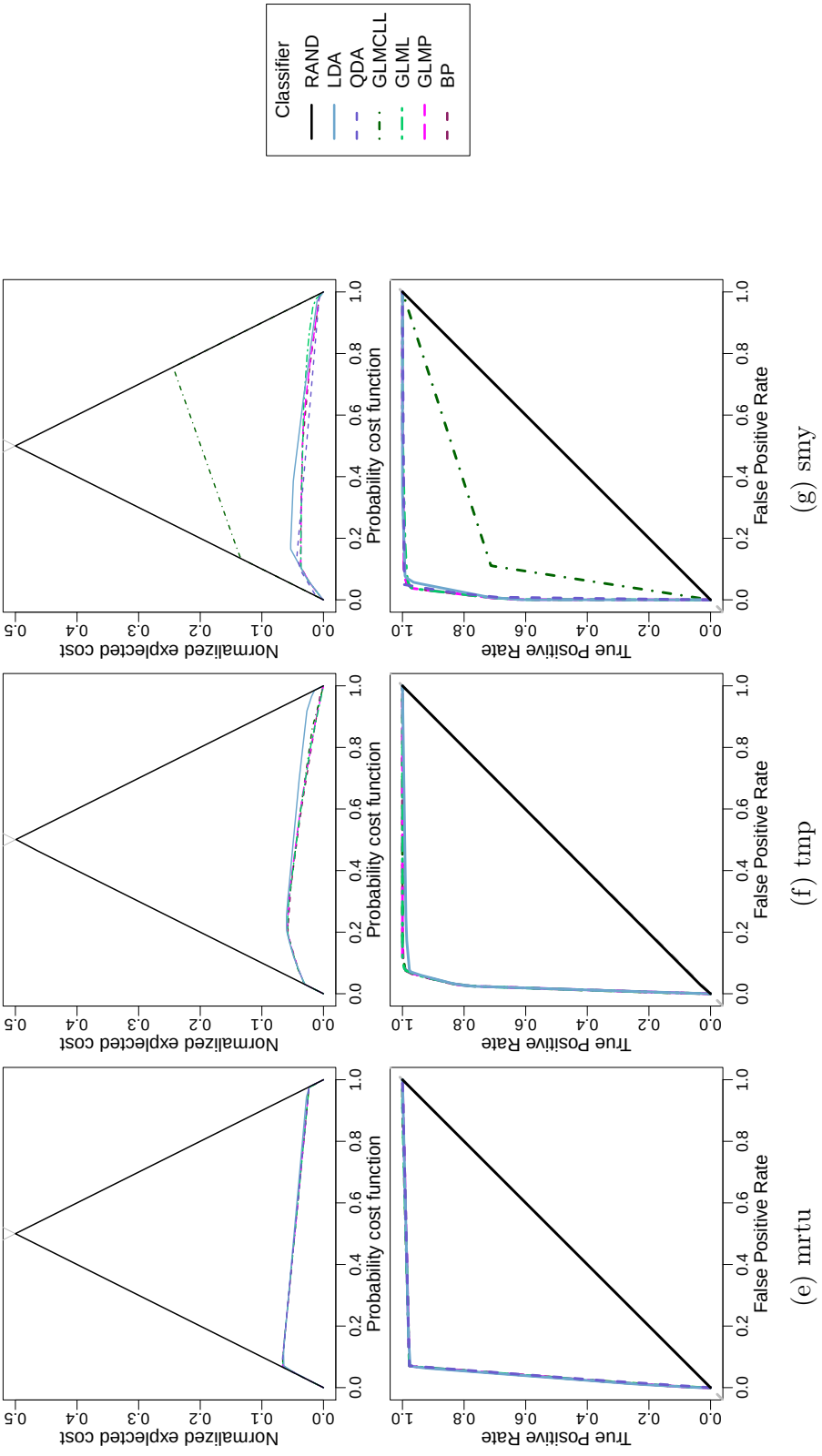


Figure B.8: Brier score by individual feature set for forum with identifier 156. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) op (b) hps (c) mrr (d) mrtm



(e) mrtu (f) tmp (g) smy

Figure B.9: Lower envelope and ROC convex hulls for individual feature sets corresponding to forum with identifier 156.

B.4 Forum 418: very low activity

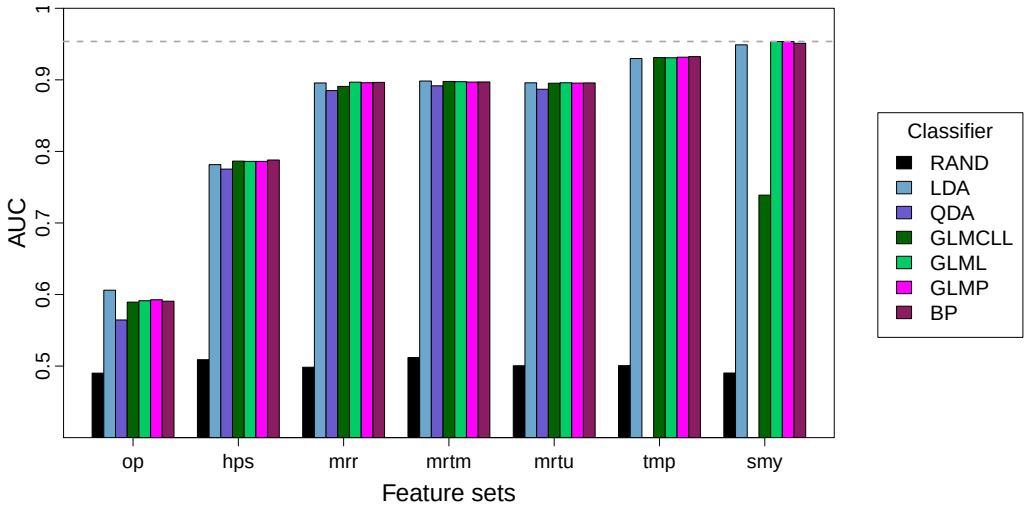


Figure B.10: Area Under (ROC) Curve (AUC) by individual feature set for forum with identifier 418. The horizontal dashed grey line marks the highest (best) AUC measure observed.

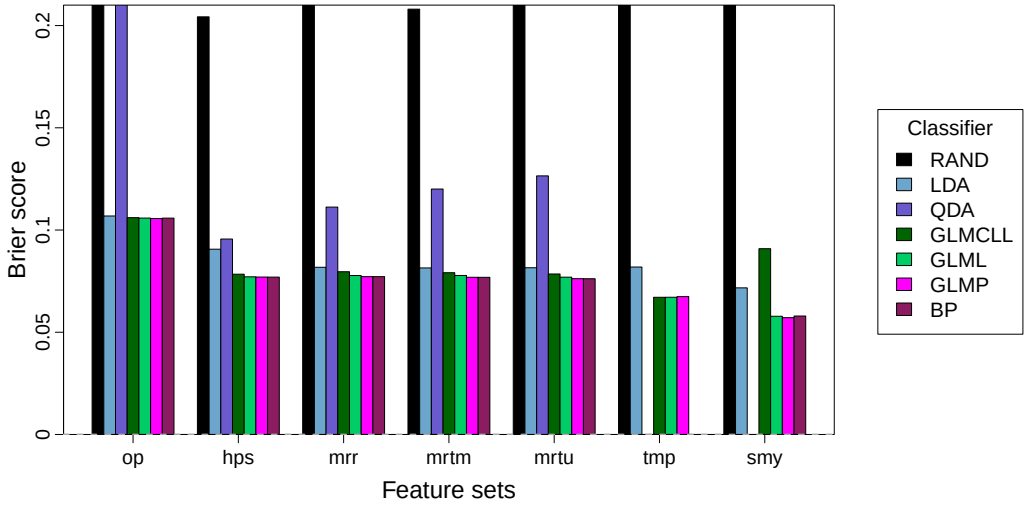
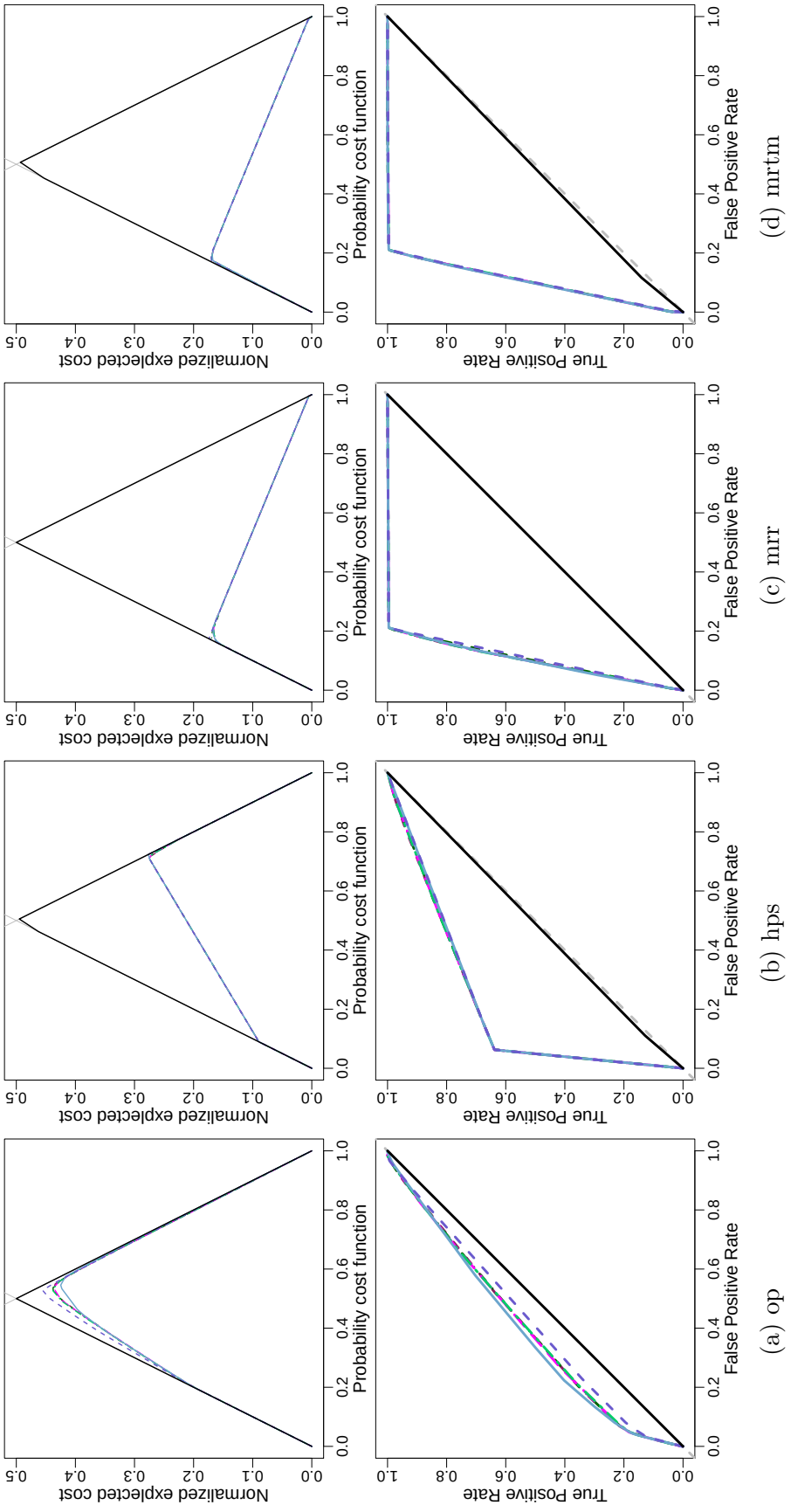
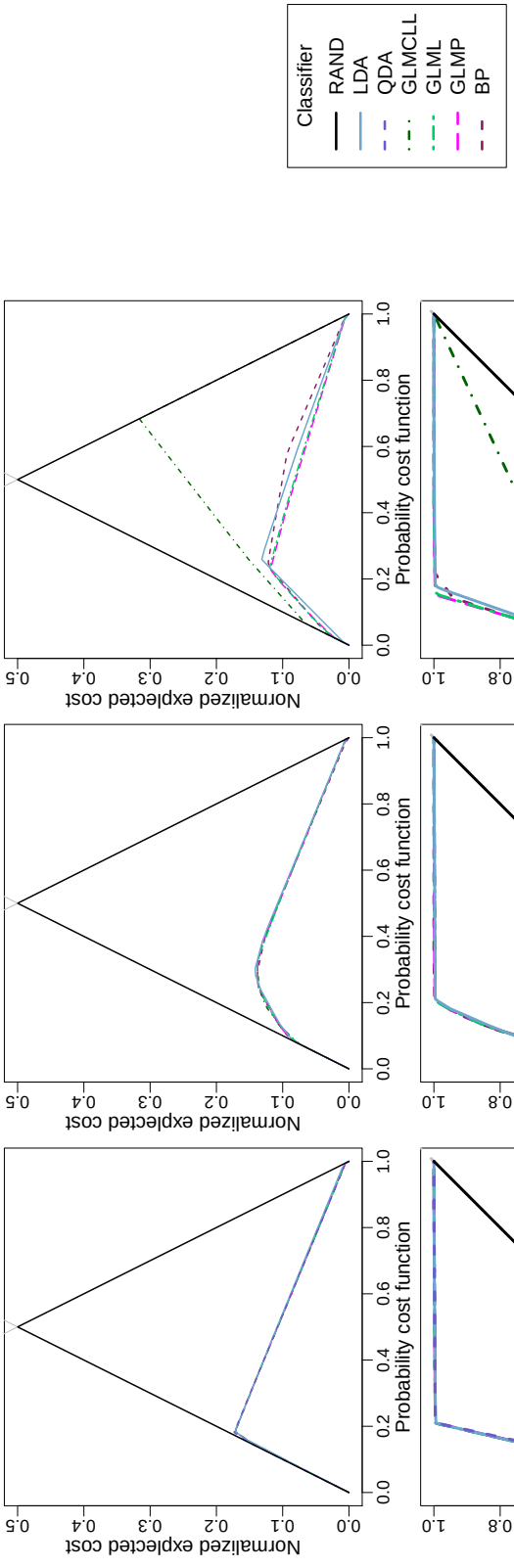


Figure B.11: Brier score by individual feature set for forum with identifier 418. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) op (b) hps (c) mrr (d) mrtm



(e) mrtu (f) tmp (g) smy

Figure B.12: Lower envelope and ROC convex hulls for individual feature sets corresponding to forum with identifier 418.

Appendix C

Modelling Questioner Satisfaction: Additive features

C.1 Forum 142: very high activity

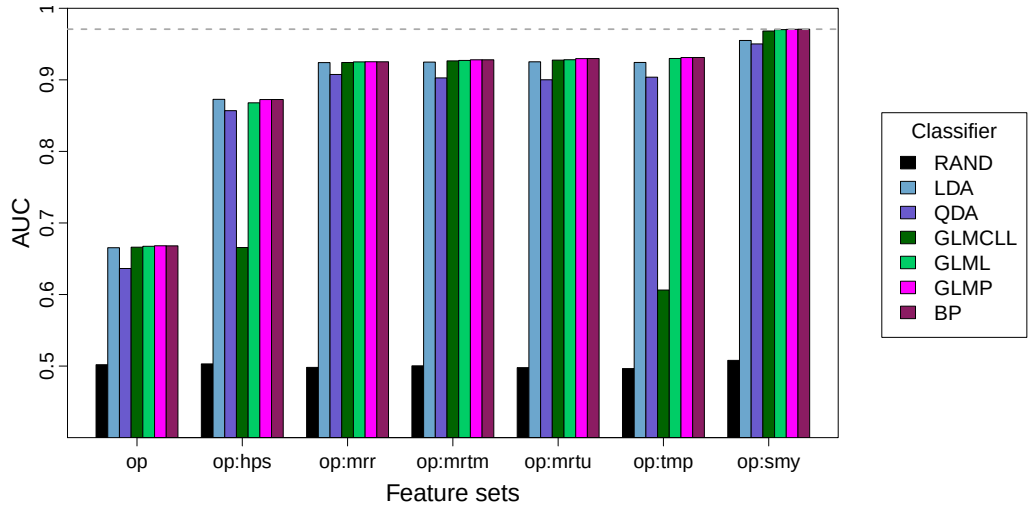


Figure C.1: Area Under (ROC) Curve (AUC) by additive feature sets for forum with identifier 142. The horizontal dashed grey line marks the highest (best) AUC measure observed.

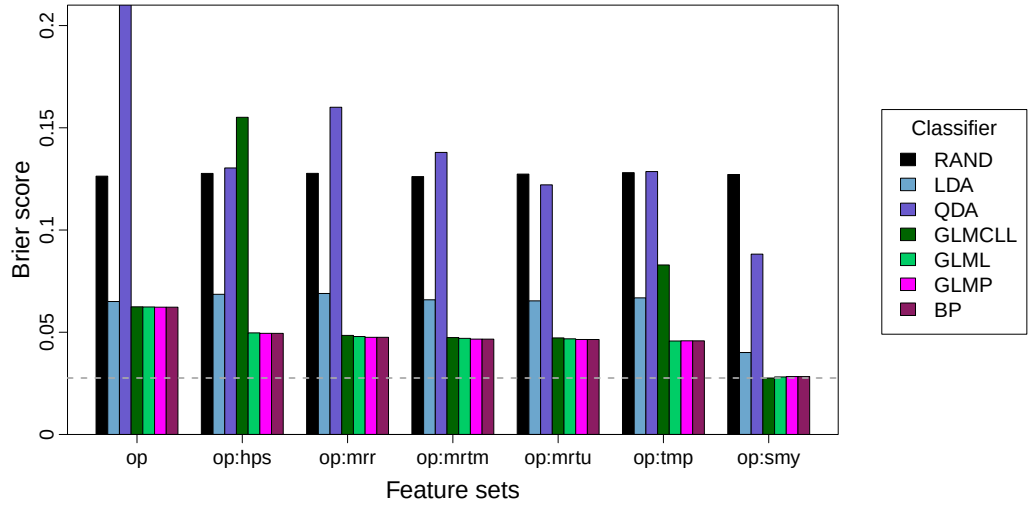
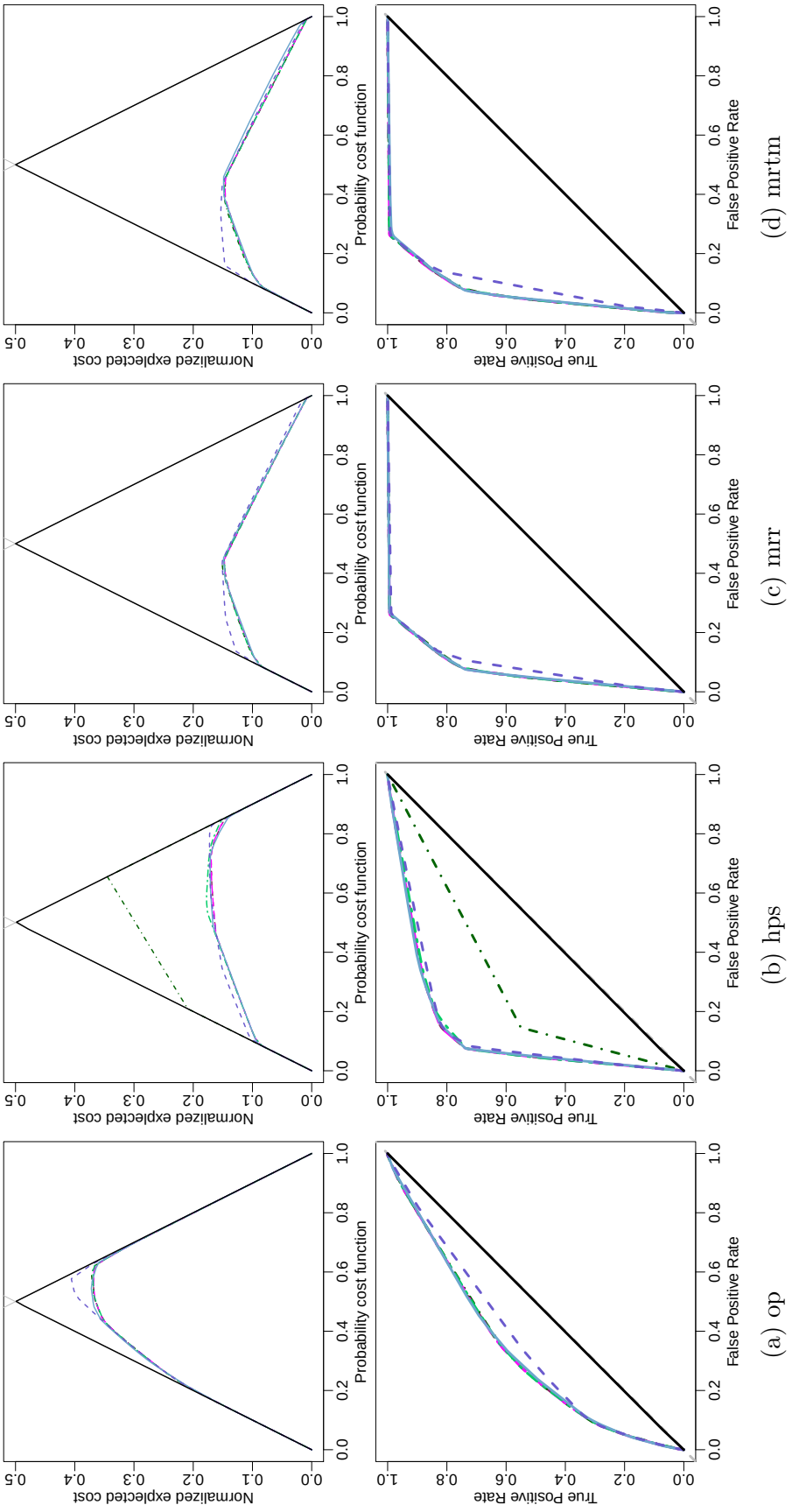
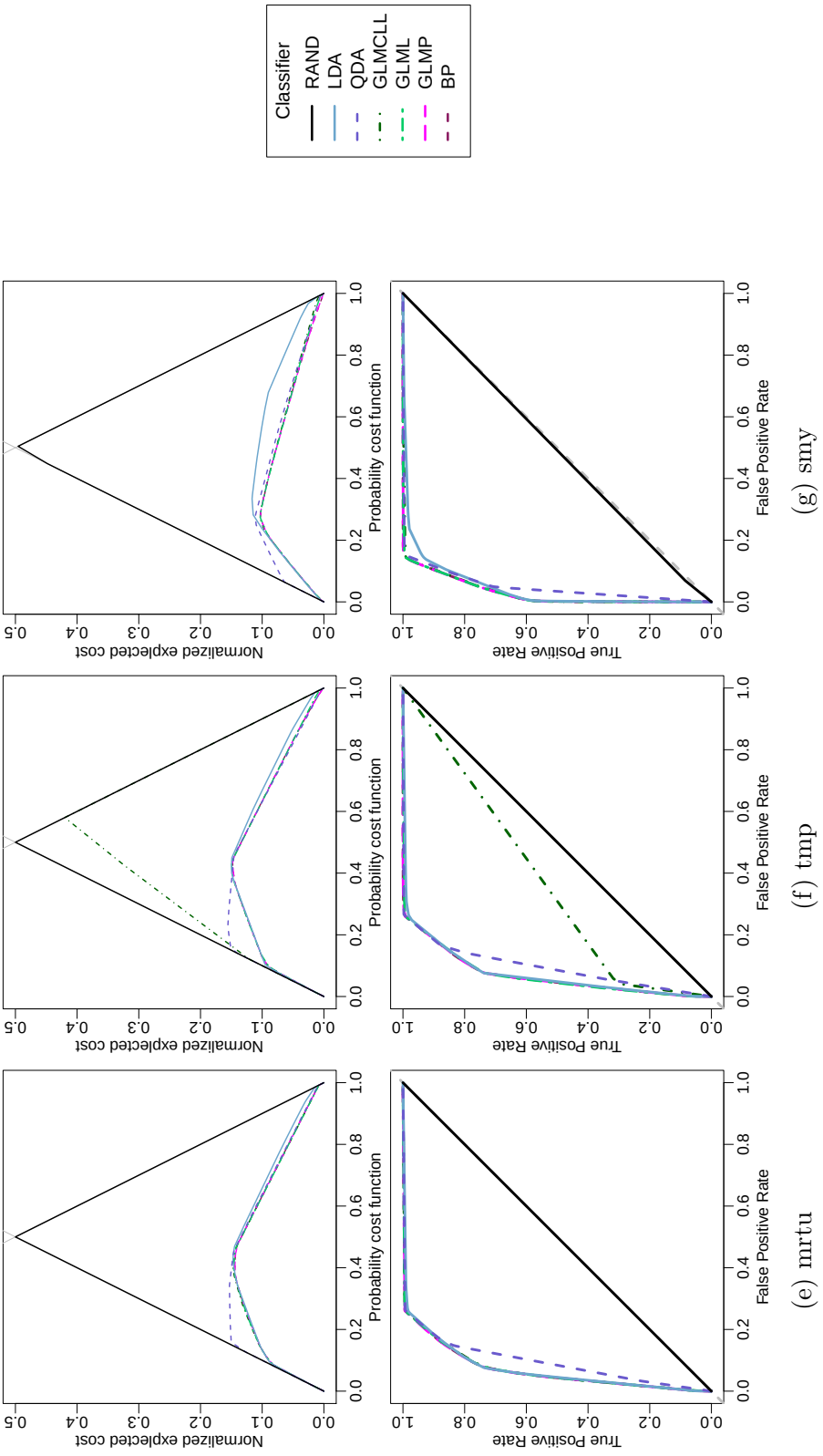


Figure C.2: Brier score by additive feature sets for forum with identifier 142. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) op (b) hps (c) mrr (d) mrtm



(e) mrtu (f) tmp (g) smy

Figure C.3: Lower envelope and ROC convex hulls for additive feature sets corresponding to forum with identifier 142.

C.2 Forum 141: high activity

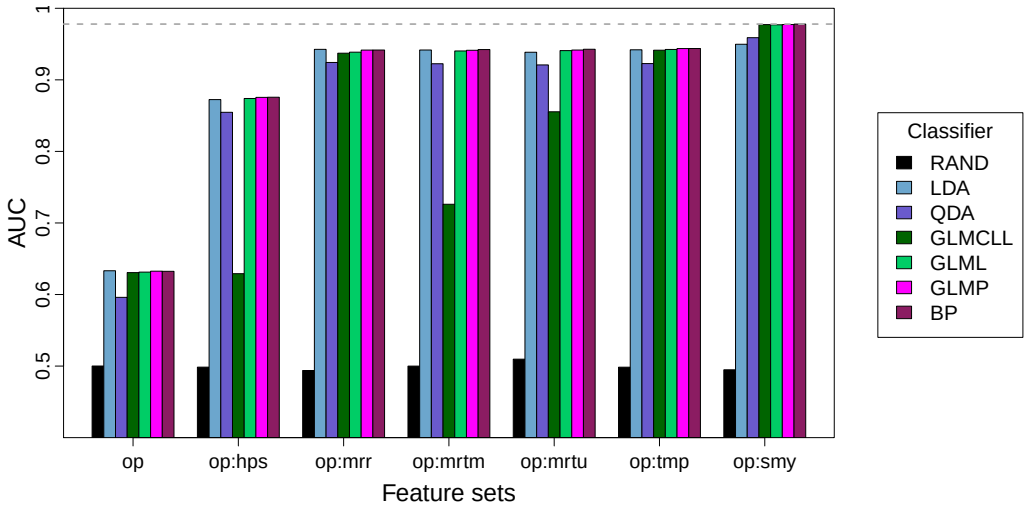


Figure C.4: Area Under (ROC) Curve (AUC) by additive feature sets for forum with identifier 141. The horizontal dashed grey line marks the highest (best) AUC measure observed.

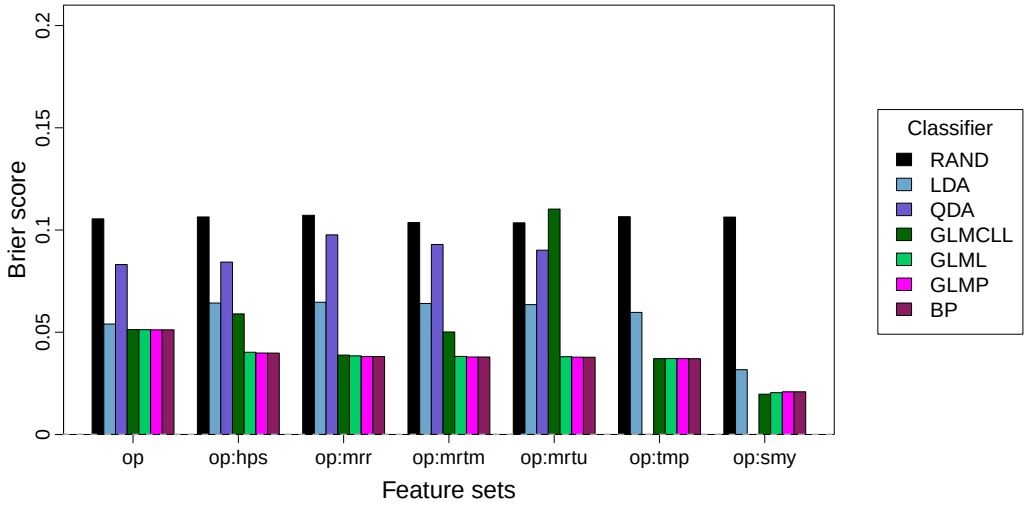
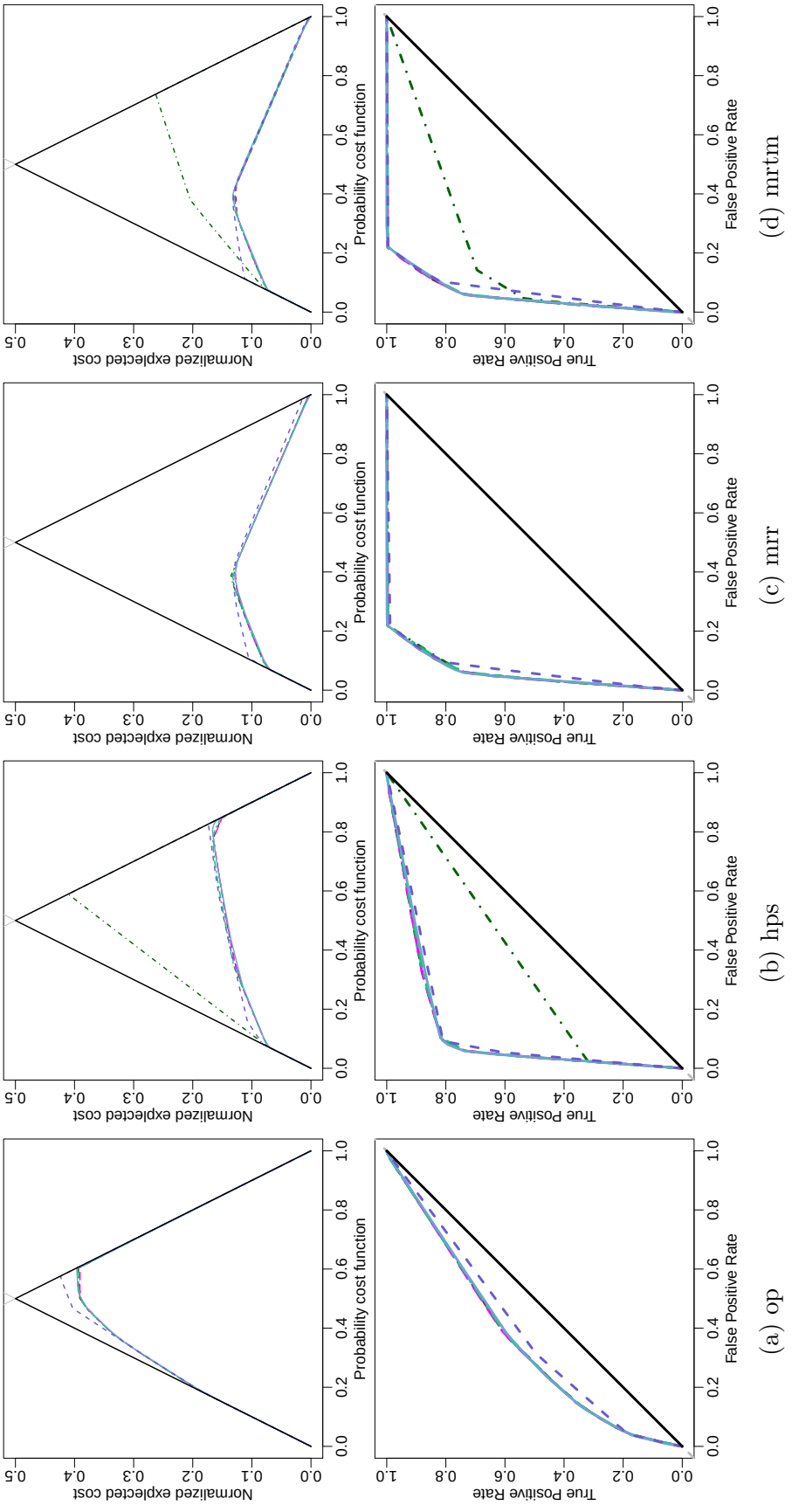
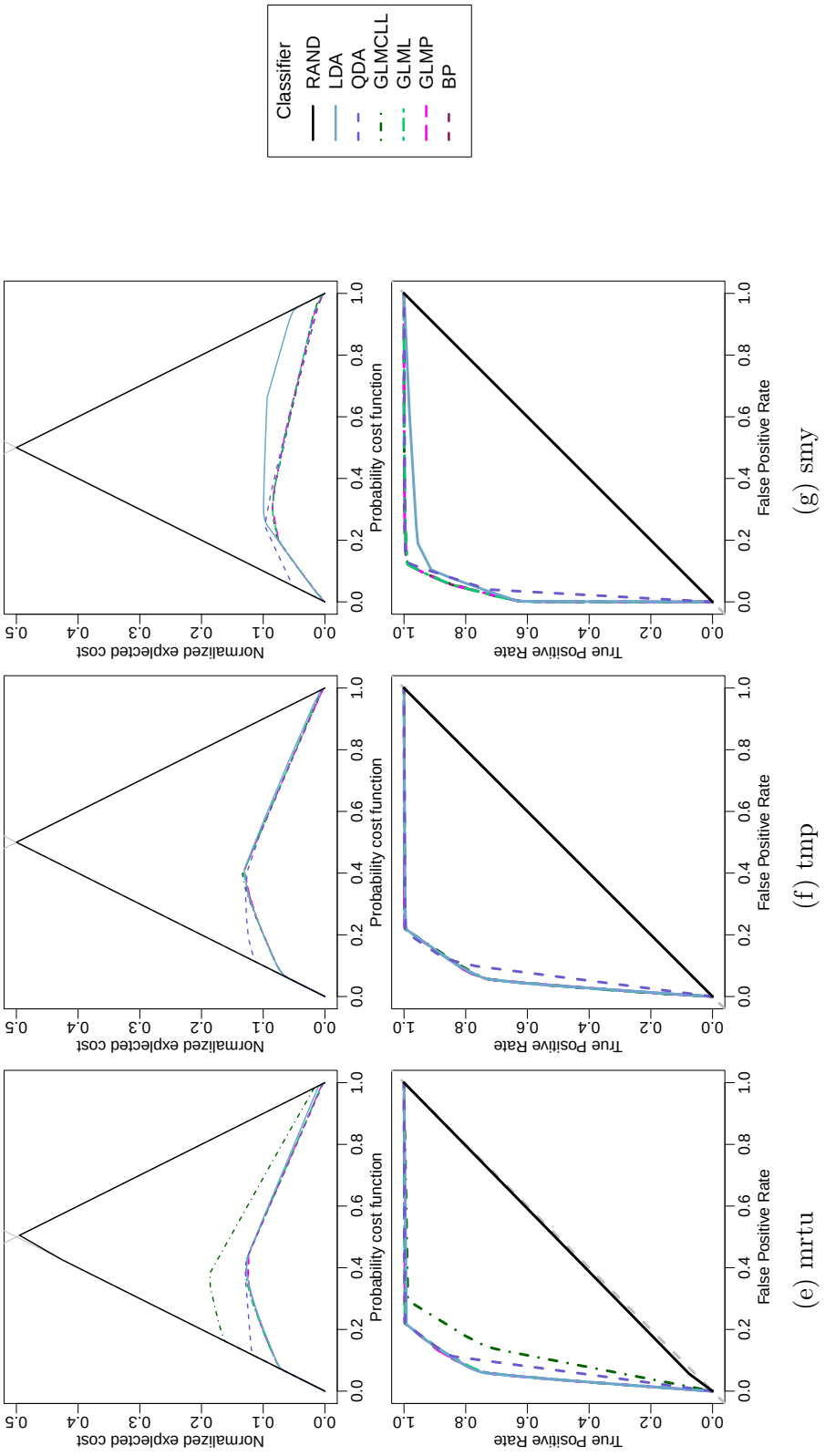


Figure C.5: Brier score by additive feature sets for forum with identifier 141. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) op (b) hps (c) mrr (d) mrtm



(e) mrtu (f) tmp (g) smy

Figure C.6: Lower envelope and ROC convex hulls for additive feature sets corresponding to forum with identifier 141.

C.3 Forum 156: low activity

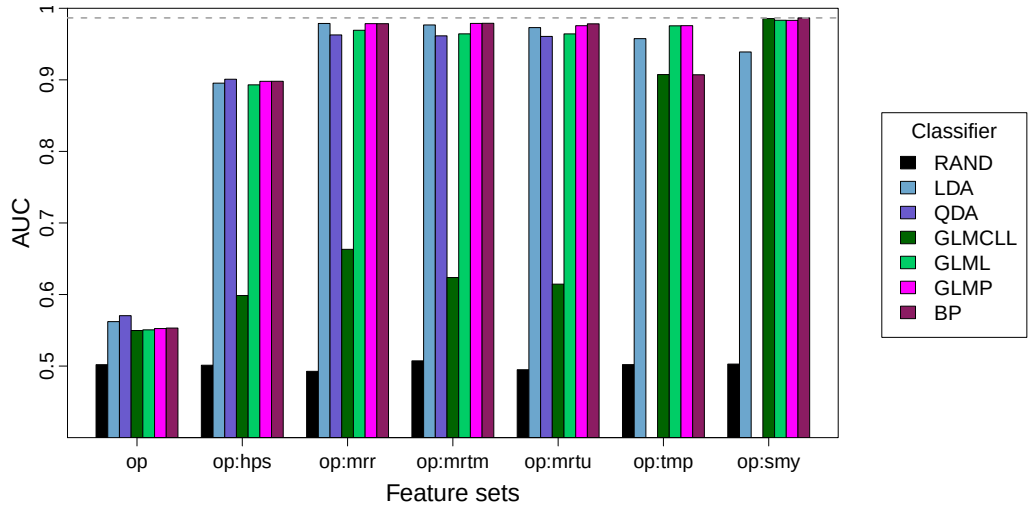


Figure C.7: Area Under (ROC) Curve (AUC) by additive feature sets for forum with identifier 156. The horizontal dashed grey line marks the highest (best) AUC measure observed.

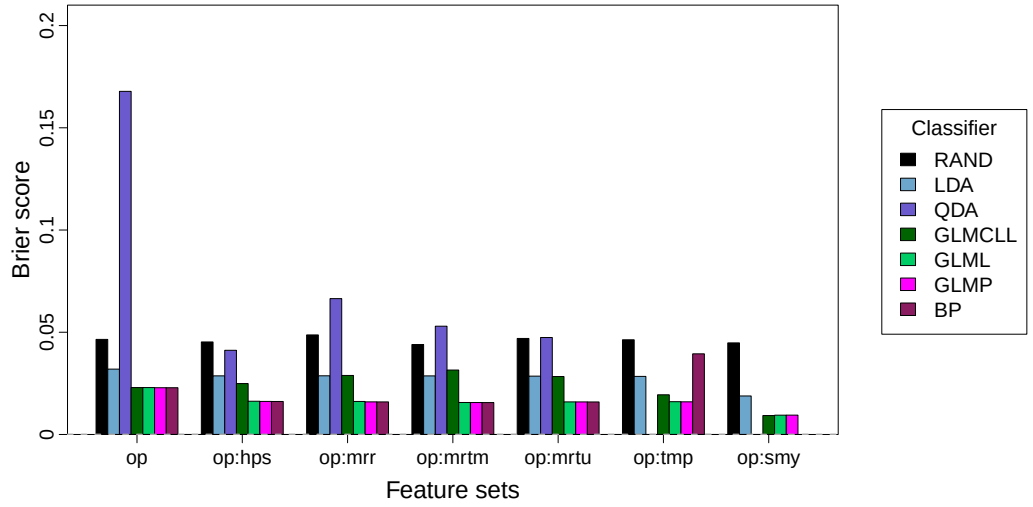
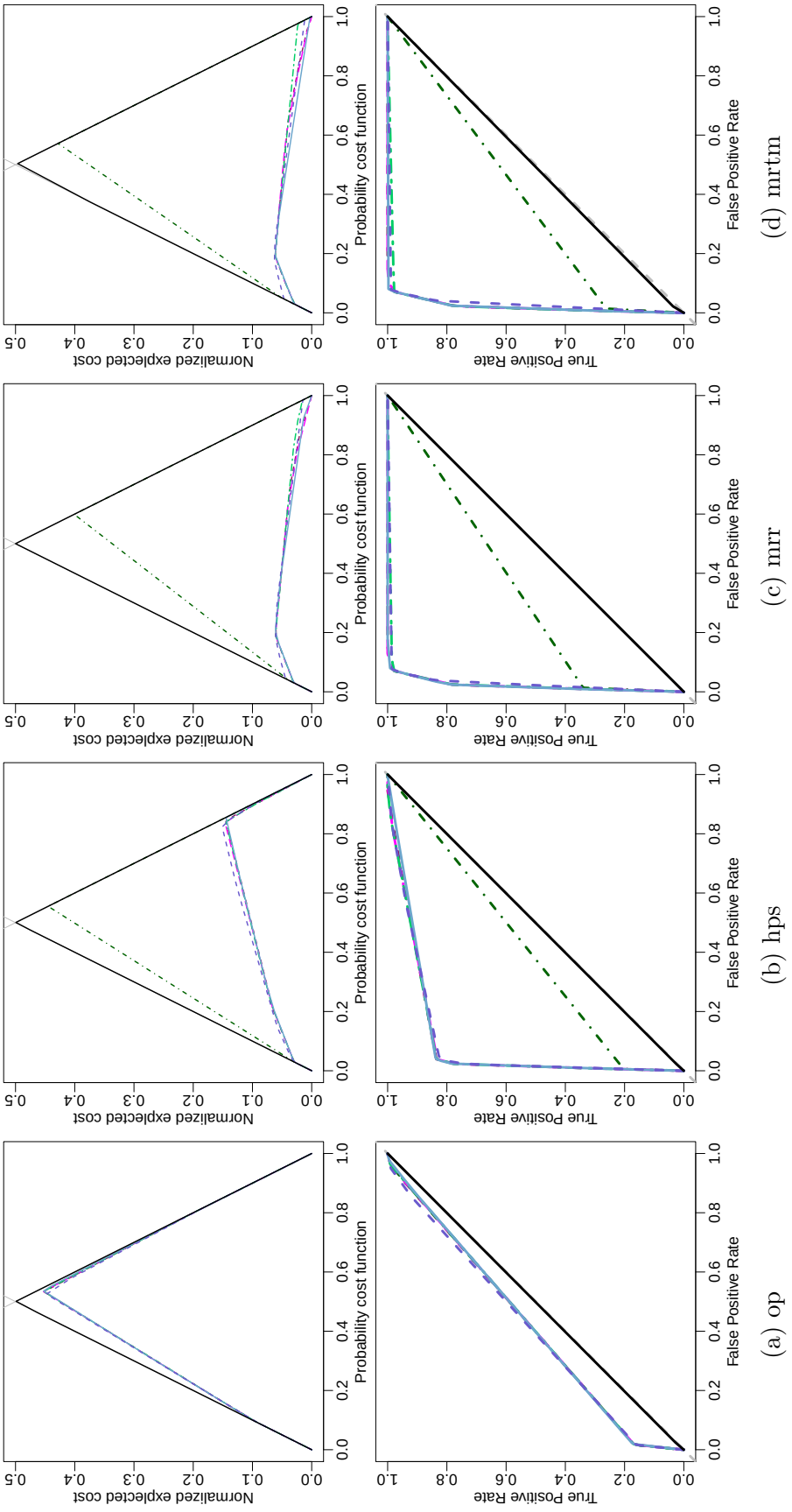
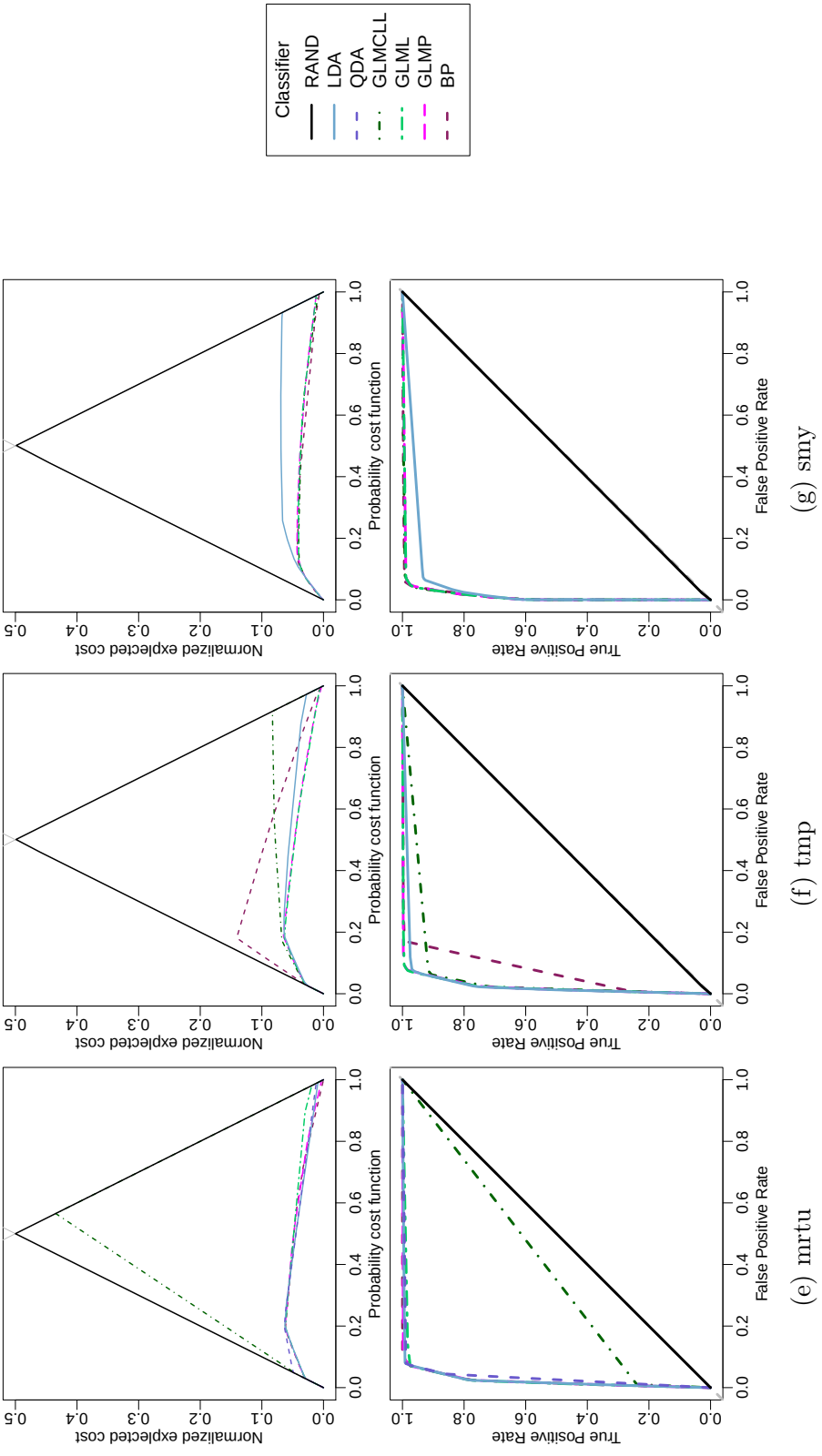


Figure C.8: Brier score by additive feature sets for forum with identifier 156. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) op (b) hps (c) mrr (d) mrtm



(e) mrtu (f) tmp (g) smy

Figure C.9: Lower envelope and ROC convex hulls for additive feature sets corresponding to forum with identifier 156.

C.4 Forum 418: very low activity

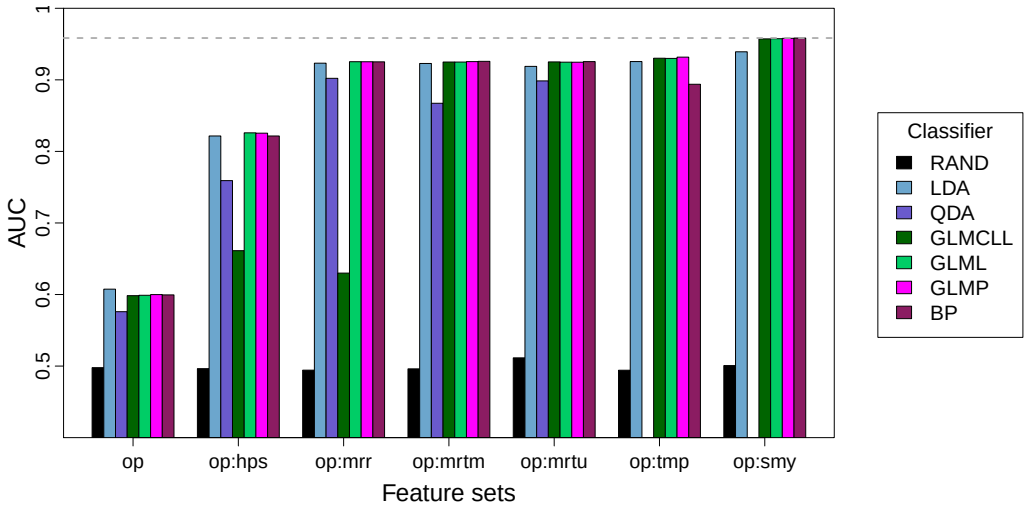


Figure C.10: Area Under (ROC) Curve (AUC) by additive feature sets for forum with identifier 418. The horizontal dashed grey line marks the highest (best) AUC measure observed.

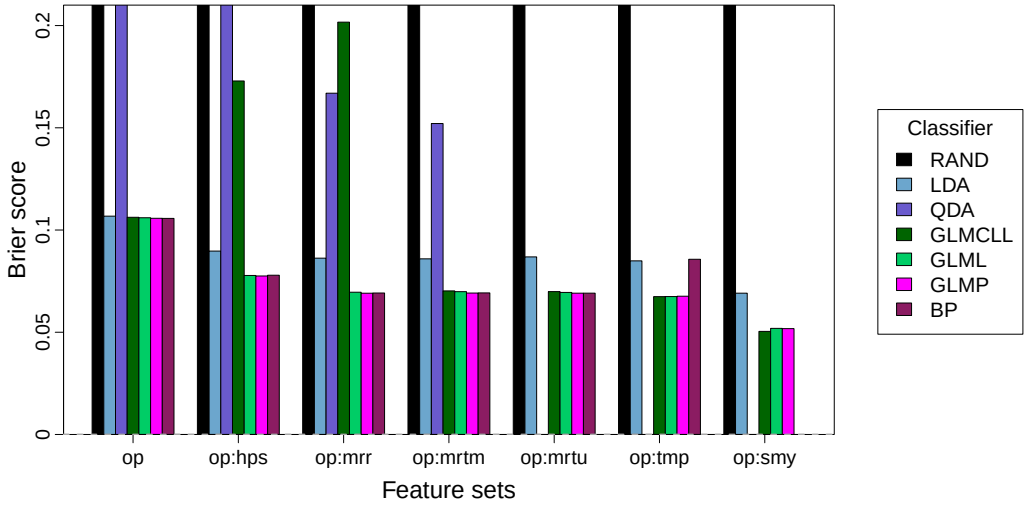
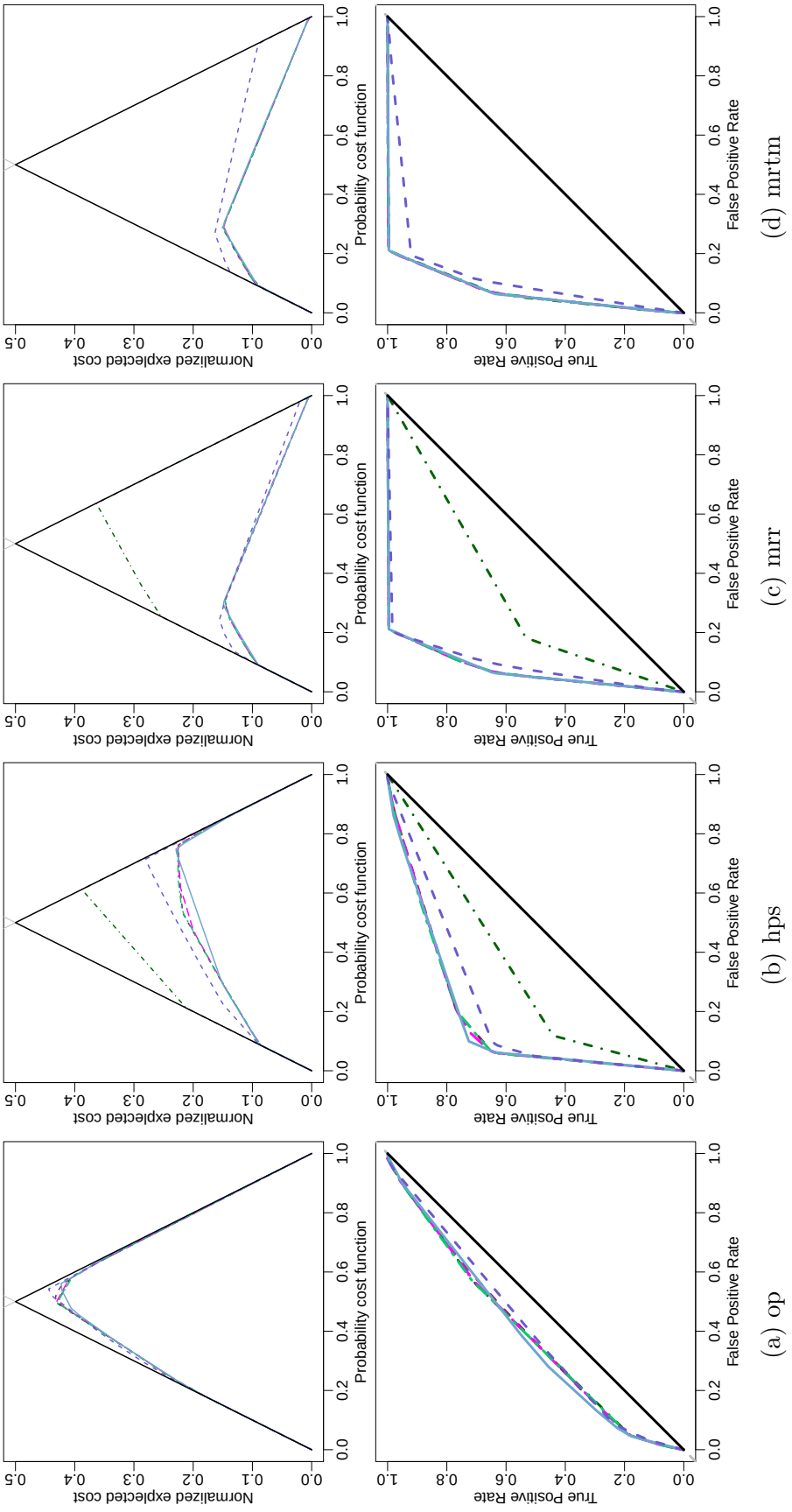
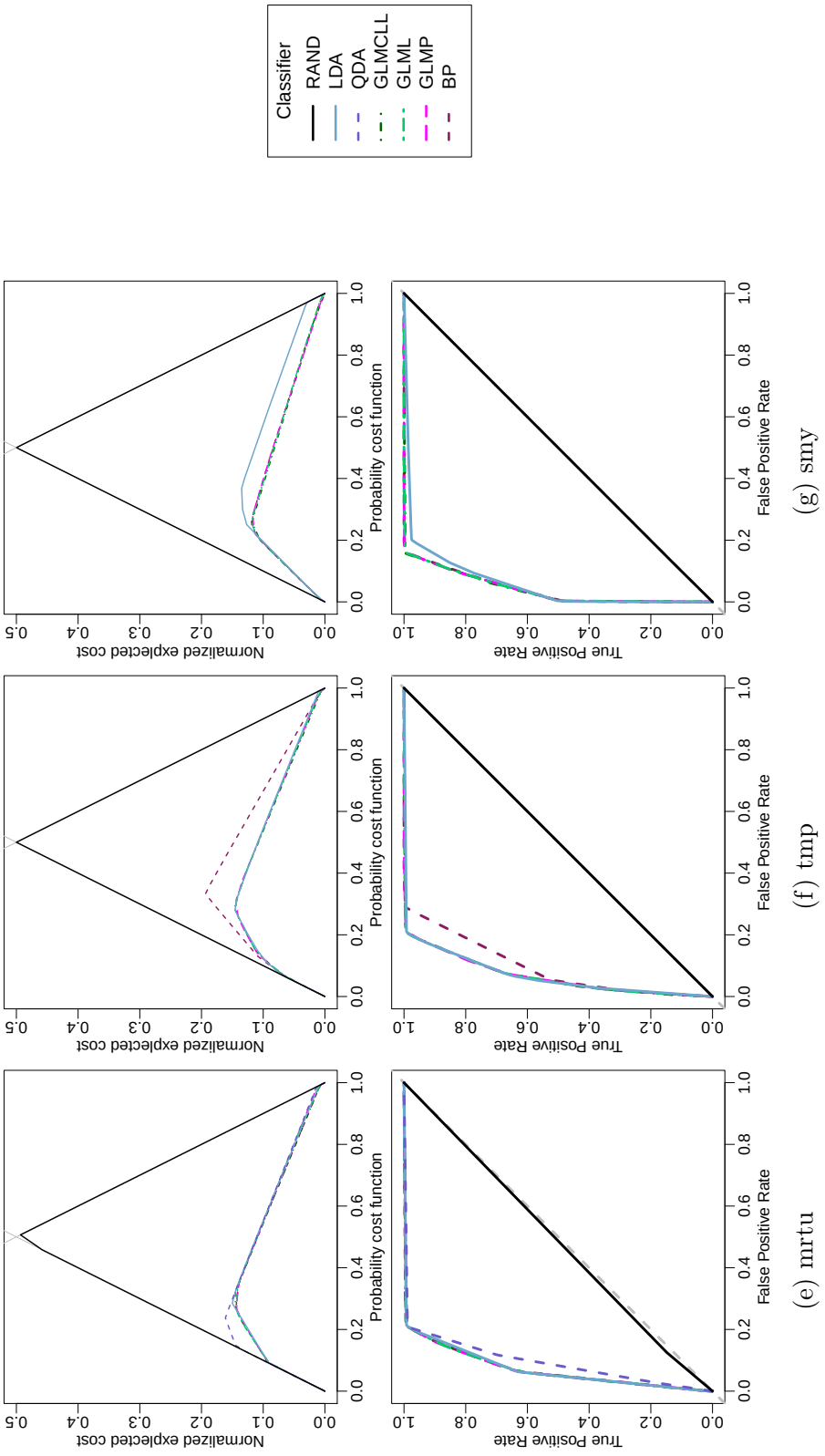


Figure C.11: Brier score by additive feature sets for forum with identifier 418. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) op (b) hps (c) mrr (d) mrtm



(e) mrtu (f) tmp (g) smy

Figure C.12: Lower envelope and ROC convex hulls for additive feature sets corresponding to forum with identifier 418.

Appendix D

Modelling Individual User Churn: churn window of 1 week

D.1 Forum 142: very high activity

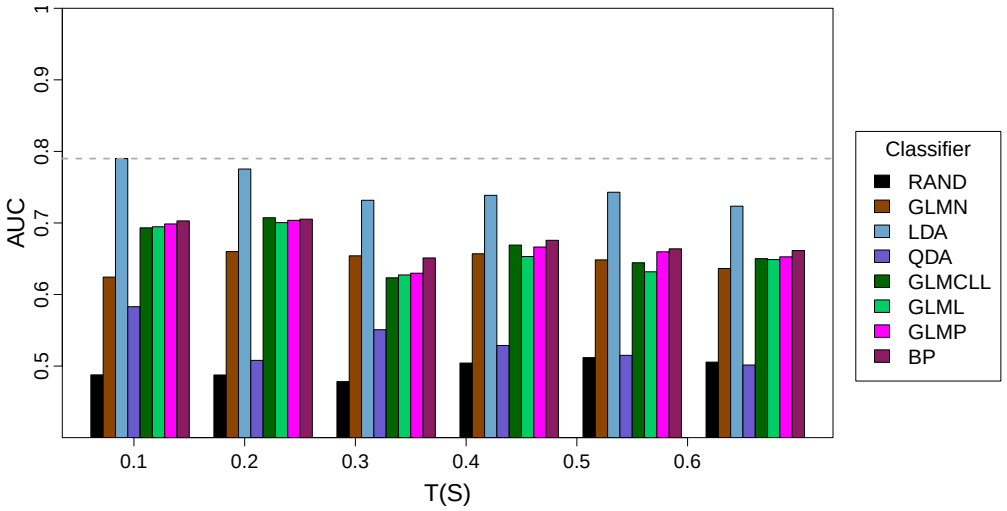


Figure D.1: Area Under (ROC) Curve (AUC) by churn threshold $T(S)$ for forum with identifier 142 when churn window is one week. The horizontal dashed grey line marks the highest (best) AUC measure observed.

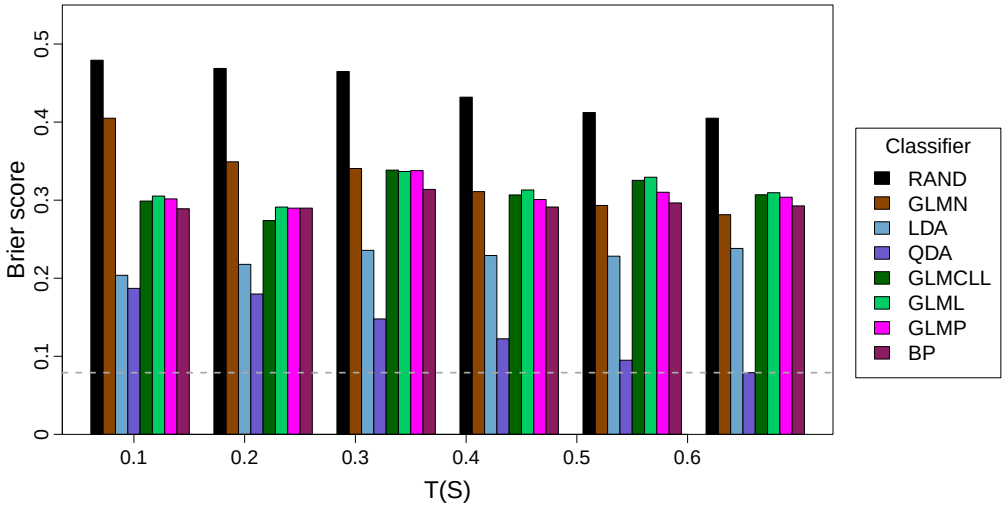
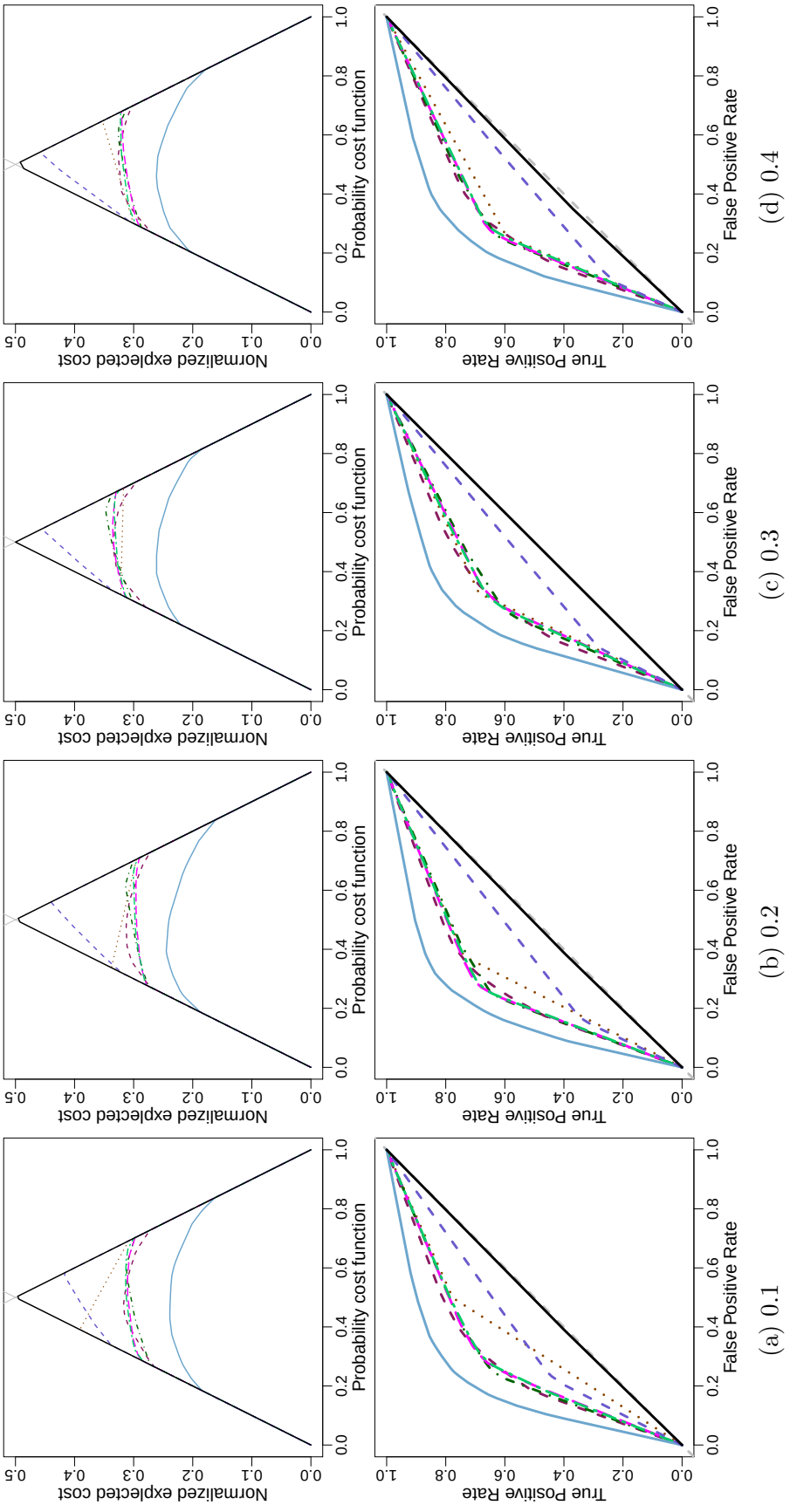
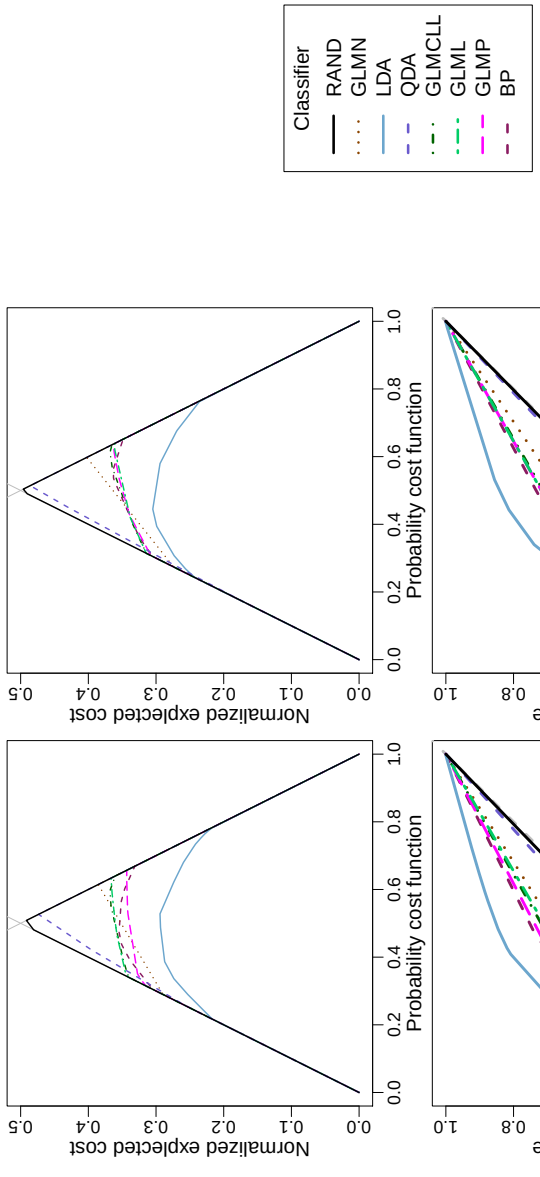


Figure D.2: Brier score by churn threshold $T(S)$ for forum with identifier 142 when churn window is one week. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure D.3: Lower envelope and ROC convex hulls for varying churn threshold $T(S)$ corresponding to forum with identifier 142 when churn window is one week.

D.2 Forum 141: high activity

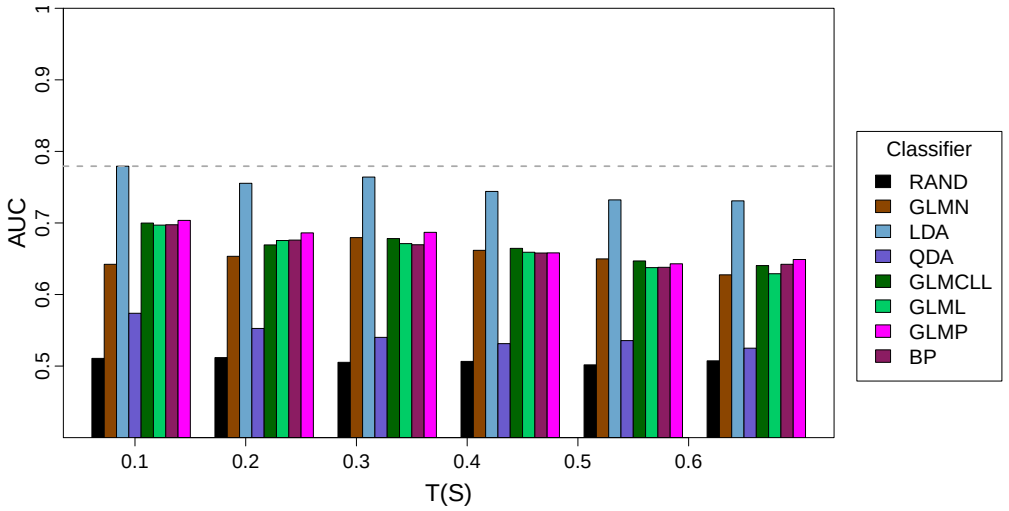


Figure D.4: Area Under (ROC) Curve (AUC) by churn threshold $T(S)$ for forum with identifier 141 when churn window is one week. The horizontal dashed grey line marks the highest (best) AUC measure observed.

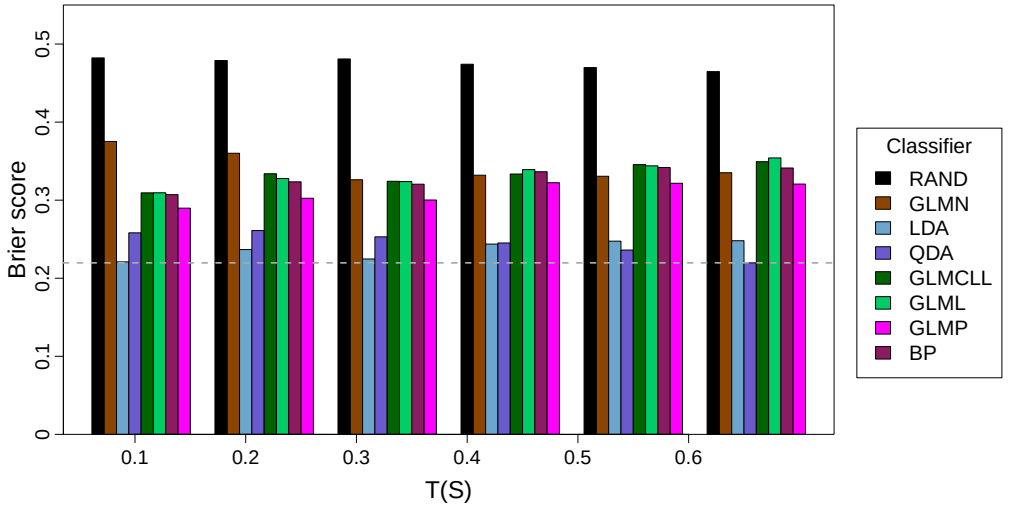
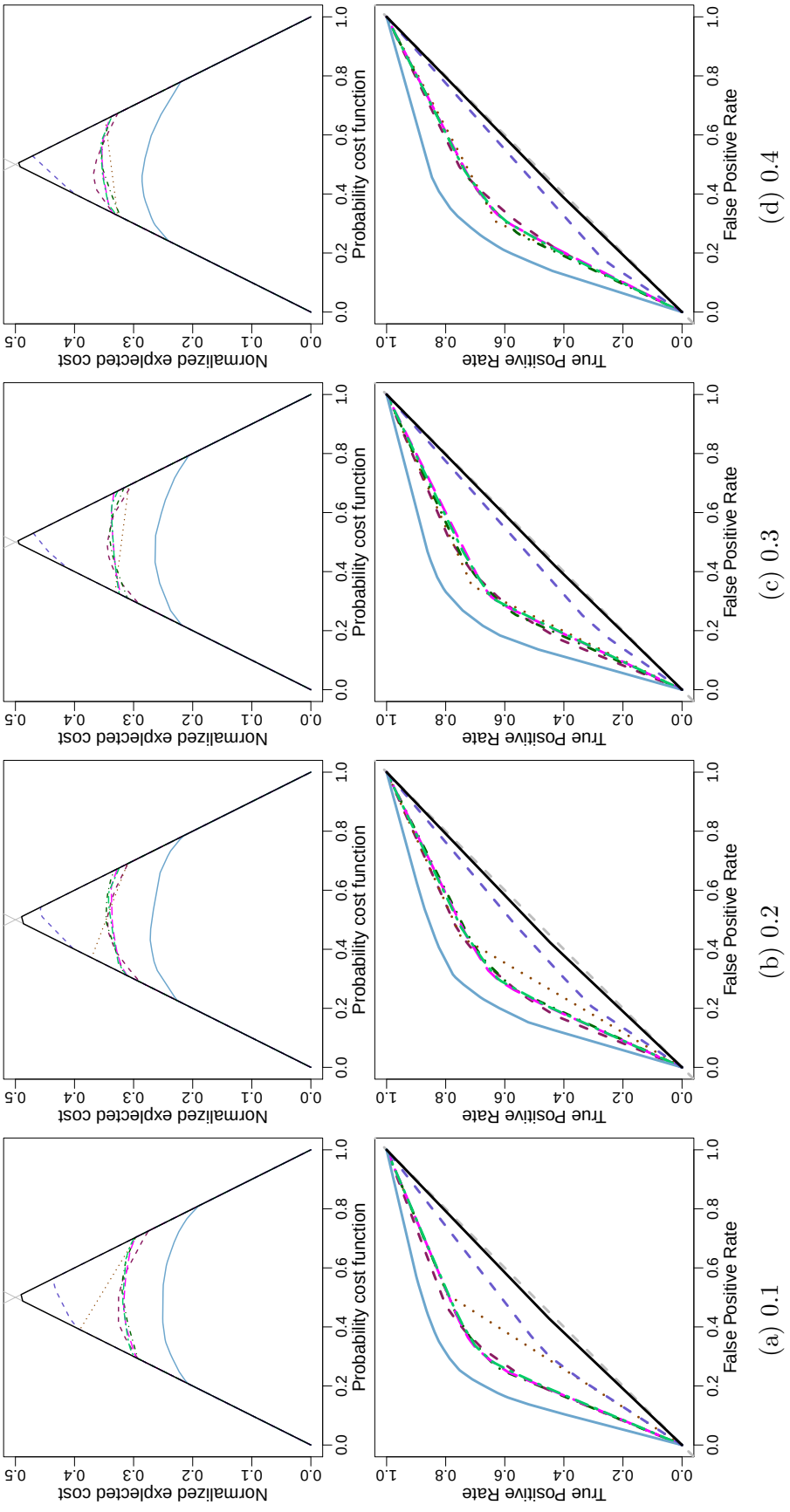
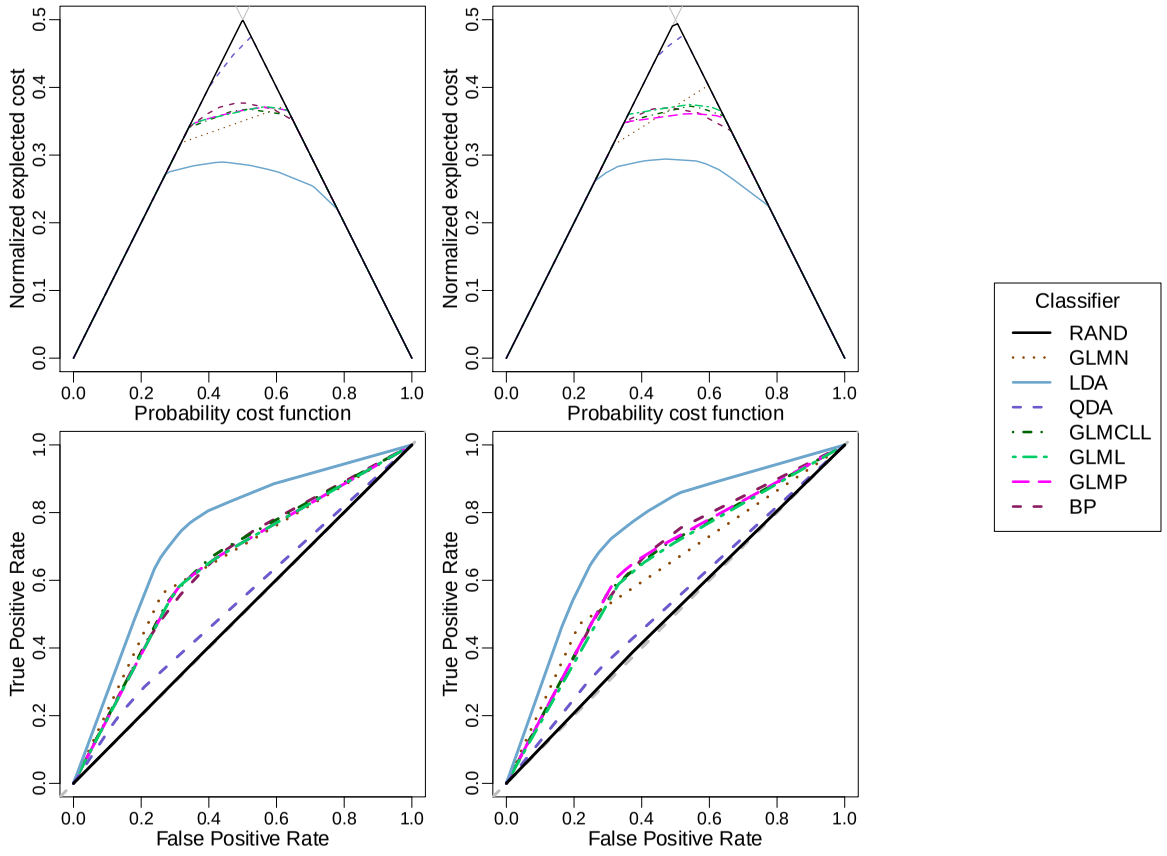


Figure D.5: Brier score by churn threshold $T(S)$ for forum with identifier 141 when churn window is one week. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure D.6: Lower envelope and ROC convex hulls for varying churn threshold $T(S)$ corresponding to forum with identifier 141 when churn window is one week.

D.3 Forum 156: low activity

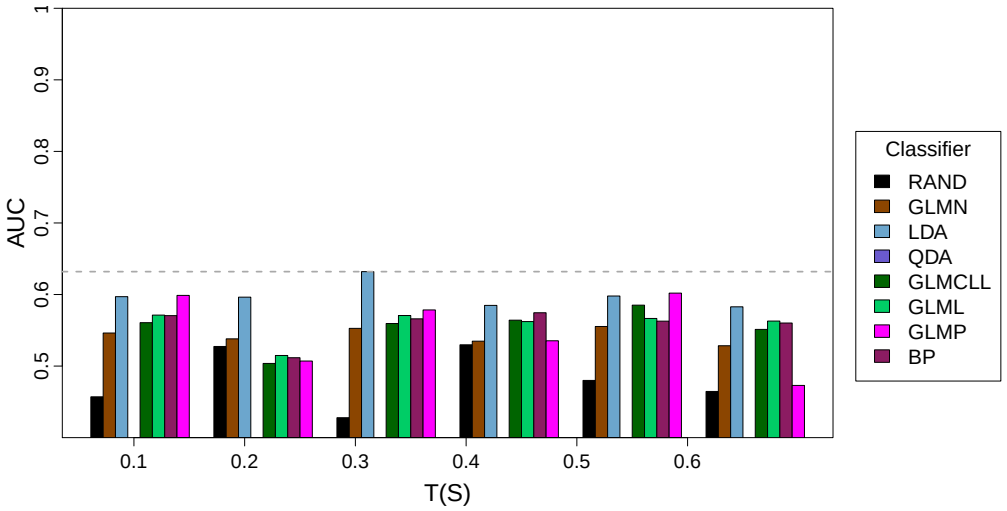


Figure D.7: Area Under (ROC) Curve (AUC) by churn threshold $T(S)$ for forum with identifier 156 when churn window is one week. The horizontal dashed grey line marks the highest (best) AUC measure observed.

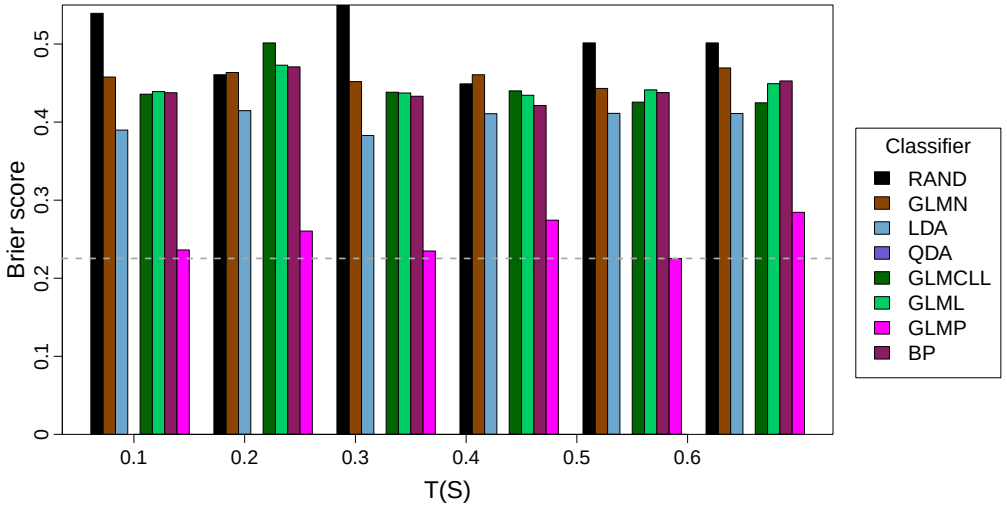
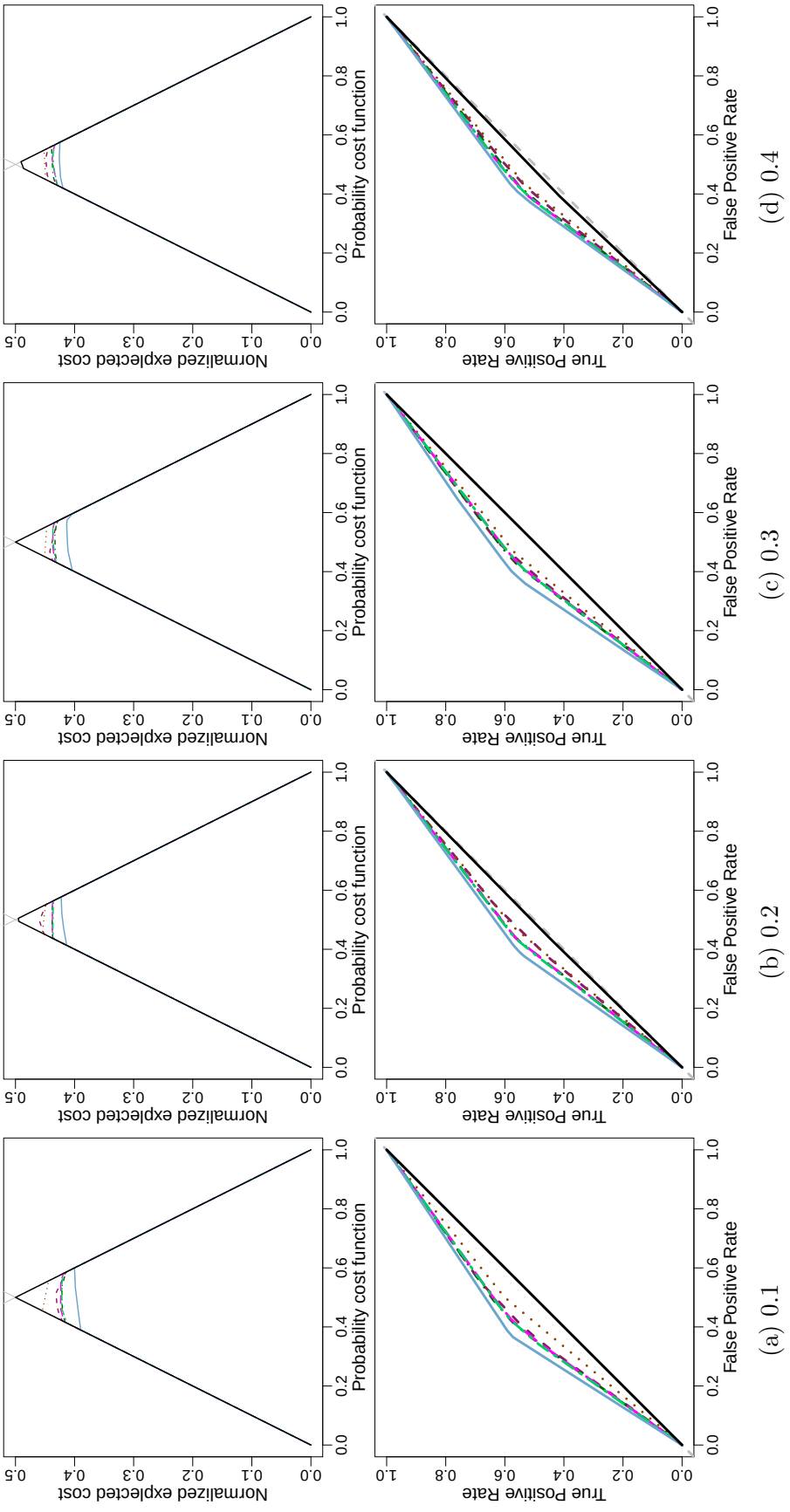
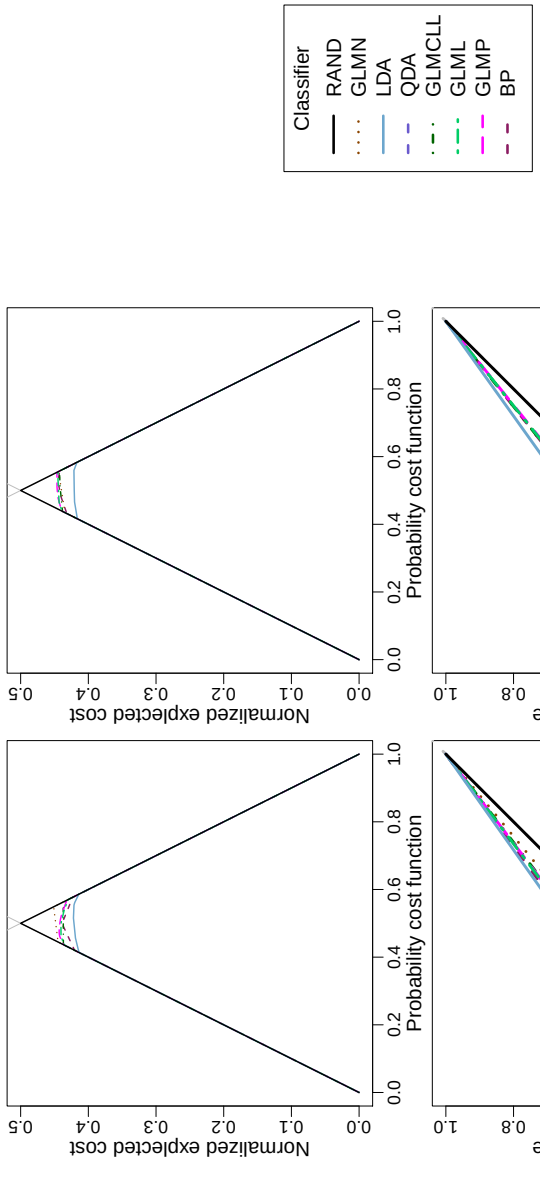


Figure D.8: Brier score by churn threshold $T(S)$ for forum with identifier 156 when churn window is one week. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure D.9: Lower envelope and ROC convex hulls for varying churn threshold $T(S)$ corresponding to forum with identifier 156 when churn window is one week.

Appendix E

Modelling Individual User Churn: churn window of 4 weeks

E.1 Forum 50: bursty activity

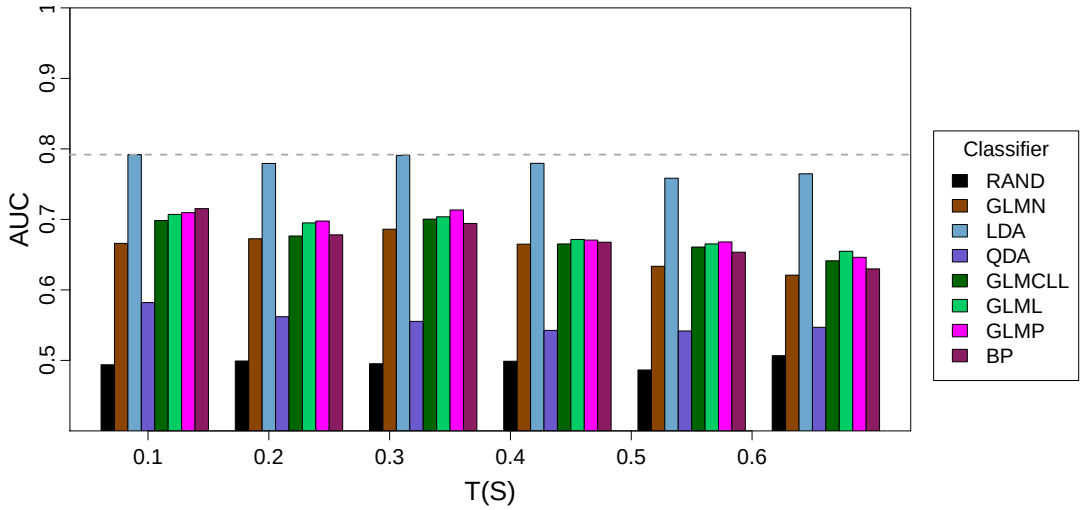


Figure E.1: Area Under (ROC) Curve (AUC) by churn threshold $T(S)$ for forum with identifier 50 when churn window is four weeks. The horizontal dashed grey line marks the highest (best) AUC measure observed.

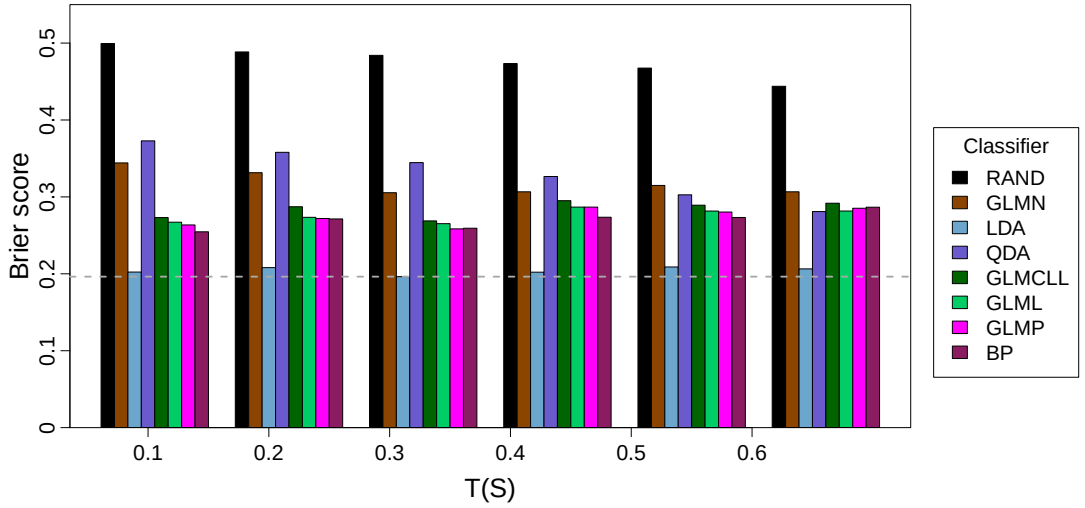
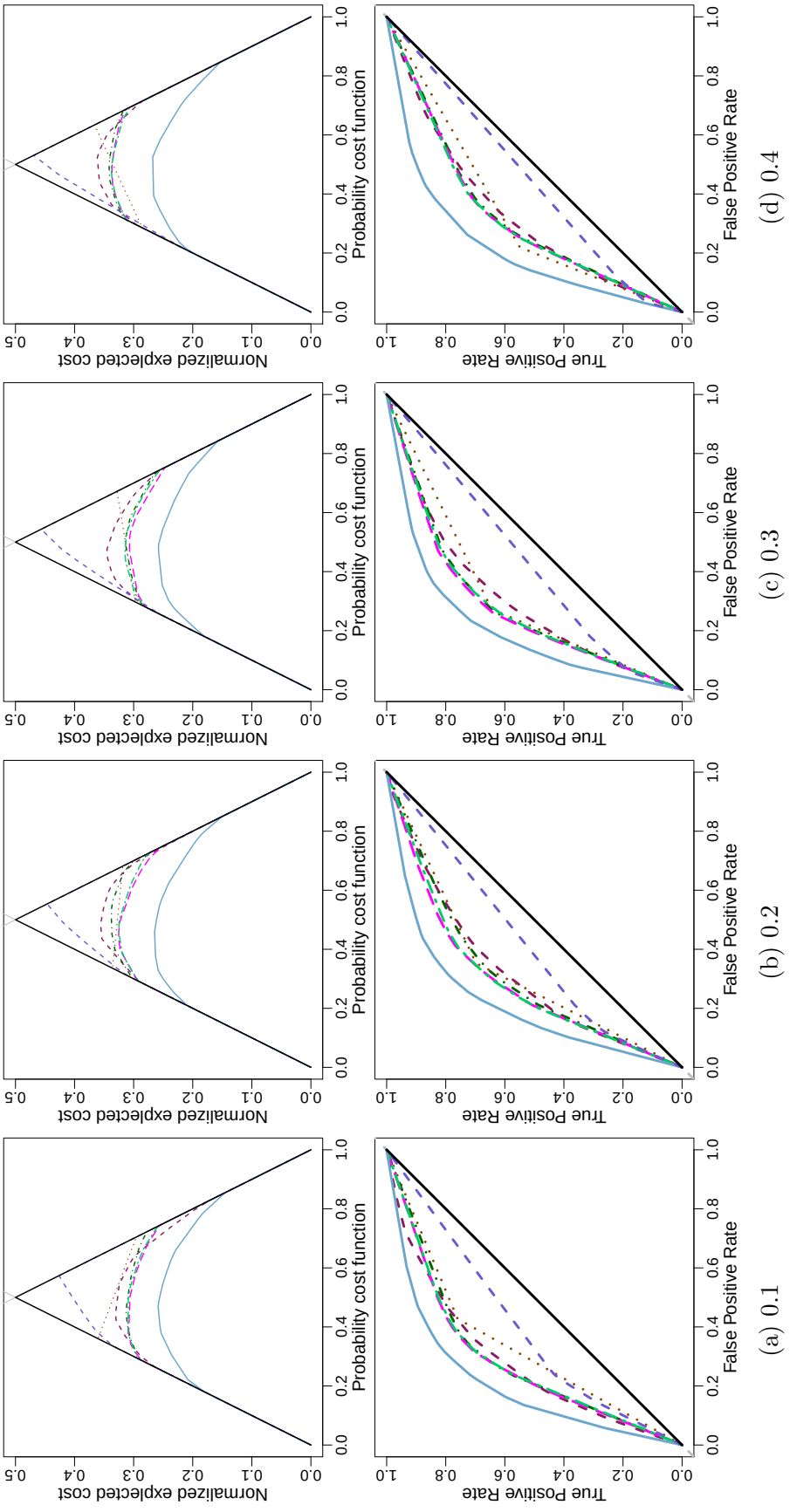
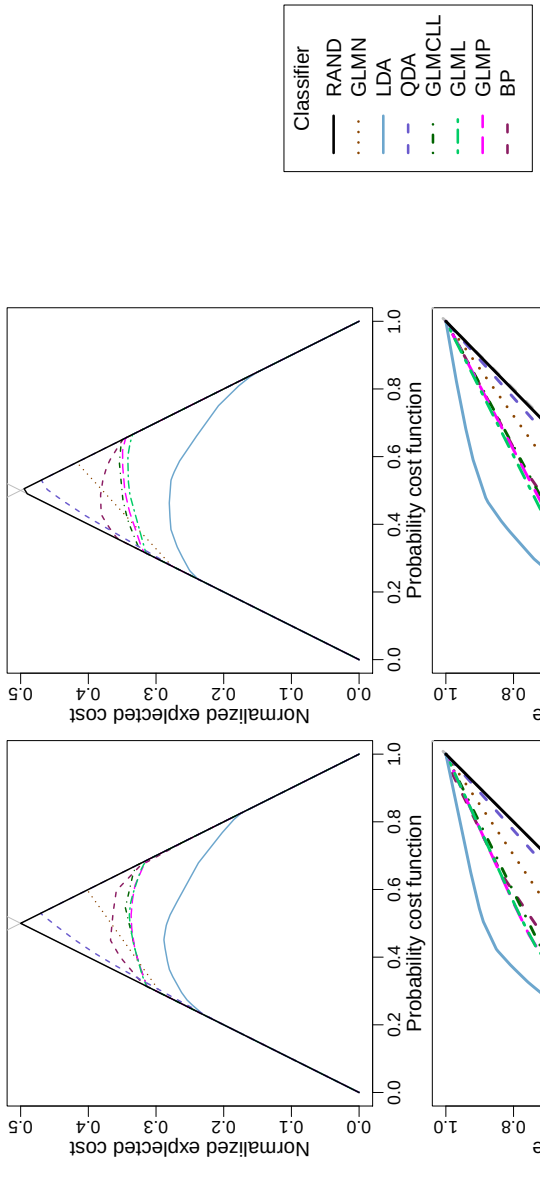


Figure E.2: Brier score by churn threshold $T(S)$ for forum with identifier 50 when churn window is four weeks. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure E.3: Lower envelope and ROC convex hulls for varying churn threshold $T(S)$ corresponding to forum with identifier 50 when churn window is four weeks.

E.2 Forum 142: very high activity

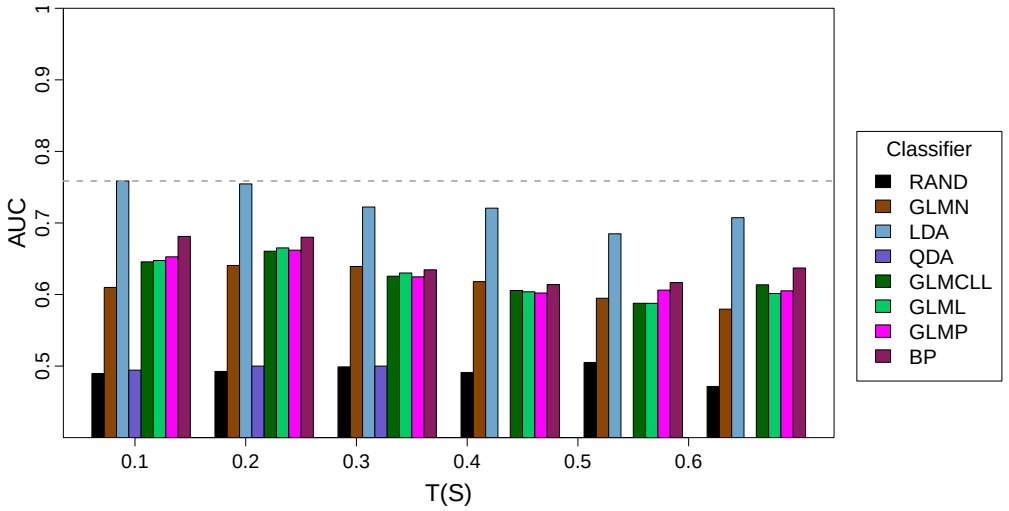


Figure E.4: Area Under (ROC) Curve (AUC) by churn threshold $T(S)$ for forum with identifier 142 when churn window is four weeks. The horizontal dashed grey line marks the highest (best) AUC measure observed.

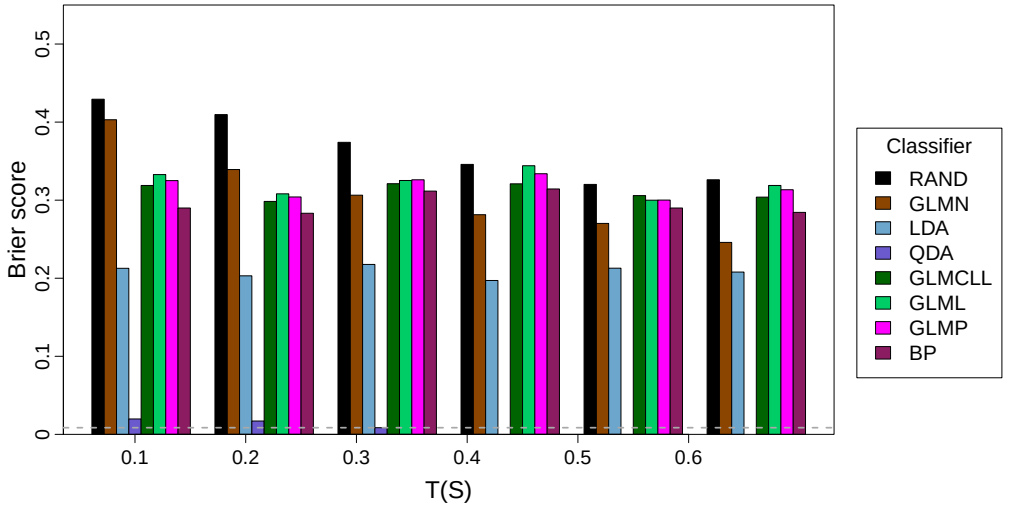
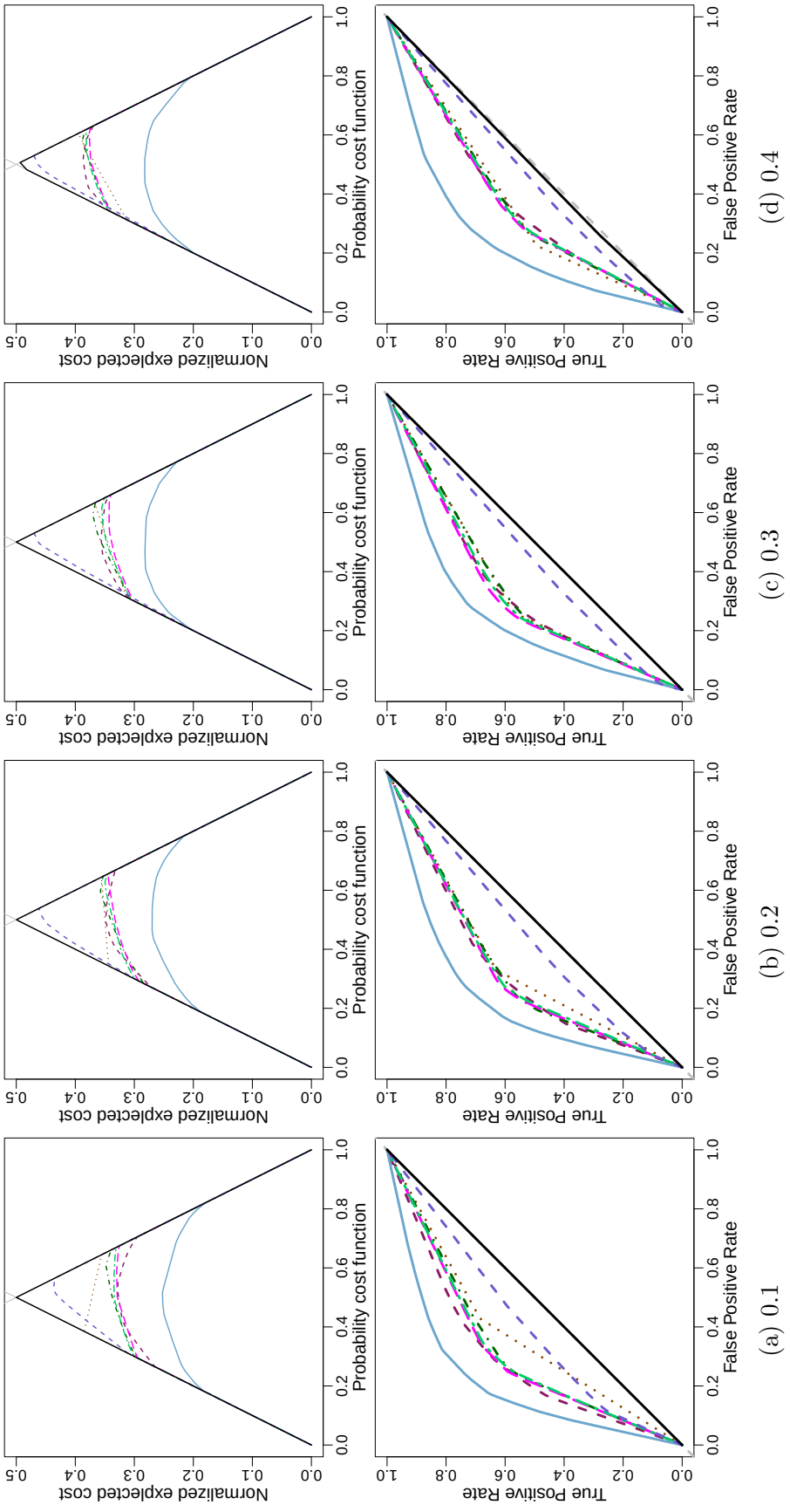
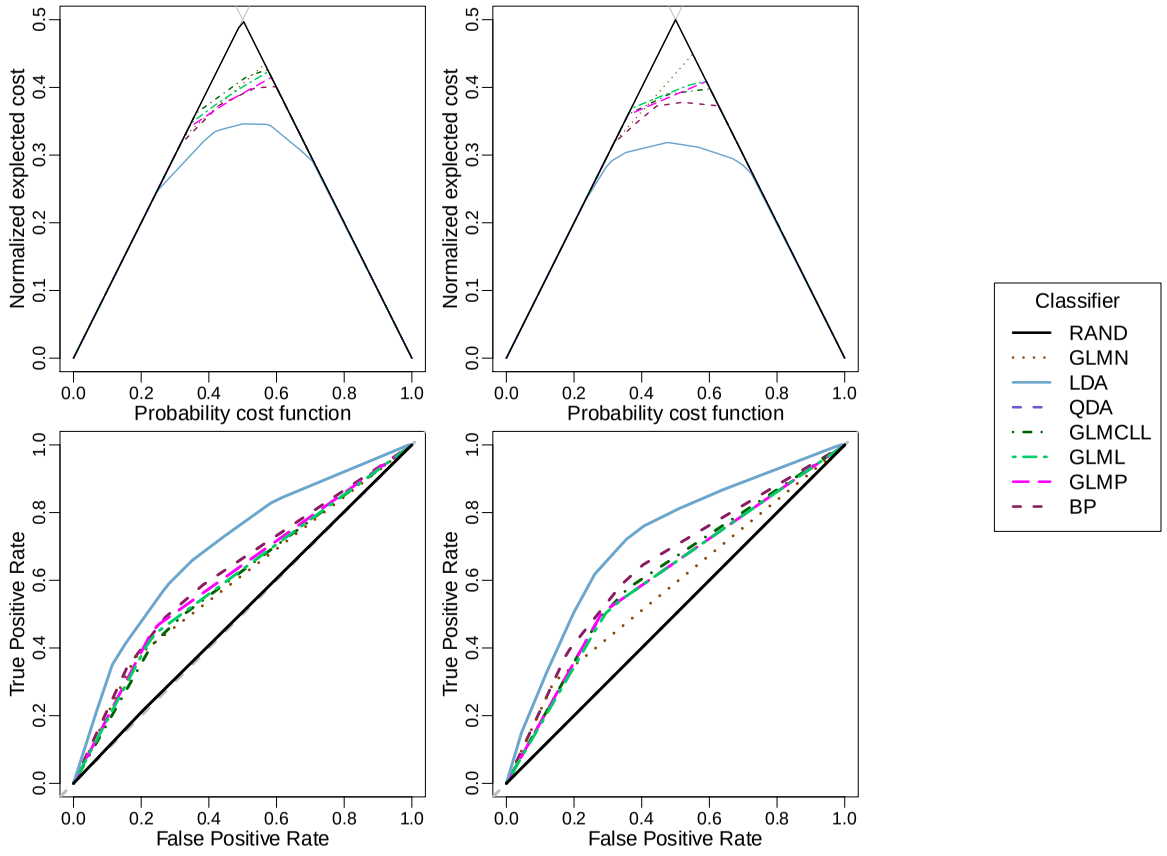


Figure E.5: Brier score by churn threshold $T(S)$ for forum with identifier 142 when churn window is four weeks. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure E.6: Lower envelope and ROC convex hulls for varying churn threshold $T(S)$ corresponding to forum with identifier 142 when churn window is four weeks.

E.3 Forum 141: high activity

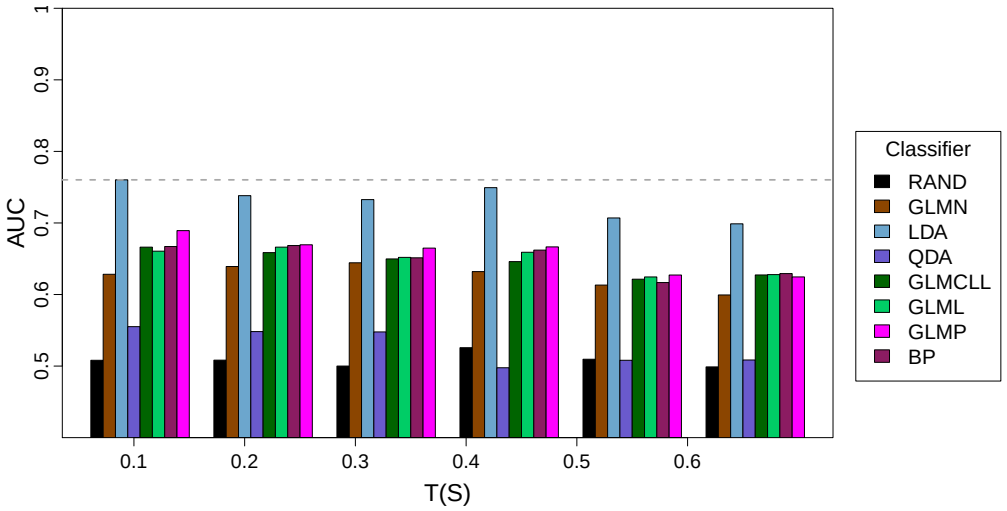


Figure E.7: Area Under (ROC) Curve (AUC) by churn threshold $T(S)$ for forum with identifier 141 when churn window is four weeks. The horizontal dashed grey line marks the highest (best) AUC measure observed.

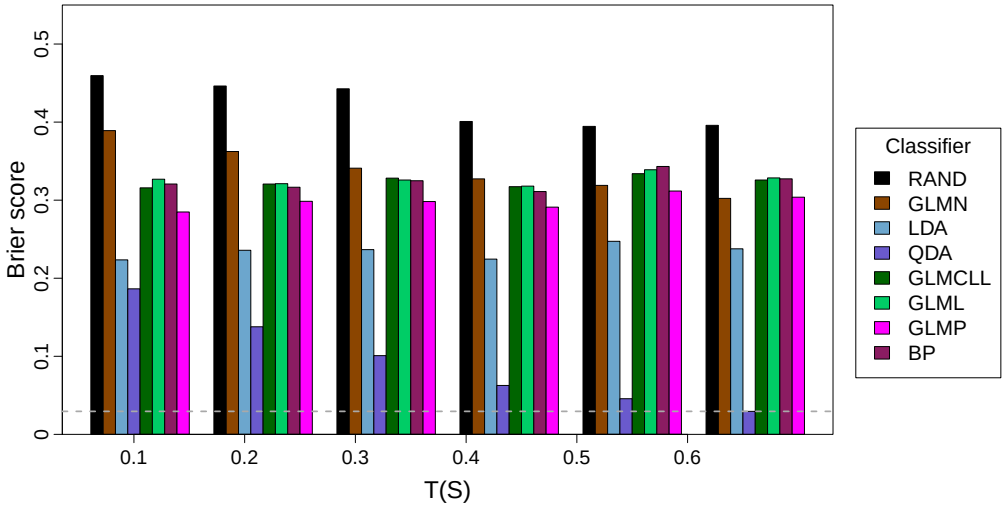
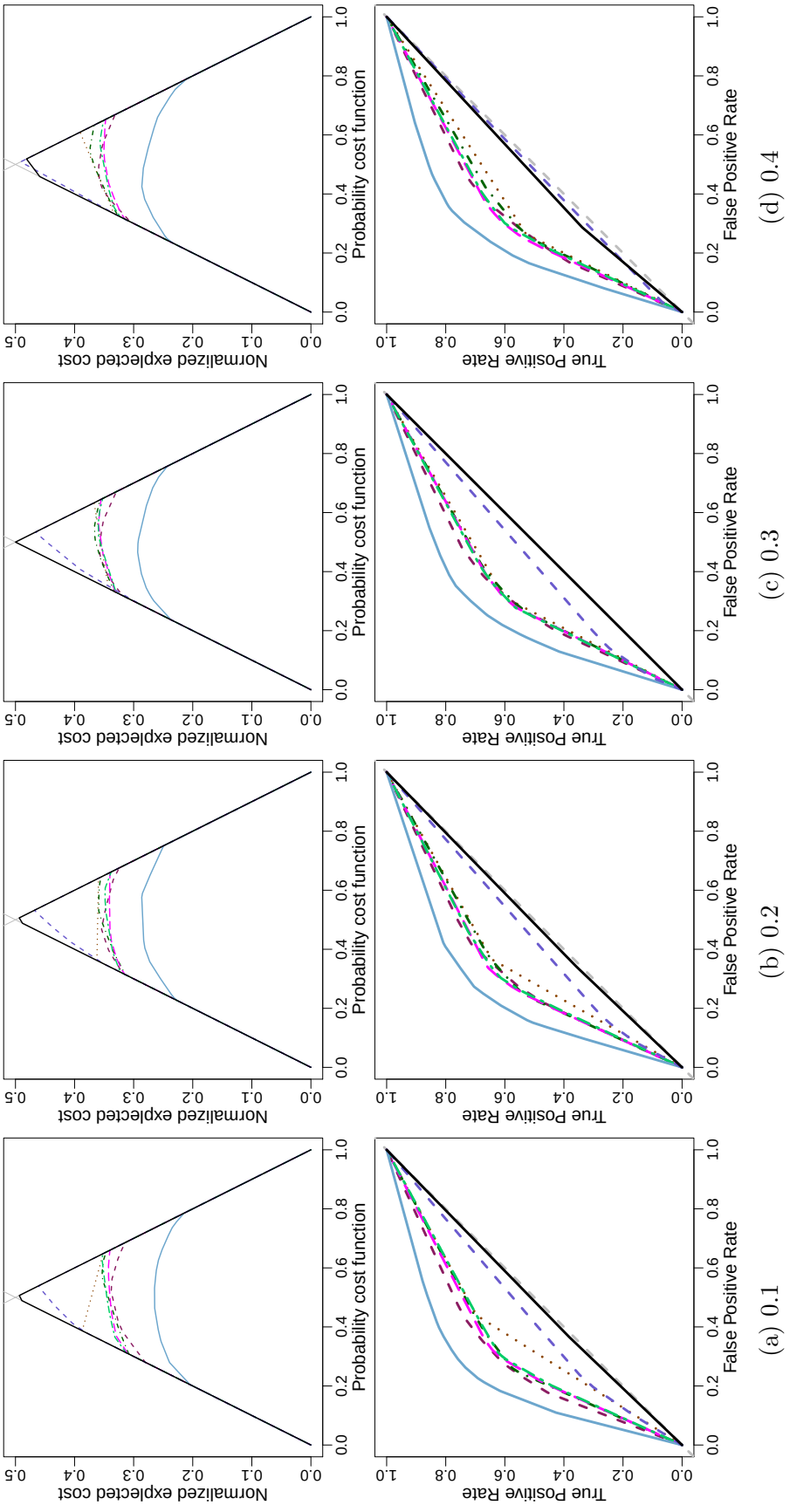
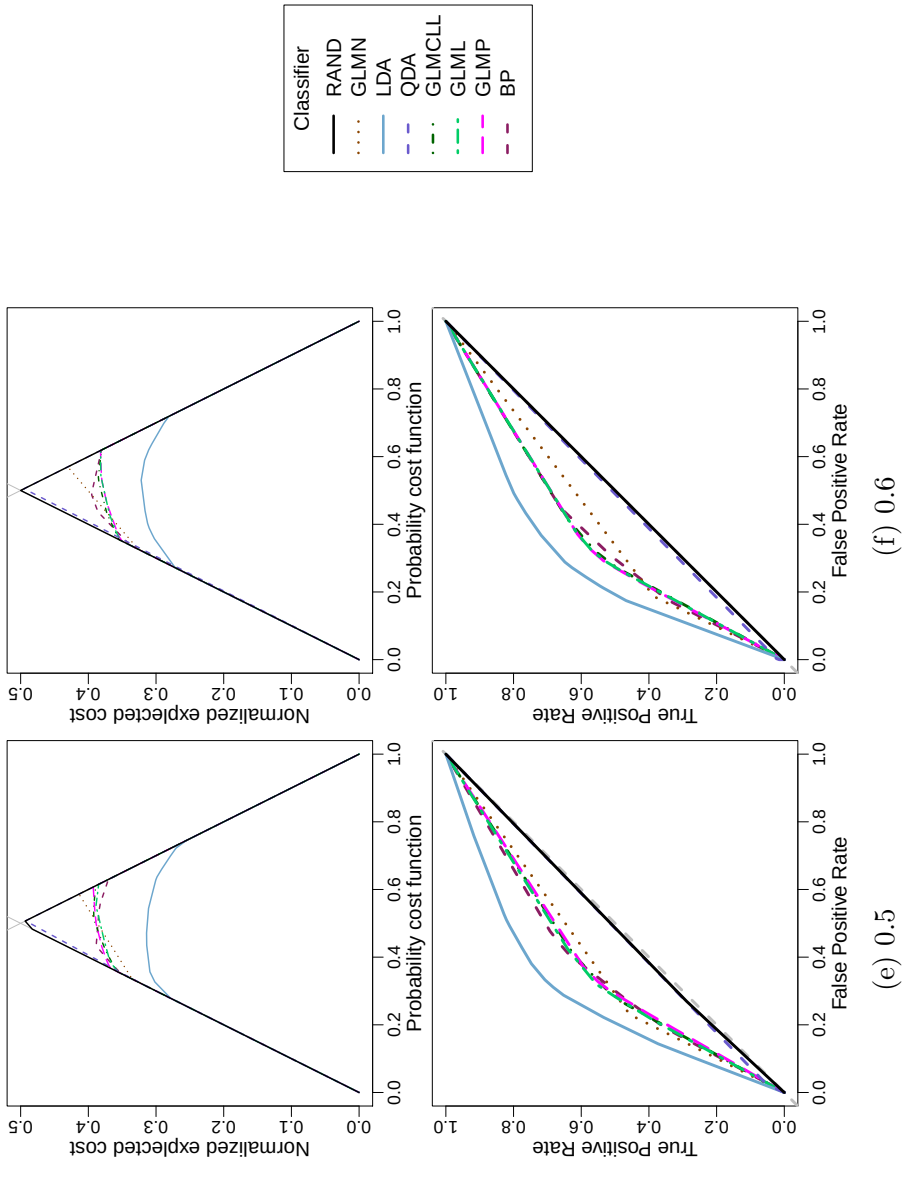


Figure E.8: Brier score by churn threshold $T(S)$ for forum with identifier 141 when churn window is four weeks. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure E.9: Lower envelope and ROC convex hulls for varying churn threshold $T(S)$ corresponding to forum with identifier 141 when churn window is four weeks.

E.4 Forum 156: low activity

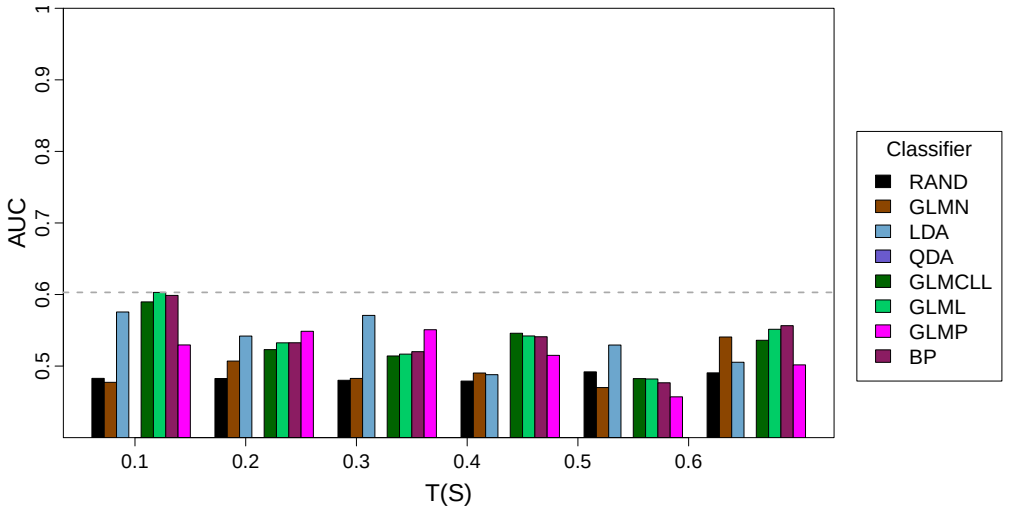


Figure E.10: Area Under (ROC) Curve (AUC) by churn threshold $T(S)$ for forum with identifier 156 when churn window is four weeks. The horizontal dashed grey line marks the highest (best) AUC measure observed.

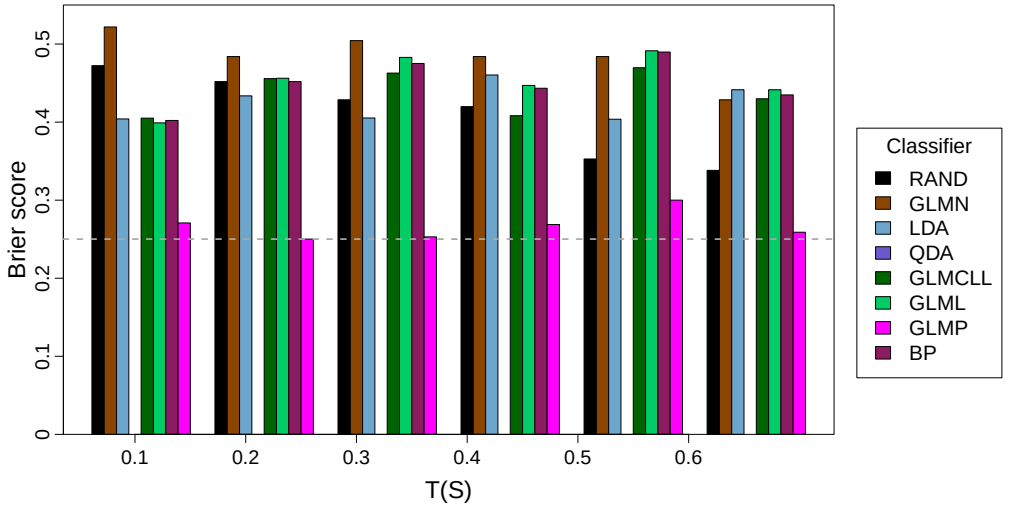
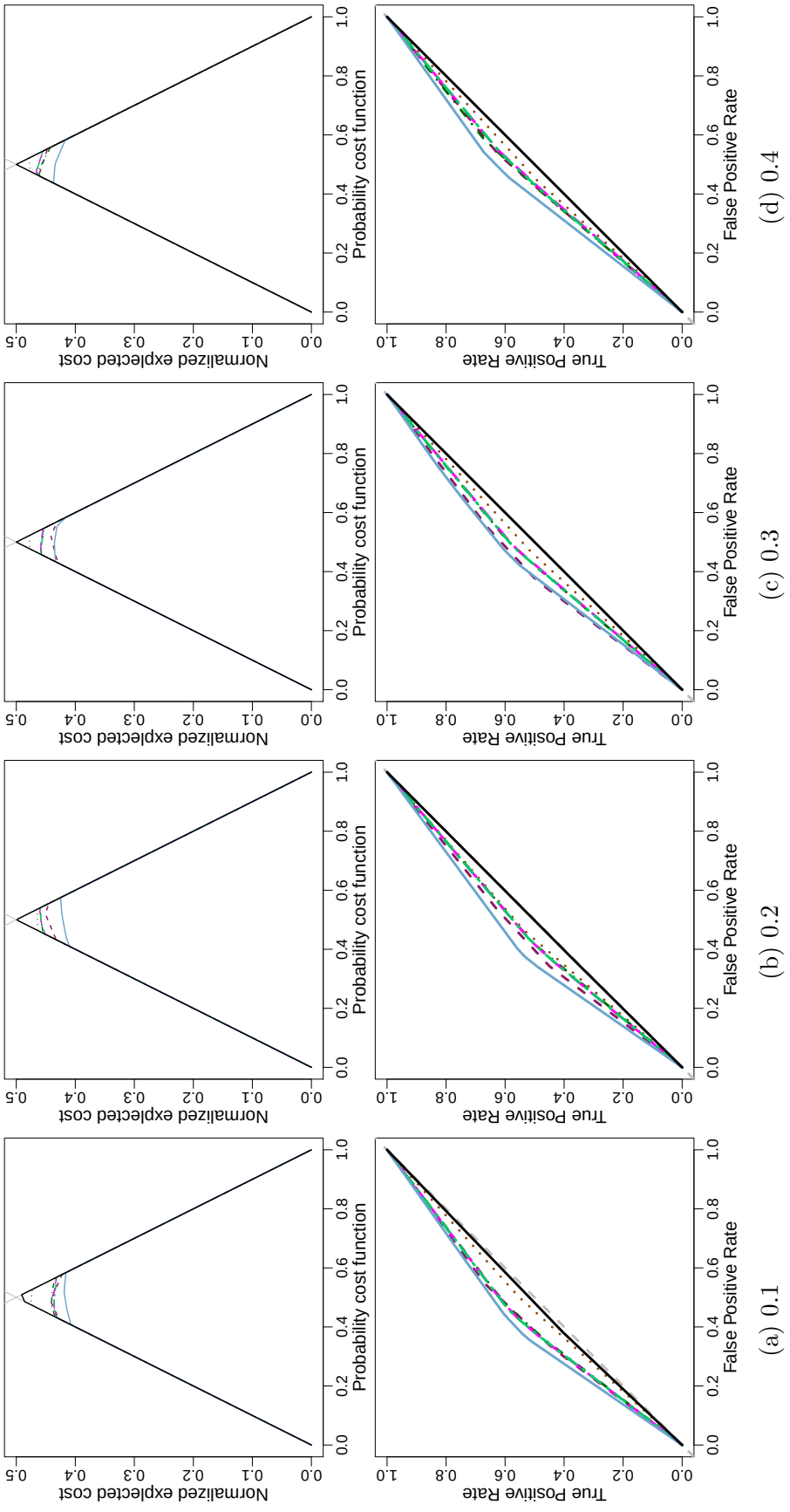
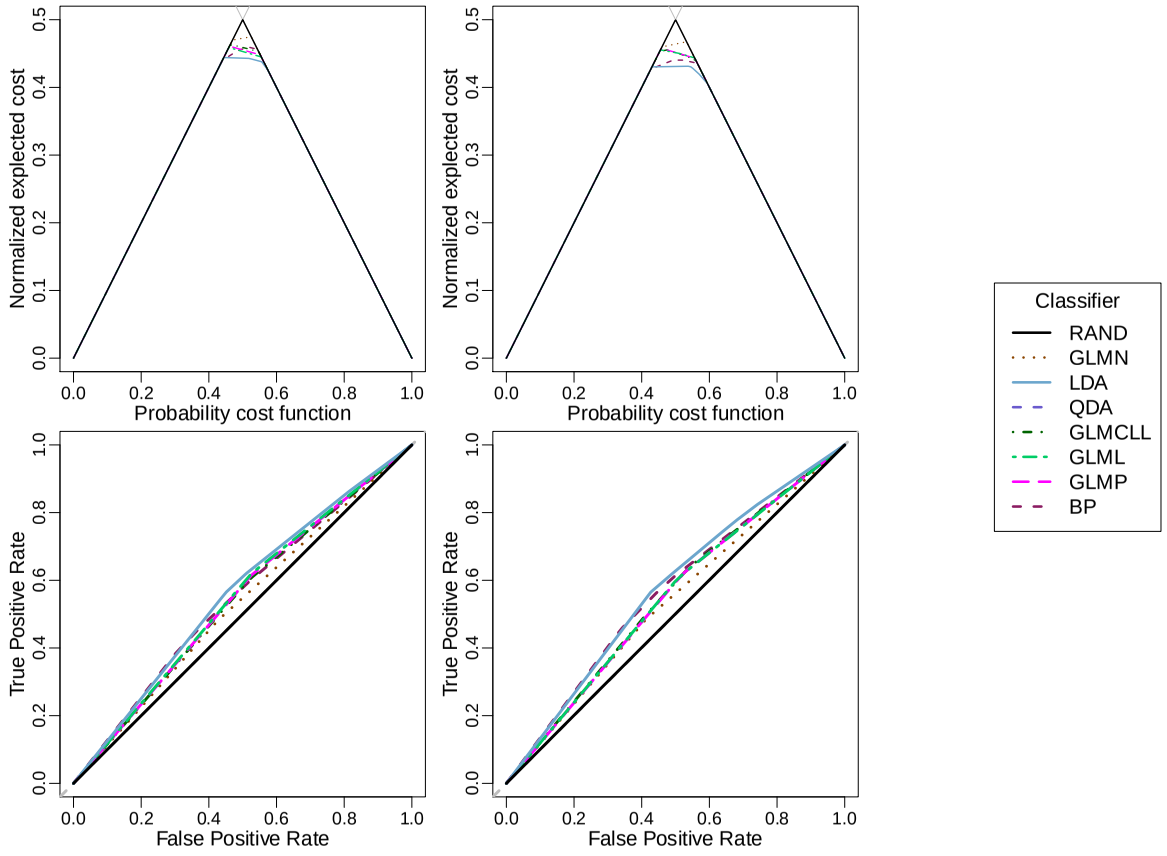


Figure E.11: Brier score by churn threshold $T(S)$ for forum with identifier 156 when churn window is four weeks. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure E.12: Lower envelope and ROC convex hulls for varying churn threshold $T(S)$ corresponding to forum with identifier 156 when churn window is four weeks.

E.5 Forum 418: very low activity

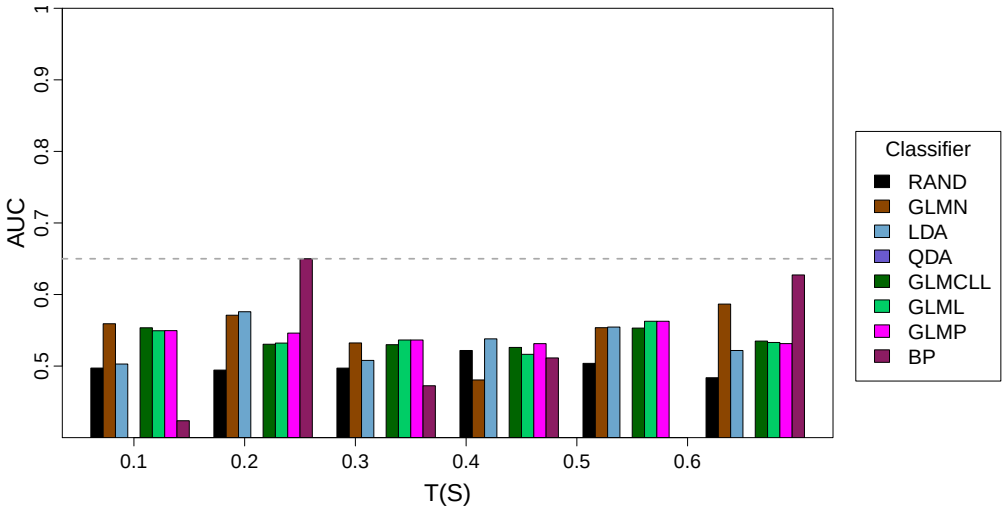


Figure E.13: Area Under (ROC) Curve (AUC) by churn threshold $T(S)$ for forum with identifier 418 when churn window is four weeks. The horizontal dashed grey line marks the highest (best) AUC measure observed.

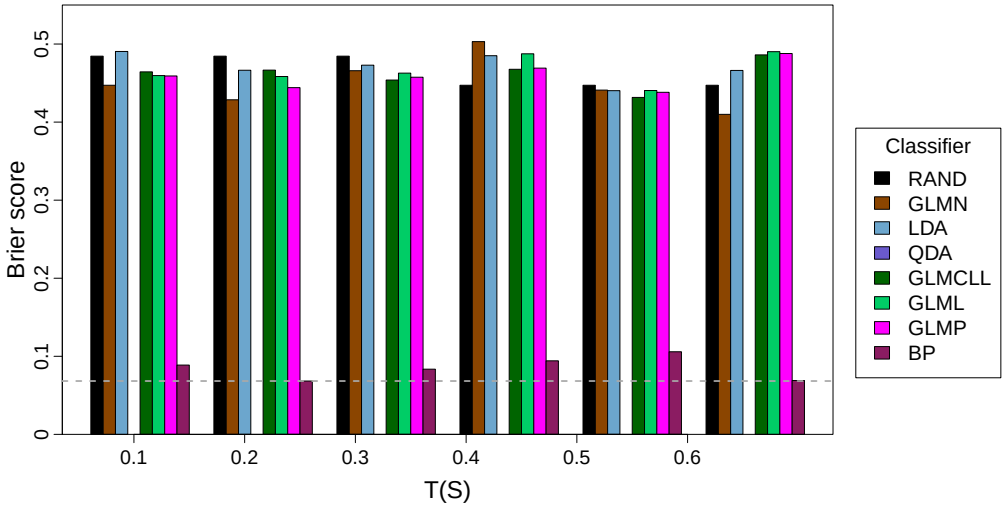
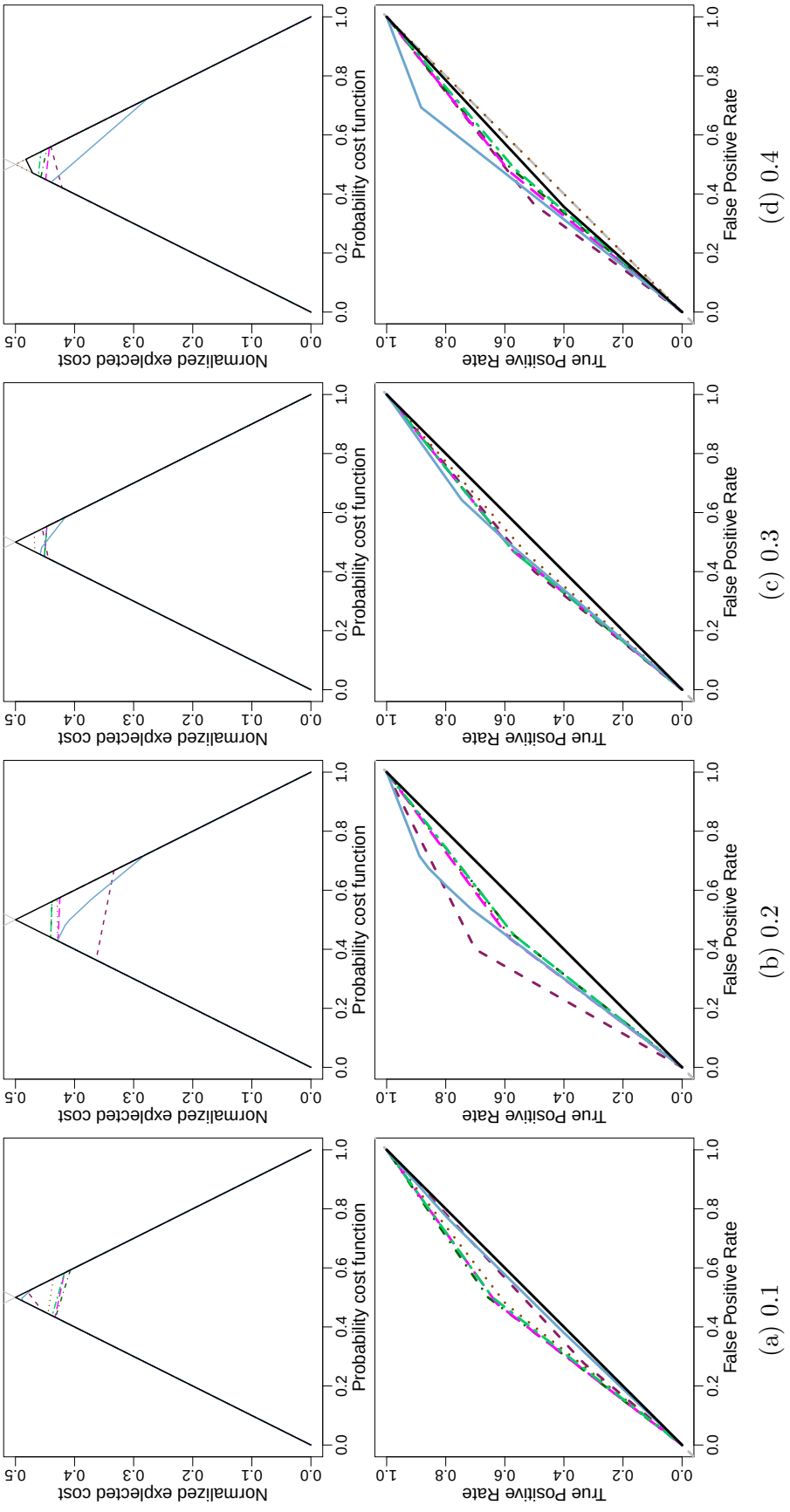
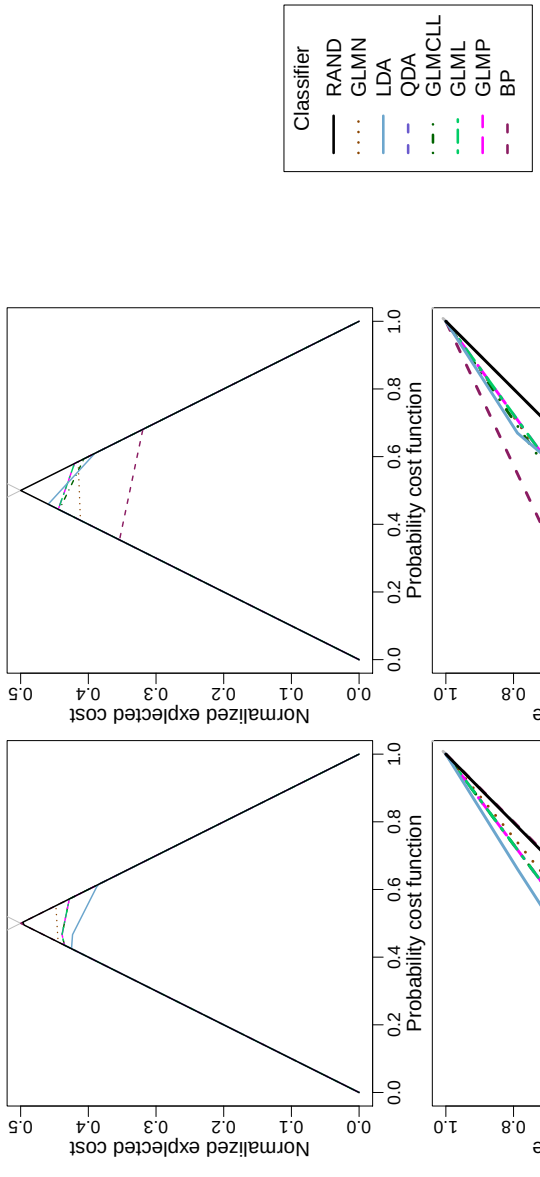


Figure E.14: Brier score by churn threshold $T(S)$ for forum with identifier 418 when churn window is four weeks. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure E.15: Lower envelope and ROC convex hulls for varying churn threshold $T(S)$ corresponding to forum with identifier 418 when churn window is four weeks.

Appendix F

Alternative Churn Formulations

This appendix provides two additional formulations of the individual user churn event that could be considered in the future. The first of these attempts to account for differences in user activity resulting in the same empirical probability of churn occurrence (Section F.1). The second additional formulation provides an entirely novel formulation of churn by conceiving it as a truly continuous response (Section F.2).

F.1 Accounting for differences in user activity

Our formulation of the churn event in the main body of this thesis can be criticised as it cannot discriminate between high and low activity users who have the same relative percentage decrease (2.3). By the definition of the probability of churn in (2.5), two users k and ℓ may have the same probability of churning whilst $\mu_C(v_k) >> \mu_C(v_\ell)$. This definition of churn therefore implies that the risk of a well respected user is no higher than a lesser respected user.

We propose to address this by multiplying the probability of churn occurrence (given in (2.8.1)) by the average reputation gained in the previous activity window $\mu_{PA}(v_i)$. We may mathematically write this response as:

$$z_i^* = \begin{cases} 0 & \text{if } \mu_C(v_i) \geq \mu_{PA}(v_i), \\ \left(1 - \left(\frac{\mu_C(v_i)}{\mu_{PA}(v_i)}\right)\right) \mu_{PA}(v_i) & \text{otherwise.} \end{cases}$$

But this is equivalent to:

$$z_i^* = \begin{cases} 0 & \text{if } \mu_C(v_i) \geq \mu_{PA}(v_i), \\ \mu_{PA}(v_i) - \mu_C(v_i) & \text{otherwise} \end{cases}$$

which is merely the difference between user activity in the previous activity window and user activity in the churn window. An unfortunate consequence of this definition is that the response is no longer a viable probability measure, or continuous over all real numbers.

The change to the churn response must also be accounted for in the churn threshold. We propose multiplying the churn threshold by the average of the average activity over all reputable users in our sample population

$$T(S)^* = T(S) \cdot \frac{1}{N} \sum_{i=1}^{N} \mu_{PA}(v_i).$$

F.1.1 Preliminary results

Due to time constraints, we are only able to present results for the fora with numerical identifiers 50, 142 and 418. In Chapter 6.5 we discussed how the results for Forum 418 were not robust, given that the majority of the classification methods considered cannot be fitted to such small sample sizes. As we are again considering the same sample population as in Section 6.5, our earlier observations regarding the sample size also holds for the results of this alternative definition of churn. We therefore disregard the results for forum 418 and observe linear discriminant analysis to again be the globally superior classifier for fora 50 and 142 over all churn thresholds. In addition, we again see no significant benefit in specifying the choice of the churn threshold $T(S)^*$.

The question is, whether this alternative definition of churn fairly represents the churn of a highly active respondent as different to the churn of a sporadically active respondent? Our definition of churn set out above directly addresses the case where two users k and ℓ have a similar probability of churning whilst k is significantly more active than ℓ. However, as the corresponding response $\mathbf{z}_i^*$ is equal to the observed decrease in activity, it is not relative unlike the formulation in the main document. Therefore, this alternative churn event views two users k and ℓ to have the same level of churn if they have the same decrease in activity from the previous activity window to the churn window. That is, if $\mu_{PA}(v_k) = 500$ and

$\mu_{PA}(v_\ell) = 50$ but $\mu_C(v_k) = 450$ and $\mu_C(v_\ell) = 0$ then our alternative formulation of churn says that the users k and ℓ have churned equally. We argue that this is a far more unfair representation of the disparity between the churn of reputable respondents with different reputations.

If we compare the figures in this section with their counterparts that were analysed in Section 6.5.1, there are no significant dissimilarities to the extent that even the cost and ROC curves appear similar in shape and curvature. That is, the classification methods seems to perform comparably for both the formulations of churn we have given thus far. However, comparing the formulation of churn given here with that outlined in Chapter 2, we are reasonably certain that some users would be classified differently. This implies that the classification methods must be performing well/poorly for a different set of users with regard to the different formulations. The next step would be to ascertain whether we are accurately predicting the same kinds of reputable respondents or whether we are identifying the more reputable/important ones. If further analysis of the types of reputable respondents which we are accurately predicting for both our formulations of the churn event did not reveal any worthwhile difference, we suggest three options that could be followed:

- explore additional alternative ways of interpreting the churn event (response);
- consider modelling churn as a continuous response (see Section F.2);
- use the formulation corresponding to Definition 2.2, given that it is a fairer representation.

Forum 50: bursty activity

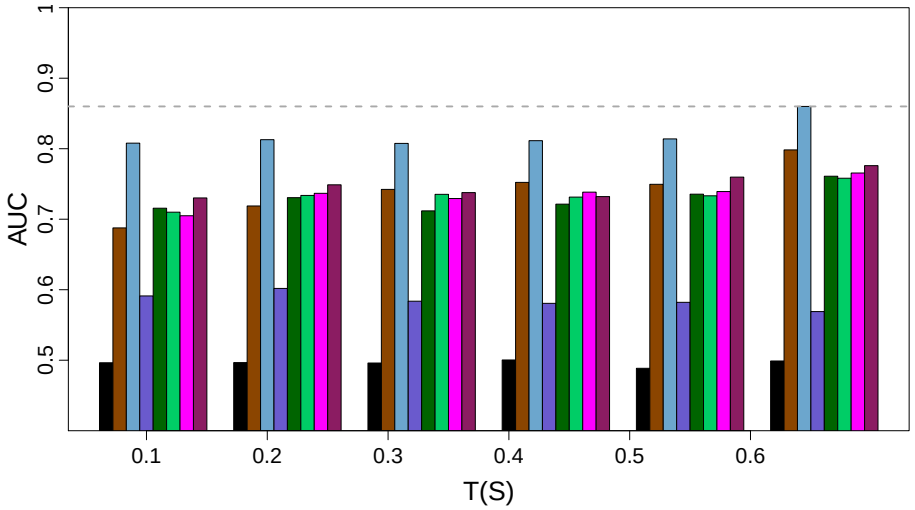


Figure F.1: Area Under (ROC) Curve (AUC) by churn threshold $T(S)^*$ for forum with identifier 50. The horizontal dashed grey line marks the highest (best) AUC measure observed.

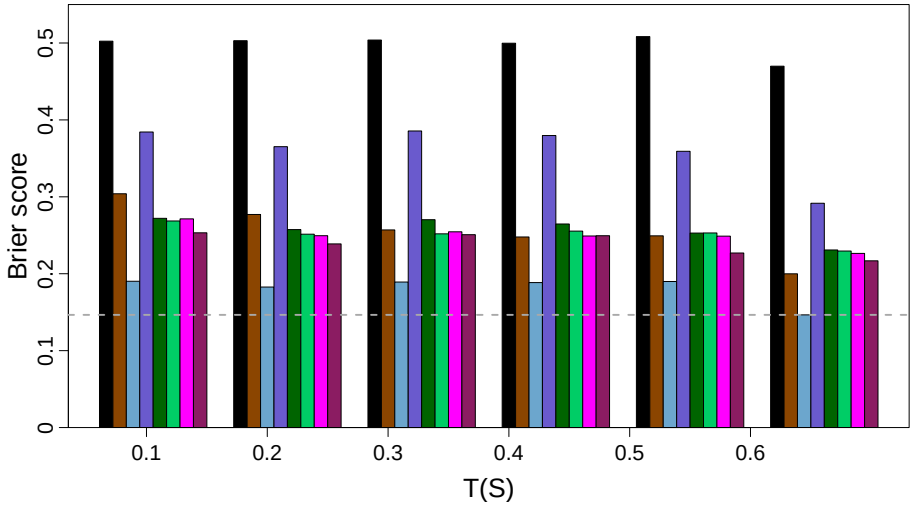
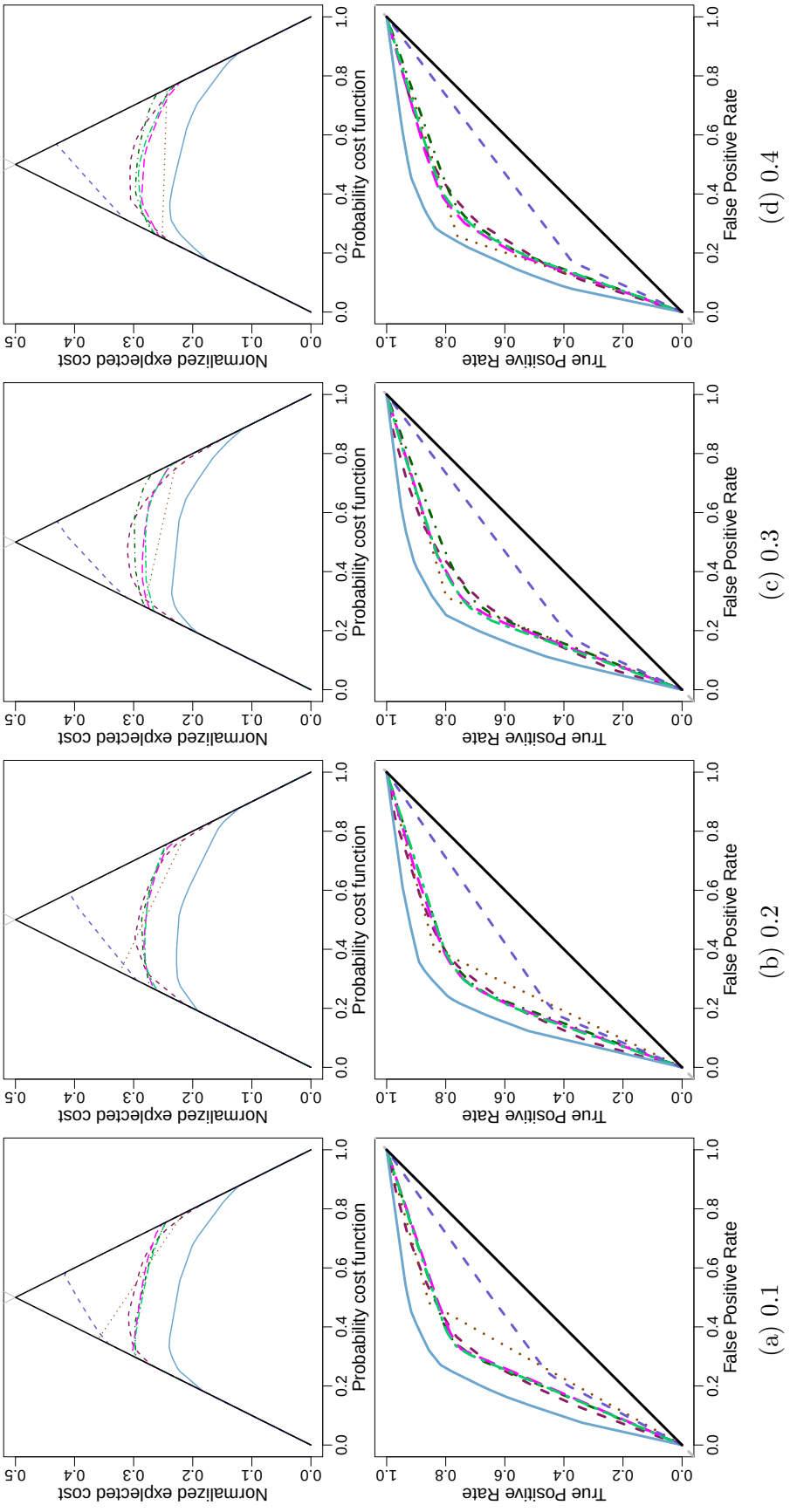
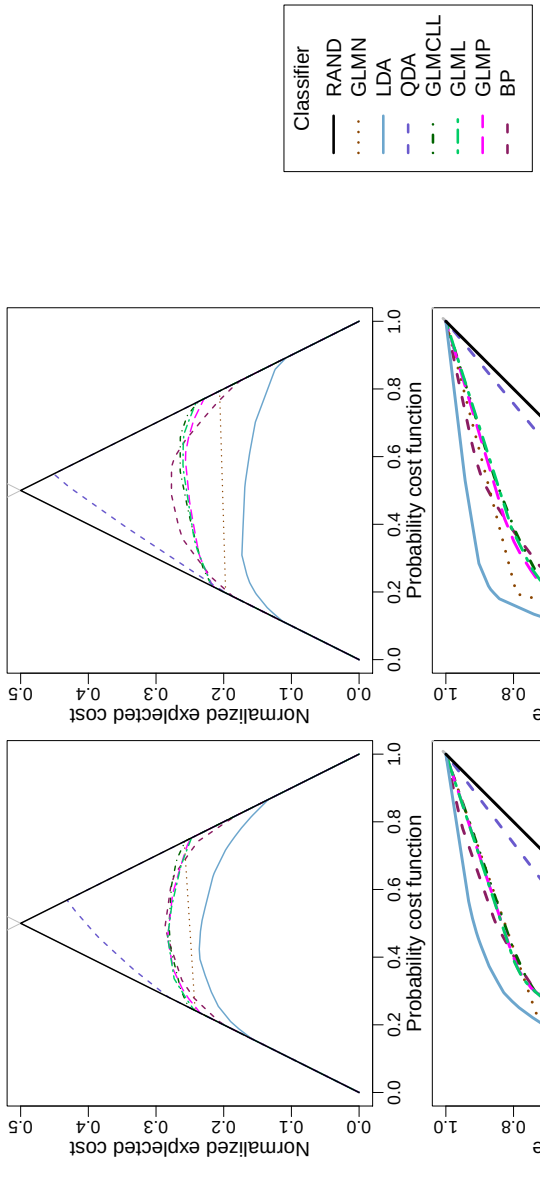


Figure F.2: Brier score by churn threshold $T(S)^*$ for forum with identifier 50. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure F.3: Lower envelope and ROC convex hulls for varying churn threshold $T(S)^*$ corresponding to forum with identifier 50 when churn window is one week.

Forum 142: very high activity

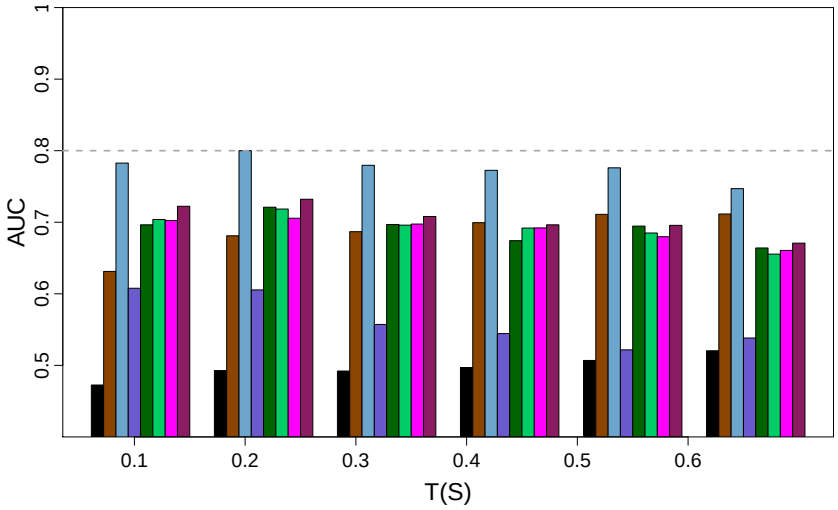


Figure F.4: Area Under (ROC) Curve (AUC) by churn threshold $T(S)^*$ for forum with identifier 142. The horizontal dashed grey line marks the highest (best) AUC measure observed.

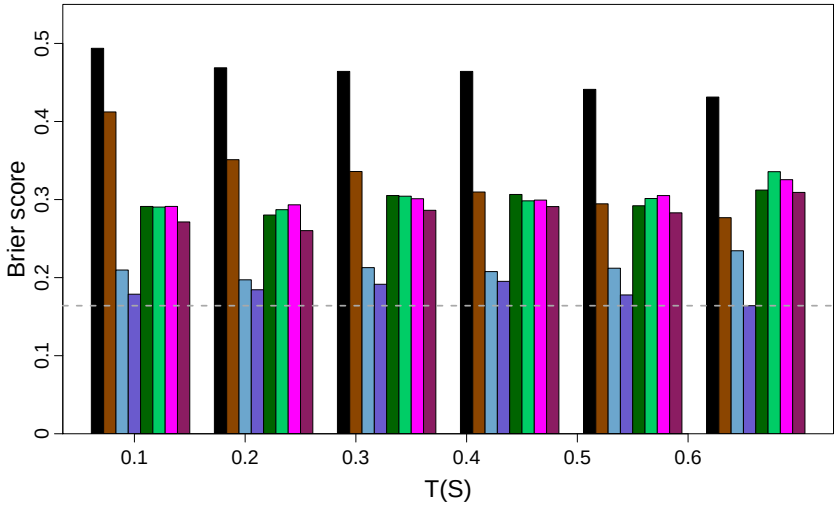
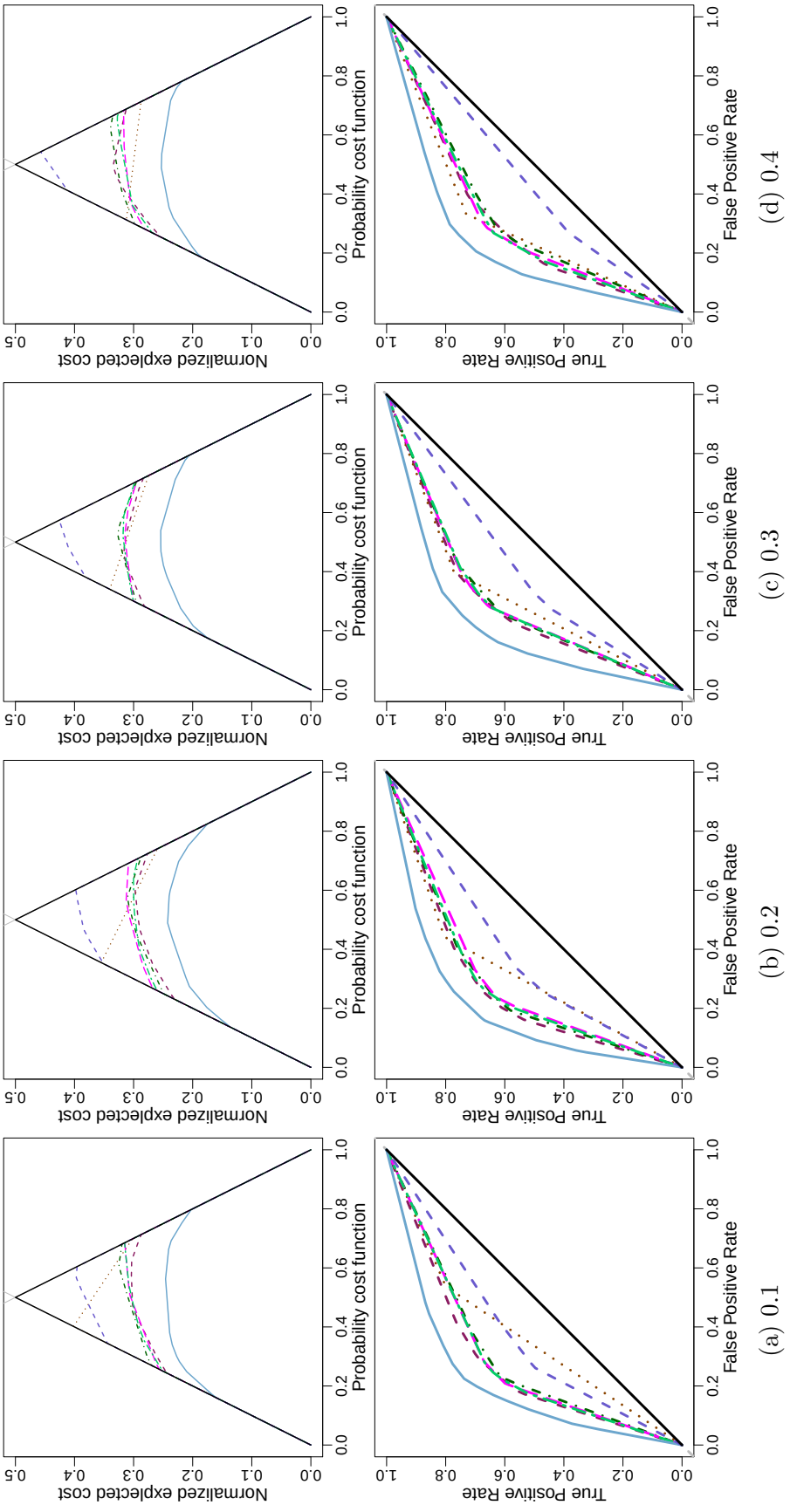
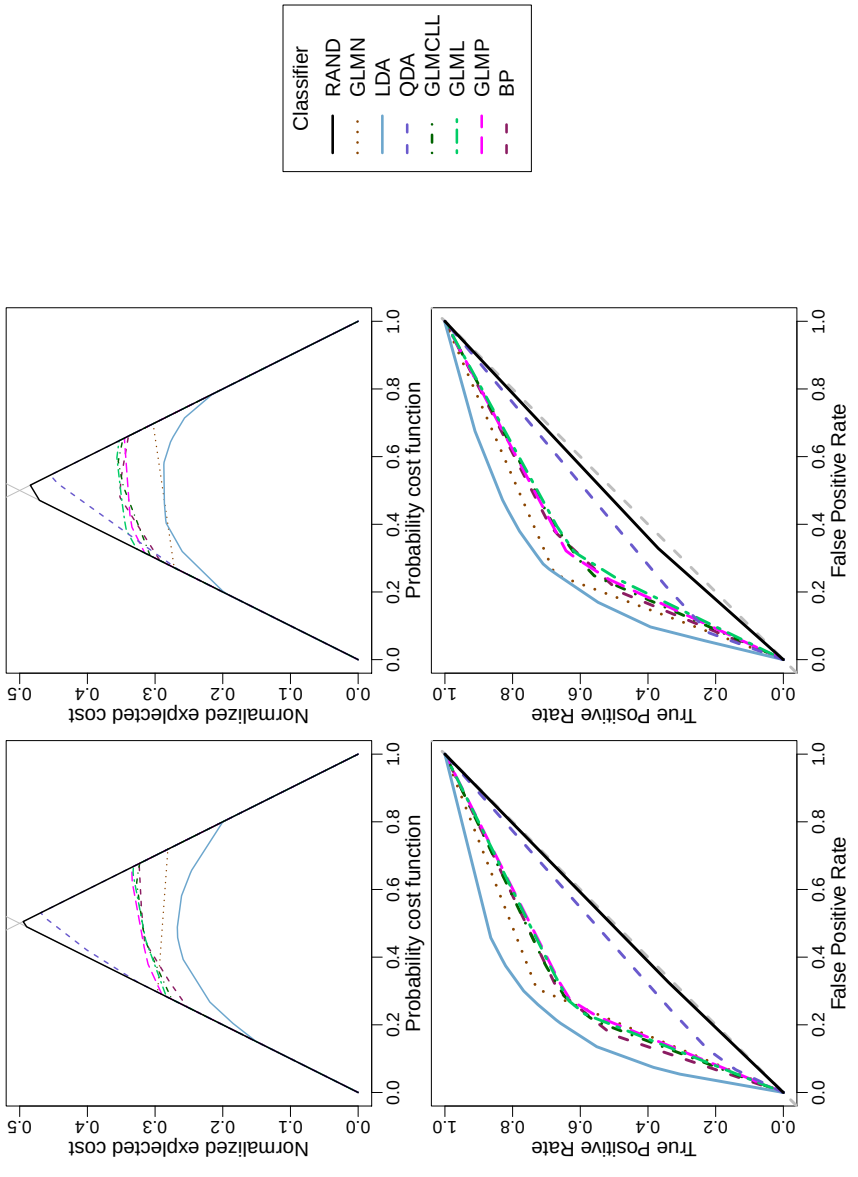


Figure F.5: Brier score by churn threshold $T(S)^*$ for forum with identifier 142. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure F.6: Lower envelope and ROC convex hulls for varying churn threshold $T(S)^*$ corresponding to forum with identifier 142 when churn window is one week.

Forum 418: very low activity

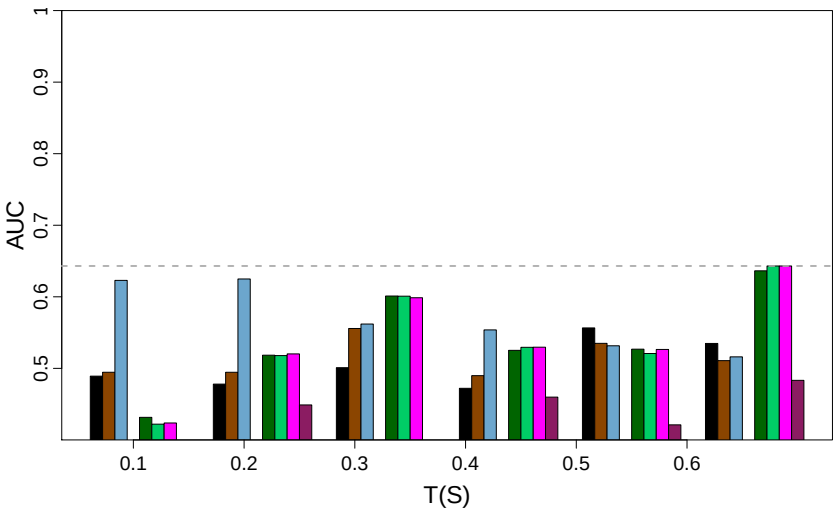


Figure F.7: Area Under (ROC) Curve (AUC) by churn threshold $T(S)^*$ for forum with identifier 418. The horizontal dashed grey line marks the highest (best) AUC measure observed.

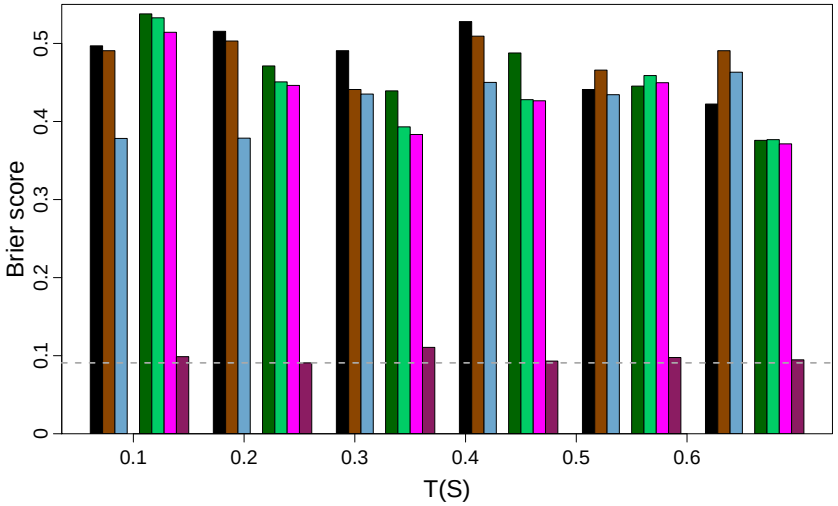
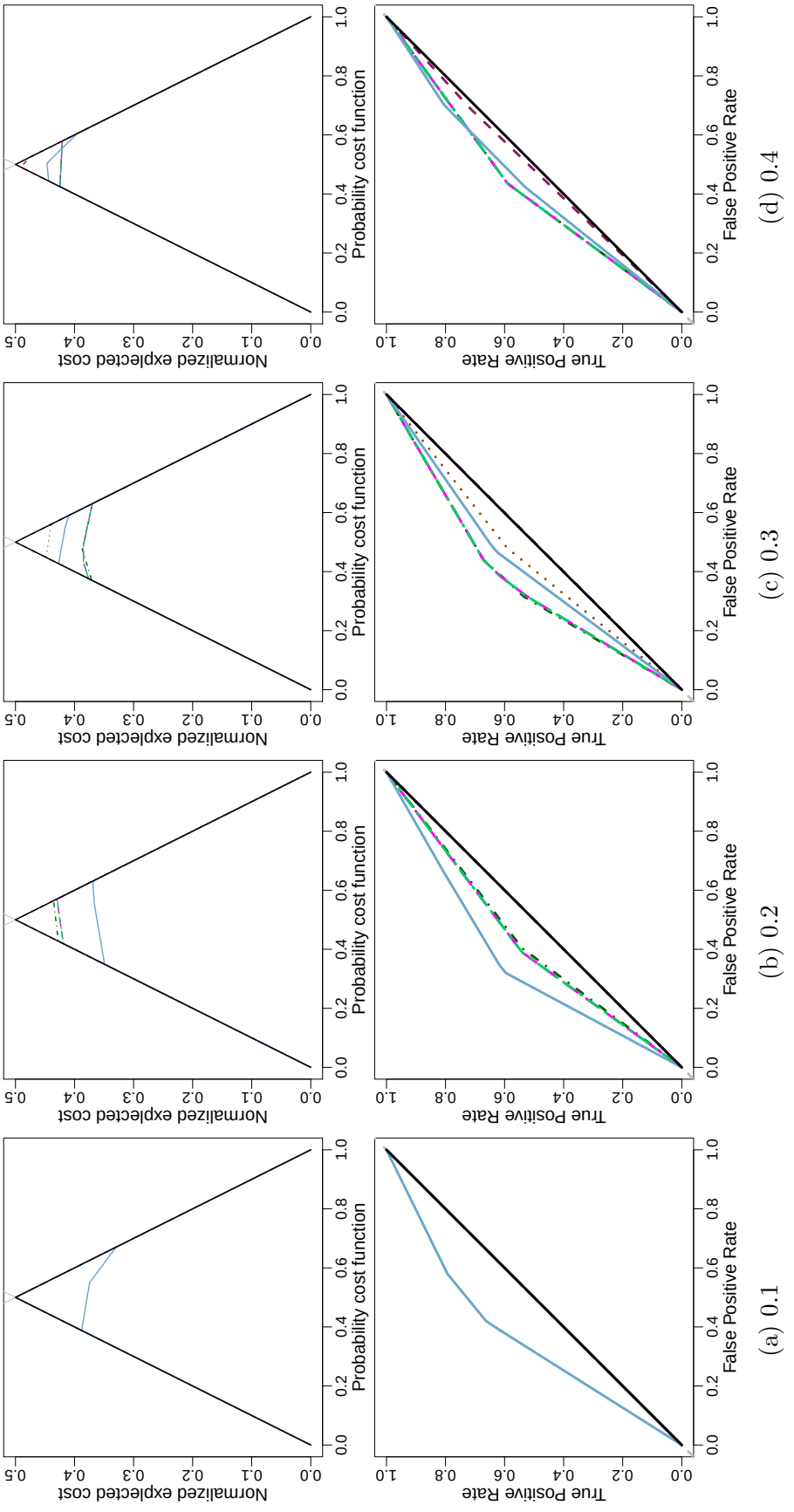
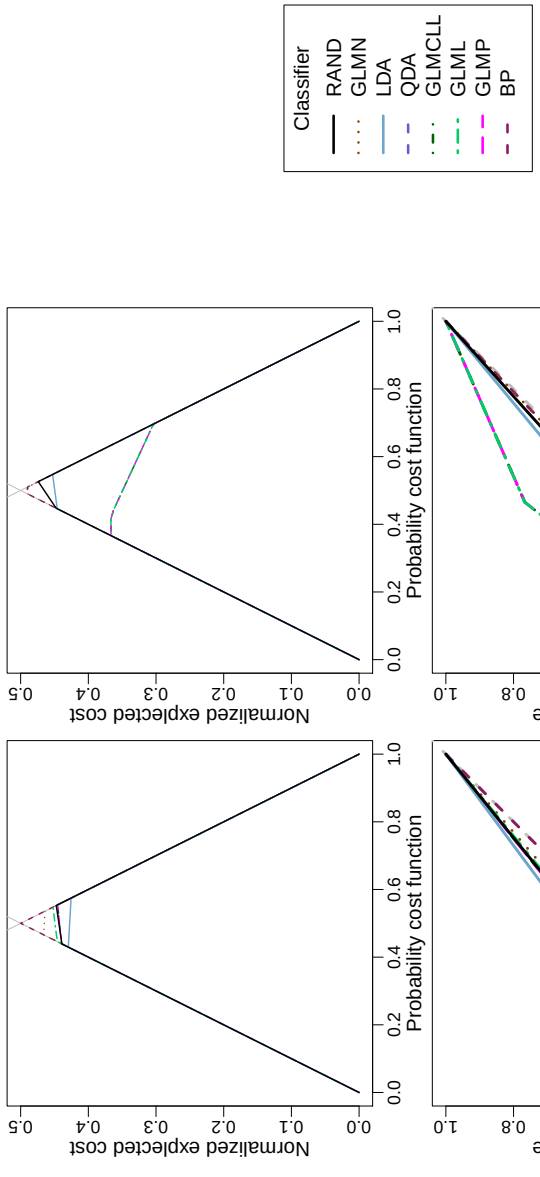


Figure F.8: Brier score by churn threshold $T(S)^*$ for forum with identifier 418. The horizontal dashed grey line marks the lowest (best) Brier score observed.



(a) 0.1 (b) 0.2 (c) 0.3 (d) 0.4



(e) 0.5 (f) 0.6

Figure F.9: Lower envelope and ROC convex hulls for varying churn threshold $T(S)^*$ corresponding to forum with identifier 418 when churn window is one week.

F.2 Churn as a continuous response

All previous analysis of churn formulates the event as a binary response and only classification methods are suitable to model binary response problems. If the event was defined by a continuous response, we might assume the response to be (approximately) normally distributed and consequently we would no longer be restricted to modelling churn as a classification problem.

To consider this further, we define the relative change in activity of the i^{th} user to be

$$q_i^* = \frac{\mu_C(v_i)}{\mu_{PA}(v_i)},$$

for $i = 1, \ldots, N$. We force q_i^* to be greater than zero by considering only those reputable users with $\mu_{PA}(v_i)$ and $\mu_C(v_i)$ strictly positive. We may then take the natural logarithm of q_i^*, and using the Taylor series expansion of $\ln(1+x)$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots,$$

we then have

$$\begin{aligned}
\ln(q_i^*) &= \ln\left(\frac{\mu_C(v_i)}{\mu_{PA}(v_i)}\right) = \ln\left(1 + \left(\frac{\mu_C(v_i)}{\mu_{PA}(v_i)} - 1\right)\right) \\
&= \frac{\mu_C(v_i)}{\mu_{PA}(v_i)} - 1 + \mathcal{O}\left(\left(\frac{\mu_C(v_i)}{\mu_{PA}(v_i)} - 1\right)^2\right) \\
&= \frac{\mu_C(v_i) - \mu_{PA}(v_i)}{\mu_{PA}(v_i)} + \mathcal{O}\left(\left(\frac{\mu_C(v_i) - \mu_{PA}(v_i)}{\mu_{PA}(v_i)}\right)^2\right)
\end{aligned}$$

which leads to

$$\begin{aligned}
\ln(q_i^*) &= -\frac{\mu_{PA}(v_i) - \mu_C(v_i)}{\mu_{PA}(v_i)} + \mathcal{O}\left(\left(\frac{\mu_C(v_i) - \mu_{PA}(v_i)}{\mu_{PA}(v_i)}\right)^2\right) \\
&= -q_i + \mathcal{O}\left(q_i^2\right) \qquad \text{(F.1)}
\end{aligned}$$

for q_i as defined in (2.3). Where q_i approaches zero the error term in (F.1) is negligible in comparison to the magnitude of q_i and $\ln(q_i^*)$ is a good approximation of the percentage decrease.

We thus define the continuous response for the i^{th} user, for $i = 1, \ldots, N$, as

$$z_i = \ln(q_i^*).$$

Therefore, assuming outlying values of q^* are not common, the associated random variable Z may be approximated by a normal distribution. This formulation of churn as a continuous response problem with a truly continuous response variable is revolutionary and is an avenue that should be explored.

Glossary

churn In the telecommunication industry, churn prediction is described as the identification of customers on the verge of transferring their custom to an alternative operator (Richter et al., 2010). This definition is accepted by some researchers in other areas but Karnstedt et al. (2010a) chooses to redefine churn specifically for analysing online communities. The main motivation is the difference in customer behaviour. In the telecommunication industry, leaving a service provider is a definite event; the customer is financially motivated to cancel their contract. However, in online communities, a user has little incentive to unsubscribe before leaving and there is therefore no definite event. Thus Karnstedt et al. defines user churn by the proportion of decrease in some measure of user activity. 20

outdegree the number of outgoing edges from vertex v_i in the directed graph G. 156

PageRank PageRank is a global importance rating of a vertex in some directed graph and is consequently a relative measure, for more details see Appendix A.2 or (Page et al., 1999). 23, 45

role In accordance with Merton (1968, p.41), the term *role* "refers to the behaviour of status-occupants that is oriented towards the patterned expectations of others". 20

Acronyms

ABAP Advanced Business Application Programming. 43

AUC Area Under (ROC) Curve. viii–xii, 37, 103, 110, 111, 113, 115, 122, 127, 135, 139, 150, 158, 161, 164, 167, 172, 175, 178, 181, 186, 189, 192, 196, 199, 202, 205, 208, 214, 217, 220

CORMSIS Centre of Operational Research, Management Sciences and Information Systems. 117

FP7 Seventh Framework Programme. 2

GLM Generalised Linear Model. iv, 64, 65, 67, 69, 71, 73

HPS Highest Point Scorer. 12, 29, 32

ICT Information Communication Technology. 2

LDA Linear Discriminant Analysis. iv, 62

MRR Most Reputable Respondent. 12, 28, 32–34

MRTM Most Responded To Message. 11, 29, 30, 32

MRTU Most Responded To User. 11, 30–32

OP Original Poster. 10–12, 27–30, 33, 34, 46–48, 121, 152

QDA Quadratic Discriminant Analysis. iv, 64

ROBUST Risk and Opportunity management of huge-scale BUSiness communiTy cooperation. iv, v, viii, xiii, 2–6, 10, 14, 21, 24, 42, 50, 51, 53, 78, 79, 81, 90, 91, 113–120, 147–149, 151

ROC Receiver Operating Characteristic. iv, viii–xii, 97, 98, 100–102, 104–111, 113, 115, 122–124, 126, 127, 130–132, 137–139, 142, 150, 160, 163, 166, 169, 174, 177, 180, 183, 188, 191, 194, 198, 201, 204, 207, 210, 213, 216, 219, 222

ROCH Receiver Operating Characteristic convex Hull. iv, viii, 102, 103

SAP Systems, Applications & Products in Data Processing. 9, 50, 51, 120, 148, 151

SCN SAP Community Network. iii, vii, xiii, 1, 7, 9–15, 17, 18, 22, 25, 26, 34, 39, 41, 43, 50, 51, 113, 114, 120, 148, 151

SQL Structured Query Language. 114, 120

TS Thread Solver. 12, 29, 47, 48

WP Work Package. 42, 116–118, 120

References

Agichtein, E., Gabrilovich, E., and Zha, H. (2009a). The Social Future of Web Search: Modeling, Exploiting, and Searching Collaboratively Generated Content. *IEEE Data Engineering Bulletin*, 32(2):52–61.

Agichtein, E., Liu, Y., and Bian, J. (2009b). Modeling Information-seeker Satisfaction in Community Question Answering. *ACM Transactions on Knowledge Discovery from Data*, 3(2):1–27.

Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons, Hoboken, NJ, USA, third edition.

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679.

Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2012). Discovering Value from Community Activity on Focused Question Answering Sites. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, page 850, New York, New York, USA. ACM Press.

Angeletou, S., Rowe, M., and Alani, H. (2011). Modelling and analysis of user behaviour in online communities. In Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., and Blomqvist, E., editors, *The Semantic Web – ISWC 2011*, volume 7031 of *Lecture Notes in Computer Science*, pages 35–50. Springer Berlin Heidelberg.

Armstrong, A. and Hagel, J. I. (1999). The real value of on-line communities. In Tapscott, D., editor, *Creating Value in the Network Economy*, pages 173–185. Harvard Business School Press, Boston, MA.

Baesens, B., Verstraeten, G., Van den Poel, D., Egmont-Petersen, M., Van Kenhove, P., and Vanthienen, J. (2004). Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. *European Journal of Operational Research*, 156(2):508–523.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC.

Barker, A. (1965). Monte Carlo Calculations of the Radial Distribution Functions for a Proton-Electron Plasma. *Australian Journal of Physics*, 18(2):119.

Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, 39(227):357–365.

Berrar, D. and Flach, P. (2012). Caveats and Pitfalls of ROC Analysis in Clinical Microarray Research (and How to Avoid Them). *Briefings in bioinformatics*, 13(1):83–97.

Bliss, C. I. (1934). The Method of Probits. *Science*, 79(2037):38–39.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. Wiley.

Bradley, A. P. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, 30(7):1145–1159.

Breiman, L. (1992). The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-fixed Prediction Error. *Journal of the American Statistical Association*, 87(419):738–754.

Breiman, L. and Spector, P. (1992). Submodel Selection and Evaluation in Regression. The x-Random Case. *International statistical review/revue internationale de Statistique*, 60(3):291—319.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly weather review*, 78(1):1–3.

Brooks, S. (1998). Markov Chain Monte Carlo Method and its Application. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):69–100.

Burez, J. and Van den Poel, D. (2009). Handling Class Imbalance in Customer Churn Prediction. *Expert Systems with Applications*, 36(3):4626–4636.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, New York.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Wadsworth Group, USA, second edition.

Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174.

Chen, M.-H. and Shao, Q.-M. (1999). Properties of Prior and Posterior Distributions for Multivariate Categorical Response Data Models. *Journal of Multivariate Analysis*, 71(2):277–296.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *Journal of the American Statistical Association*, 49(4):327–335.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall Ltd, London, UK.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., and Joshi, A. (2008). Social Ties and Their Relevance to Churn in Mobile Telecom Networks. In *Proceedings of the 11th international conference on Extending database technology Advances in database technology - EDBT '08*, page 668, New York, New York, USA. ACM Press.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, Cambridge, UK.

DeLone, W. H. and McLean, E. R. (1992). Information Systems Success: The Quest for the Dependent Variable. *Information Systems Research*, 3(1):60–95.

Delone, W. H. and McLean, E. R. (2003). The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. *Journal of Management Information Systems*, 19(4):9–30.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag.

Doyle, S. (2007). The Role of Social Networks in Marketing. *Journal of Database Marketing & Customer Strategy Management*, 15(1):60–64.

Drummond, C. and Holte, R. C. (2000). Explicitly Representing Expected Cost. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 198–207, New York, New York, USA. ACM Press.

Drummond, C. and Holte, R. C. (2006). Cost Curves: An Improved Method for Visualizing Classifier Performance. *Machine Learning*, 65(1):95–130.

Efron, B. (1975). The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. *Journal of the American Statistical Association*, 70(352):892–898.

Ermann, L., Chepelianskii, A. D., and Shepelyansky, D. L. (2012). Toward two-dimensional search engines. *Journal of Physics A: Mathematical and Theoretical*, 45(275101).

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fawcett, T. and Provost, F. (1997). Adaptive Fraud Detection. *Data Mining and Knowledge Discovery*, 1(3):291–316.

Finney, D. J. (1947). The Estimation from Individual Records of the Relationship Between Dose and Quantal Response. *Biometrika*, 34(3-4):320–334.

Fisher, D., Smith, M., and Welser, H. (2006). You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, pages 59b–59b. IEEE.

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 222(594-604):309–368.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188.

Flach, P. A. (2010). ROC Analysis. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning*, chapter R, pages 869–875. Springer US, Boston, MA.

Fliege, J., Aumayr, E., Engen, V., Fernandez, M., Hiscock, P., Hromic, H., Karnstedt, M., Mocan, A., Nasser, B., Rowe, M., Schwagereit, F., and Tye, E. (2012). D1.2: Risk Monitoring and Tracking in Online Communities. Technical report, University of Southampton.

Franke, N. and Shah, S. (2003). How Communities Support Innovative Activities: An Exploration of Assistance and Sharing Among End-Users. *Research Policy*, 32(1):157 – 178.

Franke, N., von Hippel, E., and Schreier, M. (2006). Finding Commercially Attractive User Innovations: A Test of Lead-User Theory*. *Journal of Product Innovation Management*, 23(4):301–315.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American statistical association*, 85(410):398–409.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6):721–741.

Geweke, J. (1989). Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica*, 57(6):1317–1339.

Glady, N., Baesens, B., and Croux, C. (2009). Modeling Churn Using Customer Lifetime Value. *European Journal of Operational Research*, 197(1):402–411.

Goldstein, M. and Dillon, W. R. (1978). *Discrete Discriminant Analysis*. John Wiley & Sons, USA.

Gottron, T. (2010). Report D10.1: Project Fact Sheet. Technical report.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley & Sons Inc.

Grover, V. and Segars, A. H. (1996). Information Systems Effectiveness: The Construct Space and Patters of Application. *Information & Management*, 31(4):177–191.

Haberman, S. (1977). Maximum Likelihood Estimates in Exponential Response Models. *The Annals of Statistics*.

Hadden, J., Tiwari, A., Roy, R., and Ruta, D. (2007). Computer Assisted Customer Churn Management: State-Of-The-Art and Future Trends. *Computers & Operations Research*, 34(10):2902–2917.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. Wiley, New York.

Hand, D. J. (2005). Good Practice in Retail Credit Scorecard Assessment. *Journal of the Operational Research Society*, 56(9):1109–1117.

Hand, D. J. (2006). Classifier Technology and the Illusion of Progress. *Statistical Science*, 21(1):1–14.

Hand, D. J. (2009). Measuring Classifier Performance: A Coherent Alternative to the Area Under the ROC Curve. *Machine Learning*, 77(1):103–123.

Hand, D. J. and Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45(2):171–186.

Hanley, J. A. and McNeil, B. J. (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143(1):29–36.

Hao, T. and Agichtein, E. (2012). Finding Similar Questions in Collaborative Question Answering Archives: Toward Bootstrapping-Based Equivalent Pattern Learning. *Information Retrieval*, 15(3-4):332–353.

Hardin, J. M. and Hilbe, J. (2001). *Generalized Linear Models and Extensions*. Stata Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition.

Hautz, J., Hutter, K., Fuller, J., Matzler, K., and Rieger, M. (2010). How to Establish an Online Innovation Community? The Role of Users and Their Innovative Content. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–11. IEEE.

Hernández-Orallo, J., Flach, P. A., and Ferri, C. (2012). A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss. *Journal of Machine Learning ...*, 13:2813–2869.

Hiltz, S. R. (1985). *Online Communities: A Case Study of the Office of the Future*. Ablex Publishing Corporation, USA.

Hiscock, P. A., Avramidis, A. N., and Fliege, J. (2013). Predicting Micro-Level Behavior in Online Communities for Risk Management. In *European Conference on Data Analysis, ECDA 2013*, Luxembourg. Springer.

Hobert, J. P. and Casella, G. (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association*, 91(436):1461–1473.

International Telecommunication Union (2013). The World in 2013: ICT Facts and Figures. Technical report, Geneva, Switzerland.

International Telecommunication Union (2014). http://www.itu.int/. Accessed: 10/04/2014.

Iriberri, A. and Leroy, G. (2009). A Life-Cycle Perspective on Online Community Success. *ACM Computing Surveys*, 41(2):1–29.

ISO31000:2009 (2009). Risk Management — Principles and Guidelines. Technical report, International Standards Organisation, Geneva.

Iversen, G. (1984). *Bayesian Statistical Inference*. Sage Publications, Beverly Hills, US, volume 07 edition.

Jadhav, R. J. and Pawar, U. T. (2011). Churn Prediction in Telecommunication Using Data Mining Technology. *International Journal of Advanced Computer Science and Applications(IJACSA)*, 2(2):17–19.

Janik, R., Sieprawski, M., Longosz, D., and Mazurek, M. (2013). Report D6.2.2 Integration of ROBUST Components - Version 2. Technical report, Software Mind S.A.

Jeon, J., Croft, W. B., and Lee, J. H. (2005). Finding Similar Questions in Large Question and Answer Archives. In *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05*, page 84, New York, New York, USA. ACM Press.

Jeon, J., Croft, W. B., Lee, J. H., and Park, S. (2006). A Framework to Predict the Quality of Answers with Non-Textual Features. In *Proceedings of the 29th*

annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06, page 228, New York, New York, USA. ACM Press.

Jones, Q., Ravid, G., and Rafaeli, S. (2004). Information Overload and the Message Dynamics of Online Interaction Spaces: A Theoretical Model and Empirical Exploration. *Information Systems Research*, 15(2):194–210.

Jørgensen, B. (1987). Exponential Dispersion Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(2):127–162.

Joyce, E. and Kraut, R. E. (2006). Predicting Continued Participation in Newsgroups. *Journal of Computer-Mediated Communication*, 11(3):723–747.

Karnstedt, M., Hennessy, T., Chan, J., Basuchowdhuri, P., Hayes, C., and Strufe, T. (2010a). Churn in Social Networks. In Furht, B., editor, *Handbook of Social Network Technologies and Applications*, pages 185–220. Springer US, Boston, MA.

Karnstedt, M., Hennessy, T., Chan, J., and Hayes, C. (2010b). Churn in Social Networks: A Discussion Boards Case Study. In *2010 IEEE Second International Conference on Social Computing*, pages 233–240. IEEE.

Karnstedt, M., Rowe, M., Chan, J., Alani, H., and Hayes, C. (2011). The Effect of User Features on Churn in Social Networks. In *Proceedings of the 3rd International Web Science Conference on - WebSci '11*, pages 1–8, New York, New York, USA. ACM Press.

Keegan, B. and Gergle, D. (2010). Egalitarians at the Gate. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*, page 131, New York, New York, USA. ACM Press.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Koh, J., Kim, Y.-G., Butler, B., and Bock, G.-W. (2007). Encouraging Participation in Virtual Communities. *Communications of the ACM*, 50(2):68–73.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*. Chapman & Hall/CRC.

Kubat, M., Holte, R. C., and Matwin, S. (1998). Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning*, 30(2-3):195–215.

Lemmens, A. and Croux, C. (2006). Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research*, 43(2):276–286.

Lin, H.-F. and Lee, G.-G. (2006). Determinants of Success for Online Communities: An Empirical Study. *Behaviour & Information Technology*, 25(6):479–488.

Liu, Q., Agichtein, E., Dror, G., Gabrilovich, E., Maarek, Y., Pelleg, D., and Szpektor, I. (2011). Predicting Web Searcher Satisfaction with Existing Community-Based Answers. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, page 415, New York, New York, USA. ACM Press.

Liu, Y., Bian, J., and Agichtein, E. (2008). Predicting Information Seeker Satisfaction in Community Question Answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, page 483, New York, New York, USA. ACM Press.

Lorenz, M. O. (1905). Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association*, 9(70):209–219.

Mahalanobis, P. C. (1936). On Generalized Distance in Statistics. *Proceedings of the National Institute of Sciences (India)*, 12:49–55.

Maxwell Harper, F., Raban, D., Rafaeli, S., and Konstan, J. A. (2008). Predictors of Answer Quality in Online Q&A Sites. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, page 865, New York, New York, USA. ACM Press.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall/CRC, 2 edition.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2001). *Generalized, Linear, and Mixed Models*. Wiley.

McWilliam, G. (2000). Building Stronger Brands through Online Communities. *Sloan management review*, 41(3):43–54.

Merton, R. K. (1968). *Social Theory and Social Structure*. The Free Press, New York, NY, USA, enlarged e edition.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087.

Montana, D. J. (1990). Empirical Learning Using Rule Threshold Optimization for Detection of Events in Synthetic Images. *Machine Learning*, 5(4):427–450.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., and Kaushansky, H. (2000). Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 11(3):690–6.

Nam, K. K., Ackerman, M. S., and Adamic, L. A. (2009). Questions In, Knowledge In? In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, page 779, New York, New York, USA. ACM Press.

Nasser, B., Engen, V., Crowle, S., and Walland, P. (2013a). A Novel Risk-based Approach for Online Community Management. In *ICIW 2013, The Eighth International Conference on Internet and Web Applications and Services*, pages 25–30.

Nasser, B., Engen, V., Jacyno, M., Crowle, S., Sahani, B., Fliege, J., Avramidis, T., Tye, E., Hiscock, P., and Rowe, M. (2011). D1.1: Representation of Risks in Online Communities. Technical report, University of Southampton.

Nasser, B., Engen, V., Meacham, K., Tye, E., and Hiscock, P. (2013b). D1.3: Beta Real-Time Risk Management Framework. Technical report.

Natarajan, R. and McCulloch, C. E. (1995). A Note on the Existence of the Posterior Distribution for a Class of Mixed Models for Binomial Responses. *Biometrika*, 82(3):639–643.

Natarajan, R. and McCulloch, C. E. (1998). Gibbs Sampling with Diffuse Proper Priors: A Valid Approach to Data-Driven Inference? *Journal of Computational and Graphical Statistics*, 7(3):267–277.

Ngonmang, B., Viennet, E., and Tchuente, M. (2012). Churn Prediction in a Real Online Social Network Using Local Community Analysis. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 282–288. IEEE.

Nonnecke, B., Andrews, D., and Preece, J. (2006). Non-Public and Public Online Community Participation: Needs, Attitudes and Behavior. *Electronic Commerce Research*, 6(1):7–20.

Nonnecke, B. and Preece, J. (2000). Lurker Demographics. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '00*, pages 73–80, New York, New York, USA. ACM Press.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web.

Pal, A., Farzan, R., Konstan, J. A., and Kraut, R. E. (2011). Early Detection of Potential Experts in Question Answering Communities. pages 231–242.

Petter, S. and McLean, E. R. (2009). A Meta-Analytic Assessment of the DeLone and McLean IS Success Model: An Examination of IS Success at the Individual Level. *Information & Management*, 46(3):159–166.

Phadke, C., Uzunalioglu, H., Mendiratta, V. B., Kushnir, D., and Doran, D. (2013). Prediction of Subscriber Churn Using Social Network Analysis. *Bell Labs Technical Journal*, 17(4):63–75.

Preece, J. (2000). *Online Communities: Designing Usability and Supporting Sociability*. John Wiley & Sons., Chichester, UK.

Preece, J. (2001). Sociability and Usability in Online Communities: Determining and Measuring Success. *Behaviour & Information Technology*, 20(5):347–356.

Preece, J., Nonnecke, B., and Andrews, D. (2004). The Top Five Reasons for Lurking: Improving Community Experiences for Everyone. *Computers in Human Behavior*, 20(2):201–223.

Press, S. J. and Wilson, S. (1978). Choosing Between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association*, 73(364):699–705.

Provost, F. and Fawcett, T. (1997). Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pages 43–48, Menlo Park, CA. AAAI Press.

Provost, F. and Fawcett, T. (1998). Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pages 43–48, Menlo Park, CA. AAAI Press.

Provost, F. and Fawcett, T. (2001). Robust Classification for Imprecise Environments. *Machine Learning*, 42(3):203–231.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rheingold, H. (1994). *The Virtual Community: Homesteading on the Electronic Frontier*. MIT Press, Cambridge, Mass., US, rev. ed. edition.

Richter, Y., Yom-Tov, E., and Slonim, N. (2010). Predicting Customer Churn in Mobile Networks through Analysis of Social Groups. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 732–741.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, second edition.

Roberts, G. O. and Smith, A. F. M. (1994). Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms. *Stochastic Processes and their Applications*, 49(2):207–216.

ROBUST Consortium (2009). Risk and Opportunity management of hige-scale BUSiness communiTy cooperation (ROBUST) Integrating Project (IP) Proposal - FP7-ICT-2009-5 (PART B). Technical report.

Rowe, M., Fernandez, M., Angeletou, S., and Alani, H. (2013). Community Analysis through Semantic Rules and Role Composition Derivation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 18(1):31–47.

Rubin, D. B. (1987). Comment. The Calculation of Posterior Distributions by Data Augmentation: Comment: A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm. *Journal of the American Statistical Association*, 82(398):pp. 543–546.

Schall, D. and Skopik, F. (2011). An analysis of the structure and dynamics of large-scale q/a communities. In Eder, J., Bielikova, M., and Tjoa, A., editors,

Advances in Databases and Information Systems, volume 6909 of *Lecture Notes in Computer Science*, pages 285–301. Springer Berlin Heidelberg.

Scott, M. J. J., Niranjan, M., and Prager, R. W. (1998). Realisable Classifiers: Improving Operating Performance on Variable Cost Problems. In *Proceedings of the British Machine Vision Conference*, pages 306–315. BMVA Press.

Smith, A. and Roberts, G. (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1(11):3–23.

Smithson, S. and Hirschheim, R. (1998). Analysing Information Systems Evaluation: Another Look at an Old Problem. *European Journal of Information Systems*, 7(3):158–174.

Stewart, L. (1979). Multiparameter Univariate Bayesian Analysis. *Journal of the American Statistical Association*, 74(367):684–693.

Stone, M. (1974a). Cross-Validation and Multinomial Prediction. *Biometrika*, 61(3):509–515.

Stone, M. (1974b). Cross-Validatory Choice and Assessment of Statistical Predictions (with discussion). *J. R. Statist. Soc. B*, 36:111 – 147.

Stutzbach, D. and Rejaie, R. (2006). Understanding Churn in Peer-to-Peer Networks. In *Proceedings of the Sixth ACM SIGCOMM on Internet measurement - IMC '06*, page 189, New York, New York, USA. ACM Press.

Swets, J. (1988). Measuring the Accuracy of Diagnostic Systems. *Science*, 240(4857):1285–1293.

Tagarelli, A. and Interdonato, R. (2013). "Who's Out There?". In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, pages 215–222, New York, New York, USA. ACM Press.

Tanner, M. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions (hbk.)*. Springer, third edition.

Tanner, M. and Wong, W. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American statistical Association*, 82(398):528–540.

Tausczik, Y. R. and Pennebaker, J. W. (2011). Predicting the Perceived Quality of Online Mathematics Contributions from Users' Reputations. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, page 1885, New York, New York, USA. ACM Press.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701–1728.

Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S*. Springer, 4 edition.

Wang, G., Gill, K., Mohanlal, M., Zheng, H., and Zhao, B. Y. (2013). Wisdom in the social crowd: an analysis of quora. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1341–1352. International World Wide Web Conferences Steering Committee.

Webb, A. (2002). *Statistical Pattern Recognition*. John Wiley & Sons Ltd, Chichester, UK, second edition.

Wedderburn, R. (1976). On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models. *Biometrika*.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, Cambridge, UK.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., third edit edition.

Xue, X., Jeon, J., and Croft, W. B. (2008). Retrieval Models for Question and Answer Archives. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, page 475, New York, New York, USA. ACM Press.

Zellner, A. and Rossi, P. E. (1984). Bayesian Analysis of Dichotomous Quantal Response Models. *Journal of Econometrics*, 25(3):365–393.

Zhao, Y., Li, B., Li, X., Liu, W., and Ren, S. (2005). Customer Churn Prediction Using Improved One-Class Support Vector Machine. In Li, X., Wang, S., and Dong, Z., editors, *Advanced Data Mining and Applications*, pages 300–306. Springer Berlin Heidelberg, Berlin Heidelberg.

Zhu, T., Wang, B., Wu, B., and Zhu, C. (2011). Role Defining using Behavior-Based Clustering in Telecommunication Network. *Expert Systems with Applications*, 38(4):3902–3908.

Zweig, M. H. and Campbell, G. (1993). Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical chemistry*, 39(4):561–77.