School of Mathematics & Statistics

Assigning reasons for decisions based on model scores

Student:
Amy Hetherington

Supervisor:
Dr. Philip Ansell

2013-2014

Abstract

The aim of the project is to explore methods into assigning reasons to a "bad" credit score. This exploration will give credit risk companies the ability to give their customers a specific reason as to why they receive a certain treatment. Using modelling techniques we will be able to predict if an account holder will miss a payment in the next three months.

Contents

Chapter 1

Introduction

Credit risk scoring is a method used to determine the creditworthiness of an account holder. Decisions made in credit risk are typically based upon scores provided by scorecards. Scorecards are statistical models built using a large number of variables. These variables are chosen from internal company data and also credit bureau information. The risk scores calculated then drive a certain action, depending on the models cut-off point. Generally, higher valued scores suggest a more creditworthy account. The models used in credit risk companies have a large number of variables and this makes it difficult to assign a score to a particular reason. The objective is to explore methods into assigning reasons to a "bad" score such that credit risk companies can explain to their customers why they receive a certain treatment. With the use of modelling techniques we will be able to predict if a customer will miss a payment in the next three months. From the model(s) found an optimal cut-off point will be determined which will provide the threshold used to define accounts as accepted or declined. The model(s) found will represent a short-term risk model.

1.1 Credit risk scoring

In order to define credit risk scoring, we must first look at the two components separately; "credit" and "scoring", using the definitions found in Anderson [2007].

1.1.1 Credit

Credit is a term used at present to describe the analogy "buy now, pay later". It originates from the Latin word 'credo', which has the meaning 'trust in'

or 'rely on'. Credit can be thought of as lending to an individual whilst trusting them to honour the obligation. Borrowers' must abide by terms to ensure this trust is upheld, for example agreeing to a risk premium. This provides security for the lender in case the borrower does not repay. This is where the term credit risk arises. Credit risk is the potential financial impact, on the lender, when a change in the borrowers' creditworthiness occurs, i.e. their ability to repay. When the credit risk is high, lenders typically increase charges or premiums to counterbalance the risk. However, due to lenders current vast data collection trust can be built upon using the borrowers' financial details. Thus, the risk of lending is lowered.

1.1.2 Scoring

Scoring is a ranking order tool. This ordering occurs according to a real or regarded quality used to discriminate between the ranks and make impartial, consistent decisions. For example, companies can be ranked based upon their performance to help decide which company to sell. The data available is collected and incorporated into a single value representing quality. This is thought of as a score. Scoring is universally used where predictions are needed. Predictive scoring models use prior data and past events to predict the likelihood of a future event occurring.

1.1.3 Credit scoring

Credit scoring involves the transformation of collected data into numerical measures, using statistical models. These models influence credit decisions. Credit scoring helps to determine whether borrowers' repay the lender. Credit risk scoring used in industry, particularly finance, classes an account "good" or "bad" based upon their credit report. This determination is used to filter if an account holder poses a risk to lend to. If an account is deemed "good" then the credit score is used by the lender to decipher if they qualify for the loan, what interest rate they should be charged and what credit limits should be set. Dependent on the level of "bad" score the account could be rejected.

1.2 Decision making

Lenders make decisions for different scenarios based upon credit scores. The subsequent action that occurs is also dependent on the score, for example accept/reject a customers request for an increase in loan amount, interest

rate lowered etc. Credit scores can be categorised into different types of score, with the most common scores used being; [Anderson, 2007]

- **Application score**: Used in new businesses. It combines data from the customer, past dealings and the credit bureau.
- **Behavioural score**: Used in account management, e.g. limit setting. It focuses on individual accounts' behaviour.
- **Collection score**: Used in collection processes, e.g. call centres. It combines behavioural, collections and bureau data.
- **Customer score**: Combines behaviour on a number of accounts. It is used for both account management and cross-sales to existing customers.
- **Bureau score**: A score given by the credit bureau. It is usually a number of missed payments or bankruptcy predictor that summarises the data held by them.

These definitions of scores describe the borrowers'/customer behaviour. A high number of lenders use a hybrid of their own scores and bureau scores. Credit scoring is used in fields such as: [Anderson, 2007]

- **Unsecured**: credit cards, personal loans, overdrafts.
- **Secured**: Home loan mortgages, motor vehicle finance.
- **Store credit**: Clothing, furniture, mail order.
- **Service provision**: Phone contracts, municipal accounts, short-term insurance.
- **Enterprise lending**: Working-capital loans, trade credit.

We will focus on unsecured credit scoring, particularly credit cards. Credit scores trustworthiness varies according to data used in the development of them. Decisions, in credit scoring, tend to be made using a scorecard. A scorecard is a statistical model derived, using a large number of variables, to predict the likelihood of an event occurring. These variables are chosen from internally collected data and also credit bureau information. An example of a decision made using a scorecard would be; "Do we accept this account for a loan request?". Custom-made scorecards can be adapted for a lender or product to produce the best results. If this is not feasible, generic scorecards can be applied. The vast number and diversity of the variables used in the development of scorecards creates a problem when assigning reasoning to a score.

Chapter 2

Data

The data collected by credit risk companies, both internally and externally, make up the variables used in their modelling techniques. Companies usually have around 8000 different variables available to create their models. The data referred to throughout this document is a real dataset provided by a financial company. The entries involved represents data from a 6 month statement. There are variables available for each statement cycle and collective data for the entire 6 months. The data is comprised of 34 000 accounts and 155 different variables. These 155 variables can be split into four categories:

1. Current general account information e.g. account age
2. Raw information for the last 6 months e.g. average daily balance on the account
3. Transformations of the raw data e.g. average balance on the account over the last 6 months
4. Application variables. These are variables only available for accounts less than 12 months old and are taken from the initial application process.

The dataset contains categoric, continuous and binary variables. The response variable that will be estimated with modelling techniques is a binary variable. It is named *outcome* and represents if an account holder misses a payment in the next three months. The outcome value of 0 represents the account holder not missing a payment and 1 is that they do. An example of a categoric variable is their employment status. The variable which represents employment, *employ*, can be one of seven characters defined below.

- EM: employed

- SE: self employed
- HO: home maker
- RE: retired
- ST: student
- UN: unemployed
- O: other

The employment status of an account holder is only available for accounts less than 12 months old, as it is classed as an application variable. A number of variables will be used and referred to throughout. To ensure understanding a list of them and their meanings are given below.

- *outcome*: The response variable used to estimate the data. It represents if an account holder misses a payment within the next three months.
- *delinq*6: Delinquency value over the 6 month period. Delinquency is the total number of missed payments over a given period.
- *late*6: The total number of late fees over the 6 month period.
- *delinq*1: Delinquency value for the first period/month.
- *over*6: The number of over-limit fees over the 6 month period.
- *age*: Age of the account in months.
- *young*: An indicator variable to show an account who is less than 6 months old. If the account is younger than 6 months old then it takes a value 1 and 0 otherwise.
- *employ*: The employment status for accounts it is available for. The status' it can take were defined previously.
- $fee.bal3$: The fee balance of the account at the end of the statement cycle 3, i.e. month 3.

2.1 Missing data

When using a real dataset numerous procedures need to be implemented before necessary analysis can take place. This includes dealing with missing data. In the dataset some of the entries are missing, which is not uncommon due to the raw nature of the data. Appropriate analysis cannot be performed accurately when entries for variables are missing. A possible solution could be to delete any observations where there are missing data entries, however this should be considered as a final expedient for a number of reasons. We may discard important information which might be useful in future explorations and introduce bias. Misleading interpretation of results may also be made as a consequence of deleting observations. Bayesian techniques, like data augmentation, was considered to account for the missing data however the missing entries can be explained when looking carefully at the variables to which they were missing for. Had the data been missing at random (MAR) then data augmentation would have been an appropriate method to use. Generally, the missing data is for an account that has an age of only 1 month. This is because many of the variables depend on a previous month or cycle in their calculation. Looking at Equation (2.1) the calculation of Payment Percent 1 depends upon the previous month. Payment Percent 1 represents the payment made as percent of the total balance from the previous statement. If the account is only 1 month old, this information is not available.

$$\text{Payment Percent (1)} = \frac{\text{Payment Percent (0)}}{\text{Total Balance (1)}} \tag{2.1}$$

Another explanation for missing data is due to division by zero in the calculation of variables. Again, looking at Equation (2.1), if Total Balance 1 is zero then NA would be produced in the calculation.

The solution of missing data is dealt with by imputation. Missing data imputation is a way of taking missing data and replacing it with an appropriate value. There are many ways to impute data, with varying degrees of complexity. The imputation method choice is a fairly simplistic one due to the small proportion of missing data and clear reasoning as to why it is missing. Had there been a larger number of missing entries, then a different imputation method would have been considered. The replication value is dependent on the type of data it is. If the missing value is continuous then it is replaced with the mean of that variable and if it is categoric data then another level is created and the missing values are put into this category. This means standard credit scoring procedures can now be followed to model the data.

2.2 Data behaviour

In credit risk scoring it is common to split the data into manageable segments. This grouping of data, i.e. segmentation, is usually done by looking at its behaviour and intuitively choosing the grouping. When looking at credit risk assessment the purposes of segmentation is to improve assessment and ensure customers are offered the appropriate product, e.g. correct loan rate for the level of risk they pose. The number of segments chosen should always be kept to a minimum due to the increased risk of complication when building multiple models. The validation and analysis of the models built also increases, which could be costly for the company. When looking at the dataset it is evident that a high proportion of account holders can intuitively be classed as "good" accounts. The total of missed payments over the 6 month period, i.e. the delinquency value, is a good example. Figure 2.1 depicts this behaviour well.

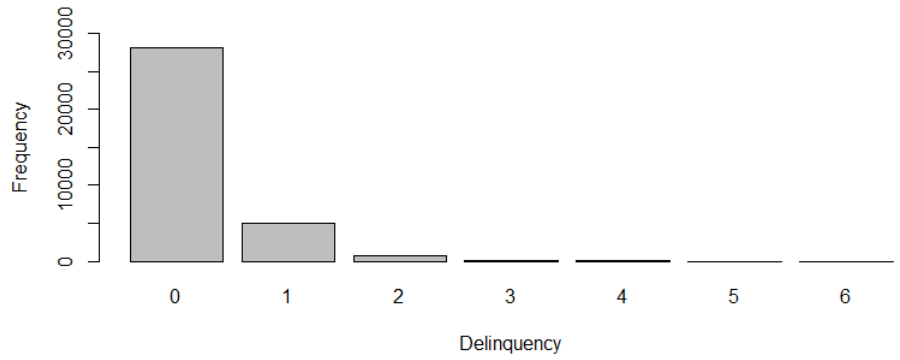

Figure 2.1: Delinquency values over 6 month period

A high number of account holders never miss a payment over the time period, so it is acceptable to assume that these account holders behave similarly. With this assumption in mind, the data is segmented with accounts who have a total delinquency value of zero grouped together and then the rest grouped together. To see if this segmentation is beneficial "quick" models are built and then segmented and non segmented models are compared. Not segmenting the data is the better option despite Figure 2.1 demonstrating a divide in the data. The choice to segment or not is based upon looking at the mean square error for the models built. The mean square error is a technique used in statistical modelling to determine how well the model fits

the data. It can be used to determine if the model needs simplifying. It is a quick way to compare models when segmenting. A lower MSE value is better as it implies a better model fit to the data. When the data is grouped by total delinquency value being zero the mean square error is slightly smaller compared to the full sized dataset. However the remaining group, of total delinquency value 1-6, has a much larger mean square error. In this instance segmenting the data is not greatly beneficial when model building. To ensure the chosen grouping size isn't too small, the data is split with one group having a total delinquency value 0 or 1 and then total value 2-6. This produces even greater mean square errors than previously. The non segmented model is a better choice and is graphically shown by ROC curves which will be discussed later in §3.3.2.

2.3 Variable reduction

The number of variables is too large to perform analysis straight away. To reduce the number of variables two methods will be explored; an ad-hoc basis and decision trees. Other methods used by credit risk companies include principal component analysis (PCA) and stochastic gradient boosting. Decision trees are the preferred method over PCA because it is believed to be more time efficient. Both methods are universally used to reduce the number of variables. Intuition based variable reduction is used as a comparative baseline to the decision tree output. The intuitive chosen variables do give a significant model, however it is not practical to use this method. The decision tree method can easily be adapted for other datasets, using a pre-built `R` package. To reduce the number of variables an initial model is constructed, using all 155 variables, to use in the R package `rpart`. This package produces the decision tree, Figure 2.2, which shows only two variables classed as important within the model and how they can be classified into groups. For example the first node has two branches, left and right. The path given by the left branch is for accounts who have a delinquency value of less than 0.5, classed by group 0. The right path takes us to a decision node where the question "is the total number of late fees less than 1.5?" is posed. If so the account is classed as group 0 again, otherwise the account is grouped as 1. The package also prints the variable importance determined from the saturated model. The output printed by `R` is stated in Figure 2.3. The variables classed as important, by the variable reduction, can now be used in regression analysis to find an appropriate model.

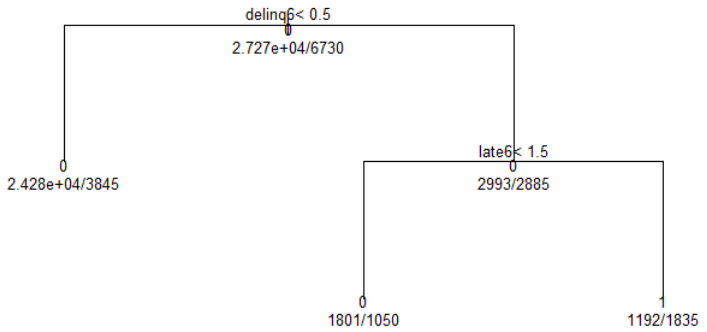

Figure 2.2: Decision tree

```
Variable importance
delinq6
late6
flate.ind
late2
delinq1
late3
late4
late1
fee.bal3
employ
cashadv.ind
tenure
```

Figure 2.3: Variable importance output

Chapter 3

Regression analysis

Regression analysis is used to build models and find the most parsimonious model to estimate the relationship between the response variable and explanatory variables (covariates). Box et al. [2013] suggested that the 'principle of parsimony' gives rise to a model with the smallest possible number of parameters for adequate representation of the data. Parsimony is a desirable trait when credit risk modelling as it reduces cost, keeps interpretations simplistic and the errors small. The response variable may be the likeliness of an event occurring and the covariates are associated variables which either add or deter from the likeliness of the event.

3.1 Linear regression

In linear regression the response variable modelled is assumed to be continuous. For example, the response variable might be the risk of a patient having a heart attack and the explanatory variables may include age, blood pressure and genetic issues. Once a model is built, the explanatory variables are used to predict a person's risk of having a heart attack. Regression analysis estimates the conditional expectation of the response variable given the independent variables. Denoting the response variable, Y_i, and the explanatory variables $X_{1,i}, X_{2,i}, \ldots, X_{m,i}$ the conditional expectation can be expressed by Equation (3.1) where it is linear in β_k, for k in $0, 1, \ldots, m$, where m is the number of covariates.

$$E(Y_i|X_{1,i} = x_{1,i}, \ldots, X_{m,i} = x_{m,i}) = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_m x_{m,i} \tag{3.1}$$

The error of the model is a random variable and a classical assumption for linear regression analysis is that the errors are distributed Normally with mean zero and variance which is constant. This normality assumption is checked

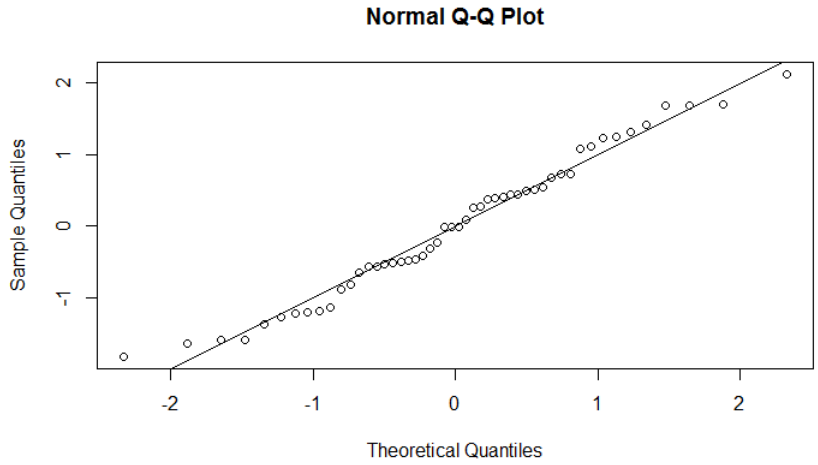

Figure 3.1: Example of a QQ plot

using Q-Q plots. An example of a Q-Q plot that satisfies this assumption is shown in Figure 3.1.

Using linear regression to model the response variable, *outcome*, gives the diagnostic plots in Figures 3.2 and 3.3. From both plots we see that the linear regression model is not a good model choice. The main issues with the diagnostic plots are the errors are not Normally distributed and the fitted values cannot be derived, as the response can only take the values 0 or 1.

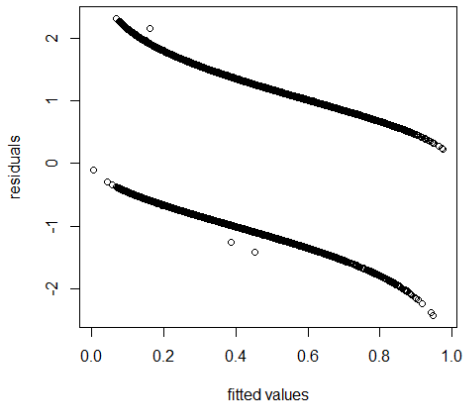

Figure 3.2: Residual plots

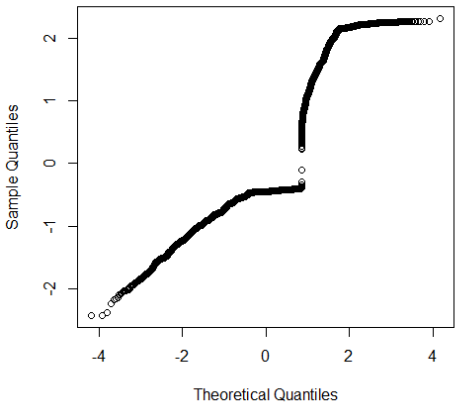

Figure 3.3: Q-Q plot for residuals

These issues are due to the nature of the response variable being binary. Therefore a better suited distribution for binary data needs to be found, as the Normal is not a good fit.

3.2 Logistic regression

A logistic regression model is differentiable from a linear regression model when the response variable is binary or dichotomous as opposed to Gaussian. [Hosmer Jr and Lemeshow, 2004] When the data being modelled is Binomially distributed, the response variable takes one of two values characterising success and failure. The response variable modelled throughout, named *outcome*, is whether an account holder misses a payment within the next three months. The outcome is either yes or no, indicated by 1 or 0 respectively. If we let the response variable, outcome, be denoted by Y_i, then it is classed as a Bernoulli variable with parameter π_i, or alternatively a Binomial variable;

$$Y_i \sim Bin(m_i, \pi_i), \tag{3.2}$$

with $m_i = 1$ and where π_i is the probability of an account holder missing a payment in the next three months for $i = 1, 2, \ldots, n$, where n is the number of accounts. Therefore, $E(Y_i) = \pi_i$ and $Var(Y_i) = \pi_i(1 - \pi_i)$. In linear regression analysis the conditional expectation is expressed by Equation (3.1) however this model is unsuitable when modelling binary data. The problem arises due to the right hand side of Equation (3.1) being able to take any value, where as $E(Y_i|\mathbf{X}_i) = \pi_i$ is a probability thus restricting the values to $[0, 1]$. The solution is to find a function, denoted h, which maps the real line to $[0, 1]$ i.e. $h : \mathbb{R} \to [0, 1]$, such that

$$E(Y_i|\mathbf{X}_i) = \pi_i = h(\beta_0 + \beta_1 x_{1,i} + \ldots + \beta_m x_{m,i}) \tag{3.3}$$

or likewise

$$g(\mu_i) = g(\pi_i) = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_m x_{m,i} \tag{3.4}$$

where $\mu_i = E(Y_i|\mathbf{X}_i) = \pi_i$ and $g(\cdot) = h^{-1}(\cdot)$ is the inverse function of $h(\cdot)$. The function $g(\cdot)$ is called the $link function$. The role of the link function is to allow the response variable to be used whilst associating it with the linear model. For data which is distributed Binomially, a prominent choice of link function is derived from Equation (3.5) giving Equation (3.6), where $\eta = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_m x_{m,i}$ is called the linear predictor.

$$h(\eta) = \frac{exp(\eta)}{1 + exp(\eta)} \tag{3.5}$$

$$g(\mu) = h^{-1}(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \tag{3.6}$$

The link function given by Equation (3.6) is called the logit link function, and hence the model is called the logistic model. To estimate the unknown parameters $\beta_0, \beta_1, \ldots, \beta_N$ we use maximum likelihood estimation. However, these estimates cannot be found analytically and this is why `R` is used. To model the response variable, *outcome*, modelling techniques are implemented in `R` using the `glm` command. This command carries out inference for Generalized Linear Models (GLMs) due to logistic models being a type of GLM. A typical `glm` model used in `R` has a standard format, shown in Equation (3.7).

$$glm(formula, family = (), data) \tag{3.7}$$

For this dataset the family used is Binomial due to the response variable modelled being a probability of success or failure. The chosen link function we will be using to relate models to the linear model is logit, where μ is the probability of an account holder missing a payment in the next three months. Logistic regression is a modelling choice used in 80 to 90% of credit companies to build and develop scorecards. [Anderson, 2007]

3.2.1 Model selection

We will be using the variables that are classed important in the decision tree analysis, listed in Figure 2.3, to perform forward step-wise regression. This is a favourable method when the number of variables trying to be modelled is large. Forward step-wise regression involves initially modelling with no variables and testing the significance after each variable addition. This process is repeated until no more additions improve the model. Analysis of Variance methods and Akaike information criterion will be analysed after a variable is added in order to select a favourable model. In credit scoring a high predictive model is favoured over probabilistic power, due to rank ordering being classed as a more important model property.

Kullback-Leibler information (K-L information)

The Kullback-Leibler information is a way to measure the difference between two models. It is particularly useful to look at how close an approximating model, g, is to the true model, f. The K-L information between f and g is defined for discrete distributions by Equation (3.8),

$$I(f, g) = \sum_{i=1}^{k} p_i \cdot \log\left(\frac{p_i}{\pi_i}\right), \tag{3.8}$$

where k are the possible outcomes, p_i is the true probability of the ith outcome and $\pi_1, \pi_2, \ldots, \pi_k$ represents the approximating model. For discrete distributions, both p_i and π_i lie in the range $(0, 1)$ and represent f and g respectively. The information lost when g is used to estimate f is denoted by $I(f, g)$ in Equation (3.8). The Kullback-Leibler information is only available when the true model, f, is known. [Burnham and Anderson, 2002]

Akaike information criterion (AIC)

The Akaike information criterion (AIC) is way of looking at information lost when a particular model is used by estimating the Kullback-Leibler information (K-L information). [Posada and Buckley, 2004] It is an estimate of the K-L information because the true model, f, is unknown. The AIC is only classed a valid estimate asymptotically, i.e. for a large sample. The AIC is thought of as information lost when models are used, hence the lower the AIC value computed the better the model is considered. However, the Akaike information criterion cannot be solely regarded in model selection as it is not a qualitative test. In theory if only the AIC was observed, the best model found may have a low AIC value but perhaps not be a good fit to the data. The Akaike information criterion is defined by Equation 3.9, where l is the log-likelihood of the model and p is the number of estimable parameters.

$$AIC = -2l + 2p \tag{3.9}$$

Burnham and Anderson [2002] said Equation (3.9) can be thought of as a trade off between bias and variance or a trade off between over-fitting and under-fitting. The term $-2l$ decreases as more estimable parameters are added to the model and $2k$ increases as more parameters are added which helps to avoid over-fitting.

Analysis of Variance (ANOVA)

Analysis of Variance is a way to analyse models and perform statistical hypothesis testing. It is readily used to make decisions, especially in model selection. Let the current model found, with q covariates, be denoted Model C and the current model with a new variable added be denoted, Model C_A. Model C_A has p covariates such that $q < p$. Therefore Model C is classed as a nested model of Model C_A, thus the likelihood ratio test can be performed to calculate the Deviance, D^*. The deviance is a measure of how close the model is to a perfect fit and has an asymptotic chi-squared distribution. The difference in deviance between Model C and Model C_A, $D^*_C - D^*_{C_A}$, is used to compare the two models, i.e. testing H_0: Model C versus H_1: Model C_A

and forming an Analysis of Deviance table. H_0 is rejected if the p-value is small, i.e. $p \leq 0.05$, where

$$p\text{-value} = Pr(\chi^2_{p-q} > D^*_C - D^*_{C_A}).$$

Thus, to compare two models in `R` an Analysis of Deviance table is obtained using the `anova` function, where `test="Chisq"` is specified. A variable added to the current model is classed insignificant if it has a p-value≥ 0.05. At each step of regression performed, the significance of the variable added is checked using Analysis of Variance (ANOVA). The variables found significant gives the model in Equation (3.10).

$$\begin{aligned} outcome \sim & delinq6 + late6 + flate.ind + delinq1 + late1 + over6 + age + young \\ & + fee.bal3 + employ + cashadv.ind + tenure \end{aligned} \tag{3.10}$$

Now that the single terms are found to be significant, the interaction terms between the variables are added. Similarly, the interaction terms are removed in order of significance. The terms with the highest p-value are removed and the model is then fitted again. This method is repeated until a final model is found, where all terms are found to be significant. This model is represented by Equation (3.11).

$$\begin{aligned} outcome \sim & delinq6 + late6 + delinq1 + over6 + age + young + employ \\ & + fee.bal3 + delinq6 * late6 + delinq6 * over6 \\ & + age * young + late6 * over6 + delinq6 * age \end{aligned} \tag{3.11}$$

3.3 Model adequacy

In credit risk scoring, companies want to know the predictive power a model has. To test how powerful a model is at predicting if an account holder will miss a payment we will look at two things; the Somers' D value it yields and the area under an ROC curve.

3.3.1 Somers' D

Somers' D is used by credit risk companies to determine the discriminative strength of a model. It is a key aspect used in rank statistics. Somers' D is usually defined in terms of Kendall's tau coefficient, τ_a, [Kendall and Gibbons, 1990] which tests the strength of association between two ordinal variables.

Definition 1 (Kendall's τ_a coefficient)**.** *Kendall's τ_a coefficient is defined as,*

$$\tau_a = \frac{n_c - n_d}{n_0}, \qquad -1 < \tau_a < 1 \tag{3.12}$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs and n_0 is total number of pair combination given by $n(n-1)/2$. For a set of joint random variables X and Y, the set of observations is given by $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Any pair of variables, $(x_i, y_i)$and (x_j, y_j), are said to be concordant if both $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$ is true. Similarly, they are classed as discordant if $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$. There is said to be no association between the pair if $x_i = x_j$ or $y_i = y_j$.

Kendall's τ_a can be redefined as

$$\tau_{XY} = E[sign(X_1 - X_2)sign(Y_1 - Y_2)],$$

where (X_1, Y_1) and (X_2, Y_2) are bivariate random variables independently sampled from a population. This is the difference between the probability of two pairs X, Y being concordant and discordant. Thus Somers' D is defined by Newson [2006], given in Equation (3.13), where D_{YX} is the difference between the two conditional probabilities, τ_{XY} and τ_{XX}, given that the two X values sampled are different.

$$D_{YX} = \frac{\tau_{XY}}{\tau_{XX}} \tag{3.13}$$

It looks at the strength of a relationship between pairs of variables. In terms of Somers' D a value of -1 means all the pairs completely disagree and 1 they completely agree. It can also be expressed in terms of Harrell's c index, such that $D = 2c - 1$. Somers' D is sensitive to event rates, meaning a segmented model cannot be compared to a model built using the full sized dataset. The value of Somers' D is calculated in R using the `Hmisc` package. It calculates the rank correlation between the fitted values for the model and the binary response variable. This gives a Somers' D value of 0.4442 which is an acceptable value in terms of model association strength.

3.3.2 ROC curves

Receiver operating characteristic curves (ROC) were originally used in statistical decision making. Objects were assumed to belong to a known category and based upon the information on the objects they were ideally assigned to the correct category. Examples of two-group classification tasks include:

- determining incoming emails as spam or not,
- diagnosing of a patient having a particular disease,

- analysing credit card expenditures to decide if its fraudulent or genuine behaviour.

Credit risk companies use receiver operating characteristic curves to determine whether an account holder will miss a payment in the next three months. They can also be used to test the discriminative power of models found. An ROC curve is useful in the analysis of logistic models found because it demonstrates the performance of a binary classification system. It plots sensitivity vs 1-sensitivity where they are defined by Equation (3.14) and (3.15) respectively. These definitions arose from signal detection theory where the aim was to detect a signal and assign each event into the signal or noise group.

$$\text{sensitivity} = \text{probability of detecting true signal} \tag{3.14}$$

$$\text{specificity} = 1 - \text{sensitivity} = \text{probability of detecting false signal} \tag{3.15}$$

According to Anderson [2007], the sensitivity, S_{TP}, and specificity, S_{FP}, can be thought of more generally as the ability to mark true positives and the ability to identify true negatives respectively. The ROC curve is a plot of X versus Y, defined by Equations (3.16) and (3.17), where the chosen cut-off is varied.

$$X = \Pr[S_{FP} \leq S_{\text{cut-off}}] \tag{3.16}$$

$$Y = \Pr[S_{TP} \leq S_{\text{cut-off}}] \tag{3.17}$$

To test the discriminative power of a model, the area under a receiver operating characteristic curve (AUROC), or the c-statistic is analysed, stated by Equation (3.18). The equation states that the area under the curve is equal to the probability that the true positive rate is less than the true negative rate, with the addition of 50% of the probability that the two rates are equal.

$$\text{AUROC}_{c_{P,N}} = \Pr[S_{TP} < S_{TN}] + 0.5\Pr[S_{TP} = S_{TN}] \tag{3.18}$$

The idealistic area under the curve for a model found is 1. As the curve tends to an area of 1, the more powerful the model is at correctly classifying if an account holder will miss a payment or not. An area under the ROC curve of 0.5 implies the model is like making a random guess at classifying. Figure 3.4 depicts a number of ROC curves. The red dotted curve shows the "perfect" ROC curve, due to its AUROC being 1. The green line is classed as the least powerful classifying model and as the lines tend to the red dotted curve they increase in discriminating power.

The logit model obtained, given by Equation (3.11), is an estimation of the data and will not correctly classify 100%. Thus an acceptable range for the area under the ROC curve is considered to be between 0.5 and 1.

Using the R package, `ROCR`, to plot an ROC curve for the chosen model gives an area of 0.7221. This value lies in the acceptable range and implies the model found is a fairly powerful discriminator. The subsequent plot, shown in Figure 3.5, depicts the ROC curve produced for the logit model obtained. The red dotted curve shows a perfect discriminating model whose AUROC is equal to 1. The black line is the ROC of the model found. The y-axis of the graph, true positive rate, is defined by Equation (3.14) and similarly the x-axis, false positive rate is defined by Equation (3.15). Figure 3.6, shows graphically the strength of the segmented model and the non segmented model. It is clear that the choice not to segment the data is a good one. The curve which depicts the non segmented model is more powerful as it has an area under the ROC greater than the segmented curve.

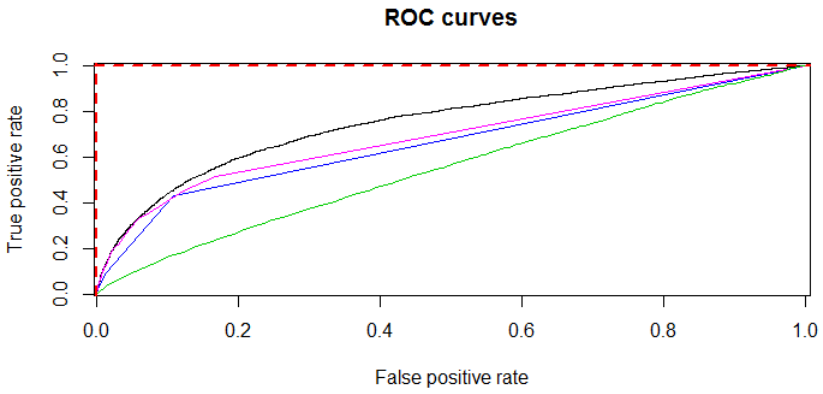

Figure 3.4: ROC curve examples

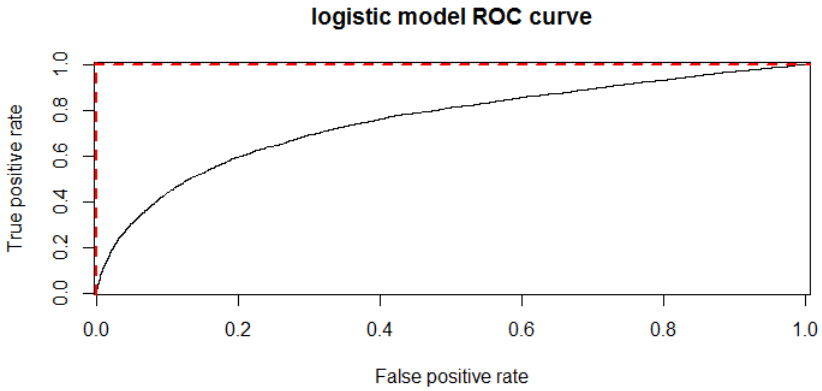

Figure 3.5: ROC curve for model found

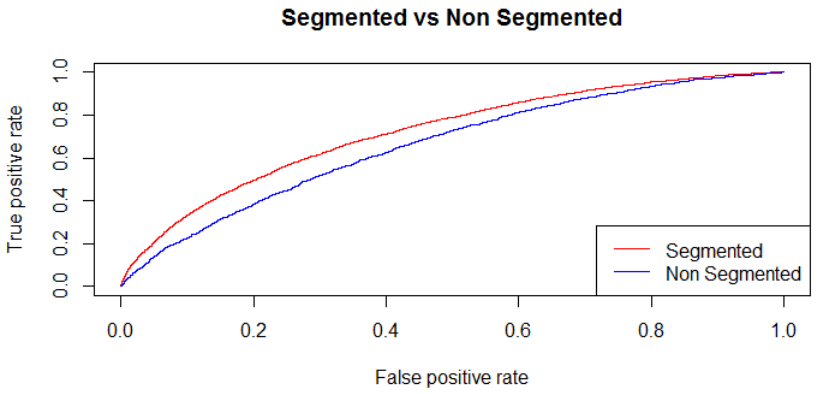

Figure 3.6: Segmented model versus non segmented model

Chapter 4

Defining an account

From the logistic model found, the outcome will be transformed using the logit link function, expressed in Equation (3.6), to calculate the predicted probability of missing a payment. Thus the probability of missing a payment is given by

$$p = \frac{\exp(outcome)}{1 + \exp(outcome)}. \tag{4.1}$$

By developing an R function we can take each individual and calculate the probabilities from the transformed link function. From such probabilities a cut-off point can be chosen, which determines if an account holder defaults on a payment. The chosen cut-off point should not be too low such that account holders are declined when they should not have been. Alternatively, the cut-off point should not be too high to ensure no accounts are classified as declined. Either situation would not be financially beneficial for the credit company. An optimal cut-off point will be chosen by considering cost-benefit analysis.

This cut-off point is found by looking at a number of plots. The cut-off values in Figures 4.1, 4.2 and 4.3 represent $1-p$, where p is the chosen cut-off point in the `R` function. Thus, if an account has a probability of missing a payment in the next three months greater than p it is classed as defaulting. The `R` package `ROCR` produces Figure 4.1, which shows the sensitivity and the specificity versus cut-off values. The sensitivity is thought of as the true positive rate, *tpr*, at which accounts are correctly classified and alternatively the specificity is the false positive rate, *fpr*, that represents the rate accounts are misclassified. The point at which the curves intersect is a good cut-off point as it is where the rate at which accounts are correctly classified as defaults is at its maximum and the rate at which accounts are misclassified is minimised, simultaneously. To reduce financial cost, misclassification will tried to be avoided.

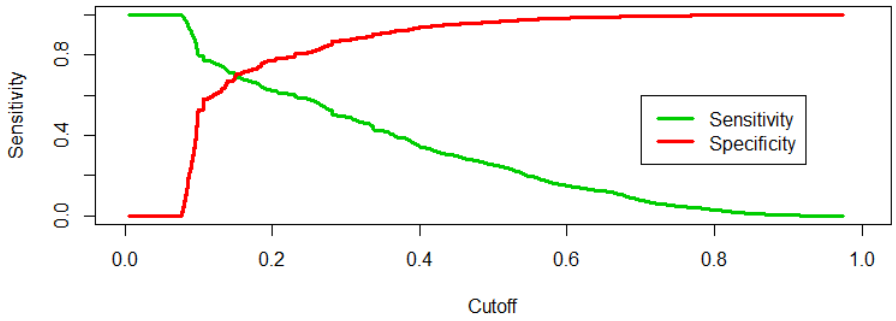

Figure 4.1: True and false positive rate vs cut-off

Another useful plot is the ROC curve for the model colourised according to the cut-off value, shown by Figure 4.2. The gradient colour legend represents the change in cut-off values. This is a useful plot as it shows that as the cut-off point is lowered both the true and false positive rate increases. As the false positive rate is increased to its maximum, almost all accounts will be deemed defaults, thus the true positive rate is maximised. Alternatively, if the cut-off is too high then very few accounts are classified as defaults, resulting in the *tpr* and *fpr* to become negligible. The choice of cut-off value should be a point in Figure 4.2 where the percentage increase in true positive rate is much greater than that of the false positive rate. As the cut-off value in the plot decreases the relationship between the two rates becomes almost linear. A lower cut-off value implies a high percentage of the sample classified as defaulting accounts, which gives effect to the true positive rate increasing. If a larger percentage of the sample is classed as defaulting the rate at which accounts are misclassified will also increase alongside the *tpr*. If we look at a cut-off value around 0.25 in Figure 4.2, i.e. $p \geq 0.75$, where accounts are classed as defaults in `R` the change in true positive rate is much greater than the false positive rate which is desirable. However, the change in true positive rate should not completely outweigh the false positive rate because it would not be economically viable for credit companies. There should still be a margin in which the company can profit whilst minimising wrongful classification.

The accuracy of the model, in terms of discriminative strength, is also calculated and plotted. The accuracy of the final model is plotted against possible cut-off values shown by Figure 4.3. The accuracy is shown to be

maximised around a cut-off value of 0.3. Therefore, if the probability of an account holder missing a payment is greater than 0.7, the accuracy of being correctly classified should be maximum.

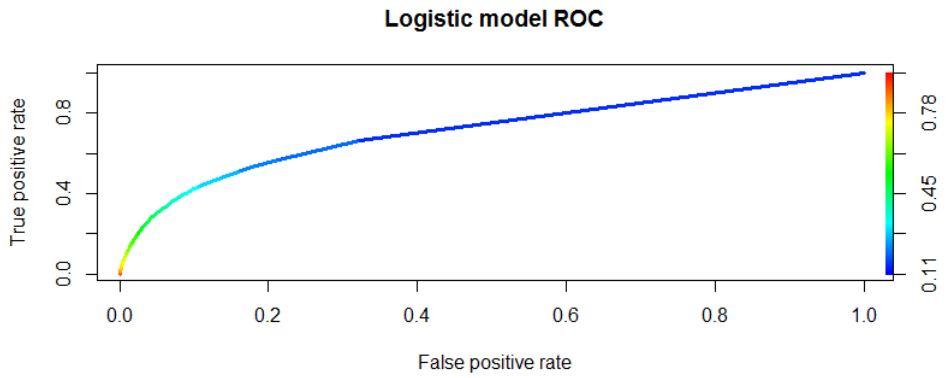

Figure 4.2: Colourised ROC curve

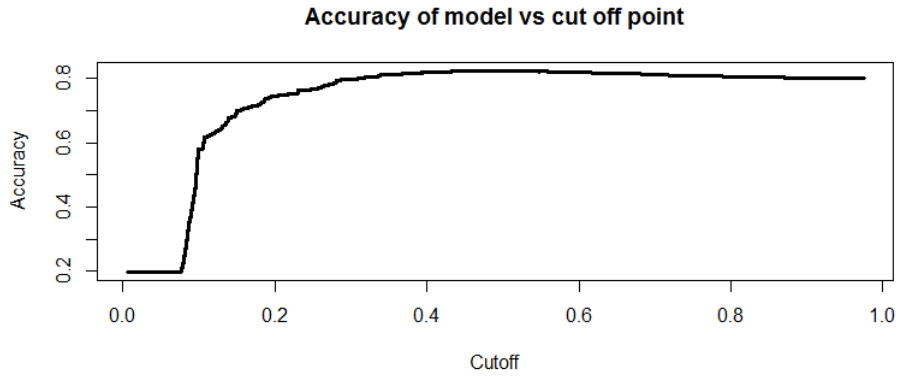

Figure 4.3: Accuracy versus cut-off

Chapter 5

Decline reasoning

We now test the logistic model found, expressed in Equation (3.11), to an acceptable predictive power. Using this model the estimated probability of missing a payment in the next three months is calculated for all accounts. Analysis is implemented to find an optimal cut-off probability. This cut-off point is where accounts are defined as one of two categories; default or non-default. If they are default they are statistically thought to miss a payment in the next three months. Once this categorisation has occurred, decline reasoning will be produced. This decline reasoning gives the customer an idea as to why they were treated a certain way. For example why they were declined for a certain action dependent on them being classed as defaulting. This decline reasoning is a legal requirement for credit companies in the United States, but in the UK it is currently optional.

5.1 Decline reasoning for whole data

To predict a chosen number of reasons, k, that contribute to decline reasoning for the whole dataset an `R` function needs to be created. It will predict the coefficients of the variables in the final logistic model, and sort the coefficient values in order of importance. It will then match the coefficient value to the variable and give the top k variables. The number one reason, for decline, changes dependent upon the cut-off point chosen in §4. This change in decline reason for different cut-off points is depicted by Figure 5.1. It can be seen that for the four different cut-off values, the main decline reason is the total number of late fees over the 6 month period, $late6$. Now that the top decline reason has been predicted, the second and third most common decline reason can also be predicted. For the whole dataset the three most common decline reasons are described by the delinquency value over 6 months, number of late

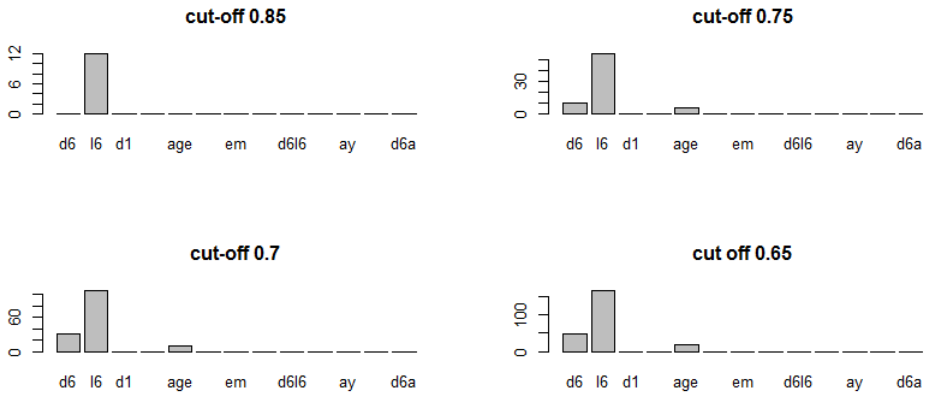

Figure 5.1: Decline reasons for different cut-off points

fees over 6 months and the age of the account in months. The second most common top decline reason is the age of the account, shown in Figure 5.2. It seems that the younger the account age poses more of a risk when lending. This is a sensible deduction, as older accounts can be thought of as loyal, paying customers who are low risk of defaulting. If older accounts missed payments and were a high risk to lend to, from a business perspective credit companies would try to minimise the risk, possibly by terminating high risk accounts. It would not be profitable for credit companies to keep accounts for a significant period of time if the account holder defaults regularly. The third, most common, top decline reason is the age of the account.

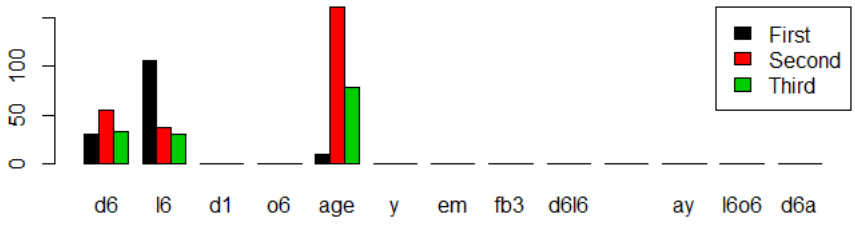

Figure 5.2: Top 3 decline reasons for whole dataset

5.2 Decline reasoning for declined accounts

The three most common top decline reason will then be looked at for accounts who are classified as defaults. The chosen cut-off point for the analysis is 0.75. The number of different top decline reason increases when predicted for accounts that are already classified as defaulted. Having accounts that are a low risk of defaulting, masks decline reasons for high risk accounts. This is beneficial to credit companies, as it gives them more insight as to which variables contribute to accounts defaulting. A high number of defaulted accounts are declined by three common top reasons; delinquency value for the first payment period, number of over-limit fees for the 6 months and the interaction term between the number of late and over-limit fees for the 6 months. The other top three most common decline reasons are employment status, fee balances for payment cycle 3 and young accounts. This is shown in Figure 5.3. The `R` function also outputs the top k variable combinations

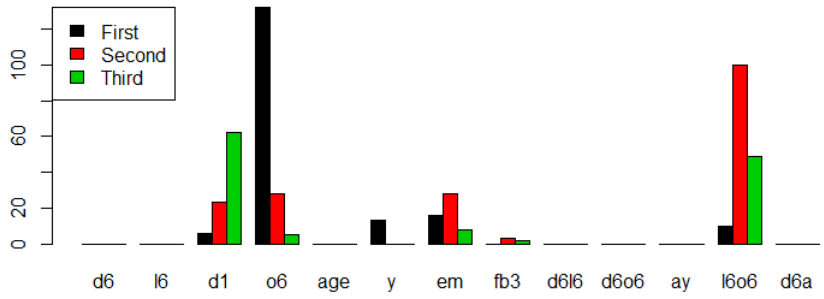

Figure 5.3: Top 3 decline reasoning for declined accounts

for decline reasoning. For example, the top 3 reasons for decline. For the complete dataset the majority of the accounts are described to be declined by the variable combination; $age * young, deling6 * late6, late6 * over6$. This combination is a reasonable prediction of decline reasoning. If the account holder is a young account and has high numbers of over-limit fees and delinquency interacted with a high number of late fees, the account is discernibly going to pose a high risk of defaulting in the future three month period. The credit company can explain to the account holder they were declined for an action due to their unreliability with payments in relation to them being a juvenile account. A high delinquency value over 6 months may not necessar-

ily be an issue if the account is older. This may be because in past cycles the older accounts have a low risk profile. However, when an account is classed as young, there is no past information of how they behaved. Thus, the 6 month period, in which the analysis has taken place, young accounts behaviour is impressionable to the company. The top 3 reasons for decline change when the function is applied to accounts pre-determined as defaults. There are 286 possible combinations that could produced decline reasons. Out of the 286 possibilities, 13 combinations account for around 70% of defaulting accounts. These top decline reasons, grouped alphabetically, and their frequencies are shown in Figure 5.4. The alphabetized groups in Figure 5.4 represent the

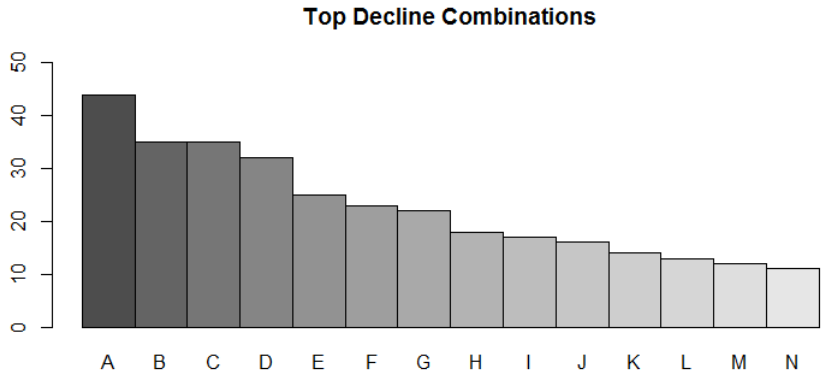

Figure 5.4: Decline combinations for declined accounts

combination of variables, produced for decline reasons. The first five groups are defined in Table 5.1, where behavioural characteristics can be deduced. It is unsurprising that the most popular decline reason is a combination involving high values of delinquency, late fees and over-limit fees over the 6 month period. The 3 variables are high in significance in the final logistic model. In a credit risk viewpoint, a delinquency value of 3 or more is an indicator of a "bad" account. Group B demonstrates that the primary decline reason for young aged accounts is due to the number of over-limit fees in their first 6 months and their delinquency value in the first month. It can be thought that for young accounts, a missed payment in their first month is crucial as to whether they are declined for future actions.

Group	Combination
A	$delinq6 * over6; late6 * over6; delinq6$
B	$over6; age * young; delinq1$
C	$delinq6 * over6; late6 * over6; age * young$
D	$over6; delinq6 * late6; age * young$
E	$over6; age * young; late6$

Table 5.1: Combination classes

5.3 Decline reasoning for accepted accounts

We will now consider that accounts who are accepted can also be incorrectly defined as declined, especially if they have a probability of defaulting close to the chosen cut-off point. Decline reasoning can also be produced, using the same prediction function as previously, for accepted accounts. The number one reason for possible decline is looked at first, with a chosen cut-off value as 0.75, depicted in Figure 5.5. This cut-off point is chosen to be consistent with the declined accounts analysis. The number one reason for decline, on accepted accounts, is the indicator variable *young*. This suggests that if an accepted account is classed as young, i.e. less than or equal to 6 months, it is most likely to be falsely declined due to this reason.

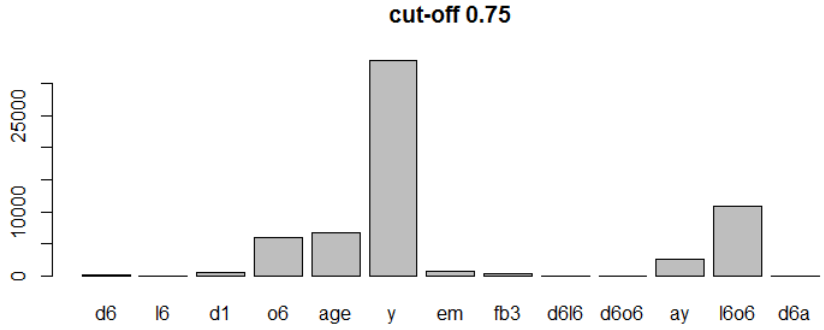

Figure 5.5: Top decline reasoning for accepted accounts

The top 3 possible decline reasons cna now be looked at for accounts classed as accepted. It is shown in Figure 5.6 that when the top 3 decline reasons are predicted for accepted accounts, the top reason for decline changes to the age of the account. After age, the variables $delinq6, late6$ and $over6$ have a similar frequency for the top decline reason. This is unsurprising as they are considered highly significant terms in the logistic model. For both accepted and declined accounts the variable $over6$ has a high frequency for the top decline reason, suggesting no matter what the status of the account a high number of over-limit fees is a general decline reason. For the second and third most common decline reason, the interaction term between the delinquency value over 6 months and the number of late fees over 6 months has the highest frequency. The interaction term between the delinquency value over 6 months and the number of over-limit fees over 6 months also has a high frequency. This is not unexpected, as all three terms separately can be considered risky behaviour when high in value. A high number of missed and late payments largely suggests that the account holder will miss another payment in the next three months. If the number of over-limit fees is high, it may be a possibility that the account holder is spending more than their means, which poses a risk of them being unable to make a payment also. Similarly, with accounts who are pre-determined as defaults, the com-

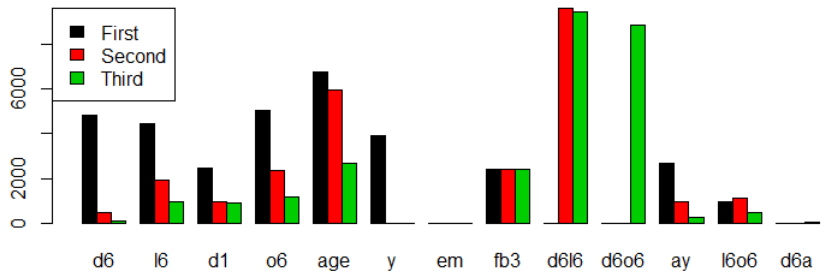

Figure 5.6: Top 3 decline reasons for accepted accounts

bination of possible decline reasons can be produced for accounts who are classed as accepted. There are still 286 possible combinations that could give potential decline reasons. Out of these combinations, 6 of them account for around 70% of accepted accounts. The top decline reasons are categorised, and their subsequent frequencies plotted in Figure 5.7. The categories give

the combination of variables that attribute to decline reasoning. The definitions of these categories are given in Table 5.2, where the decline reasoning can be inferred. It is not unexpected that the top combination of decline reasons, in category U, is $late6 * over6; delinq6 * late6; delinq6 * over6$. All three interaction terms are undesirable for credit companies who are lending. A company wants an individual who repays on time and does not always over spend, not a person who is over-spending and unreliable on payments. The category W is very similar, in decline reasons, to category C for declined accounts. This reiterates that the spending and repayment behaviour of an account who is young is very important. Thus, if a young account has high values in delinquency and high numbers of over-limit and late fees it is very likely that is the reason for them being declined whether they are predicted to default or not. This is helpful for the credit company, as they may charge younger accounts a higher interest who do have high numbers of fees to compensate for potential missed payments. An interesting decline combination is category X, where the reasons are due to the age of the account and the account holder's employment status. This may imply that a young account where the account holder is unemployed is the reason they were declined. Alternatively, an older account may be risky to lend to if they are retired, with a limited income. This category gives a broad range of decline reasoning due to the different types of employment status.

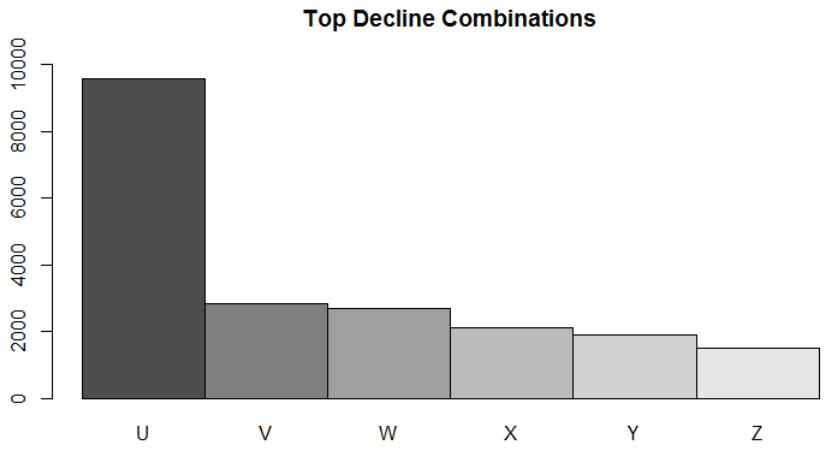

Figure 5.7: Decline combinations for accepted accounts

Group	Combination
U	$late6 * over6; delinq6 * late6; delinq6 * over6$
V	$age * young; late6 * over6; delinq6 * late6$
W	$over6; late6 * over6; delinq6 * late6$
X	$age * young; age; employ$
Y	$age; employ; late6 * over6$
Z	$late6; delinq6; delinq6 * age$

Table 5.2: Combination classes

5.4 Risk-based pricing

Regardless of an account missing a payment in the next three months, there is always risk when lending. This is where the concept risk-based pricing (RBP) is useful. Once a credit company can predict who will miss a payment and have an idea as to why, they can implement appropriate interest rates and fees charged to compensate for the specific risks. Anderson [2007] stated that Edelman [2003] found practical issues associated with risk based pricing and its implementation. He indicated that lenders progress from a flat rate to RBP through the following stages:

1. Increase in rates for high risk customers who were near the cut-off.
2. Decreased standard rate and lower rates for low risk customers.
3. Decrease in cut-off and acceptance of accounts previously declined.

The lenders use RBP to vary prices according to the predicted risk. Cost-recovery pricing is the most obvious way to do risk-based pricing. This tries to allocate costs to every single application and charge accordingly to the risks. Anderson [2007] suggests that high risk accounts should still be considered, because they could become loyal, profitable customers. An example of this may be students; initially they are a high risk to lend to but once they graduate and become earners they could bring high revenue to the credit company. The cut-off point which is used to determine if accounts are accepted/rejected can be moved by the lender to maximise their profits. Lowering the cut-off point gives more decline reasons, shown in Figure 5.1, which gives the lender more variables to consider. Thus, by lowering the cut-off the lender could price a loan higher due to the risk posed by the variables considered decline reasons. For example, lenders could use RBP to charge account holders high interest rates should they have high delinquency values in the categories given in Table 5.1. This use of RBP helps to outweigh any costs of miss-classification of accounts.

Chapter 6

Conclusions

In credit risk scoring statistical models are built, namely scorecards, which produce risk scores. These scores are used by credit companies to make decisions. The decisions made then drive some sort of action, for example the lowering of a credit card spending limit. The problem with these statistical models is the vast number of variables used to produce them. Due to the high number of variables it makes it difficult for the company to assign a reason to a score. The objective of the analysis was to explore methods into assigning reasons to a "bad" score such that the company could give an explanation to their customers as to why they receive the treatment they do.

Before any statistical analysis could take place a number of possible issues with the data had to be dealt with, the first of which was missing data entries. It was unsurprising that a number of data entries were missing due to the large dataset being real, raw data. Upon careful inspection it could be seen that the missing entries were not missing at random and could be explained as to why they were missing. As a result of this and the small proportion of missing entries, the chosen solution to deal with the missing data was imputation. Once the missing data issue was dealt with, the behaviour of the data was investigated. In credit risk modelling it is common to split or segment the data into groups. When looking at the data it could be argued that there was a possible segmentation for the total delinquency value over the 6 month period. There was a large proportion of customer accounts that had a delinquency value of zero. The data was segmented into two groups; one for a total delinquency value of zero and one for a total delinquency value of 1-6. Using the two groups "quick" models were built and analysed to see if segmentation was worthwhile. The benefit of segmentation did not outweigh the issues of analysing more than one model, thus the data was not segmented in further model building.

The number of variables needed to be reduced before logistic regression

could be applied. The chosen method to reduce the number of variables was the use of decision trees. Once the number of variables were reduced to a manageable size standard logistic regression analysis could take place. Forward step-wise regression was used where models were chosen by looking at the AIC value it yielded and significance deduced by ANOVA tests. A model was found, where all the terms were considered significant. The logistic model found then needed to be tested to ensure its predictive power was sufficient. This was done by looking at two methods used in credit risk scoring; the area under ROC curves and Somers' D. Both the area under the ROC curve and the Somers' D value, for the model, were found to be an acceptable value.

The logistic model found was able to produce probabilities as to whether an account holder would miss a payment in the next three months. These probabilities were used to determine a cut-off point where accounts were classified as accepted or rejected. If the account holders probability of defaulting on a payment was higher than the cut-off point, they were deemed declined. The chosen cut-off point was a value of 0.75. This was determined by looking at the effect the chosen cut-off point had on the accuracy of the model, the rate of correctly classifying an account and the rate of falsely classifying an account. Once a cut-off point was chosen the model could then be used to predict decline reasons. The main decline reasons for accounts involved a combination of total delinquency value, over-limit fees and late fees for the 6 month period. This was unsurprising as the three variables were classed as highly significant in the logistic model and are seen as undesirable lending traits.

Bayesian techniques could have been applied had their been more time for in depth analysis. An example of applicable Bayesian analysis would have been the use of Bayesian networks to look at the relationships between the available variables. It may have been a better tool to reduce the number of variables, before logistic regression took place. Further analysis could have taken place to try and group accounts into different risk categories so that risk-based pricing could be applied. If analysis were to continue, other techniques for assigning reasons could be explored with the intention of minimising miss-classification, especially when accounts where near the cut-off threshold.

References

R. Anderson. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, 2007.

G. E. Box, G. M. Jenkins, and G. C. Reinsel. *Time series analysis: forecasting and control*. John Wiley & Sons, 2013.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2002.

A. R. T. Donders, G. J. van der Heijden, T. Stijnen, and K. G. Moons. Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006.

D. Edelman. Working with some behavioural aspects of risk-based pricing, 2003.

D. W. Hosmer Jr and S. Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004.

F. E. H. Jr, with contributions from Charles Dupont, and many others. *Hmisc: Harrell Miscellaneous*, 2013. URL `http://CRAN.R-project.org/package=Hmisc`. R package version 3.12-2.

M. G. Kendall and J. D. Gibbons. Rank correlation methods, 1990.

R. Newson. Confidence intervals for rank statistics: Somers' d and extensions. *Stata Journal*, 6(3):309, 2006.

D. Posada and T. R. Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5):793–808, 2004.

T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):7881, 2005. URL `http://rocr.bioinf.mpi-sb.mpg.de`.

T. Therneau, B. Atkinson, and B. Ripley. *rpart: Recursive Partitioning*, 2013. URL `http://CRAN.R-project.org/package=rpart`. R package version 4.1-3.