European Journal of Operational Research

How can lenders prosper? Comparing machine learning approaches to identify profitable peer-to-peer loan investments

--Manuscript Draft--

Manuscript Number:	EJOR-D-19-00396R3
Article Type:	Innovative Application of OR
Section/Category:	(B) Investment analysis
Keywords:	Investment analysis,P2P lending, predictive modelling, ensemble learning, credit scoring
Corresponding Author:	Trevor Fitzpatrick University of Southampton Southampton, UNITED KINGDOM
First Author:	Trevor Fitzpatrick
Order of Authors:	Trevor Fitzpatrick
	Christophe Mues
Abstract:	Successful Peer-to-Peer (P2P) lending requires an evaluation of loan profitability from a large universe of loans. Predictions of loan profitability may be useful to rank potential investments. We investigate whether various types of prediction methods and the types of information contained in loan listing features matter for profitable investment. A range of methods and performance metrics are used to benchmark predictive performance, based on a large dataset of P2P loans issued on Lending Club. Robust linear mixed models are used to investigate performance differences between models, according to whether they assume linearity, whether they build ensembles, and which types of predictors they use. The main findings are that: linear methods perform surprisingly well on several (but not all) criteria; whether ensemble methods perform better than individual methods is measure dependent; the use of alternative text-based information does not improve profit scoring outcomes. We conclude that P2P lenders could potentially increase their investment returns by applying linear methods that directly predict the internal rate of return instead of other dependent variables such as loan default.
Suggested Reviewers:	
Opposed Reviewers:	

Highlights (for review)

- Successful Peer-to-Peer (P2P) lending requires an evaluation of loan profitability
- We investigate whether prediction methods and information matter for investment
- We find linear methods perform surprisingly well on several (but not all) criteria
- Ensemble methods outperformance depends on the training measure used
- Using alternative text-based information does not improve profit scoring outcomes
- Higher investment returns can be achieved by using linear profitability prediction

How can lenders prosper? Comparing machine learning approaches to identify profitable peer-to-peer loan investments

Trevor Fitzpatrick[1,*]

Southampton Business School, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

Risk Analysis & Data Analytics Division, Central Bank of Ireland, North Wall Quay, Dublin 1, Ireland

Christophe Mues

Southampton Business School, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

Abstract

Successful Peer-to-Peer (P2P) lending requires an evaluation of loan profitability from a large universe of loans. Predictions of loan profitability may be useful to rank potential investments. We investigate whether various types of prediction methods and the types of information contained in loan listing features matter for profitable investment. A range of methods and performance metrics are used to benchmark predictive performance, based on a large dataset of P2P loans issued on Lending Club. Robust linear mixed models are used to investigate performance differences between models, according to whether they assume linearity, whether they build ensembles, and which types of predictors they use. The main findings are that: linear methods perform surprisingly well on several (but not all) criteria; whether ensemble methods perform better than individual methods is measure dependent; the use of alternative text-based information does not improve profit scoring outcomes. We conclude that P2P lenders could potentially increase their investment returns by applying linear methods that directly predict the internal rate of return instead of other dependent variables such as loan default.

[*] Corresponding author.

Email addresses: Trevor.Fitzpatrick@soton.ac.uk (Trevor Fitzpatrick), C.Mues@soton.ac.uk (Christophe Mues)

[1] Ph+35312246000. The views expressed in the paper are those of the authors and do not represent the views of the Central Bank of Ireland or the European Central Bank/Single Supervisory Mechanism. The authors would like to thank the reviewers for their constructive suggestions, Lyn Thomas, Leonard Roseman, Justin Silk, Jonathan Crook, and participants at the 2015 Edinburgh Credit Scoring and Credit Control XIV and the 2017 Computational and Financial Econometrics (CFE) Conference for useful discussions on earlier versions of material contained in this paper. They would also like to thank Matt Stevenson for assistance with transfer learning word embedding methods.

Keywords: Investment analysis, P2P lending, predictive modelling, ensemble learning, credit scoring

1. Introduction

Peer-to-Peer (P2P) lending is a type of crowdfunding in which an online platform enables borrowers to obtain credit from a large number of individual lenders. Unlike other types of crowdfunding, which may be for altruistic motives, in P2P lending the lender has a financial return motive. The growth in this type of lending has been spurred by technological advances, changing consumer habits, higher costs of and lower access to bank finance for borrowers, and lower returns for investors from traditional investments (Vallee and Zeng, 2018). At present, the two largest P2P platforms in the US, Prosper and Lending Club, together lent over $76 billion by the end of 2019. In the Asia-Pacific region including China, lending by alternative finance providers (including P2P lenders) amounted to $ 221 billion at the end of 2018; in Europe, the total amount lent was just under $ 6.6 billion by end 2018.[2] In this paper, data from the Lending Club (LC) platform is used, as it is one of the largest P2P lenders currently operating in the US.

Similarly to traditional retail credit scoring, P2P loan platforms screen potential borrowers against their own acceptance criteria. For example, borrower identity verification requirements, a minimum credit bureau score, and other criteria may need to be met. After acceptance, borrowers are scored and allocated to a certain grade based on their characteristics, the requested loan amount, and their credit history. The loan is then listed on the platform. At this point, the decision whether to lend lies with the investors, as do the associated return and credit risk — if the borrower defaults on their payment obligations, the investor takes a loss. This is in contrast to bank lending, where once a borrower is accepted, credit is advanced by the bank and it is the bank itself that bears the risk and makes the return. To make this investment decision, P2P investors must weigh the importance of various attributes in determining whether a loan may present a profitable investment. However, it is not feasible for an investor to manually assess the large volume of listings. Nonetheless, the potential gains of a systematic assessment could be significant as, in recent years, advertised returns for this type of investment are comparable to those earned on high-yield bond portfolios.

[2] Based on Lending Club and Prosper website data, SEC filings, and Ziegler et al. (2020).

This prospect has attracted various types of investors. In the early years of P2P investments, they mostly consisted of retail investors funding individual loans. In recent years, institutional investors have become important in this market as well.[3] For some platforms, recent research has suggested that active or “loan-picking” strategies may yield more than passive institutional strategies (Balyuk and Davydenko, 2018). Therefore, an algorithmic approach that can produce loan-level predictions of (risk-adjusted) loan returns could be useful to rank potential investments. A comprehensive assessment is both timely and relevant because there are a wide range of prediction models and algorithms to choose from, various types of predictors, and different experimental settings to judge the effectiveness of such methods. The main goal of this paper is to provide this assessment.

In so doing, the paper makes three main contributions. First, we contribute to the emerging P2P literature (Vallee and Zeng, 2018; Jagtiani and Lemieux, 2018) and profit-scoring literature (Garrido et al., 2018; Verbraken et al., 2014), by assessing whether a profit-scoring approach is more useful to investors than one solely focused on avoiding loan default. We examine three differing alternative performance metrics from classification, ranking, and regression. This may help investors choose a suitable approach for loan selection.

Second, we contribute to the literature on the empirical assessment of machine learning models through using a variety of performance measures and a specific experimental framework to compare profit scoring methods. Given the relative success of non-linear and ensemble prediction methods in other application settings, we augment the standard testing framework to test the importance of these factors for performance. This broadens the literature to include factors associated with the variability of performance across methods, rather than solely identifying differences using the standard methods of omnibus tests for differences across methods.

Third, we investigate whether alternative text-based information provided along with the loan listing for three year loans has predictive value. This adds to the emerging research area of the use of alternative data for scoring in this alternative form of financial intermediation. If additional sources of information have predictive content, then it may provide more profitable investment opportunities.

[3] On the Prosper platform, over 90% of loans were provided through institutional channels (Balyuk and Davydenko, 2018)

The paper is organised as follows. The next section reviews related work and formulates the research questions. Sections 3 and 4 describe the data and methods, respectively. Section 5 then outlines the experimental design. The results of the experiments are reported in Section 6. Section 7 provides further discussion and elaborates on some of the robustness checks carried out. Section 8 concludes.

2. Related work and research questions

Against the backdrop of an evolving P2P lending market, a body of literature on P2P loan profit scoring is emerging. This work cuts across two different research communities: the Operations Research (OR) community, which tends to focus on P2P loan scoring methods, and finance, which studies specific aspects of P2P lending and its implications for risk and return.

A first perspective is provided by the OR literature on credit scoring for P2P lending (Malekipirbazari and Aksakalli, 2015; Emekter et al., 2014). Using the Random Forests algorithm, Malekipirbazari and Aksakalli (2015) find that credit history variables and score/grade application information are the most important determinants for Lending Club (LC) loans that default. The paper by Emekter et al. (2014) uses a logistic and a Cox proportional hazards model to investigate determinants of default. They find that credit grade, the borrower's debt-to-income ratio, FICO credit score band, and revolving credit utilisation rate are significant predictors of default.

Although default risk is indeed a concern for investors, they are primarily interested in identifying high-return loans, i.e. those loans that present a good trade-off between default risk and interest returns. Hence, a profit scoring strategy may be more appealing to them. In the P2P context, loan selection based on estimated profitability is particularly important, since a P2P investor, unlike a traditional bank, cannot benefit from the risk diversification of taking on large portfolios of loans and, on most platforms, they cannot set risk-adjusted prices.

The current P2P literature on profit scoring methods, however, is limited. Both Serrano-Cinca and Gutiérrez-Nieto (2016) and Guo et al. (2016) find that various profit scoring approaches are useful to generate returns for investors. They considered a limited selection of regression and non-parametric methods such as CART, logistic and kernel-based regressions, respectively. These are valid approaches. However, other methods such as deep learning (Kim et al., 2019; Sirignano et al., 2016) and ensemble methods such as random forests (which build not just one model but combine multiple estimates) have been found to be competitive in various related tasks. These include profit scoring applications (Verbraken et al., 2014), credit scoring (Lessmann et al., 2015), and other related applications (Fuster et al., 2018; Lessmann and Voß, 2017; Kim et al., 2019). This suggests a need for a more systematic comparison, in particular one that comprises both non-linear and ensemble methods and assesses their ability to improve predictive performance in the P2P profit scoring setting.

A second perspective on P2P lending is provided by the finance community. Their research considers various aspects of P2P financial intermediation. These include how investors adapt to specific changes in platform operation and available information (Miller, 2015), as well as broader considerations of how this type of lending could change financial intermediation mechanisms (Balyuk and Davydenko, 2018; Vallee and Zeng, 2018; Jagtiani and Lemieux, 2017). As well as assessing the impact of more traditional factors linked to creditworthiness (e.g. credit score, grades, debt ratios), this body of literature has also focused on alternative or “soft” information available in the P2P context, such as appearance and text descriptions (Duarte et al., 2012; Hertzberg et al., 2016; Jagtiani and Lemieux, 2018).

Whereas “hard” information can be easily compressed to numerical values or attributes (Liberti and Petersen, 2017), alternative information may be unverifiable/costly to verify, or based on some non-standard format like images or free text. Emerging research points towards some role for alternative or soft information, once processed appropriately, in predicting the probability of attracting P2P funding and subsequent credit risk performance (Duarte et al., 2012; Lin, 2016; Dorfleitner et al., 2016; Jagtiani and Lemieux, 2018). However, it is unclear whether this finding is platform-specific as most of this research, with the exception of Dorfleitner et al. (2016), is based on the Prosper platform. Analysing data collected from German P2P platforms, Dorfleitner et al. (2016) instead find that the list text may influence the funding probability but does not appear to be informative for default prediction.For credit risk assessment of small business loans, Stevenson et al. (2020) find that text does not add any predictive power to the other (hard) variables. In any event, further work is needed to establish whether soft information has any added value for profitability scoring.

Based on these gaps in the literature, the main goal of this paper is to investigate empirically which types of methods and sources of data are able to provide investors with more accurate predictions of P2P loan profitability. This is a broad objective; hence, it is useful to distinguish three specific research questions.

The first question relates to how different methods characterise the relationship between the profitability measure and the predictors. Linear methods, such as penalised linear regression approaches, may suffice if the underlying relationship between loan profitability and its predictors is linear. However, if the underlying relationship is non-linear, then methods originating from the machine learning community could provide a significant edge over linear regression based methods. Given that there is little theory to guide the selection of either of these approaches in this application setting, this forms the basis of the first research question: *Are non-linear models better at predicting P2P loan profitability than linear models?*

Second, having seen some evidence in the credit scoring and related literature that ensemble methods tend to perform better than single models (Lessmann et al., 2015; Lessmann and Voß, 2017), it is natural to ask whether this finding extends to P2P profit scoring as well. Hence, the second research question is: *Do ensemble methods predict P2P loan profitability better than individual models?*

Third, and finally, while certain forms of soft (e.g. free text based) information appear to matter for the likelihood of being funded or for default prediction, the predictive power of alternative information remains to be assessed for profit scoring. Given that this source of information is becoming more prevalent as platforms grow, understanding its relevance for investment decisions is also becoming more important. The third research question therefore is: *Does including alternative information into predictive models lead to more accurate and more profitable P2P investments than solely using hard information?*

3. Data

The data are from Lending Club's statistical information on application and subsequent payment data for loans originated from its platform. The application data all relate to loans with a 36-month maturity, originated between October 2008 and January 2014. The payment data for these loans start in October 2008 and end in March 2017. All of the loans are closed – they have either been paid off early (i.e. prepaid), paid off at maturity, or the borrower defaulted. The loan-level predictors are a combination of loan, borrower, credit risk and text-derived characteristics; we further added macroeconomic variables to this dataset.

The loan characteristics include loan amount and purpose. Credit risk attributes include the sub-grade assigned by Lending Club at issuance and the FICO credit score band. Borrower characteristics include previous inquiries in the past six months, adverse public records, and delinquencies within the past two years. They also include months of credit history, total open accounts, revolving balance on other credit lines, utilisation of revolving lines, monthly loan instalment to total income, annual income (after borrower incomes below/above the 0.01%/99.99% quantiles are given these respective quantile values), and overall debt-to-income for the borrower. The categorical variables indicate whether: the borrower's length of employment is unknown; their employment title is missing in the listing; their income is verified; and the borrower is a home owner.

The listing text for each loan is included as a series of features. The text is a concatenation of two free-text fields: the listing title and the description provided by the borrower. The text per listing is relatively short with two sentences on average and an average sentence length of 6 words. While there are a variety of possible approaches to including the text as features (word embeddings), not all of these may be effective here as these short texts suffer from sparsity, i.e. limited word co-occurrence in each listing. We discuss this problem further in Section 7 where we check our findings using embeddings from transfer learning methods.

We use a method adapted to short texts called a Biterm Topic Model (BTM) (Yan et al., 2013), which considers word co-occurrences (i.e. biterms) in the whole corpus of all training listing texts rather than at the individual listing level. In so doing, it tackles the sparsity problem associated with having short texts in single loan listings. This method has been found to perform better on a variety of text datasets than traditional topic models or more complex methods (Jipeng et al., 2019). The downside of this approach is that, as it produces listing-level probabilities for a series of topics derived across all listings, it could limit how expressive the model is when there are a large variety of diverse topics. Following the details in the annex we fit a BTM for a total of 18 topics. These topic probabilities were then included as features in the models and the resulting performance differences tested as before.

Two controls for prevailing macroeconomic conditions are the state-wide unemployment rate, lagged two quarters before issuance of the loan and the year-on-year change in the OFHEO house price index, lagged two quarters prior to the issuance quarter of the loan.

We consider three different dependent variables (see Section 5.3). The proposed profitability measure is the Internal Rate of Return (IRR).[4] This is the discount rate that equates the present

[4] See Brealy and Myers (2001) for further detail on this measure.

value of a loan's monthly cash inflow to the face value of the loan. Formally, the IRR is defined as the value δ for which:

$$Amount_{t0} = \sum_{t=1}^{36} \frac{CF_t}{(1+\delta/12)^t} \tag{1}$$

The cash flows, CF_t, are positive as a borrower pays back the loan. If the borrower fails to pay back a loan for four periods, the loan is charged-off/defaulted, and the cash flows are terminated at that point. The IRR is chosen as a dependent variable because loan-level cash flow data are readily available and, as the IRR incorporates the actual payments made by borrowers, it is a direct measure of return for investors. This helps with comparisons to the literature, where IRR has been one of the main ways of measuring returns. P2P IRRs can be easily benchmarked against returns on alternative investment assets such as consumer credit card Asset-Backed Securities (ABS). To solve for the IRR, a root-finding algorithm is used. Note that, for this problem, the solution of this numerical procedure is unique as there are no irregular repayment cash-flows.[5]

4. Methods

Based on the literature, a representative set of regression methods of varying complexity were selected to predict profitability. They can be grouped into two main classes: individual and ensemble. Individual methods or models produce IRR estimates based on a single model. Ensemble methods use multiple instances of a base estimator, e.g. regression trees, combined in different ways.

As summarised in Figure 1, there are six individual methods specifying a linear relationship between the response variable and predictors. These individual methods are an implementation of a regularised glm based on elastic net (Zou and Hastie, 2005; Friedman et al., 2010), lasso (Tibshirani, 1996), ridge regression (Hoerl and Kennard, 1970), partial least squares (Mevik and Wehrens, 2007), and linear Support Vector Machines (SVM). The L2 linear regression is from (Fan et al., 2008; Helleputte, 2017).

The elastic net is a generalisation of lasso and ridge regression, combining regularisation via the

[5] We removed 504 loans that were repaid over a period of more than 36 months, that defaulted but were not charged off, or that were recorded as in default but were actually fully paid. A further 184 loans with zero payments were set to a -100% IRR. This means that cash flows after origination are always positive or zero.

ridge penalty and feature selection via the Lasso penalty. The relative weighting between the two penalties is determined adaptively from the data. Lasso and ridge are special cases of this. Partial least squares forms a linear combination of predictors, chosen in a way to summarise the variation in the predictors themselves and correlated with the response. The linear SVM was chosen to reduce computational complexity (Karatzoglou et al., 2004).

Figure 1, specifies a second group of individual non-linear methods including Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991), a simple neural network, and a deep learning model (Candel et al., 2020). MARS is a non-linear regression method that uses an additive piecewise linear representation of the original predictors to approximate a non-linear relationship with the dependent variable (Friedman, 1991). The neural network is a very simple single layer feed-forward neural network, with weight regularisation. The deep learning method is a multi-layer feed-forward neural network with two hidden layers of 40 units each with drop out, input drop out, and regularisation. The effect of the hidden layer, input drop out, and regularisation is help constrain overfitting. Deep learning has been successfully applied in several finance applications, such as bond return forecasting (Bianchi et al., 2018).[6]

The ensemble methods selected for the experiments are: random forests, bagged trees, gradient boosted trees, and five stacked models, as illustrated in the lower right-hand side of Figure 1. Random Forests (RF) are an ensemble method developed by Breiman (2001) which uses the Classification and Regression Trees (CART) recursive partitioning algorithm as a base learner. Many such trees are grown from bootstrapped training samples of the data, the predictions of which are averaged. Each time a split variable is chosen for an individual tree node, the RF algorithm only chooses from a small subset of *mtry* predictors instead of trying all available predictors. This process is repeated over many trees to create a forest. This has the effect of reducing correlation among the trees in the RF, thus reducing variance when averaging the trees; this typically results in improved predictive ability compared to CART or bagged trees. The latter can be thought of as a special case of random forests where *mtry* is set to the number of predictors. RFs have been applied successfully in a variety of domains including credit scoring (Lessmann et al., 2015).

Gradient boosted trees use a sequence of base learners that minimise a chosen loss function by the stage-wise addition of a new tree that leads to the largest reduction in loss, given the tree size.

[6] We are grateful to a reviewer who suggested the inclusion of additional methods.

With a squared loss function, the focus at each of these steps is on the residuals, i.e. the variation in the response not yet explained by the terms in the ensemble up to that step.

Finally, stacked ensembles use a library or set of first-level models to make a combined prediction. The first-level base models are meant to be a reasonably diverse group. A second-level model, referred to as a "metalearner", learns the optimal combination of these base learners. In this paper, the meta-learners that were tried were a simple average of the first level models, linear (stacked ridge, stacked L2liblinear) or non-linear (stacked gbm, stacked mars).

All of the methods have tuning parameters to optimise their predictive performance. The range of settings considered for each of the methods are summarised in Table 2.

The software used for all experiments is R. The following packages were used to implement the methods: *mlr* (Bischl et al., 2016); an implementation of MARS in a package called *Earth* (Millborrow, 2018); random forests/bagged trees using the *Ranger* package (Wright and Ziegler, 2017); partial least squares using the *pls* package (Mevik and Wehrens, 2007); *h2o* (Aiello et al., 2019) for the regularised glm, neural network, and deep learning; *glmnet* for ridge and lasso (Friedman et al., 2010); *LiblineaR* for the L2 linear regression (Fan et al., 2008; Helleputte, 2017) *kernlab* for linear SVMs (Karatzoglou et al., 2004); *XGBoost* for the gradient boosted trees (Chen and Guestrin, 2016); and *gbm* (Ridgeway, 2012).

5. Experimental design

This section describes the overall process flow for the experiments, outlining the choices made at each step of the setup. The prediction problem is to estimate a chosen profitability measure, y_i, for each P2P loan, i, from a vector of selected predictors, $\boldsymbol{x_i}^\top$. A range of individual models/algorithms and ensembles are trained to produce these estimates. As the form of this regression function is unknown, model tuning/selection is guided by optimising a suitable performance measure on the training data.

The various steps and choices in the experimental design are summarised in Figure 1. Details are described in the following subsections.

5.1. Predictor selection

The first step involves making a selection from the two groups of predictors outlined in Section 3 i.e., hard and soft. Either all data (including alternative/soft information, such as text-based

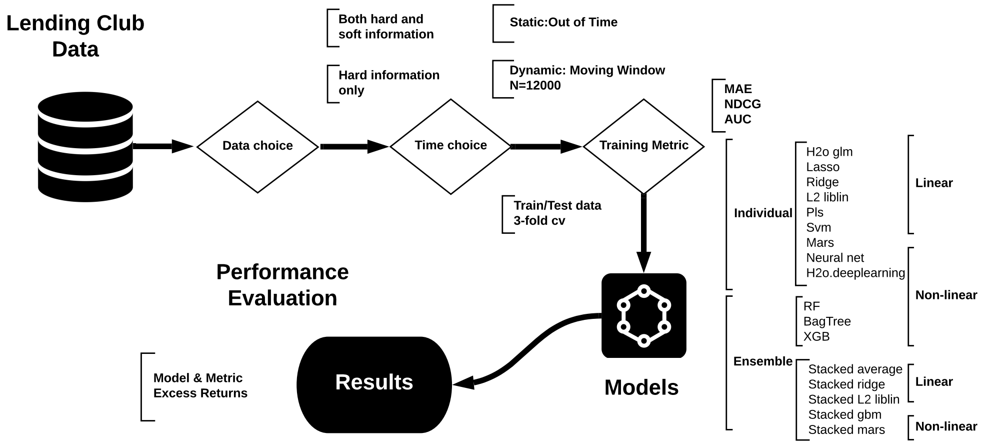

Figure 1: Experiment workflow

predictors) or only hard information (excluding the text-based predictors) are selected. The added value of soft predictors can be tested later on in the workflow.

5.2. Moving-window and out-of-time tests

In the next step, a series of training/test samples are created, either using a moving window or out-of-time test framework. Moving window experiments can be useful to investigate how changes in time periods/sample size may affect performance and allow more robust answers to the research questions. Note that in previous work (Serrano-Cinca and Gutiérrez-Nieto, 2016), calendar periods were used instead, which has the disadvantage that results could be specific to one period or may not generalise to other periods, even if careful selection of calendar periods is carried out (Butaru et al., 2016).

Second, for advanced prediction methods to be useful to investors, and as part of a comprehensive empirical approach, an out-of-time test design is added. Using only data on completed loans available at the time of the investment decision, this can provide a more realistic assessment of the performance of various methods.

Both approaches are illustrated in Figure 2. For the moving window approach, a window size, n, is selected. The first n observations (according to origination time) are then used as training data; the next n are test data. In the subsequent step, the previous test data now become the training data and a new test set is selected. This continues until the full data set is exhausted.

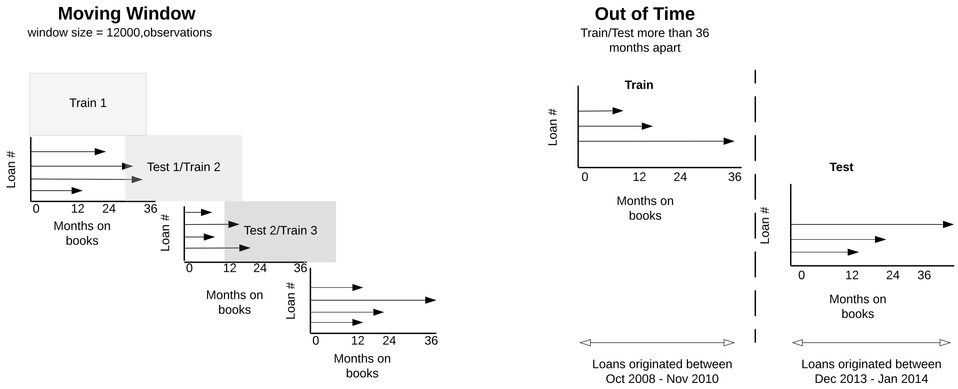
Moving Window
window size = 12000,observations
Train 1
Test 1/Train 2
Test 2/Train 3
Loan #
0
12
24
36
Months on books
Out of Time
Train/Test more than 36 months apart
Train
Test
Loans originated between Oct 2008 - Nov 2010
Loans originated between Dec 2013 - Jan 2014

Figure 2: Experiment data structure schematic

The same window size for train and test is chosen for simplicity and to not introduce another experimental variable. The window size is set to 12,000 observations. Alternative window sizes of 6,000-30,000 were used and the results did not differ markedly. For the out-of-time test, the training data used consist of loans with an origination date from October 2008 to November 2010. Given that a 36-month gap is required to observe the returns for the most recent of these training loans, the test data are loans that originated between December 2013 to January 2014. In the out-of-time framework, the training set has 12,799 observations; the test set has 10,658 observations.

5.3. Choice of dependent variable and performance evaluation metrics

The next step is to choose the type of dependent variable and appropriate performance measure for model tuning and evaluation. Previous work on P2P profit scoring focused on just one error metric and a limited range of models. However, in most profit scoring settings, the loans are ranked according to predicted profitability and a decision is made to invest in some proportion of the top ranked loans. There are particular challenges in such an application setting to pick one single metric. Therefore, three different evaluation measures are used – the Mean Absolute Error (MAE), the Normalised Discounted Cumulative Gain (NDCG), and the Area Under the ROC Curve (AUC) (see Table 1)

The MAE is the absolute residual between the predicted IRR and the actual IRR of each loan,

Table 1: Types of models and respective evaluation metrics

Dependent variable	Performance metric
IRR	MAE
Rank-transformed IRR	NDCG
Default $(y\|n)$	AUC

averaged over all n observations.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - f(x_i)| \tag{2}$$

Although MAE has an intuitive interpretation and is a robust measure of prediction accuracy, its use may not necessarily lead to higher returns. Investors have limited budgets and may only be interested in the top-k loans. Hence, an alternative approach is to focus on the relative loan return, i.e., using a rank-transformed IRR as the dependent variable. An additional rationale for transforming IRRs is given by the non-normality of the distribution of IRR which may lead to non-normal residuals in a standard regression model. In this setting, it is natural to turn to an Information Retrieval (IR) metric. One such metric, which is widely used in the learning to rank literature (Liu, 2011), is the Normalised Discounted Cumulative Gain (NDCG). This metric is useful when there are non-binary relevance scores in a complete ranked order, such as ranked IRRs. In this paper, loans are evaluated over the top k results; e.g., with k=100, the accuracy of each method is evaluated by how well the predicted rank of the top 100 loans compares to their actual ranking. The value chosen for k (e.g. 100) reflects one way of comparing to the profit scoring literature.

NDCG is calculated by using the predicted relevance score $R(m)$ for item m – here the predicted rank (equation 3). In this paper, a loan with a higher IRR receives a lower numerical rank. This is divided by a discount factor to reward better predictions of the rank of items at the top of the list compared to further down the list. In the literature, the discount typically is $log(1+m)$ or $log_2(1+m)$. This results in the Discounted Cumulative Gain (DCG). The term Z_k is a normalisation term to scale the ranking from 0 to 1.[7] A higher value indicates a better ranking; i.e., a value of 90% or 0.9 deviates by 10% from the ideal ranking.

[7] It is calculated by assuming that R_{perf} is the perfect relevance or ranking order score, and discounting by the same discount term. Dividing the calculated DCG by the ideal DCG leads to the Normalised Discounted Cumulative Gain (NDCG).

$$NDCG(k) = Z_k \sum_{m=1}^{k} \frac{R(m)}{log_2(1+m)} \tag{3}$$

Since borrower default would be a key event turning any potential profit into a loss, and given that default prediction is the standard scoring practice in consumer lending, we also build models to predict default. In this case, the dependent variable is binary with 1 indicating a default event, and 0 indicating no default. To measure the predictive ability of this third series of models, we use a widely used metric: the AUC, which is short for the Area Under the Receiver Operating Characteristic Curve (ROC). The ROC curve is created by evaluating event probabilities produced by the model across a range of cut-off or threshold values. For each threshold value, the true positive rate (sensitivity) and the false positive rate (1 - specificity) are plotted against each other. The AUC is the area underneath this curve; the higher the AUC, the better the model is able to discriminate between default and non-default.

Other measures such as Expected Maximum Profit (Verbraken et al., 2014) or the h-measure (Anagnostopoulos and Hand, 2012) could be used. EMP is a useful measure but would likely need further adaption to the present setting, and we have left this for further research. Following a reviewer's suggestion, we used the h-measure but the results were very similar to the AUC results and are therefore not included.

In summary, we built models for three different choices of dependent variable: IRR, rank-transformed IRR, and a binary variable representing whether the loan defaulted. To train and evaluate those models, we used the following three performance metrics: MAE, NDCG, and AUC, respectively.

5.4. Model training

The meta parameters are shown in Table 2. Each method or model is trained using random search and three-fold Cross Validation (CV). For the moving window test, three-fold CV is carried out for each window. For the out-of-time test, we take five bootstrap samples (0.632 fraction) of the training data to train the methods also using three-fold CV. The same tuning parameter ranges are applied when the methods are trained and evaluated using the three performance measures.

Table 2: Parameters for regression and classification methods

Name	Meta parameters	Values
h2o glm	alpha	alpha =(0.0001,. . . ,0.5)
lasso	alpha, lamba	lamba =(0.0001,. . . ,1); alpha=1
ridge	alpha, lamba	lamba =(0.0625,. . . ,4,); alpha=0
l2liblin	cost	cost = (0.0001,. . . ,10)
pls	num principal components	number=(1,. . . ,10)
svm	cost	cost=$2^{(-5,\ldots,2.2)}$
mars	degree, nprune, nk	degree=(1,2); nprune=(15,. . . ,40); nk=(10,. . . ,30)
bagged trees	ntrees	ntrees =(100,500,1000); min node size=5
rf	mtry	ntrees=1000; min node size=3; mtry=(3,. . . ,9)
xgboost	eta, max depth, sub-sample, lambda	nrounds=1000; min.child.weight=3; eta=(0.0075,0.01); max depth=(3,4,5,6); sub-sample=(0.5,0.632,0.75); lambda = $2^{(-10,\ldots,-1)}$
neural net	size, l2	size =3; l2= (0.0001,. . . ,0.5)
h2o deep learning	l1, l2, epochs	epochs =(10,20,30); l1=l2= (0.00001,. . . ,0.001); input dropout =0.05; hidden layers =(40,40,40); hidden drop out =(0.5,0.5,0.5)
sl.avg	none	none
sl.ridge	lambda	lambda=0.0625
sl.liblin	cost	cost=0.1
sl.mars	degree, nprune, nk	degree=2; nprune=5; nk=10
sl.gbm	ntrees, shrinkage, train fraction	ntree=500; shrinkage=0.01; train fraction=0.75

5.5. Statistical testing framework

To answer the three research questions outlined in section 2, a suitable statistical framework must be chosen. This is a different type of exercise than conventional model benchmarking, where the goal is to identify which methods significantly outperform which others; there, a common methodology is to apply a Friedman test to the observed differences in rank performance, followed by post-hoc tests controlling for multiple comparisons (see e.g. Lessmann et al. (2015)). The Friedman test, however, is single factor and tests if there are differences between methods. In this paper, we want to determine whether there are differences between the predictive performance of different (groups of) methods and what role the three specific factors associated with the research questions play:

1. whether a linear or nonlinear method is used;
2. whether an individual model or ensemble is used;
3. whether soft information is added to the predictors used in the model.

The experimental factors are approximately balanced for linear vs non-linear (9 vs 8 models) and ensemble vs individual (8 vs 9 models) and balanced for including or excluding soft information. One option to address the questions above is a repeated measures ANOVA in which the model performance metric is the dependent variable and the three experimental factors are the between (linear/non-linear, individual/ensemble) and within (no soft/with soft information) variables.

However, in the current application, some of the assumptions required for ANOVA may not hold. These include that the performance measures be drawn from a normally distributed population and that the variances in performance across methods are assumed to be equal (sphericity assumption). A second challenge to non-normality lies in the nature of the performance measures themselves.

The MAE is left-bounded at zero; both NDCG and AUC are bounded between zero and one, with the AUC typically between 0.5 and 1. The first challenge is likely to be more relevant than the second, as models rarely produce an AUC/NDCG of 0 or 1, or an MAE of 0 or a large positive real number.

To deal with these challenges, a Robust Linear Mixed Model (RLMM) is used to produce the results presented in the main text (Koller, 2016). This approach has two advantages in this experimental design. First, the method can cope reasonably well with non-normality and outlier observations, allows for differences in error variance and incorporates random effects to account for repeated measures. Second, it allows testing of the experimental factors of interest.

There are some downsides: inference using RLMMs is not yet well developed. Therefore, only t-statistics are referred to in the text. The robust linear mixed model was estimated using *rlmer* (Koller, 2016).

The results for the first two research questions are presented in one set of regression tables. For the third research question addressing the effect of adding soft information predictors, the result on the information coefficient from these regressions are contained in a separate set of tables. This question requires treating both the model and information type as within factors; i.e., each model experiences both levels of the information factor excluding/including the soft information.[8]

6. Results

Because of the two types of experiments conducted, the moving window and out-of-time results are discussed in separate sub-sections. Each sub-section presents the results in three ways. First, results are presented in a table summarising the performance of each method averaged over all model runs. Second, a graph is shown in which the methods are ranked according to their mean performance on each individual metric (note that ranks are used here as the original metrics are on different scales). Third, each set of results is then subjected to the statistical procedure outlined in sub-section 5.5 to determine whether there are significant differences in performance related to the research questions.

[8] The linear mixed model fit is a two factor within-subjects repeated-measures model. The first factor is *info* with two levels (hard information only, both types); the second factor is *modname* - the seventeen different model types.

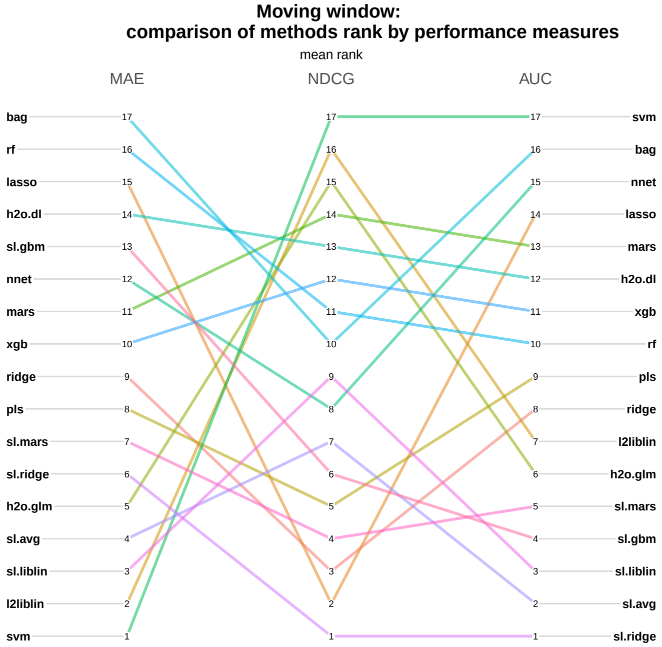

Figure 3: Performance ranked over metrics: moving window

6.1. Moving window

To help compare their predictive performance on the moving window experiments, the slopegraph in Figure 3 shows the performance ranking of the different methods, for each choice of dependent variable and corresponding performance metric. A lower numerical rank (lower on the y-axis) reflects better performance. In other words, a lower numerical rank for NDCG and AUC corresponds to a larger NDCG and AUC (see bottom-middle and bottom-right sections of the figure, respectively); a lower numerical rank for MAE means a lower absolute error value (see bottom-left section). The performance values used to produce these rankings are listed in Table 3. Each value represents the average test sample performance over the moving window of 12,000 observations.

Overall, three stacked ensembles (stacked ridge, stacked average, stacked liblinear) have the best average rank across the three dependent variables and associated performance metrics. These are followed by stacked mars and ridge regression. For the (binary) default prediction models evaluated using AUC, stacked methods (ridge, average, liblinear) constitute the top three best performing methods; using NDCG, using NDCG, stacked ridge and two linear individual methods (lasso, ridge) are the top three. For both AUC and NDCG, there is very little difference between the top three

Table 3: Rolling window: mean performance by metric

linear or non-linear	ensemble or individual	method	MAE	NDCG	AUC
linear	individual	h2o.glm	14.45	0.74	0.66
linear	individual	ridge	14.48	0.83	0.66
linear	individual	lasso	14.55	0.84	0.64
linear	individual	svm	9.10	0.69	0.54
linear	individual	pls	14.47	0.82	0.66
linear	individual	l2liblin	9.11	0.71	0.66
non-linear	individual	mars	14.50	0.76	0.64
non-linear	individual	nnet	14.54	0.82	0.64
non-linear	individual	h2o.dl	14.55	0.80	0.64
non-linear	ensemble	rf	14.75	0.81	0.66
non-linear	ensemble	bag	15.59	0.81	0.63
non-linear	ensemble	xgb	14.50	0.81	0.65
linear	ensemble	sl.avg	13.59	0.82	0.67
linear	ensemble	sl.liblin	9.12	0.82	0.67
linear	ensemble	sl.ridge	14.45	0.84	0.67
non-linear	ensemble	sl.mars	14.46	0.82	0.66
non-linear	ensemble	sl.gbm	14.54	0.82	0.67

models.[9] For the MAE measure, svm, l2liblinear, and stacked l2liblinear are the top three methods, followed by average stacked ensemble and h2o.glm.

A closer look at the results reveals that the performance ranking of some of the methods varies extensively depending on the choice of dependent variable and related performance measure. For example, svm is ranked as the top performer when using the MAE criterion to predict IRR, but ranked 17th for AUC (predicting default Y/N) and NDCG (rank transformed IRR), respectively. Stacked ridge is the best method using the AUC and NDCG criteria, yet has a much lower ranking when trained using the MAE. This variability could potentially be linked to the chosen tuning strategy, as the parameter range used to optimise the performance metric is not varied across MAE, NDCG, or AUC. However, we believe that the added complexity of further varying the tuning strategies for each performance measure is not required to answer the chosen research questions.

[9] The AUC values are in the range reported by Malekipirbazari and Aksakalli (2015) but lower than the best performing random forest found by those authors. This is likely because the results in Table 3 are averages over moving windows, and unlike Malekipirbazari and Aksakalli (2015) are not based on static samples.

6.1.1. Robust LMM: moving window

Table 4 contains the test results for the moving window approach. The variables representing the research questions are categorical. Following the discussion in sub-section 5.5, the reference category for the two-level factor *lin.nonlin (non-linear/linear)* is non-linear; the reference category for the two-level factor *ensemble (ensemble/individual)* is ensemble. Dividing the lin.nonlin variable coefficient estimates by their standard error gives t-statistics of -8.43, -0.95, 6.47, respectively, for the MAE, NDCG, and AUC criteria. This means using linear methods compared to non-linear methods tends to improve performance in this setting, lowering the MAE of IRR estimates and increasing the AUC of predicted default risk. For the ensemble variable, there is a small t-statistic for MAE (-0.2) and larger values for NDCG (-4.76), and AUC (-5.59); i.e., using individual methods as opposed to ensemble methods reduces performance in this setting.

Table 4: Rolling window: effects of linearity/non-linearity and ensemble/individual factors on performance

	MAE	NDCG	AUC
Intercept	14.6955	0.8214	0.6533
	(0.2049)	(0.0055)	(0.0015)
lin.non = linear	−1.9465	−0.0060	0.0111
	(0.2332)	(0.0063)	(0.0017)
ensemble = individual	−0.0474	−0.0300	−0.0095
	(0.2365)	(0.0063)	(0.0017)
Num. obs.	442	442	442

Standard errors in parentheses

Finally, to test the role of information on model performance, Table 5 extracts the information coefficient from the within subjects regression of model and information. The magnitude of the coefficient on the variable gives a t-statistic for MAE of 0.24 and 0.17 for NDCG, i.e., the effect of adding soft information on MAE and NDCG is likely insignificant compared to excluding these predictors. For AUC, the coefficient is -0.0069; the t-statistic is -3.45, suggesting a small negative effect to adding text-based information predictors on AUC compared to excluding it.

Table 5: Rolling window: effects of inclusion or exclusion of soft information on performance

metric	Estimate	Std. Error	t value
MAE	0.0207	0.0875	0.2365
NDCG	0.0025	0.0146	0.1712
AUC	-0.0069	0.0020	-3.4598

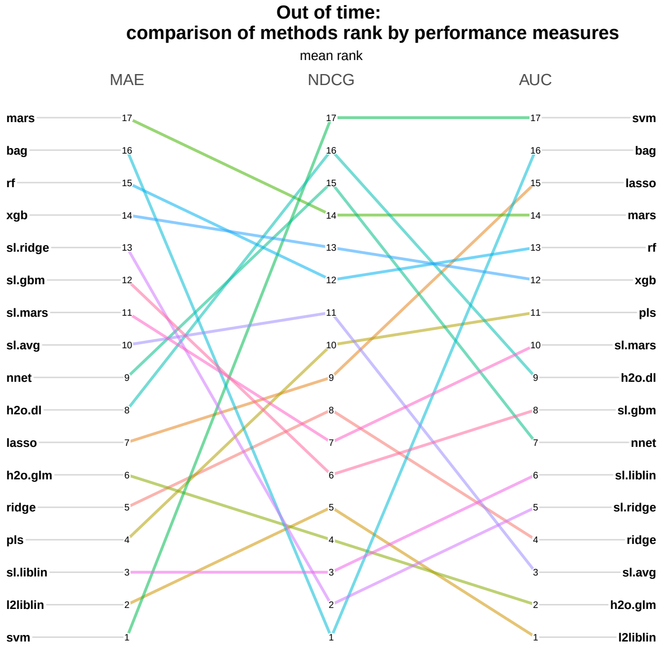

Figure 4: Performance ranked over metrics: out of time test

6.2. Out of time

The out-of-time setting is a sterner test of each method's predictive ability, in which we expect some deterioration in predictive performance. This is because, at a minimum, at least 37 months have elapsed between the origination dates in the training and test samples (see Figure 2). This setting may be more informative to investors who would only use data on closed loans to build their predictive models.

Figure 4 and Table 6 indicate that performance is more variable compared to the rolling window. In the out-of-time setting, individual models (l2liblin, h2o.glm) and stacked liblin are the best performers averaged over the three measures. For the MAE performance measure, svm, l2liblin, and stacked liblin are the top three; for AUC, l2liblin, h2o.glm, and average of stacked models perform best. Finally for NDCG, bagged trees, stacked ridge, and stacked liblin are the top three. In some instances, there is a substantial variability in performance. Just as in the rolling window results reported earlier, SVM performs well on one of the three measures (MAE) but poorly on the NDCG and AUC criteria.

Table 6: Out of time: mean performance by metric

linear or non-linear	ensemble or individual	method	MAE	NDCG	AUC
linear	individual	h2o.glm	13.79	0.77	0.64
linear	individual	ridge	13.69	0.76	0.63
linear	individual	lasso	14.10	0.76	0.58
linear	individual	svm	9.11	0.71	0.56
linear	individual	pls	13.67	0.75	0.62
linear	individual	l2liblin	9.14	0.77	0.64
non-linear	individual	mars	27.28	0.73	0.59
non-linear	individual	nnet	14.62	0.73	0.63
non-linear	individual	h2o.dl	14.28	0.72	0.63
non-linear	ensemble	rf	16.96	0.75	0.60
non-linear	ensemble	bag	21.85	0.80	0.58
non-linear	ensemble	xgb	16.36	0.74	0.61
linear	ensemble	sl.avg	15.04	0.75	0.64
linear	ensemble	sl.liblin	9.23	0.78	0.63
linear	ensemble	sl.ridge	15.80	0.78	0.63
non-linear	ensemble	sl.mars	15.11	0.76	0.62
non-linear	ensemble	sl.gbm	15.19	0.76	0.63

6.2.1. Robust LMM: out of time

This sub-section reports the results of the robust LMM applied in the out-of-time setting. The results in Table 7 indicate performance differences across measures between linear and non-linear methods for MAE, NDCG, and AUC. For the coefficient of the factor *lin.nonlin*, the t-statistics are -7.37, 2.49, 4.79, respectively. When the MAE is used as a performance measure, the average reduction in MAE from using linear methods instead of non-linear methods is -2.94, other factors unchanged. When NDCG and AUC are used, linear methods are associated with a modest performance gain. Overall, as they did in the moving window, linear methods improve performance compared to non-linear methods in the out of time setting.

The t-statistics also suggest differences between ensemble and individual methods for the MAE and NDCG performance measures (t-statistics of -4.36 and -4.29, respectively), with ensemble methods outperforming individual methods on NDCG (i.e., individual methods perform worse) and the opposite for MAE. There are no detectable differences for the AUC. Finally, as shown in Table 8, including soft information reduces MAE (a low t-statistic of -1.41), improves NDCG (t-statistic = 2.79), and has no apparent effect for AUC. This means that in this out-of-time test, when using the NDCG metric, a performance gain is observed; in a regression (MAE) and binary prediction setting (AUC), there is essentially no difference in performance.

Table 7: Out of time: effects of linearity/non-linearity and ensemble/individual factors on performance

	MAE	NDCG	AUC
Intercept	16.7943	0.7622	0.6123
	(0.3513)	(0.0158)	(0.0052)
lin.non = linear	−2.9468	0.0137	0.0139
	(0.4000)	(0.0055)	(0.0029)
ensemble = individual	−1.7671	−0.0240	0.0014
	(0.4057)	(0.0056)	(0.0030)
Num. obs.	170	170	170

Standard errors in parentheses

Table 8: Out of time: effects of inclusion or exclusion of soft information on performance

metric	Estimate	Std. Error	t value
MAE	-0.5561	0.3951	-1.4074
NDCG	0.0452	0.0162	2.7977
AUC	0.0021	0.0054	0.3941

7. Robustness checks and discussion

7.1. Robustness checks

Several robustness checks have been carried out. The first is a consistency check on the moving window and out-of-time results by rank-transforming the dependent variable in the robust LMM to check that any non-normality in the residuals does not lead to invalid inference.

The results for this alternative test for the two factors linear/non-linear and ensemble/individual are shown in appendix (see Table 11 and Table 13, for the moving window and out-of-time setting, respectively). Comparing these results with Table 4 and Table 7 leads to similar conclusions, with the exception of NDCG in the out of sample set-up which now has a small t-statistic. A similar comparison of Tables 5 and 8 with Tables 12 and 14 suggest that the effect of soft information is broadly similar across settings and performance measures.

The second set of checks involves how the input data are represented. Following a suggestion from one of the reviewers, we investigated alternatives for handling text using transfer-learning based language models and the use of categorical embeddings for some of the categorical variables. The language models are BERT, ELMO, and USE described in Devlin et al. (2018), Peters et al. (2018), Cer et al. (2018), respectively. As these models are trained on a vast corpus of text and incorporate contextual similarity, they can overcome drawbacks of standard word embeddings and

Biterm topic models. Using each of these three approaches, we extracted the word embedding matrix, and post-processed by applying Principal Components Analysis of the embedding matrix to reduce dimensionality further. These principal components were incorporated as features, and the analysis re-run. Overall, this did not change the conclusions found using the topic model.

Regarding categorical embeddings, three of the features with the highest cardinality (Fico grade, Lending Club sub-grade, loan purpose) were pre-processed using categorical embedding techniques (Guo and Berkhahn, 2016). This approach may help some of the machine learning methods such as deep learning to perform better. Replacing the categorical features with these representations for all of the models did not change the results materially. However, this does not rule out that in other datasets with higher cardinality features, this representation of features may improve predictive performance.

A third analysis considers the extent to which, in the out-of-time set, superior performance with regards to an evaluation metric is also linked to greater returns. As an example, each method's excess returns are calculated by selecting the top 100 most attractive loans based on that method's predictions and comparing their average IRR return against the mean return rate in the whole test set.[10] We then rank the methods from largest excess returns (rank 1) to smallest excess returns (rank 17). This IRR rank can now be compared against the same method's performance rank according to MAE, NDCG or AUC (lower rank numbers again indicating better performance).

For each method, Figure 5 plots the rank based on the performance measure against the return-based measure (IRR rank), for each individual performance measure as well as for the mean rank over the three performance measures. The most appealing methods are those that perform consistently well on both criteria (both have a low numerical rank) and thus appear on the lower left-hand side of each panel. The bottom-right corner is where good performance on the metric does not correspond to good performance on IRR rank.

A large difference between performance measure and IRR rank suggests inconsistent performance; i.e., in those cases, better/worse performance on the evaluation metrics may not translate to larger/smaller excess returns, relative to the other methods. For example, in the figure, linear regularised methods (h2o.glm, stacked liblin) have a reasonably good ranking compared to svm

[10] This example uses 100 loans to compare with the literature and because the minimum buy-in for retail investors on LC's platform is $25. This means in practice that investment in a large numbers of loans (e.g., a 1000) require substantial resources.

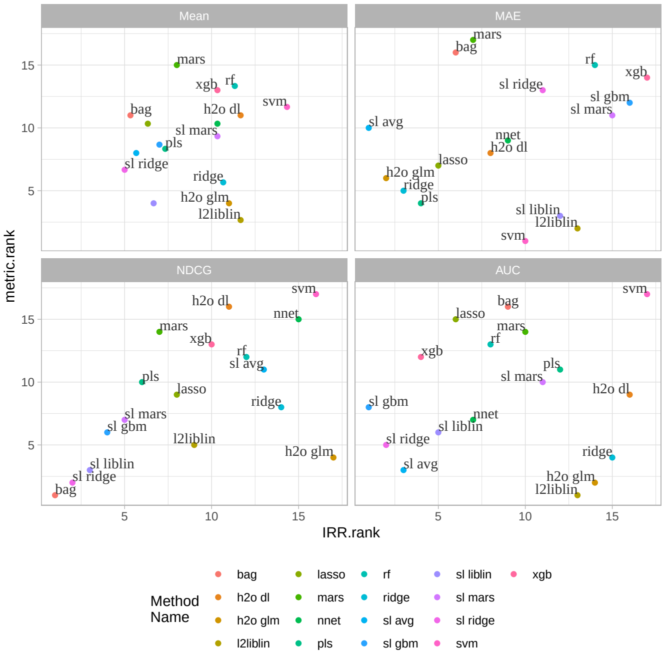

Figure 5: Out-of-time difference in rank performance (metric vs. excess returns)

for the mean ranks across performance measures. The figure illustrates the point that a method that minimises MAE or maximises AUC/NDCG for all loans does not necessarily correspond to an investment strategy that delivers excess returns (over the test mean) of the top 100 loans.

As real-world investors may be interested in those methods and strategies linked to greater excess returns, we examine the relationship between excess returns and two sets of variables – the experimental factors (i.e. *lin.nonlin* and *ensemble*) and the choice of dependent variable and tuning strategy (i.e. whether we build models to predict IRR, rank-transformed IRR or default Y/N, using MAE, NDCG and AUC as respective training metrics). To do so, we again estimated a robust linear mixed model with a random effect for inclusion of soft information.[11] The results are shown in the first column (‘ALL’) of Table 9. The respective reference categories for the factors *lin.nonlin* and *ensemble* are non-linear and ensemble; the reference category for metric is MAE. Next to these pooled results, the other three columns in the table assess the impact of the experimental factors separately for each choice of dependent variable and corresponding tuning metric.

[11] Information is treated as a within-factor with two levels: *both* and *hard.only*.

The results for the column ALL produce a t-statistic of 0.77 for the difference between linear and non-linear methods, so there is little evidence to suggest either leads to higher excess returns or the top 100 loans. However, looking at the breakdown for the individual performance metrics (columns 2-4), this result is largely down to linear methods performing worse for AUC and NDCG; for the IRR models trained with MAE, the coefficient is 0.84, with a t-statistic of 2.5, whilst for AUC and NDCG, the signs are reversed. This means linear methods are associated with larger excess returns when loans are ranked using the methods trained with the MAE as the performance criterion. Overall, individual methods are associated with lower excess returns compared to ensemble methods (coefficient = -0.28) but do not have a large t-statistic (t-statistic =-1.42). This means that there is no systematic difference in terms of excess returns from using ensemble or individual models when pooling all three performance metrics (MAE, NDCG, AUC) together (column ALL).

Another important finding relates to the choice of dependent variable and tuning strategy: relative to MAE, NDCG and AUC are associated with significantly reduced excess returns (t-statistics = -6.95; -9.0). In other words, the best modelling strategy from an average profit perspective is to predict IRR directly, with MAE as the tuning metric. Controlling for the other factors, this strategy produces larger excess returns than focusing on the IRR ranking (NDCG) or the traditional scoring approach of picking loans based on the risk of default (AUC). In terms of relative magnitude, this effect outweighs that of the two other experimental factors. In this setting, how to model the dependent variable (IRR, default class, or a ranking) and the tuning metric matters more than the type of learning method.

7.2. Discussion

The variable nature of performance across the three evaluation measures and the test setting used, and its non-trivial relationship with returns suggest that findings recommending specific methods in the existing profit scoring literature may not be generalised easily. The results may be dependent on these factors, in addition to the usual considerations such as the application domain and data used.

In this paper, a successful profit scoring approach is associated with positive excess returns (Table 9). The findings suggest that while it pays to model using profitability directly (i.e. using IRR as the dependent variable), performance depends on the methods adopted, the performance measure itself, and the type of information used. This first finding is in line with Serrano-Cinca and

Table 9: Effects of linearity/non-linearity and ensemble/individual factors on excess returns in out of time setting

	ALL	MAE	NDCG	AUC
Intercept	1.6897	0.8540	0.2228	0.1525
	(0.2415)	(0.2940)	(0.2699)	(0.4128)
lin.non = linear	0.1508	0.8390	−0.0797	−0.3067
	(0.1964)	(0.3347)	(0.3073)	(0.3643)
ensemble = individual	−0.2832	0.4978	−0.3883	−0.9453
	(0.1991)	(0.3395)	(0.3116)	(0.3695)
metric = NDCG	−1.6450			
	(0.2366)			
metric = AUC	−2.1283			
	(0.2366)			
Num. obs.	510	170	170	170

Standard errors in parentheses

Gutiérrez-Nieto (2016) for P2P lending and other profit scoring literature (Garrido et al. (2018), Verbraken et al. (2014)).

Given the range of methods, and prediction problems, it is not straightforward to identify one set of reasons or features that is associated with better predictive performance. Each of the individual models represent the data in different ways, depending on the type of prediction (continuous, ranking, or binary) and performance measure. For MAE, some models like ridge regression identify credit risk focused variables like utilisation, balance, unemployment and debt to income ratios, as well as categorical information like Fico score and LC sub-grade; linear SVMs for MAE identify grades as some of the most important factors, whereas using AUC as the performance measure, they identify a range of very different factors.

Our ability to produce positive excess returns suggests that the predictors may contain information not directly incorporated into Lending Club's grading system during the sample period in this paper. Studying a different research question, Jagtiani and Lemieux (2018) come to similar conclusions for a sample period covering much of the same period as in this paper.

The implications for platform pricing and investing are more nuanced. The information and methods in this study are public and the returns are ex-post, based on closed loans from a specific period. Therefore, one cannot be overly optimistic about excess returns in future. Platforms like Lending Club do not bear the credit risk; their main income comes from receiving a small fraction of the monthly repayments on all loans. Adjustments to pricing/grading models are one of several considerations for this type of business model, in addition to platform growth due to the supply of new listings. These types of questions could be explored using models that explicitly characterise concept drift.[12]

The negligible to negative impact of soft information may give pause for thought. There may be limitations in this study in the sense that text data has been represented through using a type of topic model adapted for short text. In an earlier version of this paper, we represented the text as certain features such as the fraction of complex words and measures of lexical diversity, as well as using Word2Vec word-embedding approach (Mikolov et al., 2013), and obtained similar results. Using more advanced language models, following a suggestion by one of the reviewers, did not change the research question results. Overall, it is likely that the result is negative in this case because the listing text was sparse. The finding that these features do not help improve predictive performance for profit scoring contrasts with other studies using data from the Prosper platform, where the text is richer. Instead, our results concur with Dorfleitner et al. (2016) who modelled default in German P2P loans. This provides further indication that results in the literature could be platform-dependent. Where richer texts are available, the new generation of language models may improve predictive performance.

8. Conclusions

This study explored three research questions motivated by a P2P investment setting. First, we compared whether non-linear methods could provide improved profitability predictions compared with linear methods. Second, drawing on findings in Lessmann et al. (2015), we investigated whether ensemble methods gave better performance than individual methods. Third, as new types of data including soft information in the form of text become available through these platforms, we also assessed their relevance for P2P investment decisions.

In our experiments, we find empirical evidence supporting a profit-scoring approach instead of modelling default risk. Specifically, we find that linear methods were actually often associated with improved predictive performance, although the magnitude of the effect varied with the performance measure. For example, linear methods produced greater excess returns on the top-100 loans than non-linear methods, provided that they are used to predict IRR and using the MAE as the training

[12] We are grateful to a reviewer for this point.

metric. Ensemble methods outperformed individual methods on some metrics (e.g. NDCG but not MAE). We did not find significantly better performance by including soft information in the predictor set.

The results add to the findings on P2P lending, and specifically contribute to the empirical assessment of P2P profit scoring. Considering the research findings, the results suggest that relatively straightforward approaches such as tuning on MAE and linear models provide good performance as well as potentially positive out-of-time excess returns, at least for this sample period. A binary classification approach that models default and uses a performance criterion such as AUC results in some excess returns out-of-time, though not as much as using the MAE. Using a ranking performance measure such as NDCG is a reasonable approach on paper but results in lower excess returns on average than using the MAE or AUC.

A relatively consistent result regardless of the performance criterion is that the inclusion of soft information either makes little difference or makes the model perform slightly worse than when trained with only hard information. However, incorporating unstructured data from text and other sources and its utilisation in predictive modelling contexts is an evolving area of research and other representations could provide better predictive ability. Specifically, soft information from P2P platforms with more abundant sources of text-based information could be incorporated using other methods than those considered in this paper. Finally, alternative sources of information such as digital footprint information could be explored further for predictive modelling of P2P loans (Berg et al., 2019).

References

Aiello, S., Eckstrand, E., A. Fu, A., Landry, M., and Aboyoun, P. (2019). *h2o: R Interface for H2O*. R package version 3.26.0.11.

Anagnostopoulos, C. and Hand, D. (2012). *hmeasure: The H-measure and other scalar classification performance metrics*. R package version 1.0.

Balyuk, T. and Davydenko, S. A. (2018). Reintermediation in Fintech: Evidence from Online Lending. *SSRN*, pages 1–54.

Berg, T., Burg, V., Gombović, A., and Puri, M. (2019). On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *Review of Financial Studies*, 33(7):2845–2897.

Bianchi, D., Büchner, M., and Tamoni, A. (2018). Bond Risk Premia with Machine Learning. *SSRN*, pages 1–84.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z. M. (2016). mlr: Machine learning in R. *Journal of Machine Learning Research*, 17(170):1–5.

Brealy, R. A. and Myers, S. C. (2001). *Principles of Corporate Finance*. McGraw-Hill Press, 3 edition.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., and Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking and Finance*, 72(C):218–239.

Candel, A., LeDell, E., Arora, A., and Parmar, V. (2020). *Deep Learning with H2O*.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal Sentence Encoder. *arxiv*, pages 1–7.

Chen, T. and Guestrin, C. (2016). XGBoost. In *The 22nd ACM SIGKDD International Conference*, pages 785–794, New York, New York, USA. ACM Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arxiv*, pages 1–16.

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., and Kammler, J. (2016). Description-text related soft information in peer-to-peer lending: evidence from two leading European platforms. *Journal of Banking and Finance*, 64(C):169–187.

Duarte, J., Siegel, S., and Young, L. (2012). Trust and Credit: The Role of Appearance in Peer-to-peer Lending. *Review of Financial Studies*, 25(8):2455–2484.

Emekter, R., Tu, Y., Jirasakuldech, B., and Lu, M. (2014). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, 47(1):54–70.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9(Aug):1871–1874.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–141.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2018). Predictably Unequal? The Effects of Machine Learning on Credit Markets. *SSRN Electronic Journal*, pages 1–73.

Garrido, F., Verbeke, W., and Bravo, C. (2018). A robust profit measure for binary classification model evaluation. *Expert Systems With Applications*, 92:154–160.

Guo, C. and Berkhahn, F. (2016). Entity embeddings of categorical variables. *arxiv*, pages 1–9.

Guo, Y., Zhou, W., Luo, C., Liu, C., and Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research*, 249(2):417–426.

Helleputte, T. (2017). *LiblineaR: Linear Predictive Models Based on the LIBLINEAR C/C++ Library.* R package version 2.10-8.

Hertzberg, A., Liberman, A., and Paravisini, D. (2016). Adverse Selection on Maturity: Evidence from Online Consumer Credit. In *Financial Innovation Online Lending to Households and Small Businesses*, pages 1–66.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–14.

Jagtiani, J. and Lemieux, C. (2017). Fintech Lending: Financial Inclusion, Risk Pricing, and Alternative Information . Technical report, Federal Reserve Bank of Philadelphia, Research, Philadelphia, PA.

Jagtiani, J. and Lemieux, C. (2018). The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the Lending Club Consumer Platform. Technical report, Federal Reserve Bank of Philadelphia, Research, Philadelphia, PA.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). Kernlab: An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11(9):1–20.

Kim, A., Yang, Y., Lessmann, S., Ma, T., Sung, M. C., and Johnson, J. E. V. (2019). Can deep learning predict risky retail investors? A case study in financial risk behavior forecasting. *European Journal of Operational Research*, pages 1–18.

Koller, M. (2016). robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. *Journal of Statistical Software*, 75(1):1–24.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

Lessmann, S. and Voß, S. (2017). Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy. *International Journal of Forecasting*, 33(4):864–877.

Liberti, J. M. and Petersen, M. A. (2017). Information: Hard and soft. Working paper, Kellogg School, Northwestern University.

Lin, M. (2016). Economic Value of Texts: Evidence from Online Debt Crowdfunding. In *Financial Innovation Online Lending to Households and Small Businesses*, pages 1–37.

Liu, T.-Y. (2011). *Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition*, volume 7. Springer, Berlin, Heidelberg.

Malekipirbazari, M. and Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems With Applications*, 42(10):4621–4631.

Mevik, B.-H. and Wehrens, R. (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, 18(1):1–23.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems NIPS*, pages 3111–3119.

Millborrow, S. (2018). *earth: Multivariate Adaptive Regression Splines*. R package version 4.6.0.

Miller, S. (2015). Information and default in consumer credit markets: Evidence from a natural experiment. *Journal of Financial Intermediation*, 24(1):45–70.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

Ridgeway, G. (2012). Generalized boosted models: a guide to the gbm package. Technical report. [Online; accessed 10-July-2014].

Serrano-Cinca, C. and Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*, 89(C):113–122.

Sirignano, J., Sadhwani, A., and Giesecke, K. (2016). Deep Learning for Mortgage Risk. *arxiv*, pages 1–83.

Stevenson, M., Mues, C., and Bravo, C. (2020). The value of text for small business default prediction: A deep learning approach. *arxiv*, pages 1–25.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288.

Vallee, B. and Zeng, Y. (2018). Marketplace Lending: A New Banking Paradigm? pages 1–60.

Verbraken, T., Bravo, C., Weber, R., and Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, 238(2):505–513.

Wright, M. N. and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++and R. *Journal of Statistical Software*, 77(1):1–17.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A Biterm Topic Model for Short Texts. In *WWW '13 Publication WWW Proceedings of the nd international conference on World Wide Web*, pages 1445–1456.

Ziegler, T., Shneor, R., Wenzlaff, K., Wanxin, B. W., Kim, J., Odorovic, A., Paes, F., Suresh, K., Zhang, B., and Johanson, D. (2020). The Global Alternative Financing Benchmarking Report . Technical report, Cambridge Center for Alternative Finance, Judge Business School, University of Cambridge, Cambridge.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Online Appendix 1: Overview of Text Features

This appendix summarises the text-based features and the approach used to fit the Biterm topic model.

Pre-processing and Summary Statistics

The text is comprised of the title of the listing text and the listing text itself. The text is a concatenation of these two text fields. The provision of this text is voluntary and this text description was discontinued on Wednesday, March 19th, 2014. Compared to other P2P platforms, the text on Lending Club is relatively short. Some summary statistics of the text information is contained in Table 10. This indicates that it is short with an average length of two sentences, and an average sentence length of six words.

The main steps in pre-processing included removing whitespace, non-ASCII characters, removing HTML tags, and other artefacts related to the platform such as "Borrower added on <Date >". This is applied to the title and description text of the listing. The title and listing text are concatenated as some borrowers provide short listing titles like "card" or "move", and longer listing text. Other borrowers do the opposite providing long titles with details that other borrowers have provided in the listing text, and short or no listing text. The following pre-processing steps are taken:

- Following merging of payment and application information, loans are selected that are issued from October 2008 - January 2014.

- Town/city and state fields are concatenated to a string and geo-coded to longitude and latitude.

- Listings with title or description texts with fewer than 4 characters in length were removed (474 loans)

- Convert numbers to words, remove punctuation, alphanumeric characters, trims strings, encode strings as UTF-8-MAC, remove any remaining non-ASCII characters, convert to lower case, remove stop words.

Table 10: Summary statistics for text features

var	min	median	mean	max	sd
number of words	1.00	4.00	20.55	819.00	38.98
number of sentences	1.00	1.00	2.18	97.00	2.26
sentence length	1.00	3.00	6.13	141.00	5.82
number frequency	0.00	0.00	0.45	11.00	1.44
complex.words	0.00	1.00	2.34	148.00	4.23

Biterm Topic Model

The short listing texts presented a challenge for feature construction or representations of the text as feature vectors as there are a limited number of words per listing. One way to deal with this is to use topic modelling. However, topic models for standard length text still face the sparsity of text within individual listings. A biterm topic model is a short-text topic model that is based on global word co-occurrence (i.e., across all listing texts) to overcome sparsity of individual documents texts (Yan et al., 2013). A biterm is an unordered pair of words from a text string. Biterms can be extracted using local word co-occurrence so that words that are within a window size are used, and words that occur outside of this window (i.e.,are too far apart), are not.

The main steps in fitting the topic model are to prepare the text input into tokens, i.e., one word per row, per listing. The key parameters are the window size (2 words), the default priors alpha and beta for the Bayesian estimation of the model (beta=0.01; alpha=50/k where k is the number of topics), and 1000 iterations of the Gibbs sampling procedure. The number of topics k was chosen by searching over 1-20 topics, recording the resulting log-likelihood as well as assessing the top 5 words within each topic for each set of iterations to ensure there were distinct topics. This resulted in 18 topics being chosen, which can then be used to obtain probabilities that a given listing has a certain topic.

For the out-of-time setting, feature generation involved using the text from the training data and scoring both the train and test data with the resulting model. For the rolling window, the tokenisation process was carried out for each iteration of the moving window of 12000 observations to ensure that only biterms present in those texts were used. This is to prevent data leakage among windows and involved fitting the biterm model separately for each slice of the moving window data. For simplicity, we kept k=18 topics as searching for different numbers of topics within each window slice would be computationally expensive and would introduce more variability within this part of the experiments.

Online Appendix 2: Supplementary Estimations

This appendix includes supplementary information on the rolling window and out-of-time experiments.

Alternative Testing Approaches for Research Questions

This section contains an alternative approach considered in exploring the research questions in which the response (i.e., the MAE, NDCG, or AUC performance measure value) was rank transformed and then used as a dependent variable in a linear mixed model. The results broadly confirm those of the main text. The exception is the inclusion of hard and soft information. This now has no detectable effect on the rank performance.

In Table 11 for MAE and AUC there are large t-statistics for the linear versus non-linear (lin.nonlin) similar to Table 4 in the main text. For ensemble, there are large t-statistics for AUC. This is similar to Table 4 in the main text. With the ranked transformation, for NDCG ensemble has a larger t-statistic than in Table 4.

The results contained in Table 12 are similar to those in Table 5 except the t-statistic for the type of information is no longer large for AUC. In Table 5 the significant effect means including soft information decreases AUC.

For the out of time setting, the results in Table 13 are similar to those in the main text in Table 7 for both linear/non-linear and ensemble/individual. The results in Table 14 are similar to Table 8 except the t-statistic on the type of information NDCG is now much lower, as well as MAE and AUC, suggesting additional text information is not important for performance.

Table 11: Rolling window: effects of linearity/non-linearity and ensemble/individual factors on performance (rank transform)

	MAE	NDCG	AUC
(Intercept)	11.6013	7.7179	8.6139
	(0.4244)	(0.4725)	(0.3355)
lin.nonlin = linear	−4.9632	−0.2587	−4.4475
	(0.4832)	(0.5379)	(0.3819)
ensemble = individual	−0.0397	2.5336	4.5629
	(0.4901)	(0.5456)	(0.3874)
Num. obs.	442	442	442

Standard errors in parentheses

Table 12: Rolling: effects of inclusion or exclusion of soft information on performance (rank transform)

metric	Estimate	Std. Error	t value
MAE	-0.0769	0.9798	-0.0785
NDCG	-0.3515	1.4950	-0.2351
AUC	0.3766	0.6158	0.6116

Table 13: Out of time: effects of linearity/non-linearity and ensemble/individual factors on performance (rank transform)

	MAE	NDCG	AUC
(Intercept)	14.5563	9.6189	11.4084
	(0.4407)	(0.7065)	(0.6355)
lin.nonlin = linear	−5.9760	−3.4869	−4.4505
	(0.5017)	(0.8044)	(0.7236)
ensemble = individual	−4.0198	2.0930	−0.3984
	(0.5088)	(0.8158)	(0.7339)
Num. obs.	170	170	170

Standard errors in parentheses

Table 14: Out of time: effects of inclusion or exclusion of soft information on performance (rank transform)

metric	Estimate	Std. Error	t value
MAE	0.4000	1.2449	0.3213
NDCG	-2.0228	2.2632	-0.8938
AUC	-0.2000	1.1414	-0.1752