Model #101: Credit Card Default Model

Model Development Guide

Juan A. Campos

Northwestern University

Masters of Science in Data Science

MODEL # 101: CREDIT CARD DEFAULT MODEL

1. Introduction

1.1 Overview

Four predictive models were constructed and evaluated to determine which was the most accurate and effective at predicting credit card default risk. The models were constructed using a Random Forest algorithm, an Extreme Gradient Boost algorithm, Logistic Regression, and a simple Artificial Neural Network. This study found that the model's overall classification accuracy did not correlate with the model's predictive value. In this regard, the models that achieved higher overall accuracy were not the models that achieved the best performance in predicting default and non-default cases. The Accuracy Paradox made an appearance in this study. The model with the lowest overall accuracy was also the model that most accurately and effectively predicted default risk.

1.2 Statement of Problem

Generally, default occurs when an individual has been delinquent (missed payments) for six consecutive months. After default has occurred, the banks writes-off the outstanding credit card balances and incur a financial loss. Although the banks conduct rigorous screenings to issue credit cards to qualified applicants, defaults still occur. Because the real probability of default is unknown, having a valid and up-to-date model or models to detect credit card default risk is crucial for risk management and to lessen financial losses resulting from defaults.

1.3 Purpose and Scope

This study is being conducted to explore the accuracy and effectiveness of four machine learning algorithms as it relates to detection and prediction of credit card default risk. Four models will be constructed using Random Forest, Extreme Gradient Boost, Logistic Regression, and an Artificial Neural Network. The constructed models will function as base models. Accordingly, activities to improve the accuracy of the models, such as implementing grid-search or parameter tuning, will not be conducted. The activities that will be addressed include data quality check, data cleaning and preprocessing, feature engineering, traditional and model-based exploratory data analysis, and model construction and evaluation. The metrics that will be explored to gauge the performance of the models will include the Accuracy Score, the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) Curve, and Confusion Matrix derived metrics such as Sensitivity, Specificity, and the F1 score. The models will not be simple classification models but instead will compute odds probabilities of default. A classification cut-off threshold point will be established using the ROC Curve. The odds probabilities values and the cut-off threshold will determine the default outcome (good = Non_Default; bad = Default). The R programming language and the R Studio environment will be used for this analysis. Several R packages will be employed including InformationValue, OneR, WoeBinning, Rattle, Psych, and SummaryTools. The Caret machine learning package will be used to construct the models.

2. The Data

2.1 Data Overview

A modified version of the *Default of Credit Card Clients* (Default dataset) dataset from the University of California's UCI Machine Learning Repository was used in this study. The dataset describes the revolving credit activities of credit card clients in Taiwan for the period April 2005 through September 2005. The dataset contains information on demographic factors, payment history, credit data, bill statements and information indicating whether or not the client defaulted. The dataset is comprised of 30,000 observations and 25 explanatory variables, including the binary response variable DEFAULT. The response variable has two categories (0 – No (non-default); 1 – Yes (default)). Each observation corresponds to an individual credit card user/customer and is assigned to a default outcome - bad or good. All 25 variables are continuous/integer variable types. A data dictionary of the dataset was on hand and was used to assist with the analysis of the variables. The data consists of the following variables:

- X1: ID (Identification Number)
- X2: Default (0 = No; 1 = Yes)
- X3: Balance Limit. Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X4: Gender (1 = Male; 2 = Female)
- X5: Education (1 = Graduate; 2 = University; 3 = High School; 4 = Other)
- X6: Marital Status (1 = Married; 2 = Single; 3 = Others)
- X7: Age (Year)
- X8 – X13: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: PAY_1 = the repayment status in September, 2005; PAY_2 = the repayment status in August, 2005; . . .; PAY_6 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- X14 – X19: Amount of bill statement (NT dollar). BILL_AMT1 = amount of bill statement in September, 2005; BILL_AMT2 = amount of bill statement in August, 2005; . . .; BILL_AMT6 = amount of bill statement in April, 2005.
- X20 – X25: Amount of previous payment (NT dollar). PAY_AMT1 = amount paid in September, 2005; PAY_AMT2 = amount paid in August, 2005; . . .; PAY_AMT6 = amount paid in April, 2005.

The dataset obtained for this study was a modified version that includes five additional variables that were appended to the original Default dataset. Based on random selection, these five variables were used to assign and flag each observation into one of three categories: Test/Train/Validate. These variables are the basis for splitting the Default dataset into three separate datasets corresponding to their assigned category. The five appended variables are listed below.

- X26: U: Random number assigned to variable (basis for the train/test/validate split)
- X27: Train: (0 = Not assigned to train data set; 1 = Assigned to train dataset)
- X28: Validate: (0 = Not assigned to validate data set; 1 = Assigned to validate dataset)
- X29: Test: (0 = Not assigned to test data set; 1 = Assigned to test dataset)
- X30: data.group: (1 = Test dataset; 2 = Validate dataset; 3 = Test dataset)

Table 1.1 summarizes the number of observations that were split into the Train dataset, the Test dataset, and the Validation dataset.

Table 1.1: Dataset Split

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
Train	15,180	50.60	50.60	50.60	50.60
Test	7,497	24.99	75.59	24.99	75.59
Validate	7,323	24.41	100	24.41	100
Total	30,000	100	100	100	100

2.2 Data Survey

A broad overview of the Default dataset was conducted at the outset to obtain a general sense of the variables and their values; to identify if the needed information was on hand to build the predictive model; and to obtain a general sense of the quality and completeness of the data. It was noted that the dataset did not contain metrics that are traditionally used to measure the income status, repayment ability and the character of the individuals. Such metrics include monthly/yearly income, total debt, debt-to-income ratio, employment status, number of years employed, etc. The data also does not contain credit worthiness scores such as FICO or credit bureau credit scores. Information describing the monthly use of the revolving credit or credit adjustment transactions was also not included in the data set. Nevertheless, the variables that are available in the dataset can be used to extract insights similar to those provided by the commonly used evaluation metrics; or can be used to develop substitutes for such metrics. Based on the overview of data, it was concluded that the dataset contains the correct and necessary information to construct the models.

2.3 Data Quality Check and Data Preprocessing

In performing the overview of the data set, it was noted that none of the dataset observations were missing values (NA). It was also observed that some variables contained a zero (0) or less as the minimum value. Zero or negative values were deemed appropriate in some cases because sporadic use of the revolving credit, inactive accounts, and credit adjustments may result in variables containing zero or negative values. In addition, the data dictionary specified that negative values were appropriate, such as with the repayment status variables (PAY_*).

Errors and discrepancies were observed in the data as well. The goal was not to remove these observations from the dataset because of the inconsistencies or errors. The removal of these observations could result in the loss of information and their exclusion could adversely impact the accuracy of the constructed models. Instead, the appropriate remedies were applied to correct the variables. Finally, it was also noted that many variables contained outliers. The outliers also needed to be addressed. Details of the observed errors and discrepancies, and any applied solutions, are discussed in the sections below.

2.3.1 Inconsistent Variable Name Labeling

The monthly payment history variables (PMT_AMT*), bill statement history variables (BILL_AMT*), and the monthly repayment status history variables (PAY_*) variables were sequentially numbered in descending order from the number six (e.g. PAY_- April 2005) to the number one (e.g. PAY_1 - September 2005). However, the September 2005 repayment status history variable was labeled with a zero value (PAY_0) instead of the number one. To maintain consistency and alignment with the other variables, the label for this observation was changed to PAY_1.

2.3.2 Undocumented Variable Values

The repayment status history variables (PAY_*) contained discrepancies between the data dictionary defined values and the values that were found in the dataset. Referring to the data dictionary (see Section 2.1), the attribute guide establishes that the valid category values for these variables are -1; and 1 and above. However, many observations were assigned the category values of -2 or 0. These two values were used frequently to categorize the monthly repayment status activities. The category 0 is assigned to approximately 50% of the observations in each of the six months while the value -2 was assigned to at least 12% of the observations each month. The high-volume use of the two category values suggests that the they are not erroneous but for some reason were not documented in the data dictionary. It is unclear what the two categories may represent.

For the purpose of this study, it was determined that the meanings of the categories -2 or 0 were not relevant. The category -1 indicates that the clients are paying their credit balances duly. This is interpreted as the clients being compliant with the credit card terms and that they are submitting the minimum required payment amounts on a timely basis. It is likely that if the credit card accounts in question were delinquent, they would be scored with a category value of 1 or above to indicate that accounts were delinquent. Based on this, it is estimated that the values -2 and 0 do not indicate that the clients are delinquent and that the two codes are at minimum a derivative of the category -1-paying duly. Accordingly, the observations with the values -2 and 0 were rolled-up into the category -1-paying duly.

The Education and Marriage features also contained values that were not documented in the data dictionary. The values 0,5, and 6 were found in the Education variable; and the value 0 was found in the Marital Status variable. These values were assigned to a category labeled ‘Unknown’.

2.3.3 Erroneous or Inconsistent Variable Values

The repayment status history variables (PAY_*) employ sequential number values not only to track that individuals have missed a payment, but also to track how long the accounts have been in delinquent. Accordingly, it is expected that for each month that an account is outstanding, the assigned values will increase in value. However, it was noted that value one (1), indicating that repayment was one month late, does not appear in any of the 30,000 observations during the five-month period April 2005 through August 2005 (PAY_6 through PAY_2). However, the value 2, which indicates that the repayment status is two months late, appears frequently during these months although the preceding month did not contain the value 1. No action was taken to correct this issue. The precise number of months that the individuals have been in default is not as important to this study as the fact that the individuals have defaulted. Also, to naively reduce the assigned values by one value to ensure that the value 1 is accounted for during the five- month period in question greatly increases the risk of analyst error being introduced into the models.

It was also noted that the repayment status history contained values of 1 or higher in instances in which the credit balances were $0. The instances in which this error occurred were corrected by modifying the existing PAY_* value to a -1-paying duly value.

2.3.4 Outliers

Outliers due to errors can be treated as missing values. Although there are extremely high and low values in the data, it is not clear if the values are erroneous. For the purpose of this analysis, it is assumed that the values are reasonable and valid. To attempt to center the data, the Outliers were significantly reduced by capping them to the upper 95th percentile and capping the lower bounds at the 5^{th} percentile interval. Even within the caps, some outliers persisted. It should be noted that not all outliers are bad for modeling. Some can provide important information. Regardless of the actions taken above to center the data, some observations remained with zero values due to inactivity of the credit accounts.

2.3.5 Statement Balances Exceeding Credit Limits

It is expected that card holders will not have the ability to exceed the credit limits that were assigned to them by the credit card issuer. Nevertheless, several instances were noted in which the bill statement balances exceeding the credit balance limits. As there can be several valid reasons for this to occur (prior-period adjustments, over-spending approvals by the card issuer, credit limits being reduced to a lower amount than existing charges, etc.), no action was taken to modify the statement values that exceed the credit limits.

3. Feature Engineering

New features were created using the domain knowledge of the Default dataset. The raw data of the dataset was transformed into features that better represent the underlying problem of identifying Default Risk. The engineered features were designed to provide insight and a score/weight on the credit card customer's financial stability, character, and ability to repay the credit card debt. It should also be noted that the raw monthly payment (PAY_AMT*) and statement amount (BILL_AMT*) time-series data could not be used for modeling. The use of these variables would significantly reduce the predictive accuracy of the constructed models. The models would learn the weights of the individual months. As there is much variance in the credit card monthly activities for each account, the learned weights would not be applicable to future months. Instead, to better represent Default Risk, these variables were smoothed over the six-month period and their values were normalized by creating features that aggregated the time-series variables into averages and/or ratios. Engineered Features also include variables that were trimmed to account for outliers. The engineered features can be divided into two categories: Financial Indicators and Non-Financial Indicators.

3.1 Financial Indicator Variables

Financial indicator variables that are traditionally used to assess Default Risk are not available in the Default dataset. Such variables include income, debt, debt ratio, employment status, etc. Nevertheless, variables were available that can serve as substitutes for the unavailable financial indicator variables. For instance, the credit limit variable (LIM_BAL) and the monthly payment history variables (PAY_AMT*) are proxies for an individual's income level. Banks award credit limits based on several factors including income level, employment status, capital and debt levels, and creditworthiness scores (i.e. FICO, etc.). Accordingly, it is believed that the awarded credit limit is relative to the individual's income level. The payment history variables represent the customer's ability to pay a certain amount each month. As such, the repayment history variables are also a proxy for income level. Features that quantify averages and ratios, such as average monthly statement amounts, average monthly payments, average monthly payment ratios, average monthly estimated charges, and average utilization represent a smoothed score/weight of the individual's ability to repay the credit card debt.

As a credit card holder's financial circumstance can change without the Bank's knowledge, and their ability to repay the debt may suffer as a result, several variables were engineered to gauge the card holders' immediate financial status. Such variables include the Current Pay Ratio (R_PMT_RATIO1_imp_5.95) and Current Utilization (R_UTIL1_imp_5.95). Variables associated with Utilization Growth (use of available credit) over 6-months, and Balance Growth over 6-months were engineered to gauge an individual's ability to repay the debt over time. These variables can also provide insight and trigger red-flags if a customer is in current financial stress.

3.2 Non-Financial Indicator Variables

The Default dataset contains several variables that are categorical in nature. Accordingly, these variables were converted from the integer to the factor variable type. The variables that were converted to factor type includes the demographic features such as SEX, MARRIAGE, and EDUCATION and the credit card repayment history behavior related features (PAY_*). The variables were then segmented into various levels. For Example, the Marriage Status feature (MARRIAGE) was sectioned into four levels (Married, Single, Others, Unknown). These variables can provide insight on the individual's character and credit worthiness. In addition, these features can be proxies for traditional variables (ex. FICO, Credit Bureau scores, etc.) that measure and quantify credit worthiness. In addition, the levels of the discretized variables serve as a form of segmentation. The segmentations can provide valuable insight on particular demographics and their Default Risk.

3.3 Engineered Feature Preparation

The prefix "R_" was affixed to all engineered variables. A suffix was also appended to each constructed variable based on its criteria. For example, the suffix "_m" was appended to variables that were created to capture modifications or corrections to the original features. The suffix "_bin" was added to variables that were discretized using analyst judgement. The Weight of Evidence (WOE) binning and the One Rule (OneR) binning algorithms were used to define alternative grouped versions of the discretized variables. The suffix "woe_bin" or "OneR_bin" was appended to the features based on these algorithms. Any variables that were constructed based on Decision Tree results were labeled with the suffix "_DT_bin". The response variable was transformed into a categorical type variable and was labeled "R_DEFAULT".

A preliminary Exploratory Data Analysis was conducted during the Data Quality Check. This examination uncovered that the monthly time-series statement history (BILL_AMT*) and payment history (PAY_AMT*) related variables; and the credit limit variable (LIM_BAL) were heavily skewed. For this analysis, it was elected to trim these variables at their bottom 5th and top 95th percentile. The trimmed variables include the suffix "_imp_5.95" to indicate the trim boundaries. This suffix was added to all the numeric engineered variables to flag them as being constructed based on the trimmed variables. Monthly payments (PAY_AMT*) were aligned with their corresponding monthly bill statement amounts (BILL_AMT*) when developing payment ratio variables. Finally, all engineered features that were based on averages/ratios were multiplied by the value 100 so that all of the features would be scaled similarly.

3.4 Feature Selection

Thirty-two variables were selected for modeling and are listed in Table 3.1. Appendix A lists the model variables. The complete list of the engineered variables can be found in Appendix A.

Table 3.1: Model Data

Variable Label	Description
R_DEFAULT	Discretized variable: (0; 1)
LIMIT_BAL_imp_5.95	(Credit) Balance Limit capped at 5% and 95 % quantiles.
R_SEX	Discretized variable: (Male; Female)
R_EDUCATION	Discretized variable: (Graduate; University; High School; Other; Unknown)
R_MARRIAGE	Discretized variable: (Married, Single, Others, Unknown)
AGE1_bin	Discretized variable: ('NA', '18 to 25 years'; '26 to 40 years'; '41 to 65 years'; '66 to 100 years')
AGE2_bin	Discretized variable: ('NA', '20s', '30s', '40s', '50s', '60s', '70 or older')
LIMIT_BAL_bin	Discretized variable: ('NA'; 'Low' [<= 39,999]; 'Low-Med' [40,000 - 139,999]; 'Med' [140,000 - 239,000]; 'Med-High' [240,000 – 359,999]; 'High' [360,000 – 499,999]; 'Very High' [500,000 <=])
R_AVG_BILL_AMT_imp_5.95	Average bill amount (BILL_AMT*) over the six months
R_PMT_RATIO1_2_5.95	(Current) Payment Ratio (PAY_AMT*_imp_5.95/BILL_AMT*_imp_5.95)
R_AVG_PAY_AMT_imp_5.95	Average pay amount (PAY_AMT*) over the six months
R_AVG_PMT_RATIO_imp_5.95	Average Payment Ratio (R_PMT_RATIO*_5.95) over the six months
AVGPAYRATIO_DT_bin	Discretized variable: ('NA'; '<11'; '> 11')
AVGPAYRATIO_OneR_bin	Discretized variable: ('NA'; '< 2.47'; '> 2.47')
R_UTIL1_imp_5.95	(Current) Utilization. Current Balance/Credit Limit (BILL_AMT*_imp_5.95/LIMIT_BAL_imp_5.95)
R_AVG_UTIL_imp_5.95	Average monthly utilization (R_UTIL*_imp_5.95) over the six months
R_BAL_GROWTH_6MO_imp_5.95	Balance Growth Over 6 Months (BILL_AMT61_imp_6.95 - BILL_AMT6_imp_5.95)
R_BAL_GROWTH_MIN_imp_5.95	Minimum Balance Growth Value Over 6 Months
R_GROWTH_6MO_LIMIT_imp_5.95	Growth/Credit Limit Ratio (R_UTIL_GROWTH_6MO_imp_5.95/LIMIT_BAL_imp_5.95)
R_GROWTH_MIN_LIMIT_imp_5.95	Minimum Growth/Credit Limit Ratio (R_UTIL_GROWTH_6MO_imp_5.95/LIMIT_BAL_imp_5.95)
R_UTIL_GROWTH_6MO_imp_5.95	Utilization Growth Over 6 Months
R_UTIL_GROWTH_MIN_imp_5.95	Minimum Utilization Growth Over 6 Months
R_UTIL_GROWTH_6MO_LIMIT_imp_5.95	Utilization Growth Over 6 Months/Credit Limit Ratio
R_UTIL_GROWTH_MIN_LIMIT_imp_5.95	Minimum Utilization Growth Over 6 Months/Credit Limit Ratio
R_MAX_BILL_AMT_imp_5.95	Maximum Bill Amount Over 6 Months
R_MAX_PMT_AMT_imp_5.95	Maximum Payment Amount Over 6 Months
R_MAX_DLQ	Maximum Delinquency (max of the PAY_AMT*_DUM variables)
MAXDLQ_woe_bin	Discretized variable: ('NA'; '= 0'; '> 0')
R_AVG_CHARGE_imp_5.95	Average of Estimated Monthly Charges Over 6 Months
R_AVG_CHARGE_LIMIT_imp_5.95	Average of Estimated Monthly Charges/Credit Limit Ratio Over 6 Months
R_BILL_LIMIT_BAL_VAR_MAX	Maximum Statement Amount Variance to Credit Limit Over 6 Months
BILL_EXCEEDS_LIMIT_BAL_FLAG	Bill Amount Exceeds Credit Limit Flag (0 = No; 1 = Yes)

4. Exploratory Data Analysis (EDA)

4.1 Traditional EDA

4.1.1 Response Variable: Default (R_DEFAULT)

The distribution of the response variable (R_DEFAULT), is materially imbalanced with class 0 (Non-Default) representing 78% of the observations and class 1 (Default) representing the remaining 22% (Table 4.1).

Table 4.1: Response Variable Frequency

	Freq	% Total	% Total Cum.
0 - No Default	23,364	0.7788	0.7788
1- Default	6,636	0.2212	1.0000
Total	30,000	1.0000	1.0000

Some classifiers, like the Logistic Regression method that is being used in this study, find it challenging to discriminate between two classes when their distributions are highly imbalanced. The classifiers tend to be biased in favor of the class that is most represented. When considering the odds, the majority class has the highest probability of being selected than the minority class. However, with the goal of achieving the greatest accuracy, the classifiers might always only predict the majority class. Despite the models maximizing their accuracy and possibly achieving excellent classification accuracy scores in this manner, the models may suffer from Accuracy Paradox. The models may be unable to control Type II Error (false negatives) and Type 1 Error (false positives). As it relates to this study, this means that an individual may default on making credit card repayments, but being influenced by the majority class, the models would predict that the person would not default.

There are several methods that can help overcome the challenges caused by the imbalanced distribution of the classes. However, the solutions are not foolproof and their application may also introduce other challenges and issues to the problem. These approaches, which include Up-Sampling, Down-Sampling, and a hybrid SMOTE Sampling were not implemented in this study. Ultimately it is a question of class separability - how well do the classes separate? An imbalanced class distribution does not mean that the classes are not well separable and that the models will favor the majority class. The models may be able to sufficiently distinguish between the classes to make predictions proportionate to the train data distribution.

4.1.2 Correlation Analysis

The Pearson Correlation Coefficient values were calculated for each of the numeric features selected for modeling against the response variable R_DEFAULT and are displayed in Table 4.2. Figure 4.1 is a bar graph displaying the correlation magnitude of the variables against the response variable R_DEFAULT. It should be noted that the relationship between the predictor variables and the response variable does not necessarily represent causation between them. Nevertheless, the relationship between the variables describes an existing pattern.

Table 4.2: Correlation Matrix

Variable	Correlation
LIMIT_BAL_imp_5.95	-0.16
R_AVG_BILL_AMT_imp_5.95	-0.02
R_AVG_PAY_AMT_imp_5.95	-0.16
R_PMT_RATIO1_2_5.95	-0.005
R_AVG_PMT_RATIO_imp_5.95	-0.01
R_UTIL1_imp_5.95	0.08
R_AVG_UTIL_imp_5.95	0.11
R_BAL_GROWTH_6MO_imp_5.95	-0.03
R_BAL_GROWTH_MIN_imp_5.95	-0.07
R_GROWTH_6MO_LIMIT_imp_5.95	-0.02
R_GROWTH_MIN_LIMIT_imp_5.95	-0.02
R_UTIL_GROWTH_6MO_imp_5.95	-0.02
R_UTIL_GROWTH_MIN_imp_5.95	-0.02
R_UTIL_GROWTH_6MO_LIMIT_imp_5.95	-0.01
R_UTIL_GROWTH_MIN_LIMIT_imp_5.95	0.02
R_MAX_BILL_AMT_imp_5.95	-0.05
R_MAX_PMT_AMT_imp_5.95	-0.15
R_MAX_DLQ	0.38
R_AVG_CHARGE_imp_5.95	-0.10
R_AVG_CHARGE_LIMIT_imp_5.95	-0.01
R_BILL_LIMIT_BAL_VAR_MAX	0.15

The majority of the variables have a Pearson Correlation value close to zero (0), indicating that they have weak to no correlation with the response variable R_DEFAULT. Only four of the 17 numeric variables have a positive correlation with R_Default, with R_MAX_DLQ (Maximum Delinquency) having the highest (but weak) correlation with the response variable with a value of .38. The remaining 13 variables have a negative correlation with R_Default.

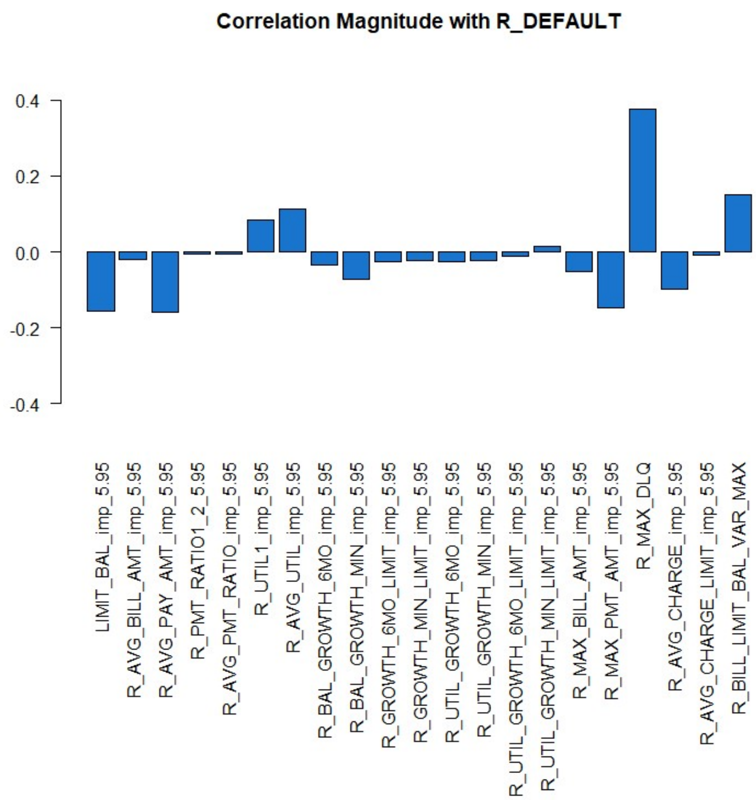


Figure 4.1: Correlation Magnitude Graph

For the most part, the positive and negative relationships between the predictor variables and R_DEFAULT make sense. For example, the positive relationship that R_MAX_DLQ and R_BILL_LIMIT_BAL_VAR_MAX (Maximum amount difference between the credit line limit and the monthly billing statements) have with R_DEFAULT indicates that with unit value increases in each of the two variables, the probability of the credit holder defaulting increases. In both cases, the amount of debt continues to grow, possibly beyond the individual's ability to repay. Due to the negative relationship, as the unit values for the average pay amounts (R_AVG_PAY_AMT_imp_5.95) increases or the credit limit (LIM_BAL_imp_5.95) increases, the probability of the individual defaulting decreases. In the first case, the outstanding amount is lowered reducing the financial stress on the card holder. In the second case, high credit limits indicate that the individuals have high income and/or other financial resources that increases their ability to repay the credit card loans.

To further understand the relationship between the stock variables, a correlation heat map plot (Figure 4.2) matrix was constructed. Each variable in the matrix is correlated with each of the other variables in the matrix. The Correlation plot will allow for the visualization of all pairwise correlations in the data. Positive correlations are displayed in blue and negative correlations are displayed in red color. Color intensity and the size of the circle are proportional to the correlation coefficients. A legend can be found at the bottom of the plot. The legend color shows the correlation coefficients and the corresponding colors. As can be seen per the plot, the variables have positive or negative correlation to varying degrees. The variables associated with Utilization and/or Growth have mid to high correlation with each other.

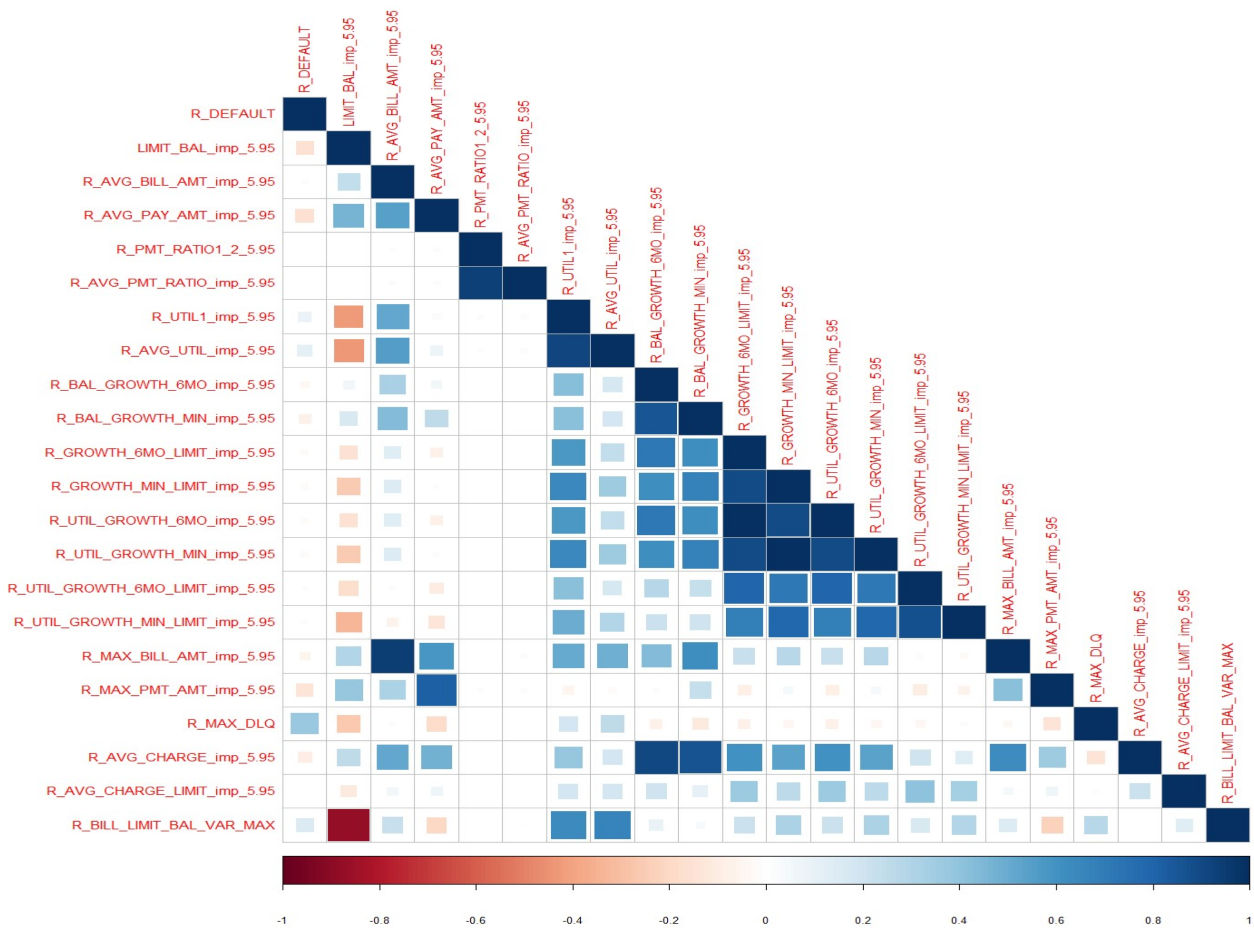


Figure 4.2: Correlation Heat Map

The Information Value metric was calculated for each the categorical variables to gauge whether or not they may be good predictors of Default. Table 4.3 list the Information Values for the variables.

Table 4.3: Categorical Variable Information Value (IV)

VARIABLE	IV SCORE	HOW GOOD
R_SEX	0.0092	Not Predictive
R_EDUCATION	0.0375	Somewhat Predictive
R_MARRIAGE	0.0054	Not Predictive
AGE1_bin	0.0180	Not Predictive
AGE2_bin	0.0089	Not Predictive
LIMIT_BAL_bin	0.1695	Highly Predictive
AVGPAYRATIO_DT_bin	0.1148	Highly Predictive
AVGPAYRATIO_OneR_bin	0.0475	Somewhat Predictive
MAXDLQ_woe_bin	0.6786	Highly Predictive
BILL_EXCEEDS_LIMIT_BAL_FLAG	0.0093	Not Predictive

According to their Information Values, the variables MAXDLQ_woe_bin, AVGPAYRATIO_DT_bin, and LIMIT_BAL_bin have a strong relationship with the response variable R_DEFAULT and may be highly predictive of Default. However, the Information Value for the variable MAXDLQ_woe_bin is suspicious because it is greater than 0.5. The characteristics of this variable may be over-predicting its relationship with R_DEFAULT. The variables R_EDUCATION and AVGPAYRATIO_OneR_bin have a mid-level strength relationship with R_DEFAULT and are somewhat predictive of Default. The remainder of the variables have a weak relationship with R_DEFAULT and are not predictive of Default.

To determine the relative risk of the categorical variable groups/bins with the response variable R_DEFAULT, their Weight of Evidence (WOE) was calculated. WOE measures the strength of the set of bins and separates events from non-events. This is accomplished by dividing the Distribution of "Goods" by the Distribution of "Bads". For this study, "Goods" refers to the individuals that did not default, and the "Bads" refers to the individuals that defaulted. Table 4.4 lists the computed WOE values for each group in the categorical variables. Generally, a negative WOE value indicates that the group may be high risk and a positive WOE value corresponds to low risk of Default. However, due to the manner in which the R software package calculated the WOE values, in this case a positive WOE number indicates high risk and a low number corresponds to low risk. For instance, for the Male group in the R_SEX variable has a relative high risk of Default while the Female group has a low relative risk.

Table 4.4: WOE Table

VARIABLE	CATEGORY	WOE	IV
R_SEX	Female	-0.0798	0.0038
	Male	0.1152	0.0054
R_EDUCATION	Graduate School	-0.1761	0.0104
	High School	0.1685	0.0049
	Other	-1.5490	0.0061
	University	0.0914	0.0040
	Unknown	-1.2484	0.0122
R_MARRIAGE	Married	0.0768	0.0027
	Other	0.0844	0.0001
	Single	-0.0706	0.0026
AGE1_bin	18 to 25 years	0.2467	0.0084
	26 to 40 years	-0.1078	0.0067
	41 to 65 years	0.0979	0.0027
	66 to 100 years	0.2510	0.0002
AGE2_bin	20s	0.0415	0.0006
	30s	-0.1119	0.0045
	40s	0.0489	0.0005
	50s	0.1527	0.0019
	60s	0.3312	0.0013
	70 or older	0.3142	0.0001
LIMIT_BAL_bin	Low	0.6768	0.0734
	Low-Med	0.1975	0.0143
	Med	-0.2692	0.0174
	Med-High	-0.4747	0.0283
	High	-0.5432	0.0203
	Very-High	-0.8111	0.0159
AVGPAYRATIO_DT_bin	< 11	0.2807	0.0465
	> 11	-0.4128	0.0683
AVGPAYRATIO_OneR_bin	< 2.47	0.7457	0.0437
	> 2.47	-0.0639	0.0037
MAXDLQ_woe_bin	= 0	-0.6631	0.2576
	> 0	1.0837	0.4210
BILL_EXCEEDS_LIMIT_BAL_FLAG	No	-0.0153	0.0002
	Yes	0.6053	0.0090

4.1.2 Numeric Variable Summary

Table 4.5 lists the data summaries for the numeric variables.

Table 4.5: Summary Statistics for Default Model Data

	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
R_DEFAULT*	0.42	1	1.15	0	1	2	1	1.34	-0.2	0.002
LIMIT_BAL_imp_5.95	120,832.00	140,000	151,607.00	133,434	20,000	430,000	410,000	0.71	-0.54	697.6
R_SEX*	0.49	1	1.37	0	1	2	1	0.42	-1.82	0.003
R_EDUCATION*	1.39	2	2.64	1.48	1	5	4	-0.06	-1.81	0.01
R_MARRIAGE*	0.99	3	2.1	0	1	3	2	-0.15	-1.96	0.01
AGE1_bin*	0.63	3	3.18	0	2	5	3	-0.05	-0.37	0.004
AGE2_bin*	0.97	3	2.98	1.48	2	7	5	0.66	-0.12	0.01
LIMIT_BAL_bin*	1.26	4	3.68	1.48	2	7	5	0.63	-0.19	0.01
R_AVG_BILL_AMT_imp_5.95	49,053.00	21,060.00	31,280.00	28,377.00	0	180,976.00	180,976.00	1.51	1.34	283.2
R_AVG_PAY_AMT_imp_5.95	3,312.00	2,397.00	2,948.00	2,368.00	0	17,397.00	17,397.00	1.56	2.47	19.12
R_PMT_RATIO1_2_5.95	2,568.00	6.51	22.63	8.17	0	444,433.00	444,433.00	172.8	29,895.00	14.83
R_AVG_PMT_RATIO_imp_5.95	1,647.00	8.41	21.61	7.86	0	266,684.00	266,684.00	149.1	23,369.00	9.51
AVGPAYRATIO_DT_bin*	0.5	2	2.44	0	2	3	1	0.19	-1.96	0.003
AVGPAYRATIO_OneR_bin*	0.25	3	3	0	2	3	1	-3.5	10.25	0.001
R_UTIL1_imp_5.95	39.56	30.95	37.61	44.83	0	645.5	645.5	0.82	2.47	0.23
R_AVG_UTIL_imp_5.95	33.76	27.81	32.65	39.04	0	536.4	536.4	0.72	1.29	0.19
R_BAL_GROWTH_6MO_imp_5.95	34,433.00	1,177	7,868.00	10,896.00	-161,912	201,203.00	363,115.00	1.32	8.47	198.8
R_BAL_GROWTH_MIN_imp_5.95	30,654.00	3,246.00	10,591.00	4,812.00	0	201,203.00	201,203.00	2.99	11.08	177
R_GROWTH_6MO_LIMIT_imp_5.95	28.45	0.87	6.72	11.08	-161.4	413.3	574.7	1.59	7.99	0.16
R_GROWTH_MIN_LIMIT_imp_5.95	25.77	2.64	9.57	3.91	0	413.3	413.3	2.59	11.32	0.15
R_UTIL_GROWTH_6MO_imp_5.95	28.45	0.87	6.72	11.08	-161.4	413.3	574.7	1.59	7.99	0.16
R_UTIL_GROWTH_MIN_imp_5.95	25.77	2.64	9.57	3.91	0	413.3	413.3	2.59	11.32	0.15
R_UTIL_GROWTH_6MO_LIMIT_imp_5.95	0.08	0.0005	0.01	0.01	-0.81	2.07	2.87	4.28	65.01	0.0004
R_UTIL_GROWTH_MIN_LIMIT_imp_5.95	0.07	0.002	0.01	0.003	0	2.07	2.07	6.47	83.56	0.0004
R_MAX_BILL_AMT_imp_5.95	58,406.00	31,209.00	44,647.00	38,743.00	0	201,203.00	201,203.00	1.26	0.51	337.2
R_MAX_PMT_AMT_imp_5.95	6,228.00	5,000	6,801.00	5,031.00	0	19,004.00	19,004.00	0.71	-0.97	35.95
R_MAX_DLQ	1.46	0	0.65	0	0	9	9	1.42	1.48	0.01
MAXDLQ_woe_bin*	0.45	2	2.23	0	2	3	1	0.95	-1.09	0.003
R_AVG_CHARGE_imp_5.95	7,742.00	3,163.00	4,475.00	4,157.00	-29,123.00	54,331.00	83,454.00	1.67	5.49	44.7
R_AVG_CHARGE_LIMIT_imp_5.95	14.44	1.42	2.37	2.37	-149.1	502	651	3.97	76.7	0.08
R_BILL_LIMIT_BAL_VAR_MAX	756,049.00	-456,467.00	-620,561.00	596,185.00	-4,774,000	1,413,016	6,187,016	-1.16	0.74	4,365.00
BILL_EXCEEDS_LIMIT_BAL_FLAG*	0.14	1	1	0	1	2	1	6.64	42.11	0.001

As can be seen in Table 4.5, several of the engineered variables are highly skewed. Of particular concern are the Current Payment Ratio (R_PMT_RATIO1_2_5.95) and the Average Payment Ratio R_AVG_PMT_RATIO_imp_5.95). The skewness value of 1 indicates a normal distribution. These two variables have a skew value greater than 100. The skewness is a result of outliers in the data. The range between minimum and maximum values for many of the variables is also very high. Some variables also have negative minimum values. Negative values could correspond to prior-period adjustments. The skewness of the data, the very wide range values, and the negative values may adversely impact the predictive accuracy of the models. Nevertheless, aside from the outlier trimming that was discussed in section 3.1, additional attempts to further center the data were not made.

4.1.2 Default Probabilities by Segments

Appendix B displays frequency boxplots of the categorical variables by group and default status. The frequency data is useful to evaluate which segment had the highest rate of defaults. However, the information provided by frequency tables and graphs does not necessarily correspond to those that may default in the future. The frequency simply shows which group had the highest levels of defaults in the data sample. To estimate which group might default, the odds probabilities for each segment must be calculated. Table 4.6 lists the default probabilities for each segment. The Male group has a higher probability of defaulting than the Female group. The individuals that fall within the Low credit limit balances segment (LIMIT_BAL_bin LOW) have the highest probability of defaulting based on credit limits. This makes sense as a low credit limit suggests that the individual has low income that may further lead to a low repayment ability if the credit is overused.

Table 4.6: Probability of Default by Category Table

Variable	CATEGORY	PROBABILITY OF DEFAULT
R_SEX	Female	0.2078
	Male	0.2417
R_EDUCATION	Graduate School	0.1923
	High School	0.2516
	Other	0.0569
	University	0.2373
	Unknown	0.0754
R_MARRIAGE	Married	0.2347
	Other	0.2361
	Single	0.2093
AGE1_bin	18 to 25 years	0.2666
	26 to 40 years	0.2032
	41 to 65 years	0.2386
	66 to 100 years	0.2674
AGE2_bin	20s	0.2284
	30s	0.2025
	40s	0.2297
	50s	0.2486
	60s	0.2834
	70 or older	0.2800
LIMIT_BAL_bin	Low	0.3585
	Low-Med	0.2571
	Med	0.1783
	Med-High	0.1502
	High	0.1416
	Very-High	0.1121
AVGPAYRATIO_DT_bin	< 11	0.2733
	> 11	0.1582
AVGPAYRATIO_OneR_bin	< 2.47	0.3745
	> 2.47	0.2104
MAXDLQ_woe_bin	= 0	0.1277
	> 0	0.4564
BILL_EXCEEDS_LIMIT_BAL_FLAG	No	0.2186
	Yes	0.3422

4.2 Model- Based EDA

Four models were constructed for EDA: a naïve Logistic Regression model; a Decision Tree model; a Random Forest model; and a One Rule model. Other than a couple of variables being derived from the EDA models, the models did not provide much more insight beyond what was already obtained from the Traditional EDA. Some of the model-based engineered variables were excluded from the model data because the Logistic Regression model indicated that they contained high collinearity with other variables and did not assign coefficient values to them. The One Rule model selected one variable, R_MAX_DLQ, as having the highest predictive value of default. This model indicated that by using this variable alone, the model would achieve 78.83% overall accuracy in predicting default.

The Decision Tree provided several rules/conditions that will result in an individual to defaulting on their credit card debt. What is noteworthy is that all four models selected the repayment behavior related variables as the most significant. Additionally, these models indicated that the Non-Financial Indicator variables were the least important and have a weak to no value in predicting default risk. Appendix C lists the Logistic Regression EDA results. Appendix D lists the One R model results. Appendix E displays the Decision Tree and its associated rules. It should be noted the Decision Tree represents a complexity that resulted in a graph that could be displayed in print. A more complex Decision Tree generates a Tree graph that is ineligible in print. Finally, Appendix F displays the feature importance values generated by the Random Forest model.

5. Predictive Modeling: Methods and Results

Four models were constructed using the R programming language Caret machine learning package. The four models include 1) a Random Forest model; 2) an Extreme Gradient Boost model; 3) a Logistic Regression model with Stepwise automatic variable selection; and 4) a simple feedforward Neural Network. The default parameters for each model type per the Caret package were used in constructing the models to establish them as base models.

As was explained in section 3.3, the sequential monthly payment and bill statement variables were trimmed to allow for the Logistic Regression and Neural Network models to converge. During test runs, it was discovered that the Neural Network would not converge unless the data was transformed by being centered and scaled. Centering the data is accomplished by subtracting the mean from the values. The scale transformation involves dividing the values by their standard deviation. Using the Caret model parameters, the data was centered and scaled and applied to all four models to ensure that they were all processing the same data and to allow for the Neural Network model to converge.

All four models predicted the odds probabilities of default by each individual. A cut-off classification threshold for each model was established using the Receiver Operating Characteristic (ROC) Curve. If the odds exceeded the cut-off threshold, the observation was classified as 1-Yes (Default). Otherwise, the observation was classified as 0-No (Non-Default).

5.1 Random Forest (RF) Model

Random Forest (RF) is an ensemble machine learning algorithm that consists of a large number of individual decision trees. In classification problems, each individual decision tree in the RF makes class predictions and the algorithm outputs the class with the most occurrences (votes). The accuracy of the Random Forest depends on the strength and dependence of the individual trees. Random Forest is a black box algorithm, understanding the decision process of each individual tree is not possible.

5.1.1 Random Forest Important Features

One of the advantages of RF is that it can be used to determine the importance of the features in explaining the response variable. The RF model estimates the predictive value of the variables and examines how the model performs. Figure 5.1 lists the ranking of the features that were considered as having the highest predictive value by the RF model that was constructed for this study.

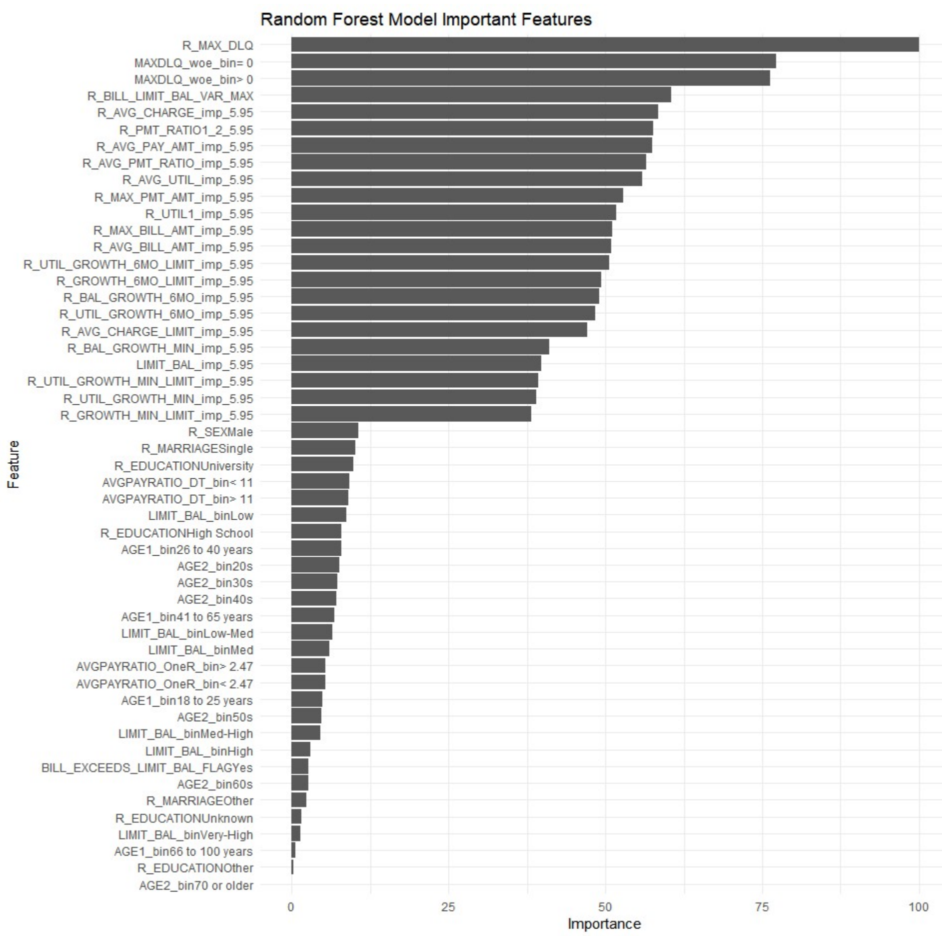


Figure 5.1: Random Forest Model Important Features

The variables considered as the most important in explaining the response variables (R_DEFAULT) by the RF model were the behavioral related variables Maximum Delinquency (R_MAX_DLQ) and Maximum Delinquency greater than or equal to zero (0) (MAXDLQ_woe_bin). Maximum Delinquency achieved a perfect Importance score of 100. After these two variables, the model estimated that many of the Financial Indicator variables have the most predictive value. The model considered the Non-Financial Indicator features as being the least important and having the least impact in making predictions.

5.1.2 Random Forest Predictive Accuracy - In-sample Performance

As can be seen in Table 5.1, the model appears to be achieving good separation of the data as it's classification distribution split of 77% for class 0 and 23% for class 1 nearly matches the Default dataset's actual distribution. The RF model has a fair overall accuracy score of 74.04%. At best, the model is poor at predicting Default Risk. The model correctly predicted 2,077 or 60.68% of the instances of default (True Positives), but misclassified 1,346 or 39.32% of the cases (False Negatives). This means that the model is predicting that nearly 40% of the individuals that will actually default, will not default on their credit card debt. The model is achieving fair performance in predicting non-default cases. The model correctly predicted 9,163 or 77.94% of the instances of non-default (True Negatives), but misclassified 2,594 or 22.06% of the cases (False Positives). This means that the model is predicting that 22% of the individuals that did not default, will default. The mid-range Recall score of 60.68% reflects the correct predictions for class 1 (True Positives) and indicates that the model missed 39.32% of the default cases. The low Precision score of 44.47% signifies that the model misclassified a high number of class 0 cases as class 1. The model AUC score 74.60 (Figure 5.2) indicates that there is a 74.60% chance that the model will be able to distinguish between class 0 (non-default) and class 1 (default).

Model #1: Random Forest													
Actual Class	Predicted Class		Totals	Total %	Actual Class	Predicted Class		TP	0.6068	TP+TN	1.3861	AUC	74.60
	0	1				0	1	TN	0.7794	Precision	0.4447	Sensitivity	0.6068
0	9,163	2,594	11,757	0.7745	0	0.7794	0.2206	Type I Error	0.2206	Recall	0.6068	Specificity	0.7794
1	1,346	2,077	3,423	0.2255	1	0.3932	0.6068	Type II Error	0.3932	F1	0.6641	Accuracy	0.7404

Table 5.1: Confusion matrix and classification metrics for Model #1: Random Forest - In-Sample Data Predictive Accuracy.

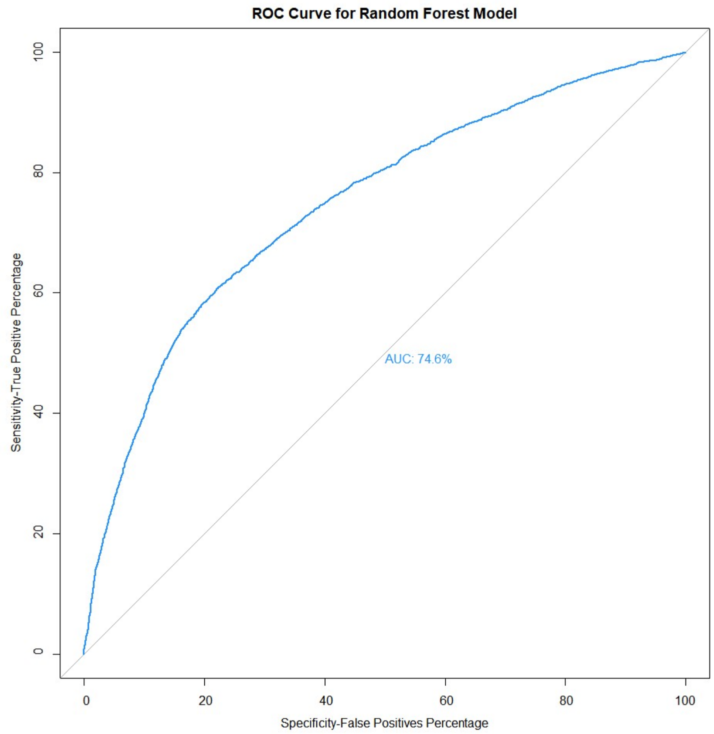


Figure 5.2: Random Forest Train Data ROC Curve

The ROC based classification threshold for in-sample testing is .1590. If the probabilities of Default were greater than .1590, the observation was classified as class 1 – Default. Otherwise, the observation was classified as class 0 – Non-Default.

5.1.3 Random Forest Predictive Accuracy – Out-of-sample Performance

A defining feature of predictive modeling is assessing model performance out-of-sample using data the model has not seen. As can be seen in Table 5.2, using the Test dataset for out-of-sample testing, the model appears to be achieving good separation of the data as it's classification distribution split of 79% for class 0 and 21% for class 1 is not much different than the Default dataset's actual distribution. The RF model has a fair overall accuracy score of .73.07%. At best, the model is poor at predicting Default Risk. The model correctly predicted 1,023 or 65.70% of the instances of default (True Positives), but misclassified 534 or 34.30% of the cases (False Negatives). This means that the model is predicting that 34% of the individuals that will actually default, will not default on their credit card debt. The model is achieving fair performance in predicting non-default cases. The model correctly predicted 4,328 or 75.06% of the instances of non-default (True Negatives), but misclassified 1,438 or 24.94% of the cases (False Positives). This means that the model is predicting that 25% of the individuals that did not default, will default. The mid-range Recall score of 65.70% reflects the correct predictions for class 1 (True Positives) and indicates that the model missed 34.30% of the default cases. The low Precision score of 41.57% signifies that the model misclassified a high number of class 0 cases as class 1. The model AUC score 75.50 (Figure 5.3) indicates that there is a 75.50% chance that the model will be able to distinguish between class 0 (non-default) and class 1 (default).

Model #1: Random Forest Model													
Actual	Predicted Class		Totals	%	Actual	Predicted Class		TP	0.6570	TP+TN	1.41	AUC	75.50
Class	0	1			Class	0	1	TN	0.7506	Precision	0.4157	Sensitivity	0.6570
0	4,328	1,438	5,766	0.7874	0	0.7506	0.2494	Type I Error	0.2494	Recall	0.6570	Specificity	0.7506
1	534	1,023	1,557	0.2126	1	0.3430	0.6570	Type II Error	0.3430	F1	0.6893	Accuracy	0.7307

Table 5.2: Confusion matrix and classification metrics for Model #1: Random Forest - Out-of-Sample Data Predictive Accuracy.

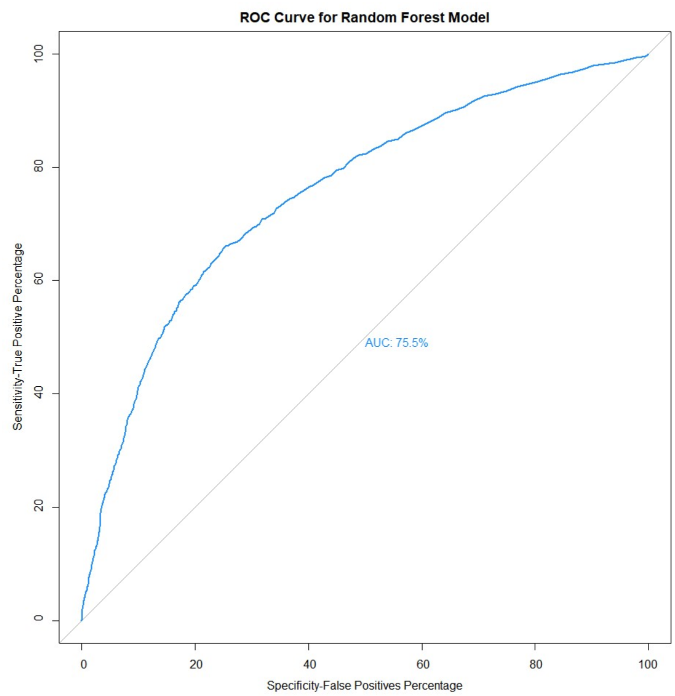


Figure 5.3: Random Forest Test Data ROC Curve

The ROC based classification threshold for out-of-sample testing is .139. If the probabilities of Default were greater than .139, the observation was classified as class 1 – Default. Otherwise, the observation was classified as class 0 – Non-Default.

5.1.4 Random Forest Comparison of In-Sample and Out-of-sample Performance Results

When compared to in-sample performance results, Table 5.3 reflects that the model lost nearly 1% of overall accuracy using the out-of-sample test dataset. However, the model's ability to correctly predict Default Risk improved slightly when testing with the Test dataset. The misclassification rate for class 1 (Default) decreased by 5%. This is also reflected by the AUC score which indicates that the model's chances of predicting class 1 increased by 1%. But there was a trade-off. The model's ability to correctly classify class 0 cases (non-Default) lessened. Misclassification of class 0 increased by nearly 3%.

Table 5.3: Random Forest Model
Metrics Comparison (Train vs Test)

Metric	Train Set	Test Set	Difference
AUC	74.60	75.50	0.90
Class 0 Misclassification	0.2206	0.2494	0.0288
Class 1 Misclassification	0.3932	0.3430	(0.0502)
Accuracy	0.7404	0.7307	(0.0097)

5.2 Extreme Gradient Boosting (XGB) Model

Extreme Gradient Boosting (XGB) is an ensemble machine learning algorithm that consists of a large number of weak prediction models, usually decision trees. XGB is a scalable implementation of gradient boosting. XGB is also a black box algorithm, understanding the decision process of each individual tree is not possible. Although XGB can output a decision tree that optimizes the final model predictions, the decision tree could not be produced for print. The XGB tree that was very complex and in order to fit on the screen, the tree was miniaturized to the point of being illegible.

5.2.1 Extreme Gradient Boosting Important Features

The XGB model estimates the predictive value of the variables and examines how the model performs. Figure 5.4 lists the ranking of the features that were considered as having the highest predictive value by the XGB model that was constructed for this study. The variables considered as the most important in explaining the response variables (R_DEFAULT) by the XGBoost model are the behavioral related variables Maximum Delinquency (R_MAX_DLQ) and Maximum Delinquency greater than or equal to zero (0) (MAXDLQ_woe_bin). Maximum Delinquency achieved a perfect Importance score of 100. After these two variables, the model estimated that many of the Financial Indicator variables have the most predictive value. The model considered the Non-Financial Indicator features as being the least important and having the least impact in making predictions.

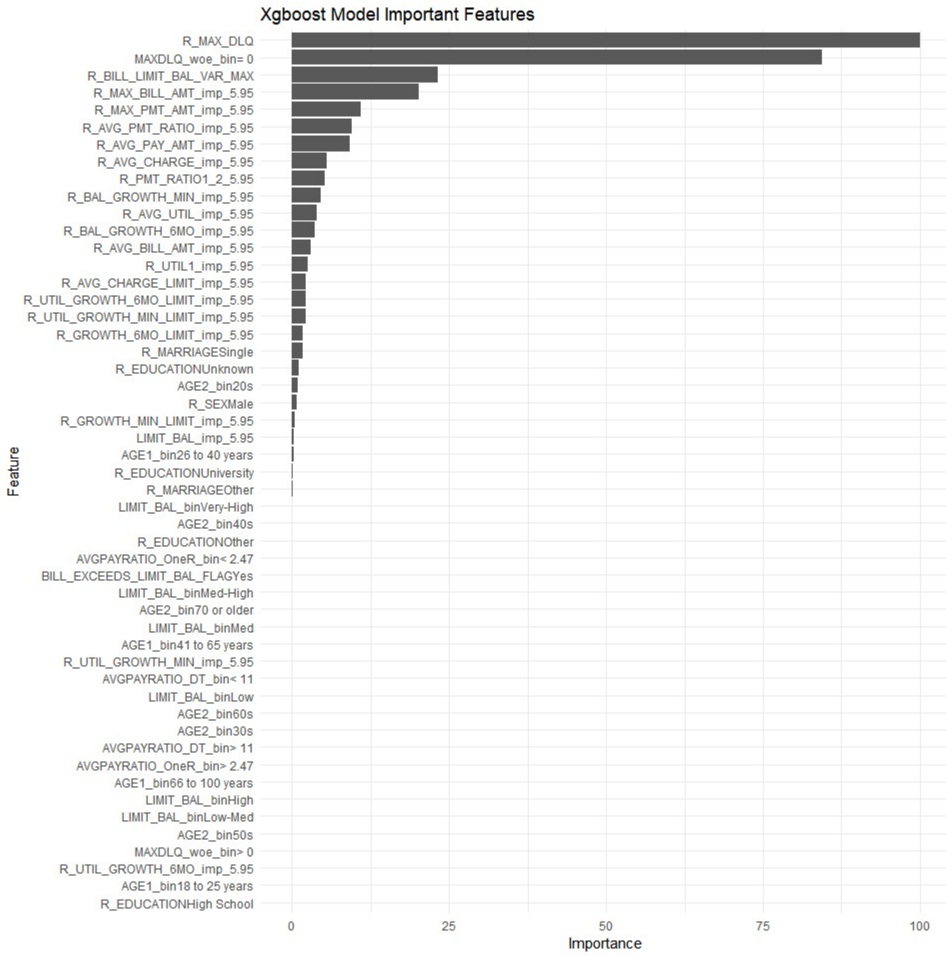


Figure 5.4: XGBoost Model Important Features

5.2.2 Extreme Gradient Boosting Predictive Accuracy - In-sample Performance

As can be seen in Table 5.4, the model appears to be achieving good separation of the data as it's classification distribution split of 77% for class 0 and 23% for class 1 nearly matches the Default dataset's actual distribution. The XGB model has a fair overall accuracy score of 74.94%. At best, the model is poor at predicting Default Risk. The model correctly predicted 2,282 or 66.67% of the instances of default (True Positives), but misclassified 1,141 or 33.33% of the cases (False Negatives). This means that the model is predicting that 33% of the individuals that will actually default, will not default on their credit card debt. The model is achieving fair performance in predicting non-default cases. The model correctly predicted 9,094 or 77.35% of the instances of non-default (True Negatives), but misclassified 2,663 or 22.65% of the cases (False Positives). This means that the model is predicting that 23% of the individuals that did not default, will default. The mid-range Recall score of 66.67% reflects the correct predictions for class 1 (True Positives) and indicates that the model missed 33.33% of the default cases. The low Precision score of 46.15% signifies that the model misclassified a high number of class 0 cases as class 1. The model AUC score 79.00 (Figure 5.5) indicates that there is a 79% chance that the model will be able to distinguish between class 0 (non-default) and class 1 (default).

Model #2: XGBoost Model														
Actual Class	Predicted Class		Totals	Total %	Actual Class	Predicted Class		TP	0.6667	TP+TN	1.4402	AUC	79.00
	0	1				0	1	TN	0.7735	Precision	0.4615	Sensitivity	0.6667
0	9,094	2,663	11,757	0.7745	0	0.7735	0.2265	Type I Error	0.2265	Recall	0.6667	Specificity	0.7735
1	1,141	2,282	3,423	0.2255	1	0.3333	0.6667	Type II Error	0.3333	F1	0.7043	Accuracy	0.7494

Table 5.4: Confusion matrix and classification metrics for Model #2: XGBoost - In-Sample Data Predictive Accuracy.

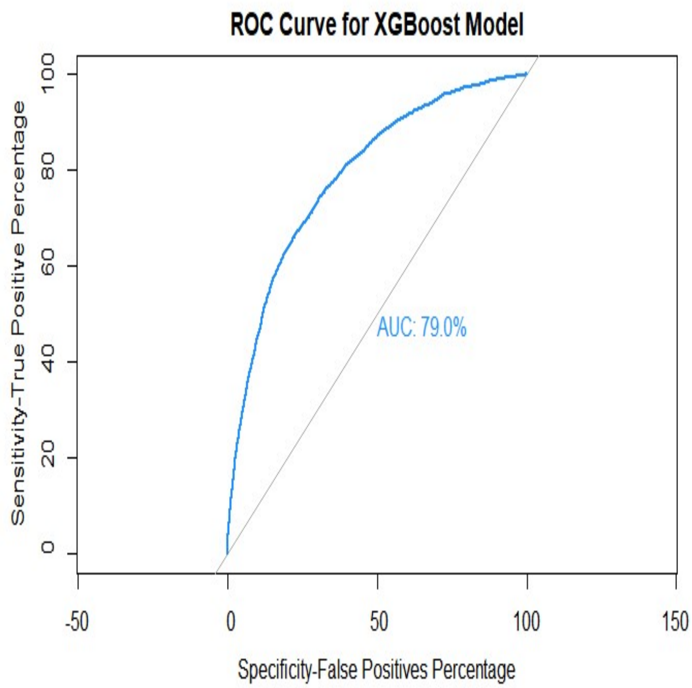


Figure 5.5: XGBoost Train Data ROC Curve

The ROC based classification threshold for in-sample testing is .2392. If the probabilities of Default were greater than .2392, the observation was classified as class 1 – Default. Otherwise, the observation was classified as class 0 – Non-Default.

5.2.3 Extreme Gradient Boosting Accuracy – Out-of-sample Performance

As can be seen in Table 5.5, using the Test dataset for out-of-sample testing, the model appears to be achieving good separation of the data as it's classification distribution split of 79% for class 0 and 21% for class 1 is not much different than the Default dataset's actual distribution. The XGB model has a fair overall accuracy score of 76.70%. The model is poor at predicting Default Risk. The model correctly predicted 956 or 61.40% of the instances of default (True Positives), but misclassified 601 or 38.60% of the cases (False Negatives). This means that the model is predicting that 39% of the individuals that will actually default, will not default on their credit card debt. The model is achieving good performance in predicting non-default cases. The model correctly predicted 4,661 or 80.84% of the instances of non-default (True Negatives), but misclassified 1,105 or 19.16% of the cases (False Positives). This means that the model is predicting that 19% of the individuals that did not default, will default. The mid-range Recall score of 61.40% reflects the correct predictions for class 1 (True Positives) and indicates that the model missed 38.60% of the default cases. The low Precision score of 46.39% signifies that the model misclassified a high number of class 0 cases as class 1. The model AUC score 77.10 (Figure 5.6) indicates that there is a 77.10% chance that the model will be able to distinguish between class 0 (non-default) and class 1 (default).

Model #2: XGBoost Model														
Actual Class	Predicted Class		Totals	%	Actual Class	Predicted Class		TP	0.6140	TP+TN	1.42	AUC	77.10
	0	1				0	1	TN	0.8084	Precision	0.4639	Sensitivity	0.6140
0	4,661	1,105	5,766	0.7874	0	0.8084	0.1916	Type I Error	0.1916	Recall	0.6140	Specificity	0.8084
1	601	956	1,557	0.2126	1	0.3860	0.6140	Type II Error	0.3860	F1	0.6801	Accuracy	0.7670

Table 5.5: Confusion matrix and classification metrics for Model #2: XGBoost - Out-of-Sample Data Predictive Accuracy.

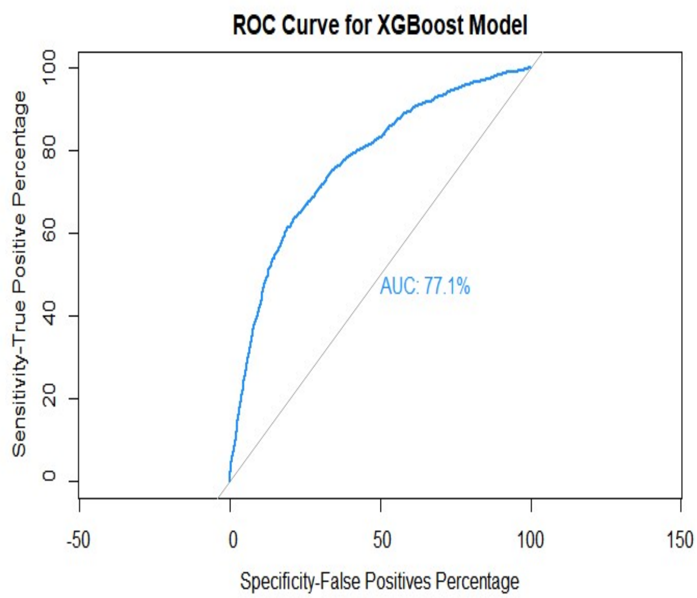


Figure 5.6: XGBoost Test Data ROC Curve

The ROC based classification threshold for out-of-sample testing is .2910. If the probabilities of Default were greater than .2910, the observation was classified as class 1 – Default. Otherwise, the observation was classified as class 0 – Non-Default.

5.2.4 Extreme Gradient Boosting Comparison of In-Sample and Out-of-sample Performance Results

When compared to in-sample performance results, Table 5.6 reflects that the model increased by nearly 2% in overall accuracy (from 74.94% to 76.70%) using the out-of-sample test dataset when compared to the in-sample performance results. Typically, the Test data accuracy is less than of the Train data accuracy. Higher train accuracy may indicate that the model is over-fitting the data. K-fold cross-validation may solve overfitting. The model's ability to correctly predict Default Risk deteriorated slightly when fitting never before seen data. The misclassification rate for class 1 (Default) increased by 5%. This is also reflected by the AUC score which indicates that the model's chances of predicting class 1 decreased by nearly 2%. The model's ability to correctly classify class 0 cases (non-Default) improved. Misclassification of class 0 decreased by nearly 3.5%.

Table 5.6: XGBoost Model
Model Metrics Comparison (Train vs Test)

Metric	Train Set	Test Set	Difference
AUC	79.00	77.10	(1.90)
Class 0 Misclassification	0.2265	0.1916	(0.0349)
Class 1 Misclassification	0.3333	0.3860	0.0527
Accuracy	0.7494	0.7670	0.0176

5.3 Logistic Regression (LR) with Variable Selection Model

Logistic Regression examines the relationship between a categorical dependent variable and the independent variables and produces a probabilistic method of classification. The Logistic Regression (LR) Model was constructed using a Stepwise Automatic Feature Selection method. The automated variable selection procedure was implemented to allow for the software to identify variables that had potential to add to the predictive value of the model. The resulting constructed model was:

R_DEFAULT ~ LIMIT_BAL_imp_5.95 + R_SEXMale + R_EDUCATIONOther + R_EDUCATIONUnknown + R_MARRIAGEOther + R_MARRIAGESingle + AGE2_bin20s + LIMIT_BAL_binLow + LIMIT_BAL_binLow-Med + LIMIT_BAL_binMed + LIMIT_BAL_binMed-High + LIMIT_BAL_binHigh + R_AVG_BILL_AMT_imp_5.95 + R_AVG_PAY_AMT_imp_5.95 + AVGPAYRATIO_OneR_bin< 2.47 + R_BAL_GROWTH_6MO_imp_5.95 + R_MAX_DLQ + MAXDLQ_woe_bin= 0 + R_AVG_CHARGE_imp_5.95 + R_BILL_LIMIT_BAL_VAR_MAX

MODEL # 101: CREDIT CARD DEFAULT MODEL

The model statistics are listed in Table 5.7.

Table 5.7: Logistic Regression Model

	Estimate	Pr(>|z|)	
(Intercept)	-1.4638	0.0000	***
LIMIT_BAL_imp_5.95	0.8707	0.0000	***
R_SEXMale	0.0569	0.0080	**
R_EDUCATIONOther	-0.0447	0.1490	
R_EDUCATIONUnknown	-0.1013	0.0007	***
R_MARRIAGEOther	-0.3942	0.0670	.
R_MARRIAGESingle	-0.0857	0.0004	***
AGE2_bin20s	-0.0457	0.0614	.
LIMIT_BAL_binLow	-0.2034	0.1152	
`LIMIT_BAL_binLow-Med`	-0.3500	0.0310	*
LIMIT_BAL_binMed	-0.3321	0.0043	**
`LIMIT_BAL_binMed-High`	-0.2779	0.0002	***
LIMIT_BAL_binHigh	-0.1543	0.0362	**
R_AVG_BILL_AMT_imp_5.95	-0.2408	0.0004	***
R_AVG_PAY_AMT_imp_5.95	-0.6258	0.0000	***
`AVGPAYRATIO_OneR_bin< 2.47`	0.1774	< 2e-16	***
R_BAL_GROWTH_6MO_IMP_5.95	-0.4619	0.0812	.
R_MAX_DLQ	0.5890	0.0000	***
MAXDLQ_woe_bin= 0'	-0.1238	0.0698	.
R_AVG_CHARGE_imp_5.95	0.4777	0.1128	
R_BILL_LIMIT_BAL_VAR_MAX	1.0777	0.0000	***

Table 5.7 lists the model coefficients and the log odds that each variable predicting the response variable R_DEFAULT. For instance, for one unit increase of the Maximum Delinquency (R_MAX_DLQ), the chances of the card holder defaulting will increase by .5890. However, there are variables that have negative coefficients such as for individuals in their 20's (AGE_vin20s). This suggests that with every customer that is in their 20s decreases the odds of default by .0457.

5.3.1 Logistic Regression Important Features

The p values (Pr(<|z|) listed in Table 5.7 indicates whether or not the variable is important in predicting the default class. The variable's significance levels are denoted by the star symbols ("***"; "**"; "*") with three stars ("***") being the most significant. The LR model has also determined that the behavior related R_MAX_DLQ is an important variable. But unlike the RF and XGB models, the LR model has included Non-Financial Indicator variables (such as R_EDUCATION and R_MARRIAGE) as important variables. Nevertheless, the majority of the important variables are Financial Indicator variables.

5.3.2 Logistic Regression Predictive Accuracy - In-sample Performance

As can be seen in Table 5.8, the model appears to be achieving good separation of the data as it's classification distribution split of 77% for class 0 and 23% for class 1 nearly matches the Default dataset's actual distribution. The LR model has a fair overall accuracy score of 73.41%. At best, the model is poor at predicting Default Risk. The model correctly predicted 2,228 or 65.09% of the instances of default (True Positives), but misclassified 1,195 or 34.91% of the cases (False Negatives). This means that the model is predicting that 35% of the individuals that will actually default, will not default on their credit card debt. The model is achieving fair performance in predicting non-default cases. The model correctly predicted 8,916 or 75.84% of the instances of non-default (True Negatives), but misclassified 2,841 or 24.16% of the cases (False Positives). This means that the model is predicting that 24% of the individuals that did not default, will default. The mid-range Recall score of 65.09% reflects the correct predictions for class 1 (True Positives) and indicates that the model missed 34.91% of the default cases. The low Precision score of 43.95% signifies that the model misclassified a high number of class 0 cases as class 1. The model AUC score 75.90 (Figure 5.7) indicates that there is a 76% chance that the model will be able to distinguish between class 0 (non-default) and class 1 (default).

Model #3: Logistic Regression Model													
Actual Class	Predicted Class		Totals	Total %	Actual Class	Predicted Class		TP	0.6509	TP+TN	1.4092	AUC	75.90
	0	1				0	1	TN	0.7584	Precision	0.4395	Sensitivity	0.6509
0	8,916	2,841	11,757	0.7745	0	0.7584	0.2416	Type I Error	0.2416	Recall	0.6509	Specificity	0.7584
1	1,195	2,228	3,423	0.2255	1	0.3491	0.6509	Type II Error	0.3491	F1	0.6879	Accuracy	0.7341

Table 5.8: Confusion matrix and classification metrics for Model #3: Logistic Regression - In-Sample Data Predictive Accuracy.

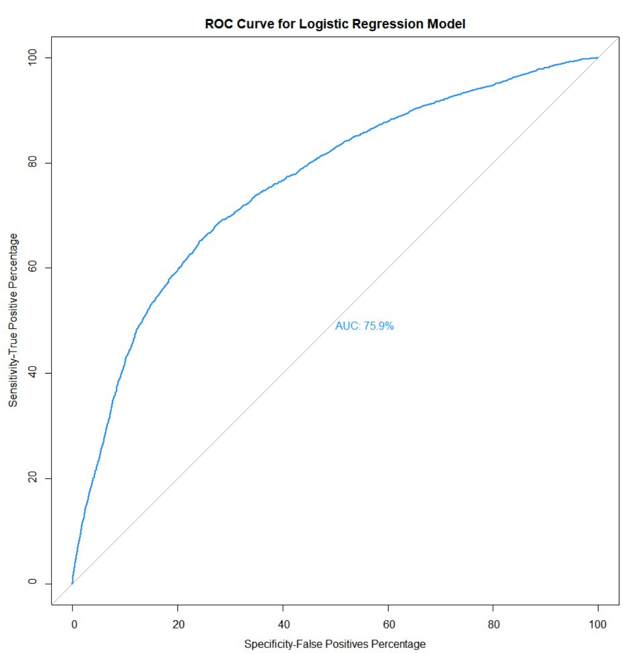


Figure 5.7: Logistic Regression Train Data ROC Curve

The ROC based classification threshold for in-sample testing is .2133. If the probabilities of Default were greater than .2133, the observation was classified as class 1 – Default. Otherwise, the observation was classified as class 0 – Non-Default.

5.3.3 Logistic Regression Accuracy – Out-of-sample Performance

As can be seen in Table 5.9, using the Test dataset for out-of-sample testing, the model appears to be achieving good separation of the data as it's classification distribution split of 79% for class 0 and 21% for class 1 is not much different than the Default dataset's actual distribution. The LR model has a fair overall accuracy score of 72.05%. The model is poor at predicting Default Risk. The model correctly predicted 1,060 or 68.08% of the instances of default (True Positives), but misclassified 497 or 31.92% of the cases (False Negatives). This means that the model is predicting that 32% of the individuals that will actually default, will not default on their credit card debt. The model is achieving fair performance in predicting non-default cases. The model correctly predicted 4,216 or 73.12% of the instances of non-default (True Negatives), but misclassified 1,550 or 26.88% of the cases (False Positives). This means that the model is predicting that 27% of the individuals that did not default, will default. The mid-range Recall score of 68.08% reflects the correct predictions for class 1 (True Positives) and indicates that the model missed 31.92% of the default cases. The low Precision score of 40.61% signifies that the model misclassified a high number of class 0 cases as class 1. The model AUC score 75.80 (Figure 5.8) indicates that there is a 75.80% chance that the model will be able to distinguish between class 0 (non-default) and class 1 (default).

Model #3: Logistic Regression Model													
Actual Class	Predicted Class		Totals	%	Actual Class	Predicted Class		TP	0.6808	TP+TN	1.41	AUC	75.80
	0	1				0	1	TN	0.7312	Precision	0.4061	Sensitivity	0.6808
0	4,216	1,550	5,766	0.7874	0	0.7312	0.2688	Type I Error	0.2688	Recall	0.6808	Specificity	0.7312
1	497	1,060	1,557	0.2126	1	0.3192	0.6808	Type II Error	0.3192	F1	0.6984	Accuracy	0.7205

Table 5.9: Confusion matrix and classification metrics for Model #3: Logistic Regression - Out-of-Sample Data Predictive Accuracy.

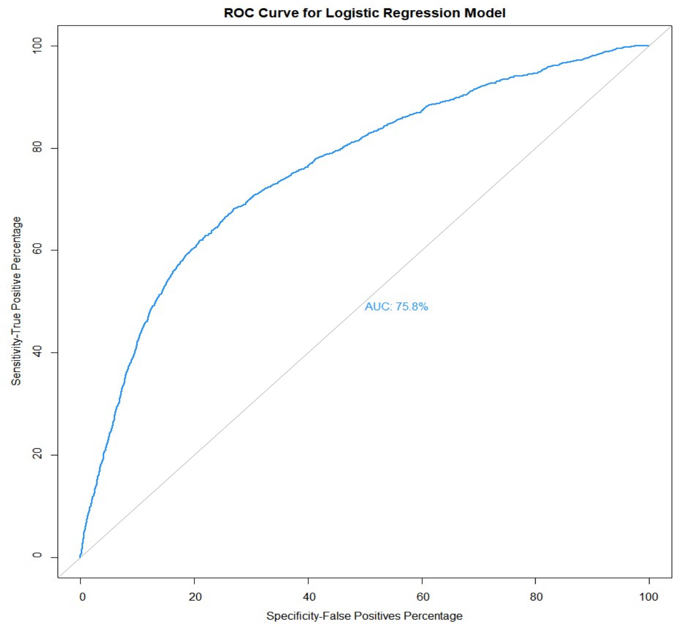


Figure 5.8: Logistic Regression Test Data ROC Curve

The ROC based classification threshold for out-of-sample testing is .1981. If the probabilities of Default were greater than .1981, the observation was classified as class 1 – Default. Otherwise, the observation was classified as class 0 – Non-Default.

5.3.4 Logistic Regression Comparison of In-Sample and Out-of-sample Performance Results

When compared to the in-sample performance results, Table 5.10 reflects that the model overall accuracy decreased by just over 1% (from 73.41% to 72.05%) using the out-of-sample test dataset. The model's ability to correctly predict Default Risk deteriorated slightly when fitting never before seen data. The misclassification rate for class 1 (Default) increased by 3%. The model's ability to correctly classify class 0 cases (non-Default) improved. Misclassification of class 0 decreased by nearly 3%.

Table 5.10: Logistic Regression Model
Model Metrics Comparison (Train vs Test)

Metric	Train Set	Test Set	Difference
AUC	75.90	75.80	(0.10)
Class 0 Misclassification	0.2416	0.2688	0.0272
Class 1 Misclassification	0.3491	0.3192	(0.0299)
Accuracy	0.7341	0.7205	(0.0137)

5.4 Neural Network (NN) Model

Artificial Neural Networks are a set of algorithms that are loosely modeled after the human brain and are used to develop relationships between the input and the output. Figure 5.9 displays the architecture of the Neural Network. The NN has 51 input nodes, five hidden notes, two bias activation input nodes, and one output node for the negative Default class. It should be noted that the discretized variable bins were treated as individual inputs. Being a feedforward model, the bias nodes were added by the algorithm to help them learn patterns. The bias nodes always produce the constant value 1 and are not connected to the preceding layers.

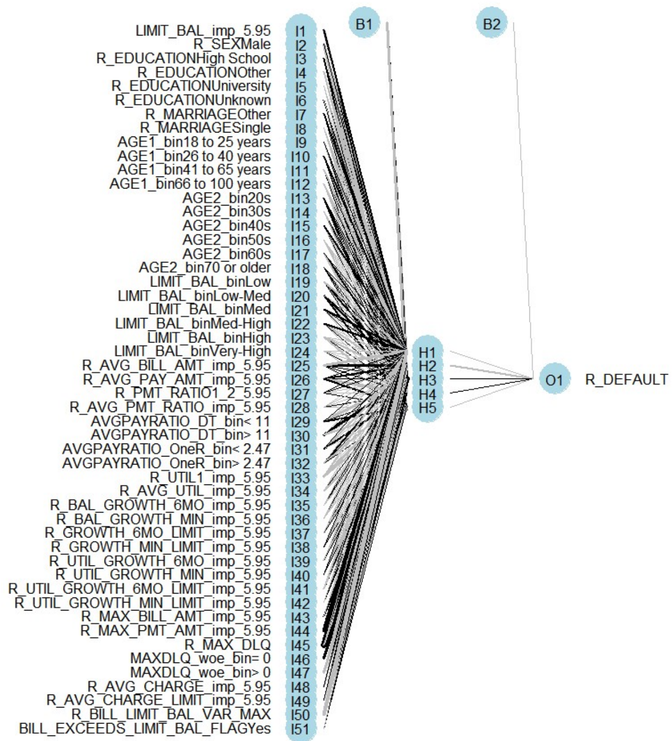


Figure 5.9: Neural Network Architecture

5.4.1 Neural Network Important Features

The Neural Network can be used to determine the importance of the features in explaining the response variable. The NN model estimates the predictive value of the variables and examines how the model performs. Figure 5.10 lists the ranking of the features that were considered as having the highest predictive value by the NN model. The variables considered as the most important in explaining the response variables (R_DEFAULT) by the NN model are the behavioral related variables Maximum Delinquency (R_MAX_DLQ) and Maximum Delinquency greater than or equal to zero (0) (MAXDLQ_woe_bin). Maximum Delinquency achieved a perfect Importance score of 100. After these two variables, the model estimated that many of the Financial Indicator variables have the most predictive value. The model considered the Non-Financial Indicator features as being the least important and having the least impact in making predictions.

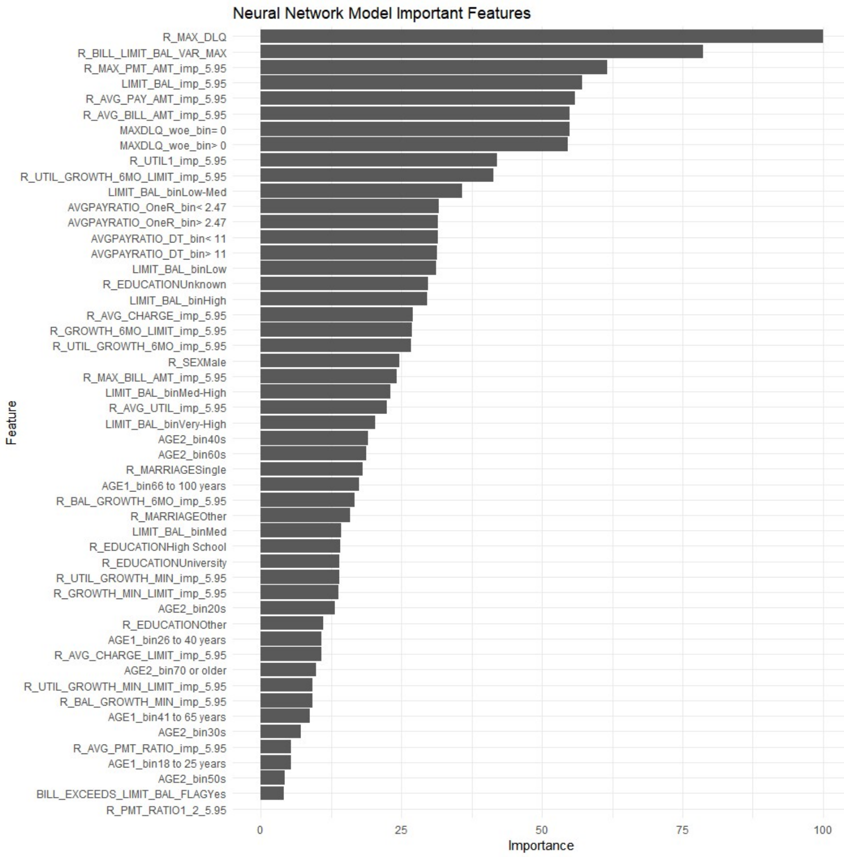


Figure 5.10: Neural Network Model Important Features

5.4.2 Neural Network Predictive Accuracy - In-sample Performance

As can be seen in Table 5.11, the model appears to be achieving good separation of the data as it's classification distribution split of 77% for class 0 and 23% for class 1 nearly matches the Default dataset's actual distribution. The XGB model has a fair overall accuracy score of 73.93%. At best, the model is poor at predicting Default Risk. The model correctly predicted 2,365 or 69.09% of the instances of default (True Positives), but misclassified 1,058 or 30.91% of the cases (False Negatives). This means that the model is predicting that 31% of the individuals that will actually default, will not default on their credit card debt. The model is achieving fair performance in predicting non-default cases. The model correctly predicted 8,857 or 75.33% of the instances of non-default (True Negatives), but misclassified 2,900 or 24.67% of the cases (False Positives). This means that the model is predicting that 24% of the individuals that did not default, will default. The mid-range Recall score of 69.09% reflects the correct predictions for class 1 (True Positives) and indicates that the model missed 30.91% of the default cases. The low Precision score of 44.92% signifies that the model misclassified a high number of class 0 cases as class 1. The model AUC score 78.70 (Figure 5.10) indicates that there is a 78.70% chance that the model will be able to distinguish between class 0 (non-default) and class 1 (default).

Model #4: Neural Network Model													
Actual Class	Predicted Class		Totals	Total %	Actual Class	Predicted Class		TP	0.6909	TP+TN	1.44	AUC	78.70
	0	1				0	1	TN	0.7533	Precision	0.4492	Sensitivity	0.6909
0	8,857	2,900	11,757	0.7745	0	0.7533	0.2467	Type I Error	0.2467	Recall	0.6909	Specificity	0.7533
1	1,058	2,365	3,423	0.2255	1	0.3091	0.6909	Type II Error	0.3091	F1	0.7132	Accuracy	0.7393

Table 5.11: Confusion matrix and classification metrics for Model #4: Neural Network - In-Sample Data Predictive Accuracy.

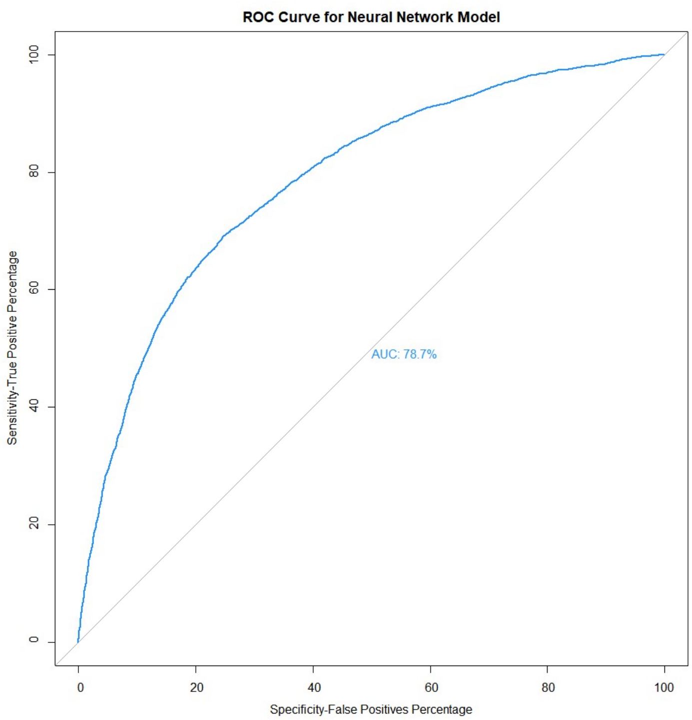


Figure 5.11: Neural Network Train Data ROC Curve

The ROC based classification threshold for in-sample testing is .2185. If the probabilities of Default were greater than .2185, the observation was classified as class 1 – Default. Otherwise, the observation was classified as class 0 – Non-Default.

5.4.3 Neural Network Accuracy – Out-of-sample Performance

As can be seen in Table 5.12, using the Test dataset for out-of-sample testing, the model appears to be achieving good separation of the data as it's classification distribution split of 79% for class 0 and 21% for class 1 is not much different than the Default dataset's actual distribution. The NN model has a fair overall accuracy score of 74.83%. The model is poor at predicting Default Risk. The model correctly predicted 981 or 63.01% of the instances of default (True Positives), but misclassified 576 or 36.99% of the cases (False Negatives). This means that the model is predicting that 37% of the individuals that will actually default, will not default on their credit card debt. The model is achieving fair performance in predicting non-default cases. The model correctly predicted 4,499 or 78.03% of the instances of non-default (True Negatives), but misclassified 1,217 or 21.97% of the cases (False Positives). This means that the model is predicting that 22% of the individuals that did not default, will default. The mid-range Recall score of 63.01% reflects the correct predictions for class 1 (True Positives) and indicates that the model missed 36.99% of the default cases. The low Precision score of 43.64% signifies that the model misclassified a high number of class 0 cases as class 1. The model AUC score 76.20 (Figure 5.11) indicates that there is a 76.20% chance that the model will be able to distinguish between class 0 (non-default) and class 1 (default).

Model #4: Neural Network Model													
Actual Class	Predicted Class		Totals	%	Actual Class	Predicted Class		TP	0.6301	TP+TN	1.41	AUC	76.20
	0	1				0	1	TN	0.7803	Precision	0.4364	Sensitivity	0.6301
0	4,499	1,267	5,766	0.7874	0	0.7803	0.2197	Type I Error	0.2197	Recall	0.6301	Specificity	0.7803
1	576	981	1,557	0.2126	1	0.3699	0.6301	Type II Error	0.3699	F1	0.6812	Accuracy	0.7483

Table 5.12: Confusion matrix and classification metrics for Model #4: Logistic Regression - Out-of-Sample Data Predictive Accuracy.

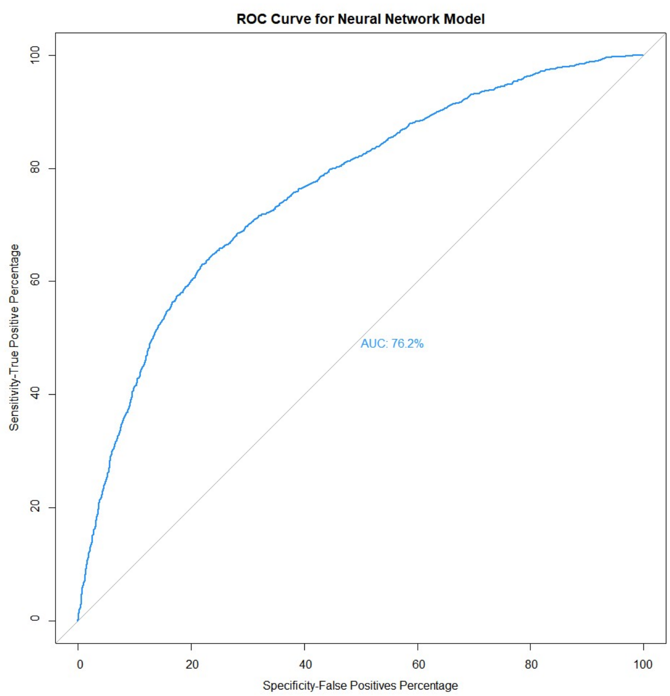


Figure 5.12: Neural Network Test Data ROC Curve

The ROC based classification threshold for out-of-sample testing is .2424. If the probabilities of Default were greater than .2424, the observation was classified as class 1 – Default. Otherwise, the observation was classified as class 0 – Non-Default.

5.4.4 Neural Network Comparison of In-Sample and Out-of-sample Performance Results

When compared to the in-sample performance results, Table 5.13 reflects that the model overall accuracy increased by nearly 1% using the out-of-sample test dataset. This suggests that the model was slightly overfitting the data. The model's ability to correctly predict Default Risk deteriorated when fitting never before seen data. The misclassification rate for class 1 (Default) increased by 6%. The model's ability to correctly classify class 0 cases (non-Default) improved. Misclassification of class 0 decreased by nearly 3%.

Table 5.13: Neural Network Model
Model Metrics Comparison (Train vs Test)

Metric	Train Set	Test Set	Difference
AUC	78.70	76.20	(2.50)
Class 0 Misclassification	0.2467	0.2197	(0.0270)
Class 1 Misclassification	0.3091	0.3699	0.0608
Accuracy	0.7393	0.7483	0.0091

6. Comparison of Results

Accuracy appears to be a very popular index to measure the performance of classification models. But one of its major weaknesses is that this index only works well if the distribution of the binary class is balanced, which is not the case with the Default dataset. The Accuracy rate might be more or less skewed by the number of positive classes in the sample. For instance, if 98% of the samples have a positive classification, the Accuracy score will calculate at 98%. Another limitation of the Accuracy score is that it doesn't provide a complete picture of how the model performed overall. The Accuracy score only measures how accurately the model predicted positive class samples. It does not measure how many predications were misclassified (False Positives and False Negatives). It would be impossible to measure how well the model can distinguish between two classes solely using the Accuracy score. The real danger is the misuse of the Accuracy score as it may give a false sense that the model is achieving high accuracy in distinguishing between the binary classes. Similarly, the Area Under the Curve (AUC) ROC Curve score will suffer from the same limitations as the Accuracy score because it is also not sensitive to an imbalanced dataset. Tables 6.1 and 6.2 illustrate the limitations of using the Accuracy and AUC scores to gauge the performance of the models that were constructed with the Default dataset.

Table 6.1: In-Sample and Out-of-Sample Performance Results Metrics

Model	Sample	Accuracy	AUC	Sensitivity	Specificity	Precision	Recall	F1	Type I Error	Type II Error
RF	Train	0.7404	74.60	0.6068	0.7794	0.4447	0.6068	0.6641	0.2206	0.3932
XGB	Train	0.7494	79.00	0.6670	0.7735	0.4615	0.6667	0.7043	0.2265	0.3333
LR	Train	0.7341	75.90	0.6509	0.7584	0.4395	0.6509	0.6879	0.2416	0.3491
NN	Train	0.7393	78.70	0.6909	0.7533	0.4492	0.6909	0.7132	0.2467	0.3091
RF	Test	0.7307	75.50	0.6570	0.7506	0.4157	0.6570	0.6893	0.2494	0.3430
XGB	Test	0.7670	77.10	0.6140	0.8084	0.4639	0.6140	0.6801	0.1916	0.3860
LR	Test	0.7205	75.80	0.6808	0.7312	0.4061	0.6808	0.6984	0.2688	0.3192
NN	Test	0.7483	76.20	0.6301	0.7803	0.4364	0.6301	0.6812	0.2197	0.3699

Table 6.2: Ranking of In-Sample and Out-of-Sample Performance Results

Model	Sample	Accuracy	AUC	Sensitivity	Specificity	Precision	Recall	F1	Type I Error	Type II Error
RF	Train	2nd	4th	4th	1st	3rd	4th	4th	1st	4th
XGB	Train	1st	1st	2nd	2nd	1st	2nd	2nd	2nd	2nd
LR	Train	4th	3rd	3rd	3rd	4th	3rd	3rd	3rd	3rd
NN	Train	3rd	2nd	1st	4th	2nd	1st	1st	4th	1st
RF	Test	3rd	4th	2nd	3rd	3rd	2nd	2nd	3rd	2nd
XGB	Test	1st	1st	4th	1st	1st	4th	4th	1st	4th
LR	Test	4th	3rd	1st	4th	4th	1st	1st	4th	1st
NN	Test	2nd	2nd	3rd	2nd	2nd	3rd	3rd	2nd	3rd

Table 6.1 lists the performance metrics of the four models using both the in-sample dataset (Train) and the out-of-sample dataset (Test). Table 6.2 ranks the model's performance under each metric. As can be seen, the XGB model achieved the highest Accuracy and AUC scores with both the Train and Test datasets. However, upon further analysis of how well the model distinguished between the Default and Non-Default classes, it becomes apparent that the model did not perform as might be expected based on the higher Accuracy and AUC scores. Based on the lowest Type I or Type II Error scores, when fitting the model using the in-sample data, the XGB model ranked 2nd place among the four models in predicting the two classes. But when using the out-of-sample data, data that the model had not seen before, the model ranked in 1st place in predicting Non-Default cases - but it scored last among the four models in predicting the Default class. Despite the model achieving the highest scores for both the Accuracy and AUC metrics, the XGB model did not perform as well as the other models in predicting Default. Based on the above, the Accuracy and AUC metrics are not ideal to measure the performance of the models that were constructed for this study.

The Sensitivity metric measures how well the models were able to predict the Default class. The Type II Error is its counterpart and measures the Default class misclassifications. Similarly, the Specificity metric measures the Non-Default classifications and the Type 1 Error measures the Non-Default misclassifications. The ideal model will have both high Sensitivity and high Specificity. Such a model will achieve a low negative class misclassification rate (Type II Error) and will also be optimized to achieve low positive class (Type I Error) misclassification rate.

Referring to the Sensitivity scores in Table 6.1, all of the models performed poorly in predicting the Default class using both the Train and Test datasets. With the exception of XGB, all of the models achieved fair performance in predicting the positive Non-Default class. XGB performed good in predicting Non-Default. While these two metrics provide good insight on how well the models were able to predict Default, it becomes tricky in determining which model performed the best using these metrics. With both the Train and Test datasets, the model that performed the best in predicting the negative Default class, also performed the worst in predicting the Non-Default class. And the models that performed the best in predicting the positive Non-Default class, also performed the worst in predicting Default. For example, with the Train sample, the XGB model achieved the lowest Type I Error score but also achieved the highest Type II Error score. On the other hand, the LR model achieved the highest Type I Error score, but also achieved the lowest Type II Error.

Because of the imbalance of the Default dataset classes; the models achieving varying ranks in predictive performance (Table 6.2); and from an interpretation standpoint, the F1 metric is the metric of importance to measure the model's predictive accuracy. F1 is the harmonic average of the Precision and Recall metrics and is not influenced by the imbalance of the class distribution in the Default dataset. Being a harmonic average, the F1 smooths the value imbalances in the Precision and Recall metrics. Precision is a metric of the exactness and accounts for the number of false positives misclassifications. A low Precision score may indicate that a large number of false positives were predicted by the model. A high Precision score suggests that there are a low number of false positives. Recall is a metric of completeness. A low Recall number indicates that there are many false negatives. A high Recall score suggests a low number of false negatives.

Based on the F1 criteria, the model that achieved the highest F1 score would be considered the better model to predict Default. The NN model performed the best with the in-sample dataset with an F1 score of .7132. The NN model was the most accurate model in predicting the negative Default class but it was also the least accurate in predicting the Non-Default class. In terms of overall Accuracy, the NN model ranked in third place with an Accuracy score of .7393. Although the statistics may suggest that the NN model and the other models potentially have predictive value, this cannot be ascertained until a new set of data is applied to the predictive models.

The four models were cross-validated using out-of-sample data. As can be seen in Tables 6.1 and 6.2, the LR model computed the highest F1 score of .6984. Similar to the NN model with the Train dataset, the LR model was the model that most accurately predicted the negative class but was also the least accurate in predicting the Non-Default class. The LR model also ranked last among the four models in overall Accuracy and ranked third using the AUC score. The LR model is an example of the Accuracy Paradox. The model achieved the lowest overall Accuracy among the four models, but based on the other metrics, it performed the best at classifying the Default and Non-Default classes.

7. Conclusion

7.1 Discussion

In spite of bank efforts to award credit cards to the most qualified applicants, the banks still experience customer defaults of credit card debt resulting in financial losses for the banks. Because the real probability of default is unknown, having a valid and up-to-date model or models to detect credit card default risk is crucial for risk management and to lessen financial losses resulting from defaults. This study set out to construct and explore the predictive value of four models. The models were constructed to predict the probability of default risk. All four models performed poorly in predicting defaults. Nonetheless, it may be possible to improve the model's predictive performance. The results of the analysis also showed that the models with the highest overall accuracy may not necessarily be the models that most accurately classify the negative and positive classes. Instead, the Logistic Regression model that achieved the lowest overall accuracy score using out-of-sample data, performed the best in distinguishing between the Default and Non-Default classes. All four models also selected the repayment behavior related variables as the most predictive of defaults. Once a card holder has become delinquent, the probabilities of default increase. Additionally, the models indicated that variables associated with Financial Indicators are more predictive of the response variable than Non-Financial Indicator variables. The demographic Non-Financial Indicator variables provide insight on the potential repayment ability of the individuals. But these variables cannot account for the card holder's current financial situation or circumstances. The Financial Indicators can grasp if an individual is currently experiencing financial stress that may lessen their ability to repay the credit card debt. Consequently, the Financial Indicator will have more predictive value than the Non-Financial Indicator variables.

7.2 Recommendations

All four models are adequate to be used to predict and forecast financial losses due to defaults. However, in their current state, the models are not fit to be used to take corrective actions on potential defaulters or in screening campaigns to accept qualified applicants or to reject unqualified applicants. The models perform poorly in distinguishing between the Default and Non-Default classes. The models generate a high percentage of false positives and false negatives. This means that the models will reject many qualified candidates but will accept many unqualified candidates. If the models are used to take proactive action to minimize default risk, there is a high probability that customers in good standing may be penalized while many in delinquency will be missed. The models can never be used in screening efforts to accept applicants. Because of their design, the models are incapable of detecting fraud and untruths on the applications. However, if the model's improvement efforts result in models with excellent (>= .90) Sensitivity and Specificity scores, the models can be used to screen and reject unqualified applicants.

The imbalance in the distribution of the classes discussed in section 4.1.1 does not seem to have adversely impacted the accuracy of the models. Based on the results of this analysis, it appears that the classes separate well as none of the four models favored the class with the most samples. Based on this, it is with hesitance that the implementation of sampling approaches that are employed to resolve class imbalances is recommended. Because it appears that the classes have well separation, it is believed that the sampling approaches will not have a positive impact on the accuracy of the models. However, several other issues that may impair the model's accuracy might be introduced by implementing the sampling methods. Depending on the sampling approach employed, information can be lost or knowledge of the actual class imbalanced distribution will be lost by the models. If either of these issues occur, the models will have no prediction value when introduced to new data that has the imbalanced distribution or the lost information. Nevertheless, to understand the impact that the sampling methods will have on the models, they should be used in future testing.

MODEL # 101: CREDIT CARD DEFAULT MODEL

The data is exaggeratingly skewed even after efforts were made to center it by trimming outliers to the 5% and 95% quantiles. It is believed that the skewness is adversely impacting the accuracy of the models. Several more cut-off boundary points should be tested to attempt to improve the model's accuracy in predicting default.

The models should also be tested with fewer features. The study results indicate that the model accuracy declines when larger quantities of features are used. The features that were denoted as less important/significant should be removed from the modeling data. In addition, model improvement processes such as grid-search and parameter tuning should be utilized to attempt to improve the accuracy of the models.

More complex Neural Network architectures should be implemented in future tests. For instance, utilizing backpropagation may provide better performance results than the simple feedforward model that was used in this study. Deep Learning methods such as Recurrent Neural Network (RNN) and its variants should also be tested. RNNs are designed to work with sequential data. Such models provide for the use of the raw monthly payment and bill statement amounts.

Bibliography

1. Yap, Bee & Ong, Seng-Huat & Huselina Mohamed Husain, Nor. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Syst. Appl. 38. 13274-13283. 10.1016/j.eswa.2011.04.147.

2. Yeh, Ivy & Lien, Che-Hui. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications. 36. 2473-2480. 10.1016/j.eswa.2007.12.020.

3. Bijak, K., & Thomas, L.C. (2012). Does segmentation always improve model performance in credit scoring? Expert Syst. Appl., 39, 2433-2442.

4. Thomas, Lyn. (2009). Consumer Credit Models: Pricing, Profit and Portfolios. Oxford University Press. 10.1093/acprof:oso/9780199232130.001.1.

Appendix A

Engineered Variables

Variable Label	Description
PAY_6_m	PAY_* variable Modified with corrections
PAY_5_m	PAY_* variable Modified with corrections
PAY_4_m	PAY_* variable Modified with corrections
PAY_3_m	PAY_* variable Modified with corrections
PAY_2_m	PAY_* variable Modified with corrections
PAY_1_m	PAY_* variable Modified with corrections
LIMIT_BAL_imp_5.95	(Credit) Balance Limit capped at 5% and 95 % quantiles.
BILL_AMT1_imp_5.95	BILL_AMT* variable capped at 5% and 95% quantiles.
BILL_AMT2_imp_5.95	BILL_AMT* variable capped at 5% and 95% quantiles.
BILL_AMT3_imp_5.95	BILL_AMT* variable capped at 5% and 95% quantiles.
BILL_AMT4_imp_5.95	BILL_AMT* variable capped at 5% and 95% quantiles.
BILL_AMT5_imp_5.95	BILL_AMT* variable capped at 5% and 95% quantiles.
BILL_AMT6_imp_5.95	BILL_AMT* variable capped at 5% and 95% quantiles.
PAY_AMT1_imp_5.95	PAY_AMT* variable capped at 5% and 95% quantiles.
PAY_AMT2_imp_5.95	PAY_AMT* variable capped at 5% and 95% quantiles.
PAY_AMT3_imp_5.95	PAY_AMT* variable capped at 5% and 95% quantiles.
PAY_AMT4_imp_5.95	PAY_AMT* variable capped at 5% and 95% quantiles.
PAY_AMT5_imp_5.95	PAY_AMT* variable capped at 5% and 95% quantiles.
PAY_AMT6_imp_5.95	PAY_AMT* variable capped at 5% and 95% quantiles.
R_SEX	Discretized variable: (Male; Female)
R_EDUCATION	Discretized variable: (Graduate; University; High School; Other; Unknown)
R_MARRIAGE	Discretized variable: (Married, Single, Others, Unknown)
R_DEFAULT	Discretized variable: (0; 1)
R_PAY_1	Discretized variable: (-1; 2; 3; 4; 5; 6; 7; 8;9)
R_PAY_2	Discretized variable: (-1; 2; 3; 4; 5; 6; 7; 8;9)
R_PAY_3	Discretized variable: (-1; 2; 3; 4; 5; 6; 7; 8;9)
R_PAY_4	Discretized variable: (-1; 2; 3; 4; 5; 6; 7; 8;9)
R_PAY_5	Discretized variable: (-1; 2; 3; 4; 5; 6; 7; 8;9)
R_PAY_6	Discretized variable: (-1; 2; 3; 4; 5; 6; 7; 8;9)
AGE1_bin	Discretized variable: ('NA', '18 to 25 years'; '26 to 40 years'; '41 to 65 years'; '66 to 100 years')
AGE2_bin	Discretized variable: ('NA', '20s', '30s', '40s', '50s', '60s', '70 or older')
LIMIT_BAL_bin	Discretized variable: ('NA'; 'Low' [<= 39,999]; 'Low-Med' [40,000 - 139,999]; 'Med' [140,000 - 239,000]; 'Med-High' [240,000 – 359,999]; 'High' [360,000 – 499,999]; 'Very High' [500,000 <=])
AGE1_woe_bin	Discretized variable: ('NA'; '18-30'; '31-40'; '41-65'; 'Over 65')
AGE2_woe_bin	Discretized variable: ('NA'; '20s + 40s + 50s'; '30s'; '60s + 70s')
AGE3_woe_bin	Discretized variable: ('NA'; '25 & Under'; '26-40'; '41 & Over')

MODEL # 101: CREDIT CARD DEFAULT MODEL

Variable Label	Description
R_AVG_BILL_AMT_imp_5.95	Average bill amount (BILL_AMT*) over the six months
R_AVG_PAY_AMT_imp_5.95	Average pay amount (PAY_AMT*) over the six months
BILL_AMT6_dum_5.95	Dummy Variable to hold Monthly BILL_AMT* negative values converted to zero (0) value
BILL_AMT5_dum_5.95	Dummy Variable to hold Monthly BILL_AMT* negative values converted to zero (0) value
BILL_AMT4_dum_5.95	Dummy Variable to hold Monthly BILL_AMT* negative values converted to zero (0) value
BILL_AMT3_dum_5.95	Dummy Variable to hold Monthly BILL_AMT* negative values converted to zero (0) value
BILL_AMT2_dum_5.95	Dummy Variable to hold Monthly BILL_AMT* negative values converted to zero (0) value
BILL_AMT1_dum_5.95	Dummy Variable to hold Monthly BILL_AMT* negative values converted to zero (0) value
R_PMT_RATIO5_6_5.95	Payment Ratio (PAY_AMT*_imp_5.95/BILL_AMT*_imp_5.95)
R_PMT_RATIO4_5_5.95	Payment Ratio (PAY_AMT*_imp_5.95/BILL_AMT*_imp_5.95)
R_PMT_RATIO3_4_5.95	Payment Ratio (PAY_AMT*_imp_5.95/BILL_AMT*_imp_5.95)
R_PMT_RATIO2_3_5.95	Payment Ratio (PAY_AMT*_imp_5.95/BILL_AMT*_imp_5.95)
R_PMT_RATIO1_2_5.95	(Current) Payment Ratio (PAY_AMT*_imp_5.95/BILL_AMT*_imp_5.95)
R_AVG_PMT_RATIO_imp_5.95	Average Payment Ratio (R_PMT_RATIO*_5.95) over the six months
AVGPAYRATIO_DT_bin	Discretized variable: ('NA'; '<11'; '> 11')
AVGPAYRATIO_OneR_bin	Discretized variable: ('NA'; '< 2.47'; '> 2.47')
R_UTIL1_imp_5.95	Utilization. Current Balance/Credit Limit (BILL_AMT*_imp_5.95/LIMIT_BAL_imp_5.95)
R_UTIL2_imp_5.95	Utilization. Current Balance/Credit Limit (BILL_AMT*_imp_5.95/LIMIT_BAL_imp_5.95)
R_UTIL3_imp_5.95	Utilization. Current Balance/Credit Limit (BILL_AMT*_imp_5.95/LIMIT_BAL_imp_5.95)
R_UTIL4_imp_5.95	Utilization. Current Balance/Credit Limit (BILL_AMT*_imp_5.95/LIMIT_BAL_imp_5.95)
R_UTIL5_imp_5.95	Utilization. Current Balance/Credit Limit (BILL_AMT*_imp_5.95/LIMIT_BAL_imp_5.95)
R_UTIL6_imp_5.95	Utilization. Current Balance/Credit Limit (BILL_AMT*_imp_5.95/LIMIT_BAL_imp_5.95)
R_AVG_UTIL_imp_5.95	Average monthly utilization (R_UTIL*_imp_5.95) over the six months
R_BAL_GROWTH_6MO_imp_5.95	Balance Growth Over 6 Months (BILL_AMT61_imp_6.95 - BILL_AMT6_imp_5.95)
R_BAL_GROWTH_MIN_imp_5.95	Minimum Balance Growth Value Over 6 Months
R_GROWTH_6MO_LIMIT_imp_5.95	Growth/Credit Limit Ratio (R_UTIL_GROWTH_6MO_imp_5.95/LIMIT_BAL_imp_5.95)
R_GROWTH_MIN_LIMIT_imp_5.95	Minimum Growth/Credit Limit Ratio (R_UTIL_GROWTH_6MO_imp_5.95/LIMIT_BAL_imp_5.95)
R_UTIL_GROWTH_6MO_imp_5.95	Utilization Growth Over 6 Months
R_UTIL_GROWTH_MIN_imp_5.95	Minimum Utilization Growth Over 6 Months

MODEL # 101: CREDIT CARD DEFAULT MODEL

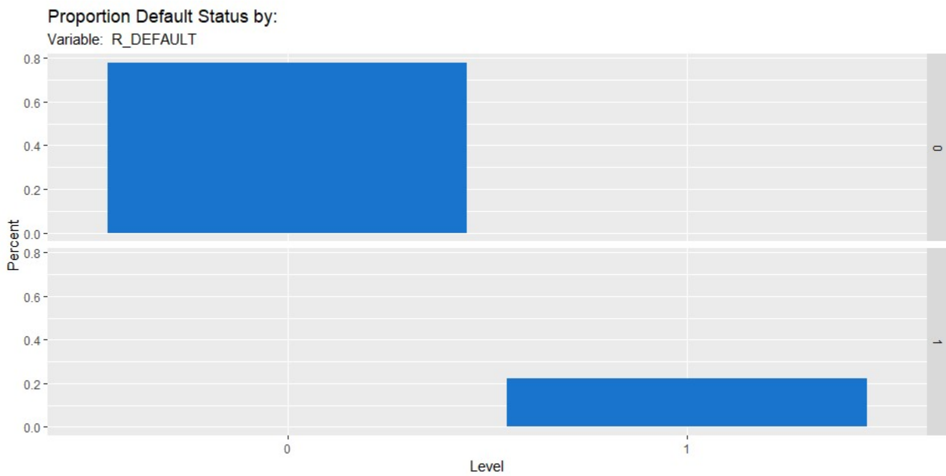
Variable Label	Description
R_UTIL_GROWTH_6MO_LIMIT_imp_5.95	Utilization Growth Over 6 Months/Credit Limit Ratio
R_UTIL_GROWTH_MIN_LIMIT_imp_5.95	Minimum Utilization Growth Over 6 Months/Credit Limit Ratio
R_MAX_BILL_AMT_imp_5.95	Maximum Bill Amount Over 6 Months
R_MAX_PMT_AMT_imp_5.95	Maximum Payment Amount Over 6 Months
PAY_AMT1_DUM	Dummy Variable to hold conversion R_PAY* values of -1 and -2 to zero (0)
PAY_AMT2_DUM	Dummy Variable to hold conversion R_PAY* values of -1 and -2 to zero (0)
PAY_AMT3_DUM	Dummy Variable to hold conversion R_PAY* values of -1 and -2 to zero (0)
PAY_AMT4_DUM	Dummy Variable to hold conversion R_PAY* values of -1 and -2 to zero (0)
PAY_AMT5_DUM	Dummy Variable to hold conversion R_PAY* values of -1 and -2 to zero (0)
PAY_AMT6_DUM	Dummy Variable to hold conversion R_PAY* values of -1 and -2 to zero (0)
R_MAX_DLQ	Maximum Delinquency (max of the PAY_AMT*_DUM variables)
MAXDLQ_woe_bin	Discretized variable: ('NA'; '= 0'; '> 0')
MAXDLQ_OneR_bin	Discretized variable: ('NA'; '= 0'; '> 0')
R_CHARGE6_imp_5.95	Estimated Monthly Charges
R_CHARGE5_imp_5.95	Estimated Monthly Charges
R_CHARGE4_imp_5.95	Estimated Monthly Charges
R_CHARGE3_imp_5.95	Estimated Monthly Charges
R_CHARGE2_imp_5.95	Estimated Monthly Charges
R_AVG_CHARGE_imp_5.95	Average of Estimated Monthly Charges Over 6 Months
R_CHARGE6_LIMIT_imp_5.95	Estimated Monthly Charges/Credit Limit Ratio (R_CHARGE6_imp_5.95/LIMIT_BAL_imp_5.95)
R_CHARGE5_LIMIT_imp_5.95	Estimated Monthly Charges/Credit Limit Ratio (R_CHARGE6_imp_5.95/LIMIT_BAL_imp_5.95)
R_CHARGE4_LIMIT_imp_5.95	Estimated Monthly Charges/Credit Limit Ratio (R_CHARGE6_imp_5.95/LIMIT_BAL_imp_5.95)
R_CHARGE3_LIMIT_imp_5.95	Estimated Monthly Charges/Credit Limit Ratio (R_CHARGE6_imp_5.95/LIMIT_BAL_imp_5.95)
R_CHARGE2_LIMIT_imp_5.95	Estimated Monthly Charges/Credit Limit Ratio (R_CHARGE6_imp_5.95/LIMIT_BAL_imp_5.95)
R_AVG_CHARGE_LIMIT_imp_5.95	Average of Estimated Monthly Charges/Credit Limit Ratio Over 6 Months
R_BILL_LIMIT_BAL1_VAR	Statement Amount Variance to Credit Lim (BILL_AMT1 - LIMIT_BAL)
R_BILL_LIMIT_BAL2_VAR	Statement Amount Variance to Credit Lim (BILL_AMT1 - LIMIT_BAL)
R_BILL_LIMIT_BAL3_VAR	Statement Amount Variance to Credit Lim (BILL_AMT1 - LIMIT_BAL)
R_BILL_LIMIT_BAL4_VAR	Statement Amount Variance to Credit Lim (BILL_AMT1 - LIMIT_BAL)

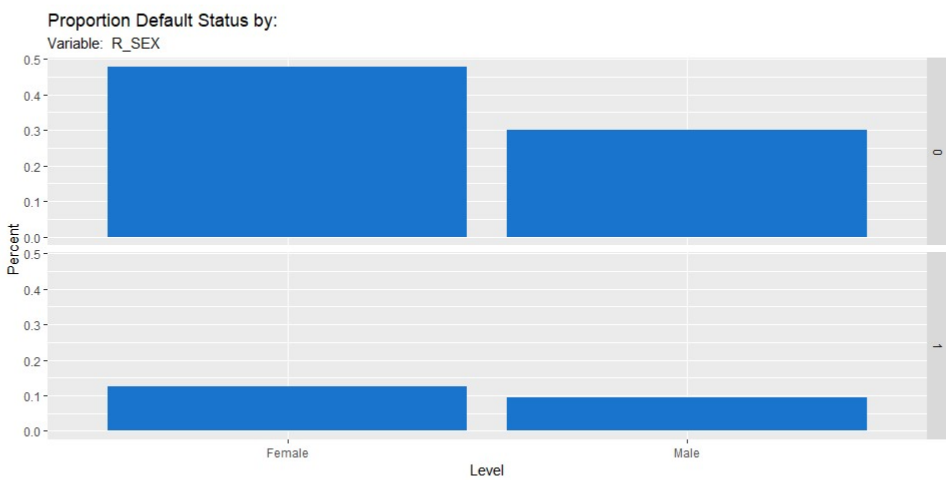
MODEL # 101: CREDIT CARD DEFAULT MODEL

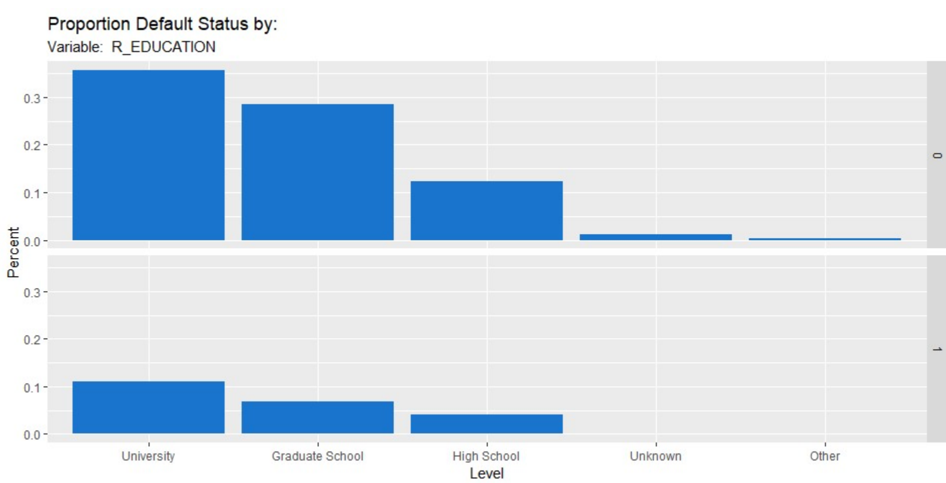
Variable Label	Description
R_BILL_LIMIT_BAL5_VAR	Statement Amount Variance to Credit Lim (BILL_AMT1 - LIMIT_BAL)
R_BILL_LIMIT_BAL6_VAR	Statement Amount Variance to Credit Lim (BILL_AMT1 - LIMIT_BAL)
R_BILL_LIMIT_BAL_VAR_MAX	Maximum Statement Amount Variance to Credit Limit Over 6 Months
BILL_EXCEEDS_LIMIT_BAL_FLAG	Bill Amount Exceeds Credit Limit Flag (0 = No; 1 = Yes)

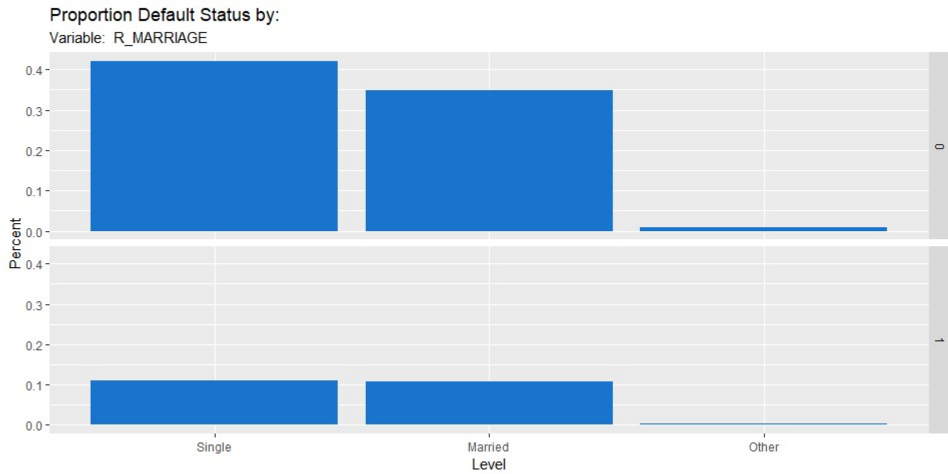
Appendix B

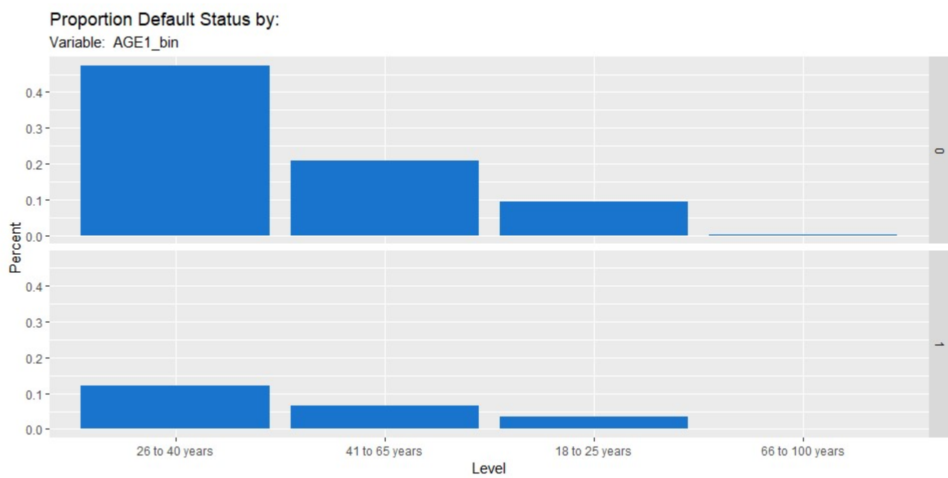
Categorical Variable Frequency Graphs by Segments

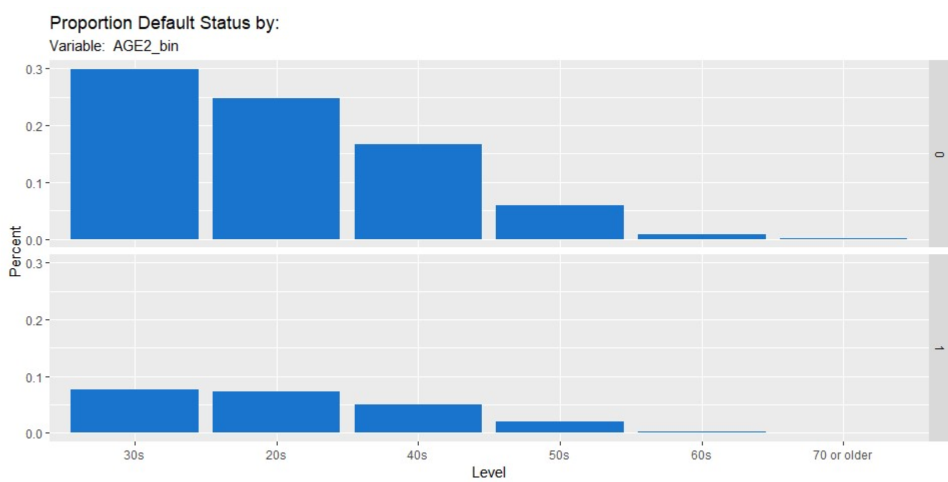


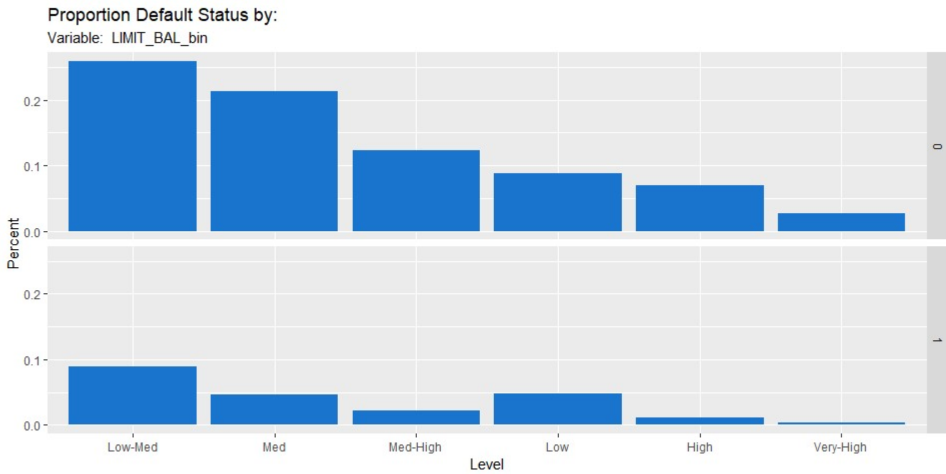


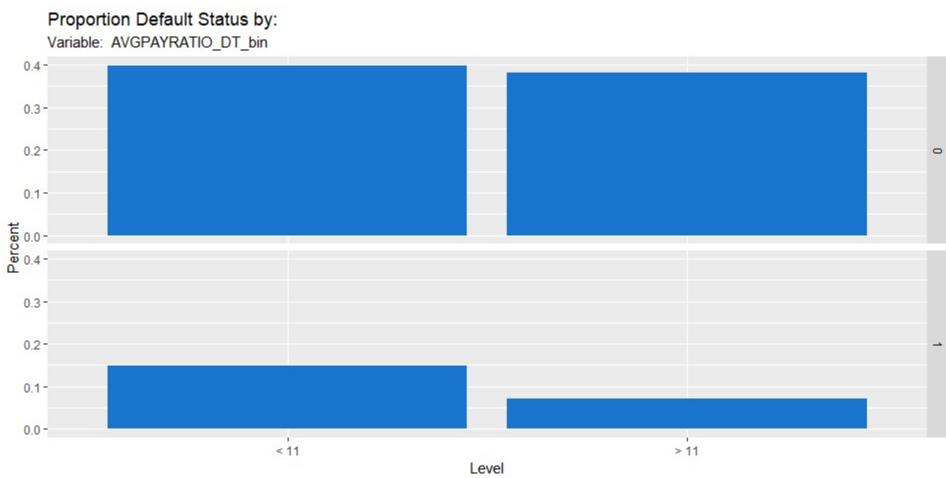


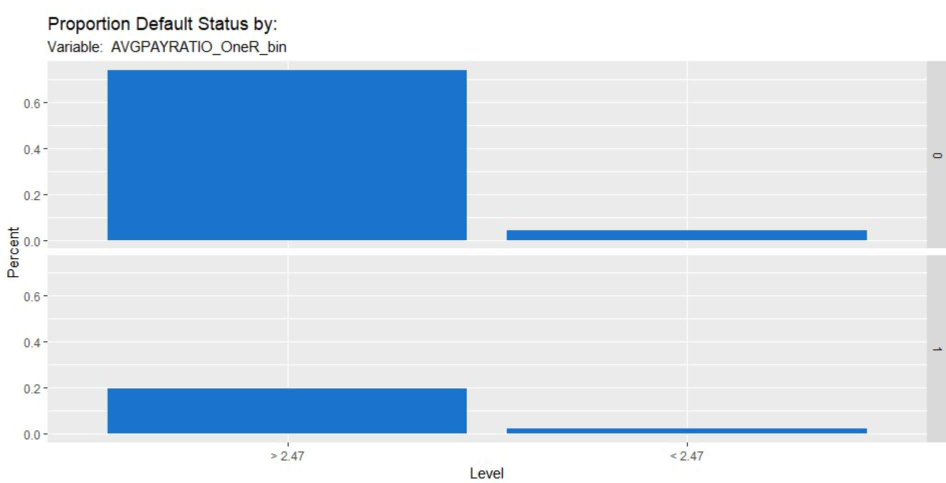

Proportion Default Status by:
Variable: R_MARRIAGE
Percent
0.0
0.1
0.2
0.3
0.4
0
1
Single
Married
Other
Level

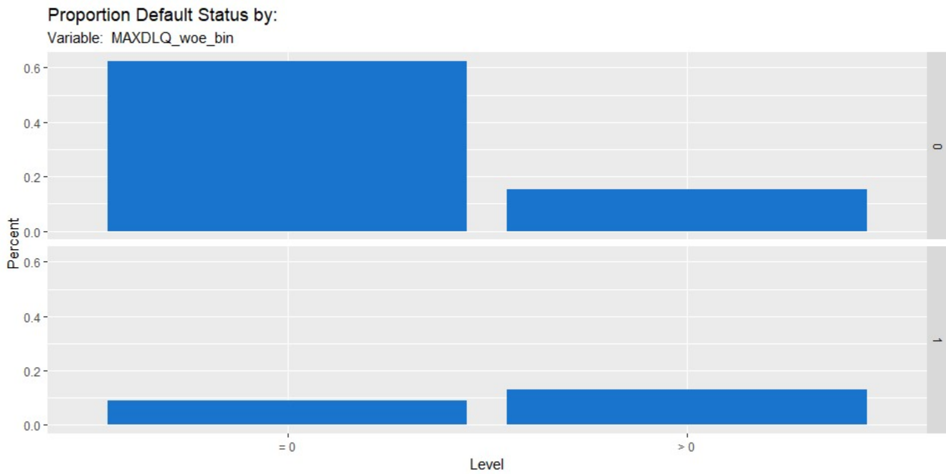

Proportion Default Status by:
Variable: AGE1_bin
Percent
0.0
0.1
0.2
0.3
0.4
0
1
26 to 40 years
41 to 65 years
18 to 25 years
66 to 100 years
Level

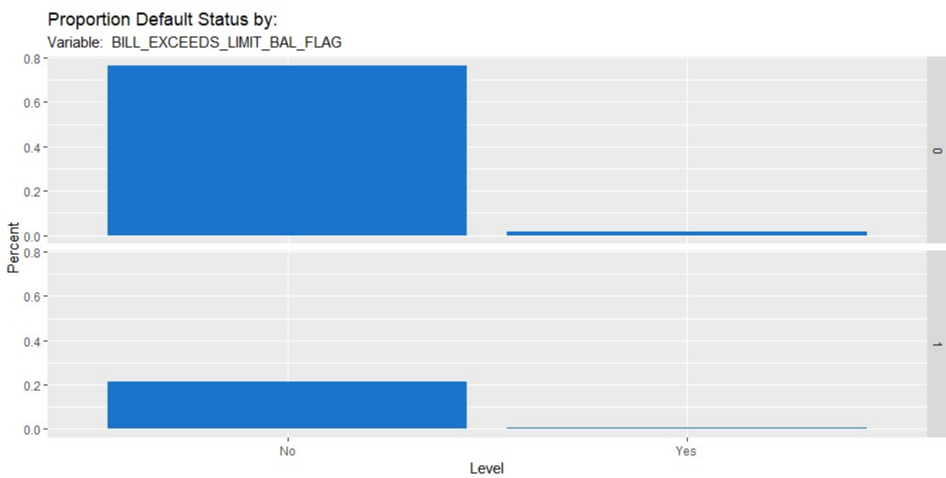

Proportion Default Status by:
Variable: AGE2_bin
Percent
0.0
0.1
0.2
0.3
0
1
30s
20s
40s
50s
60s
70 or older
Level


Proportion Default Status by:
Variable: LIMIT_BAL_bin
Percent
0.0
0.1
0.2
0
1
Low-Med
Med
Med-High
Low
High
Very-High
Level


Proportion Default Status by:
Variable: AVGPAYRATIO_DT_bin
Percent
0.0
0.1
0.2
0.3
0.4
0
1
< 11
> 11
Level


Proportion Default Status by:
Variable: AVGPAYRATIO_OneR_bin
Percent
0.0
0.2
0.4
0.6
0
1
> 2.47
< 2.47
Level


Proportion Default Status by:
Variable: MAXDLQ_woe_bin
Percent
0.6
0.4
0.2
0.0
0
1
= 0
> 0
Level


Proportion Default Status by:
Variable: BILL_EXCEEDS_LIMIT_BAL_FLAG
Percent
0.8
0.6
0.4
0.2
0.0
0
1
No
Yes
Level

Appendix C

Logistic Regression Naive Model

	Dependent variable:
	R_DEFAULT
LIMIT_BAL_imp_5.95	0.0000***
R_SEXMale	0.11***
R_EDUCATIONHigh School	-0.09*
R_EDUCATIONOther	-1.00**
R_EDUCATIONUniversity	-0.03
R_EDUCATIONUnknown	-1.08***
R_MARRIAGEOther	-0.17
R_MARRIAGESingle	-0.19***
AGE1_bin26 to 40 years	-0.09
AGE1_bin41 to 65 years	-0.09
AGE1_bin66 to 100 years	-0.09
AGE2_bin30s	0.05
AGE2_bin40s	0.14
AGE2_bin50s	0.12
AGE2_bin60s	0.25
AGE2_bin70 or older	0.34
LIMIT_BAL_binLow-Med	-0.11
LIMIT_BAL_binMed	-0.15
LIMIT_BAL_binMed-High	-0.14
LIMIT_BAL_binHigh	-0.01
LIMIT_BAL_binVery-High	0.51*
R_AVG_BILL_AMT_imp_5.95	0
R_AVG_PAY_AMT_imp_5.95	-0.0003***
R_PMT_RATIO1_2_5.95	0
R_AVG_PMT_RATIO_imp_5.95	0
AVGPAYRATIO_DT_bin> 11	-0.02
AVGPAYRATIO_OneR_bin> 2.47	-0.67***
R_UTIL1_imp_5.95	-0.01**
R_AVG_UTIL_imp_5.95	0.01**
R_BAL_GROWTH_6MO_imp_5.95	-0.0000***
R_BAL_GROWTH_MIN_imp_5.95	0
R_GROWTH_6MO_LIMIT_imp_5.95	0.0003
R_GROWTH_MIN_LIMIT_imp_5.95	0.003
R_UTIL_GROWTH_6MO_imp_5.95	
R_UTIL_GROWTH_MIN_imp_5.95	
R_UTIL_GROWTH_6MO_LIMIT_imp_5.95	0.9
R_UTIL_GROWTH_MIN_LIMIT_imp_5.95	-1.08
R_MAX_BILL_AMT_imp_5.95	0
R_MAX_PMT_AMT_imp_5.95	0.0000**
R_MAX_DLQ	0.32***
MAXDLQ_woe_bin> 0	0.55***
R_AVG_CHARGE_imp_5.95	0.0001***
R_AVG_CHARGE_LIMIT_imp_5.95	-0.004***
R_BILL_LIMIT_BAL_VAR_MAX	0.0000***
BILL_EXCEEDS_LIMIT_BAL_FLAGYes	-0.03
Constant	-0.60***
Observations	30,000
Log Likelihood	-13,506.99
Akaike Inf. Crit.	27,101.97
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix D

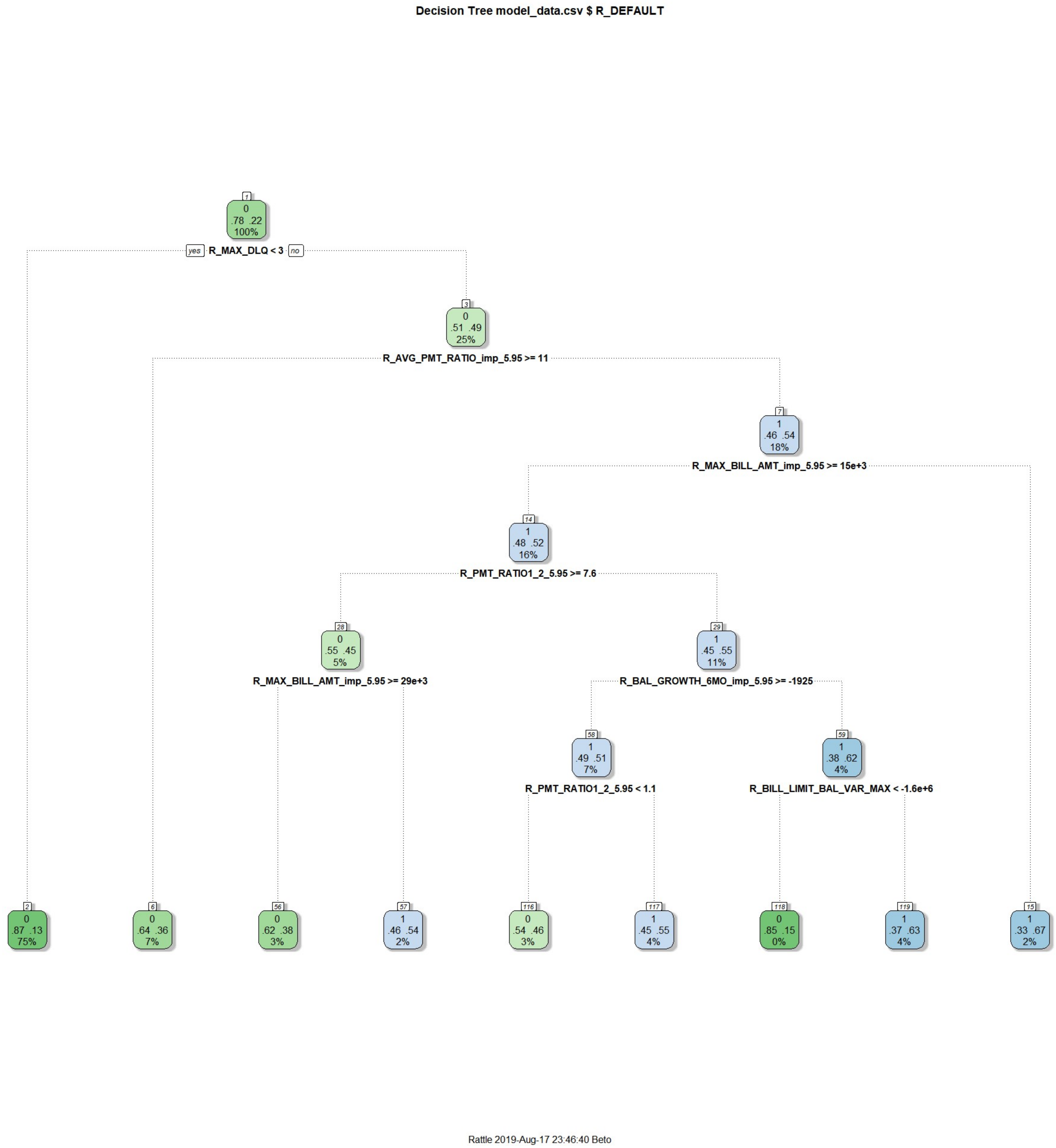
One Rule Model

```
Attribute                         Accuracy
1 * R_MAX_DLQ                         78.83%
2   LIMIT_BAL_imp_5.95                77.88%
2   R_SEX                             77.88%
2   R_EDUCATION                       77.88%
2   R_MARRIAGE                        77.88%
2   AGE1_bin                          77.88%
2   AGE2_bin                          77.88%
2   LIMIT_BAL_bin                     77.88%
2   R_AVG_BILL_AMT_imp_5.95           77.88%
2   R_AVG_PAY_AMT_imp_5.95            77.88%
2   R_PMT_RATIO1_2_5.95               77.88%
2   R_AVG_PMT_RATIO_imp_5.95          77.88%
2   AVGPAYRATIO_DT_bin                77.88%
2   AVGPAYRATIO_OneR_bin              77.88%
2   R_UTIL1_imp_5.95                  77.88%
2   R_AVG_UTIL_imp_5.95               77.88%
2   R_BAL_GROWTH_6MO_imp_5.95         77.88%
2   R_BAL_GROWTH_MIN_imp_5.95         77.88%
2   R_GROWTH_6MO_LIMIT_imp_5.95       77.88%
2   R_GROWTH_MIN_LIMIT_imp_5.95       77.88%
2   R_UTIL_GROWTH_6MO_imp_5.95        77.88%
2   R_UTIL_GROWTH_MIN_imp_5.95        77.88%
2   R_UTIL_GROWTH_6MO_LIMIT_imp_5.95 77.88%
2   R_UTIL_GROWTH_MIN_LIMIT_imp_5.95 77.88%
2   R_MAX_BILL_AMT_imp_5.95           77.88%
2   R_MAX_PMT_AMT_imp_5.95            77.88%
2   MAXDLQ_woe_bin                    77.88%
2   R_AVG_CHARGE_imp_5.95             77.88%
2   R_AVG_CHARGE_LIMIT_imp_5.95       77.88%
2   R_BILL_LIMIT_BAL_VAR_MAX          77.88%
2   BILL_EXCEEDS_LIMIT_BAL_FLAG       77.88%
---
Chosen attribute due to accuracy
and ties method (if applicable): '*'


Call:
OneR.formula(formula = R_DEFAULT ~ ., data = data, verbose = TRUE)
```

MODEL # 101: CREDIT CARD DEFAULT MODEL

```
Rules:
If R_MAX_DLQ = (-0.009,1.8] then R_DEFAULT = 1
If R_MAX_DLQ = (1.8,3.6]    then R_DEFAULT = 1
If R_MAX_DLQ = (3.6,5.4]    then R_DEFAULT = 0
If R_MAX_DLQ = (5.4,7.2]    then R_DEFAULT = 0
If R_MAX_DLQ = (7.2,9.01]   then R_DEFAULT = 0

Accuracy:
23648 of 30000 instances classified correctly (78.83%)
```

Appendix E



Note: green = No Default; blue = Default

MODEL # 101: CREDIT CARD DEFAULT MODEL

Rule number: 15 [R_DEFAULT=1 cover=478 (2%) prob=0.67]

R_MAX_DLQ>=2.5

R_AVG_PMT_RATIO_imp_5.95< 10.7

R_MAX_BILL_AMT_imp_5.95< 1.537e+04

Rule number: 119 [R_DEFAULT=1 cover=887 (4%) prob=0.63]

R_MAX_DLQ>=2.5

R_AVG_PMT_RATIO_imp_5.95< 10.7

R_MAX_BILL_AMT_imp_5.95>=1.537e+04

R_PMT_RATIO1_2_5.95< 7.556

R_BAL_GROWTH_6MO_imp_5.95< -1925

R_BILL_LIMIT_BAL_VAR_MAX>=-1.585e+06

Rule number: 117 [R_DEFAULT=1 cover=775 (4%) prob=0.55]

R_MAX_DLQ>=2.5

R_AVG_PMT_RATIO_imp_5.95< 10.7

R_MAX_BILL_AMT_imp_5.95>=1.537e+04

R_PMT_RATIO1_2_5.95< 7.556

R_BAL_GROWTH_6MO_imp_5.95>=-1925

R_PMT_RATIO1_2_5.95>=1.087

Rule number: 57 [R_DEFAULT=1 cover=410 (2%) prob=0.54]

R_MAX_DLQ>=2.5

R_AVG_PMT_RATIO_imp_5.95< 10.7

R_MAX_BILL_AMT_imp_5.95>=1.537e+04

R_PMT_RATIO1_2_5.95>=7.556

R_MAX_BILL_AMT_imp_5.95< 2.892e+04

Rule number: 116 [R_DEFAULT=0 cover=676 (3%) prob=0.46]

R_MAX_DLQ>=2.5

R_AVG_PMT_RATIO_imp_5.95< 10.7

R_MAX_BILL_AMT_imp_5.95>=1.537e+04

R_PMT_RATIO1_2_5.95< 7.556

R_BAL_GROWTH_6MO_imp_5.95>=-1925

R_PMT_RATIO1_2_5.95< 1.087

Rule number: 56 [R_DEFAULT=0 cover=543 (3%) prob=0.38]

R_MAX_DLQ>=2.5

R_AVG_PMT_RATIO_imp_5.95< 10.7

R_MAX_BILL_AMT_imp_5.95>=1.537e+04

R_PMT_RATIO1_2_5.95>=7.556

R_MAX_BILL_AMT_imp_5.95>=2.892e+04

Rule number: 6 [R_DEFAULT=0 cover=1443 (7%) prob=0.36]

R_MAX_DLQ>=2.5

R_AVG_PMT_RATIO_imp_5.95>=10.7

Rule number: 118 [R_DEFAULT=0 cover=13 (0%) prob=0.15]

R_MAX_DLQ>=2.5

R_AVG_PMT_RATIO_imp_5.95< 10.7

R_MAX_BILL_AMT_imp_5.95>=1.537e+04

R_PMT_RATIO1_2_5.95< 7.556

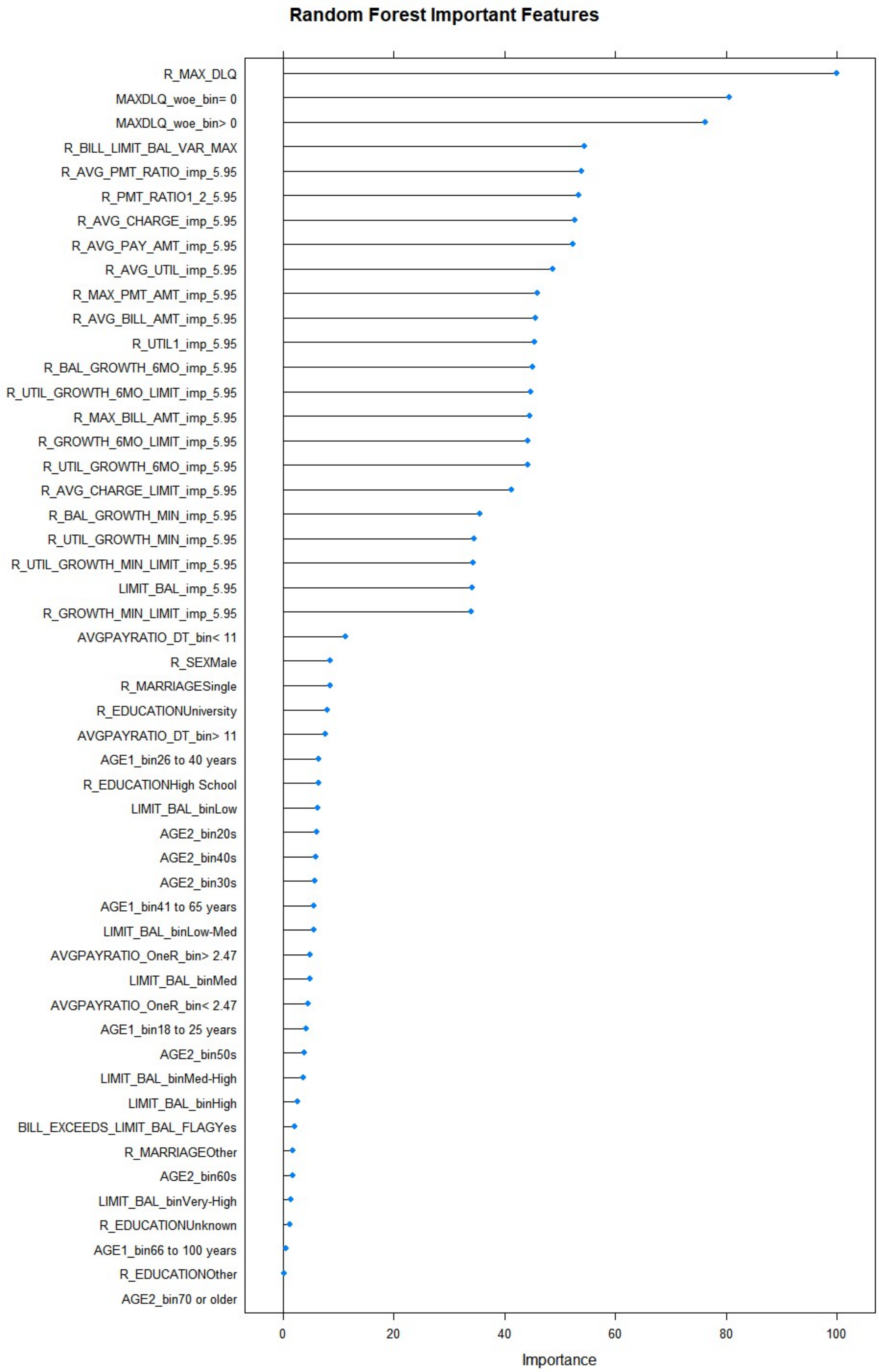
R_BAL_GROWTH_6MO_imp_5.95< -1925

R_BILL_LIMIT_BAL_VAR_MAX< -1.585e+06

Rule number: 2 [R_DEFAULT=0 cover=15775 (75%) prob=0.13]

R_MAX_DLQ< 2.5

Appendix F


Random Forest Important Features
R_MAX_DLQ
MAXDLQ_woe_bin= 0
MAXDLQ_woe_bin> 0
R_BILL_LIMIT_BAL_VAR_MAX
R_AVG_PMT_RATIO_imp_5.95
R_PMT_RATIO1_2_5.95
R_AVG_CHARGE_imp_5.95
R_AVG_PAY_AMT_imp_5.95
R_AVG_UTIL_imp_5.95
R_MAX_PMT_AMT_imp_5.95
R_AVG_BILL_AMT_imp_5.95
R_UTIL1_imp_5.95
R_BAL_GROWTH_6MO_imp_5.95
R_UTIL_GROWTH_6MO_LIMIT_imp_5.95
R_MAX_BILL_AMT_imp_5.95
R_GROWTH_6MO_LIMIT_imp_5.95
R_UTIL_GROWTH_6MO_imp_5.95
R_AVG_CHARGE_LIMIT_imp_5.95
R_BAL_GROWTH_MIN_imp_5.95
R_UTIL_GROWTH_MIN_imp_5.95
R_UTIL_GROWTH_MIN_LIMIT_imp_5.95
LIMIT_BAL_imp_5.95
R_GROWTH_MIN_LIMIT_imp_5.95
AVGPAYRATIO_DT_bin< 11
R_SEXMale
R_MARRIAGESingle
R_EDUCATIONUniversity
AVGPAYRATIO_DT_bin> 11
AGE1_bin26 to 40 years
R_EDUCATIONHigh School
LIMIT_BAL_binLow
AGE2_bin20s
AGE2_bin40s
AGE2_bin30s
AGE1_bin41 to 65 years
LIMIT_BAL_binLow-Med
AVGPAYRATIO_OneR_bin> 2.47
LIMIT_BAL_binMed
AVGPAYRATIO_OneR_bin< 2.47
AGE1_bin18 to 25 years
AGE2_bin50s
LIMIT_BAL_binMed-High
LIMIT_BAL_binHigh
BILL_EXCEEDS_LIMIT_BAL_FLAGYes
R_MARRIAGEOther
AGE2_bin60s
LIMIT_BAL_binVery-High
R_EDUCATIONUnknown
AGE1_bin66 to 100 years
R_EDUCATIONOther
AGE2_bin70 or older
0
20
40
60
80
100
Importance