A Study On Statistical Models Of Credit Risk

Ajay Pirabaharan, Alexander Watson, Jake Denton, Jinghan XI, Liping Wen, Mingyang LI

Department of Mathematical Sciences
University of Nottingham
United Kingdom

Abstract

The aim of this report is to study models which predict and measure the credit risk of different individuals. Logistic Regression, Basic Decision Tree and Random Forest Models are fitted using a dataset with 150000 observations of 11 variables. It is found that the models are capable of identifying a large proportion of individuals with serious delinquencies (between 70 and 85%) and similar proportions of those without serious delinquencies in the past two years. If those predicted not to have serious delinquencies were to be offered a loan, between 1.5 and 2% would be classified incorrectly. Payment history and amounts currently owed are found to be the two most significant aspects for making predictions, which is the same as many models currently used in finance. The models could be improved by considering more individuals who have had delinquencies (the data is highly imbalanced) and by considering other variables which may be influential in measuring credit risk.

Contents

Chapter 1

Introduction

1.1 Context

A credit score is a metric which acts as a measure of an individual's risk of defaulting on a loan. The term default refers to when an individual fails to repay their loan. In the past, a lender simply decided based upon a consumer's character (often based off the word of neighbours and people in the community) whether or not to trust them with a loan. This was obviously not the fairest way to do things, and societal prejudice underpinned the process. Notably, the credit bureau Equifax (a massive company which began in 1899) was subject to heavy criticism for storing personal information such as race, gender, and marital status. This subsequently led to the Fair Credit Reporting Act, so that credit bureaus decided to contract tech companies to build impartial algorithms for the task.

One of these tech companies was FICO, began by an engineer Bill Fair and a mathematician Earl Isaac. They designed a credit-scoring algorithm which is still widely used today. This algorithm assigns a 3-digit number in the interval 300-850, where a higher score indicates someone with low risk of default. [3]

Everyone over the age of 18 has a credit score, which not only determines whether or not a loan application is approved but also how much it will cost to borrow the money (interest). A score is used whenever you enter a contract with a mobile phone company, whenever you seek a loan to purchase a car, and also influences how much your down payment is on a flat or a house.

In the UK, in contrast to other countries, most lenders have their own internal scoring mechanisms to assess a consumer. Often statistical techniques such as Logistic Regression are used. Oddly, if a lender rejects an applicant, they are not obliged to give the reason why!

So, what factors are considered when a credit score is evaluated? The FICO score is determined by evaluating five sections of an individual's credit report, which are weighted as follows[2]:

- Payment history (reliability with past credit) - 35%,
- Amounts owed (other loans in the process of repayment) - 30%,
- Length of credit history (age of the accounts) - 15%,
- Credit mix (types of credit an individual possesses) - 10%,
- New credit (lines of credit opened recently) - 10%.

In our dataset more variables are considered, some outside the scope of those utilised by FICO. Later in this report the models produced will be compared with the FICO model.

1.2 Motivation

Before starting the investigation, it is important to understand the rationale behind using certain methods or defining the response variable. Firstly, organisations need to define the specific set of rules that classifies a loan as "bad". The definition should be easy to interpret and allow for performance tracking. Financial institutions commonly use PAR90 (the number of times a person experienced 90 days past due delinquency), with the Basel Committee on Banking Supervision defining default essentially as a delinquency stage of 90 days or more.[1] Therefore, from the given variables it is clear that SeriousDlqin2yrs shall be defined as our response variable.

1.3 Structure

This report aims to investigate different models which can be used to predict the risk of offering an individual a loan. Firstly, given variables are studied in order to get a preliminary understanding of how they might be involved in the modelling, then the process in which the data is cleaned is detailed. After this, the models themselves are fitted and analysed. The models considered include Logistic Regression, Basic Decision Trees and Random Forests. The results of this analysis are discussed, with attention to how these models compare to existing models used in industry, how different models may be appropriate depending on the requirements of the company and how the models could be improved. All of these models are based off classification of the variable SeriousDlqin2yrs, which takes the value 0 if an individual has not had a serious delinquency in the past 2 years or the value 1 otherwise. . In this report the observations with value 1 are referred to as 1s or delinquents. The term delinquency refers to the event that a borrower misses a loan payment whilst delinquent (in this context) is someone who has had a serious delinquency (and so has value 1 for the SeriousDlqin2yrs variable). For all the models in this report, if an individual is predicted to have value 0 then they are offered a loan, otherwise the individual is rejected.

Chapter 2

Understanding the Data

2.1 Introduction to the dataset

The dataset contains information from 150000 consumers in 11 variables: 7 integer variables, a percentage, a ratio, one real variable and a binary variable. These variables are described below:

1. **SeriousDlqin2yrs** (Binary) - the number of times a person experienced 90 days past due delinquency or worse.
2. **RevolvingUtilizationOfUnsecuredLines** (Ratio) - the total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits.
3. **Age** (Integer) - the age of borrower in years.
4. **NumberOfTime30-59DaysPastDueNotWorse** (Integer) - the number of times a borrower has been 30-59 days past due but no worse in the last 2 years.
5. **DebtRatio** (Integer) - the monthly debt payments, alimony, living costs divided by monthy gross income.
6. **MonthlyIncome** (Real) - a person's monthly income.
7. **NumberOfOpenCreditLinesAndLoans** (Integer) - the number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)
8. **NumberOfTimes90DaysLate** (Integer) - the number of times a borrower has been 90 days or more past due.
9. **NumberRealEstateLoansOrLines** (Integer) - the number of mortgage and real estate loans including home equity lines of credit
10. **NumberOfTime60-89DaysPastDueNotWorse** (Integer) - the number of times a borrower has been 60-89 days past due but no worse in the last 2 years.
11. **NumberOfDependents** (Integer) - the number of dependents in family excluding themselves (spouse, children etc.)

2.2 Preliminary Investigation and Data Cleaning

2.2.1 Missing values in the data

Variable	1	2	3	4	5	6	7	8	9	10	11
Min	0	0	0	0	0	0	0	0	0	0	0
1st Q	0	0.03	41	0	0.2	3400	5	0	0	0	0
Median	0	0.15	52	0	0.4	5400	8	0	1	0	0
Mean	0.06	6.05	52.3	0.42	353	6670	8.45	0.27	1.02	0.24	0.76
3rd Q	0	0.56	63	0	0.9	8249	11	0	2	0	1
Max	1	50708	109	98	329664	3008750	58	98	54	98	20
N/A's	0	0	0	0	0	29731	0	0	0	0	3924

Table 2.1: The Summary Data for all variables.

From Table 2.1, one can observe that the mean for the response variable SeriousDlqin2yrs is 0.06. Since this variable takes values of 0 or 1 for all observations, it is clear that our data is heavily imbalanced, meaning there are many more observations with value 0 than 1. This has consequences which need to be considered in the modelling section.

There are 120269 observations for which values are available for every variable, leaving 29731 observations which have at least one variable missing. The table reveals that all the missing data belongs to the columns representing monthly income and the number of dependents. 3924 observations have both monthly income and number of dependents missing. There are a few ways to deal with these missing values. Firstly, the two variables which have missing values could be removed, so that none of the observations are removed. This way, all the delinquents remain in the dataset however the potentially crucial information provided by both these variables is lost. Instead, the observations with missing values could be removed. This would have the downside that almost a fifth of the dataset is lost along with 1669 delinquents, but the remaining observations are complete. Finally, the missing values could be replaced by the median value for each variable. The main benefit of doing this is that all of the data is preserved. The median is chosen as a sensible prediction for missing values as it isn't skewed by unusually small/large data. Using this method to deal with the missing values is standard practice (there are many similar examples online) and won't affect the modelling too much, especially if the variables involved are not found to be of any significance.

2.2.2 Variable distributions and larger observations

The summary data for the age variable reveals a minimum value at age 0. Due to the fact that credit scores and loan applications can only be made by people over the age of 18 this is clearly an error so it is removed from our dataset for the modelling. The histogram (Figure 2.1) shows a good spread of ages which you might expect from an unbiased sample of the population, with the majority of the consumers being aged between 21 and 80. There are 13 observations over 100 years old with a maximum age of 109, which is nothing unexpected when data is obtained from a representative sample of the population.

For monthly income, the maximum value is 3008750. This might be extreme and need to be removed, however studying the largest 50 observations in descending order shows that there are other monthly income values with the same order of magnitude. Looking at a wider range of observations reveals that monthly income decreases in reasonable steps (there is no point where the magnitude suddenly drops as you might expect if the value was accidentally interpreted

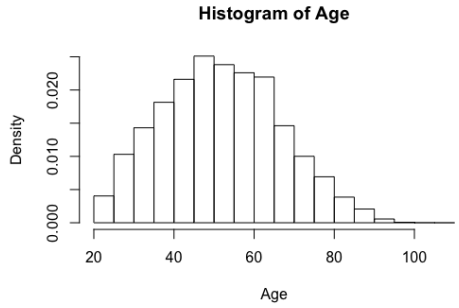

Figure 2.1: Histogram of Age.

as yearly income). As a result, these large values will not be removed. As a consequence of these large values, the logarithm is applied before a histogram is produced (Figure 2.2). Most observations are between $e^7(\frac{1096.63amonth}{13152ayear})$ and $e^9(\frac{8103.08amonth}{97237ayear})$. This is a range that should be considered as 'normal' earnings, as for example, in the UK the national living wage is £15269 (near the lower amount in the above), and anywhere between this and £100000 (near upper bound in the above) is achievable in certain professions.

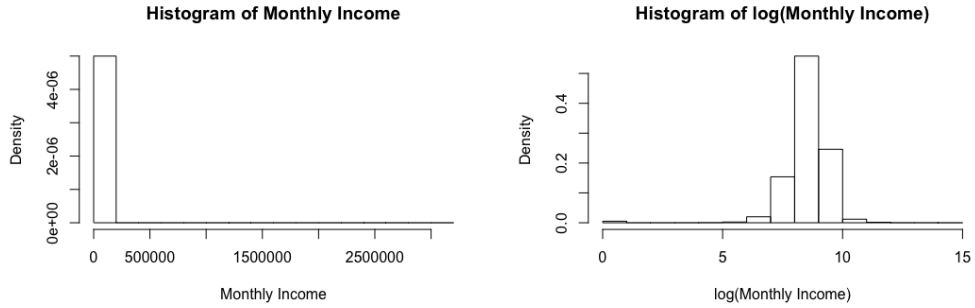

Figure 2.2: Histogram of Monthly Income vs log(Monthly Income).

The revolving utilisation of unsecured lines variable needs a bit of explanation before the values can be interpreted. Personal lines of credit are amounts of money that a lender agrees to make available, similar to a credit card where you can borrow up to a limit. The credit limit is the maximum amount the lender is willing to make available to you. Therefore, this variable represents the proportion of the full amount that a consumer is currently borrowing/using. Sorting the data in decreasing order and looking at the first 50 consumers shows that there are consumers that have values much higher than 1 for these variables, contradicting the description given above which implies an upper bound of 1. Selecting the data where the value is above 1 (3321 observations) and considering the serious delinquency variable reveals that 37.2% of the observations in this subset have had a serious delinquency in the past two years. This also shows that these observations contain over a thousand 1s, so instead of removing these odd observations and losing the information they provide, they are replaced with value 1 and kept in the dataset. Since the subset contained many 1s this variable may be influential in the modelling. The histogram (Figure 2.3) is shown below, and shows that there are many observations which

use hardly any of their available credit as you might expect, leading to the left peak, whilst there are also individuals who use almost the entirety of the credit available to them, near the value 1.

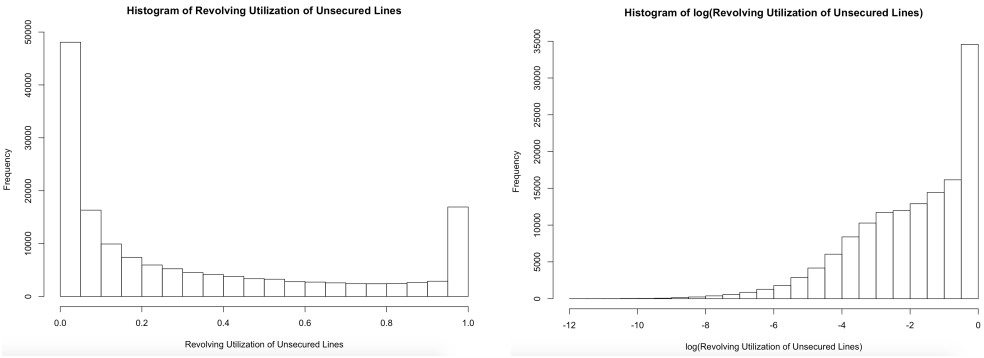

Figure 2.3: Histogram of Revolving Utilization of Unsecured Lines vs log(Revolving Utilization of Unsecured Lines).

The last real variable to consider is the debt ratio. Someone with a high debt ratio might have a higher chance of delinquency as the debt repayments they make are a large proportion or greater than their monthly gross income (sum of all earnings). Note that values for debt ratio of above 1 are of course acceptable as the denominator depends on monthly income, which may be bigger or smaller than monthly debt payments. The summary data shows some incredibly high values and studying the mean and median suggests a positive skew. Strangely, the histogram (Figure 2.4) is bi-modal (two peaks/modal values), the first where we might expect at a debt ratio between 0 and 1, but the second peak is between 1096 and 2980. It might have been expected that the histogram would be quite similar to the one for the unsecured lines of credit variable, and in fact for values between 0 and 1 there is a similar shape. However, for higher values there are exceptionally more people with a higher debt ratio.

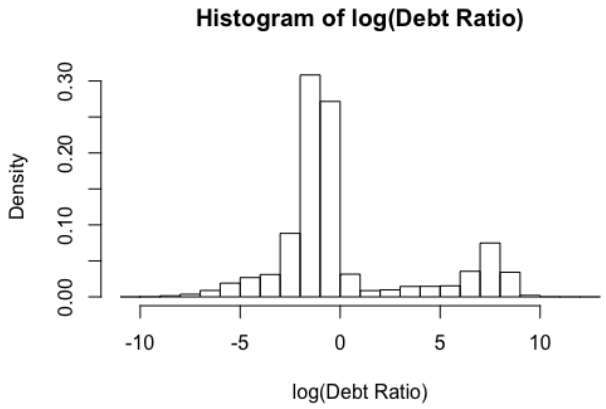

Figure 2.4: Histogram of log(Debt Ratio).

High debt ratio means that a person owes a lot more money through debt (one might also assume that they have a high revolving utilisation) than what they earn, which may have been

caused by high interest rates, high living costs or owing a proportion of their income to their spouse via alimony. If somebody has debt through borrowing and owes alimony/rent each month then if they have low monthly income this debt ratio can easily become large quickly. Taking the consumers with a debt ratio above one, it can be seen that this subset makes up 22.85% of those who had serious delinquencies in the last 2 years. One possible explanation for the second peak might be that a lot of people in this range decide to ask the creditor to freeze their interest/charges because they're in arrears, or these people have recently lost their job, so their debt ratio has skyrocketed.

The integer variables need to be checked to make sure that there is nothing odd. The number of times someone is 30-59 days past due has a lot of values at the maximum value 98, and then suddenly there is a drop to 14 or less, and by the 1000th observation, we have a value of 4. This needs to be investigated further. When these 264 values of 98 are considered, more than half of these had serious delinquencies in the past two years, which you would expect if these people missed their due date so often. Despite this, since there is no reasonable explanation for why they are all the same value, and an individual having 294 late payments in 2 years is implausible, they need to be removed from the data. On top of this, later in this report it will be seen that this cluster of observations has a large impact on the correlation, which supports the decision to remove them.

Moving on, the number of real estate loans shows nothing too unrealistic and has a similar distribution to the number of open loans, the first doesn't give too much more information than the latter as one might expect from their descriptions. There might be some correlation/collinearity between the two which is later investigated.

Finally, the number of dependents has some large values, but nothing that should not be expected when you consider that the dataset is a random sample of 150000 people. There is one observation with 20 dependents, but the vast majority of the data has between 0 and 5 dependents (5 dependents could be a spouse and 4 children) which is realistic.

2.2.3 Similar variables and a parsimonious model

Some of these variables encapsulate much of the same data. One of the aims of model fitting is to find the simplest model to describe the data. With this in mind, it might be possible to pick just one or two variables from each of the groups below for parts of the modelling. These groupings are chosen as the variables have an above 0.4 linear correlation, and also due to the descriptions of each variable which might lead to some collinearity, so are deserved of further investigation.

Group	**Variables**
1	NumberOfTimes30-59Days/60-89/90Late
2	DebtRatio/NumberRealEstateLoansOrLines
3	NumberOfOpenCreditLinesAndLoans/NumberRealEstateLoansOrLines

Table 2.2: The groupings of similar variable types.

2.2.4 Investigating relationships between variables

The Pearson correlations between each of the variables are summarised in the table below, with yellow highlights for those values that will be studied further (note: the variable numbers are as defined in section 2.1):

Variables	1	2	3	4	5	6	7	8	9	10	11
1	1	0.24	-0.1	0.12	0.06	-0.02	-0.03	0.11	-0.02	0.1	0.04
2	0.24	1	-0.26	0.11	0.06	-0.03	-0.17	0.1	-0.08	0.09	0.08
3	-0.1	-0.26	1	-0.05	-0.08	0.03	0.18	-0.05	0.06	-0.04	-0.21
4	0.12	0.11	-0.05	1	-0.03	-0.01	-0.05	0.98	-0.03	-0.98	0
5	0.06	0.16	-0.08	-0.03	1	-0.05	0.35	-0.05	0.52	-0.05	0.1
6	-0.02	-0.03	0.03	-0.01	-0.05	1	0.09	-0.01	0.14	-0.01	0.06
7	-0.03	-0.17	0.18	-0.05	0.35	0.09	1	-0.07	0.42	-0.06	0.04
8	0.11	0.1	-0.05	0.98	-0.05	-0.01	-0.07	1	-0.04	0.99	-0.01
9	-0.02	-0.08	0.06	-0.03	0.52	0.14	0.42	-0.04	1	-0.04	0.13
10	0.1	0.09	-0.04	0.98	-0.05	-0.01	-0.06	0.99	-0.04	1	-0.01
11	0.04	0.08	-0.21	0	0.1	0.06	0.04	-0.01	0.13	-0.01	1

Table 2.3: The Pearson correlation between each of our variables.

1) NumberOfTimes30-59DaysPastDueNotWorse with 60-89 days and 90 days late

Producing scatterplots at first for these variables doesn't reveal much as all these variables share the cluster of large values at 98 which were discussed previously and subsequently removed. After removal, the plots become a little clearer (see Figure 2.5) and a recalculation of the correlation coefficient gives 0.295, significantly lower. As a result of this and reasons discussed earlier, the observations with 98 are removed for all of the modelling.

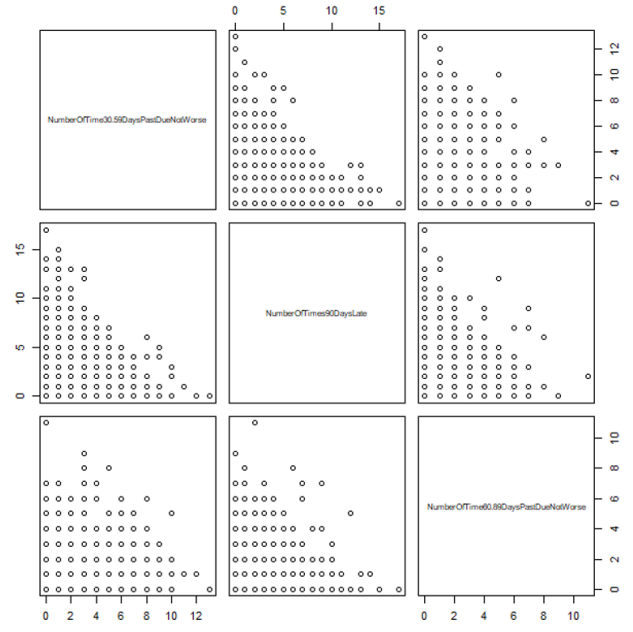

Figure 2.5: Scatterplot for Payment History Variables.

Are all three variables necessary for the modelling? Looking at the summary data, it is clear that more people have had at least one 30–59-day late payment than for the other two. This

could be due to people not realising they are overdue, or perhaps they had to wait for payday before they could meet their payments, so it may not really give us an indicator whether or not someone will default. The 60–89-day late payment variable might provide more indication that an individual carries risk, but it has the lowest mean out of the three and also the lowest range. Finally, the 90 days late variable has the largest spread of values, you would expect all these people to be aware that they are late (in contrast to 30-59 days late), so this could be the most significant variable to use of the three. Delving further, the observations with values greater than 5 were selected for each variable and the number of delinquencies counted. The subset for the 90-day variable showed 2/3 of this subset had delinquencies in the past two years with 121 observations. For 30-59 days there were 120, whilst the 60-89 days variable had just 29 observations of people with serious delinquencies in the past two years. In summary, we might expect that the 90-day late variable will be the most useful in the modelling followed by 30-59-days late variable.

2) DebtRatio and NumberRealEstateLoansOrLines

An initial scatterplot (left side of Figure 2.6) between the two shows that the observations of extremely high debt ratio tend to have few real estate loans/lines, whilst the very high number of real estate loans/lines tend to have low debt ratio. To see the relationship more clearly, a second scatterplot (right side of Figure 2.6) is produced which considers debt ratio up to 5000 and 20 or less real estate loans/lines. When this is done, it tends to be the case that as the number of real estate loans/lines increase, the maximum debt ratio observed for a fixed number of real estate loans also increases. However, for the same fixed number of real estate loans there is a large range of possible values for debt ratio, so that if you had an observation with say 4 real estate loans, you could not accurately predict their debt ratio. The third group also contains the number of real estate loans/lines, so this too should be considered before coming to a verdict on how the variable might be involved in the modelling.

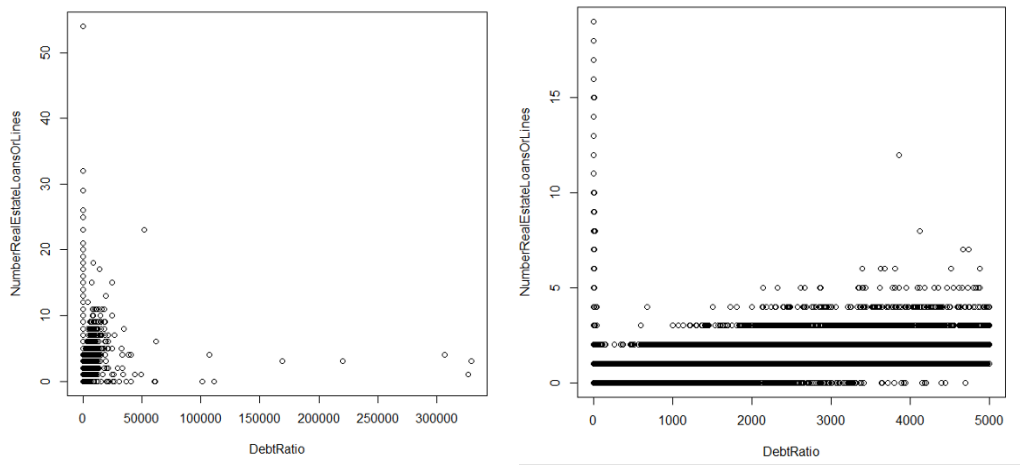

Figure 2.6: Scatterplot for DebtRatio/RealEstateLines.

3) NumberOfOpenCreditLinesAndLoans and NumberRealEstateLoansOrLines

From the descriptions of these variables, the number of open credit lines and loans includes the number of real estate lines or loans as well as car loans, credit cards and other lines of credit. This is translated into the plot (Figure 2.7) as the number of open lines of credit is always greater than the real estate loans/lines. The difference in height from the line with the number of open lines of credit equal to the number of real estate lines or loans are the other lines of credit are mentioned above. The number of these other lines of credit vary between individuals. These variables have some dependence, so naturally we might only include one for the modelling depending on the assumptions of the model. Since the number of open credit lines and loans has more information, this is the one that might be preferred in the modelling.

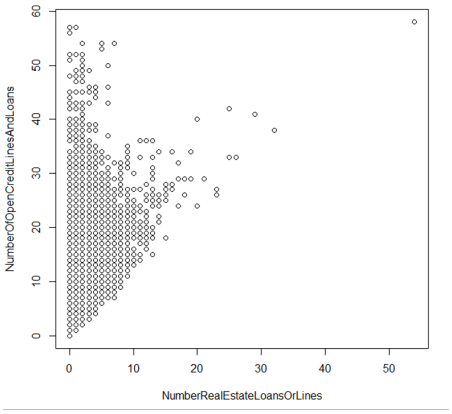

Figure 2.7: Scatterplot for Open/RealEstate loans.

2.2.5 Summary of Data Cleaning

The first few parts of this section made it clear that certain data points should either be removed or replaced. In regard to the missing data, it was suggested that one standard way to deal with this problem is to replace them with the median for whichever variable they represent. In all the modelling that follows in this report, the missing values are replaced with the median.

There are also a number of other odd observations in the dataset. Firstly, there is an observation with an age of 0 which cannot be true, so this is removed. On top of this, when we looked at the variables representing the number of times someone pays a specified number of days past due, a cluster of observations with value 98 were discovered which are common for all three variables. Since there is no obvious explanation for these points and these observations have a large effect on the correlation, they are removed too before any modelling is performed. Lastly, when the revolving utilisation variable was considered, a number of observations had a value above one which is not allowed considering the description of the variable. These observations contain many 1s, so instead of removing them their value for this variable is replaced by the maximum allowed value from the definition, which is 1.

Chapter 3

Modelling

Now that we are at a point where we understand the data and we know that odd observations have been dealt with, we can proceed with the statistical modelling.

3.1 Logistic Regression

3.1.1 Introduction

Logistic regression is one of the most widely used models for categorical response data. It is an example of a generalized linear model which is able to estimate the probability that an event occurs or not based on a number of predictor variables. Logistic regression is used for many applications including biomedical studies, social science research, marketing and finance.[5] We will be using the logistic model to assess the probability that a customer is creditworthy (i.e. able to meet a financial obligation in a timely manner).

3.1.2 Notation and Definitions

First some key notation must be introduced:

Let z_{ij} represent the value of the jth variable for the ith observation, where $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p$. Let Z denote a $n \times p$ matrix whose (i, j)th element is z_{ij}. z_i represents all the information for one individual, included as a row vector in Z possessing length p. Y is the response variable, y_i denotes the value of the response variable for the ith individual. Let β_j be the coefficient of the jth variable in the logistic regression model.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \tag{3.1}$$

Then using this notation, the logistic regression model for predicting a binary response Y using p predictors is defined by:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}, \tag{3.2}$$

where $X = (X_1, X_2, \cdots, X_p)$ are the p predictors, and p(X) is the probability of default given X, i.e. p(X)=Pr(Y=1|X). Rearranging the equation above and taking the logarithm of both sides gives:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{3.3}$$

The left-hand side is called the $log-odds$ or $logit$. This equation shows that the logit is linear in X. The coefficients $\beta_0, \beta_1, \cdots, \beta_p$ are estimated using maximum likelihood, where the likelihood function is given by:

$$l(\beta) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})). \tag{3.4}$$

3.1.3 Assumptions Of Logistic Model

The key assumptions/requirements for a valid model [6] are as follows:

1. The observations should be independent of each other (must not be repeated measurements/matched data).
2. There should be little or no multicollinearity between the independent variables (i.e the independent variables should not be highly correlated with each other).
3. The independent variables should be linearly related to the log odds, as can be seen from the definition of the model.
4. A large sample size.

These assumptions need to be checked in order for the logistic model to be valid. It can be shown that each of these is satisfied for our given dataset:

1. Each individual data point is a separate study of a unique individual. In this way, each observation is independent of another.
2. In order to check that there is little or no multicollinearity among the independent variables one can study table 4.1 which can be found in the appendix. The updated table for the cleaned data reveals that there is no strong correlation between the variables. Another way to test this is using the Variance Inflation Factor (VIF). A VIF value that exceeds 5 indicates a problematic amount of collinearity.[7] The VIF value for each variable is between 1 and 2 in our case, further supporting the assumption (shown in Table 6.2).
3. Studying the scatterplots in Figure 6.1 (see appendix) reveals that for two examples with the variables Debt Ratio and Monthly Income there are in fact linear relationships as required. This can be shown for all variables.
4. The clean dataset is extremely large and the training data is a large proportion of it, so the final requirement is met.

3.1.4 Training/Testing Data

The clean dataset is split into a training set and a testing set. 70% of the initial dataset is used as a sample for modelling the initial data in the training set, with the other 30% used for predictions as the testing set.

3.1.5 Variable significance

Two different methods have been used in order to determine which of the explanatory variables are significant:

- Hypothesis tests using the z-statistic: H_0: β_i=0 vs. H_1: $\beta_i \neq 0$.

  The null hypothesis implies that the probability of default does not depend on the variable X_i $(i = 1, 2, \cdots, p)$ when all the other variables are included in the model. According to the z-statistic associated with the predictors, the p values will determine whether a variable is significant to the model or not. In other words, the test will conclude whether there that is an association between the probability of default and each predictor variable.

- The Akaike information criterion (AIC).

  AIC tackles the trade-off between the goodness of fit of the model and the model complexity, providing a means for variable selection. Stepwise regression is conducted for this logistic regression model, in each step of which a variable is considered for addition or subtraction from the current set of explanatory variables based on AIC value.[8]

The logistic regression model is built with all explanatory variables (2-11) using the training data. The table below shows the outputted results:

Variables	**Estimate**	**Z Value**	**Odds Ratio**	**P-Value**
(Intercept)	-3.407	-50.698	0.033	<2e-16
RevolvingUtilizationOfUnsecuredLines	2.409	49.017	7.759	<2e-16
Age	-0.018	-16.177	0.982	<2e-16
NumberOfTime30.59DaysPastDueNotWorse	0.426	31.809	1.531	<2e-16
DebtRatio	-3.184e-05	-2.471	0.99997	0.01347
MonthlyIncome	-2.259e-05	-6.185	0.99998	6.21e-16
NumberOfOpenCreditLinesAndLoans	0.031	10.108	1.032	<2e-16
NumberOfTimes90DaysLate	0.702	34.766	2.017	<2e-16
NumberRealEstateLoansOrLines	0.097	7.486	1.102	7.12e-14
NumberOfTime60.89DaysPastDueNotWorse	0.598	21.782	1.818	<2e-16
NumberOfDependents	0.036	3.007	1.037	0.00264

Table 3.1: Summary findings of our fitted model from Logistic Regression.

To judge the fit of the new model, a likelihood ratio test is conducted between the null model (where all the coefficients except the intercept are zero) and the fitted model. The null hypothesis for this test is that the two models are statistically equivalent. A p-value of much less than 0.001 is obtained, indicating that the fitted model fits the data significantly better than the null model and consequently the fitted model, containing all variables, has some power in explaining the variability of the response variable. Also, according to the importance of the predictors, the predictor DebtRatio is the least important one. When a likelihood ratio test is conducted between the reduced model (i.e. excluding DebtRatio in the fitted model), the p-value is 0.009494<0.05, so the reduced model is statistically different form the fitted model, which gives more evidence that the fitted model is not bad. Moreover, the result is consistent with the predictor selection that all variables should be kept in the model.

3.1.6 Explaining the fitted model

Coefficients For Each Variable

The "Estimate" in the table is the coefficient estimate found using the maximum likelihood.The coefficient being less than zero or the odds ratio being greater than 1, indicates that an increase in the predictor is associated with an increase the probability of default. For example, the coefficient of RevolvingUtilizationOfUnsecuredLines is 2.049. An increase in this predictor is associated with an increase in the probability of default (more likely that an individual has value 1 for the response). Each one-unit increase in RevolvingUtilizationOfUnsecuredLines is associated with an increase in the log-odds by 2.049 units. Rearranging the log-odds ratio, we find that when RevolvingUtilizationOfUnsecuredLines increases from 0 to 1 with other variables fixed, the odds of delinquency increases by a factor of 7.759.

Variable Importance

To see how the predictors are influencing the results, variable importance for logistic regression models can be measured using the absolute value of Z from the table. The importance of the predictors using this metric is shown below:

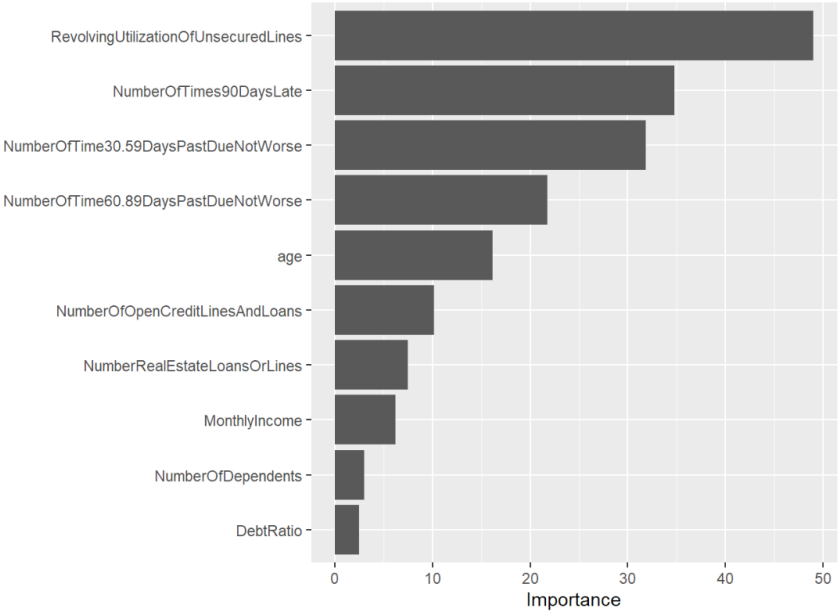

Figure 3.1: Plot showing variable importance

The revolving utilisation variable is the most important, followed by the variables representing the number of times someone is late on a payment. This trend in variable importance is shared with the random forest model which will be discussed later in this report.

3.1.7 Predictive Ability

Probability Thresholds For Prediction

The model can be used to predict the probability of an individual having a serious delinquency in the past two years using the testing data. To assess the performance of the model, the number of correct and incorrect classifications are considered. Since the output is a probability for this model, a threshold needs to be chosen to determine whether an observation should be predicted to be a delinquent. In this way, the probability can be transformed into binary outcomes 0 or 1, which are the classes of the response variable.

Confusion Matrices and Other Measures

Given an observation with known response and prediction, there are 4 possible outcomes. An observation which possesses value 0 for the delinquency variable and is predicted as value 0 is a true negative (TN); conversely if it is classified as 1, it is a false positive (FP). Given an observation with a response of 1, if it is classified as 1 it is counted as a true positive(TP) and if it is classified as 0, it is a false negative(FN).The 4 possibilities can be summarised into a 2×2 table (Table 3.2), which is named the confusion matrix. The diagonal elements of the matrix are correct predictions, while the off-diagonal elements are incorrectly classified.

	True 0	True 1
Predict 0	TN	FN
Predict 1	FP	TP

Table 3.2: The Confusion Matrix

Using this confusion matrix, many important performance metrics can be calculated. Performance is characterized using the terms sensitivity and specificity. The sensitivity is the percentage of the true defaulters that are identified. The specificity is the percentage of non-defaulters that are correctly classified. A perfect classification has a sensitivity/specificity of 100%. Furthermore, it is possible to set a threshold value to classify all the values greater than threshold as 1 and lesser then that as 0. That's how the probability of defaulting is predicted. The default value for threshold is usually 0.50, however by altering this threshold value it is possible to see a change in predicted values, leading to different matrices and more importantly varying TN and TP values. When these different sensitivity and specificity values are plotted on a scatter plot and a line is passed through them we get what is called an Receiver Operating Characteristic (ROC) curve.

As previously mentioned, the dataset is highly unbalanced with around 93.4% of the observations having value 0 in the response. Then, if the threshold is set to 1, all of the predictions are 0 but the accuracy is 93.4%. Using a threshold of 0.5 gives the confusion matrix in Table 3.3, which shows that only 16.2% of the true 1s are captured. Clearly this is not ideal either as if all the observations which are predicted 0 are offered a loan, more than 1 in 20 would be a high risk of default. The next section deals with picking the best threshold which balances sensitivity and specificity.

	True 0	True 1
Predict 0	41624	2481
Predict 1	331	482

Table 3.3: The Confusion Matrix for a threshold of 0.5.

ROC curves

One method of finding better thresholds is by using an ROC curve. The ROC curve is a plot illustrating the diagnostic ability of a binary classifier system as its discrimination threshold is varied. As previously described, the ROC curve is created by plotting sensitivity against 1-specificity for various thresholds.

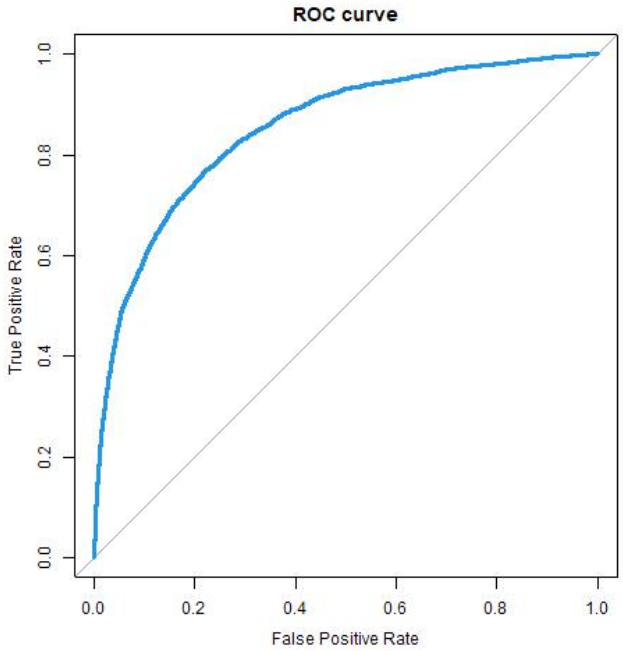

Figure 3.2: The ROC Curve For Our Model.

The figure above shows the ROC curve for the model. The lowest left point (0,0) represents all of the response being predicted as 0s, the threshold being 1. The (0,1) point is a perfect classification. The ideal ROC curve hugs the top left corner near this point, indicating that a threshold can be chosen with a high true positive rate and a low false positive rate.[8]

	True 0	True 1
Predict 0	32578	683
Predict 1	9377	2280

Table 3.4: The Confusion Matrix using a threshold of 0.06336881.

The threshold that minimises the distance between the ROC curve and the point (0,1) is 0.06336881, and the associated sensitivity/ specificity for this threshold are 76.9%/77.6% respectively. The corresponding confusion matrix (Table 3.4) indicates that the sensitivity has increased with this choice from 16.3% to 76.9% and the true 1s predicted correctly increase from 482 to 2280 (specificity increased). Hence, this threshold has a better performance and a better trade-off between sensitivity and specificity. Other thresholds can be found to obtain particular chosen sensitivity or specificity. For example, if the required specificity is 60%, the sensitivity can go up to 89.0% by choosing the right threshold. Unfortunately, increasing sensitivity decreases

specificity and vice versa, so we are restricted to the threshold above for the best performing model overall.

3.1.8 Improving The Model

Principal Component Analysis (PCA)

Since there are 10 predictors in the data, PCA is tried in an attempt to reduce the dimensions of the data and improve the sensitivity and specificity of the model. Doing so leaves 8 predictors and the principal component scores explain 90% of the total variance plus the coefficients are significant. The sensitivity and specificity achieved are approximately 76.2% and 78.5% respectively, which is not much better than the previous model, with the disadvantage that the model becomes much harder to interpret. In other words, the incorporation of PCA does not make that much of a difference to the model. A plot (Figure 6.2) of the ROC curves of our logistic regression model with PCA is included in the appendix for completeness.

Altering the training data set

The data is highly unbalanced, so one natural method to deal with this is to change the proportion of 1s and 0s in the training data to make it more balanced, then use the same testing data to make predictions. The data is first split into the 0s and 1s respectively, then many new training sets are formed which have a specific number of 1s and an integer multiple N of this number of 0s. Logistic models are fitted and the thresholds which minimise the distance to (0,1) on the ROC curve are found for each. The highest sensitivity achieved in this way is around 78.5% and the corresponding specificity is 76.5%, with 5 times as many 0s as 1s. The ROC curves (Figure 6.3) of the logistic regression model when altering the training data or not, which are almost the same, are included in the appendix. The confusion matrix is shown in Table 6.3. It is difficult to conclude which is better, since the false positives increase from 9377 to 9843 and the true positives increase from 2280 to 2325. This leads to a dilemma: how many false positives should be accepted in exchange for a slight increase in the number of true positives?

3.2 Decision Trees

3.2.1 Introduction to Decision Trees

Decision trees and associated ensemble/bagging methods are one family of supervised machine learning algorithms.[9] Supervised machine learning is the process of fitting a model based upon observations where the response variable is already known. In this section, the focus is on classification trees, meaning that the response variable is binary (takes only values 0 and 1).

A decision tree essentially acts as a filter, moving data where certain conditions on the predictor variables are met into different subsets with the aim of having pure subsets at the end (pure meaning that all the points in that subset have the same value for the response variable, in our case 0 or 1). The whole dataset begins at the top of the tree, which is named the root node. From here, data is split based upon a first condition into two subsets. One of these subsets has observations satisfying the condition whilst the other does not. This process of splitting based on different conditions continues until either: the tree reaches a maximum depth (user defined complexity of the tree, where each condition used to split the data creates a new level), the subsets at the end reach a user-defined maximum size (maximum leaf size), or the subsets are pure with respect to the response variable. These user-defined inputs are called hyper-parameters. The points at which the tree splinters or ends are called the nodes. The lines that connect the nodes

are called branches. When a branch ends and there are no further splits, the node is called a leaf.

When a prediction is required based on new data, each observation is considered in turn, moving from the top of the tree along different branches according to whether or not the condition at that level is satisfied, until the observation reaches a leaf. At the leaf, if the conditions have been chosen well enough so that the subsets are reasonably pure, we'd expect that the response variable of the new observation matches the majority of those at that leaf, so summing the number of each class gives a probability associated with how confident we are with predictions for a particular group. The most important question to ask at this point is how are the conditions chosen at each split? The method used in this report uses a metric known as the Gini impurity. This is a measure of how often you wrongly classify a random element in the subset based on the class distribution in the set.[10] In our case, there are two classes, either 0 (which represents someone having no serious delinquencies in the last 2 years) or 1. The class distribution is then just the probability that if you picked out an observation at random from the set, it would be either a 0 or a 1. In this way, we can calculate the Gini impurity:

$$G = \sum_{i=1}^{C} p(i)(1 - p(i)) = p(Res = 0)(1 - p(Res = 0)) + p(Res = 1)(1 - p(Res = 1)), \quad (3.5)$$

where 'Res' denotes Response. The first summation is the general form of the Gini impurity, where i represents a class of the response variable and C is the number of classes. The right equation is the form for our problem, where the classes are either the response equal to zero or one. This is a measure of the purity of the set. To see this, consider a set entirely made up of 0 observations (or entirely 1 observations), then G=0. Therefore, the closer G is to the value zero, the purer the set. A condition is then chosen based upon the Gini gain, which is found by first considering the Gini impurity before the split and calculating the Gini impurity for the two subsets after the split. The impurities for the post-split subsets are weighted according to the proportion of the original dataset they contain and summed (this represents the new Gini impurity after the split). This is then subtracted from the original impurity to give us Gini gain, which tells us how much impurity we removed by performing the split.

It is this measure that informs the tree how to choose the condition for the split. Clearly, a higher Gini gain is a better split, so the condition which splits the dataset into the purest subsets will be used in the tree. An important note is that this is not the only metric which could be considered in order to split the node, others include information gain which is based on entropy and variance reduction.

The advantages of decision trees include how intuitive they are to interpret, how they can be easily used to predict the response for new data, and how they can handle numerical and categorical data. One limitation is over-fitting. If a leaf size is chosen too small or depth chosen too large, you can quickly have a complex tree fitting the training data well, but with poor predictive power for new data. Another issue is that the trees fitted in this report use a greedy algorithm, meaning the decision at each node in turn is made to maximise Gini gain. This doesn't guarantee that the subsets obtained at the leaf nodes will always have the optimal minimised Gini impurity (which would be ideal) since these decisions are local to the nodes and not global (for the entire tree).

One extension on simple decision trees considered in this report are random forests. Random forest fits many decision trees and meshes them together to get a more stable tree. One important difference between the two methods is feature randomness. At each split, instead of considering every variable for the condition, a smaller subset of the variables is considered. This adds some diversity to the model and limits the effect that the use of the greedy algorithm has on the individual trees, which often leads to better models.

A measure of the importance of each variable can be calculated too, which tells us which variables have strong relationships with the response. The feature importance is the decrease in impurity for a condition with respect to a variable at a node, weighted by the probability of reaching that node (proportion of full dataset that reaches the node). Larger values represent more important features.

3.2.2 Basic Decision Trees

The first tree that is fitted uses training data which is representative of the whole dataset. 60% of the observations from the dataset are randomly sampled to create the training set whilst 20% of the observations will be randomly sampled for a representative test set (this will be used to test other models later too). It is important to note that when fitting the model, all variables are considered. This is acceptable since there are no assumptions on relationships between variables in this method, the greedy algorithm always picks the variable which generates the highest Gini gain at each split. Consequently, insignificant variables do not interfere with the fitting. However, the greedy algorithm used in performing splits always chooses the best variable locally so the globally optimised tree is unlikely to be obtained. On top of this, when variables have multicollinearity, the tree may lose some important information when all variables are considered. This is one of the reasons why random forest is implemented later in this report, as it reduces this effect. Once the model is fitted, it can be visualised, revealing the conditions it imposes at each split:

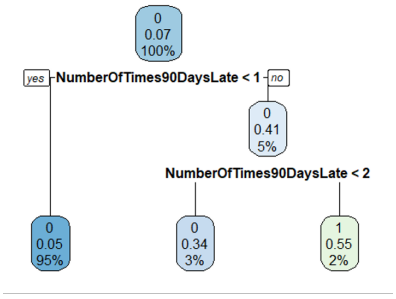

Figure 3.3: Decision Tree 1 fitted with 60% representative sample.

The numbers at each node in the figure can be interpreted as follows: the top number represents the dominating class of the response variable in the subset (are there more 0s or 1s?), the second number is the probability that if you picked a random observation from the subset you would pick a 1, and the third number is the proportion of the full dataset that the subset represents. The conditions here involve the same variable twice. The predictions that are made can be summarised as: predict 0 if observation has value of 0 or 1 for number of times more than 90 days past due on repayment, otherwise predict 1. This tree will not have accurate predictions since only one variable of the available 11 is considered for the prediction. Indeed, there is a lot of error in the leaf nodes. For example, if you take an observation from the right-hand leaf, there is only 55% chance the observation is a true 1. The trees predictive power can be observed using the test data. The following confusion matrix (Table 3.5) is obtained:

This shows that of the 27970 0s in the testing data, the tree correctly predicted 27696, giving a 99.02% accuracy. Clearly, this tree is very effective at predicting values of 0. On the other hand, of the 1976 1s in the testing data, the tree only predicted 336 correctly, representing only

	True 0	True 1
Predict 0	27696	1640
Predict 1	274	336

Table 3.5: The Confusion Matrix for Decision Tree 1.

17% accuracy. This is very poor, and when we class predictions of 0 as people worthy of a loan, 5.59% of those offered a loan under this model are 1s.

Recall that for this tree, a representative sample of the dataset was used as training data. Of the 89838 observations in this training set, just 5927 have value 1 for the serious delinquency variable. The training data has more than 10 times the number of 0s as 1s, in other words it is heavily imbalanced. This is an issue as the dataset already has low Gini impurity (there aren't many 1s), so when splits are performed there is only tiny Gini gain. This could be why the same variable is considered in both conditions. Another intuitive reason is as follows: the more observations considered, the more likely it is that similar observations (similar predictor values) can be found which have different values for the response. These similar observations create noise which impacts the tree. As a result, other trees will be fitted from more balanced training data.

3.2.3 Altering the Training Data

The training data we choose to fit a second tree will have 7500 of each class of the serious delinquency variable. This guarantees a higher Gini impurity at the root node, and also still provides a lot of data for the tree to use both for training and testing, since over 2000 1s still remain after taking the training data. The new tree is visualised below:

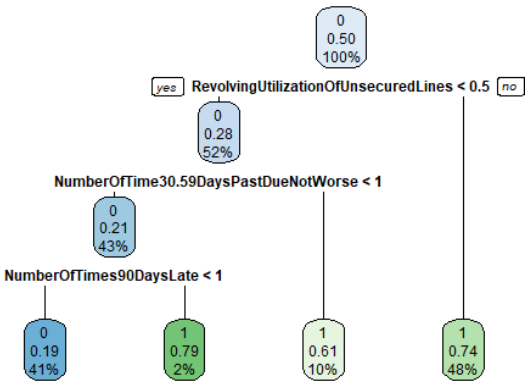

Figure 3.4: Decision Tree 2 fitted with balanced training data.

This tree has more conditions relating to other variables than the previous tree. It predicts the 1s better, as the probabilities in the leaf nodes are much higher, with 79% 1s in the second group (from left). Another thing to note is that it still picks up a large amount of the 0s in the data, as in the first group (furthest left) which holds 42% of the original dataset, 81% of the observations are 0s. Since every other group predicts a value of 1, there will be some error in the predictions as the root node started with 50% of the dataset having value 0. In regard to

the conditions present in this tree, firstly the revolving utilisation of unsecured lines variable is utilised, which is the proportion of credit available to a person that they are currently using. If a person is using at least half of the credit available to them, the tree predicts that they have had a serious delinquency in the past 2 years. These people represent 47% of the training dataset (7050 observations), and of these 75% have value 1 (approximately 5288), quite an effective split! The other 53% of the data is passed down the left-hand branch from the root node. At the second level, the condition is related to the number of times that a person is 30-59 days late on a payment. If this number is not 0 then the tree predicts that the person has had a serious delinquency in the past two years, with about 61% accuracy. This is the poorest accuracy of all the groups. The tree is being quite ruthless with these observations, and indeed if this tree were to be used to decide whether to offer a loan, it would seem quite harsh to not offer it if someone had been 31 days late to pay one time. This information is not available in the data unfortunately, but of course the variable is only considered from 2 years ago, so does not condemn the consumer for life. There is a third condition if a person is using less than 50% of their available credit and have not been more than 30 days late on a payment involving the variable which represents the number of times someone has been over 90 days late for a payment in the last 2 years. From here, if a person has had any payments over 90 days late, the tree predicts them to have had a serious delinquency, and in doing so is 79% accurate. If they haven't had any 90-day late payments, they are predicted to be 0. Most of the data moves into this second subset (the far-left group), which is 81% accurate in it's prediction of 0 within the training data. Having discussed this, balanced test data can now used to see the predictive power of this tree, the results are summarised in the table below:

	True 0	True 1
Predict 0	1358	310
Predict 1	642	1690

Table 3.6: The Confusion Matrix for Decision Tree 2 (balanced test data).

As expected, this tree is far better at predicting 1s. Of the 2000 1s in this test data, the tree correctly predicted 1690 of them (84.5%), and of the 0s, the tree correctly predicted 1358 (67.9%). This tree is in a way more 'harsh' than previous, as if we reject the people who the tree predicts have value 1, almost a third of them might be wrongly rejected when they would not have gone on to default. On the other hand, 18.59% of those who are accepted for a loan turn out to have serious delinquencies. Ideally, a company wants to minimise the proportion of people they accept that go on to default on the loan. So, the improvement in this tree compared to the last can truly be seen if the representative testing set is used on the new tree. This gives the table below:

	True 0	True 1
Predict 0	18912	291
Predict 1	9058	1685

Table 3.7: The Confusion Matrix for Decision Tree 2 (representative test data).

The improvement with the new tree is now evident, as of those who would be offered a loan only 1.5% have serious delinquencies in the past two years. However, 8784 fewer individuals are accepted who do not have serious delinquencies. The company could have profited from these individuals, so we should see if a better tree can be developed which accepts a higher proportion of the true 0s.

3.2.4 Developing a better tree

The hyper-parameters can be altered in order to fit a tree that models the training data better. This is done with the training data used for tree 2. The tree can be evolved deeper, giving an additional condition which decreases the Gini impurity at the leaf nodes even more, hopefully increasing the number of true 0s captured. A maximum depth of 5 gives the following tree:

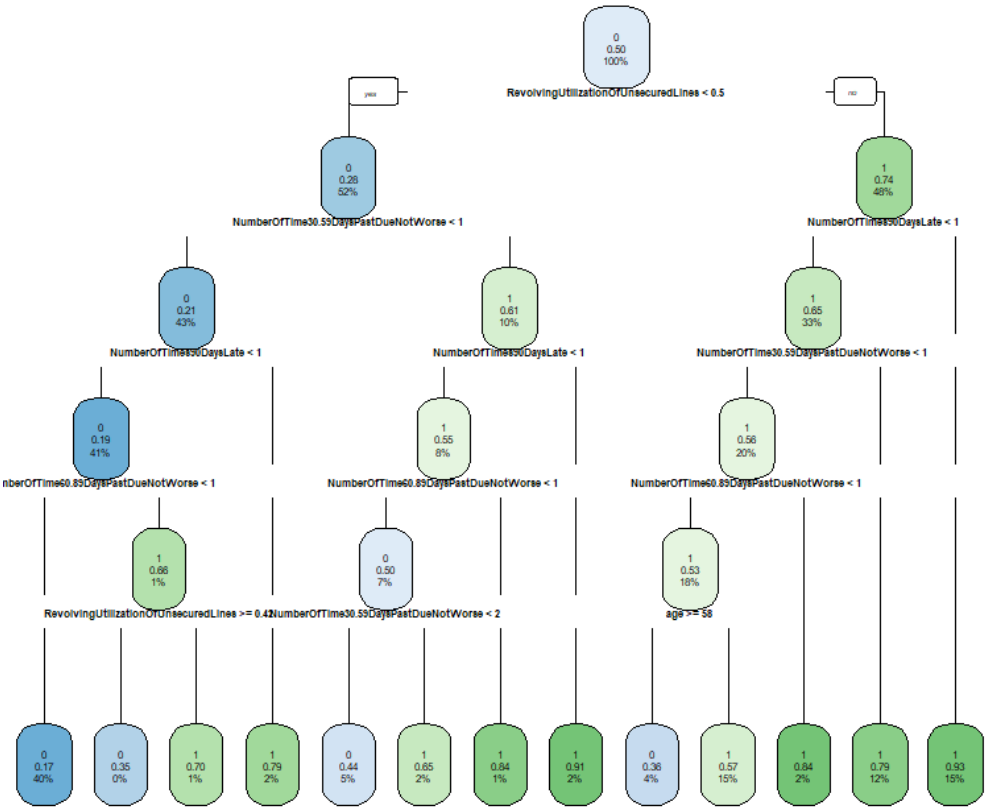

Figure 3.5: Decision Tree 3, the tuned tree.

This tree is far more complex than the previous tree, imposing two more levels of conditioning. The tree has more leaf nodes which predict value 0, which will hopefully benefit the predictive power for this value. Interestingly, one of the final conditions incorporates the age variable. At the sub-node, if an individual is at least 58 years old they are predicted value 0, otherwise they are predicted value 1. One unexpected condition is the final condition regarding the revolving variable, it does the opposite of what is expected and predicts 0 for those with a value above 0.42. This is symptomatic of an over-fitting to the training data at this sub-node, but the sub-node at which this split is performed is just 1% of the full training data, so it can be ignored. The predictive power of this tree can be tested. With the balanced testing set, this tree classifies 165 more 0s correctly (higher specificity), but wrongly classifies 125 more 1s as 0 (lower sensitivity). As a result, the proportion of those who are offered a loan who actually have a serious delinquency in the past 2 years is higher at 22.22%. A company might benefit from using this tuned tree. Using the representative testing set, the tree still has 1.79% acceptance of people who turn out to have had a delinquency whilst allowing 2471 more to have a loan who don't have a delinquency.

This is quite a significant amount of people, and a company might find that they can make more profit using this tree.

Concluding this section, three trees have been considered with added complexity between each one. The first tree predicts the 0 observations very well but has the highest acceptance rate of people who have had serious delinquencies, at 5.7%. The second and third trees offer fewer loans, but fewer of the offer-takers have had serious delinquencies (1.5 and 1.79% respectively).

3.3 Random Forest

3.3.1 Random Forest Modelling

As mentioned in the beginning of this section, one of the issues with basic decision trees is the use of the greedy algorithm. This ensures the Gini gain is maximised locally so some variables may never be used for a split. Random forest performs splits based upon random subsets of the variables which ensures more of these variables are considered. Additionally, since random forest is an amalgamation of many basic decision trees which are built on different subsets of the training data, it limits the sensitivity that comes from only fitting one decision tree.[11] As a result, random forest models are often more diverse and stable than basic decision trees. To test if there is any improvement in sensitivity using this model for the credit dataset, this section fits random forests and analyses them.

The training set used will be the same as for the final two models in the previous section, as this was effective at producing a reasonably well-fitted decision tree. 100 trees are used, each allowed up to 8 leaf nodes at first. One big issue with random forest is that it is a black box model, this means that this method does not give the user access to the internal functions as they are either hidden or very complicated. This means an easy-to-interpret plot cannot be produced for the model. The input data is passed through all the trees individually and the output is based upon the number of trees that predict value 1 or value 0 with this same input data, as if each had a vote and the class which is deemed the final prediction has the highest number of votes. This might not be ideal for a company, as if a person is rejected for a loan (predicted value 1), it will be more difficult to figure out why, as so many conditions are considered when you move through the many trees with your input data. Despite this, if the model fits well, it should still be considered, and the importance of each variable might give an indication of why an individual is rejected. The predictive power based on the representative testing data can be studied in the table below:

	True 0	True 1
Predict 0	19554	326
Predict 1	8416	1650

Table 3.8: The Random Forest 1 Confusion Matrix (representative test data.

This table can be compared to the confusion matrix for the second decision tree that was fitted (table 3.3). Of those offered the loan, a very slightly higher percentage are offered a loan (prediction of 0) truly have a serious delinquency in the past two years (1.64% vs. 1.5%), which is better than the tuned tree. 642 more loans are offered to those who have no serious delinquency in the past two years (69.9% sensitivity/ 83.5% specificity). An ROC curve for the random forest model shows that the best balance that can be achieved is 70.5% sensitivity/82.6% specificity which is very close to what we have above.

Since the second tree still has the lowest proportion of false positives with the representative test data, the random forest algorithm could be considered with the same maximum number of leaf nodes as the second tree has, to see if it can be improved in this way. Doing this gives a model with a proportion of 2.02%, why is this? One explanation is that since random forest also takes random subsets of the training data to train the trees, these subsets are much smaller than what's required to give a sensitive tree in both classes, and this is exacerbated by the fact that some of the significant variables are not considered as only a subset of the features are used. Conversely, the maximum nodes can be increased even more so that the individual trees that contribute to the random forest prediction fit their training data very well. One might expect that this leads to over-fitting, but since 100 trees are used, it can improve the predictions. For example, with 20 as the maximum number of leaf nodes per tree, we obtain a tree that has the following predictions with the representative test set:

	True 0	True 1
Predict 0	19596	289
Predict 1	8374	1687

Table 3.9: The Confusion Matrix for predictions when max. nodes=20 per tree using Random Forest.

Of the people who would be offered a loan (predict 0), 1.45% actually have value 1, this is lower than the second tree! It also has higher specificity/sensitivity (85.3%/70.06%). Despite this, there is the issue of reproducibility, as every time random forest is fitted in R it is done so randomly i.e. the same model is not fitted each time.

3.3.2 Importance Of Variables

For the final random forest model which gave the table above, the importance of variables is summarised as follows:

Variables	**Mean decrease in Gini coefficient**
RevolvingUtilizationOfUnsecuredLines	968.03
Age	103.95
NumberOfTime30.59DaysPastDueNotWorse	508.37
DebtRatio	17.58
MonthlyIncome	21.03
NumberOfOpenCreditLinesAndLoans	44.09
NumberOfTimes90DaysLate	516.99
NumberRealEstateLoansOrLines	17.13
NumberOfTime60.89DaysPastDueNotWorse	231.29
NumberOfDependents	2.26

Table 3.10: Importance of each variable with regards to the Gini coefficient.

The revolving utilisation variable is by far the most prevalent in all the trees. This is followed by the number of times someone is 90 days late on a payment, then the number of times someone is 30-59 days late on a payment. This is in line with what might be expected from the trees that were visualised, where revolving utilisation was used for the first condition, and the 90 days and 30-59 days late variables were often used for the conditions that followed.

Chapter 4

Discussions

4.1 Issues With The Modelling

One important issue with the modelling is with respect to the response variable itself. It was mentioned earlier that the Basel Committee on Banking Supervision defines default as a delinquency stage of 90 days or more. We used the serious delinquency variable in all our models, but many financial institutions instead consider if a person has been more than 90 days late with a repayment. If we wished to adapt the models to this, we could have done so by grouping the serious delinquency variable with the 90 day late variable and setting an individual to class 1 if they had either a serious delinquency or an over 90-day late payment in the past two years and assigning value 0 otherwise. Doing this may have led to more balanced data (there would be more individuals of class 1), which could lead to more powerful models. Secondly, many of these variables only take the last two years of credit history into account. There seems to be no obvious reason for choosing this particular length of time, and it doesn't make much sense that someone with potentially a number of serious delinquencies and late payments 3 years ago isn't a risk to default on a loan. Including a person's entire credit history and assigning a weight to a serious delinquency or late payment depending on the duration of time between each event could be an effective way to include all this information whilst not penalising individuals who have missed payments/had delinquencies a long time ago. Credit scores used in everyday life consider the entirety of a person's financial history. This is important, as recall that for the FICO score, payment history is the most significant aspect of the calculation. Lastly, it could be that a variable with a significant relationship to the response is not provided in our dataset. For example, information on new credit (has an individual opened a lot of new lines of credit recently?) might be useful.

4.2 How Do Our Models Compare With Existing Models?

In the introduction to this report, the five aspects of credit history and their respective weightings used to calculate the FICO score were given. The most important aspect was payment history, which in regard to our dataset represents the number of times someone is late with a payment and how long by (30-59, 60-89 or 90+ days). All of these variables are tested individually via conditions in all three of the basic decision trees that were fitted, and on top of this the importance for these three variables is below only the revolving utilisation of unsecured lines variable. This shows that payment history is of a similar if not higher significance in our modelling. Amounts

owed is the second most important aspect for the FICO score, representing a weight of 30%. This is a measure of how much money is owed through loans currently, and the variable that measures this best is the revolving utilisation of unsecured lines. This has the highest importance of any variable for both methods of modelling. Our models hence share the idea that these two features of an individual's financial history are the most important with the FICO scoring model. As for the remaining three aspects that are considered in this widely used scoring system: length of credit history is represented by age in our dataset and age has an importance that follows the variables mentioned above; credit mix is represented by the number of open lines and real estate lines, these have much lower importance in our dataset; new credit is not an aspect that can be investigated as none of the variables tell us any information about an individual's recent activity.

4.3 Which Model Is Best?

From a company's perspective, the 'best' model may depend on many factors. For example, a company may wish to have a simple-to-interpret white box model, where the pathways leading towards prediction are explicit, as in the case of the basic decision trees. They may alternatively require a quantitative risk in the form of a probability for each individual in order to reject a loan candidate based off a threshold probability/assign a maximum amount of money that they will offer. This requirement could be better met by a logistic model. Other examples of factors might revolve around profit aims and even staff numbers. All of the models accept only small percentages of individuals who turn out to have had a serious delinquency in the past two years (between 1.5 and 2%). The logistic and decision tree models are relatively simple to interpret and visualise. The logistic model assigns a percentage to each individual based off every predictor variable, which allows a company to judge risk on a case-by-case basis, whereas the decision tree model puts individuals into groups then bases risk off the number of each class of the response variable within the group and the random forest model bases its predictions off the judgements of many decision trees. The threshold in the logistic model can be changed easily to give a higher sensitivity or specificity using the ROC curve. For basic decision trees, since each group is assigned a probability, a choice of threshold is unlikely to give such flexibility, and for random forest it was found with ROC that any threshold chosen does not improve the balance achieved by the logistic model. All of these models could still be improved further if more 1s were available in the dataset. In conclusion, the logistic model provides much more information regarding a single individual's risk, and also has flexibility in the form of the threshold which allows you to pick a desired sensitivity or specificity.

4.4 Ideas For Further Research

4.4.1 Extensions On Our Modelling

The discussions section above mentions that in finance, the serious delinquency variable may not be sufficient to indicate high risk, and the number of times someone is 90 days late on a payment could be combined with this variable to create a new response. This report does not investigate the effect that this would have on the modelling, but this could be worthwhile to consider as the data is highly imbalanced for the response variable we have used. Of course, the models could be improved in other ways, for example by increasing the number of 1s in the dataset or weighting 0 observations less than the 1s (using for example grid search). As well as this, a cost function might be considered in order to see how much profit a company might make using each model/how much money they should offer each individual to maximise profit. Below we discuss

a couple of other models which could be considered, but many other models could be studied (e.g. ensemble/bagging/bayesian...)

4.4.2 Neural Networks

One example of a different method of modelling that could have been applied to our dataset is neural networking. A neural network is an algorithm inspired by the human brain, in which neurons work together to send and receive signals. An artificial neural network is a system comprised of two main structures, the node and the link. Most neural networks are made up of three layers: the input layer, the output layer, and the hidden layer. The input layer is presented with raw data i.e. the independent variables in our dataset. The information is then passed to the hidden layer where it is processed and sent to the output layer, giving the prediction. The predictors of the input are multiplied by link weights (changing the values of these 'weights' changes the behaviour of our model), which are summed and passed through an activation functionto predict if a person will default on a loan. Neural networks are used widely for these classification problems, so it could be worth training a model and seeing how well it fits our data.

4.4.3 Support Vector Machine

Support Vector Machines are similar to logistic regression in that they both attempt to find the optimal hyperplane (the plane in variable space which separates sets of points) separating two classes of the response. More specifically, Support Vector Machines search for a hyperplane that maximises the margin (this represents in a way the distance between the two classes of points, a larger distance means the model will give more accurate predictions). This could be useful for our dataset as this type of model works well in high-dimensional spaces (we have many variables), and tends to have better accuracy in results. This method is popular in this area of research, there are many papers which consider it.

Chapter 5

Conclusion

This report aimed to build and analyse machine learning algorithms which are able to predict the probability that an individual will default on a loan. Each model used the same response variable, namely whether or not an individual has had a serious delinquency in the previous 2 years. If they have, they are not to be offered a loan. In this way, the models attempted to minimise the number of people offered a loan who turn out to have had a serious delinquency. After cleaning the data, the logistic model was fitted and analysed, followed by decision trees and random forests. Out of the three, logistic regression offers the most tailored individual probability and using the ROC curve, a threshold of 0.06336881 gave a balance in sensitivity and specificity of 76.9% and 77.6% respectively. Higher sensitivities could be achieved by choosing different thresholds. A decision tree providing a sensitivity of 85.3% and 67.6% specificity was found which ensured just 1.5% of those offered a loan had serious delinquencies, along with a random forest model with 85.4% and 70.1% sensitivity/specificity respectively. For all of the above, the variable representing the proportion of credit an individual is currently utilising is the most important/useful in determining the response, followed by the variables which represent history of late payments. This is similar to models used in the financial sector, one example being the FICO score, which has payment history as the most significant aspect followed by amounts owed. On the other hand, despite being significant in the logistic model, credit mix (the number of different lines of credit an individual has) had much less impact on the measurement of risk and length of credit history was not considered until decision trees got more complex. The dataset investigated was highly imbalanced which may have affected the accuracy of the models. If the dataset were to be studied further, the variables for delinquency and very late payments could be combined to deal with this. Variables representing new credit could also be added to check for a relationship with risk, and alternatives to variables like age might be considered (being older does not guarantee you have longer credit history). Finally, other popular choices of models including support vector machines might be fitted, as this is a standard model to deal with this kind of problem.

Chapter 6

Appendix

Variables	1	2	3	4	5	6	7	8	9	10	11
1	1	0.27	-0.11	0.27	-0.01	-0.02	-0.02	0.31	0	0.27	0.05
2	0.27	1	-0.28	0.24	-0.01	-0.03	-0.16	0.24	-0.07	0.2	0.09
3	-0.11	-0.28	1	-0.07	0.02	0.03	0.14	-0.08	0.03	-0.07	-0.22
4	0.27	0.24	-0.07	1	0	0	0.08	0.22	0.04	0.31	0.07
5	-0.01	-0.01	0.02	0	1	-0.02	0.05	-0.01	0.12	0	-0.04
6	-0.02	0.03	0.03	0	-0.02	1	0.09	-0.02	0.12	-0.01	0.07
7	-0.02	-0.16	0.14	0.08	0.05	0.09	1	-0.09	0.43	-0.02	0.07
8	0.31	0.24	-0.08	0.22	0.01	-0.02	-0.09	1	-0.06	0.29	0.03
9	0	-0.07	0.03	0.04	0.12	0.12	0.43	-0.06	1	-0.02	0.13
10	0.27	0.2	-0.07	0.31	0	-0.01	-0.02	0.29	-0.02	1	0.04
11	0.05	0.09	-0.22	0.07	-0.04	0.07	0.07	0.03	0.13	0.04	1

Table 6.1: The Pearson correlation between each of our variables after cleaning the data.

Variables	**VIF Value**
RevolvingUtilizationOfUnsecuredLines	1.22
Age	1.15
NumberOfTime30.59DaysPastDueNotWorse	1.13
DebtRatio	1.06
MonthlyIncome	1.23
NumberOfOpenCreditLinesAndLoans	1.53
NumberOfTimes90DaysLate	1.12
NumberRealEstateLoansOrLines	1.45
NumberOfTime60.89DaysPastDueNotWorse	1.11
NumberOfDependents	1.06

Table 6.2: Table of VIF Values for the explanatory variables.

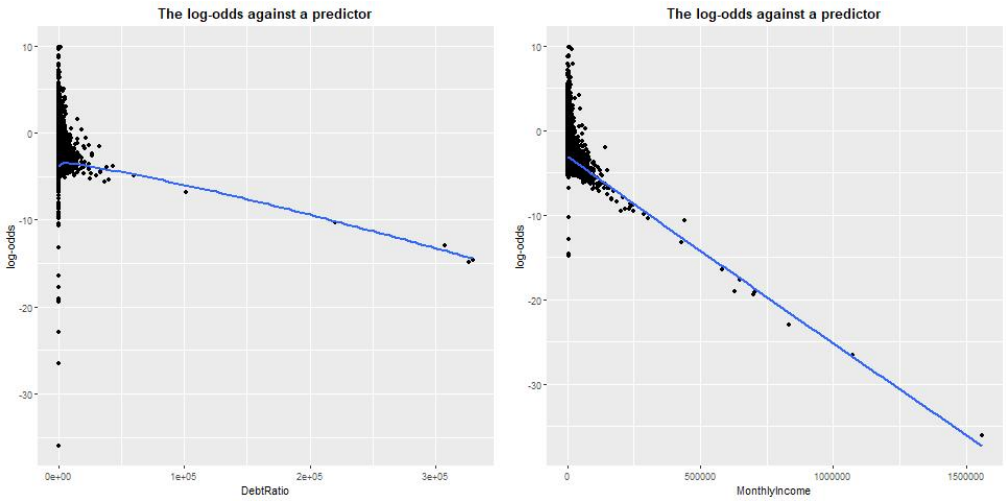

Figure 6.1: Scatterplots showing the linear relationship between the variables DebtRatio/MonthlyIncome and the log odds.

	True 0	True 1
Predict 0	32112	638
Predict 1	9843	2325

Table 6.3: The Confusion Matrix using a threshold of 0.1494658.

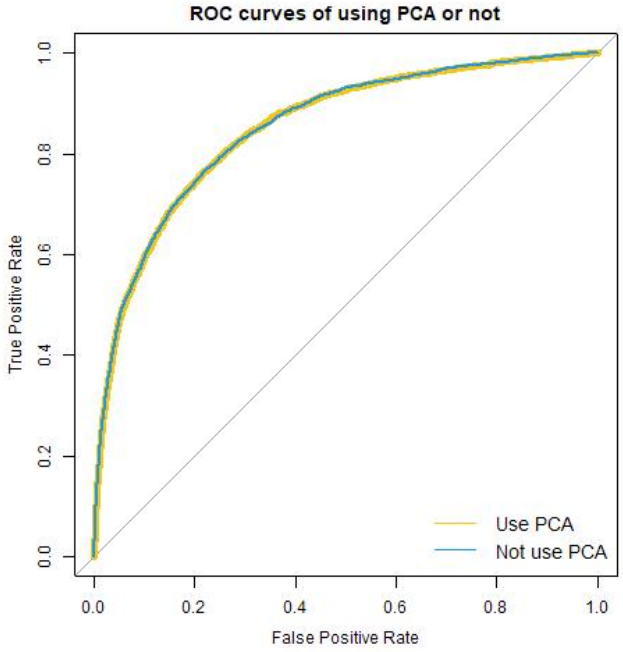

Figure 6.2: ROC curves with PCA applied or not.

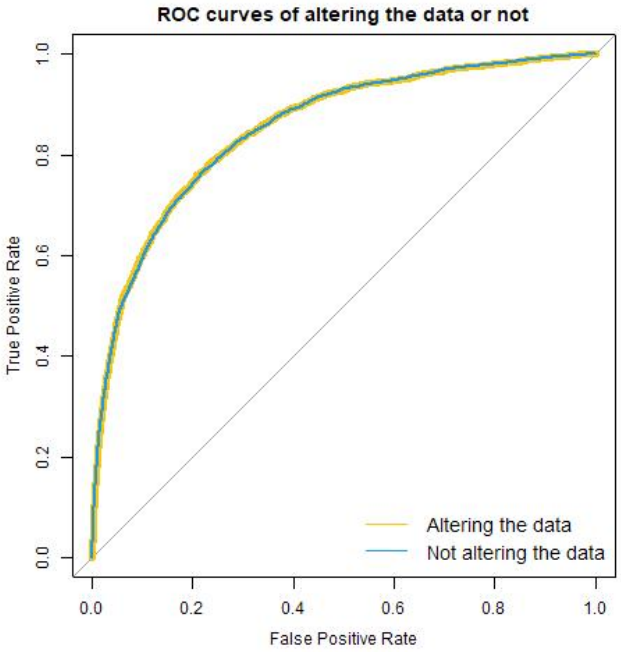

Figure 6.3: ROC curves when altering the data or not.

Bibliography

[1] Salome Tabagari. *Credit scoring by logistic regression.*
https://core.ac.uk/download/pdf/79110695.pdf

[2] *FICO Credit Score Algorithm.*
https://www.dourish.com/classes/infx161f14/slides/2014-10-28-team14-FICO.pdf

[3] Louis DeNicola. *What is a Good Credit Score?.*
https://www.experian.com/blogs/ask-experian/credit-education/score-basics/what-is-a-good-credi

[4] Maria Fernandez Vidal, Fernando Barbon. *Credit Scoring In Financial Inclusion.*
https://www.cgap.org/sites/default/files/publications/2019-07-Technical-Guide-CreditScore.pdf

[5] Gang Dong, Kin Keung Lai, Jerome Yen. *Credit scorecard based on logistic regression with random coefficients.*
https://core.ac.uk/download/pdf/81135378.pdf

[6] Nguyen Chi Dung. *An Application of Credit Scoring: Developing Scorecard Model for A Vietnam Commercial Bank.*
https://rpubs.com/chidungkt/442168

[7] Statistics Solutions. *Assumptions of Logistic Regression.*
https://www.statisticssolutions.com/assumptions-of-logistic-regression/

[8] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning.*
http://www.springer.com/series/417

[9] Lkhagvadorj Munkhdalai, Tsendsuren Munkhdalai, Oyun-Erdene Namsrai, Jong Yun Lee, and Keun Ho Ryu. *An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments.*
https://www.mdpi.com/2071-1050/11/3/699

[10] Mathworks. *Decision Tree Learning.*
https://en.wikipedia.org/wiki/Decision_tree_learning

[11] Deloitte. *Using Random Forest for credit risk models.*
https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/financial-services/deloitte-nl-fsi