It's the Economy Stupid:

Comparison of Proportional Hazards Models with Economic and Social-demographic Variables for Estimating the Purchase of Financial Products

L. Tang, L. Thomas, S. Thomas, J. Bozzetto
University of Southampton

Abstract

Relatively there is little empirical research that has been taken to understand how the underlying economy affects customers' subsequent financial product purchase behaviours. A better understanding of this influence and being able to predict the probability of purchasing are important for financial service industries. This paper undertakes an examination of the impacts of social-demographic and economic variables on the probability of purchasing financial products. In particular two most common, the Cox and Weibull, proportional hazard models are compared to examine their adequacy in terms of predictive ability. The results show that the change of external economic environment is an important source that drives customers' financial products purchasing behaviours. Furthermore, the results also that indicate Cox proportional hazard models are superior to Weibull proportional hazard models.

1. Introduction

This work considers survival analysis as a way of studying customers' financial policy purchase. It changes the objective from the traditional marketing focus of whether or not customers will purchase to estimating how long customers will wait before their next purchase (Thomas *et. al.*, 2003). This change in emphasis is also seen in consumer credit risk analysis as lenders move from default scoring (who will default) to profit scoring, which requires estimates of how long before consumers will default. The advantages of applying such analysis are impressive. Firstly, survival analysis leads to useful insights on the full span of customers' financial policy purchasing and usage processes. Secondly survival analysis can be used to estimate how long customers are likely to wait until their next purchase and what economic conditions or other observable characteristics (e.g., customers' age, financial status, and so on) affect the duration of the wait. Such information is of special interest to financial institutions in their customer relationship management modelling where the products involved, such as investment, life insurance and pension savings, can have life times which are approaching the life-times of the customer and hence will be strongly affected by the changes in economic conditions over such time scales.

The literature on the dynamics involved in the purchase and usage of financial products is quite limited. In the case of usage, research has tended to concentrate on discriminating between users and non-users of credit cards (Lindley *et. al*., 1989), (Crook 1999), (White 1975), (Carow and Staten 1999) or on predicting the amount purchased (Volker 1982), (Hirschman 1982), (Banasik *et. al.,* 2001). As to purchases, Till (Till *et. al*., 2001) investigated the number of transactions and the time between transactions using a store card. He suggested the former could be modelled by a negative binomial variable and the latter as a Weibull distribution. Andreeva (2004) used Cox proportional hazards model to look at the times between purchases of a credit card targeted at substantial purchases in three European countries. Ansell *et. al.* (2001) examined the purchasing behaviour of UK insurance company customers using a proportional hazards model to aid marketing decisions. It concentrated on the age and financial sophistication of the customer. Van der Poel (2003) looked at a similar approach but using Belgium data. None of these models though considered the impact of the state of the economy on purchases decisions and it was only very recently (Thomas *et. al.*, 2003) this has been considered. The models in that work were all proportional hazards versions of survival analysis.

There are two approaches to survival analysis for heterogeneous populations: parametric models which include both accelerated failure time (AFT) models and proportional hazard models and the non-parametric Cox proportional hazard approach (Cox 1972). In both the AFT parametric and the PH parametric approaches the Weibull distribution is the most commonly used – perhaps because it is the most general distribution that appears to satisfy both the AFT and the PH assumptions. On the other hand, though Cox's proportional hazard regression is distribution free, it does require the stronger assumption that the hazard rates for different individuals are proportional to one another over all time. This means that the same people are the most likely to have the event of interest occurring to them at all times. Relatively little empirical work appears to have been done on comparing these two approaches in the application of financial policy purchase studies. The primary purpose of this study is to estimate the impact of economic resources on financial products purchases and in particular to evaluate the performances of the two approaches from the perspective of predictive accuracy.

This study is organised as follows. In section 2, we discuss the fundamentals of our analysis, recalling the definitions of proportional hazards and accelerated life models, and indicating why we need time-dependent variables in order to estimate the probability of purchasing the next financial policy. Following that we present the model validation procedures. In section 3, we present the description of the dataset and variables to be used in our analysis. The parameter estimation in the parametric and non-parametric models, their comparison and a discussion of the results are presented in section 4. A summary of the study is given in section 5.

2. Models explanation, estimation and validation

Initially in mortality and reliability, but more recently in marketing and credit risk analysis, survival analysis – the ways of measuring the duration of a life time and leading to insights on the full span of the history process of interest – has become an important modelling methodology. In this paper we concentrate on the time between financial product purchases as the duration of interest, the "survival" of next purchasing or the length of purchase waiting. The survival analysis is mainly based on two essential concepts: the survival function and the hazard rate. The survival function, $S(t)$, gives the probability that the time until the next purchase, a random variable T, is greater than t:

$$S(t) = p(T \geq t)$$

The hazard rate, $h(t)$, captures the instantaneous rate at which duration or waiting ends in the interval $[t, t+\Delta t]$, given that the next purchase has not happened by time t. The hazard function is then defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{p(t < T \leq t + \Delta t \mid T \geq t)}{\Delta t}$$

As well as being very flexible, the hazard rate allows one to introduce explanatory variables to control the heterogeneity of the population. Here we are particularly interested in how the heterogeneity of the population and the environment may affect the customers' time until their next purchase. This heterogeneity is described by two type sets of characteristics $x_1, x_2, \ldots, x_n$ and $y_1(s), y_2(s), \ldots, y_m(s)$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is a vector of social-demographic characteristics describing the static characteristics (which will mainly be the socio-demographic information about the individual) and $\mathbf{y(s)} = (y_1(s), y_2(s), \ldots, y_m(s))$ is the vector of external economic condition variables at time s.

In proportional hazards the basic assumption is that if it is t since the last purchase which happened at time s, then the hazard rate if the individual and environment are given by x, y(s), is

$$h^{s}(t) = h_0(t)\exp(\boldsymbol{\alpha}'\mathbf{x} + \boldsymbol{\beta}'\mathbf{y}(\mathbf{s}+\mathbf{t}))$$

where $h_0(t)$ terms represents the baseline hazard, i.e., the propensity of a purchase event occurring when all independent variables equal zero. α and β denote the coefficients associated with the variables. There are two approaches in defining $h_0(.)$, α, and β - the parametric and the non-parametric (or at least the semi-parametric). In the parametric approach the lifetime distribution and hence the hazard rate is chosen to be of a certain form; e.g. if exponential then $h_0(t) = \lambda$, if Weibull with shape parameter k and scale parameter λ, $h_0(t) = k(\lambda t)^{k-1}$. The parameters of the distribution as well as α and β coefficients are then estimated from the data. The semi-parametric approach, which is more commonly used, is Cox's proportional hazard function (Cox (1972), Kalbfleisch and Prentice (1980)). In this model the coefficients of α and β can be estimated without having to assume a specific distributional form for $h_0(t)$. One can then use the Kaplan-Meier procedure to estimate $h_0(t)$.

In the purchase case considered in this paper, data is only updated monthly and so the data has a number of ties – all those customers who purchased in the same month and all those who stopped being customers in that month (and so their data was censored at that time) will have the same time given. In fact the purchase rate was so small in each month it is more appropriate in this case to take quarters – three monthly intervals – as the basic time unit. We do not distinguish when in a quarter a purchase was made and so the number of ties increases significantly and reinforces the need to use a robust but computationally tractable estimator. The standard log-likelihood estimator (Cox 1972) has difficulty dealing with the amount of computation involved with a large data set which has a lot of ties and so Breslow (Breslow 1974) and Efron (Efron 1972) suggested approximations which speeded up the calculations. Allison (1995) suggests that for data with a large number of ties, the Efron approximation gives the best result for a reasonable amount of computation.

In the parametric proportional hazards case, the most common distributional form used is the Weibull distribution, since it allows increasing, decreasing and constant hazard rates, depending on the shape parameter k. Collett (2000) discusses other distributional forms for the hazard rates, such as the log-normal, gamma, and Gompertz distributions

It is normally reported that the Weibull distribution is an example of both a proportional hazard model and an accelerated life model. This follows since if

$$S_0(t) = e^{(\lambda t)^k} \text{ and } h_0(t) = k\lambda^k t^{k-1}, \text{so } S(t) = S_0(e^{\alpha.x}t) = e^{(\lambda t e^{\alpha.x})^k}$$
$$\text{then } f(t) = k\lambda^k e^{k\alpha.x} t^{k-1} e^{(\lambda e^{\alpha.x} t)^k} \text{ and } h(t) = k\lambda^k e^{k\alpha.x} t^{k-1} = e^{k\alpha.x} h_0(t)$$

However, if the heterogeneity characteristics are time varying, as well as static, i.e. $\alpha x+\beta y(s+t)$, then this equivalence no longer holds. So in this paper we will consider the proportional hazards model version of the Weibull model where

$$h^s(t) = e^{k(\alpha.x+\beta.y(s+t))}(\lambda^k t^{k-1}) = e^{k(\alpha.x+\beta.y(s+t))} h_0(t)$$

The accelerated life interpretation of the Weibull distribution (or of any other distribution) is lost when the variables become time dependent.

The proportional hazards requirement that the relative purchase hazards of two different customers remain the same when they face identical economic conditions holds for both the Weibull and Cox models described above. This may be too strong an assumption as the relative hazard may vary as the time since the last purchase increases. Making the coefficient depend on the time since the last purchase would relax this assumption. We could also make the coefficient β do likewise but since the corresponding variable is also changing it would be difficult to be sure it is not being affected by the time dependence of the economic variables. One way of doing this is to consider characteristics $\mathbf{x}$, $t\mathbf{x}$, $\mathbf{y}(s+t)$, so the relative hazard becomes $\alpha\mathbf{x}+\beta t\mathbf{x}+\gamma\mathbf{y}(s+t) = (\alpha+\beta t)\mathbf{x}+\gamma\mathbf{y}(s+t)$. We will as part of our analysis examine how this relaxation impacts on both the parametric and non-parametric models.

Sometimes it is difficult to discriminate between Cox and Weibull regressions (Collett, 2000). The standard errors of coefficients are normally used to compare models. If the estimated standard errors for the Weibull parametric regressions are smaller than for Cox's semi-parametric regressions, then the Weibull regression is more efficient than Cox regression. However, if the standard errors for both models are similar, Cox regression is clearly of interest as it requires fewer assumptions.

Two validation procedures are applied to compare Cox and Weibull regressions. In the first validation procedure tests the relative ranking. In this procedure, we estimated the Cox proportion hazard and Weibull models based on training data sample. Then we use the estimated coefficients to predict the financial policy purchasing probabilities for the holdout sample and rank them in likelihood of purchase. We compare these ranking with who actually made purchases over different time periods. In the case of constant coefficients with no interaction variables between the economic and socio-demographic characteristics, this ranking is independent of the economic variables since the same economic conditions apply to all the consumers.

Secondly we compare the predictions from the two models of the number of purchases in the future again using the holdout sample. These predictions depend on our view of the values of the economic variables and to reduce the errors involved we will take these predicted economic values to be the actual ones that occurred in practice. In the case of non-independent coefficients and no interactions between the economic and socio-demographic variables, the ratio of the predictions of the total number of purchases in different time periods is independent of the socio-demographic variables. Thus the prediction of the total number of purchases is essentially a function of the economic variables.

We repeat these comparisons for the models with time dependent coefficients and whereas the robustness validation procedure and the purchase prediction procedure remain the same, the relative ranking classification becomes much more subtle since the relative ranking of the consumers can now change over time. Moreover the predicted total number of purchases in each time period will depend on the socio-economic variables as well as the economic ones.

Another way of ensuring both economic and socio-demographic variables play a part in both the relative propensity to purchase of the customers and the expected total number of purchases in each period is to allow interaction variables between the two groups. For example if we believe that the unemployment rate is a big factor for the middle aged (35-55 say) but less so for others, then define the variable unempl(35-55) which, if the unemployment rate is y(t) at time t, would take the value y(t) for those in that age group and 0 for those in the other age groups.

3. Data description

The dataset used in this study is provided by an international insurance company based in the UK. It covers the purchase, payment and termination history of just under 50,000 customers (24,797 male and 24,977 female), who used the direct sales channel. This history was available from January 1999 until July 2003. The advantage of this data is the detail record of accurate information for every customer financial product purchasing history. The information on customers included their gender, age, and Financial Acorn category which described their financial status. The information on their previous purchase included whether the purchase was one involving just a single payment or whether there were monthly or annual instalments and for all policies purchased one had the policy start date and the policy end date.

The outcomes of both the age and financial acorn variables were split into sets using the coarse classifying approach for survival analysis outlined in Stepanova and Thomas (2002). This involved splitting the answers into a fine classification (every 5 years for age, every category for Financial Acorn) and using a binary variable to describe inclusion in this set. A proportional hazards model is then built just using the binary variable for each finely classified category of the original variable and the coefficients of these binary variables in the proportional hazards model examined. Adjacent categories with similar coefficients are then combined into coarse classes. In this way one can allow for possible non-linearities in the relationship between the independent variables like age and the probability of purchasing. In this case this led us to split age into four groups – under 20; 20 to 35; 35 to 55; and over 55- where we have a binary variable for each of the last three categories while the first was the reference age group. For financial Acorn, a similar analysis split it into three categories into A, B, and C or D, where we use B as our reference group.

In addition to the variables that are recorded in the dataset, five external economic variables are also included in the analysis, since purchasing decisions made by customers may be influenced by external economic environment conditions. It should be noted that these economic variables are exogenous. Traditional consumer demand analysis focuses on relative goods prices and income, while saving models include variables such as interest rates, wealth, personal income and consumer sentiment. Here we chose variables to reflect the attractiveness of financial investments and the general economic and investment climate. The external UK economic variables considered are Consumer Prices, Consumer Confidence Index, Unemployment Rate, FTSE All Share Index, and Bank of England Base Interest Rate. Transformations of these variables are considered in order to conform with the macro economic literature, to avoid the problems of non-stationary time series and to have variables that relate to the way consumers perceive the economic conditions. We also looked at which variant of an economic variable is chosen by the proportional hazards model when it can only use data on that variable. In the light of this we chose the following variants

- Consumer Prices: The yearly difference of the consumer prices is used as this is representative of a price inflation that a consumer experience. Higher inflation may be considered to have a negative effect on buying savings' products
- Confidence Index: The quarterly index level is used because this is a stationary process representing the difference between those who are more or less confident about the future of the economy. Throughout this period this variable is negative and this must be remembered when considering the effect

of its coefficient. More confident customers would be expected to buy more financial products.

- Unemployment rate: The yearly difference in the unemployment rate is used as it represents the increase or decrease in jobs for consumers. It will also give information regarding the business cycle in addition to that given by the Confidence Index.
- Stock return: The impact of the stock market is measured by its return which is defined as the quarterly difference in the natural logarithm of the FTSE100 index. A buoyant stock market may encourage customers with a greater tendency to buy financial products.
- Interest Rate: The rate level at the start of the quarter is used in the model. This usually impacts consumers through the effect on the mortgage repayment rate, and hence affect disposable income available for savings. It also reflects the opportunity cost of switching from a bank deposit into a financial product.

We randomly split the whole dataset sample into training sample and holdout sample. The sample size for training sample is 39,820 customers, with 3,742 customers making further purchases during the period. The size of the holdout dataset sample is 9,954 customers of whom 935 made further financial policy purchases. We use the training data to estimate the maximum likelihood estimation (MLE) of relevant coefficients and use these estimated coefficients to predict the purchasing probabilities for customers in the holdout data sample.

4. Analysis and results

We begin by estimating several Cox and Weibull models with different combination of social-demographic and economic variables to identify the determinants of the probability of purchasing. Customers with higher confidence of current or future economy and higher stock market returns are expected to associate with higher purchasing hazards, while the rise of consumer prices or interest rates could damp customers purchasing hazards because they raise the opportunity cost. High unemployment rates are also expected to have negative impacts on customers' willingness to buy further financial products. We then complete the analysis by validating both models in terms of their predicting abilities.

The results are presented first for the comparison of the basic Cox and Weibull models (hereafter referred to as the vanilla models). Then the models allowing interactions between economic and socio-demographic variables are considered and finally the changes brought about by time dependent coefficients are considered.

Table 1 reports the estimates of the coefficients of the variables in both the Cox and Weibull vanilla models. All five economic variables are significantly different from zero (at the 5% level of confidence) so the economy plays a significant effect on purchasing behaviour. Age, gender and method of payment are the important socio-demographic and purchase characteristics. A positive coefficient on a variable means that as its value increases the hazard rate increases and so the customer is more likely to purchase. Thus in both models, all economic variables have the expected signs so that as stock returns and the confidence index goes up the probability of purchasing goes up. As unemployment, interest rates and consumer prices go up then in both models the propensity for purchasing goes down. In both models, males are significantly more likely to make a repeat purchase than females and those aged over 55 are more likely to purchase than those aged 35-55 who in turn are more likely to purchase than the under 20s. The 20-35 year old group has a different effect in the two models. In the Cox model they are as likely to purchase as the 55+ while in the Weibull model they are less likely to purchase than the 35-55 year old group. Similarly the effect of monthly payments changes between the two models but is significant in both.

Figure 1 shows the baseline hazard rate for the two models (in the Cox case this is got by using the Kaplan-Meier approach). Both show a decreasing long run likelihood to purchase as the time since the last purchase increases- not a startlingly result. What is surprising though is the sharp rise in the propensity to purchase 4 quarters after the last purchase shown in the Cox model, thus capturing the annual effect of customer purchasing behaviours. The Weibull model by definition is forced to smooth this non-monotonic behaviour away. This shows the flexibility of the Cox model compared with the robustness of the Weibull one and, to give the game away, is, we believe, the reason that the Cox model appears subsequently to be a better forecasting tool.

The coefficients of the models with interaction terms included are given in Table 2. In both models all five economic variables are significant (interest rate enters the Weibull model through its interaction terms) as are gender, age and payment pattern. Again the last is the only one which has a different effect in the two models. Looking at the significant interaction terms one sees that the Financial Acorn A group customers are less affected by rises in consumer prices than others and those aged over 55 in this group are not affected at all. Consumer confidence does seem to affect the age groups in different ways, with it having least effect on the 20-35 year olds but increasing impact on the older and younger groups. As it stands it seems as if increases in unemployment lead to increases in purchases but this is only for the under 20 age group who are really not affected by unemployment. For the other age groups the interaction terms turn the effect the other way around with those aged 35 or over being the ones most affected. Similarly rises or falls in interest rate have much more effect on the over 20s than the under 20s, making them more likely to purchase if the interest rate falls. Perhaps this is partly the effect of falls in the mortgage rate and partly that investing in cash accounts look less attractive. This may be reinforced by the fact that it is the Financial Acorn A group who are most affected. Interestingly there is no significant interaction between changes in the stock market yield and the socio-demographic variables.

When time dependent coefficients are used on the socio-demographic variables in the vanilla models, the main impact could be described as "regression to the mean" in that with the exception of one insignificant coefficient the time effect decreases the initial impact of the coefficients. The time-dependence also brings the effect of payment frequency into agreement between the Cox and Weibull models in that in both cases monthly payments make it more likely for another purchase to be made in the first 8 quarters since the last purchase but less likely thereafter. The impact of age, with the over 20s much more likely to make repeat purchases, also decreases over time though it disappears in the Weibull model after about 10 quarters and in the Cox model after about 18 quarters. Again all the economic variables are strongly significant and their effect in both models is the same as in the vanilla cases.

Finally we look at the Cox model with time dependent coefficients on the socio-demographic variables and interaction terms between these and the economic variables. The performance of the previous Weibull models meant we did not feel that approach merited such a complex extension. Table 4 shows the coefficient estimates of this extended Cox model and it shows all the economic variables are highly significant (with interest rate entering through the interaction terms) as well as age and payment frequency. Gender is not now significant and financial status just splits into the two AB or CD groups. Again in all the significant socio-demographic variables, time diminishes the value of the coefficients with the impact of monthly payments changing sign after 9 quarters but the impact of age lasting more than 20 quarters. Again it is interesting that there is no significant interaction between stock market returns and the socio-economic variable. Rises in consumer prices decrease the likelihood of purchases but this impact decreases with age and is wiped out for the Financial Acorn A group who are over 55. Increasing consumer confidence leads to increases in the probability of purchasing but this is least marked among the 20-35 age group followed by the 35-55 aged group. The impact of unemployment seems to have the wrong sign attached to it until one looks at the interaction terms. It seems that for anyone over 20 especially the 35-55 age group falls in unemployment leads to rises in purchasing and this effect is most pronounced among the Financial Acorn A group. Lastly a fall in interest rates (since there were mainly falls during this period) leads to an increasing propensity to purchase as people get older and also this effect is again more pronounced on the wealthier Financial Acorn A group.

Among all the results that are reported in Tables 1 through 4, most estimated standard errors in Cox models are noticeably smaller than those of the corresponding Weibull models, suggesting that the Cox proportional hazard model better fits the dataset. The reason why parametric accelerated failure time estimations based on Weibull distribution for purchase waiting time perform poorer than semi-parametric estimations can be attributed as follows. Firstly, the Weibull distribution assumption for the baseline hazard could be misspecified for our dataset, thus, causing inconsistent estimation of covariate coefficients. Cox models, on the other hand, avoid assuming any particular distributional forms for the underlying baseline hazard function. Secondly, we need to incorporate external economic variables into all models and these economic variables are time-dependent variables; in other words, their values keep changing over time through the process of customer purchasing history. This means that the coefficients of these economic variables are sensitive to the underlying baseline hazard.

Turning now to the validation of the different models, firstly each model was tested on the holdout sample to predict its ability to correctly rank the likelihood of purchasing. The results are shown by a series of ROC curves. In these as the cut-off probability of purchase moves, the x-axis gives the percentage of the actual non-purchasers with predicted purchase probabilities above that value and the y-axis gives the percentage of purchasers with predicted probabilities above that value. This is a standard method of assessing the power of a scorecard in credit scoring and perfect predictors would go through the point (0, 1) while ones that correspond to random predictions would trace out the diagonal line. To give an indication of robustness over time the ROC curve is shown for 9 different time periods beginning in the top left with the prediction for purchases in the next quarter; top centre gives the results for predicting purchases in the next two quarters, i.e. over the next 6 months. The results of increasing time periods are shown ending with the 9 period predictions at the right of the last row. The "fatter" the curve the better relative ranking is being given by the model. Thus comparing Figure 2a and 2b for the vanilla Cox and Weibull models it is clear that the Cox model is much better and in fact the Weibull model performs worse than randomly as the forecast period gets larger.

For this reason in Figures 2c, 2d and 2e we show only the corresponding results for the more complex Cox models with time dependent coefficients (2c), interaction terms (2d) and both interaction terms and time dependent coefficients (2e). It is clear that all are better than the vanilla version but it is the interaction terms that make the greatest improvement in the forecasting of the relative ranking. Adding time dependent coefficients if anything makes the ranking predictions worse.

The other method of validation we use is to estimate the number of purchases in the future. To do this we take the models prediction of the probability of purchase over each quarter in the time period being considered for each customer in the holdout sample and sum up all these probabilities. We then compare this expected number of purchase with the actual numbers made during that period. If we are interested in more than one quarter ahead then to make our predictions we have to estimate what the economic variables are likely to be at the start of the subsequent quarters. In order to concentrate on the model validity rather than the economic predictions we take these estimates to be the actual values that occurred. Figure 3 shows the results for predictions over time periods ranging from 1 to 19 quarters-the total time period available. What is clear is that even the basic Cox model is much superior to the basic Weibull model which essentially gets significant errors in the long run predictions. This is where the flexibility of the Cox model to allow for the non-monotonic in the hazard rate comes into its own. The time dependence of the coefficients makes little difference in the short term but improves the long term forecasts considerably. The interaction models over-estimate the number of purchases slightly in the short term but underestimate them slightly in the long term, while the original model underestimated throughout. Again if you were forced to choose a best estimator it would be the Cox model with interaction terms but no time dependent variables.

5. Conclusions

A better understanding of customers' financial product purchasing decisions and determinants of decisions could help financial service industries. The paper has investigated the use of several survival analysis models to model and forecast the propensity for customers of a financial institution to make repeated purchases. The results show the important role external economic variables have played, along with individual-specific characteristics, in determining customers purchasing behaviours. In particular, it emerges that different customers in terms of age and financial status respond differently to changes in the economy. Thus, the economic influences should not be understated.

The paper also compared semi-parametric (Cox proportional hazards) and parametric (Weibull proportional hazards) models and it appears that the flexibility in the choice of hazard function allowed by the semi-parametric models more than outweighs the robustness of the parametric models for this data set. The paper looked at introducing both socio-demographic and economic variables into the models and pointed out that in the basic model the former essentially gives the relative ranking among the customers of their propensity to purchase while the latter gives the estimates on the total number of purchases. It looked at more complex models with time – dependent coefficients and interaction terms between the economic and socio-demographic variables and the results indicate that these models are superior to the basic model. It does seem though that it is the interaction between socio-demographic and economic variables that is most important in improving both the targeting of the customers by providing the most predictive purchase rankings and in providing the accurate forecasts of future purchases.

Acknowledgements

The authors are grateful to the Institute of Actuaries and the Engineering and Physical Sciences Research Council (EPSRC) for funding this research under the Financial Mathematics initiative.

References:

Allison P.D. (1995), Survival analysis using SAS, A Practical Guide, SAS Institute, Cary.

Andreeva G (2004), Credit Risk in the Context of European Integration: assessing the possibility of Pan European Scoring, Ph.D. thesis, University of Edinburgh.

Ansell J.I., Archibald T. Harrison T. (2001), Lifestage, lifestyle and the impact on the client propensity in financial services, Working Paper 01/3, School of Management, University of Edinburgh .

Banasik J.L., Crook J.N., Thomas L.C., (2001), Scoring by usage, J. Operational Research Society 52, 997-1006.

Breslow R.E., (1974), Covariance analysis of censored survival data, Biometrics 30, 89-99.

Carow K.A., Staten M.E. (1999), Debt, credit or cash: survey evidence on gasoline purchases, J. Economics and Business 51, 409-421

Collett D., (1994) Modelling Survival data in Medical Research, Chapman and Hall, London.

Cox D.R., (1972), Regression Models and Life Tables, J. Royal Statistical Society B, 34, 187-220

Crook J.N. (1999), Who is discouraged from applying for credit, Economics Letters 65, 165-173.

Efron B. (1977), The Efficiency of Cox's likelihood function for censored data, J. American statistical Association 76, 312-319.

Hirschman E.C. (1982), Consumer payment systems: the relationship of attribute structure to preference and usage, j. Business 55,531-545.

Kalbfleisch J.D., Prentice R.L., (1980), The Statistical analysis of failure time data, Wiley and Sons, New York.

Lindley J.T., Rudolf P., Selby E.B., (1989) Credit card possession and use; change over time, J. Economics and Business 41 127-142.

Stepanova M., Thomas L.C., (2002), Survival analysis methods for personal loan data, Operations Research 50, 277-289.

Thomas L.C., Thomas S., Tang L., Gwilym O.A., (2003), The impact of demographic and economic variables on financial policy purchase timing, Working Paper ?? School of Management University of Southampton

Till R.J. (2001), Predictive behavioural models in credit scoring and retail banking, Ph.D. thesis, Imperial College, London.

Van den Poel D., Lariviere B., (2003), Customer Attrition analysis for financial services using Proportional Hazard models, Working paper, Department of marketing, Ghent University, Ghent.

Volker P.A., (1982), A note on factors influencing the utilization of bank cards, Economic Records, September, 281-289.

White K.J., (1975) Consumer choice and the use of bank credit cards: a model and cross section results, J. Consumer Research 2, 10-18

Table 1: Maximum Likelihood Estimates of Vanilla Cox & Weibull Models

Variables	Cox PH	Weibull
Constant	---	-9.0001
	(---)	(0.23)
Male	0.0716**	0.1404**
	(0.03)	(0.07)
Financial Acorn A	0.0019	-0.0028
	(0.03)	(0.08)
Financial AcornCD	-0.0636	0.3029**
	(0.06)	(0.14)
Monthly Payment Frequency	0.1865**	-0.4268**
	(0.04)	(0.09)
Age between 20 and 35	1.0145**	0.1974**
	(0.09)	(0.20)
Age between 35 and 55	0.7770**	1.4646**
	(0.09)	(0.20)
Age above 55	0.9913**	1.9758**
	(0.09)	(0.20)
Consume Price	-0.1371**	-0.7432**
	(0.02)	(0.05)
Confidence Index	0.0876**	0.2004**
	(0.01)	(0.02)
Unemployment	-0.4517**	-2.6013**
	(0.15)	(0.33)
Stock Return	2.3279**	6.4458**
	(0.21)	(0.4623)
Interest Rate	-0.6493**	-1.1661**
	(0.02)	(0.05)
Scale	---	2.4361**
	(---)	(0.03)
Weibull shape	--	0.4105**
	(---)	(0.01)
Log Likelihood	-35451.23	-32405.11

Notes: (1) The reference category for financial Acorn variable is Financial Acorn B customers. (2) The reference category for Age variable is for customers aged below 20. (3) Those numbers in parentheses are estimated standard errors. (4) Cox PH and Weibull stand for Cox proportion hazard model and accelerated failure time with Weibull distribution model, respectively. (5) ** stands for statistically significant at 95% level.

Table 2: Maximum Likelihood Estimates of Cox & Weibull Models with Interaction Terms

Variables	Cox PH	Weibull
Constant	--- (---)	-9.3375 (0.41)**
Social Demographic		
Male	0.0740 (0.03)**	0.1421 (0.07)**
Financial Acorn A	0.0553 (0.06)	0.1901 (0.13)
Financial AcornCD	-0.2795 (0.11)**	0.0032 (0.24)
Monthly Payment Frequency	0.1948 (0.04)**	-0.4246 (0.90)**
Age between 20 and 35	1.6161 (0.19)**	2.1669 (0.40)**
Age between 35 and 55	1.4024 (0.18)**	1.7662 (0.39)**
Age above 55	1.6368 (0.18)**	2.2851 (0.40)**
External Economic		
Consume Price	-0.3207 (0.12)**	-0.9943 (0.27)**
Confidence Index	0.2391 (0.06)**	0.6326 (0.14)**
Unemployment	2.4696 (0.85)**	4.4379 (1.83)**
Stock Return	2.5792 (1.04)**	8.3014 (2.40)**
Interest Rate	-0.0806 (0.11)**	-0.1892 (0.24)
Demographic*Economic		
FinAcorn A*Consume Price	0.1227 (0.05)**	0.3426 (0.11)**
FinAcorn A*Confidence Index	-0.0111 (0.02)	-0.0078 (0.04)
FinAcorn A*Unemployment	-0.6161 (0.31)**	-0.8736 (0.69)
FinAcorn A*Stock Return	-0.3245 (0.43)	-1.4529 (0.96)
FinAcorn A*Interest Rate	-0.1227 (0.04)**	-0.2895 (0.09)**
FinAcorn CD*Consume Price	0.0775 (0.08)	0.4353 (0.18)**
FinAcornCD*ConfidenceIndex	0.0017 (0.03)	-0.0426 (0.07)
FinAcorn CD*Unemployment	-0.2461 (0.56)	0.1023 (1.22)
FinAcorn CD*Stock Return	-0.1168 (0.74)	-1.5551 (1.65)
FinAcorn CD*Interest Rate	0.0231 (0.07)	-0.1653 (0.16)
Age(20-35)*Consume Price	-0.0367 (0.13)	-0.1847 (0.28)
Age(20-35)*Confidence Index	-0.1900 (0.06)**	-0.5124 (0.14)**
Age(20-35)*Unemployment	-2.2178 (0.88)**	-4.5462 (1.90)**
Age(20-35)*Stock Return	-0.5658 (1.08)	-2.5904 (2.47)
Age(20-35)*Interest Rate	-0.4616 (0.11)**	-0.6783 (0.25)**
Age(35-55)*Consume Price	0.1536 (0.13)	0.0652 (0.27)
Age(35-55)*Confidence Index	-0.1548 (0.06)**	-0.4587 (0.14)**
Age(35-55)*Unemployment	-2.8850 (0.87)**	-7.2138 (1.88)**
Age(35-55)*Stock Return	-0.2043 (1.06)	-0.8394 (2.45)
Age(35-55)*Interest Rate	-0.5612 (0.11)	-0.9071 (0.25)**
Age(>55)*Consume Price	0.2310 (0.12)**	0.1938 (0.28)
Age(>55)*Confidence Index	-0.1057 (0.06)	-0.3447 (0.14)**
Age(>55)*Unemployment	-2.7695 (0.87)	-8.2196 (1.90)**
Age(>55)*Stock Return	0.6290 (1.08)	0.6172 (2.48)
Age(>55)*Interest Rate	-0.5529 (0.11)**	-0.9216 (0.25)**
Scale	--- (---)	2.4349 (0.03)**
Weibull shape	--- (---)	0.4107 (0.01)**
Log Likelihood	-35405.28	-32350.66

Notes: (1) The reference category for financial Acorn variable is Financial Acorn B customers. (2) The reference category for Age variable is for customers aged below 20. (3) Those numbers in parentheses are estimated standard errors. (4) Cox PH and Weibull stand for Cox proportion hazard model and accelerated failure time with Weibull distribution model, respectively. (5) ** stands for statistically significant at 95% level.

Table 3: Maximum Likelihood Estimates of Cox & Weibull Models with Time Dependent Coefficients

Variables	Cox PH	Weibull
Constant	--- (---)	-5.8772 (0.13)**
Social Demographic		
Male	0.0733 (0.05)	0.3174 (0.07)**
Financial Acorn A	-0.0401 (0.06)	0.2940 (0.07)**
Financial AcornCD	-0.1188 (0.11)	-0.0074 (0.13)
Monthly Payment Frequency	0.5562 (0.07)**	1.9621 (0.08)**
Age between 20 and 35	1.2994 (0.16)**	5.6913 (0.16)**
Age between 35 and 55	1.1319 (0.15)**	5.3534 (0.16)**
Age above 55	1.4398 (0.15)**	5.2836 (0.17)**
External Economic		
Consume Price	-0.1230 (0.02)**	-0.7050 (0.03)**
Confidence Index	0.0867 (0.10)**	0.0926 (0.01)**
Unemployment	-0.3323 (0.15)**	-3.8320 (0.18)**
Stock Return	2.2632 (0.20)**	1.6451 (0.22)**
Interest Rate	-0.6489 (0.02)**	-0.2386 (0.02)**
Time Dependent Coefficient		
Male*Time	-0.0007 (0.01)	-0.0489 (0.01)**
FinAcorn A*Time	0.0080 (0.01)	-0.0469 (0.01)**
FinAcornCD*Time	0.0080 (0.02)	-0.0557 (0.02)**
MthPayFrequency*Time	-0.0679 (0.01)**	-0.2487 (0.01)**
Age between 20 and 35*Time	-0.0556 (0.02)**	-0.5606 (0.01)**
Age between 35 and 55*Time	-0.0665 (0.02)**	-0.5485 (0.01)**
Age above 55*Time	-0.0847 (0.02)**	-0.4943 (0.02)**
Scale	--- (---)	1.2373 (0.02)**
Weibull shape	--- (---)	0.8082 (0.01)**
Log Likelihood	-35404.28	-22619.07

Notes: See Table 2 Notes.

Table 4: Maximum Likelihood Estimates of Cox & Weibull Models with Interaction Terms and Time Dependent Coefficients

Variables	Cox PH	Weibull
Constant	--- (---)	
Social Demographic		
Male	0.0758 (0.05)	
Financial Acorn A	0.0341 (0.09)	
Financial AcornCD	-0.4914 (0.16)**	
Monthly Payment Frequency	0.5645 (0.07)**	
Age between 20 and 35	1.9282 (0.24)**	
Age between 35 and 55	1.8940 (0.23)**	
Age above 55	2.3716 (0.23)**	
External Economic		
Consume Price	-0.3280 (0.12)**	
Confidence Index	0.2400 (0.06)**	
Unemployment	2.4220 (0.84)**	
Stock Return	2.8781 (1.05)**	
Interest Rate	-0.0636 (0.11)	
Time Dependent Coefficient		
Male*Time	-0.0007 (0.01)	
FinAcorn A*Time	0.0037 (0.01)	
FinAcornCD*Time	0.0283 (0.02)	
MthPayFrequency*Time	-0.0679 (0.01)**	
Age between 20 and 35*Time	-0.0647 (0.02)**	
Age between 35 and 55*Time	-0.0838 (0.02)**	
Age above 55*Time	-0.1114 (0.02)**	
Demographic*Economic		
FinAcorn A*Consume Price	0.1199 (0.05)**	
FinAcorn A*Confidence Index	-0.0112 (0.02)	
FinAcorn A*Unemployment	-0.6329 (0.32)**	
FinAcorn A*Stock Return	-0.3061 (0.43)	
FinAcorn A*Interest Rate	-0.1214 (0.04)**	
FinAcorn CD*Consume Price	0.0566 (0.08)	
FinAcornCD*ConfidenceIndex	0.0019 (0.03)	
FinAcorn CD*Unemployment	-0.3633 (0.56)	
FinAcorn CD*Stock Return	-0.0549 (0.74)	
FinAcorn CD*Interest Rate	0.0464 (0.07)	
Age(20-35)*Consume Price	-0.0418 (0.13)	
Age(20-35)*Confidence Index	-0.1910 (0.06)**	
Age(20-35)*Unemployment	-2.2095 (0.87)**	
Age(20-35)*Stock Return	-0.8750 (1.09)	
Age(20-35)*Interest Rate	-0.4521 (0.11)**	
Age(35-55)*Consume Price	0.1720 (0.12)	
Age(35-55)*Confidence Index	-0.1567 (0.06)**	
Age(35-55)*Unemployment	-2.7259 (0.86)**	
Age(35-55)*Stock Return	-0.5866 (1.07)	
Age(35-55)*Interest Rate	-0.5813 (0.11)**	
Age(>55)*Consume Price	0.2807 (0.13)**	
Age(>55)*Confidence Index	-0.1062 (0.06)	
Age(>55)*Unemployment	-2.4374 (0.87)**	
Age(>55)*Stock Return	0.1829 (1.09)	
Age(>55)*Interest Rate	-0.6071 (0.11)**	
Scale	--- (---)	
Weibull shape	--- (---)	
Log Likelihood	-35349.59	

Notes: See Table 2 Notes.

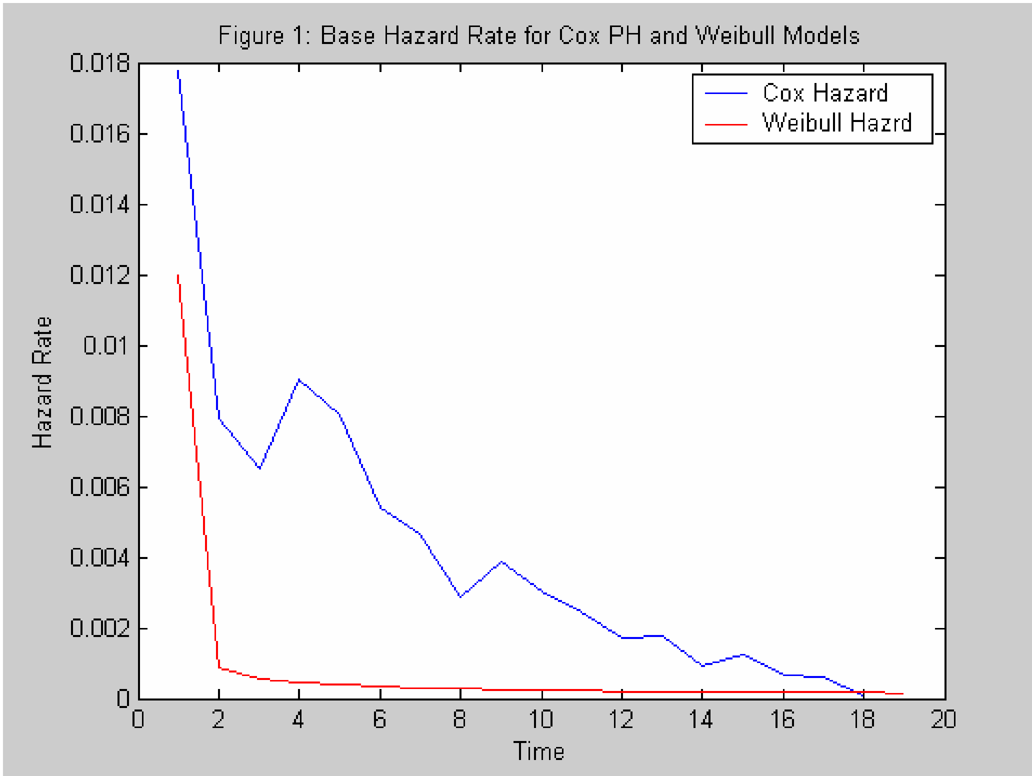
Figure 1: Base Hazard Rate for Cox PH and Weibull Models
Cox Hazard
Weibull Hazrd
Hazard Rate
Time
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
2
4
6
8
10
12
14
16
18
20

Figure 2a: ROC Curves for Cox vanilla model

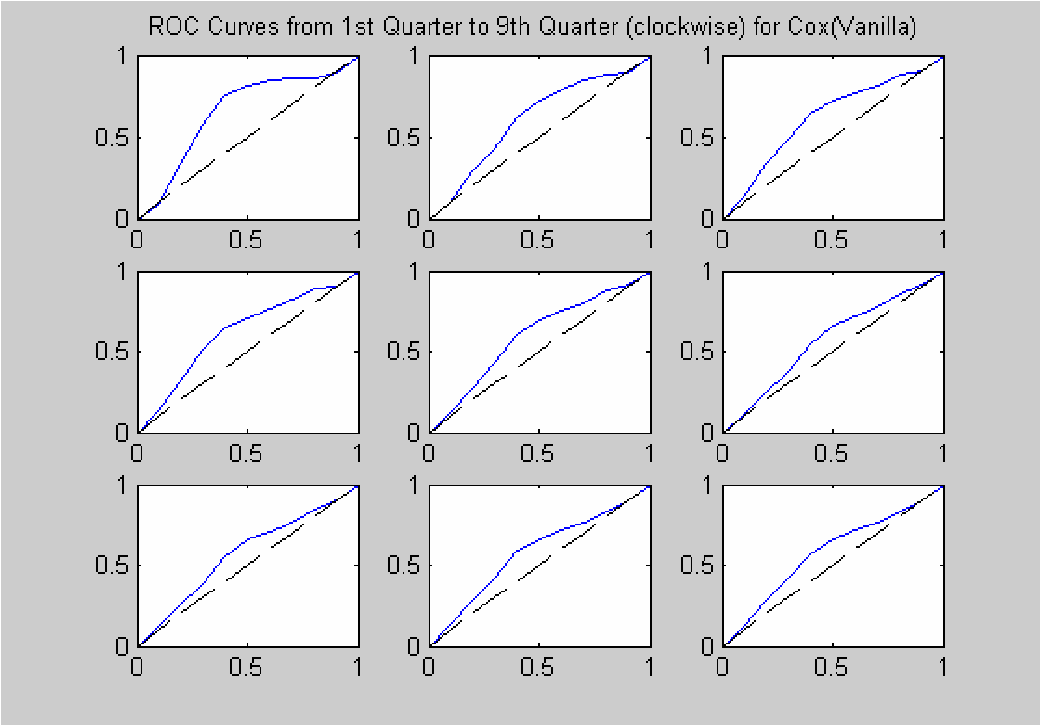

Figure 2b: ROC curve for Weibull vanilla model

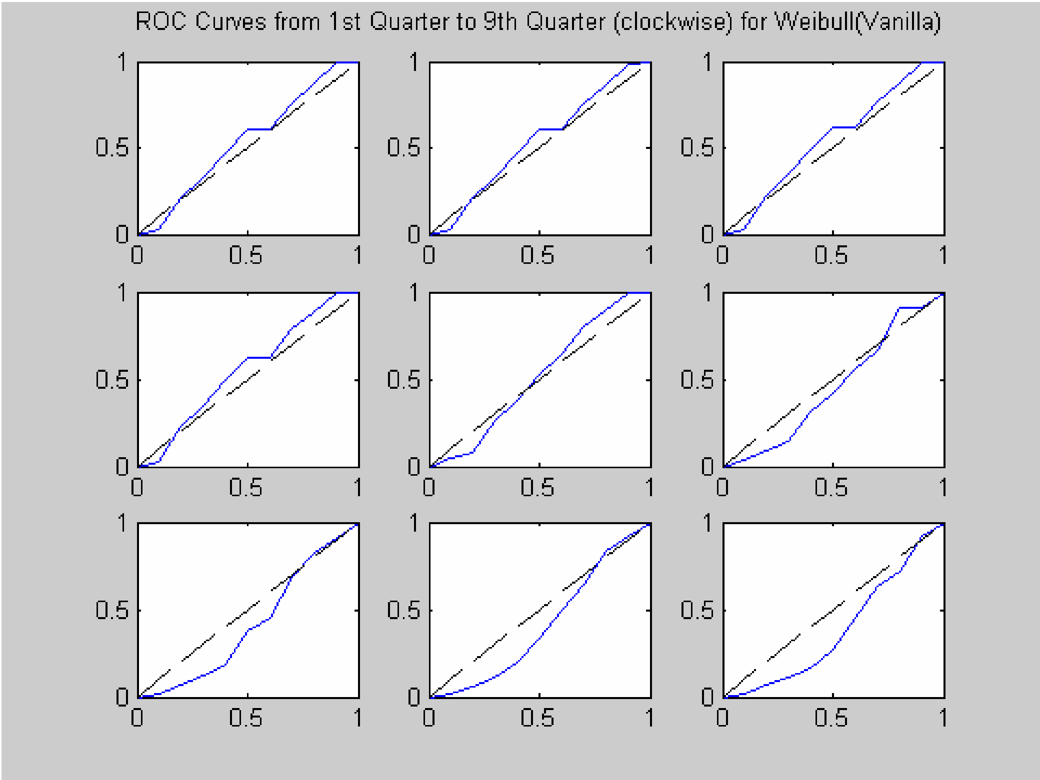

Figure 2c: ROC curve for Cox model with time dependent coefficients

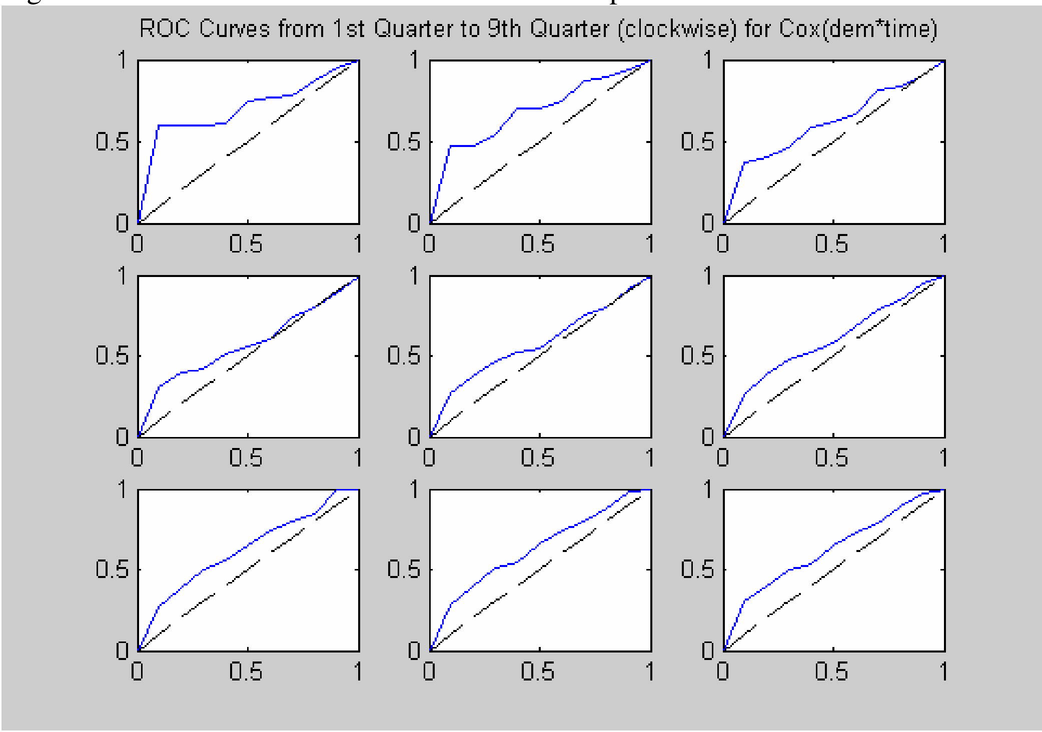

Figure 2d: ROC curve for Cox model with interaction terms

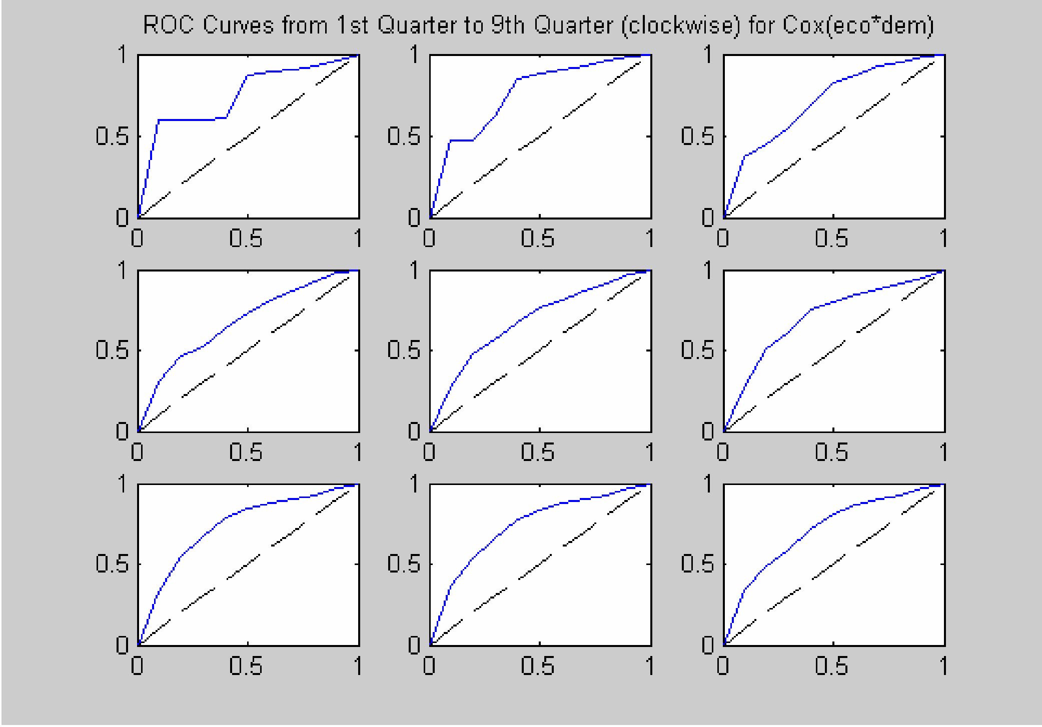

Figure 2e: ROC curve for Cox model with interaction terms and time-dependent coefficients

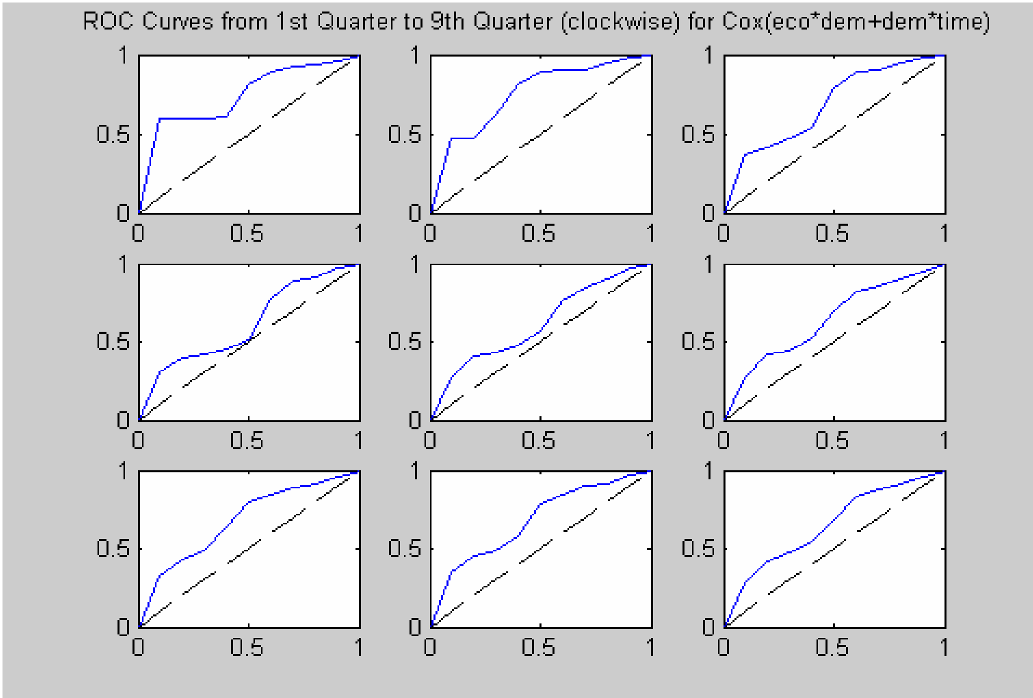

Figure 3: Predicted and actual purchases on holdout sample