LUNDS UNIVERSITET

CREDIT RISK MODELLING

An IRB & Machine Learning Approach

In collaboration with:

Grant Thornton

Author
Jan Muller

Supervisor
Magnus Wiktorsson

Faculty of Science
Department of Mathematical Statistics
Bachelor of Science Degree Project
January, 2020

Table of Contents

Abstract

The following document outlines the development process for an Internal Ratings Based (IRB) probability of default (PD) model for Prosper, a peer to peer lender during the period 2005 - 2014 . The data is prepared and analysed using suitable single factor analysis techniques accompanied with a heuristic qualitative business approach to credit risk modeling. The chosen variables undergo further manipulation using recursive partitioning followed by forward logistic regression to output the selected variables and their coefficients for the PD model. These logit scores obtained from the logistic regression are mapped to corresponding probabilities of default and an adequate score per consumer. These scores are then blocked into suitable grades which are at the discretion of the user. The most suitable choice of number of grades and grade ranges is evaluated during calibration and optimisation. Finally, the estimated capital needed to comply with Capital Requirements Regulation is calculated.

Keywords: *Credit Risk, Vasicek, Default, Expected Loss, Capital Requirements, Scorecard, Segmentation, Clustering, Calibration*

Acknowledgements

I would like to thank the people at Grant Thornton for their continued support and encouragement during my stay at the company. I felt very welcomed and a lot of time was dedicated to this project through the busy working times. In particular, special thanks to Simon Wahlstrom for his encouragement and detailed advice as my "buddy" during my stay at the firm. Finally, I would like to thank my supervisor Magnus Wiktorsson whose student focus and dedication as the director of studies ensured the project could be carried out smoothly and efficiently, with ample enough time to discuss the sensitivities of politics, nuclear reactors and solar storms.

List of Figures

List of Tables

List of Abbreviations

AIRB Advanced Internal-Ratings Based Approach

ANOVA Analysis of Variance

AUC Area under the Curve

CART Classification and Regression Trees

EAD Exposure at Default

EL Expected Loss

FIRD Foundational Internal-Ratings Based Approach

GDR Grade Default Rate

ID3 Iterative Dichotimiser 3

IG Information Gain

IV Information Value

KS Kolmogorov-Smirnov

LRA Long Run Average (Default Rate)

LGD Loss Given Default

ODR Observed Default Rate

PD Probability of Default

pdo Points to Double Offset

P2P Peer to Peer

ROC Receiver Operating Characteristic

RWA Risk Weighted Assets

SA Standard Approach

SFA Single Factor Analysis

SQP Sequential Quadratic Programming

UTP Unlikely to Pay

WCDR Worst Case Default Rate

WGSS Within-groups Sum of Squares

WoE Weight of Evidence

Introduction

Credit Risk models have a range of practical applications, raning from a financial institutions motivation and need to develop quantitative estimates of the amount of economic capital required to support their risk taking endevours to provisioning and in turn deciding how much retained profits the bank will have to offset against the risk weighted assets. Banks can further utilise these models to make credit decisions and for the pricing of loans by choosing a suitable interest rate.

For the purposes of this report, the credit risk model will be utilised to estimate the amount of economic capital required to support their risk taking endevours. This is achieved by attempting to assess current and future credit risk exposures across all asset classes by determining the chance that a borrower will be 90 days past due on the repayment of their line of credit or deemed unlikely to pay (UTP).

This risk, that a lender may endure is assessed on the basis of not receiving their interest due or the principle loaned on time. These indications, that a loan may not be fully repaid spurs the bank to create provisions or reserves to account for these defaults. This leads to an evaluation of the expected loss (EL), which can be expressed by the following equation:

$$\mathrm{EL} = \mathrm{PD} \times \mathrm{LGD} \times \mathrm{EAD} \tag{1}$$

where PD, LGD and EAD represent the probability of default, the loss given default and the exposure at default respectively. Thus, expected loss or EL is a combination of the value of the loan or EAD, the probability that the loan will default or PD and finally in an event that default occurs, an estimation of the part of the loan which will be in lost or LGD. This paper will focus on the estimation of the probability of default. This expected loss corresponds to the mean value of the credit loss distribution and as this average can be easily exceeded, the bank must have enough capital in reserve to fully cover any unexpected losses. These unexpected losses are defined as the difference between the expected loss and a 99.9% quantile obtained from the Vasicek Loan Portfolio Model [9]. This is further illustrated by the following diagram:

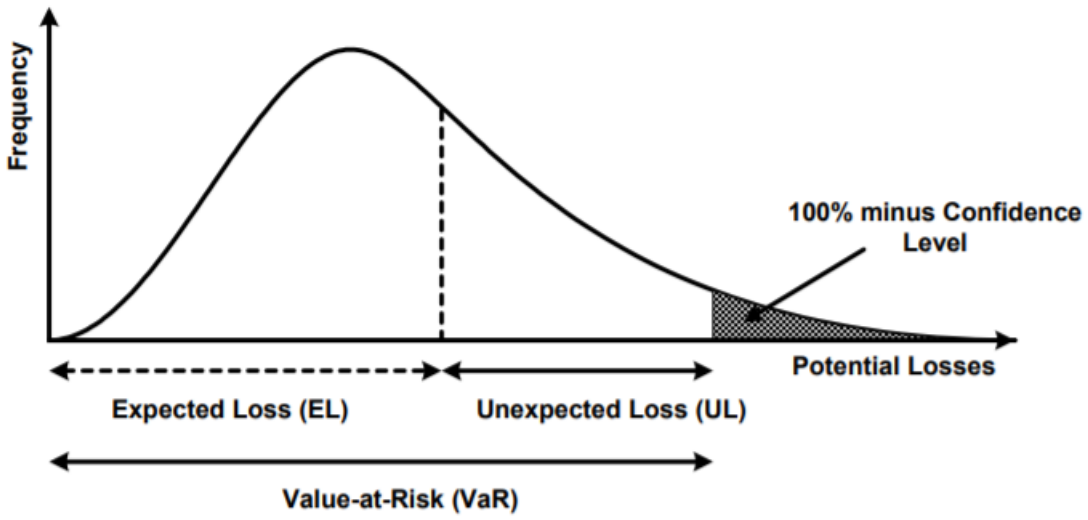

Figure 1: An illustration of the Credit Loss Distribution, which is derived from the Vasicek Loan Portfolio Model.

Furthermore, The Basel Committee on Banking Supervision sets the global standard for the regulation of banks as well as providing a forum for regular cooperation on banking supervision matters. They have provided three established models which are used to determine the capital requirement for credit risk. This

includes the Standard Approach (SA), the Foundational Internal-Ratings Based Approach (FIRD) and the Advanced Internal-Ratings Based Approach (AIRB), all of which are determined and evaluated using the following formulae:

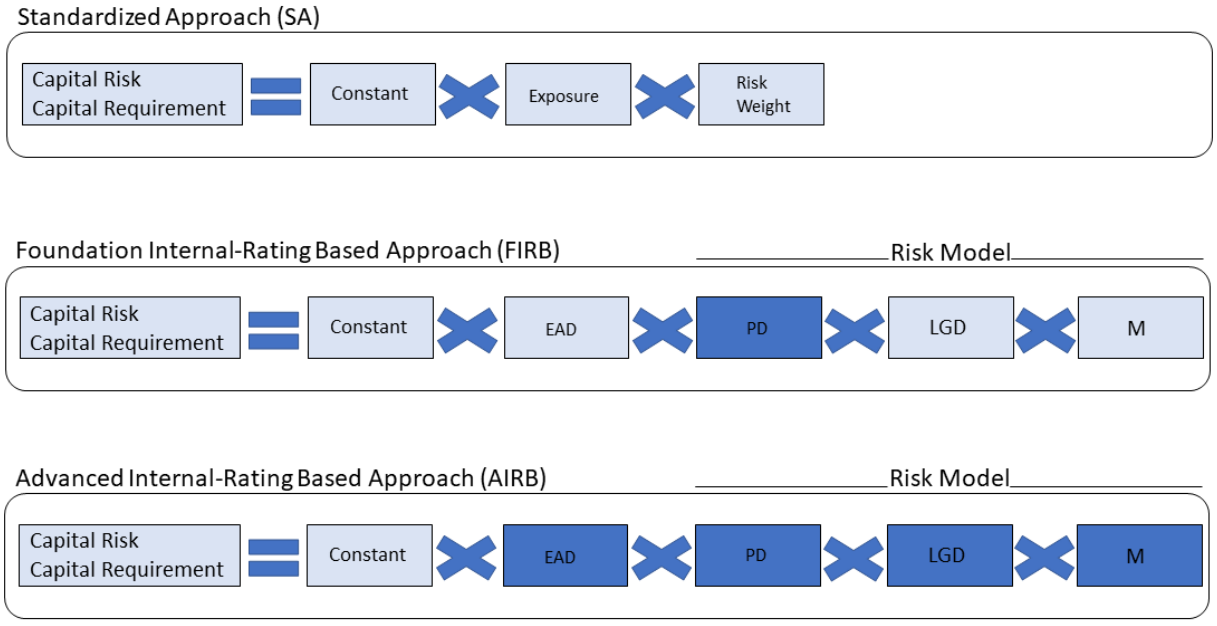

Figure 2: Outline of Internal Ratings Based critera as recommended by the Basel Committee.

where the boxes in blue represent areas of the modeling process which financial i n stitutions w i ll model themselves. For the basis of this report, which will outline the methods and strategy in developing a probability of default model will utilise the FIRB approach in determining the capital requirements for a financial institution [7].

The Dataset

Prosper lending, founded in 2005, is a peer to peer lender (P2P) based in San Francisco offering the first P2P lending marketplace in the United States. Peer to peer lending, or crowdlending, is the practice by which money is lent to individuals or businessess through online services whereby it matches suitable lenders with borrowers. Due to the fact that the company does not extend credit themselves, it allows the business to function with lower overheads and provide a cheaper service in comparison to traditional financial institutions. Thus, it is common and typical for their to be more than one lender on any given loan.

The current dataset provided by prosper contains the default status, as well as 80 other candidate variables for all borrowers between 2005 and Q1 2014. In this time, Prosper has facilitated more than $14 billion in credit through an excess of 910,000 facilities. As previously stated, borrowers are matched to lenders via an online service and are offered a fixed r a te, fi xe d te rm lo an be tw een $2 ,0 00 - $4 0, 000 re sp ectively. So me of the leading investors include Sequoia Capital, Francisco Partners, Institutional Venture Partners, and Credit Suisse NEXT Fund.

Due to the structure of the data, it was not possible to conduct a through the cycle analysis as is usually performed in practice and instead a point in time model was constructed using information gathered at origination and through the life of the borrowing facility. In accordance with regulations, a line of credit is considered defaulted once it surpasses 90 days past due [3]. Thus, for the purposes of this project and the

Prosper dataset, default was defined by acccessing the categorical factors within the variable *LoanStatus* in the following way:

> A borrower is considered defaulted if they are listed as "Defaulted", "Past Due (91-120 days)", "Past Due (>120 days)" or "Charged Off". All other class labels are considered as non-defaulted at the time the credit listing was pulled.

This indicates that there are 17,330 defaulted observations and 96,607 non-defaulted, corresponding to a 15.2% default rate at the time the credit listings were pulled.

Single Factor Analysis

Variable Selection

As previously stated, the default rate of 15.2% can be divided into defaulted and non-defaulted cases , as illustrated by the histogram below:

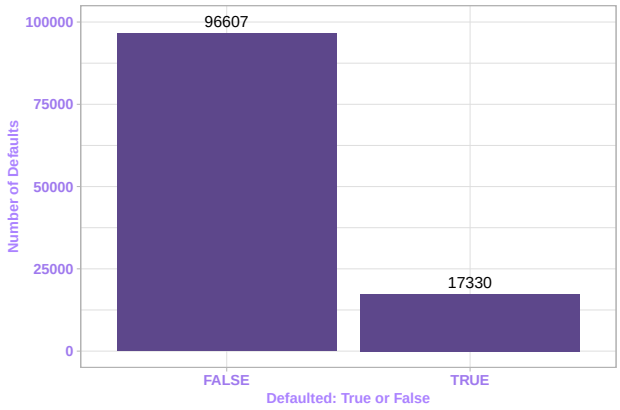

Figure 3: An Illustration of the defaulted and non-defaulted borrowers at the time the credit listing was pulled for the Prosper peer to peer lending institution.

Of the 83 eligible variables located in the dataset, the aim was to indicate the most suitable candidates which would be dependent variables in determining default in the modeling process. This was achieved through a number of operations. To start, as the dataset contains superflous future information, these variables were removed from the modeling process. In addition, there were a number of third party as well as prosper rating metrics that were similarly removed from the dataset. This was done in order to perform an independent analysis of the dataset. Furthermore, the dataset contained variables such as $ListingKey$, $LoanNumber$ and other forms of ordering variables which could not be used to indicate and model a borrowers likelihood of default and were similarly removed. Finally, from the remaining candidates, the dataset contained a lot of missing information and it was decided that variables with missing rates above 25% to be removed from the dataset. This was a judgement decision based on the interpretation that it was assumed that default predictablitiy would substantially decrease for variables with missing rates above 25%. That being said, it must be noted that 48 variables had no missing rate at all, and 35 with some missing information. The distribution of missing rates is illustrated as follows:

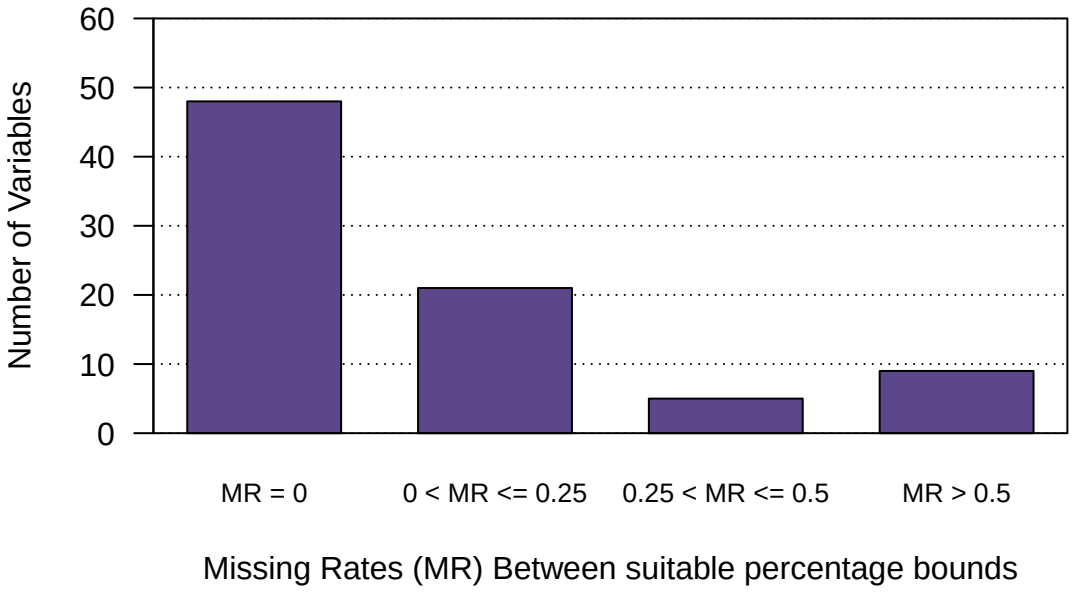

Figure 4: An Illustration of the number of variables with missing rates, expressed as a percentage, contained within certain suitable groups.

Information Value

Now that a sufficient amount of superfluous information was removed, it was possible to conduct more substantial single factor analysis. A further criterion for each model candidate was a significant information value. The information value ranks variables on the basis of their importance in a predictive model and is a particularly useful slection method when performining binary logistic regression. It is calculated in the following way:

$$\mathrm{IV} = \sum_{i=1}^{m} (PG_i - PB_i) \times \mathrm{WOE}_i, \tag{2}$$

where PG_i and PG_B are the percentage of goods (or non-defaulted cases) and bads (defaulted cases) in bucket i respectively [11]. It must be noted that the IV is sensitive and its value increases to the choice of bins and as such should be used with caution when there are very few events and non-events.

The weight of evidence (WOE) on the other hand measures the strength of a grouping relating to risk. In the case of credit risk modeling, it indicates whether a particular group state affects the risk of default. It is a measure of the distribution of goods and bads within this grouping and is calculated in the following way:

$$\mathrm{WOE}_i = \ln\left(\frac{\frac{NG_i}{TG}}{\frac{NB_i}{TB}}\right) \tag{3}$$

where NG_i and NB_i represent the number of goods and bads in bucket i, whilst TG and TB represent the total number of goods and bads respectively. The information values obtained can thus be categorized and their indicative predictiveness expressed by the following table:

Table 1: Information Value Table Identifying the range of IV's and their significance level.

Information Value	Predictive Power
x < 0.02	Insignificant
0.02 < x < 0.1	Weak
0.1 < x < 0.3	Medium
x > 0.3	Strong

It was decided that all variables with an IV greater than 0.2 were deemed as relevant and seem to be a good candidate for determining default. The information values were calculated and are displayed in the following figure:

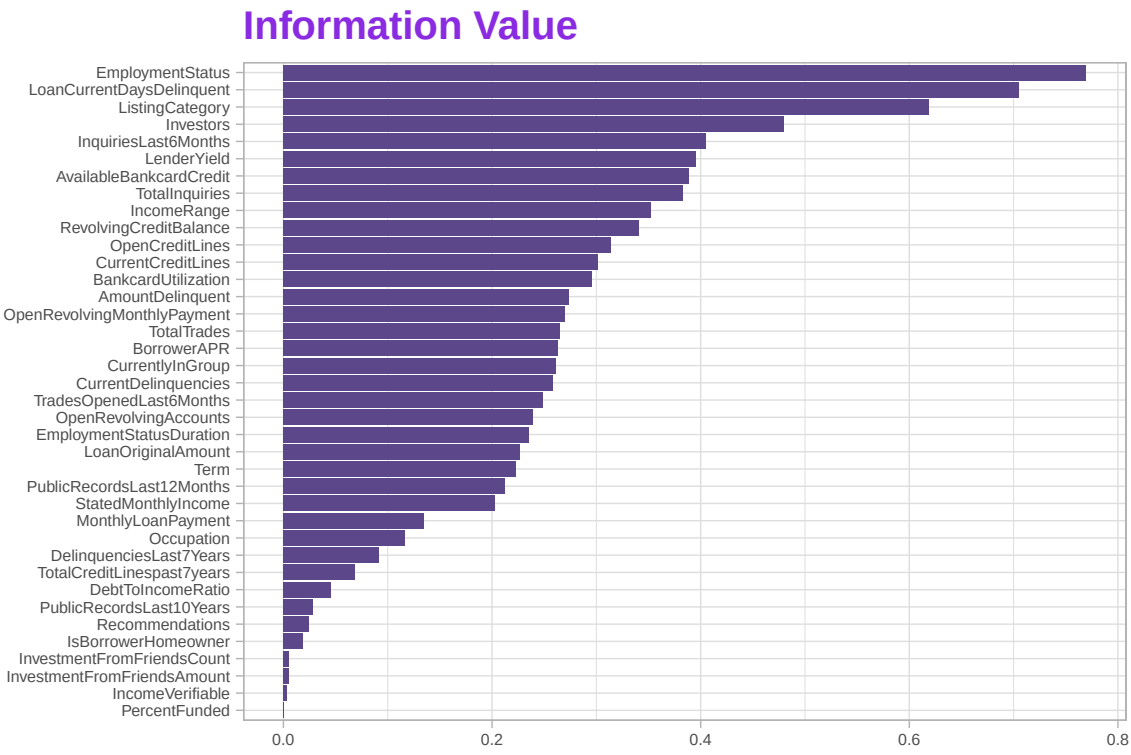

Figure 5: An illustration of the Iinformation Values for the remaining variables after having conducted Single Factor Analysis.

Correlation

Once the pertinent variables were removed, the remaining continuous variables were checked for cross-correlation. It was assumed that the categorical variables were considered independent from one another. The correlations of the remaining variables were determined, producing the following plot:

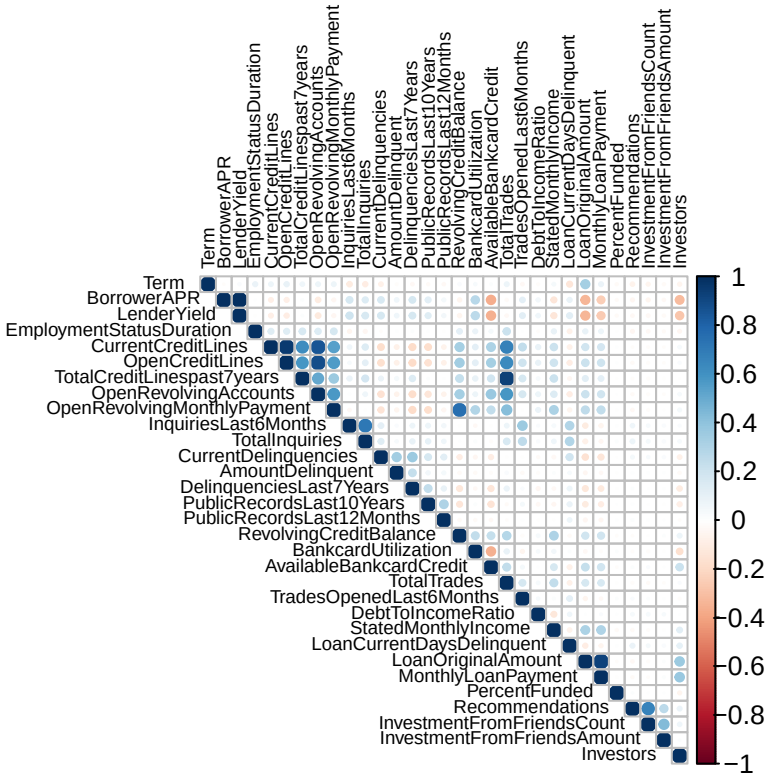

Figure 6: Determination of Correlation for the remaining continuous variables within the Prosper dataset.

From the plot above, the size of the circles as well as their colour indicate the degree of correlation.In an effort to make the model as simple as possible, variables with a high correlation, which was deemed to be the absolute value of 0.7 or higher, were analysed and compared to one another. Once the highly correlated variables were assessed, it was decided to remove variables from the dataset by comparing their missing rates and information values. In these cases, business intuition was also called upon to determine which of the potential candidates were deemed most suitable from a business perspective and the other candidates were removed from the model.

Segmentation Analysis

The majority of non-pertitent variables have now been removed from the dataset and suitable model selection variables can now be determined in order to develop a scorecard.

Scorecards are one of the most widely used credit risk analysis tools which allocate a score to each borrower based on their credit risk which is determined based on their characterstics. This can range from age, employment status, marital status etc. Thus, these scorecards can be used to distinguish between characteristics which can affect a borrowers ability to repay a line of credit.

Once the single factor analysis was completed, segmentation was performed to investigate whether it was possible to construct multiple scorecards for the dataset. If the segmentation process yielded different risk drivers, multiple scorecards could refine the different segments of the portfolio.

Thus, the goal of segmentation is to assess whether there exists groups of borrowers which have different risk characteristics than others. If found, it would provide the opportunity to produce multiple scorecards. These scorecards would more accurately depict the credit risk an institution would be taking on and would add an extra layer in refinement in determining the amount of capital which need to be held in reserves. Of course, there is a judgement call in such cases due to the limited lifetime of any current scorecard. In

addition, the risk drivers need to be substantial with an adequate population and consistent default rate to be deemed a suitable candidate.

Multiple scorecards could similarly impact risk as well as yield a higher return of investment. A number of approaches were taken in refining the dataset. These included:

- A Heuristic Approach
- K-Means Clustering
- Decision Trees

Furthermore, it must be noted that some parts of the banks financial statements are management differently, where they have different risk profiles. In these cases, apart from the statistical benefits, segmentation can be of benefit due to its benefits to capital.

Heuristic Approach

To start, a heuristic or business style approach was taken in order to determine if there were some obvious candidates present which could be used for segmentation. The borrowers' employment status, income range as well as their listing type, or $ListingCategory$, seemed like apt and reasonable initial attempts at segmentation. The data was manipulated by regrouping the categorical variables into more intuitive groups and to analyse the default rates across the different segments. This produced the following results:

Table 2: Heuristic Segmentation of Suitable Candidate Variables.
Compare Bad Rates Across different intuitive segmentation choices.

	Employed[1]	Unemployed[1]	Other[1]	Unsegmented[2]
ListingCategory				
Debt	0.1039 (0.57)	0.2606 (0.01)	0.0746 (0.02)	0.11 (0.07)
Home	0.1115 (0.06)	0.2381 (<0.01)	0.0932 (<0.01)	0.39 (0.15)
Other	0.1202 (0.18)	0.2534 (<0.01)	0.1527 (0.01)	0.12 (0.19)
Missing	0.3852 (0.08)	0.4135 (<0.01)	0.4054 (0.07)	0.1 (0.6)
IncomeRange				
Other	0.0539 (0.01)	0.1171 (0.02)	0.4054 (0.07)	0.39 (0.08)
High	0.1001 (0.56)	0.1893 (0.32)	0.3284 (<0.01)	0.1 (0.57)
Low	0.1775 (<0.01)	0.3354 (0.01)	0.2602 (0.01)	0.19 (0.35)

[1]Default Rates with Percentage of population between Employed,Unemployed and Other in brackets.
[2]Same as above but the Unsegmented column is relative to the total population.

It can be seen that there are issues regarding the percentage of population in each respective bin given the current design. In addition, in several cases where default rates could be considered relevant with a reasonable population size relative to the total poulation, the default rates are comparable to that of the unsegmented case. It must be noted that the discretization choices or variables chosen may have not been suitable and could have been replaced with other candidate variables. The conclusion is that further analysis is required.

K-Means Clustering

Once the heuristic approach bore little fruit, it was decided to utilise a variety of clustering techniques in an effort to distinguish between different characteristics of risk drivers between the borrowers. One such suitable method is that of K-means clustering.

Formal Definition

The K-Means algorithm is an unsupervised learning algorithm (there are no target labels) which attempts to identify, group or cluster data points within the dataset. The greater the similarity, or homogenitiy within each of these groups, the greater the difference between each group and as a result, would produce a more distinct clustering.

The algorithm works by randomly initilizing k starting centroids. Each data point is then assigned to its nearest centroid by a chosen metric, which in the case of this project is Euclidian distance. Following on from this, the centroids are re-evaluated by computing the mean of the data points within each respective cluster. These steps are repeated until the pre-determined stopping criteria is triggered. More formally, K-means utilises expectation-maximisation to allocate each datapoint to a specific cluster. This is illustrated mathematically as follows:

Let $X = \{x_i\}_{i=1}^{N}$ be the set of data points and $\mu_k = \{\mu_i\}_{i=1}^{k}$ be the k set of cluster centroids. The objective function J is defined as:

$$J = \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}||x_i - \mu_k||^2, \quad w_{ik} = \begin{cases} 1, & x_i \in k \\ 0, & x_i \notin k \end{cases} \tag{4}$$

where K is the number of clusters, N the total number of data points, μ_k the cluster centroid and w_{ik} an indicator function if a datapoint belongs to cluster k. Thus the methodology underlying the k-means algorithm is a minimisation problem of two parts. The E-step, which assigns data points to the closest cluster is performed by differentiating the objective function with respect to w_{ik} first to update the cluster assignments [6]. Thus:

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{N}\sum_{k=1}^{K} ||x_i - \mu_k||^2 \tag{5}$$

$$\implies w_{ik} = \begin{cases} 1, & k = \text{argmin}_k ||x_i - \mu_k||^2 \\ 0, & \text{o.w.} \end{cases} \tag{6}$$

The M-step then entails recomputing the centroids after the clustering assignments whereby:

$$\frac{\partial J}{\partial \mu_k} = 2\sum_{i=1}^{N} w_{ik}||x_i - \mu_k|| = 0 \tag{7}$$

$$\implies \mu_k = \frac{\sum_{i=1}^{N} w_{ik} x_i}{\sum_{i=1}^{N} w_{ik}}. \tag{8}$$

The K-means clustering approach offers a number of advantages. As well as being fast, robust and easy to understand, it yields the best results when datasets are distinct or well separated from each other. From a computational standpoint, it is relatively efficient with a timeline of $\mathcal{O}(tknd)$, where n is the number of objects, t the number of iterations, d the number of dimensions of each object and k the number of clusters respectively.

In contrast, the main disadvantage is the apriori specification of the number of cluster centers. Furthermore, the algorithm does not hold for a non-linear dataset, has difficulties handling noisy data as well as outliers and cannot be applied to categorical data. The latter is later circumvented with the use of decision trees.

Metrics

As previously stated, the K-means algorithm requires a pre-specification of the number of clusters. A resolution to this issue is the introduction of the elbow plot. The elbow plot is a visual description of the WGSS or within-groups sum of squares across a a range of choices for k. It is at the discretion of the user to then choose the most suitable k in comparison to the WGSS. In an ideal setting, there is a sharp decline in the WGSS at a suitable k, yielding a bend in the graph, hence the name elbow plot.

In addition to this, the average silhouette method and gap statistic was applied. The silhouette method attempts to interpret the quality of a clustering by measuring how similar an object is to its own cluster in comparison to the other clusters. The silhoutte ranges between minus one and one for each data point, where a high value illustrates that the data point is well matched to its own cluster in relation to the other cluster choices. The clustering configuration is considered appropriate when a high value is obtained for the majority of the dataset.

More formally, for each data point $i \in K_i$ where K represents the number of clusters, the distance between a data point i and all other points in the same cluster can be calculated by letting:

$$a(i) = \frac{1}{|K_i| - 1} \sum_{j \notin K_i, i \neq j} d(i, j), \tag{9}$$

where d(i,j) represents the distance between the points i and j in the cluster K_i. Thus a(i) measures how well the data point is assigned to the cluster. Following on from this, define a function b(i) which measures the mean dissimilarity of a data point i with another cluster K_n in the dataset where $K_n \neq K_i$. Thus, for each point $i \in K_i$, we can define:

$$b(i) = \min_{n \neq i} \frac{1}{|K_n|} \sum_{j \in K_n} d(i, j), \tag{10}$$

to be the smallest mean distance between a point i to all other points in neighbouring clusters. The silhouette function s(i) can thus be defined as :

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & a(i) < b(i) \\ 0, & a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & a(i) > b(i) \end{cases} \tag{11}$$

It is then clear from the above formula that $-1 \leq s(i) \leq 1$.

In contrast the gap statistic method compares the intracluster variation for different possible values of k by utilising their expected values under a reference distribution of the data under a constraint that they show no obvious clustering patterns. This reference dataset is generated using Monte Carlo simulations by computing the range for each point in the dataset and generating n points uniformly between this interval. Thus, the gap statisic can be computed for a given cluster k as follows:

$$Gap_n(k) = E_n^* \log(W_k) - \log(W_k), \tag{12}$$

where W_k is the within cluster dispersion and E_n^* is the expected value of the reference distribution with a sample size of n. This is achieved by utilising bootstrapping techniques and applying the following algorithm:

1. Computing W_k by first clustering the data into k_i clusters i times where $i = 1, ..., k_{max}$.
2. Generate N reference datasets, where they are clustered in a similar way and the gap statistic is calculated.
3. Let $\bar{\omega} = \frac{1}{N}\sum_n \log(W^*_{kn})$ and compute the corresponding standard deviation such that

$$sd_k = \sqrt{\frac{1}{n}\sum_n \log(W^*_{kn}) - \bar{\omega})^2}, \quad s_k = sd_k \times \sqrt{1 + \frac{1}{B}}. \tag{13}$$

4. Choose the most suitable number of clusters k as the smallest k whereby:

$$Gap(k) \geq Gap(k+1) - s_{k+1}. \tag{14}$$

Implementation of K-Means Metrics

In the case of the prosper dataset, the aforementioned metrics were applied in an effort to determine the most suitable choice of k. For the ease of convenience, a sample of 10,000 borrowers were segmented from the data and illustrated by the figure below. The results are as follows:

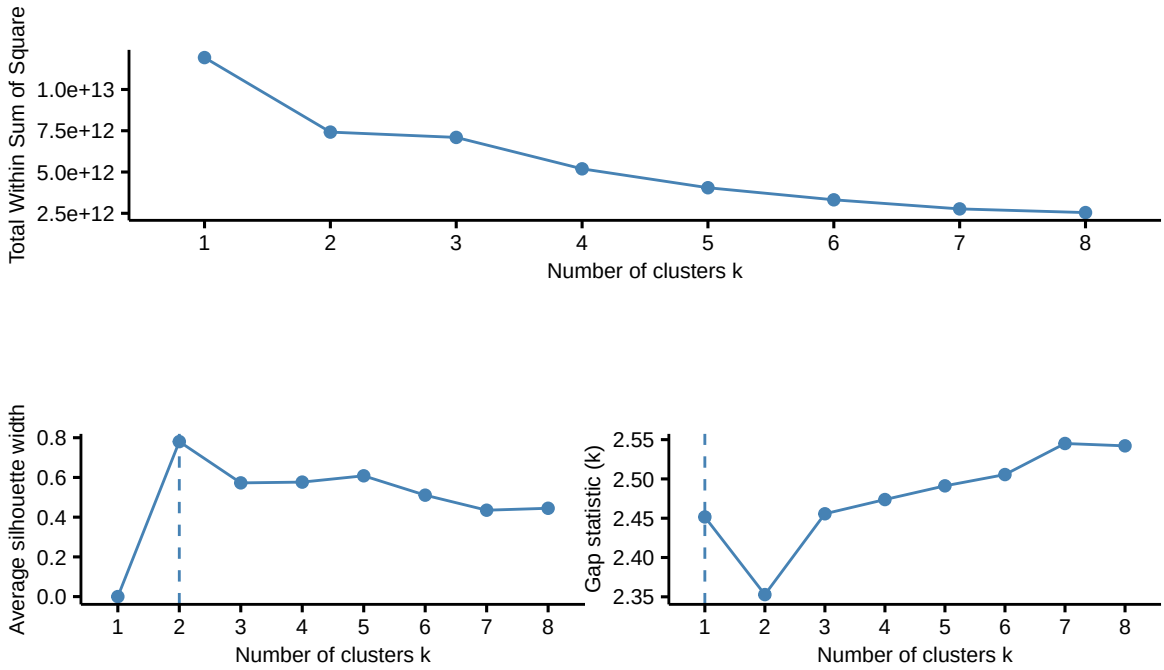

Figure 7: Utilisation of suitable K-means Clustering Metrics, including Elbow plot, Silhouette and Gap statistic at a range of cluster centers and total clusters for continuous variables in the dataset.

As can be seen, there is no distinguishable clear choice of k which can be clearly applied to the dataset. Thus, it was attempted to apply the algorithm for small choices of k, namely k = 2 and k = 3. This is due to the fact that increasing numbers for k will have an insignificant impact on the WGSS, as illustrated by the elbow plot.

The k-means was applied and for the ease of convenience, a sample of 10,000 borrowers were segmented from the results and illustrated by the figure below. It must also be noted that due to the large number of rows and columns represented in matrix form, it is difficult to visualise the space given the very large number of dimensions. Thus, in order to preserve the space and interpret the clusters, the dimensions are projected down to a more tractable two dimensions which can be easily plotted and visualised. This was achieved using principle component analysis which projected the data to the first two principle components, i.e. the dimensions which show the most variation in the data. The percentages observed at Dim1 and Dim2 account for the percentage of the variation of the princple components.

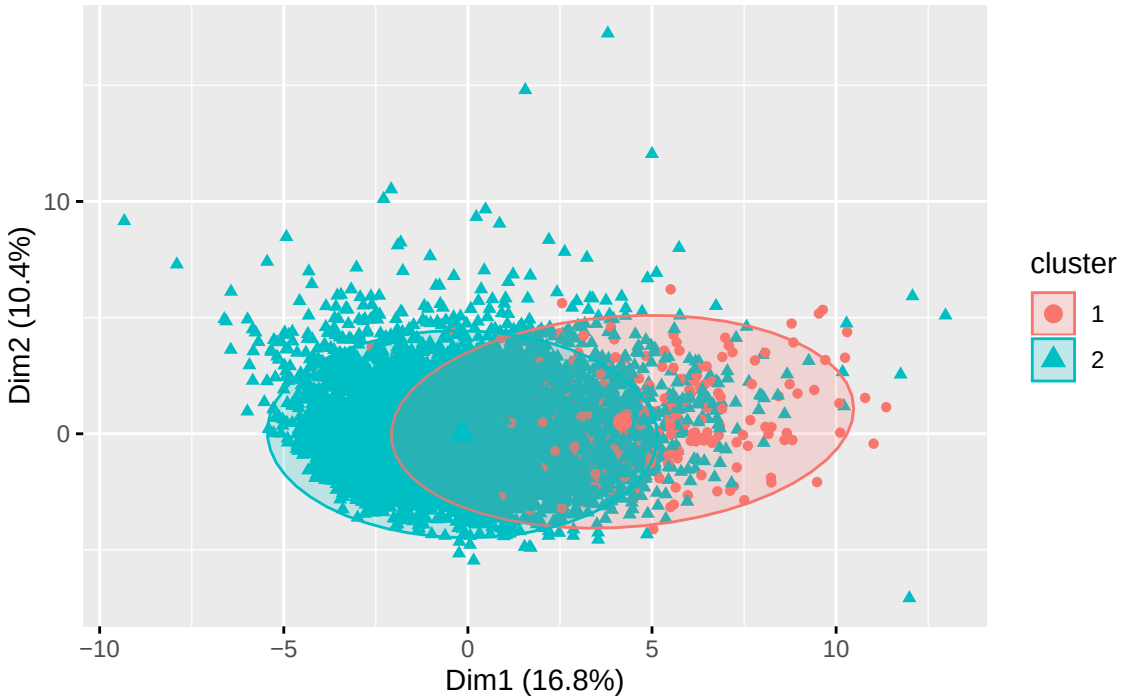

Figure 8: Illustration of the applied K-Means algorithm at K = 2 Cluster centres to identify different risk drivers between the remaining continuous variables in the Prosper dataset.

It can clearly be seen that the clusters distinctly overlap with one another. This is not ideal as it was hoped there would be a clear segmentation in potential risk drivers between defaulted and non-defaulted cases. In contrast, the risk drivers for this segmentation approach are indistinguishable.

Thus, it does not seem to be evident that there are distinguishable clusters within this dataset, especially when one considers a k = 2 setup. Therefore, a K-means algorithm is repeated utilising three clusters and the results are as follows:

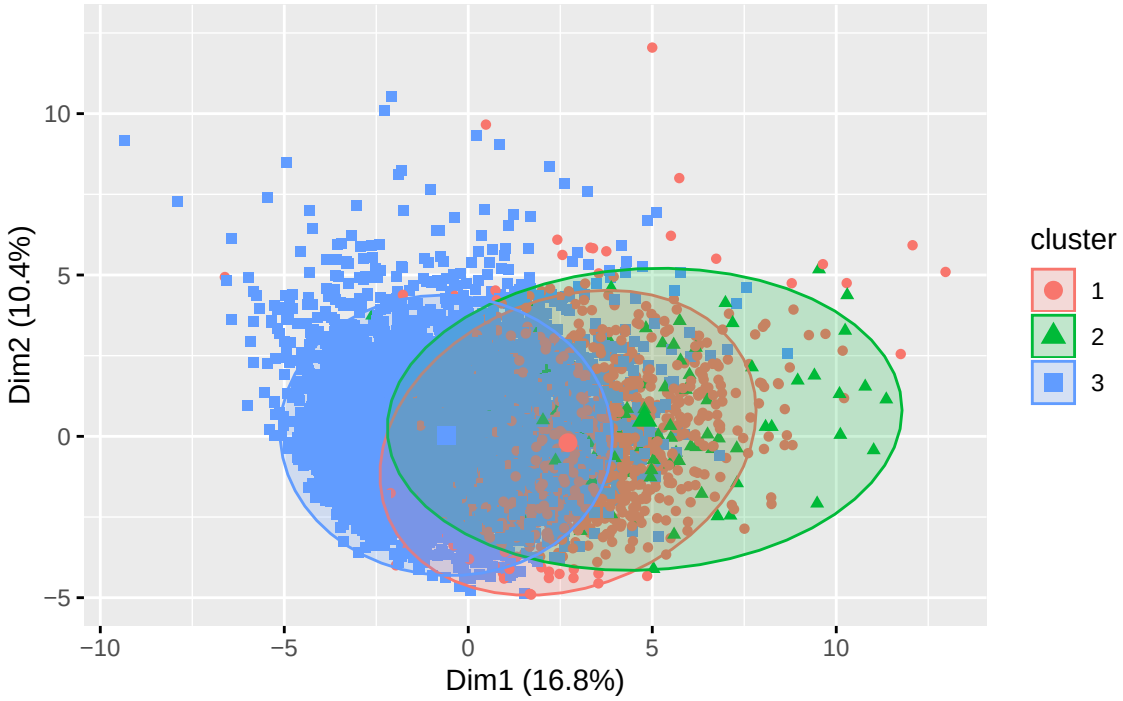

Figure 9: The K-means algorithm applied with K = 3 cluster centers to the continuous variables remaining in the Prosper dataset.

Similarly, for k = 3, there is a distinct overlap between the clusters and it is evident that the k-means approach does not indicate a clear and distinguishable segmentation opportunity. There remains a number of alternatives, one approach utilising k-medoids as well as hierarichal clustering using Decision Trees. It was decided to utilise the latter.

Decision Trees

Decision Trees take a response variable, which in the case of this report is default, and try to identify the dominant risk drivers within a hierarichal structering. The most dominant of these variables determines the root of the tree, where each variable following on from this follows one of the two branches which the root of the tree creates. The next leaves or nodes of the tree continue the pattern until there are either no more suitable segmentations or a pre-specified depth has been reached. Naturally, the criteria for possible splits can be altered and for the purposes of this report, each node would require a population of at least 500 people for a possible segmentation.

Formally, decision trees utilise recursive partitioning to the instance space. In contrast to the K-means counterpart, this method can be applied to both categorical and continuous variables. The tree consists of nodes, of which the first node is deemed the root of the tree. All other nodes stem from the root and are called leaves which can be split into two or more sub spaces according to certain specified conditions. Decision trees come in a number of forms, from the Classification and Regression Trees (CART) which use the Gini index as a metric to the Iterative Dichotimiser 3 (ID3) which use the entropy function and information gain as metrics. The CART algorithim favours larger partitions and is very simple to implement yet the ID3 approach favours partitions with very small counts yet many distrinct values [6].

The Gini impurity measures how often a randomily assigned element from the set would be incorrectly labeled according to the possible distributions of labels available in the subsets. This Gini impurity can be calculated

by letting p_i be the probability of an item with label i being chosen. Furthermore, let $\sum_{k\neq i} p_k = 1 - p_i$ be the probability of mistakingly categorizing that item. By multiplying these values together, the Gini impurity can thus be calculated. Thus, for a set of items with N classes, whereby $i \in \{x\}_{i=1}^{N}$, the Gini impurity is calculated as:

$$\text{Gini} = \sum_{i=1}^{N} \sum_{k\neq i} p_k = \sum_{i=1}^{N} p_i(1-p_i) = \sum_{i=1}^{N} (p_i - p_i^2) = \sum_{i=1}^{N} p_i - \sum_{i=1}^{N} p_i^2 = 1 - \sum_{i=1}^{N} p_i^2. \tag{15}$$

On the other hand, the ID3 alogrithims utilise the concept of entropy which is defined as follows:

$$H(T) = -\sum_{i=1}^{N} p_i \log_2 p_i, \tag{16}$$

where p_i are now fractions which sum to one and correspond to the percentage of each class which are present in the leaf nodes due to the splitting of the tree. In this scenario, the information gain is then defined as :

$$IG(T, x) = H(T) - H(T|x). \tag{17}$$

Thus, given a list of x leaf nodes or children, the information gain is the entropy of the parents minus the weighted sum of the entropy of the parent given the children. This can finally be expressed as:

$$IG(T, x) = -\sum_{i=1}^{N} p_i \log_2 p_i - \sum_{x} p(x) \sum_{i=1}^{N} -p(i|a) \log_2 p(i|a). \tag{18}$$

The resulting information gain can then be used to decide on which features it should split on at each step in the tree building process. It does so by attempting to determine the purest possible leaves using an information metric denoted as bits. These bits, or alternatively shannons, correspond to the information entropy of a binary random variable which can be either 0 or 1 with an equal probability. Finally, at each node of the tree, the information value then represents "the expected amount of information which is required to specify whether a new instance can be classified as a yes or no, given that the example reached that node".

Decision trees have a number of advantages, such as the resulting output can be easily understood and interpreted. Furthermore, as previously stated they can handle both numerical and categorical data and can perform well with large datasets. In contrast, they lack robustness and small changes in the data can consequently impact the final predictions. Finally, these trees can suffer from overfitting, though this can sometimes overcome by pruning.

Implementation of CART Decision Trees

The decision tree was applied to the dataset numerous times. The initial few iterations showed that in order to correctly distinguish a suitable tree, custom categorisation of dominant parameters would need to be applied or in an effort to determine if there were less dominant underlying categorizations, the variable was removed entirely. In order to be deemed a node, the node must contain at least 500 borrowers to ensure the significance of the node as well as to maintain the sensitivity of splits and the max depth of the leaf nodes was set to three, to prevent the tree from becoming too large and to obtain the key risk drivers. The entire decision tree iteration scheme can be seen in the Appendix. The first interesting tree, obtained at the third iteration is seen below:

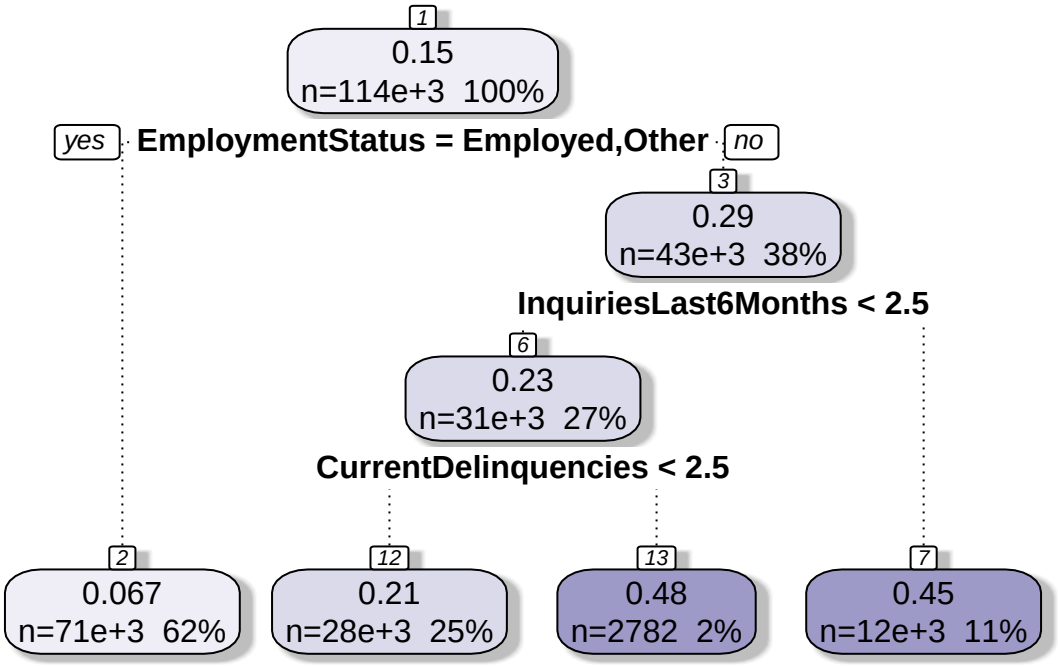

Figure 10: The first suitable decision tree determined during segmentation snalysis to determine risk drivers utilising all remaining continuous and categorical variables in the Prosper dataset.

The tree outputs three explanatory variables at each node, which are the default or bad rate, the number of borrowers which fulfill this condition and the percentage of the total population which are represented there. In addition, at each node a criteria is stated, such as $CurrentDeliquences < 2.5$. If these criteria are satisfied, they travel to a new node to the left and otherwise travel to the right node. This process repeats until the stopping criteria is met.

At this iteration, it can be seen that *EmploymentStatus* is a heavily dominant variable and it would be interesting to see, with the current specificiations, if this variable is hindering any further splits. Therefore, it was removed and or discretized into custom groups and the process was re-initiated. This mantra was repeated several times, in an attempt to find the most suitable tree. It must be noted that these trees are extremely sensitive to the grouping choice and to specific dominating variables. A further iteration involves the customisation of *IncomeRange*, to yield an intermediate decision tree for the chosen model:

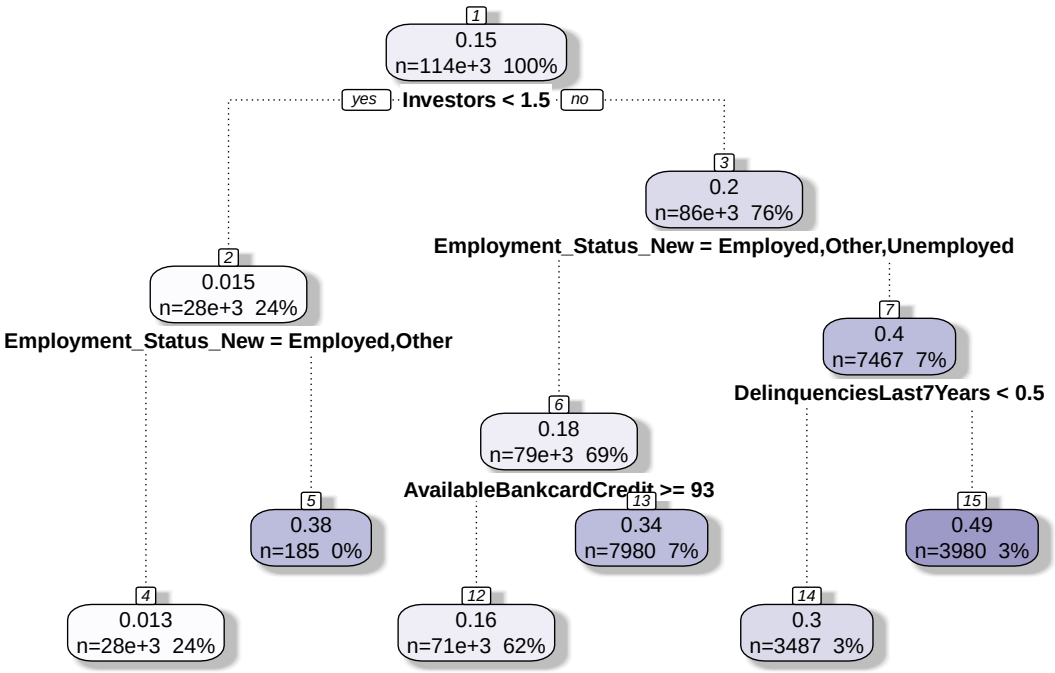

Figure 11: A further iteration of segmentation analysis utilising decision trees with several dominant variables such as EmploymentStatus and InquriesLast6Months removed and/or re-categorized to establish borrowers with varying risk drivers throughout the dataset.

As it can be seen, the current decision tree starts at a root node of the variable $Investors$ which is the dominating variable. It is open to interpretation which decisions can be made at this point of the modeling process, whether it be the removal of $Investors$ from the modeling process or the removal of $EmploymentStatusNew$ due to the similar characteristics of the leaf nodes. It was decided to do the latter to yield the final model of this stage of the segmentation analysis.

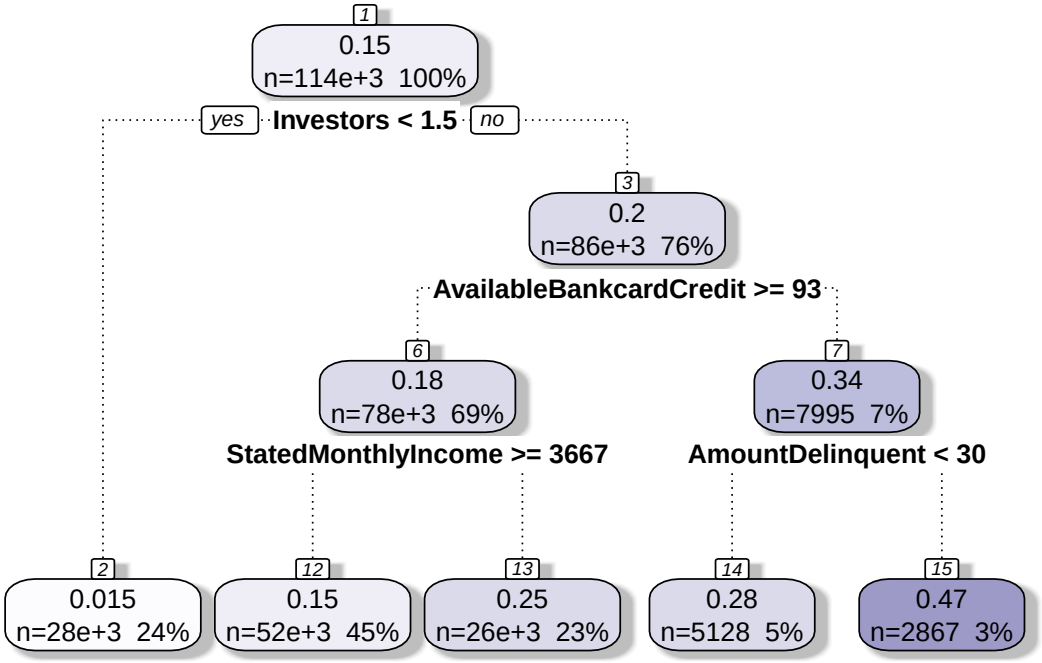

Figure 12: The Final iteration and chosen decision tree model after having categorized or removed the most dominant variables within the dataset.

As it can be seen, the removal of *EmploymentStatusNew* did not yield any more substantial splits, which could be pertitent to the determination of varying risk drivers between borrowers. Thus, It was concluded that there are no distinct eligible splits in the above plots to indicate a clear candidate for segmentation and the production of multiple scorecards. Therefore, it was decided to continue to modeling process with a single scorecard. It must be noted the variables removed or re-categorized during this stage of the modeling process were to indicate varying risk drivers between borrowers and as such were re-introduced to the dataset at this point. That being said, binary variables such as *CurrentlyInGroup* were removed permanently due to their binary nature, which from a business standpoint was deemed to not contain valuable enough information in predicting default. In addition, it was desired to generate a model which contained as few variables as possible and as such, these variables were considered surplus to requirements.

Binning and Discretisation

Following on from the conclusion that only one scorecard will be built, the variables determined during variable selection in Single Factor Analysis were brought forward to the binning or discretization stage of the model. This stage prepares the variables for logistic regression.

Binning was applied in order to determine the key groups or intervals which could be deemed strong, neutral or negative indicators for a customers likelihood to default. The process utilises recursive partitioning to categorize the numeric characteristics and works under a set of defineable conditions. This process of discretization or binning can be divided into two categories: supervised and non-supervised. Supervised binning takes into consideration the class values, i.e. dividing the number of classes from the discretization parameter. Unsupervised on the other hand requires a default number of bins or classes. The library utilised, *Smbinning* uses the former, with an entropy based approach. The entropy or information content is

determined through a number of steps [4]. To start it calculates the entropy of the target which is defined as:

$$\mathbb{E}(S) = \sum_{i=1}^{n} -p_i \log_2 p_i, \tag{19}$$

where p_i is the probability of default in class i. This is achieved by building frequency tables of all the class values for default and non-default. The next step is to calculate the entropy for the target given a bin. This is achieved by the following:

$$\mathbb{E}(S, A) = \sum_{v \in A} \frac{|S_v|}{|S|} \mathbb{E}(S_v), \tag{20}$$

where S_v is the sum of the row of values, default and non-defaulted cases, for a given bin selection and S is the total number of observations across all the bins.

Finally, the information gain is calculated using:

$$\text{Information Gain} = \mathbb{E}(S) - \mathbb{E}(S, A). \tag{21}$$

This process is utilised to determine the best interval or cutpoints for any continuous variable to return the highest gain, which in this case is the weight of evidence.

The binning was performed for both the continuous and categorical variables where there were in the case of the categorical variables, a maximum number of categories were set to 20 and in the case of the continuous variables, the minimum population percentage present in each bin was set to 5%. Unforunately, for variables such as *ListingCategory*, custom categorization was required due to the large number of categories and insufficient representation of the population within each category. This population criteria of 5% allowed for a good balance between number of partitions and information as detaled by the weight of evidence.

A number of variables were removed during this process due to there being an insignificant number of splits or too many categories. In addition, certain variables were binary and were similary removed as it was believed that splitting in this manner for an excessive number of variables would inaccurately describe the probability of default. An example of the output is seen below:

Table 3: An example of the yielded output from performing continuous binning on the variable Investors using recursive partitioning on the remaining variables in the Prosper dataset.

	Cutpoint	CntRec	BadRate	LnOdds	WoE	IV
1	<= 1	22217	0.0153	4.1672	2.4501	0.6338
2	<= 8	4921	0.1536	1.7064	-0.0108	0.0000
3	<= 35	14220	0.2404	1.1507	-0.5665	0.0604
5	> 73	33621	0.1758	1.5448	-0.1724	0.0116
6	Missing	0	NaN	NaN	NaN	NaN
7	Total	91149	0.1522	1.7172	0.0000	0.7405

where *CntRec* stands for the number of borrowers present within the defined ranges. A sample output for categorical variables is similarly displayed as follows:

Table 4: An example of the yielded output from performing categorical binning on the variable IncomeRange using recursive partitioning on the remaining variables in the Prosper dataset.

Cutpoint	CntRec	BadRate	LnOdds	WoE	IV
= '0'	503	0.3976	0.4154	-1.3018	0.0137
= '1-24999'	5859	0.2316	1.1992	-0.5179	0.0205
= '100000+'	13877	0.0764	2.4925	0.7753	0.0694
= '25000-49999'	25765	0.1735	1.5611	-0.1561	0.0073
= '50000-74999'	24768	0.1167	2.0242	0.3070	0.0230
= '75000-99999'	13529	0.0922	2.2865	0.5693	0.0393
= 'Not displayed'	6212	0.3992	0.4087	-1.3085	0.1707
= 'Not employed'	636	0.2689	1.0004	-0.7168	0.0045
Missing	0	NaN	NaN	NaN	NaN
Total	91149	0.1522	1.7172	0.0000	0.3484

It must be noted that recursive partitioning performs particularly well with continuous variables whilst categorical variables would require a combination of intuition and customized grouping for optimal re-discretization. As can be seen from the variables *Investors* and *IncomeRange* above, there is an arguement for customized discretization based on the observed ouput. On the whole, the population representation is representative within each group in *Investors* yet *IncomeRange* could be re-discretized by merging some of the groups together. This must be done with caution to not risk losing too much information whilst also attempting to maximize the weight of evidence within each group. For more accurate interpretation of the results, the categorical illustrated as follows:

Figure 13: Illustration of the Categorical WoE for a select chosen variables after having performed recursive partitioning.

Furthermore, the continuous variables were plotted in a similar fashion as follows:

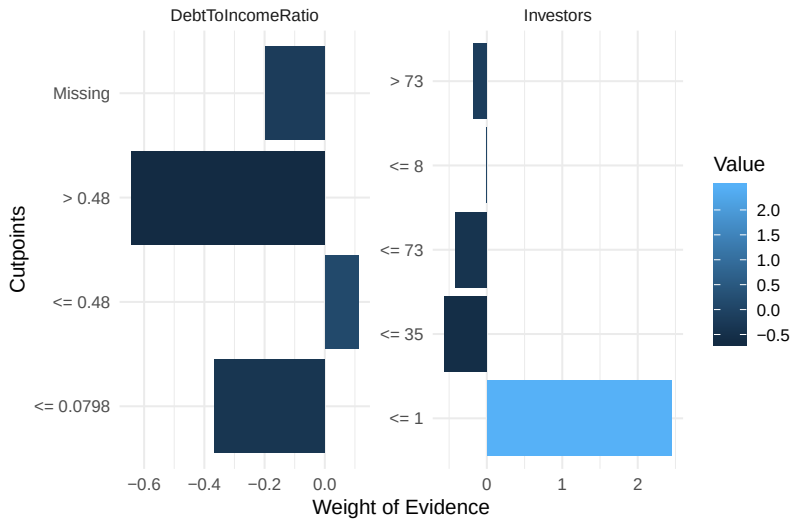

Figure 14: Illustration of the Continuous WoE for a select chosen variables after having performed recursive partitioning.

As can be seen in the above graphs of *EmploymentStatus*, *IncomeRange*, *DebtToIncomeRatio* and *Investors*, both the continuous and categorical variables are discretized into suitable groups, determined by the recursive partitioning algorithm. The graphs containing ranges, as in for example $ <= 35$ in *Investors* indicate that these borrowers have less than or equal to 35 investors yet have at least eight. It is of particular interest to note that borrowers with only a single investor yield a large WoE, indicating a strong indicator for non-default wherease higher number of investors were deemed negative indicators and a likely predictor of default. Similarly, both *EmploymentStatus* and *IncomeRange* show expected intuitive characteristics for borrowers respective incomes and employment type respectively. Finally, *DebtToIncomeRatio* indicates that a ratio between 0.0798 and 0.48 are positive characteristics whilst ratios above and below these values are negative characteristics in predicting default. The plots clearly illustrate that different grouping for pertinent variables have varying WoE characteristics. It must be noted that the above plots on the other hand omit the percentage of the total population which are present in each bin, though they do contain at least 5 percent due to the splitting critera. That being said, certain criteria may thus falsely indicate default characteristics. Due to this, certain variables were manually re-categorized using business intuition. This decision comes with its own issues, as it affects the optimum discretization which were obtained from recrusive partitioning.

Logistic Regression

Once suitable discretizational groups were chosen for the relevent variables, the corresponding WoE values for each discretized group were processed to replace their raw values. It was decided to utilise logistic regression, which is one suitable method in classifying the binary data and attempting to predict default. Within Logistic Regression, suitable options included forward, backward or stepwise logistic regression.

These methods have a number of advantages, being easy to implement, train and interpret but also giving a measure of how relevant each predictor is with a direction of association. In contrast, Logistic Regression assumes linearity between the dependent and the response variables. Furthermore, it can only be utilised to predict discrete functions, outlining the dependence of discretizing continuous variables within the dataset to categorical ones.

Apart from this, an interesting alternative is that utilising generalised extreme value theory, which is particularly effective in managing datasets with rare events, as is often the case in probability of default models.

However, for the basis of this report, it was decided to use the former.

Formal Definition

More formally, logistic regression is a classification algorithm which is used to allocate a set of observations to a discrete set of classes. There are several types such as binary (default, non-default), multi (apples, oranges, pears) or ordinal (low, medium, high) . For the basis of this report, the binary variable default will be used as the response variable. In contrast to its linear regression counterpart, which could be used to predict the loan status of each borrower, logistic regression can be used to predict whether the borrower has defaulted or not. The model furthermore offers the ability to analyze the probability scores which underly the model classifications. This is achieved by utilising the sigmoid function which is obtained from the hypothesis function to map any observation to another value between 0 and 1 whereby:

$$h_\theta(x) = g(\theta^T x), \quad g(z) = \frac{1}{1+e^z} \quad 0 \leq h_\theta(x) \leq 1. \tag{22}$$

This hypothesis function $h_\theta(x)$ is mapped by the function g to form the sigmoid or logistic function as seen in the following:

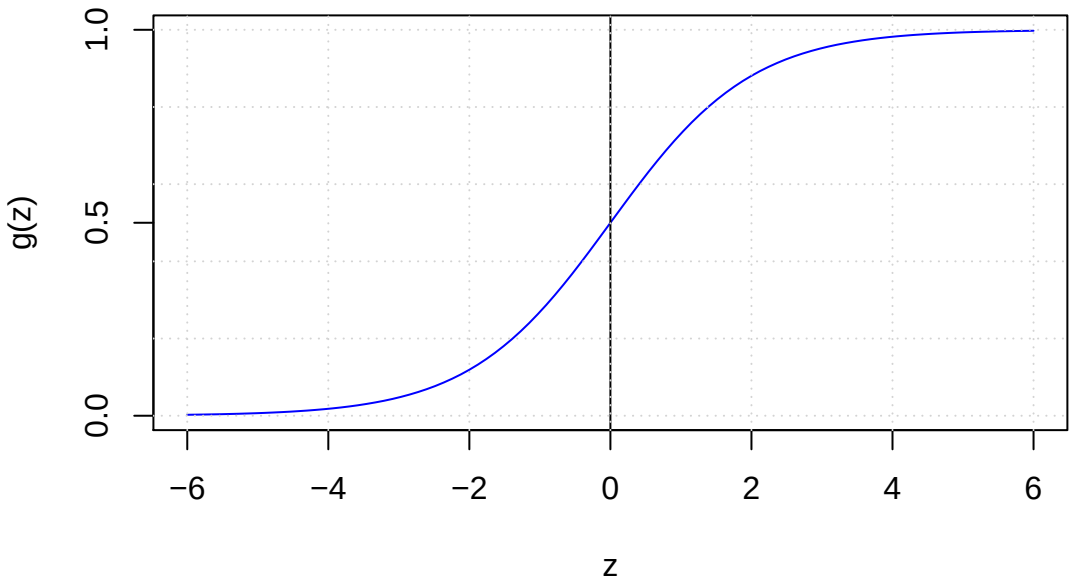

Figure 15: Illustration of a simulated Logistic Regression Sigmoid function across a suitable range of values.

The resulting regression analysis can then be used to describe the relationship between one dependant binary variable and the remaining one or more independent variables. The resulting output of the analysis can be expressed mathematically as follows:

$$\mathbb{E}(Y) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n, \tag{23}$$

where Y is the dependent response variable, the variable one wishes to investigate. Furthermore, p is the probability of that event occurring. When performing the regression, the output will yield the β coefficients, where β_0 will be the intercept and $\beta_i, i = 1, .., n$ will be the coefficients for the n independent explanatory variables [5].

However, similar to the issue encountered with decision trees, the algorithm can suffer from overfitting which can be reduced by reducing the number of variables in the model. This allows for an increase in the generalizability of the model beyond the data and the model fit and can be suitable from a business sense due to the fact that one needs to track less variables.

The degree of accuracy of the models predictability is then determined using the cost function, which can be defined as:

$$C(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)), & y = 1 \\ -\log(1 - h_\theta(x)), & y = 0. \end{cases} \tag{24}$$

This cost function thus describes the cost the algorithm pays if it were to predict a value $h_\theta(x)$ when the label turns out to be y. This grants the function convexity which is required for the gradient descent algorithm to process. This is further illustrated by the following diagram:

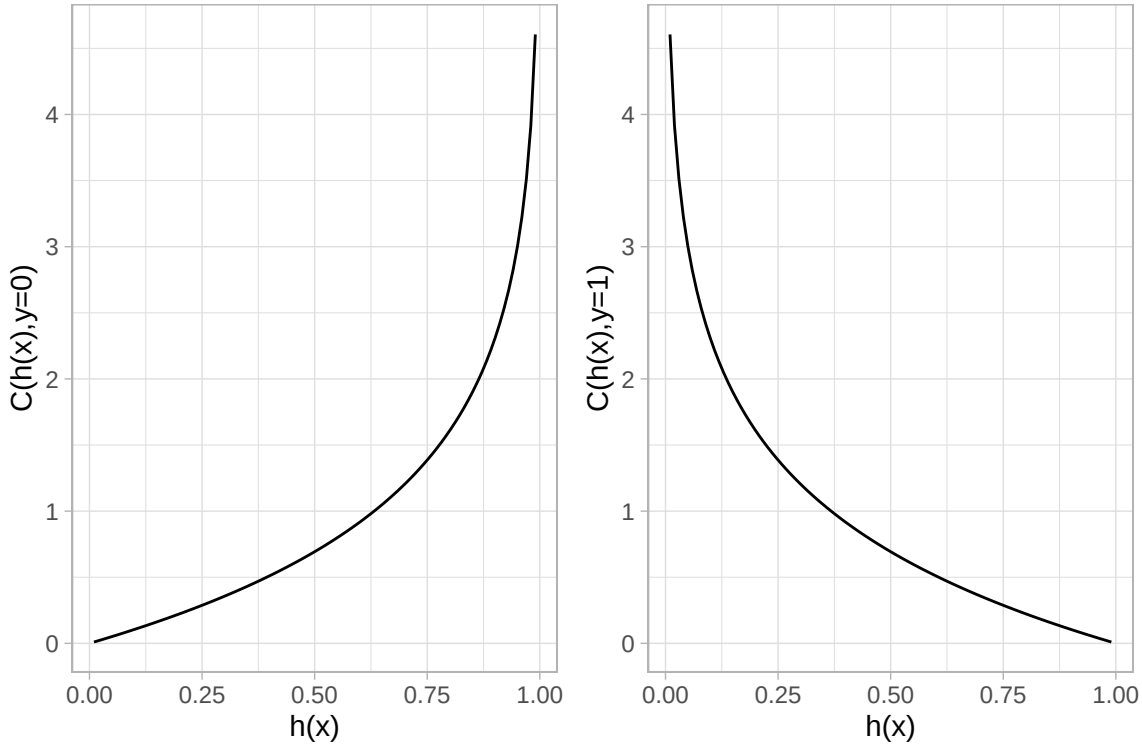

Figure 16: An illustration of the simulated cost function at both values of the binary response variable, required for the gradient descent algorithm.

Following on from this, this expression can be further simplied by merging the two cases to form a single expression:

$$C(h_\theta(x), y) = -y \log(h_\theta(x)) - (1 - y) \log(1 - h_\theta(x)). \tag{25}$$

This process is repeated for each prediction, and the optimised cost function is evaluated as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{C}(h_\theta(x^{(i)}), y^{(i)}) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\log\left(h_\theta(x^{(i)})\right) + (1-y^{(i)})\log\left(1-h_\theta(x^{(i)})\right)\right], \tag{26}$$

where m is the number of examples. This function is then minimised which will output the $\theta's$ with the least error. It does this by making use of gradient descent. Gradient descent attempts to find the local minimum by taking steps proportional to the negative of the gradient of the function at the current point. The full derivation can be found in the appendix but the resulting gradient at any point j can then be derived as :

$$\frac{\partial}{\partial\theta_j}J(\theta) = \sum_{i=1}^{m}\left(h_\theta(x^i) - y^i\right)x_j^i. \tag{27}$$

The function will now be ready to make predictions by utilising the hypothesis function:

$$h_\theta(x) = \frac{1}{1+\exp(\theta^T x)} \tag{28}$$

which is in fact the estimation of the probabilities of default (PD) obtained from the logit scores from the Logistic Regression. Therefore:

$$\text{PD} = \frac{1}{1+\exp(\alpha+\beta x)} \tag{29}$$

where α is the intercept, β are the beta coefficients from the logistic regression model and x are the input values. The logit score is the value obtained within the exponential for each customer respectively.

Implementation of Forward Logistic Regression

Once the relevant variables for regression were selected from the binning process, it was decided to use forward selection. Forward selection is a form of stepwise regression wherby one starts with an empty model and adds variables one by one. During each step of the process, based on a particular criterion such as BIC or AIC, the variable is added or omitted in the model. This process continues until the algorithm deems that the addition or removal of any variables no longer improves the model.

The forward regression was run by inputting the binned or discretized weights of evidence which yielded the following variables used for the final model and their respective β coefficients:

Table 5: Display of the remaining variables for the chosen model, obtained after performing Forward Logistic Regression.

Variables	Coefficients
(Intercept)	-1.6933647
InvestorsWoE	-0.7663794
InquiriesLast6MonthsWoE	-0.7613133
AvailableBankcardCreditWoE	-0.5230256
StatedMonthlyIncomeWoE	-0.7856571
CurrentDelinquenciesWoE	-0.6790574
DelinquenciesLast7YearsWoE	0.3172702
AmountDelinquentWoE	0.2415181
TradesOpenedLast6MonthsWoE	-0.2600109
EmploymentStatusDurationWoE	0.2010353
LoanOriginalAmountWoE	0.0995004
RevolvingCreditBalanceWoE	-0.0798988

As illustrated from the table above and the probability of default in Eq: (29), a negative β coefficient will indicate a smaller exponent which in turn will be associated with a higher probability of default. Similarly, a positive β coefficient will constitute a lower probability of default. From an intuitive point of view, the sign of the coefficients make sense, where it is postulated that the higher the number of respective investors, the lower the faith in that particular borrower and as such, the higher the risk and the probability of default. A peculiar variable is that of $StatedMonthlyIncome$ which indicates that a large income would indicate a higher probability of default. Interestingly, when looking at the weight of evidence for this particular variable, there is in fact higher likelihood for default at both the lower and higher ends of the spectrum. In contrast, borrowers in the "medium" ranges tend to show a lower probability of default for an increased income, as would be expected.

Performance Measures

Once a suitable model was chosen from forward regression, further analysis of each of the chosen variables and model as a whole were required. In order to achieve this, the logit scores were obtained and further mapped to obtain the resulting probabilities of default. It was now possible to assess this model under a number of suitable metrics to check its validity.

ANOVA Analysis

The analysis of variance (ANOVA) is used to analyse the differences amongst group means in a sample. It is a similar method of hypothesis testing which measures the significance of each variable within the model. The results of which can be seen below:

Table 6: ANOVA analysis table displaying variable significance for each variable obtained in the chosen model after applying Forward Logistic Regression.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	91148	78046.44	NA
InvestorsWoE	1	6391.863681	91147	71654.57	0.0000000
InquiriesLast6MonthsWoE	1	3689.366957	91146	67965.21	0.0000000
AvailableBankcardCreditWoE	1	2079.742585	91145	65885.46	0.0000000
StatedMonthlyIncomeWoE	1	1164.993719	91144	64720.47	0.0000000
CurrentDelinquenciesWoE	1	614.682144	91143	64105.79	0.0000000
DelinquenciesLast7YearsWoE	1	64.524603	91142	64041.26	0.0000000
AmountDelinquentWoE	1	31.096464	91141	64010.17	0.0000000
TradesOpenedLast6MonthsWoE	1	44.155344	91140	63966.01	0.0000000
EmploymentStatusDurationWoE	1	19.122059	91139	63946.89	0.0000123
LoanOriginalAmountWoE	1	17.263931	91138	63929.63	0.0000325
RevolvingCreditBalanceWoE	1	8.422819	91137	63921.20	0.0037054

It can be seen that at a 95% confidence level, each of the paramaters chosen from the forward regression process are deemed significant and will remain in the chosen model.

Correlation Matrix

In order to ensure that the final chosen variables were as independent as possible, a correlation matrix was evaluated as illustrated in the plot below:

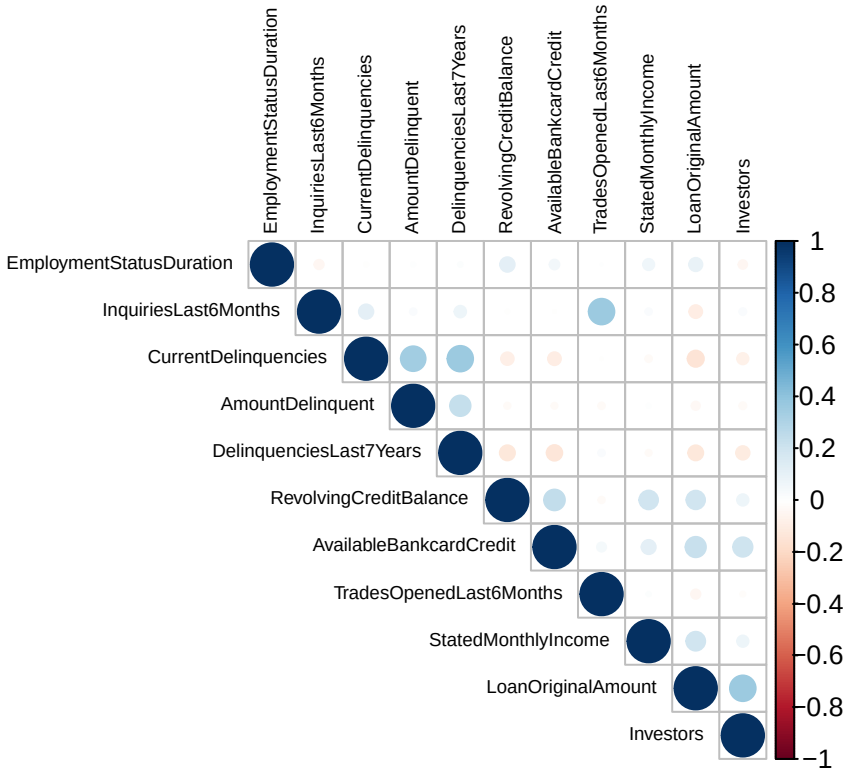

Figure 17: Correlation Matrix of the chosen model variables obtained from Logistic Regression.

As can be seen, the degree of correlation is once again visualised from positive correlations in blue to negative ones in red. The degree of correlation is based on the size of the circle within the grid. It can be seen that the variables are majorly uncorrelated, with the absolute correlations observable at ≈ 0.35, which is that of $InquiriesLast6Months$ and $TradesOpenedLast6Months$. This correlation is quite intuitive and is deemed sufficiently low to proceed without removing any further variables.

Receiver Operating Characteristic (ROC) Curve and Gini

ROC curves portray how any predictive model can distinguish between true positive and negatives. In order to do this, a model needs to not only correctly predict a positive as a positive but also a negative as a negative. The ROC curve plots the True Positive Rate against the False Positive Rate.

A model that is no better than a random guess will produce a ROC curve of a straight line running diagonally through the origin, as printed in the graph below. This is used as a base line for ROC curves and produces an Area Under the Curve (AUC) Value of 0.5.

The optimal ROC curve runs at a right angle to the y-axis. The closer the ROC curve is to the top left corner of the graph, the better the model is at predicting, i.e. the larger the AUC value [12].

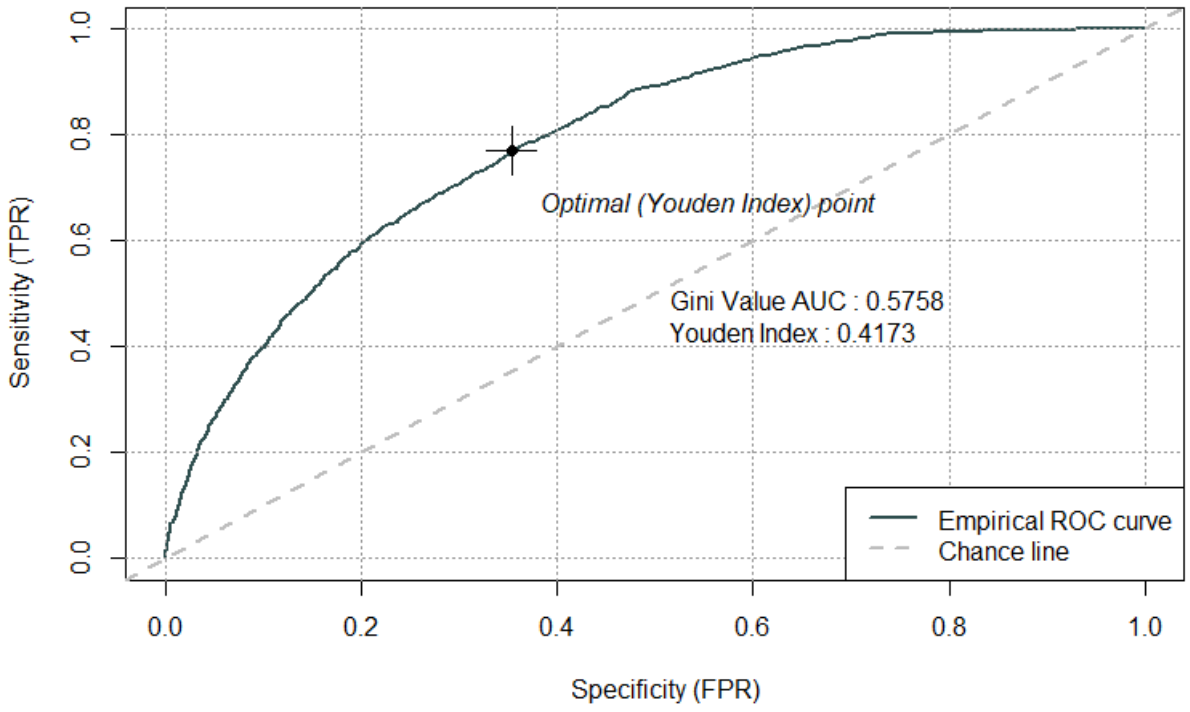

Figure 18: ROC Curve displaying the Gini Value and Youden Index for the chosen model relative to the chance line of arbitrary estimates.

The ROC curve was applied and the Gini coefficient as well as the youden index were calculated. The Youden index J is a measure of determining the effectiveness of a diagnostic marker and selects the optimal cutoff point for that marker. It summarizes the performance of a diagnostic test with ranging values between 0 and 1, where a 1 signifies that there are no false positives or false negatives and the test is perfect. The Youden Index is defined as follows:

$$J_c = \max_c(\text{Sensitivity}_c + \text{Specificity}_c - 1), \tag{30}$$

where the optimal cut-point corresponds to the point closest to (0,1) on the ROC Curve.

As can be seen from the ROC curve above, the Gini coefficient obtained is ~0.6 indicating strong predictive power between both the training and test sets. The Gini is obtained via the following formula:

$$\text{Gini} = (\text{AUC} * 2) - 1. \tag{31}$$

The Gini Index describes the global quality of the predictive model. It is widely used to describe the quality of a scoring function. It takes values between -1 and 1. The ideal model, i.e., a scoring function that perfectly separates good and bad clients, has a Gini index equal to 1. On the other hand a model that assigns a random score to the client has a Gini index equal to 0. Negative values correspond to a model with reversed meanings of scores. The Gini coefficient can similarly be calculated as:

$$\text{Gini} = \frac{N_c - N_d}{N_c + N_d + T_p}, \tag{32}$$

where N_c, N_d and T_p are the number of concordant and discordant pairs and T_p are the total number of pairs with tied ranks on the dependent variable. In addition, the Gini for each individual variable in the model is calculated and displayed in the following table :

Table 8: The Gini Coefficients obtained from Logistic Regression by independent analysis in predicting default of each variable.

Variable	Gini
Investors	0.325
InquiriesLast6Months	0.332
AvailableBankcardCredit	0.324
StatedMonthlyIncome	0.249
CurrentDelinquencies	0.220
DelinquenciesLast7Years	0.148
AmountDelinquent	0.241
TradesOpenedLast6Months	0.232
EmploymentStatusDuration	0.224
LoanOriginalAmount	0.264
RevolvingCreditBalance	0.298

From the table above it is clear that there are some key driving factors which provide a strong indication of default. It is interestnig to note that all the variables show some predicting power with the majority between 0.2 and 0.3. This manages to shed some light and provide a numeric approximation to the previously evaluated variable importance factors evaluation by outlining their predicting power in this metric.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov or KS test is a non-parametric test used to determine whether two datasets differ significantly. It is quite convenient as it makes no assumption about the distributions. The KS test is defined as :

$$D = \max_s |F_n(s) - F(s)|, \tag{33}$$

where $F_n(s)$ is the empirical distribution function for n i.i.d ordered observations defined as :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]} X_{(i)}, \tag{34}$$

where I represents the indicator function, being equal to one within the interval and zero otherwise[5].

The KS test was applied to the model and the results are as follows:

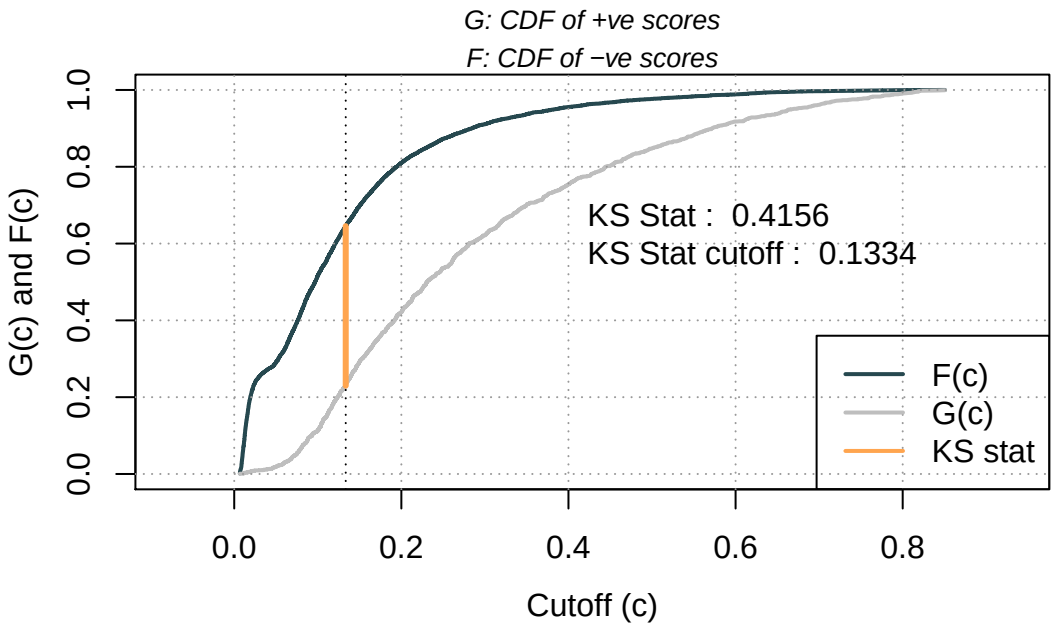

Figure 19: A two-sided Kolmogorov-Smirnov Analysis for the comparison of the distribution of defaulted and non-defaulted cases for the chosen model.

The maximum vertical distance between both distributions F(c) and G(c) corresponds to the D-statistic. This value is compared to a K-S Test P-Value table to determine the hypothesis of whether both distributions stem from the same distribution, to which it is determined that the distributions are different and the null hypothesis is rejected. In this particular case, the D-statistic is ≈ 0.41 (in a range between 0 and 1) displaying that in fact the defaulted and non-defaulted cases are distributed differently. As can be seen in the kolmorogov-smirnov figure above, the defaulted G(c) distribution has a fatter lower tail in comparison to its non-defaulted counterpart.

Benchmark

By performing an independent analysis on the dataset, it was thus possible to compare this obtained model to Prospers own internal rating system and a third party institution. The resulting ROC curve for each of the other metrics was plotted and the results are seen below:

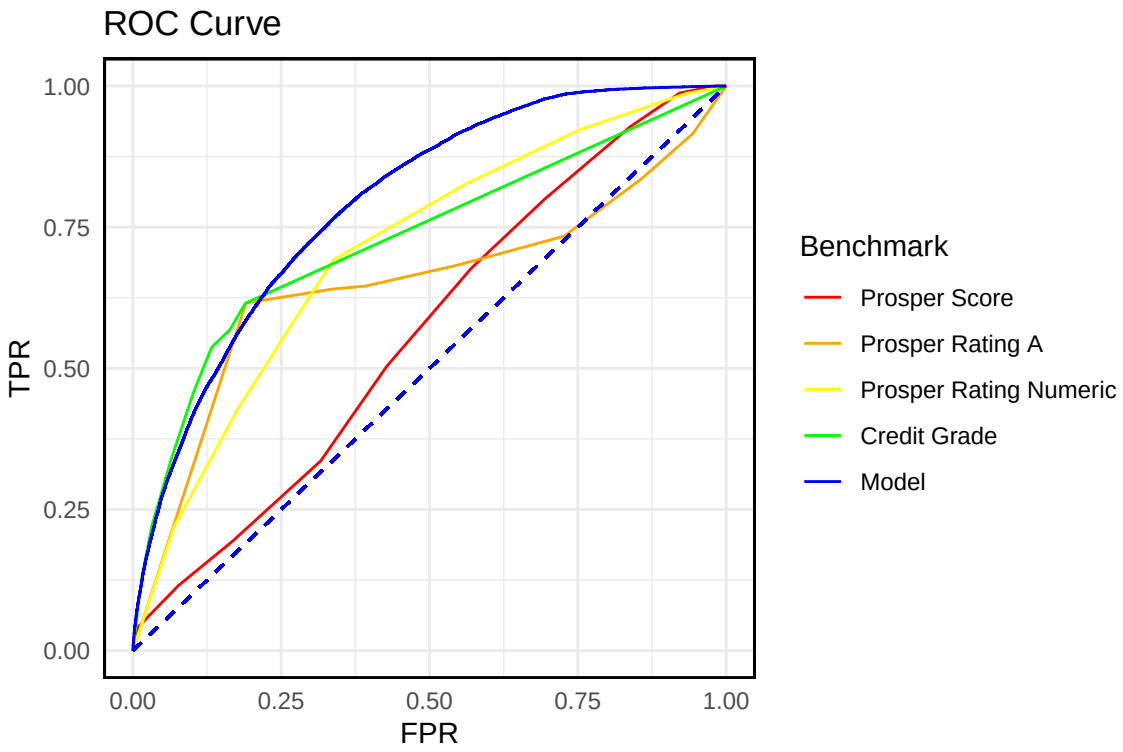

Figure 20: A ROC Curve of the False Positive Rate against True Positive Rate, used to benchmark the chosen model against metrics performed by Prosper and third parties.

As with the chosen model, these AUC values were mapped to a corresponding Gini and the following results were obtained:

Table 9: Comparison between the predictive power of the Chosen Model against that of Prosper Grading metrics and other Third Party Analysis, where the dotted line represents a random model with a Gini of zero.

	Prosper Score	Prosper Rating Alpha	Prosper Rating Numeric	Credit Grade	Model
Gini	0.127	0.297	0.417	0.461	0.576

It is noted that the chosen Model outperforms all of the scoring metrics by both Prosper and the third party instutition, indicating that the model does show relative strength in predictability in relation to its peers and can be deemed a suitable choice at this stage of the modelling process. It must be noted however that a number of the metrics by Prosper contained a large number of missing values which are likely causes of the poor Gini coefficients.

From the results of the performance measures, it can be indicated that the chosen model has a reasonable predicting power as indicated from the Gini for the entire model and the weighted average of their individual components.

Scorecard

Once the model was selected the probabilities of default were mapped to scores, with low scores indicating a high credit risk and high scores a lower credit risk and thus a lower probability of default. This was achieved in the following way:

$$\text{Score} = \text{Offset} - (\text{Factor} \times \text{Logit Scores}) \tag{35}$$

$$\text{Offset} = \text{Target Score} - \text{Factor} \times \log(\text{Target Odds}) \tag{36}$$

$$\text{Factor} = \frac{\text{pdo}}{\log(2)}. \tag{37}$$

The pdo or points to double offset, which has been set to 20, indicates the amount of points required to a doubling of the good/bad odds. The target odds are set to 50 and the target score is set to 600. The scaling does not affect the predictive capabilities of the scorecard and are just the recommended and chosen parameters for the company in question. Prior to performing this mapping, it was pertinent to provide some summarizing information pertaining to the performance measures. The results are as follows:

Table 10: Summary of the selected model variables with suitable descriptory metrics.

Variable	Beta Coefficients	Individual Gini	VIF	Modal WoE	Modal Population	Percentage of Total
AmountDelinquentWoE	0.2332521	0.325	6.317	0.3243	90219	0.792
AvailableBankcardCreditWoE	-0.5313672	0.332	21.474	0.3429	33414	0.293
CurrentDelinquenciesWoE	-0.6761031	0.324	24.417	0.2858	89742	0.788
DelinquenciesLast7YearsWoE	0.3194860	0.249	8.266	0.2286	76439	0.671
EmploymentStatusDurationWoE	0.1819834	0.220	5.015	0.2539	39726	0.349
InquiriesLast6MonthsWoE	-0.7592853	0.148	46.925	0.6237	50005	0.439
InvestorsWoE	-0.7667456	0.241	43.496	-0.1734	41922	0.368
LoanOriginalAmountWoE	0.0995816	0.232	4.204	0.5653	48933	0.429
RevolvingCreditBalanceWoE	-0.1019192	0.224	2.908	0.4858	59543	0.523
StatedMonthlyIncomeWoE	-0.7887664	0.264	32.839	0.5266	41581	0.365
TradesOpenedLast6MonthsWoE	-0.2606844	0.298	7.908	0.3364	54249	0.476

With this information in mind, the resulting logit scores were mapped to scores and the results are as follows:

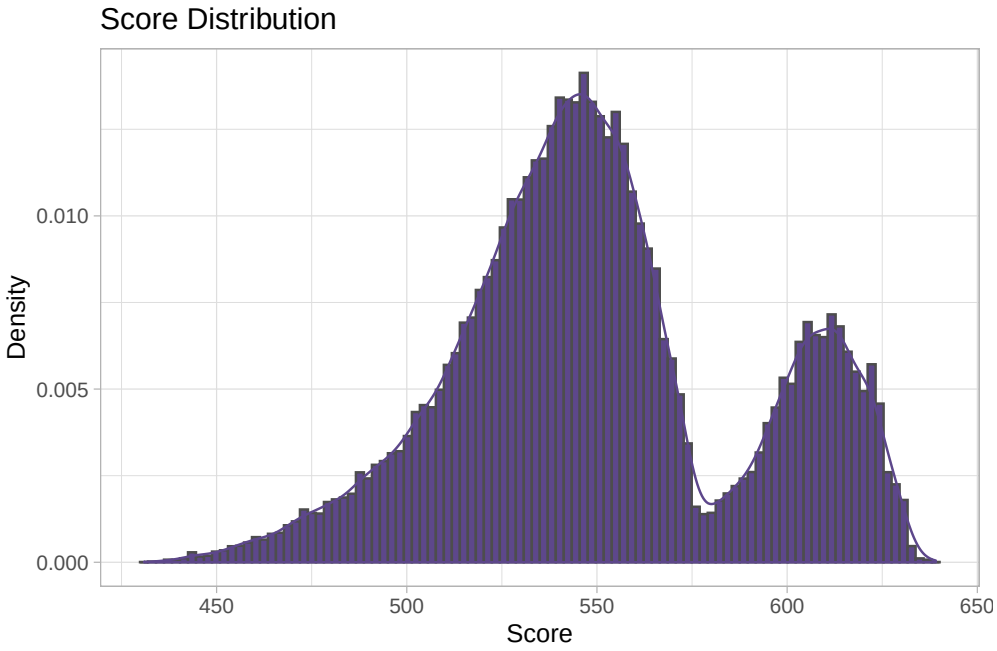

Figure 21: The Distribution of Scores obtained for each respective borrower, defaulted or non-defaulted, after mapping their respective logit scores from Forward Logistic Regression to a score.

As can be seen, the distribution of the scores has a skew whereby most individuals harbour to the left of the target score of 600. This is not worrying as the "punishment" due to the model selection to consumers will

be more of a benefit to the bank. The conservative approach will mean that any further loan agreeements will be stricter and accepted loans will be of lower risk. Thus, this decrease in lending would be of benefit to the banking sector and could offer more financial stability.

The higher concern is the bi-modal distribution. This is certainly not ideal given a symmetric distribution about the target odds would be desired. Furthermore, the valley between the score range of 550 - 600 shows that it is difficult for borrowers to obtain scores in this range. This indicates that certain variables dominate and are overly or underly representative in these sections. It was established that the variable *Investors* is the culprit for this bi-modal nature. The variable was removed and the the process was re-iterated which yielded a Gini of ≈ 0.52. This drop in predictability was deemed too great and it was thus decided to leave this variable within the model. Furthermore, the variable *Investors* performed strongly in predictability throughout a number of other metrics. It does open up the possibility to re-visit segmentation and to develop two scorecards with those borrowers having only one single investor and those with multiple. To further visualise the score distribution, a violin plot was constructed as follows:

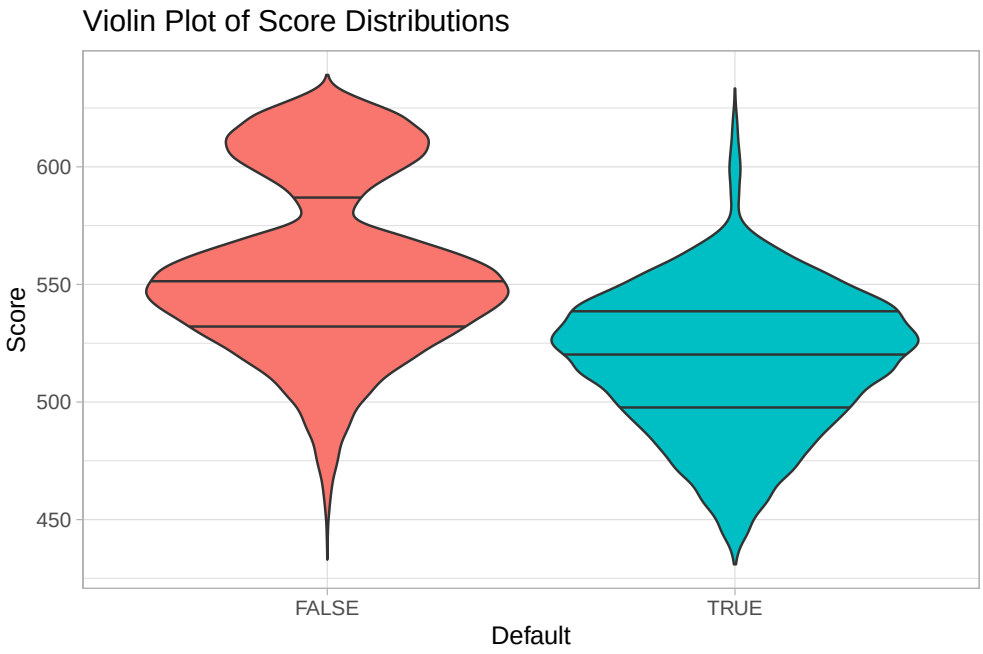

Figure 22: A Violin Plot outlining the distribution of Scores of the respective borrowers, with quantiles displayed at 25%,50% and 75% respectively.

The violin plot above contains a distribution of the defaulted and non-defaulted cases across the score range. In addition, it contains three quantile plots present at 25, 50 and 75 percent of the total population respectively. As with the previous visualisation of the score distribution, the violin plot indicates that a large percentage of defaulted cases harbour below a range of 550. This is in fact desireable, indicating that the model does show some strong distinguishable factors in determining a borrowers likelihood to default. That being said, there still remains up to 50 percent of non-defaulted cases which also harbour within this region. The hope is that calibrating these scores into grades and performing a new mapping of score ranges which help to address this issue and further bolster the model quality.

Calibration

Once the distribution of scores was analysed and deemed satisfactory, it was time for model calibration. During this stage of the process, the scores were mapped to grades, which can be determined at the discretion

of the user. This would be the figure which each borrower would ultimately receive to denote their credit score when applying for a line of credit. It was decided to utilise 8 grades, between the scales of our score distribution, namely 400 - 800 at equal intervals. This was the inital procedure, which was modified manually several times to obtain a more suitable score range which adequately depicted the distribution of scores. Furthermore, an optimisation algorithm was utilised to re-assess the grade range selection and find a more suitable distribution of grades based on specific constraints. Once each customer had been allocated to a grade, the model was ready to be optimised. The calibration was performed and a number of metrics to assess the quality of calibration is displayed below:

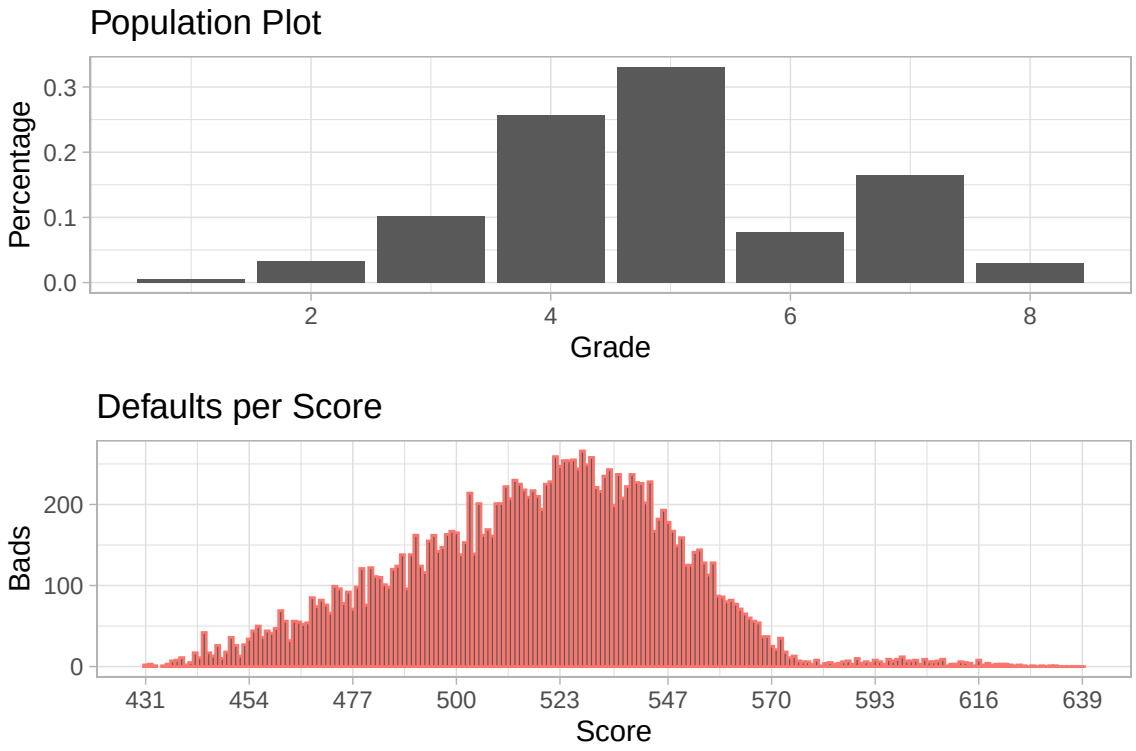

Figure 23: Analysis of the grade selection by determination of the population present within each grade and the number of defaults per score respectively, performed during the calibration process of mapping score ranges to suitable grades1.

As can be seen, the current grade selection is far from optimal. The population plot illustrates that the grade choice is not representative of the total population due to the large percentage of people harbouring between the scales four to seven. In contast, the defaults per score plot does show a decreased relationship of defaults as we progress from those likely to default, at lower scores to the lower risk clients at the higher scores.

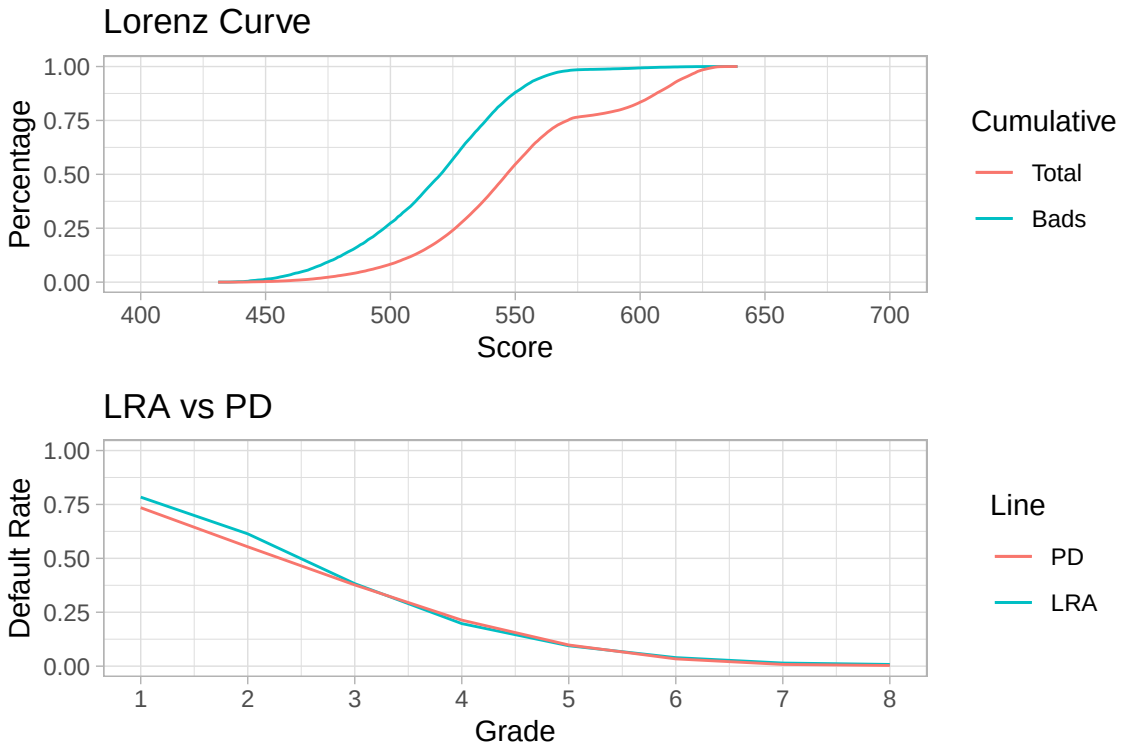

Figure 24: Analysis of the grade selection, by utilisation of the Lorenze Curve and LRA vs PD comparison, performed during the calibration process of mapping score ranges to suitable grades.

The long run average default rate or LRA measures the risk of the loan and plays an important role in the regulatory framework. In this context it is a measure of :

$$LRA = \frac{GDR_i - GDR_{i+1}}{GDR_{i+1}}, \tag{38}$$

where GDR_i represents the grade default rate in bucket i.

Thus, The LRA curve shows the actual observed long run average default rate for given grades. In an ideal scenario, PD should always be equal to or higher than the LRA for any grade though this can depend on calibration. With that in mind, it can be seen that grade 1 and 2 have a higher LRA in comparison to PD, which is certainly not ideal. This means that for these given grades, the probability of default model is under estimating the underlying risk. It must be noted that the population of people present within these grades is quite small, which is in itself its own issue and must be taken care of during calibration.

The Lorenz curve can similarly be used to show the discriminatory power of the scoring function, usually represented as an empirical cumulative distribution function of good against bad clients. The above graph shows a modified lorenz curve which compares the cumulative bads against the cumulative total as a function of score [2]. It can further be utilised to see at which scores the number of bads are starting to increase.

To compensate for this, as well as the peculiar distribution of the scores, analysis was performed on the grade selection. The function split the distribution of scores into groups, calculating its modal weight of evidence and comparing it to that of the global modal weight of evidence. This was performed for each variable in the model and the results of which are illustrated below:

Figure 25: Grade analysis determining whether, for a particular variable, there exists a different modal WoE between a grade in comparison to the modal WoE of the entire dataset.

This was performed to ensure that each grade in the final grade selection would display characteristics representative of the whole data set. It furthermore added a further net to catch anomolies which may have been present in this large data set.

As can be seen, there are a number of issues whereby different risk drivers seem to be present at different ranges of the distribution. This could be due to the grade selection and/or the score range. This issue will be addressed further during the optimisation process by altering the number of grades or by more suitable grade intervals.

Table 11: Calibrated Scores Per Grade

Grade	Score Upper	Score Lower	Population	Number of Defaults	Grade Default Rate
1	457.5	430.0	630	463	0.7349
2	485.0	457.5	3728	2064	0.5536
3	512.5	485.0	11641	4385	0.3767
4	540.0	512.5	29335	6260	0.2134
5	567.5	540.0	37696	3700	0.0982
6	595.0	567.5	8860	298	0.0336
7	622.5	595.0	18715	150	0.0080
8	650.0	622.5	3332	10	0.0030

Optimisation

Once reasonable grade choices have been made, it is time to optimise their selection using the NLcOptim package which uses sequential quadratic programming (SQP) to find the solution for a general non-linear op-

timisation problem. SQP is an interative method for non-linear optimisation whereby the objective function and the constraints are twice continuously differentiable.

It was attemtped to determine the logit score bounds for each of these grades using this algorithm under the following constraints:

$$0.01 \leq N \leq 0.3, \tag{39}$$

where N represents the percentage of the total population represented in each grade. In addition, this was an attempt to minimise the observed default rate (ODR) and the probability of default (PD).

The optimised grades were chosen as follows:

Table 12: The Optimised Upper and Lower Scores obtained for the desired number of grades under suitable constraints.

Grades	Initial Score Lower	Initial Score Upper	Optimised Scores Lower	Optimised Scores Upper
1	430.0	457.5	430.0000	457.5000
2	457.5	485.0	457.5000	488.4397
3	485.0	512.5	488.4397	512.5000
4	512.5	540.0	512.5000	530.1509
5	540.0	567.5	530.1509	567.5000
6	567.5	595.0	567.5000	594.9890
7	595.0	622.5	594.9890	622.5000
8	622.5	650.0	622.5000	650.0000

Unfortunately, as can be seen from the above table, the optimiser was not suitably sensitive to the upper and lower bounds provided when performing this analysis. It is unclear what is preventing the algorithm from optimising the grade selection. A number of score ranges were attempted to verify whether a local minima was obtained, yet the algorithm failed to produce any satisfactory results. It is hypothesized to add further constraints to investigate the nature of the error. For the determination of Capital Requirements, the final stage of the modeling process, the un-optimised grade selections were chosen.

Capital Requirements

In June 2004, the Basel Committee issued a Revised Framework on International Convergence of Capital Measurement and Capital Standards (also known as "Revised Framework" or Basel II). This paper takes into account these new developments in the measurement and management process using the internal ratings based (IRB) approach [1].

This approach allows for the use of internal measures to identify their own key risk drivers of credit risks as their own primary arguments for capital calculation. This is of course subject to meeting certain regulatory conditions and supervisory approval. All institutions using the IRB approach are permitted to determine their own metrics to estimate default probabilities. Furthermore, those using the advanced IRB approach are permitted to rely on internal estimates of loss given default and exposure at default on an exposure by exposure basis. These calculations are then mapped into risk weighted assets and regulatory capital requirements by formulas specified by the Basel Committee [8].

Expected Losses

Thus, a financial institution can forecast their average level of credit losses which it can reasonably expect to experience at any given time, which is denoted as the Expected Losses or EL. These expected losses are a

combination of several other factors, namely PD, EAD and LGD which stand for the probability of default, exposure at default and loss given default respectively. This is illustrated by the following formula:

$$\text{EL} = \text{PD} \times \text{EAD} \times \text{LGD}. \tag{40}$$

Loss Given Default(LGD)

The loss given default is the value of credit which a bank or other financial institutions are exposed to when a borrower defaults on their line of credit and is depicted as a percentage of the total exposure at the time at which default has occured. In the case of this report, due to Prosper being a P2P lending facility, it is estimated that this exposure is high and an estimate of 75% will be utilised.

Exposure at Default (EAD)

The exposure at default or credit conversion factor is the total value a lender is exposed to when a loan defaults. In the case of this report, it is the sum of the expected value of each loan outstanding at the time the credit listing was pulled. In general, this value is dynamic and changes with time as the borrowers repay their credit.

The exposure at default can be estimated in the following way:

$$A = P\frac{r(1+r)^n}{(1+r)^n - 1}, \tag{41}$$

where P is the initial principal (loan amount), r is the interest rate per period and n is the number of payments. The value of A, the current loan outstanding, is summed across all the borrowers to obtain the EAD at the time the credit profile was pulled. It must be noted that the EAD calculations were a rough estimate and did not take into account those people who were past due at varying levels but not yet considered defaulted. As there were only a small fraction of people within this range, the result will have an insignificant impact.

The exposure for each individual borrower was calculated and aggregated to obtain the EAD, which was valued at ~ \$563 Million out of a total of ~ \$697 Million or ~ 81%, of the total loan origination amount for the borrowers.

Table 13: Expected Loss Calculation Table Based on Suitable Estimates.

Variable	Estimates
PD	Borrower Dependent
EAD	~ \$ 563.108 Million
LGD	0.75
EL	~ \$ 64.741 Million

Risk Weighted Assets (RWA)

As part of the Basel II IRB risk weighted functions, specific values which are used within the IRB formulas are asset class dependent. This is due to the fact that they show different degrees of freedom of dependence on aggregate macro-financial conditions. Thus, institutions must categorize these exposures into a choice of five different asset classes which are corporate, soverign, bank, retail and equity respectively. For corporate, soverign and bank exposures, the value of K or unexpected loss can be calculated as follows:

$$K = \text{LGD} \times (\text{WCDR} - \text{PD}) \times \text{M}, \tag{42}$$

where LGD is the loss given default, PD is the probability of default, M is the adjusted maturity and WCDR is the worst case default rate respectively. This formula can be more explicity expressed as:

$$K = UL = \left[\text{LGD} \cdot N\left[\frac{\sqrt{\rho}N^{-1}(0.999) + N^{-1}(PD)}{\sqrt{1-\rho}}\right] - \text{PD} \cdot \text{LGD}\right] \times \frac{1 + (M - 2.5) \times b}{1 - 1.5 \times b}, \tag{43}$$

where N and N^{-1} represent the gasussian and inversed distribution function respectively. M is the average portfolio effective maturity and b is the maturity adjustment. This maturity adjustement can be calculated as :

$$b = (0.11852 - 0.05478 \times \log(\text{PD}))^2. \tag{44}$$

The minimum capital requirement is then set such that unexpected losses do not exceed the banks capital up to a 99.9% confidence level. In addition, the average portfolio maturity is assumed to be 2.5 years. Maturities with exposures beyond that time will necesitate holding more capital. In the case of this project ρ is estimated by the asset correlations. The asset correlations in accordance with capital requirements regulations can be estimated as follows:

$$\text{Asset Correlations } \rho = 0.03 \times \frac{1 - e^{-35 \times \text{PD}}}{1 - e^{-35}} + 0.16\left[1 - \frac{1 - e^{-35 \times \text{PD}}}{1 - e^{-35}}\right], \tag{45}$$

which permits the range for ρ to be between 3 - 16%. The risk weighted assets (RWA) can then be calculated as :

$$\text{RWA} = K \times 12.5 \times \text{EAD}. \tag{46}$$

The RWA was calculated and compared to the probability of default. The results are as follows:

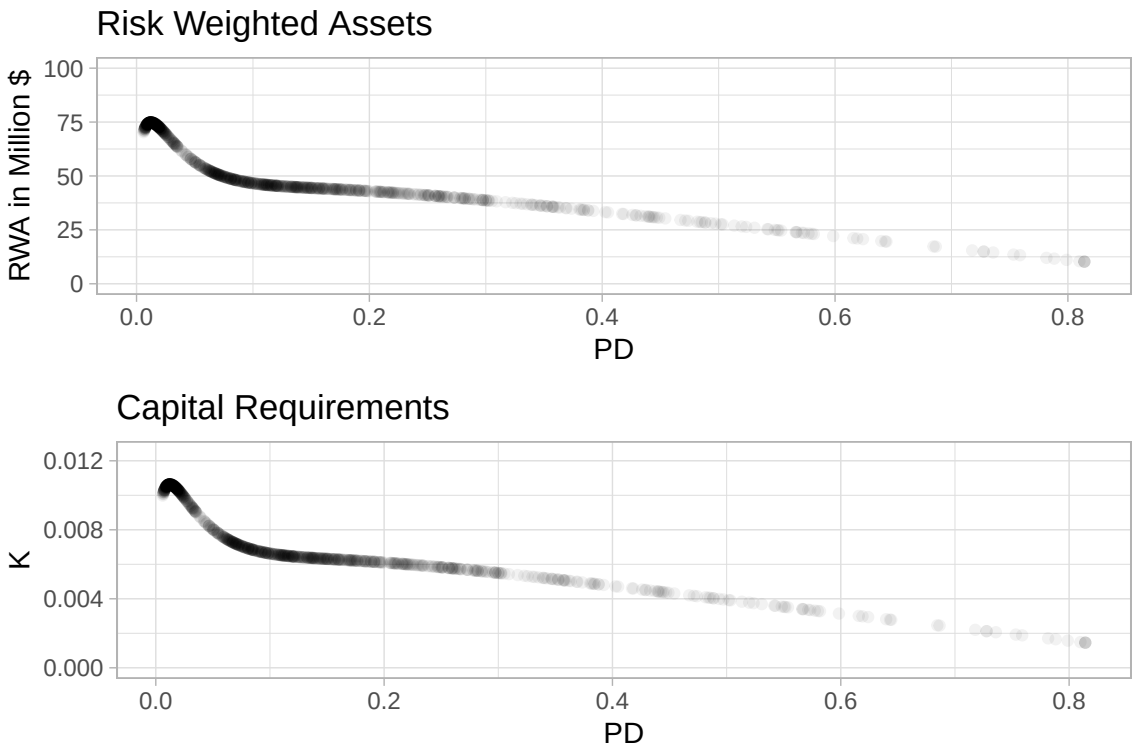

Figure 26: Comparison of Capital Requirements and Risk weighted Assets based on the borrowers probability of default and population of people over the range of default probabilities.

As can be clearly seen in the figures above, an increase in PD corresponds to a decrease in capital requirements and risk weighted assets. This naturally seems unintuitive but it must be noted that these requirements are scaled based on the population that is present at these default rates. To further illustrate this effect, the exposure at default is illustrated for each respective grade as follows:

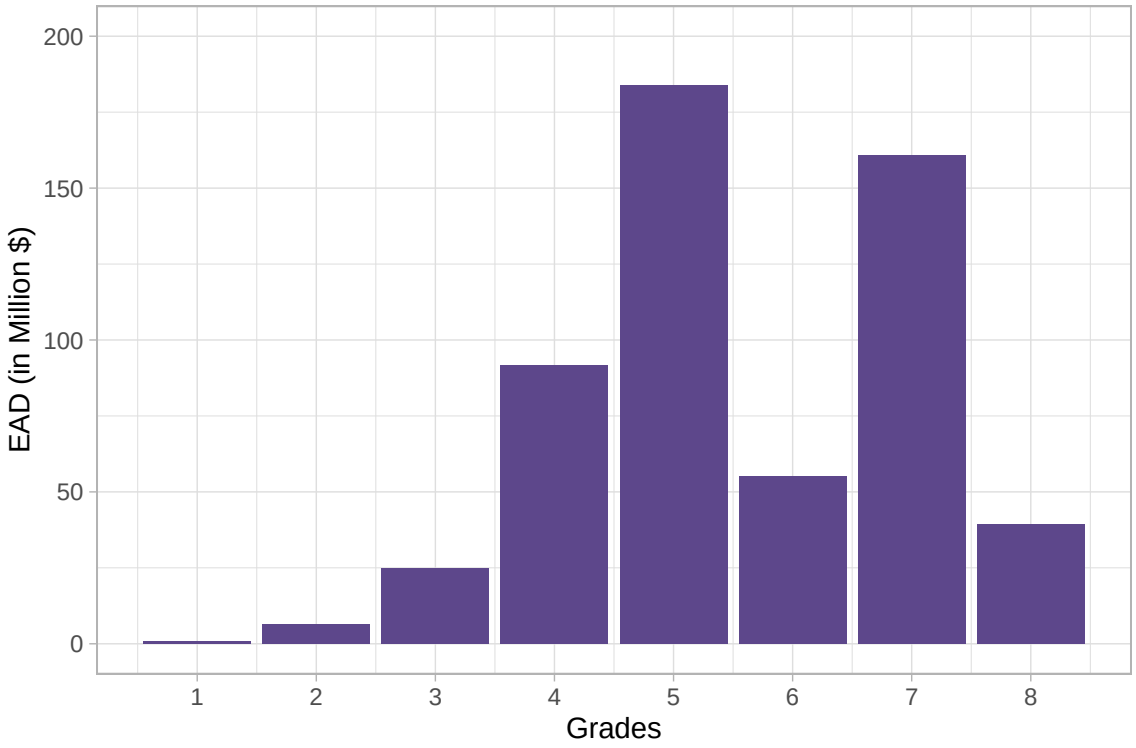

Figure 27: Exposure at Default in Million Dollars based on each respective Grade.

The above graph illustrates the exposure at default for each respective grade. This was calculated by determining the grade and independent EAD of each borrower and summing up those values for each grade. Intuition would assess that the lowest grades would have the highest exposure due to their natural higher probabilities of default, yet this exposure is also determined by the number of borrowers represented within each grade. Thus, the highest exposure is seen at the most prevelant grades, where this bi-modal structure is once again visualized as it was in the score distribution.

By utilising the EAD per grade, it was possible to determine the the RWA which must be kept by banks and other lenders in order to reduce their risk of insolvency. A summary of the factors and values in this evaluation is outlined in the following table:

Table 14: Expected Loss Calculation Table Based on Suitable Estimates. The RWA and EAD are displayed in dollars.

Grade	K	RWA	EAD	Population	Population Percentage
1	0.0016870	19670.25	932768	630	0.0055
2	0.0030304	246267.08	6501296	3728	0.0327
3	0.0048686	1520431.21	24983657	11641	0.1022
4	0.0060898	6973867.18	91614339	29335	0.2575
5	0.0067844	15580043.93	183717303	37696	0.3308
6	0.0088644	6100034.24	55051839	8860	0.0778
7	0.0104994	21114720.09	160882647	18715	0.1643
8	0.0103556	5103213.46	39423936	3332	0.0292

Following on from this, the RWA was calculated and the results are displayed below:

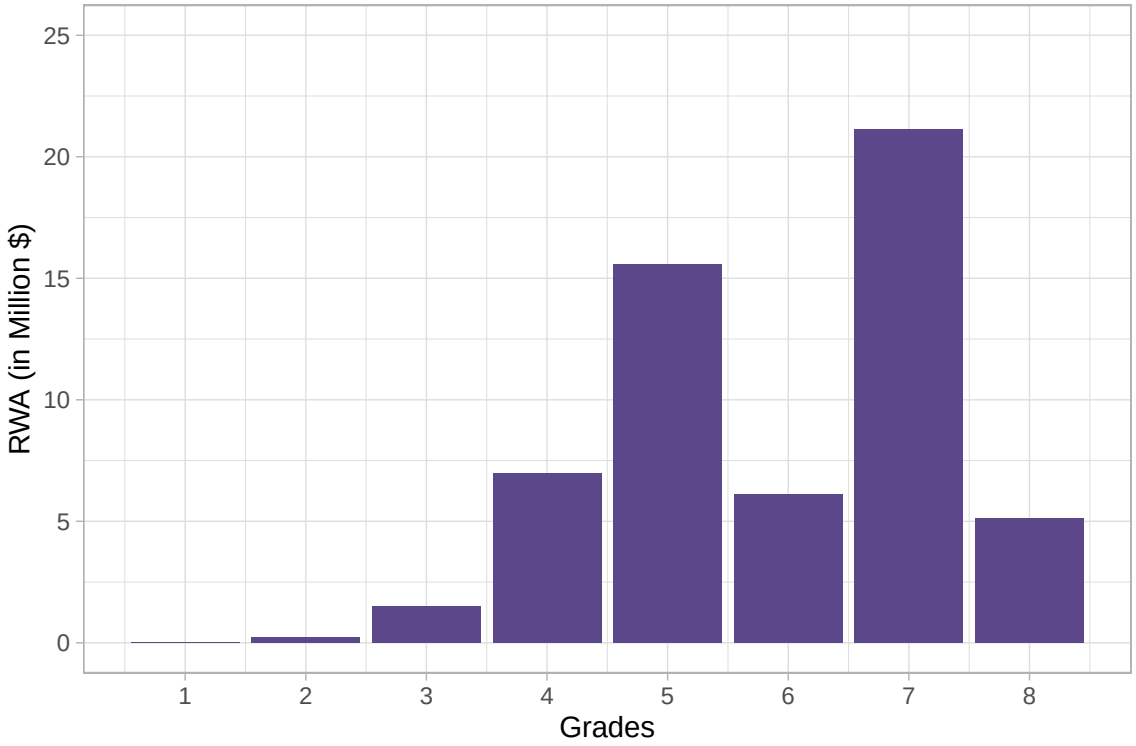

Figure 28: Calculation of RWA for the borrowers within each grade respectively.

As can be seen, there is a direct linear correlation between the RWA and EAD, as would be expected. The total RWA for a financial institution is then determined by aggregating the RWA for each grade respectively. This amounts to 56.69 million dollars of the total 697.20 in credit offered, or 8.13% respectively.

Discussion

The methodical step by step approach in this project resulted in a number of interesting results. The segmentation analysis stood out particularly. Though there was no concrete indicator for developing multiple scorecards, there is still room for manouevre using more advanced segmentation techniques. In addition, custom grouping of the categorical variables may lead to more distinct leaf nodes and suitable risk drivers which have yet to be explored. Furthermore, the presence of the bi-modal score distribution is certainly not ideal from a practical perspective, where a symmetrical distribution would be preferable. This particular issue is caused by the variable *Investors*, where it is postulated that borrowers with a large number of investors show an increased probability of default. This postulation was further validated from the observable coefficient obtained during the forward regression process. Thus, this gives rise to the possibility of segmenting borrowers with a single investor and those with multiple investors to produce multiple scorecards.

Due to the large number of variables present in the dataset, it sets the scene for the utilisation of a random forest. Random forests are an ensemble of decision trees which may more accurately provide suitable segmentation opportunitites. They are similarly useful for compensating for overfitting, which is a common problem associated with decision trees.

This gives rise to alternative options during the predictive process. Once the single factor analysis had been completed and the model had been narrowed to a more suitable selection of variables, it was decided to utilise forward regression instead of backward regression due to the desire to have a model with as few variables as possible, which is more suitably obtained by the forward regression methods. Another alternative is that of stepwise regression which could more aptly manage the large amount of potential predictor variables. Logistic regression relies heavily on data from the past and due to the relative rareness of defaults, could underestimate the probability of a default occurence . Statistical techniques such as generalized extreme value regression as illustrated by Calabrese and Osmetti [10] could provide an interesting alternative. The GEV approach could circumvent the drawbacks of logistic regression and more suitably describe the tail behaviour of the distributions. One further comment pertains to the use of K-fold cross validation, which is a resampling procedure to evaluate machine learning models on a limited data sample. It is a popular alternative to the classic train/test split to divide the data into K number of groups and to fit a model on one group independently, evaluating it on the remaining test set. The model is evaluated, discarded and the process is repeated for the remaining K-1 groups. This generally provides a less biased and less optimistic model though it must be noted that there is a sensitivity to the choice of K. K must be chosen such that the test and training sets are large enough to be representative of the dataset and one must also consider the bias and variance trade-off imposed by this implementation.

This leads to questioning the discretization which was performed utilising recursive partitioning. Though this may yield the most optimal results from a statistical standpoint, it may be interpreted negatively from a business one. Therefore, a more intuitive categorization may be more suitable prior to feeding these values through the regression model process. This was in fact performed, yet only utilised in a few key scenarios and only regarding discrete data. There would be a trade off and the model would lose some of its predicting power yet would be a more apt description of the risk drivers.

That being said, the final chosen model which yielded a Gini of ≈ 0.59 showed a significant predicting power across the training and test sets. A common rule of thumb is a Gini index above 0.4 are considered suitable model choices which thus indicated that a viable model choice was chosen in modeling the borrowers probability of default. As was previously mentioned, it must be noted that the score distribution obtained from this model did show an uncharacteristic double bump within the range of scores. The presence of this valley in between the two peaks is certainly not ideal, indicating that a score of 550 was difficult to obtain and there is a cluster of borrowers with high credit scores afterwards. In an ideal world, the distribution would tail off. As was previously stated, the variable responsible for this was that of *Investors*, which due to the peer to peer lending characteristics of the dataset corresponded to the total amount of investors a borrower had when requesting their line of credit. It is therefore believed that an increase in the number of investors indicates a sign of weakness for the borrower and an increase in their perceived likelihood to default. Once *Investors* was removed from the model, the distribution resembled that of a gaussian distribution which certainly shows certain desireable characteristics yet it similarly resulted in a substantially decreased Gini of approximately 0.52. It was decided that this decrease was far too substantial and *Investors* remained in the final model. The presence of this issue does require further investigation, whether there is cause for a secondary scorecard for borrowers with one sole investor and those with greater than one. Customised discretization would similarly be utilised, yet is undesireable due to the issues which were listed above.

One major room for improvement is the process of optimisation, once the grade selection was performed. Unfortunately, the optimisation algorithms lackluster attempt to optimise did not yield statistically significant results. It is postulated that the algorithm is sensitive to the choice of grade selection, especially in grades where there is a small percentage of the population present. In these scenarios, the algorithm immediately converges and does not alter the upper nor lower bound of the score ranges. An alternative option to investigate could be to structure the grades in a manner whereby each grade contained equal portions of the population or equal default rates, as was desired. Furthermore, these issues could be circumvented by imposing additional and more strict constraints in an effort to find more viable optima. Of course, the manual approach to this issue is one resolving solution, though does leave little to the imagination and does not follow the automated and general structure outlined through the majority of the report.

Another interesting avenue to consider is to only study the effects of the borrower default rate post the financial crisis of 2008. This is due to the fact that the 2008 financial crisis was an extreme economic event

which would have affected all borrowers, causing some to default where under normal circumstances they would have most likely repaid their line of credit. Events such as this may have distorted the predictive capabilities during the modeling process by falsely interpreting certain characteristics as those likely to default, when in fact these characteristics would have only been temporary predictor variables and are not representative of long term predictors of default.

Though there were a number of issues encountered in the project, it was possible to succesfully model the probability of default using some machine learning techniques and obtain viable reserves which a financial institution should hold which met capital requirement regulations.

Appendix

Variable Removal

Table 15: Variable Removal Table and brief description Outlining the reasons for removal.

Variable	Stage	Removal
LenderYield	SFA	Future Information
EstimatedLoss	SFA	25.5% Missing Rate
BorrowerAPR	SFA	Future Information
TotalCreditLinespast7years	SFA	Lower IV compared to TotalTrades
OpenCreditLines	SFA	Lower IV compared to CurrentCreditLines
OpenRevolvingAccounts	SFA	Lower IV compared to CurrentCreditLines
MonthlyLoanPayment	SFA	Lower IV than LoanOriginalAmount
OpenRevolvingMonthlyPayment	SFA	Lower IV than Revolving Credit Balance
TotalInquiries	SFA	Split Decision - Similar MR and IV to InquiriesLast6Months
LoanMonthsSinceOrigination	SFA	Completed Loan % is Substantial
DebtToIncomeRatio	SFA	Individual Gini too Low
PublicRecordsLast10Years	SFA	Individual Gini too Low
ListingKey	SFA	Contains No Predictive Information
Term	SFA	IV too Low
ListingNumber	Binning	Contains No Predictive Information
ListingCreationData	Logistic Regression	Contains No Predictive Information
CreditGrade	Logistic Regression	Third Party Analysis
IsBorrowerHomeowner	Logistic Regression	Binary Data
CurrentlyInGroup	Logistic Regression	Binary Data
IncomeVerifiable	Logistic Regression	Binary Data

[a] Any other Variable which is not present in this table was removed during pre-processing and single factor analysis (SFA).

Decision Tree Iterations

The initial Decision tree was compiled without any ammendments, requiring only a minimum of 500 individuals within each node. The results are as follows:

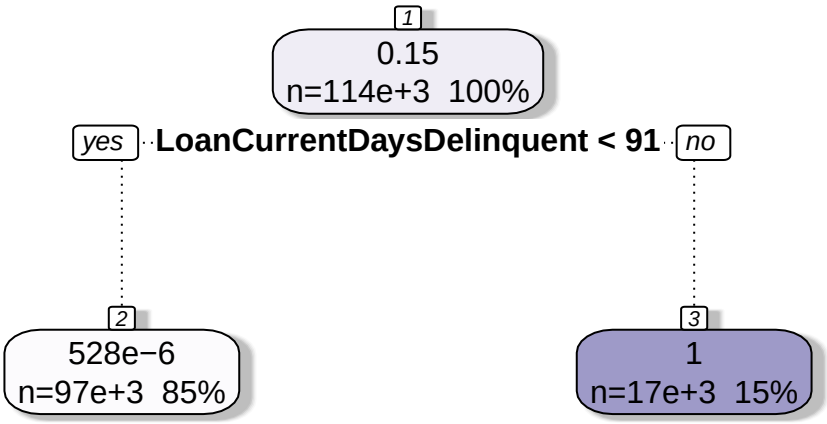

Figure 29: The Initial Decision Tree Obtained during segmentation, prior to any ammendments or manipulation.

The above graph then shows the distribution of borrowers across the chosen risk drivers. Each node contains the percentage of population, the corresponding number of people within that node (n) and finally the observed default rate within the node.

As it can be seen, *LoanCurrentDaysDelinquent* is a dominant variable and is inhibiting any further nodes from being created.Thus, this variable along with *BorrowerState* and *LoanMonthsSinceOrigination* inhibit further leaves and are removed from the dataset. The decision tree is iterated further and the results are as follows:

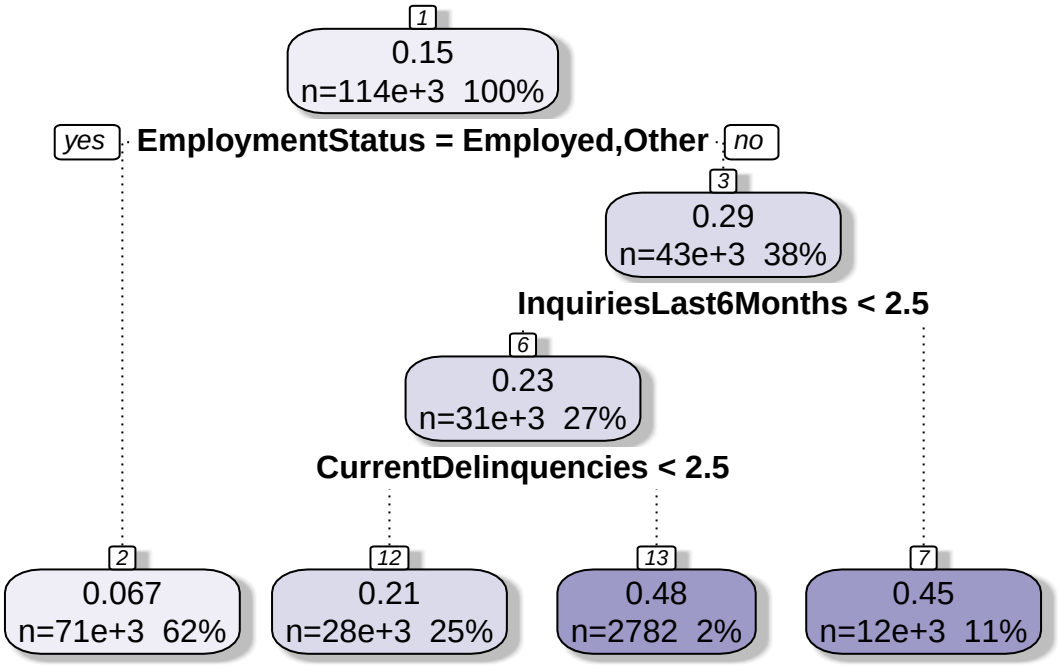

Figure 30: The Decision Tree obtained after removal of dominant variables 'LoanCurrentDaysDelinquent', 'BorrowerState' and 'LoanMonthsSinceOrigination'.

It can be seen to start that the dominant variable was that of EmploymentStatus, which is to be expected due to its high information value. Due to the dominance of this variable, custom categorisation was applied to allow for some more suitable leaf nodes. Thus, EmploymentStatus was re-categorized into Employed, Unemployed, Missing and Other. The results are as follows:

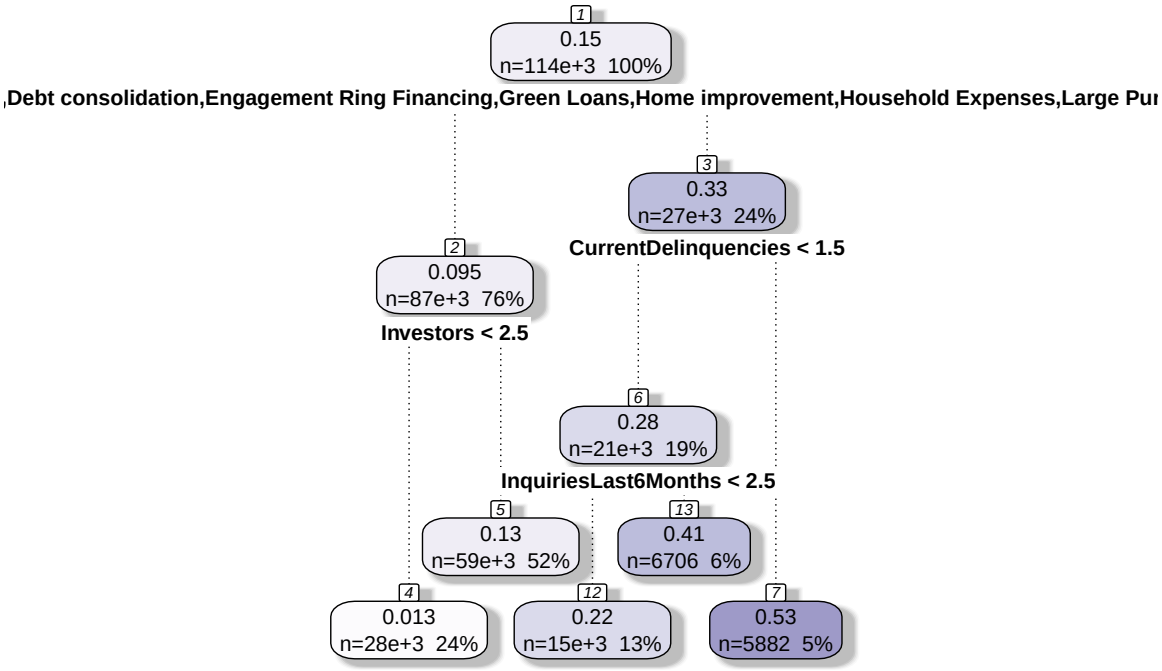

Figure 31: Decision Tree Visualisation After Customized EmploymentStatus.

From the tree above, a key distinguishable leaf node is that of ListingCategory. ListingCategory is currently categorizable into 17 types of listings and in order to allow for more suitable segmentation, this variable was re-categorized into Debt and Consolidation, Missing, Home and Other. The results are as follows:

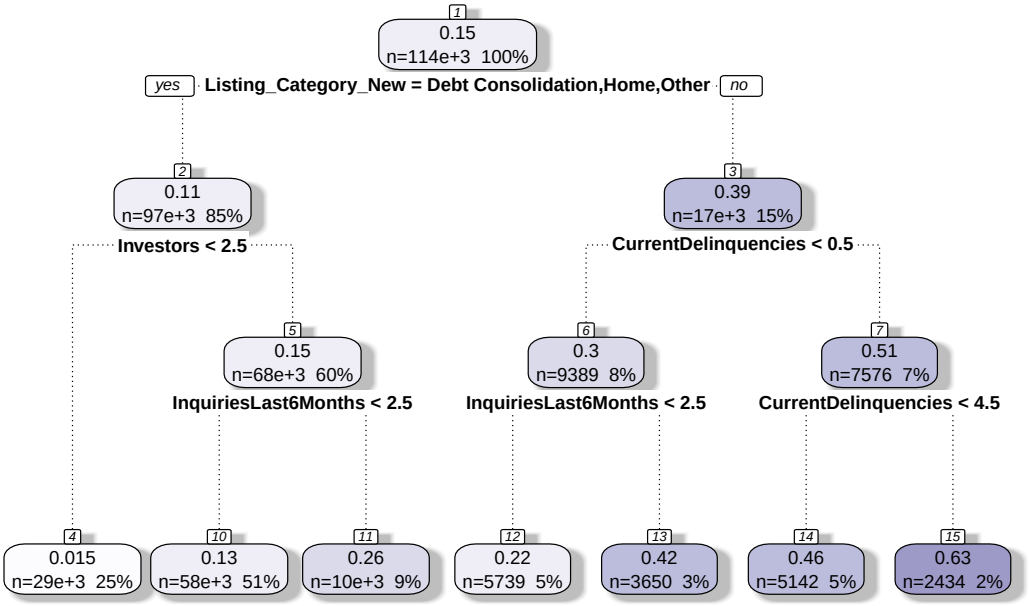

Figure 32: Decision Tree Visualisation after the removal of Listing Category.

By similar logic, ListingCategory_New is removed as both nodes exhibit similar risk drivers. This yields the following:

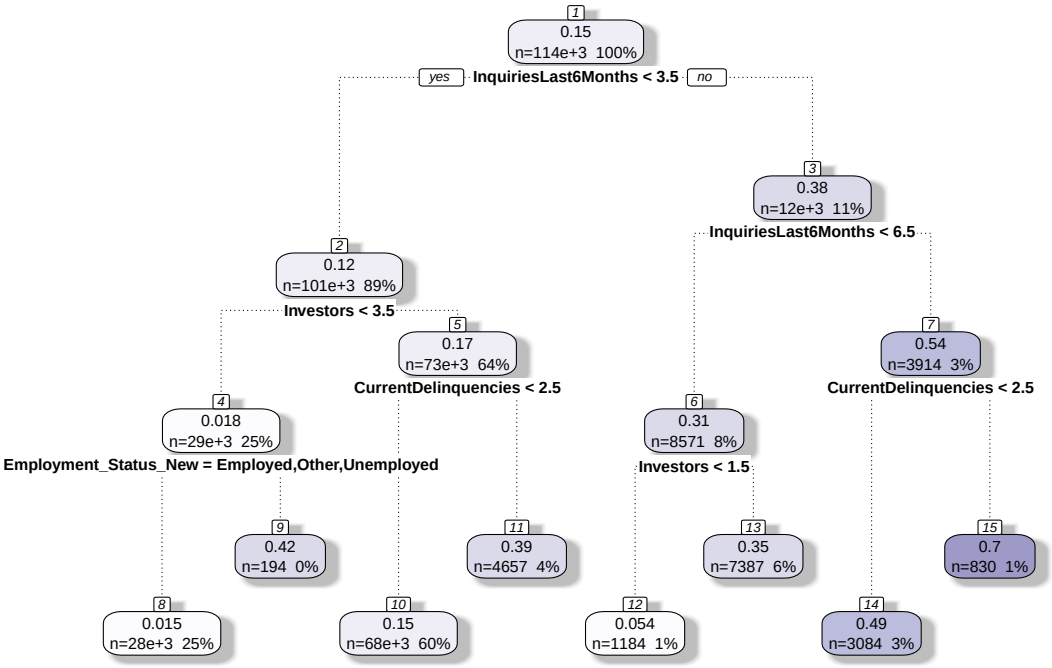

Figure 33: Decision Tree Visualisation after the removal of ListingCategoryNew.

This tree was iterated again with the removal of $InquiriesLast6Months$, $CurrentlyInGroup$ and $CurrentDelinquencies$. They were removed due to their similar risk driving characteristics across both nodes. In addition $CurrentlyInGroup$ is a binary variable and it was deemed to not be a suitable splitting candidate. These variables were removed and the results are as follows:

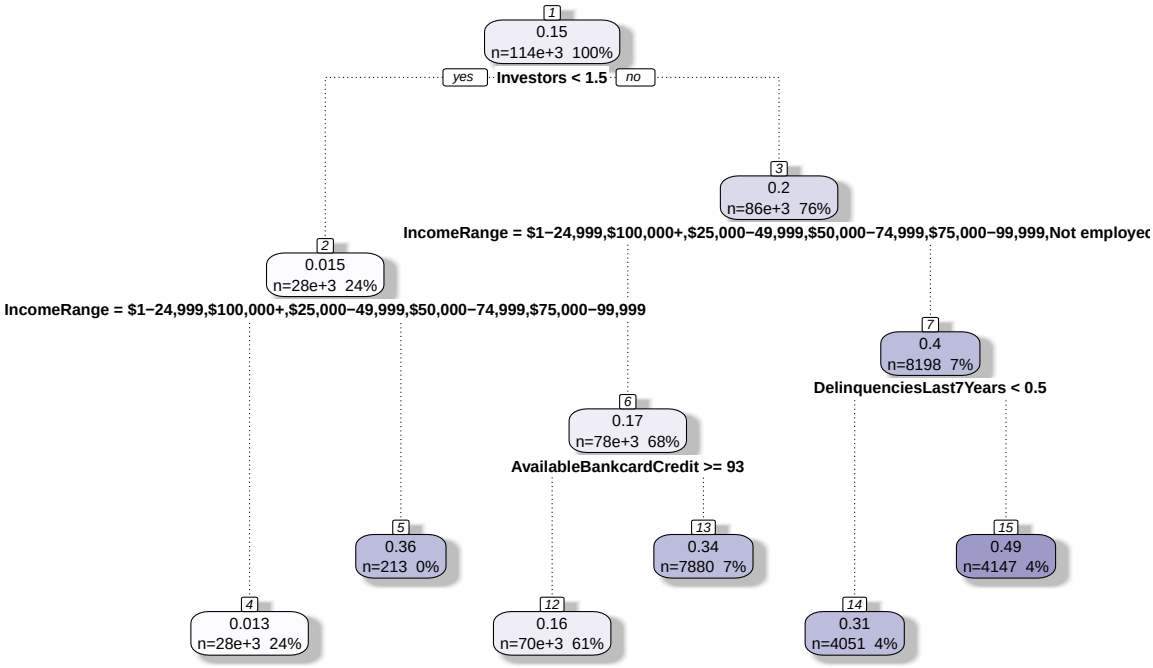

Figure 34: Decision Tree Visualisation after the removal of InquiriesLast6Months, CurrentDeliquencies and CurrentlyinGroup.

In this particular case, rather than removal, *IncomeRange* is customised into a new grouping scheme. These are Low, Medium, High and Other yielding the following decision tree:

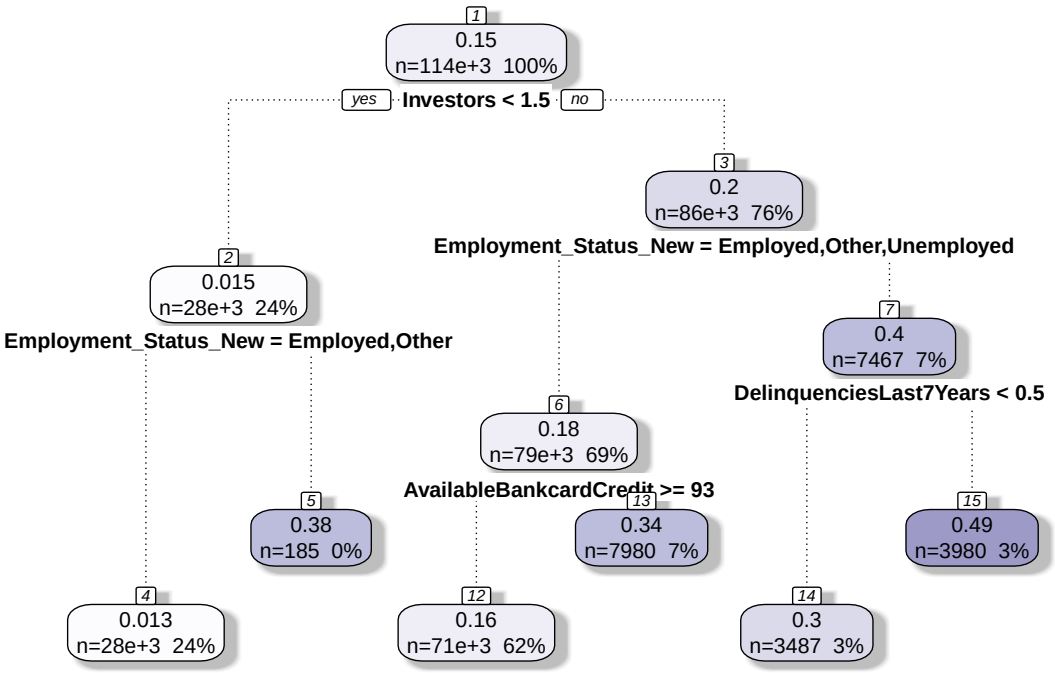

Figure 35: Decision Tree Visualisation after the customization of IncomeRange.

Finally, it was seen that *EmploymentStatus_New* was constituted as a dominant variable. It was removed to establish if other leaf nodes would flourish in its absence. This was the final iteration, leading to the proposed model and the deduction that it was not possible to adequately determine a potential segmentation candidate with the use of decision trees.

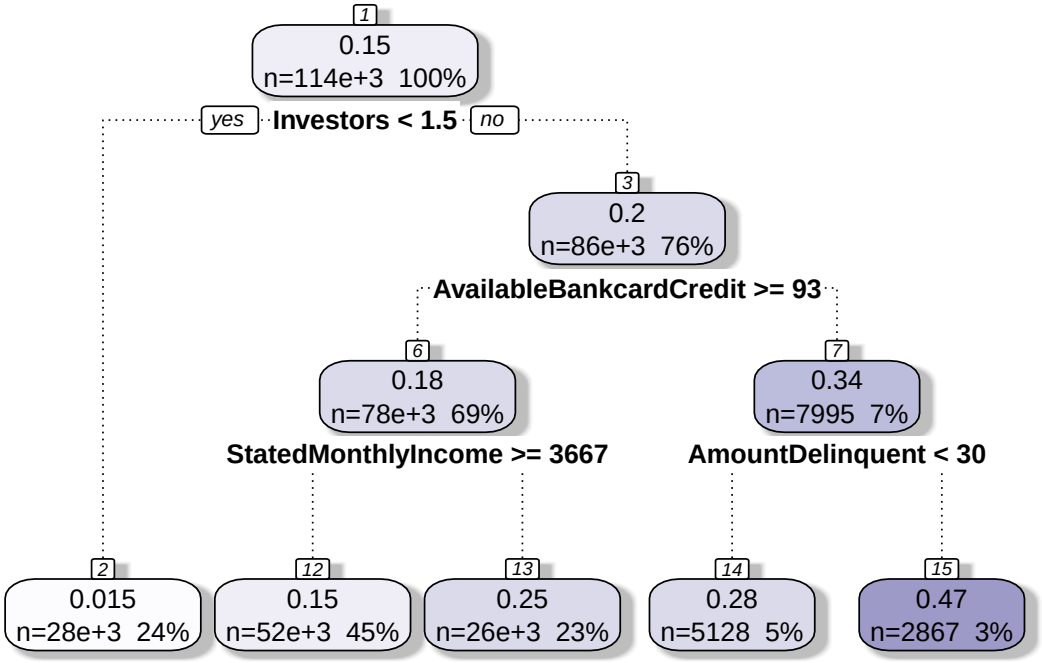

Figure 36: Decision Tree Visualisation after the removal of EmploymentStatusNew

Derivation of the Loss function using gradient descent.

The loss function can be written as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} C(h_\theta(x^{(i)}), y^{(i)}) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log\left(h_\theta(x^{(i)})\right) + (1-y^{(i)}) \log\left(1-h_\theta(x^{(i)})\right)\right]. \tag{47}$$

By utilising the definition of the hypothesis function:

$$h_\theta(x) = \frac{1}{1+e^{\theta x}}, \tag{48}$$

it is possible to fill in this expression into the cost function to obtain the following:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} -y^{(i)} \log\left(1+e^{\theta x^i}\right) + \left(1-y^{(i)}\right)\left(\theta x^i - \log\left(1+e^{\theta x^i}\right)\right)\right], \tag{49}$$

which can be further simplified as:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} -y_i \theta x^i + \theta x^i - \log\left(1+e^{\theta x^i}\right)\right] = \frac{1}{m}\left[\sum_{i=1}^{m} y_i \theta x^i + \log\left(1+e^{-\theta x^i}\right)\right]. \tag{50}$$

The second equality follows from the fact that:

$$\theta x^i - \log\left(1 + e^{\theta x^i}\right) = \log\left(e^{\theta x^i}\right) - \log\left(1 + e^{\theta x^i}\right) = \log\left(\frac{e^{\theta x^i}}{1 + e^{\theta x^i}}\right) = -\log\left(1 + e^{-\theta x^i}\right). \tag{51}$$

By computing the partial derivatives with respect to θ, the gradient at a point can be expressed as:

$$\frac{\partial}{\partial \theta_j} - \log\left(1 + e^{-\theta x^i}\right) = \frac{x_j^i e^{-\theta x^i}}{1 + e^{-\theta x^i}} = x_j^i h_\theta(x^i), \tag{52}$$

$$\frac{\partial}{\partial \theta_j} y_i \theta x^i = y_i x_j^i. \tag{53}$$

References

[1] "European Central Bank". *Guide for the Targeted Review of Internal Models (TRIM)*. 2017. URL: https://www.bankingsupervision.europa.eu/ecb/pub/pdf/trim_guide.en.pdf.

[2] František ŘEZÁČ" "Martin ŘEZÁČ. "How to Measure the Quality of Credit Scoring Models". In: *Czech Journal of Economics and Finance* 61.5 (2011).

[3] "Prosper". *What does it mean for a loan to be charged-off?* URL: https://prosper.zendesk.com/hc/en-us/articles/208500546-What-is-a-charge-off-.

[4] Achim Zeileis" "Torsten Hothorn. "A Modular Toolkit for Recursive Partytioning in R". In: *Journal of Machine Learning Research* 16 (2015), pp. 3905–3909.

[5] Alan Agresti. *An Introduction to Categorical Data Analysis*. Vol. 2. Wiley, 2007.

[6] Trevor Hastie et al. *The Elements of Statistical Learning, Data Mining, Inference and Prediction*. Vol. 2. Springer, 2009.

[7] *An Explanatory Note on the Basel II IRB Risk Weight Functions*. Basel Committee on Banking Supervision. Bank for International Settlements, July 2005.

[8] European Banking Authority. *Interactive Single Rulebook : Capital Requirements Regulation*. URL: https : / / eba . europa . eu / regulation - and - policy / single - rulebook / interactive - single - rulebook/ - /interactive-single-rulebook/toc/504.

[9] Somnath Chatterjee. *Modeling Credit Risk*. Bank of England. Centre for Central Banking Studies, Bank of England, Threadneedle Street, London, EC2R 8AH, 2015.

[10] Silvia Angela Osmetti Raffaella Calabrese. "Generalized Extreme Value Regression for Binary Rare Events Data: An Application to Credit Defaults". MA thesis. University College Dublin, University Cattolica del Sacro Cuore, 2011.

[11] Naeem Siddiqi. *Developing and Implementing Intelligent Credit Scoring*. Wiley, 2006.

[12] Wei Xia Steve Satchell. "Analytic Models of the ROC Curve: Applications to Credit Rating Model Validation". MA thesis. University of Cambridge, University of London, 2007.