Stockholms universitet

Survival of Deposit Accounts Using Logistic Regression

Felix Wahl

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2014:7
Matematisk statistik
Juni 2014

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2014:7**
http://www.math.su.se

Survival of Deposit Accounts Using Logistic Regression

Felix Wahl*

June 2014

Abstract

The objective of this thesis is two-fold. The first is to find a predictive model for the probability of a deposit account, held at a Swedish bank, being closed or emptied within a year. The second is to describe how a change in the interest rate of the deposit account affects this probability. The data consists of monthly observations of the maximum account balance for each deposit account along with a set of explanatory variables. We use the explanatory variables from the first month in each year. The remaining months are used to determine whether the account stayed open or closed. When the data is of this form we have approximately 200 000 yearly observations. We use logistic regression along with a set of different algorithmic selection procedures. A number of model validation statistics are used and the conclusion is that no model is completely satisfactory in regard to predictive capabilities. Nevertheless, we find that the coefficient for the interest rate is robust, i.e. does not change considerably, between models. Together with the fact that the coefficients in a logistic regression model always are the log-odds ratios, even if the model does not t the data, we find that the interest rate coefficient is interpretable. Note that the data and results used and obtained in this thesis are confidential and hence only an overview is presented.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail:felixwahl91@gmail.com . Supervisor: Tom Britton.

Preface

This paper constitutes a bachelor's thesis of 15 ECTS in Mathematical Statistics at the Department of Mathematics at Stockholm University. This thesis has been carried out in collaboration with SBAB Bank AB.

I would like to thank my supervisor Tom Britton for all the help and advice and also my external supervisor Peter Svensén, CRO at SBAB Bank AB, for the opportunity to conduct this work. I would also like to thank all the people at SBAB Bank AB who have helped me throughout this thesis, especially Fredrik Lundgren.

Contents

1 Introduction

In 2007 SBAB Bank AB started utilizing deposit accounts. In recent years they have grown their deposit volume quite a bit. The total volume now amounts to about 50 billion SEK and is predominantly from private individuals. The strong increase in deposits is thought to come from SBAB maintaining a significantly higher interest rate than competing banks combined with advertisement. Now it is of interest to determine how sensitive the current customers are to changes in the interest rate. It is hypothesized that customers leaving the large banks, who generally have quite low interest rates, go looking for a bank with high rates. Once they've changed bank they might however not be that sensitive to rate changes.

We will in this thesis describe the steps taken to find a predictive model, that is, a model that predicts the survival of deposit accounts on a yearly basis. The effect of a change in the interest rate on the survival will also be examined.

Note that the data on which this model building occurs is confidential and thus also the results to some extent. Hence more weight is placed on the steps up to finding a model rather then interpreting the results.

2 Background

2.1 The Problem

The objective of this thesis is two-fold. The first objective is to try to find a model with acceptable forecasting abilities. That is, a model with which it is possible to estimate, with good accuracy, future probabilities of an account being closed within a year. The second objective is to estimate how a change in the interest rate affects the survival of accounts.

2.2 The Data

The major part of this thesis has consisted of data processing, that is building the data set to be analyzed from several data sets consisting of only one or two variables. Also, some variables have taken some time to collect from sources outside of the data warehouse in the bank. In short, a data warehouse is a central database where data from different departments are stored, often uploaded automatically.

The data consists of monthly observations of the maximum account balance for each deposit account together with a set of explanatory variables x_1-x_{24}. The values of all explanatory variables are either from the month of the observation or backwards in time, this is since we cannot use future data to predict the future, obviously. We cannot disclose what variables have been used, but the response variable is account status over intervals

of one year. Hence only the first month in each year for the account of the monthly observations are used, meaning not always January but the first month from opening of the account. The other months are used to determine whether the account "survived". When the data is of this form we have approximately 200 000 observations. Note that if an account did not close the first year our sample will contain more than one observation from that same account. For example, if an account closed on its third year, there will be three observations corresponding to that account.

The status of the account is defined as closed/inactive if the maximum balance is below a certain cut off value during three consecutive months. This is so that we do not count accounts that are still in use, but might not receive deposits for a short period of time, as closed/inactive. Without this time period we might overestimate the true probability of an account closing.

The cut off value is defined by two criteria, firstly it is a set value c and secondly it is the maximum balance during the year times the interest rate and taxes. This is since an individual might remove all money from the account but then receive an interest payment (minus taxes) resulting in an account balance greater than the set value c.

Some variables are inherently bad to use for forecasting future events, especially variables that have an inherent need for extrapolation to predict future events. One example of such a variable is time, or more specifically time since SBAB started using deposit accounts. Every year, when trying to predict the next year, one would have to enter a value into the model that it has not been built upon. Hence, the model has inherent issues with extrapolation. Therefore, variables with the need to extrapolate will be left out of the model building (variables x_5, x_{10} and x_{15}).

As mentioned, we look at accounts on a yearly basis starting at day one for each account. This is so that we use accounts from their opening day to minimize the bias resulting from truncation. We also make sure that all accounts in the later time intervals are such that they have had the possibility to be open for a year, otherwise we would have a problem with censoring and hence overestimate the probability of an account closing.

Note that when referring to the interest rate from now on we are actually talking about the difference between the interest rate at SBAB compared to the maximum interest rate of four competing banks. These four banks also hold a significantly higher interest rate than the market average.

3 Theory

In this section we will describe the theory needed and used to perform the modeling.

We will show and discuss tools used in conjunction with the logistic regression model, such as the model selection procedures and model validation.

3.1 Logistic Regression

Let Y be a binary response variable and $x_1, x_2, ..., x_n$ be explanatory variables. Define $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|\ X = x)$. The logistic regression model is then defined as (Agresti, 2013, p. 163)

$$\pi(x) = \frac{\exp(\alpha + \sum_{i=1}^{n} \beta_i x_i)}{1 + \exp(\alpha + \sum_{i=1}^{n} \beta_i x_i)} = \frac{\exp \boldsymbol{\beta X}}{1 + \exp \boldsymbol{\beta X}},$$

where $\beta_1, ..., \beta_n$ are the regression coefficients. This yields the linear relationship in the logit

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \boldsymbol{\beta X}. \tag{1}$$

The regression coefficients are usually estimated using maximum likelihood estimation. The null hypothesis $H_0 : \beta_i = \beta_0$ is tested for each coefficient using a Wald test, which is of the form

$$\frac{(\hat{\beta} - \beta_0)^2}{\sqrt{Var(\hat{\beta})}}.$$

It is compared to a chi-squared distribution with 1 degree of freedom. For our test we set each β_0 to zero.

3.1.1 Interpretation of Coefficients

The explanatory variables in a logistic regression model are linear in the logit (1). By exponentiating (1) we see that the odds ratio is an exponential function of the explanatory variables. Assume we have a logistic regression model with one explanatory variable x. The odds ratio at $x + \Delta x$ is

$$\frac{\pi(x + \Delta x)}{1 - \pi(x + \Delta x)} = e^{\alpha + \beta(x + \Delta x)} = e^{\beta \Delta x} \frac{\pi(x)}{1 - \pi(x)}.$$

That is, for an increase (or decrease) Δx the odds ratio multiplies by $e^{\beta \Delta x}$. Or simply, for each unit increase of x the odds ratio multiplies by e^{β}.

3.2 Model Selection

When dealing with many explanatory variables it is unreasonable to expect that every possible model be tested. If we have k explanatory variables, not considering interactions, then we have 2^k possible models. For example, if $k = 10$ then we have $2^{10} = 1024$ possible models. Hence, when dealing with more than a few explanatory variables we use different algorithmic methods to either find a good model according to certain criteria or to limit the model space to more manageable sizes.

For model selection we will use a number of algorithmic methods to find candidate models: *Forward selection*, *Backward elimination*, *Stepwise regression* and *Purposeful selection*. We will also look for models that minimize certain information criteria (AIC and BIC), but these will be introduced in section 3.6.

Backward elimination starts with all variables and then, sequentially, removes the variable that is least significant and stops at a prespecified confidence level (according to the Wald test). Forward selection starts with only the intercept and then, sequentially, adds variables that are significant at the prespecified confidence level. Stepwise regression works like Forward selection but at each step checks if the variables currently in the model are still significant, if not, they are removed. Purposeful selection (Hosmer & Lemeshow, 2013 p.90-93) takes, in short (see reference for full description) the following steps:

Step 1: Identify variables with a p-value less than 0.25 in a univariate analysis for each independent variable. The rationale behind the high significance level is that we do not want to throw away important variables that, while not significant alone, might be when other variables are in the model.

Step 2: Fit a multivariate model with all variables identified for inclusion at Step 1 and then assess the importance of each variable using the Wald statistic. Variables not significant at traditional levels of significance are eliminated in a stepwise fashion, much like Backward elimination.

Step 3: The coefficients in this reduced model are compared to those of the model containing all variables identified in Step 1. If any coefficient changed its value by more than 20% we have an indication that one or more of the removed variables are important in that they provide a (perhaps) needed adjustment of the estimated effects of the other variables. Hence, if this is the case we add such variables back into the model until we are satisfied that we have a model containing important variables.

Step 4: Now we add each variable not selected in Step 1 to the current model, one at a time, and check its significance by using the Wald statistic.

This is done by the same rationale as for the 0.25 significance level in step 1.

Step 5: Before starting selection procedures we should have a list of variables that might interact. Now we add these, one at a time, to the model we got at the end of Step 4 and then examine the significance using the Wald statistic. We then add each significant interaction to the model at the same time, removing, sequentially, interactions not significant at traditional levels. No main effects are considered for removal in this step.

Step 7: We now assess the models fit to the data. This is of course not something special for Purposeful selection and must be done after fitting any model.

Bursac et al. (2008) showed via simulations that Purposeful selection identifies and retains confounders correctly at a larger rate than other selection procedures. Although they do note that this is generally for smaller data sets ($\approx 240 - 600$ observations) and that for larger data sets (≈ 1000 observations) all selection procedures except Forward selection converge to the same model.

3.3 Interactions with Variable of Interest

Since the interest rate is the variable of interest, we want to be able to interpret the effect it has on the outcome easily. Hence we will try to keep interaction terms between this variable and other variables out of the models. If we think such interactions should be in the model we will try to use them when finding a model with a good forecasting ability, but we will try to keep these away when finding models that we want to use for interpretation of the interest rates effect on the outcome.

3.4 Transforming Continuous Variables

In a logistic regression model, the covariates are assumed to be linear in the logit, see equation (1). Hence we need to check if the variables actually fulfill this. We do this by sorting each independent variable by rank and then dividing these into a number of equally sized groups. In these groups we can estimate the probability as the proportion of observations that ended in "deaths". Using this we can estimate the logit by calculating equation (1). Then we plot the average of the independent variable in each group against the estimated logit. This should yield a linear relationship if the assumption of linearity in the logit holds. If it does not, then we have to think of possible transformations, such as power or log transformations of the independent variables. We will also linearize relationships using linear splines, or rather several linear segments.

3.5 Multicollinearity

In logistic regression models, as in linear regression models, multicollinearity is a problem. Multicollinearity means that there exists a linear dependence between explanatory variables. It is a problem because when two or more variables are highly correlated with one another it can be difficult to get reliable estimates of their effects. It does not bias the maximum likelihood estimates of the coefficients, but the standard errors may get large and unstable (Allison, 1999, p. 48) since it is difficult to determine distinct effects of the collinear variables.

To see if multicollinearity exists it is often proposed that one examines the pairwise (Pearson) correlation coefficient between the explanatory variables. This is not optimal since two variables might not be highly correlated with one another, but together with all the other explanatory variables we might have a (approximate) linear dependence. Hence we will also examine the *variance inflation factors*.

Let R_j^2 be the coefficient of determination for a linear regression model with x_j as the response variable and the remaining explanatory variables as regressors in the linear model. The variance inflation factor for the explanatory variable x_j is defined as

$$\text{VIF}_j = \frac{1}{1 - R_j^2}.$$

We see that as the coefficient of determination increases towards 1 the variance inflation factor goes to infinity. Standard cutoff points often proposed is a variance inflation factor greater than 5 or 10. If any variable has a variance inflation factor that is deemed too large, than we might have to think about removing one of the problematic variables from our analysis.

We also examine the correlation coefficients between variables and conclude that if two variables have a correlations coefficient greater than 60% then one of them should be excluded from the analysis. This is a somewhat arbitrarily chosen cutoff point.

3.6 AIC and BIC

The *Akaike information criterion* (AIC) is a measure we can use when comparing models. It is based on the value of the likelihood function for a model but is adjusted by the number of parameters in the model. It can be said that AIC penalizes models with more parameters and thus prefers simpler models. AIC is defined as (Agresti, 2013, p. 212)

$$\text{AIC} = -2(\text{maximised log likelihood} - \text{number of parameters in model}).$$

It is important to note that the AIC value does in no way indicate how well a model fits the data, but rather how well it fits the data compared to other models.

The *Baysian information criterion* (BIC), also called the *Schwarz criterion* (SC), is another measure we can use when comparing models. It is defined as

$$\text{BIC} = -2\cdot\text{maximised log likelihood} + \ln(n)\cdot\text{number of parameters in model.}$$

As can be seen, it penalizes complex models greater than does AIC. It is based on a Baysian argument for finding the model with the highest posterior probability among a set of models (Agresti, 2013, p.212-213).

We will use both AIC and BIC in the model selection, favoring BIC when they disagree. This is because we would rather have a simple model than a complex one. We will not however fit all possible models and then choose the one with the smallest AIC/BIC, this would not be feasible with the amount of variables and data we have. Rather we will use them in determining if dropping or adding a certain variable or interaction from or to the model leads to a better model.

We will also do a Backwards elimination procedure including all variables and possible interactions, setting the threshold p-value for dropping a variable to zero. Then we will plot the steps in the procedure vs AIC and BIC at each step. This way we can find a model, from this new subset of all models, that is the simplest possible in the range of models with the lowest AIC or BIC.

3.7 Validation & Predictive Power

Any model must always be validated before it can be used or trusted. Here we will validate our models by utilizing a collection of tools.

Note that we want to find a model that is well calibrated to the data, that is a model for which the predicted probabilities correspond to the actual probabilities of events. Also, we would like to find models that discriminate well, that is a model that generally assigns larger probabilities to events and smaller probabilities to non-events. For instance, The Hosmer-Lemeshow test is used for calibration while the classification tables are used for discrimination as one will note in the sections below.

This section is concerned with finding a model that has good forecasting abilities. When it comes to interpretation of effects, most importantly how the interest rate affects the probability of an account getting closed, calibration is not important. To cite Hosmer and Lemeshow (2013, p. 186)

> When the focus of the study is on the $\hat{\beta}$'s (or odds ratios), calibration is not important. It is important when the estimated probabilities are meaningful and of interest to the investigator.
>
> The coefficients of a logistic regression analysis are always the log-odds ratios — whether the model fits or not. However, if the study's objective is to estimate the $\Pr(y = 1)$ then we need to assess calibration and discrimination.

3.7.1 The Hosmer-Lemeshow Test, Ungrouped Data and Effects of Sample Size

Since we have ungrouped data, it is not possible to use the regular Pearson chi-square and deviance statistics (which is the reason they are not described in this thesis). One possible goodness of fit test for ungrouped data is the *Hosmer-Lemeshow test*. It is a chi-square test formed by sorting the data by estimated probabilities and then dividing these into a number of approximately equally sized groups. Usually one forms 10 such groups and from this one can create a Pearson statistic comparing observed and fitted counts. The problem is that the power (the probability of rejecting a false model) of a chi-square test increases with sample size. Usually one is interested in finding an acceptable model, if such a model exists, but when the sample size is very large and the number of variables are small in comparison this can be problematic since the model has to be very well fitting to be "accepted" by the Hosmer-Lemeshow test. Paul et al. (2013) studied methods for specifying the number of groups so that the power would equal what one would have for a sample of size 1000 and 10 groups. They concluded that they do not recommend this kind of test for sample sizes exceeding 25,000, which we have. We will still use the Hosmer-Lemeshow statistic, but we are aware of the fact that even a model that is considered significantly different from the data according to the Hosmer-Lemeshow statistic might be a reasonable model, unless of course the chi-square value is very large.

One way around the problem of the Hosmer-Lemeshow test is to categorize continuous covariates, then one can use for example the deviance measure. This however has its fair share of problems. We will discuss this further in section 6.

3.7.2 Classification Tables

Assume that we have a model that we use for predicting future values. Sometimes it might be of use to classify individuals with an estimated probability above a certain threshold value into a high-risk group and the rest into a low-risk group. This is what we do in a *classification table*.

Let the prediction for observation k be $\hat{y} = 1$ when $\hat{\pi}_k > \pi_0$ and $\hat{y} = 0$ when $\hat{\pi}_k \leq \pi_0$, for some cutoff probability π_0. By application of the "leave-one-out" cross validation approach we estimate $\hat{\pi}_i$ based on the model fitted to the other $n - 1$ observations.

Now we define the quantities sensitivity (ability to predict survival correctly) and specificity (ability to predict death correctly) as

$$\text{sensitivity} = P(\hat{y} = 1|y = 1) \quad \text{and} \quad \text{specificity} = P(\hat{y} = 0|\ y = 0). \tag{2}$$

Predictive power can now be summarized as the proportion of correct clas-

sifications:

$$\begin{aligned} P(\text{correct classifications}) &= P(y = 1 \text{ and } \hat{y} = 1) + P(y = 0 \text{ and } \hat{y} = 0) \\ &= P(\hat{y} = 1|y = 1)P(y = 1) + P(\hat{y} = 0|\ y = 0)P(y = 0). \end{aligned}$$

For this and more, see Agresti (2013, p.223).

3.7.3 ROC curves

A *receiver operating characteristic* (ROC) curve is a plot of sensitivity as a function of (1 - specificity) for all possible π_0, see equation (2). The ROC curve hence sums up the classification table in a way such that we get an overall view of predictive power for all possible π_0. For any given specificity, the higher sensitivity we have the better the predictions, and the other way around. This means that the larger the area under the ROC curve the better the predictive power (Agresti, 2013, p. 224).

It can be shown that the area under the ROC curve is equal to the probability that the model will rank a randomly chosen event higher than a non-event. (Hosmer & Lemeshow, 2013, p.177)

In Figure 1 an example of a ROC curve, from utilizing SAS software on a simulated dataset, can be seen.

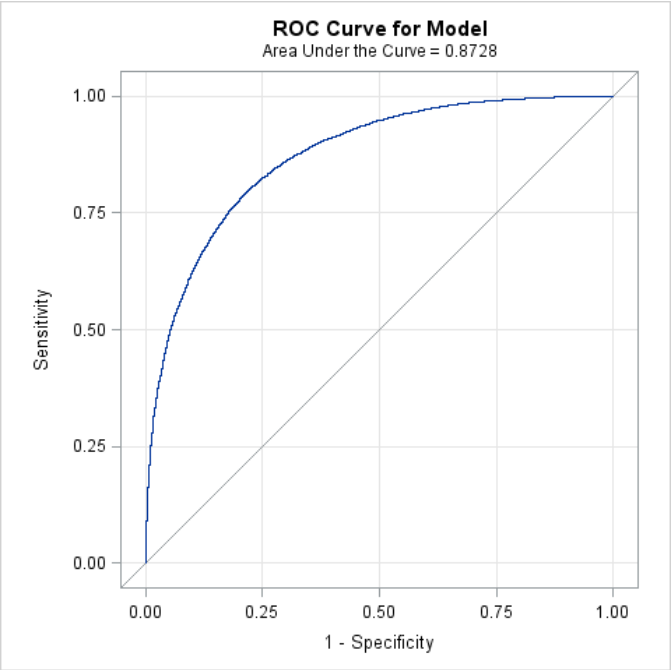

Figure 1: Example of a ROC curve. It is from a simulated dataset for illustration purposes.

3.7.4 Brier Score

The *Brier score*, proposed by Brier (1950), is defined as

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2,$$

where f_t is the predicted probability and o_t is the observed outcome. Hence a small Brier score correspond to a good calibration of predictions. As a reference, pure guessing would yield

$$BS = \frac{1}{N}\sum_{t=1}^{N}\left(\frac{1}{2} - o_t\right)^2 = \frac{1}{4}.$$

The Brier Score can be decomposed into three parts (Murphy, 1973)

$$BS = \text{Reliability} - \text{Resolution} + \text{Uncertainty}$$

where

$$\text{Reliability} = \frac{1}{N}\sum_{k=1}^{K} n_k(f_t - \bar{o}_k)^2,$$

$$\text{Resolution} = \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2,$$

$$\text{Uncertainty} = \bar{o}(1 - \bar{o}).$$

Here K is the number of unique forecasts, $\bar{o}$ is the frequency of outcomes in the total sample, $\bar{o}_k$ is the frequency of outcomes in the k:th sample of unique forecasts and n_k is the sample size of the k:th sample of unique forecasts.

Since we have ungrouped data with a lot of single unique forecasts, we will sort the estimated probabilities by rank and then split them into a number of groups of equal size, as done in the Hosmer-Lemeshow test. Then we will use these groups as groups of unique forecasts. This is done simply so that we can calculate the Reliability, the Resolution and the Uncertainty.

A short explanation of the three quantities is warranted. The Reliability measures the distance between outcome and estimated probabilities. We do of course want this quantity to be as small as possible. The Resolution measures the distance between the frequencies of the outcomes from the K "unique" forecasted probabilities and the frequency in the whole sample. A large Resolution corresponds to a greater inherent ability to discern situations where the event is likely to happen from those where the event is unlikely to happen. The Uncertainty measures the inherent uncertainty in the system. With these quantities, the Brier score is more informative. This can be seen by thinking about two different situations and models. First we think of coin tossing where the probability of success (heads) is

50%. A perfect model would always predict 50% chance of success, this would yield a Brier score the same as pure guessing (as shown above), that is $BS = 0.25$. The other example is a deterministic situation where the success event always happens given a certain condition, a perfect model would always predict 100% chance of success given this condition or 0% if the condition is unfulfilled. This model would yield $BS = 0$. Hence, two "perfect" models yield two quite different Brier scores. This is of course because of the different problems the examples above present. For example, the Uncertainty in the two examples are 0.25 and 0, respectively. That is, we would expect a larger Brier score in the first example simply because the coin tossing presents a greater uncertainty than the zero uncertainty of the always success example.

We can also construct a test for testing whether the forecasted probabilities are equal to the real probabilities. Under the null hypothesis $H_0 : f_t = p_t$, where p_t is the true probability of trial t, the expected value and variance of the Brier score are (see Appendix A for derivation)

$$E[BS] = \frac{1}{N}\sum_{t=1}^{N} f_t(1 - f_t),$$

$$Var(BS) = \frac{1}{N^2}\sum_{t=1}^{N} f_t(1 - f_t)(1 - 2f_t)^2.$$

Using this we can create the z-statistic

$$z = \frac{BS_{\text{Obs}} - E[BS]}{\sqrt{Var(BS)}}.$$

Hence we can test the null hypothesis by calculating the p-value corresponding to the observed z using the standard normal distribution.

3.7.5 Generalized R-square

The coefficient of determination, R^2, is something everyone familiar with linear regression has utilized. It would hence be nice to have a similar measure when building logistic regression models. We define the generalized coefficient of determination as

$$R^2 = 1 - \left(\frac{L(\mathbf{0})}{L(\hat{\boldsymbol{\beta}})}\right)^{\frac{2}{n}}$$

where $L(\mathbf{0})$ is the likelihood of the intercept-only model, $L(\hat{\boldsymbol{\beta}})$ is the likelihood of the model in question and n is the sample size. The generalized coefficient of determination R^2 does not achieve its maximum at 1 for discrete models, but rather at $R^2_{\max} = 1 - L(\mathbf{0})^{2/n}$. Since most people think of

R^2 as lying between 0 and 1 we might want to make an adjustment to it. Instead a rescaled coefficient, ranging from 0 to 1, is defined as

$$\tilde{R}^2 = \frac{R^2}{R^2_{\max}}.$$

For this and more see SAS/STAT® 9.3 User's Guide (2011, p.4115)

We include this description of the Generalized R^2 simply for completeness sake. Hosmer and Lemeshow explain why quite well (Hosmer & Lemeshow, 2013, p. 186)

> In general, these measures [R^2] are based on various comparisons of the predicted values from the fitted model to those from model(0), the no data or intercept only model and, as a result, do not assess goodness of fit. We think that a true measure of fit is one based strictly on a comparison of observed to expected values from the fitted model.

3.7.6 Validation Using Holdout Samples

When dealing with a very large data set, it is of no problem to split it into a training and validation set. That is, we develop a model using the training set and then we evaluate the model using the validation set. This can be done by either a random sampling or choosing the data for the last year as the validation set (since we most often want to predict the future using our model). We will mainly consider the last one.

When evaluating models on the validation set we can examine statistics such as the Hosmer-Lemeshow statistic, area under the ROC curve and the Brier score. For instance, a greatly lower area under the ROC curve for the validation set compared to the training set would indicate bad prediction power.

One also often fits separate models, containing the same variables, on the training and validation set in order to examine whether the coefficients are approximately the same between the two. This is a problem when dealing with data sets that lie separately in time. As an example, the interest rate does not vary that greatly over a year and hence the estimate for its coefficient would presumably differ by quite a bit compared to previous years. This would probably not be because of a difference in effect, but because of the scarcity of variation.

3.7.7 Plotting Estimated vs Observed Probabilities

To assess the models fit we can plot estimated vs observed probabilities. If the model performs well we should get a 45 degree line through the origin. One of the advantages this plot has is that it gives us an indication of where

a badly performing model fails. A model might generally perform well, but is badly calibrated for very low or high probabilities, for instance.

Obviously we cannot plot estimated vs observed probabilities for each observation in the data set, we will have to estimate the observed probabilities by sorting the estimated probabilities by rank and then dividing these into a number of equally sized groups in which we can calculate the observed probabilities as the proportion of "deaths". The estimated probabilities in this graph will then be the mean of the estimated probabilities from the model.

3.7.8 Robustness of Coefficients

When we are interested in how one or more variables affect the probability of survival, in order to be able to rely on the estimates to be close to the truth, we would like the coefficient to be robust to changes between models. If we for instance have four models, two from minimizing AIC and BIC, one from Purposeful selection and one univariate model, all containing the variable of interest, we would like the coefficient of interest to be approximately equal in all models. Otherwise it gets a bit more difficult to determine which coefficient to trust, if any.

This will mainly be done for the interest rate. Since this is the main variable we want to say something about, especially if we cannot find a model with an acceptable forecasting ability.

4 Analysis

In this section we will provide an overview of the analysis and the results, sadly nothing too specific since both the data and the results are confidential.

In our problem we have a binary response variable. That is, the account either became inactive (died) or it continued being used (survived). This is a situation where logistic regression is commonly used. We will hence let $\pi(x)$ in equation (3.1) be the probability of an account getting closed within one year.

It should be mentioned that one of the assumptions of the logistic regression model is independent observations. This obviously does not hold in our data set since we firstly have dependency between the same account over different years and secondly a dependency between two or more accounts that have at least one account holder in common. During this analysis we will assume independent observations.

4.1 Multicollinearity

We start by examining multicollinearity utilizing the variance inflation factors and the Pearson correlation coefficient between the explanatory vari-

ables. Doing this we see that there are variables with a high variance inflation factor (at least greater than 5). Removing one of them at a time, recalculating the variance inflation factors at each step and only removing variables that still have a high VIF, results in removal of four variables.

When choosing which variable of collinear variables to remove we consider several factors. Firstly, which has the best fit in a univariate logistic regression model. Secondly, linearity in the logit. One variable might be linear in the logit while another has a horrendous looking relationship that is hard to find a tranformation for. Thirdly, we also consider which variable is actually most reasonable to keep in the model.

After removing variables using VIF we examine the pairwise correlation coefficient of the remaining variables. If any correlation is above 60% we consider removal of one of the variables. Here we consider the same reasons for removal as above. This results in us removing two additional variables.

Note that we are now down to 18 explanatory variables.

4.2 Transforming Continuous Covariates

Next we examine the scale of the continuous covariates as discussed in section 3.4. In Figure 2 we can see one of the explanatory variables, x_1, plotted against the estimated logit using 80 groups. We can see that we have three approximately linear parts, one before the point c, one after the point k and one in between. We therefore construct a transformation that yields four new variables, three continuous variables

$$x_{1,1} = \begin{cases} x_1 & \text{if } x_1 \leq c, \\ 0 & \text{if } c < x_1 < k, \\ 0 & \text{if } k \leq x_1, \end{cases}$$

$$x_{1,2} = \begin{cases} 0 & \text{if } x_1 \leq c, \\ x_1 & \text{if } c < x_1 < k, \\ 0 & \text{if } k \leq x_1, \end{cases}$$

$$x_{1,3} = \begin{cases} 0 & \text{if } x_1 \leq c, \\ 0 & \text{if } c < x_1 < k, \\ x_1 & \text{if } k \leq x_1, \end{cases}$$

and one categorical

$$x_{1,4} = \begin{cases} 0 & \text{if } x_1 \leq c, \\ 1 & \text{if } c < x_1 < k, \\ 2 & \text{if } k \leq x_1, \end{cases}$$

that works as the intercept for the continuous variables above.

Now it should be noted that we have chosen the points c and k somewhat arbitrarily. These points could, and should, be estimated properly, but this is out of the scope for this thesis. We should not be too worried though since during this analysis the plot in Figure 2, which is based on the whole dataset, has also been plotted for random subsets of the data and for each year, yielding very similar relationships. Hence we can feel confident that the points are close to where they would be estimated to be. This holds at least for the point c, it is a little more difficult to determine where the point k should actually be. Nevertheless, we have tried to slightly adjust the positions of k and c which yielded approximately the same fit. Hence we can feel confident that the result is robust to slight errors in the positions of k and c.

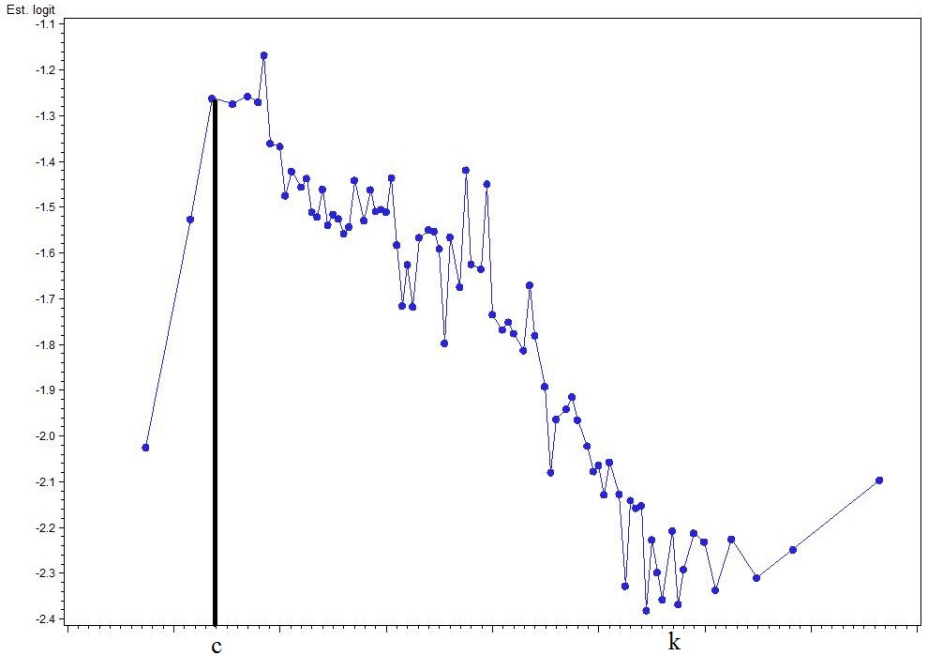

Figure 2: Plot of estimated logit against the value of one of the explanatory variables, x_1.

Another example can be seen in Figure 3 where explanatory variable x_2 is plotted against the estimated logit. Variables with this pattern in the logit will be treated as if they were linear, but with an awfully large variance, since this is the only thing we can really assume.

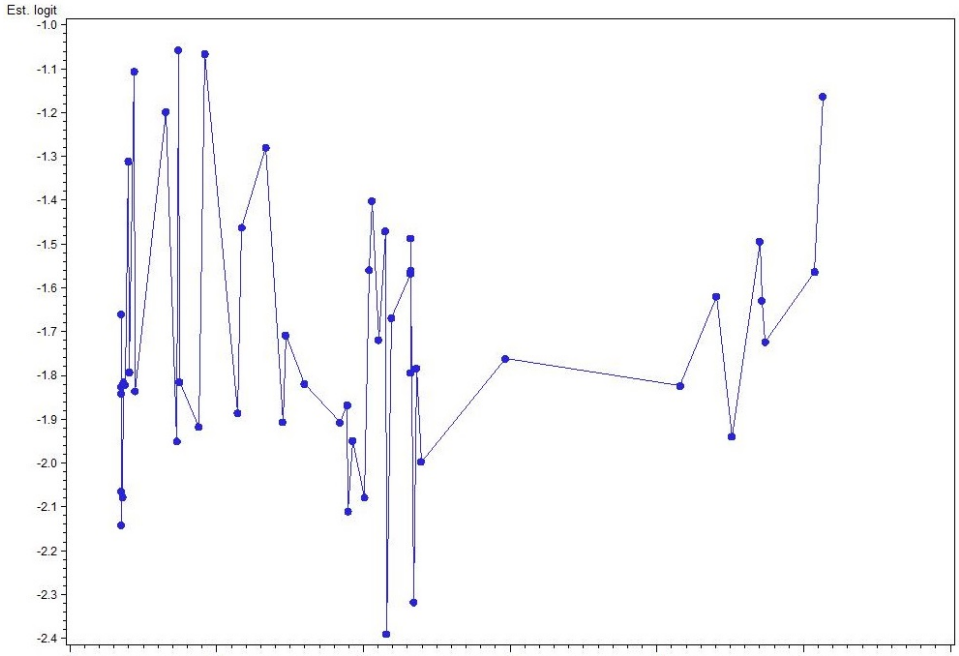

Figure 3: Plot of estimated logit against x_2.

4.3 Model Selection

Next we start building our logistic regression models. We will use Forward selection, Backward elimination, Stepwise regression, Purposeful selection and a procedure to find a model that is in the vicinity of the lowest AIC/BIC.

4.3.1 Forward Selection, Backward Elimination & Stepwise Regression

We begin by utilizing Forward selection, Backward elimination and Stepwise regression as is quite standard. This results in models containing almost all variables and interactions, regardless of how small a threshold value we choose for the p-value (0.05, 0.01 and 0.0001). This is because we are working with large amounts of data, hence almost any effect can be found significant. Therefore we conclude that we might need other methods to build sensible models. Note that a model can contain a lot of variables and interactions and still be sensible.

4.3.2 Purposeful Selection

First we build univariate logistic regression models for each explanatory variable. Since we have a lot of data, it is not surprising that every variable

except one is significant at a level below 0.001, the majority being below 0.0001. The one not significant has a p-value of 0.4422. Hence we go on and construct a multivariate model containing all variables except the one that was not significant in the univariate analysis.

In the multivariate model, some variables get large p-values. We remove these in a stepwise fashion resulting in a model with only significant variables, at least at the 0.001 level. This yields the removal of three variables of which two are part of linear spline transformations mentioned above. After this we investigate whether any coefficients changed considerably (more than 20%), which none did.

We then add the variable that was not significant in the univariate analysis to the model, still yielding a coefficient not significantly different from zero, although with a much smaller p-value this time. Hence, it is not kept. Adding this variable also yields a higher AIC and BIC.

We now see that one variable barely has any effect at all, hence we try to remove it in spite of its significant Wald-test. Removal yields a lower BIC and a higher AIC. This with the fact that it barely had any effect leads to its removal.

Now we create a list of possible pairs of variables that might interact with each other. Note that these interactions are included in the other selection procedures. There must be a reasonable reason for the interaction to exist in order for us to consider it. Then we add each interaction, one at a time, to the current model. The interactions that are significant at traditional levels are then added to the model all at the same time, actually all interactions either have a very low or a very high p-value so we really do not have to worry about what is an appropriate significance level. As before, removal occurs in a stepwise fashion if the interaction is not significant. Only one, with a p-value equal to 0.7395, is removed.

Before moving on to checking the models fit, we take some time to consider whether the variables have a clinically significant effect or if their effect can be reasonably explained. Some variables might be significant but there are no possible ways of explaining why it would have the effect it has. We remove these variables and then examine whether this yields a lower AIC/BIC. If it does, we can feel confident in our suspicion that the variable should have been removed. We remove one variable that had a low effect, yielding a lower BIC, but not a lower AIC. Since it resulted in a lower BIC together with it barely having any effect we remove it.

4.3.3 Minimizing AIC & BIC

We now build two models based on minimizing AIC and BIC. Checking the AIC and BIC of every possible model is not reasonable. As mentioned in section 3.2, if we have k explanatory variables, not considering interactions, then we have 2^k possible models. Since we have slightly fewer than 20

variables we conclude that another method is needed.

What we do is a Backward elimination with a threshold p-value of 0. At each step we get a value for AIC and BIC which we can plot against the steps. Note that the last step corresponds to the intercept only model and the first step to the one containing all variables and interactions. This can be seen in Figure 4 and 5.

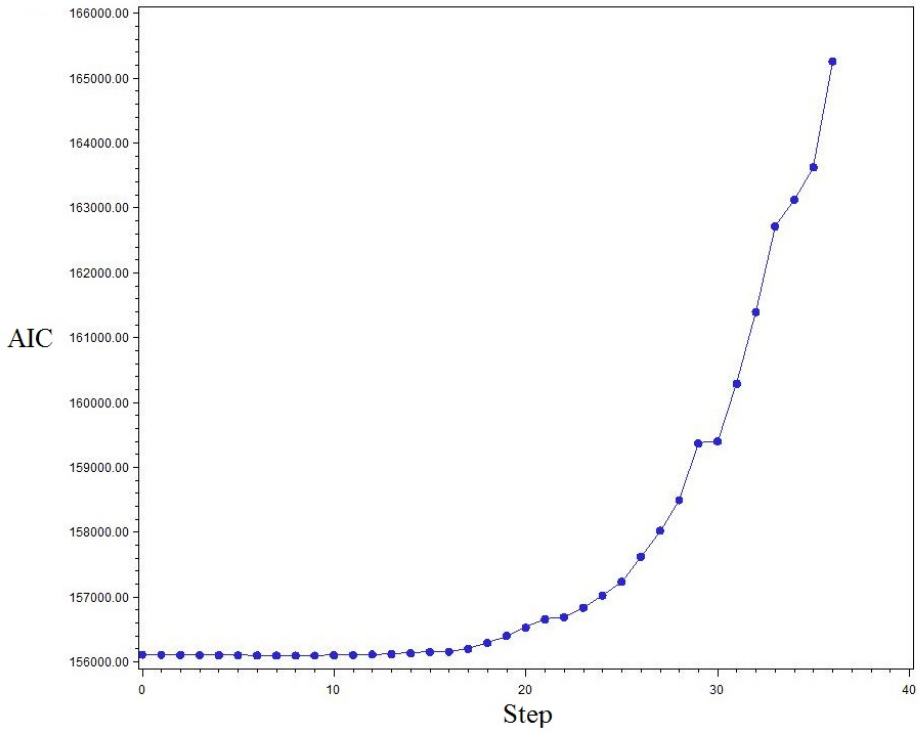

Figure 4: Plot of steps in the Backward elimination with threshold p-value of 0 against the AIC at each step.

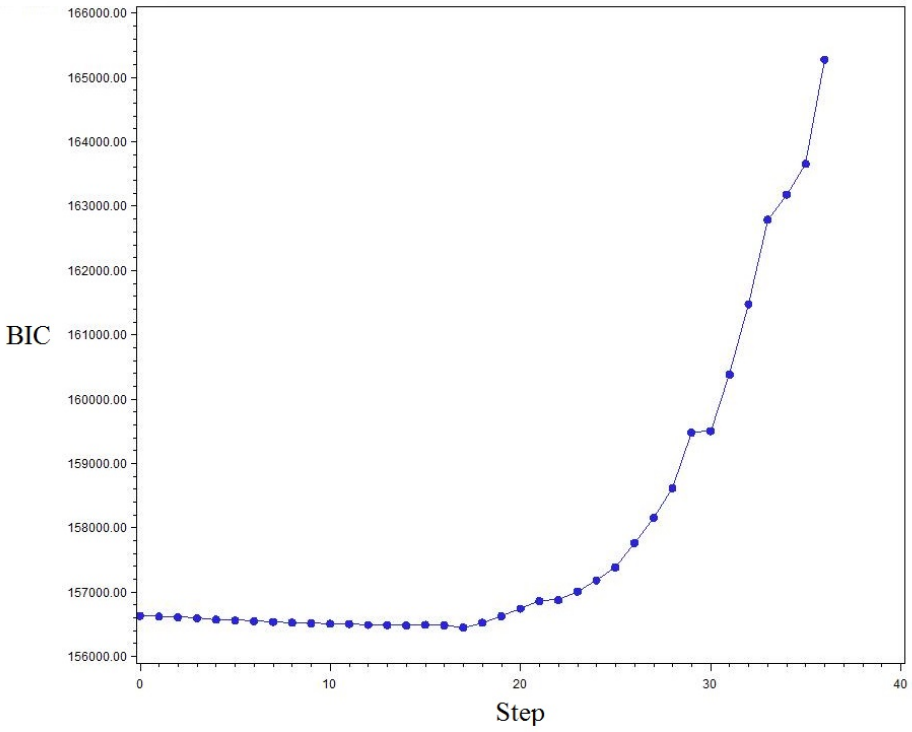

Figure 5: Plot of steps in the Backward elimination with threshold p-value of 0 against the BIC at each step.

The AIC is minimized at step 9 and the BIC at step 17. Note that we do not actually use the plots to visually find the minimal AIC/BIC, instead we check the actual values in a table.

4.4 Validation & Predictive Power

Now we have a few models to work with, the one from Purposeful selection and the two from minimizing AIC & BIC. In Table 1 and 2 the results can be seen.

Table 1: AIC, BIC, area under ROC curve (AUC), generalized coefficent of determination R^2 and max-rescaled coefficient of determination $\tilde{R}^2$, Brier score (BS) and significance test of null hypothesis $H_0 : f_k = p_k$ for the different models.

Statistic \ Model	Purposeful selection	Minimzed AIC	Minimized BIC
AIC	156256.32	156100.82	156206.40
BIC	156551.87	156518.66	156450.99
AUC	0.679	0.680	0.679
R^2	0.0516	0.0525	0.0518
$\tilde{R}^2$	0.0904	0.0919	0.0907
HL	235.6434	140.4707	152.6350
BS	0.12004	0.11997	0.12007
E[BS]	0.12040	0.12021	0.12032
Var(BS) $[10^{-7}]$	2.62180	2.61839	2.62347
z-statistic	-0.7131	-0.4680	-0.4948
Two-sided p-value	0.4758	0.6398	0.6207

Table 2: Brier score and its decomposition for the different models. Note that the Brier score here is different from the one in Table 1, this is since this Brier score is based on the grouping of estimated probabilities. This is done so that we can calculate the Reliability, Resolution and Uncertainty.

Statistic \ Model	Purposeful selection	Minimzed AIC	Minimized BIC
Brier score	0.12035	0.12021	0.12032
Reliability	0.00011	0.00007	0.00007
Resolution	0.00730	0.00739	0.00728
Uncertainty	0.12754	0.12754	0.12754

We see that all models are significantly different from the data according to the Hosmer-Lemeshow test. Note that the statistic is approximately chi-squared with 8 degrees of freedom. The area under the ROC curve is quite small for all models so we do not expect to be able to discriminate well with these models. To illustrate, Figure 6 shows the ROC curve for the minimized BIC model. The max-rescaled coefficients of determination are all incredibly small.

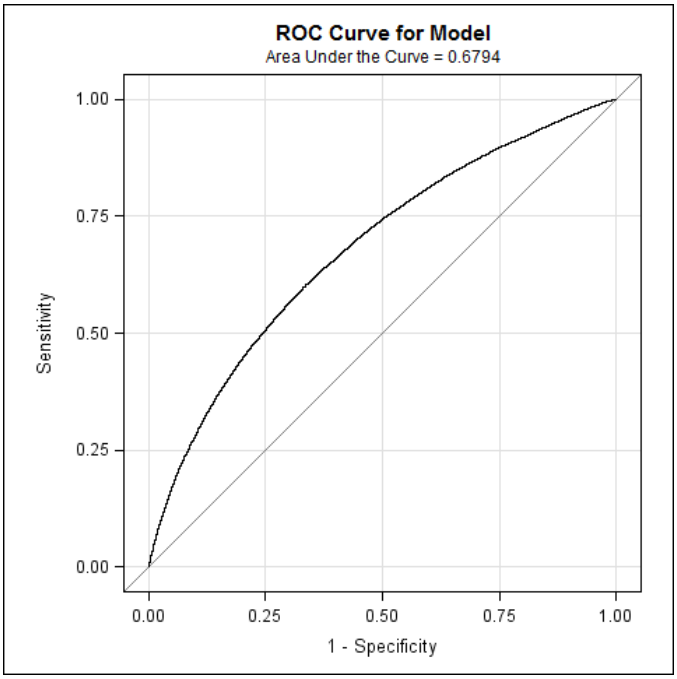

Figure 6: ROC curve for the minimized BIC model.

We do however note that the Brier score looks quite good and that we cannot reject the null hypothesis that the predicted probabilities are equal to the real probabilities. Note that the main contribution to the Brier score here is the Uncertainty. We have both small values for Reliability and Resolution. Hence our models have estimated probabilities close to the outcome but the spread of the probabilities are small. Our models are hence probably quite well calibrated but would perform poorly if used to discriminate.

In Figures 7, 8 and 9 we have plotted estimated probabilities against observed probabilities by grouping the estimated probabilities by rank and then calculating the observed probabilities as the proportion of deaths. It looks like the Purposeful selection model performs worse than the other two for higher probabilities, although they all seem to deviate slightly. We do note however that it seems as if the models perform quite well for lower probabilities.

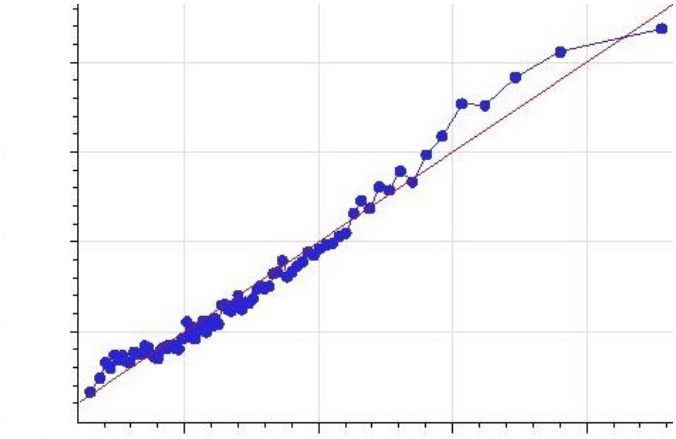

Figure 7: Estimated vs observed probabilities (calculated by 80 groups) for the Purposeful selection model.

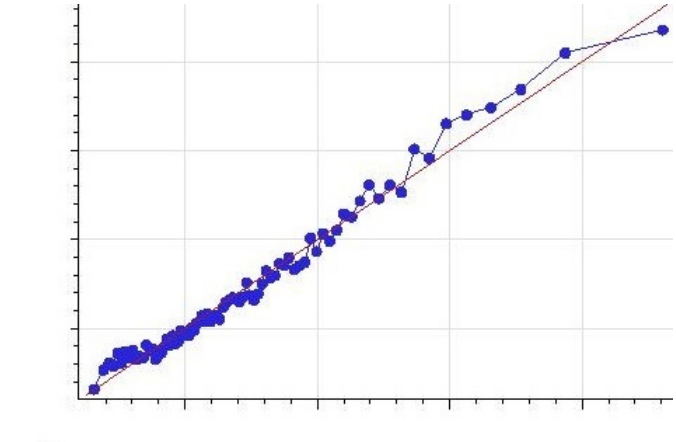

Figure 8: Estimated vs observed probabilities (calculated by 80 groups) for the minimized AIC model.

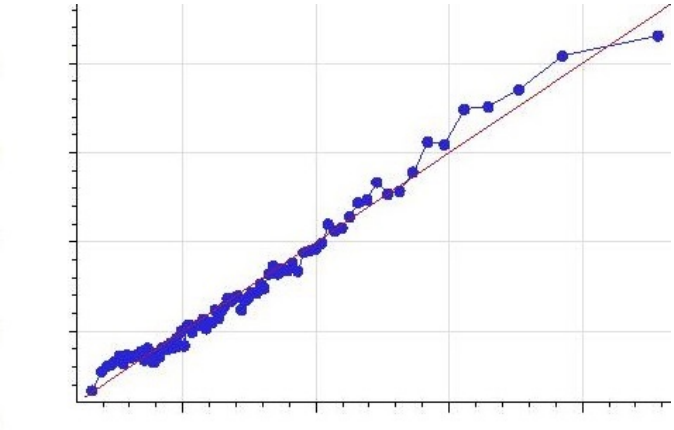

Figure 9: Estimated vs observed probabilities (calculated by 80 groups) for the minimized BIC model.

To conclude, the minimized AIC model has the lowest AIC, a BIC close to the minimized BIC model, the highest AUC and $\tilde{R}^2$ and also the lowest Brier score. Although none of these quantities are much different from the ones for the other models, this model would probably be the one to call the "best".

4.4.1 Validation Using Holdout Samples

Now we try to fit the models to a sample containing years 2007 to 2011 and then we calculate the model validation statistics on the data from year 2012. This will quantify the forecasting capabilities of our models. The results are seen in Table 3.

Table 3: Area under ROC curve (AUC), Hosmer-Lemeshow test statistic, Brier score (BS) and significance test of null hypothesis $H_0 : f_k = p_k$ for the different models. The upper value of each row is the value for the sample containing years 2007 to 2011 (which the models are fitted to) and the lower value is the value for the sample containing year 2012.

Statistic \ Model	Purposeful selection	Minimzed AIC	Minimized BIC
AUC	0.673 0.668	0.674 0.668	0.673 0.671
HL	211.4271 1242.4887	204.7158 392.8128	181.0521 1078.8727
BS	0.13042 0.09590	0.13010 0.09357	0.13056 0.09517
z-statistic	-0.8999 -34.0143	-0.9217 -18.2906	-0.7957 -31.9937
Two-sided p-value	0.3682 ≈ 0	0.3567 ≈ 0	0.4262 ≈ 0

As can be seen, all models have a lower AUC in the holdout sample, although they do not differ by a lot. The Hosmer-Lemeshow test statistic increases a lot for all models, although the minimized AIC model increases by a lot less. The Brier score is lower than expected for all models. That is, the mean squared error is quite small compared to what we would expect. Hence we reject the null that the estimated probabilities are equal to the real probabilities but we do note that the change is in the negative direction.

We conclude that if we had to choose a model for prediction it would probably be the minimized AIC model.

4.5 Interpretation of Coefficient

We now have four models, the three previously mentioned and a univariate model, that we can use to examine whether the coefficient for the interest rate is robust. That is, whether the coefficient has approximately the same value between models. And indeed, this seems to be the case.

Closer inspection tells us that the 95% confidence intervals for the coefficients in the Purposeful selection, minimized AIC and minimized BIC models, all overlap. The coefficient for the univariate model is slightly lower than the others and the 95% confidence interval only overlaps the confidence interval for the coefficient in the minimized AIC model. If we construct 99% confidence intervals, they all overlap.

Exponentiation of the 95% confidence intervals yields confidence intervals for the odds ratio. All confidence intervals upper bounds lie below 1 and hence there is a statistically significant association between the interest rate and customers leaving/staying. We also examine each model without the interest rate. That is, we remove the interest rate from each model and examine what this results in. These results can be seen in Table 4.

Table 4: Area under ROC curve (AUC), Brier score (BS) and significance test of null hypothesis $H_0 : f_k = p_k$ for the different models with the interest rate excluded from the explanatory variables.

Statistic \ Model	Purposeful selection	Minimzed AIC	Minimized BIC
AUC	0.677	0.679	0.678
BS	0.12020	0.12003	0.12022
z-statistic	-0.6920	-0.5673	-0.4765
Two-sided p-value	0.4889	0.5705	0.6337

We see that for each model we get a higher Brier score and a lower AUC. Even if the changes are minuscule, we conclude that the interest rate probably does have a small positive effect on the predictive power of the models. This together with the robustness of the coefficient tells us that the interest rate is important and can safely be interpreted. Since the estimate has been determined to be robust and we this far have considered the minimized AIC model as the "best" we will use the coefficient estimate for the interest rate from that model. Unfortunately we will not further look at the interpretation of this estimate since the results of this thesis are confidential.

5 Conclusions

The purpose of this study was twofold: Investigate whether it was possible to find a logistic regression model with good forecasting abilities and investigate how the interest rate affects the accounts survival over one year intervals. We conclude that the model that (locally) minimizes AIC was the best model both in terms of forecasting abilities, but also in terms of fitting the data. We are however not fully confident in the forecasting abilities of any model presented, this because of the rather large Hosmer-Lemeshow test statistics. Nevertheless, the Brier score and also plots of estimated against observed probabilities seem to indicate that the models are satisfactory, at least in the sense of fitting the data. It could be that the models actually fit quite well, but the large sample size gives the Hosmer-Lemeshow test such power

that any non-perfect model will be discarded.

When investigating how the interest rate affects survival we first noted that the coefficient seemed robust. That is, it was approximately the same over all constructed models. Hence we feel quite sure that the estimated coefficient corresponds well with reality. Since the minimized AIC model was deemed the "best", the coefficient from this model will be used for further interpretation, not discussed in this thesis.

6 Discussion

The biggest time consumer during this thesis has been data management. Before starting, it would have been a good idea to learn how, for instance, `PROC SQL` in SAS software works. This would have saved some time and probably made it possible to do a more thorough analysis, such as building models that take into account the dependency structure of the data or properly estimating the cut points for the linear spline transformations. Nevertheless, this has been a great opportunity to learn about how to assemble a data set and people at SBAB have been very helpful, which I am greatly thankful of. Another thing that has taken some time is actually getting hold of data that was outside of the data warehouse of the bank.

In hindsight it might have been better to model the account balances or perhaps the overall flow of deposits using multivariate linear regression, instead of modeling survival of accounts. Such models might be of more use to the bank since they can be used to calculate the expected loss or gain from increasing or decreasing the interest rate. With a model only considering survival it is more difficult to determine the actual money at stake, and also we have no indication of the flow of new customers and/or money.

In the banking sector, especially in credit scoring, it is not unusual that continuous variables are categorized This is both for ease of interpretation and implementation. Categorizing continuous variables has its share of problems, see Royston et al. (2005) for more information. For instance, the categories are often quite arbitrary and we might have a loss of information. Methods have been developed to estimate "optimal" cut points when categorizing, but these lead to a severe bias, especially the type I error rate will be very high. This might not be a problem since we are dealing with a large sample, hence the power to find well behaving categories should be quite high. Even though categorization might be a solution, we have not implemented it in this report since it is associated, as mentioned, with a bunch of problems.

The biggest problem during this thesis has been figuring out how to decide whether a model has a good fit and forecasting abilities. Remember that we want a model that, while not necessarily perfect, is acceptable. The Hosmer-Lemeshow test discards all models and the question then becomes:

Is this because of the sample size and hence the large power or the possible fact that the models simply cannot predict events in an acceptable fashion. On the other hand, according to the created test using the Brier score, we cannot discard the null hypothesis that the estimated probabilities are equal to the real probabilities. We also have a quite low area under the ROC curve and hence discriminating events from non-events is difficult. As we can see from the decomposition of the Brier score, this probably stems from the fact that there is not a lot of variation in the estimated probabilities. This only means that discrimination is bad, not calibration. We also have the fact that the plots of estimated vs observed probabilities seem to indicate quite good fitting models. If we set aside the Hosmer-Lemeshow test and blame the significance on the large sample size, it does seem as if the models are quite good. On the other hand, when trying to predict later time intervals using models built on earlier time intervals the results are a bit disconcerting. With all of the above we feel that it is difficult to determine whether the models should be used to predict future events.

We should note that it is not really surprising that the models have such a hard time when forecasting. Our set of explanatory variables are quite crude, so to speak, and hence we know very little about the customers and accounts. For instance two of the explanatory variables are age and sex. We cannot really expect these two to explain a whole lot of the variation in the data. It would be nice to have variables such as the customer's salary, what interest rate they had at their previous bank and if they are customers at other banks. It could also be possible to ask customers why they chose SBAB as their bank and then see whether this is a variable that has predictive power. If a customer chose SBAB because of their interest rate they might be more sensitive to price changes than a customer who chose SBAB because they also have a loan there.

A Expected value and variance of the Brier Score

The Brier score is

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 = \frac{1}{N}\sum_{t=1}^{N}\left(f_t^2 - 2f_t o_t + o_t^2\right)$$

where N is the sample size, f_t is the estimated probability of event t and o_t is the outcome of trial t. Under the null hypothesis $H_0 : f_t = p_t$, where p_t is the true probability of event t, the expected value of the Brier score is (note that $o_t \sim Bin(1, f_t)$)

$$\begin{aligned} E[BS] &= \frac{1}{N}\sum_{t=1}^{N}\left(f_t^2 - 2f_t E[o_t] + E[o_t^2]\right) \\ &= \frac{1}{N}\sum_{t=1}^{N}\left(f_t^2 - 2f_t^2 + f_t\right) \\ &= \frac{1}{N}\sum_{t=1}^{N} f_t\left(1 - f_t\right), \end{aligned}$$

and the variance of the Brier score is

$$\begin{aligned} Var(BS) &= Var(\frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2) \\ &= \frac{1}{N^2}\sum_{t=1}^{N} Var((f_t - o_t)^2) \\ &= \frac{1}{N^2}\sum_{t=1}^{N}\sum_{k=0}^{1}\left((f_t - k)^2 - E[(f_t - o_t)^2]\right)^2 f_t^k (1 - f_t)^{1-k} \\ &= \frac{1}{N^2}\sum_{t=1}^{N}\sum_{k=0}^{1}\left((f_t - k)^2 - f_t(1 - f_t)\right)^2 f_t^k (1 - f_t)^{1-k} \\ &= \frac{1}{N^2}\sum_{t=1}^{N}\left(\left((f_t^2 - f_t(1 - f_t)\right)^2 (1 - f_t) + \left((f_t - 1)^2 - f_t(1 - f_t)\right)^2 f_t\right) \\ &= \frac{1}{N^2}\sum_{t=1}^{N}\left(\left(2f_t^2 - f_t\right)^2 (1 - f_t) + \left((1 - f_t)(1 - 2f_t)\right)^2 f_t\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N^2} \sum_{t=1}^{N} \left(\left(4f_t^4 - 4f_t^3 + f_t^2\right)(1 - f_t) + \left((1 - f_t)(1 - 2f_t)\right)^2 f_t \right) \\
&= \frac{1}{N^2} \sum_{t=1}^{N} f_t(1 - f_t) \left(4f_t^3 - 4f_t^2 + f_t + (1 - f_t)(1 - 2f_t)^2\right) \\
&= \frac{1}{N^2} \sum_{t=1}^{N} f_t(1 - f_t) \left(4f_t^3 - 4f_t^2 + f_t + 1 + 4f_t^2 - 4f_t - f_t - 4f_t^3 + 4f_t^2)\right) \\
&= \frac{1}{N^2} \sum_{t=1}^{N} f_t(1 - f_t) \left(1 + 4f_t^2 - 4f_t)\right) \\
&= \frac{1}{N^2} \sum_{t=1}^{N} f_t(1 - f_t)(1 - 2f_t)^2
\end{aligned}
$$

The second equality of the variance calculations comes from the fact that we assume independent observations in a logistic regression model.

References

[1] Agresti, Alan. (2013) *Categorical Data Analysis* - 3rd ed., John Wiley & Sons, Inc.

[2] Allison, Paul D. (1999) *Logistic Regression using the SAS® system: Theory and Application*. Cary, NC: SAS Institute Inc.

[3] Bursac, Z., Gauss, C. H., Williams, D. K. & Hosmer, D. W. (2008) Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine* 2008, 3:17.

[4] Brier, G. W. (1950) Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1), 1-3.

[5] Murphy, A. H. (1973) A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, 12, 595-600.

[6] Hosmer, David W. & Lemeshow, Stanley. (2013) *Applied Logistic Regression* - 3rd ed., John Wiley & Sons, Inc.

[7] Paul, P., Pennell, M. L. & Lemeshow, S. (2013) Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Statistics in medicine*, 32: 67-80.

[8] Royston, P., Altman, D. G. & Sauerbrei, W. (2005) Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, 25: 127-141.

[9] SAS Institute Inc. (2011) *SAS/STAT® 9.3 User's Guide*. Cary, NC: SAS Institute Inc.