Spring 2019

Matthew Dobbin

Northwestern University

MSDS 498 Capstone Project

Modelling Development Guide

Contents

1.0 Introduction

The purpose of this report is to document the analysis of the default of credit card clients data set which was obtained from the UCI Machine Learning Repository (UCI, 2016). The data set was originally used in research aimed at predicting the probability of default payments for credit card clients in Taiwan. The first part of the report describes the data set and documents the process of validating and cleaning the data.

The second part of the report describes the predictive modelling process. Five different machine learning algorithms were implemented to predict the probability of a default payment in the following month. The methods were Random Forest, Gradient Boosted, Logistic Regression, Naïve Bayes and Support Vector Machine (SVM). The report is summarised by comparing the in sample and out of sample model performance for each of the algorithms.

2.0 The Data

The data set was originally used in research aimed at predicting the probability of default payments for credit card clients in Taiwan. It contains customer demographics, credit data, history of payments and bill statements from April 2005 to September 2005. The data set contains 30,000 observations and there are 23 predictor variables. The binary variable, DEFAULT, is the response variable. A value of one represents a default payment while a zero does not. Five additional variables ‘u’,’ train’, ‘test’, ‘validate’ and ‘data.group’ were created for the purpose of splitting the data set into training, test and validation sets. A data dictionary is shown in Table 1.

Table 1 - Data dictionary for the credit card default data set.

Variable	Description
ID	Unique ID number for each row of data.
LIMIT_BAL	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family supplementary credit.
SEX	Gender: Male = 1, Female = 2
EDUCATION	Education: Graduate School = 1, University = 2, High school = 3, Others = 4
MARRIAGE	Marital Status: Married = 1, Single = 2, Others = 3
AGE	Age in years.
PAY_0	Repayment status in September 2005

PAY_2	Repayment status in August 2005
PAY_3	Repayment status in July 2005
PAY_4	Repayment status in June 2005
PAY_5	Repayment status in May 2005
PAY_6	Repayment status in April 2005
BILL_AMT1	Amount of bill statement (NT dollar) in September 2005.
BILL_AMT2	Amount of bill statement (NT dollar) in August 2005.
BILL_AMT3	Amount of bill statement (NT dollar) in July 2005.
BILL_AMT4	Amount of bill statement (NT dollar) in June 2005.
BILL_AMT5	Amount of bill statement (NT dollar) in May 2005.
BILL_AMT6	Amount of bill statement (NT dollar) in April 2005.
PAY_AMT1	Amount of previous payment (NT dollar) in September 2005
PAY_AMT2	Amount of previous payment (NT dollar) in August 2005
PAY_AMT3	Amount of previous payment (NT dollar) in July 2005
PAY_AMT4	Amount of previous payment (NT dollar) in June 2005
PAY_AMT5	Amount of previous payment (NT dollar) in May 2005
PAY_AMT6	Amount of previous payment (NT dollar) in April 2005
DEFAULT	Response Variable. Default Payment: Yes = 1, No = 0
u	A random number generated for splitting into training/test/validate set.
train	A value of 1 indicates the observation is used in the training data set.
test	A value of 1 indicates the observation is used in the test data set
validate	A value of 1 indicates the observation is used in the validation data set.
data.group	Constructed to partition the data set in a single dimension. data.group <- 1*train + 2*test + 3*validate

The PAY_0 to PAY_6 variables track the past monthly payment records from April to September 2005. The measurement scale for the repayment status is shown in Table 2.

Table 2 - Measurement scale for PAY_0 to PAY6 variables.

-1 = Pay duly	5 = payment delay for five months
1 = payment delay for one month	6 = payment delay for six months
2 = payment delay for two months	7 = payment delay for seven months
3 = payment delay for three months	8 = payment delay for eight months
4 = payment delay for four months	9 = payment delay for nine months and above

2.1 The Train/Test Split

A uniform random number was used to split the sample into training, test and validation data sets. The training set is used for in sample model development and the test set used for out of sample model assessment. The validation data set will be used to validate the performance of the production model. The count of observations for each data set is shown in Table 3.

Table 3 – Count of observations for the training, validation and test data sets.

	Train	Test	Validation
Observation Count	15,180	7,323	7,497

2.2 Summary Statistics

Table 4 shows the summary statistics for the variables in the data set. A scan of column 'N' revealed that there were no missing values in the data set. The 'Min' and 'Max' columns were used to determine if the values were within the expected range as described by the data dictionary. EDUCATION and PAY_0 to PAY_6 variables have values that are unexpected and were investigated further in the data preparation section of the report.

Table 4 - Summary statistics for the credit card default data set.

Statistic	N	Mean	St. Dev.	Min	P(25)	Median	P(75)	Max
ID	30,000	15,000.50	8,660.40	1	7,500.8	15,000.5	22,500.2	30,000
LIMIT_BAL	30,000	167,484.30	129,747.70	10,000	50,000	140,000	240,000	1,000,000
SEX	30,000	1.60	0.49	1	1	2	2	2
EDUCATION	30,000	1.85	0.79	0	1	2	2	6
MARRIAGE	30,000	1.55	0.52	0	1	2	2	3
AGE	30,000	35.49	9.22	21	28	34	41	79
PAY_0	30,000	-0.02	1.12	-2	-1	0	0	8
PAY_2	30,000	-0.13	1.20	-2	-1	0	0	8
PAY_3	30,000	-0.17	1.20	-2	-1	0	0	8
PAY_4	30,000	-0.22	1.17	-2	-1	0	0	8
PAY_5	30,000	-0.27	1.13	-2	-1	0	0	8
PAY_6	30,000	-0.29	1.15	-2	-1	0	0	8
BILL_AMT1	30,000	51,223.33	73,635.86	-165,580	3,558.8	22,381.5	67,091	964,511
BILL_AMT2	30,000	49,179.08	71,173.77	-69,777	2,984.8	21,200	64,006.2	983,931
BILL_AMT3	30,000	47,013.15	69,349.39	-157,264	2,666.2	20,088.5	60,164.8	1,664,089
BILL_AMT4	30,000	43,262.95	64,332.86	-170,000	2,326.8	19,052	54,506	891,586
BILL_AMT5	30,000	40,311.40	60,797.16	-81,334	1,763	18,104.5	50,190.5	927,171
BILL_AMT6	30,000	38,871.76	59,554.11	-339,603	1,256	17,071	49,198.2	961,664
PAY_AMT1	30,000	5,663.58	16,563.28	0	1,000	2,100	5,006	873,552
PAY_AMT2	30,000	5,921.16	23,040.87	0	833	2,009	5,000	1,684,259
PAY_AMT3	30,000	5,225.68	17,606.96	0	390	1,800	4,505	896,040

PAY_AMT4	30,000	4,826.08	15,666.16	0	296	1,500	4,013.2	621,000
PAY_AMT5	30,000	4,799.39	15,278.31	0	252.5	1,500	4,031.5	426,529
PAY_AMT6	30,000	5,215.50	17,777.47	0	117.8	1,500	4,000	528,666
DEFAULT	30,000	0.22	0.42	0	0	0	0	1

2.3 Data Preparation and Visualisation

The data quality of the modelling variables is investigated further in the following sections. A variety of charts have been used to visualise the distributions of the variables and their relationship with the target variable.

2.3.1 LIMIT_BAL

LIMIT_BAL is a continuous variable with values between $10,000 and $1,000,000. The distribution is shown in the histogram and boxplot in Figure 1. The plots show that the distribution is highly skewed and that most of the clients have credit limits below $500,000.

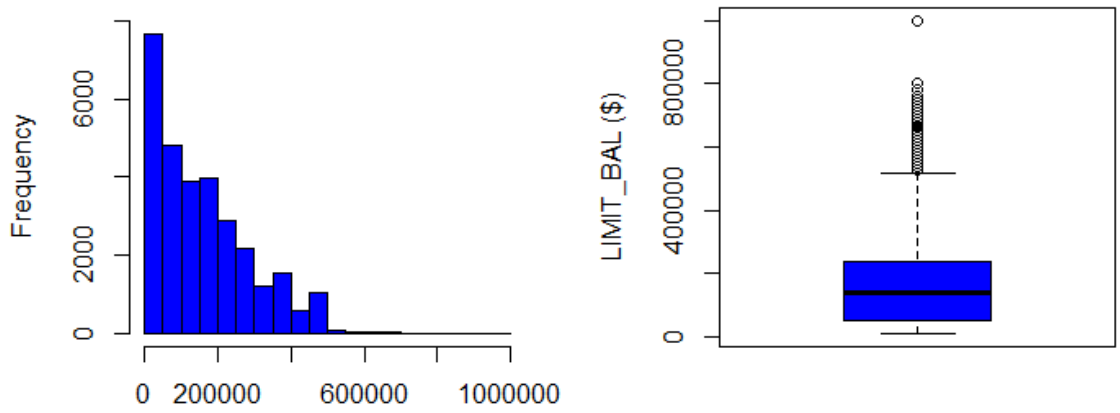

Figure 1 - Histogram and boxplot of LIMIT_BAL variable.

2.3.2 SEX

Figure 2 shows a bar plot and a mosaic plot of the SEX variable versus the response variable. Approximately 60% of the observations are female. The mosaic plot shows that there is a slight increase in the percentage of males that defaulted on payment compared to females. The indicator variable SEX_M was coded so that a value of one represents a male.

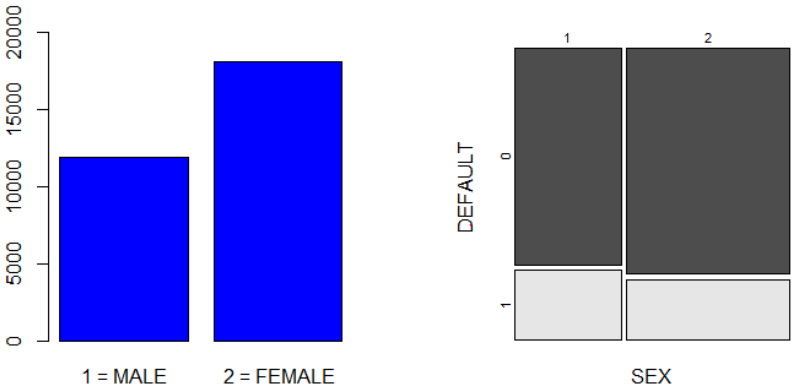

Figure 2 - Bar and mosaic plot of the SEX variable.

2.3.3 EDUCATION

The mosaic plot in Figure 3 shows that as the level of education decreases (grad school =1, high school = 3) there is a small increase in the percentage of default payments. The histogram reveals that there are unexpected values (0, 5, 6). These have been grouped with the 4-Others category which will be treated as the base variable. The indicator variables EDUCATION_GS, EDUCATION_UNI and EDUCATION_HS were created.

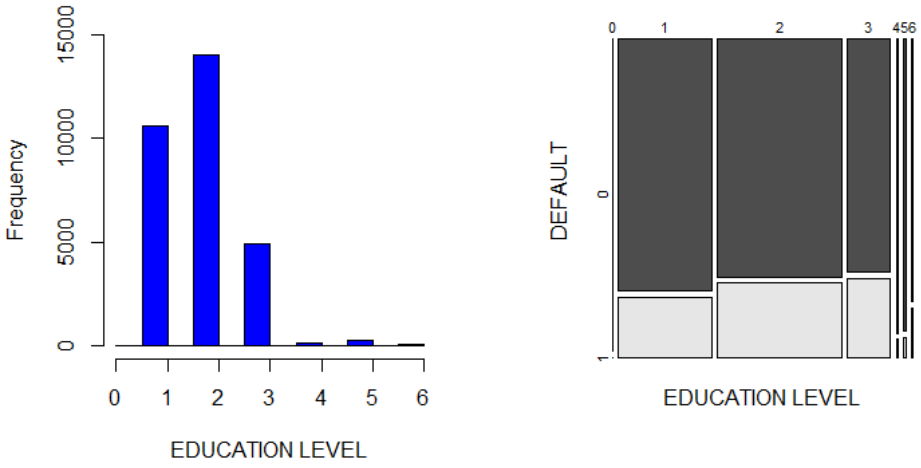

Figure 3 - Histogram and mosaic plot of the EDUCATION variable.

2.3.4 MARRIAGE

The mosaic plot in Figure 4 shows that the married category (1) has a slightly higher rate of default compared to singles (2). There is an unexpected value of zero. This has been grouped with the 3-Others category which will be treated as the base variable. The indicator variables MARRIAGE_MD and MARRIAGE_SN were created.

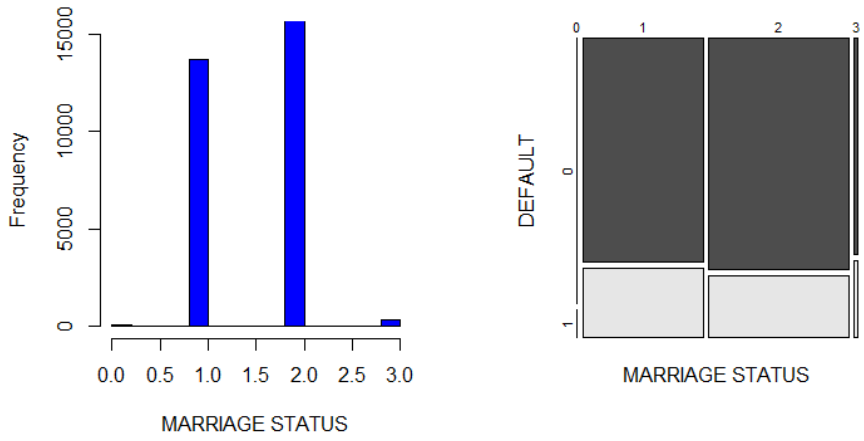

Figure 4 - Histogram and mosaic plot of MARRIAGE variable.

2.3.5 AGE

The distribution of the age variable is shown in Figure 5. The minimum and maximum values of 21 and 79 respectively are realistic ages for people making credit payments.

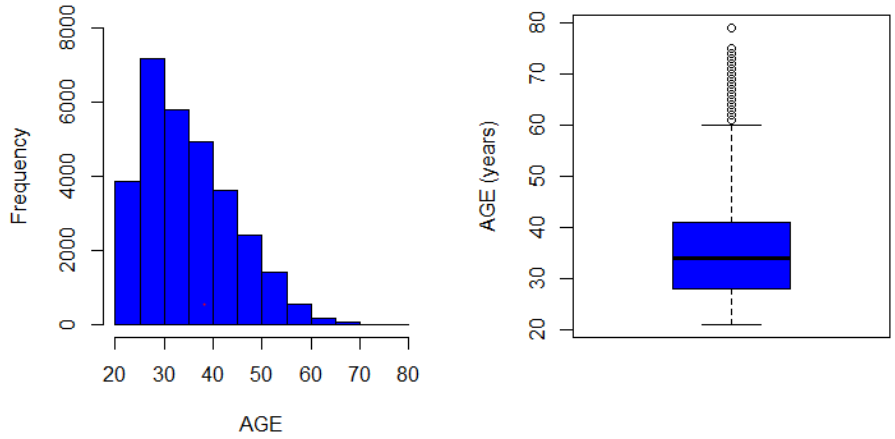

Figure 5 - Histogram and box plot of the AGE variable.

2.3.6 PAY_0 to PAY_6

Based on the information from the data dictionary the PAY_0 variable was renamed to PAY_1 so that it aligned with the BILL_AMT1 and PAY_AMT1 September variables. The histograms shown in Figure 6 revealed that there were unexpected negative values for the repayment status. This could indicate that the repayment was paid early. Using this assumption, the values of -2 and -1 were recoded as 0, which indicates it was paid on time. The recoded variables are PAY_R1, PAY_R2, PAY_R3, PAY_R4, PAY_R5 and PAY_R6.

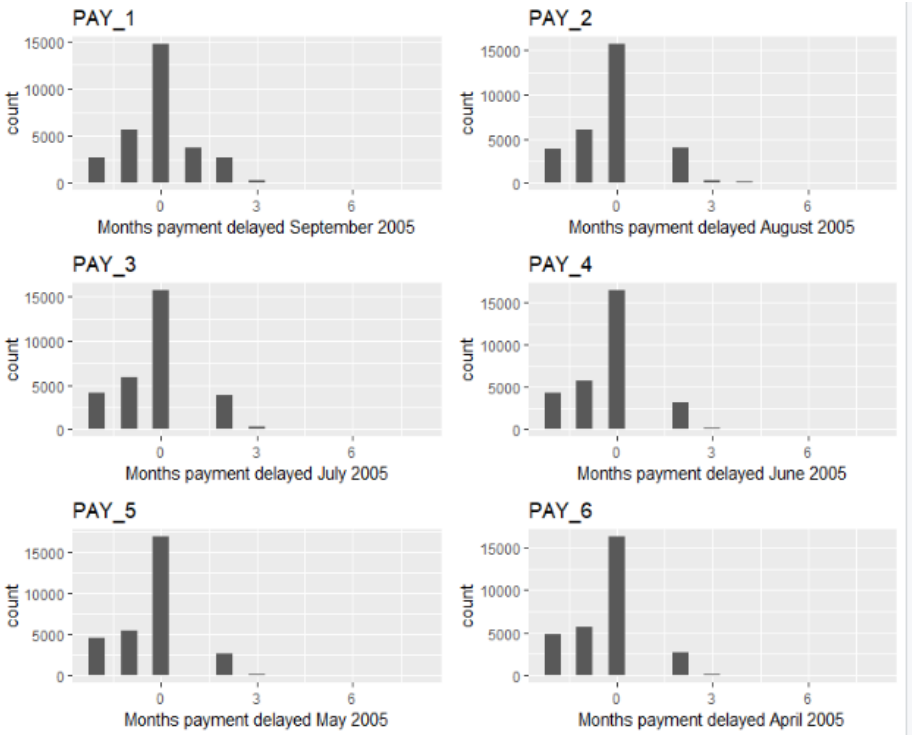

Figure 6 - Histograms of the repayment status variables.

2.3.7 BILL_AMT1 to BILL_AMT6

The boxplots in Figure 7 show that there is a wide distribution of the customer bill amounts. The maximum bill amount of approximately $1.7 million in July may be a potential outlier as the maximum balance limit for the dataset is $1.0 million. The value has been left untreated. Negative bill amounts may indicate that the client is ahead of payments.

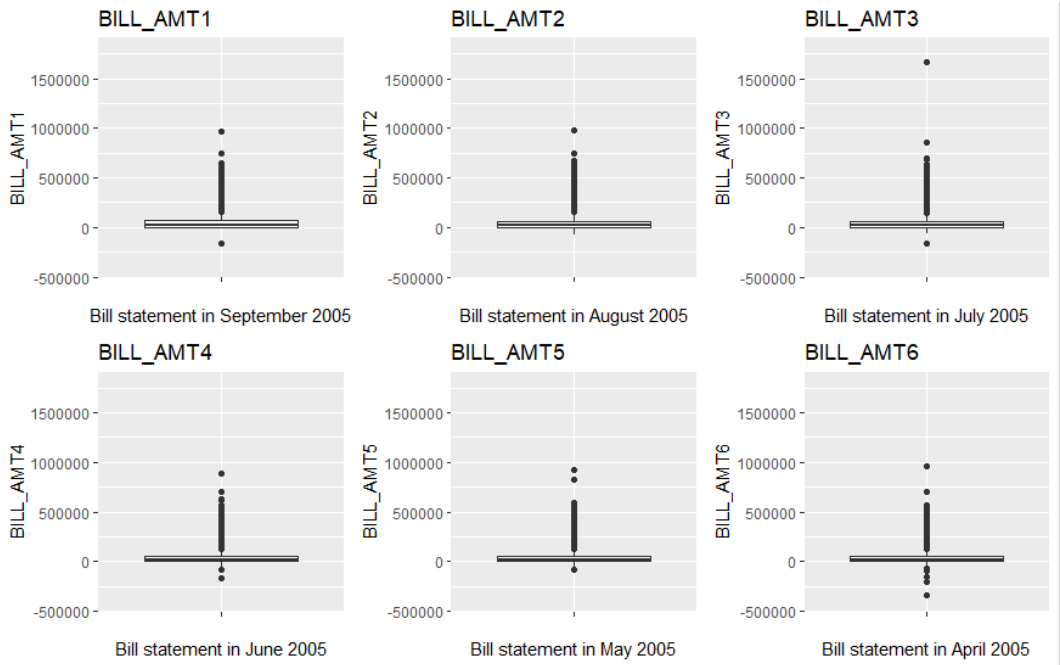

Figure 7 - Boxplots of the bill amount variables from April to September 2005.

2.3.8 PAY_AMT1 to PAY_AMT6

The boxplots in Figure 8 show the previous payment amount from April to September. The maximum previous payment of approximately $1.6 million in August 2005 aligns with the maximum bill amount of $1.7 million in July 2005.

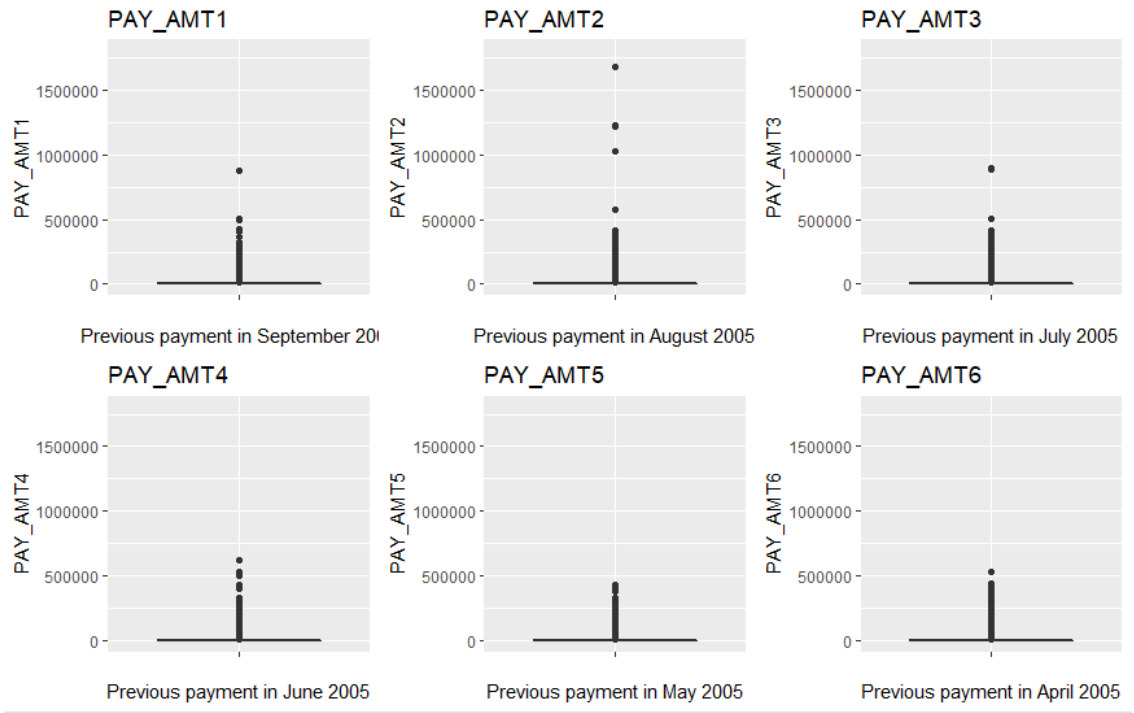

Figure 8 - Boxplots of the previous payment amount variables from April to September 2005.

2.3.9 DEFAULT

DEFAULT is the response variable for the data set. A value of one represents a default payment. Table 5 shows the split between on time and default payment observations. Approximately 22% of the clients defaulted on payment.

Table 5 - Observation count of the values for the default payments variable.

Data Set	0	1
Training	11757	3423
Test	5766	1557
Validation	5841	1656

3.0 Feature Engineering

The credit card default data set consists of demographic variables and six months of billing and payment history for each customer. Feature engineering was conducted on the billing and payment history to refine the raw variables into predictors that would be useful for training the model. The continuous variables were also discretised using the weight of evidence binning algorithm. The following sections detail how the features were engineered.

3.1 Age

The weight of evidence (WOE) supervised binning algorithm was used to determine the optimal bin sizes for the age variable. WOE can be used to create discrete variables that display separation in the response variable which leads to engineering better predictor variables. The plot in Figure 9 shows the intervals for the three age bins that were determined by the algorithm as well as the WOE scores. The indicator variables AGE_18_25 and AGE_26_40 were created and the age bin greater than 40 was treated as the base variable. The mosaic plot revealed that clients under 25 were more likely to default on payment.

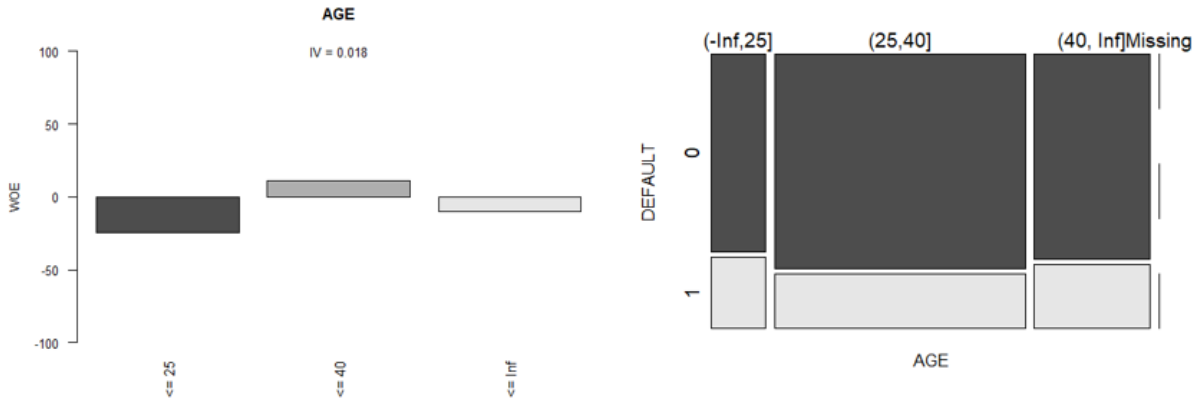

Figure 9 - Optimal age bins based on WOE binning algorithm.

3.2 Payment and Billing History

The data set contains six months of billing and payment history. The feature variables that were engineered from the raw payment and billing history variables are shown in Table 6. The payment ratio (PMT_RATIO) and utilization variables (UTIL) were scaled between [0,100]. If the bill amount was zero, then the payment ratio variable was given a score of 100.

Table 6 - Engineered features for billing and payment history variables.

Feature	Description
AVG_BILL_AMT	The average monthly bill amount over the six months.
AVG_PAY_AMT	The average monthly payment amount over the six months.
PMT_RATIO_APR PMT_RATIO_MAY PMT_RATIO_JUN PMT_RATIO_JUL PMT_RATIO_AUG	This feature looks at the amount the customer repays compared to the bill amount. Note there is a time delay as PAY_AMT is previous month payment. Payment Ratio April = Previous Month Payment May / Bill Amount April PMT_RATIO_APR = PMT_AMT5 / BILL_AMT6 The PMT_RATIO value is then scaled between [0,100].
AVG_PMT_RATIO	The average payment ratio over the five months it can be calculated for.
UTIL_APR UTIL_MAY UTIL_JUN UTIL_JUL UTIL_AUG UTIL_SEP	Determines the amount of the credit line the customer is using each month. Utilization = Current Balance / Credit Limit UTIL_APR = BILL_AMT6 / LIMIT_BAL The utilization value is then scaled between [0,100].
AVG_UTIL	The average utilization over the six months.
BAL_GROWTH_6MO	The balance growth over the six months. BAL_GROWTH_6MO = BILL_AMT1 - BILL_AMT6
UTIL_GROWTH_6MO	The utilization growth over the six months. UTIL_GROWTH_60 = UTIL_SEP – UTIL_APR
MAX_DLQ	Maximum delinquency calculated by taking the maximum of the PAY_R# variables.
MAX_BILL_AMT	Maximum billed amount over the six months.
MAX_PAY_AMT	Maximum payment amount over the six months.

4.0 Exploratory Data Analysis

The summary statistics for the engineered variables are shown in Table 7. The 'Max' and 'Min' columns show that the PMT_RATIO and UTIL variables have been scaled between 0 and 100. The minimum BAL_GROWTH_6MO value is negative which would indicate that a client's balance decreased over the six-month period.

Table 7 - Summary statistics for the engineered variables.

Statistic	N	Mean	St. Dev.	Min	P(25)	Median	P(75)	Max
AGE_18_25	30,000	0.13	0.34	0	0	0	0	1
AGE_26_40	30,000	0.60	0.49	0	0	1	1	1
AVG_BILL_AMT	30,000	44,976.95	63,260.72	-56,043	4,781.3	21,051.8	57,104.4	877,314
AVG_PAY_AMT	30,000	5,341.83	11,060.81	0.00	1,103.62	2,350.00	5,500.00	672,018.20
PMT_RATIO_APR	30,000	31.76	26.85	0	21.2	21.2	21.3	100
PMT_RATIO_MAY	30,000	97.42	1.26	0	97.1	97.1	97.1	100
PMT_RATIO_JUN	30,000	95.41	1.68	0	94.9	94.9	94.9	100
PMT_RATIO_JUL	30,000	16.04	27.32	0.00	7.14	7.14	7.16	100.00
PMT_RATIO_AUG	30,000	17.59	24.89	0.00	10.07	10.07	10.09	100.00
AVG_PMT_RATIO	30,000	51.64	13.65	26.65	46.06	46.07	46.10	100.00
UTIL_APR	30,000	33.88	6.40	0.00	28.12	31.41	38.77	100.00
UTIL_MAY	30,000	20.82	6.03	0.00	15.28	18.73	25.45	100.00
UTIL_JUN	30,000	26.59	5.65	0.00	21.30	24.79	31.32	100.00
UTIL_JUL	30,000	12.10	3.38	0.00	8.89	11.08	15.20	100.00
UTIL_AUG	30,000	23.23	5.20	0.00	18.18	21.75	28.32	100.00
UTIL_SEP	30,000	14.75	5.82	0.00	9.07	13.20	20.49	100.00
AVG_UTIL	30,000	21.90	5.01	12.88	17.03	20.63	26.27	92.44
BAL_GROWTH_6MO	30,000	12,351.57	43,922.42	-428,791	-2,963	923	19,793.8	708,323
UTIL_GROWTH_6MO	30,000	-19.13	4.79	-53.26	-21.52	-19.24	-18.07	55.83
MAX_DLQ	30,000	0.68	1.07	0	0	0	2	8

4.1 OneR

The OneR package was used to implement a one rule classifier for each of the variables. Table 8 shows that by using a single variable in the model the one rule classifier achieved a minimum accuracy of 77.88%. This accuracy score was used as a baseline to determine if the more complex models in Section 5 of the report were performing well. The PAY_R1 variable appears to be the best predictor with an accuracy of 81.96%. The best performing engineered variable was MAX_DLQ with an accuracy of 78.26%.

Table 8 - Accuracy of one rule decision tree classifier models.

Rank	Attribute	Accuracy
1	PAY_R1	81.96%
2	PAY_R2	79.65%
3	PAY_R5	78.98%
4	PAY_R4	78.71%
5	PAY_R3	78.52%
6	PAY_R6	78.36%
7	MAX_DLQ	78.26%
8	BAL_GROWTH_6MO	77.89%
9	PMT_RATIO_APR	77.88%
10	ID	77.88%
10	LIMIT_BAL	77.88%
10	SEX	77.88%
…	…	…
…	…	…
10	MARRIAGE_MD	77.88%

4.2 Correlation Plot

A correlation plot of the predictor variables and the response variable is shown in Figure 10. The plot highlights that there are several predictor variables that are highly correlated with each other. For example, MAX_DLQ is highly correlated with the PAY_R# variables. This is due to them being part of the MAX_DLQ calculation. The variable MAX_PAY_AMT appears to be negatively correlated with the response variable. The PAY_R# and MAX_DLQ variables appear to be the most strongly correlated predictors with the response variable DEFAULT.

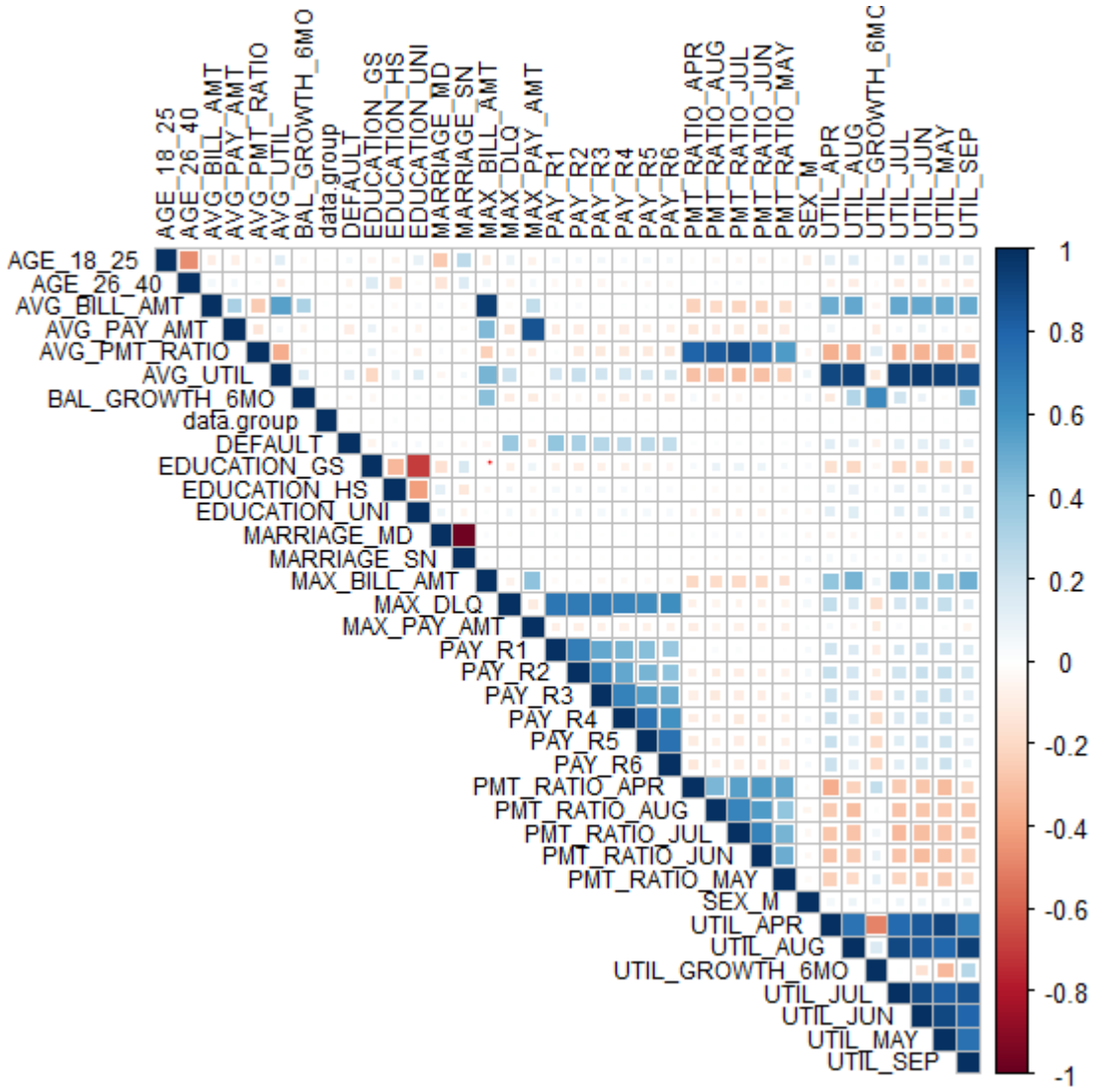

Figure 10 - Correlation plot for modelling variables.

4.3 Decision Tree

The decision tree shown in Figure 10 has been used to identify variables that should be included in the pool of predictor variables to be used in the modelling stage. Decision trees are not the most competitive predictive algorithm however they are easy to interpret. Each node shows the predicted class, the predicted probability and the percentage of observations in the node. The variables that should be included in the modelling phase are PAY_R1, MAX_DLQ, AVG_PMT_RATIO, MAX_BILL_AMT, PAY_R2, UTIL_AUG, MAX_PAY_AMT, PAY_R6, AVG_BILL_AMT, PAY_R5 and AVG_PAY_AMT.

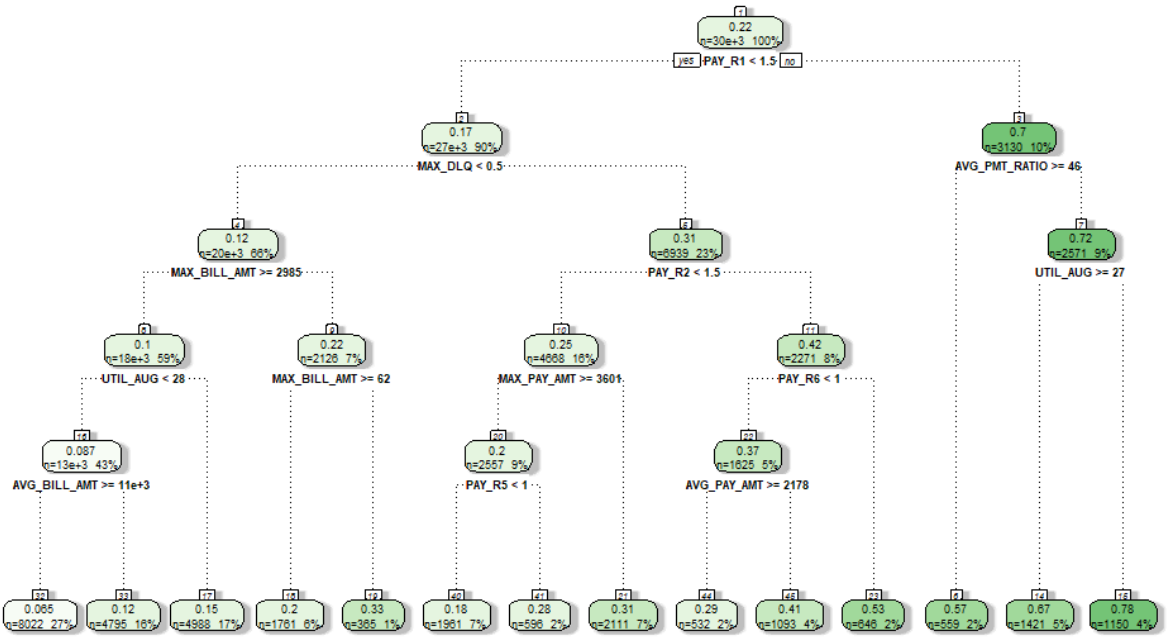

Figure 11 - Classification decision tree with DEFAULT as the target variable.

4.4 Discretisation of Continuous Variables

The WOE binning algorithm was used to discretise the continuous predictor variables that were identified as potentially being useful predictor variables. The variable PAY_R1 (which was coded in Section 2.3.6) was identified as most likely having the highest classifying potential. It tracks the number of months the client's payment is delayed. The results of the WOE binning algorithm are shown in Table 9. The variable was split into three intervals.

Table 9 - PAY_R1 discretised variables interval range based on WOE score.

Interval	WOE Score	Indicator Variable
[-Inf, 0]	57.0	PAY_R1_LE0
[0, 1]	-59.3	Base
[1, Inf]	-208.5	PAY_R1_G1

The mosaic plot in Figure 12 shows that clients with a PAY_R1 value less than or equal to zero (not behind on payments) are less likely to default on payment compared to those with a value greater than zero. Two indicator variables PAY_R1_LE0 and PAY_R1_G1 were created. This method and process was used for the other continuous variables. Details on the WOE scores and the indicator variables that were created can be found in Appendix 1.

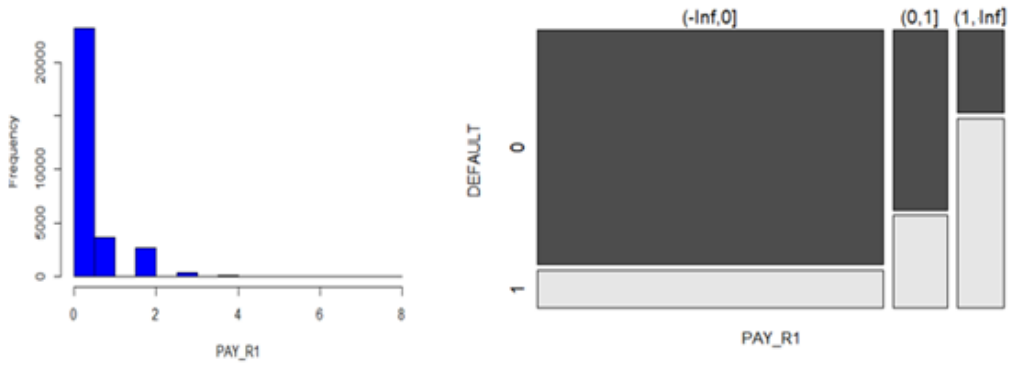

Figure 12 - PAY_R1 distribution and mosaic plot.

Table 10 summarizes the pool of predictor variables to be used in the predictive modelling. They have been split into two pools. The 'Discrete and Binned Continuous Variables' column show the indicator variables that correspond to the continuous variable they were created from.

Table 10 - Summary of variables to be used for predictive modelling.

Discrete and Continuous	Discrete and Binned Continuous Variables
SEX_M	SEX_M
EDUCATION_GS	EDUCATION_GS
EDUCATION_UNI	EDUCATION_UNI
EDUCATION_HS	EDUCATION_HS
MARRIAGE_MD	MARRIAGE_MD
MARRIAGE_SN	MARRIAGE_SN
AGE	AGE_18_25, AGE_26_40
PAY_R1	PAY_R1_LE0, PAY_R1_G1
PAY_R2	PAY_R2_G0
PAY_R3	PAY_R3_G0
PAY_R4	PAY_R4_G0
PAY_R5	PAY_R5_G0
PAY_R6	PAY_R6_G0
AVG_BILL_AMT	AVG_BILL_AMT_LE697, AVG_BILL_AMT_2861_7373, AVG_BILL_AMT_7373_31478, AVG_BILL_AMT_G31478
AVG_PAY_AMT	AVG_PAY_AMT_LE2000, AVG_PAY_AMT_G11842
PMT_RATIO_APR	
PMT_RATIO_MAY	
PMT_RATIO_JUN	
PMT_RATIO_JUL	
PMT_RATIO_AUG	
AVG_PMT_RATIO	AVG_PMT_RATIO_LE46, AVG_PMT_RATIO_46, AVG_PMT_RATIO_46_1
UTIL_APR	
UTIL_MAY	
UTIL_JUN	
UTIL_AUG	UTIL_AUG_LE18, UTIL_AUG_18_21, UTIL_AUG_G24
UTIL_SEP	UTIL_SEP_8_14, UTIL_SEP_G14
AVG_UTIL	AVG_UTIL_LE16, AVG_UTIL_16_22, AVG_UTIL_22_28, AVG_UTIL_G28
BAL_GROWTH_6MO	BAL_GROWTH_6MO_LEN21881, BAL_GROWTH_6MO_N21881_N10172, BAL_GROWTH_6MO_N10172_923

UTIL_GROWTH_6MO	UTIL_GROWTH_6MO_LEN21
MAX_DLQ	MAX_DLQ_G1
MAX_BILL_AMT	MAX_BILL_AMT_LE600, MAX_BILL_AMT_600_4079, MAX_BILL_AMT_4079_18400, MAX_BILL_AMT_18400_21034, MAX_BILL_AMT_G52496
MAX_PAY_AMT	MAX_PAY_AMT_LE168, MAX_PAY_AMT_5000_36621, MAX_PAY_AMT_G36621

5.0 Predictive Modelling: Methods and Results

The follow sections of the report describe the different classification modelling techniques that were used to predict the probability of a client defaulting on their next payment. Five different modelling algorithms were used. They were Random Forest, Gradient Boosted, Logistic Regression, Naïve Bayes and Support Vector Machine. The in sample (training) and out of sample (test) performance of the models were analysed using several metrics. These metrics were accuracy, true positive rate (TPR), false positive rate (FPR) and AUC.

The true positive rate, which is also known as sensitivity, is the proportion of actual positives that are correctly predicted. The false positive rate is the proportion of actual negatives that are predicted as positive. The classification table shown in Table 11 has been provided to visualise how these rates are calculated.

Table 11 - Classification table and formula for TPR and FPR.

		Predicted 0	Predicted 1
Actual	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

$$TPR = TP / (TP + FN)$$

$$FPR = FP / (FP + TN)$$

Accuracy is poorly suited to being used as the single performance metric for a classifier. A commonly used metric to compare out of sample performance of different classifiers is AUC. This metric is calculated by computing the area under the receiver operating characteristic (ROC) curve. ROC curves plot the true positive rate against the false positive rate for all possible cut off values (NCSS, n.d). The cut off value is used to assign each observation to a

class. Observations with a probability greater than the cut off value are assigned to the class equal to one. Observations with scores less than the cut off value are assigned to the class equal to zero. AUC measures the trade-off between selecting as many true positives as possible while avoiding false positives.

5.1 Random Forest

The first classification model that was fitted was a Random Forest model that used all of the predictor variables in the 'Discrete and Binned Continuous Variables' column in Table 10. The importance of the top 30 variables in the model are shown in Figure 13. The larger the MeanDecreaseGini value the more important the variable is. It is clear from this plot that the PAY_R1_G1 variable is by far the most important. This variable indicates that the client has delayed payment by one or more months. Two other variables that stand out are SEX_M and AGE_26_40.

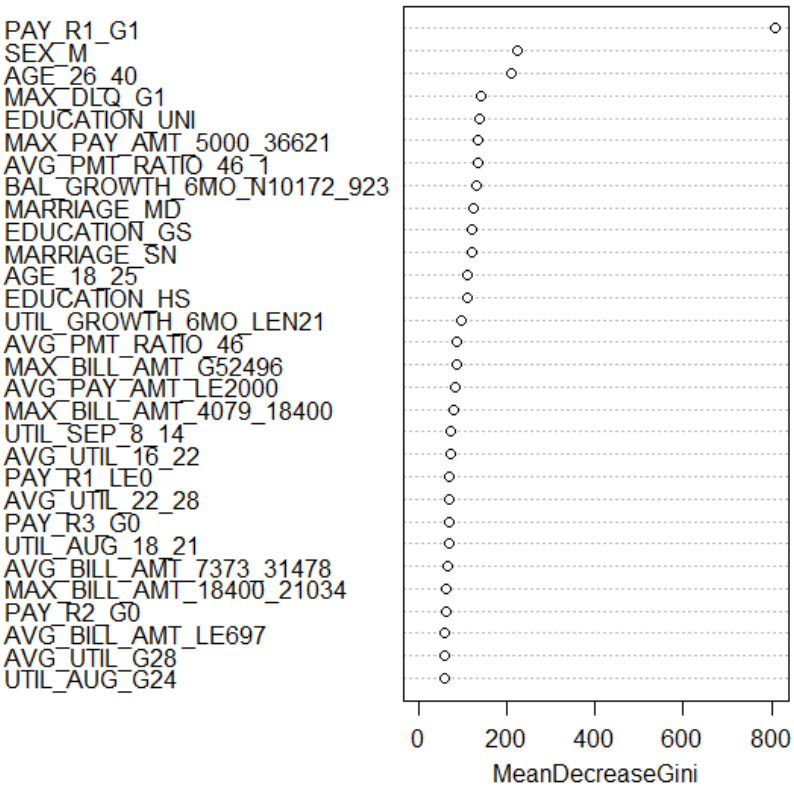

Figure 13 - Random forest model variable importance plot.

The in sample and out of sample performance of the Random Forest model are shown in Table 12. The model predictions for the training sample achieved an accuracy of 94%. However, it may have over fit as there was a large decrease in predictive accuracy on the test set where the accuracy was only 73%. The decrease in AUC between the training and test sets is clearly seen in the ROC curves shown in Figure 14.

Table 12 - Random forest model performance metrics.

	In Sample	Out of Sample
Cut off	0.339	0.269
TPR	0.856	0.600
FPR	0.034	0.237
Accuracy	0.941	0.729
AUC	0.944	0.728

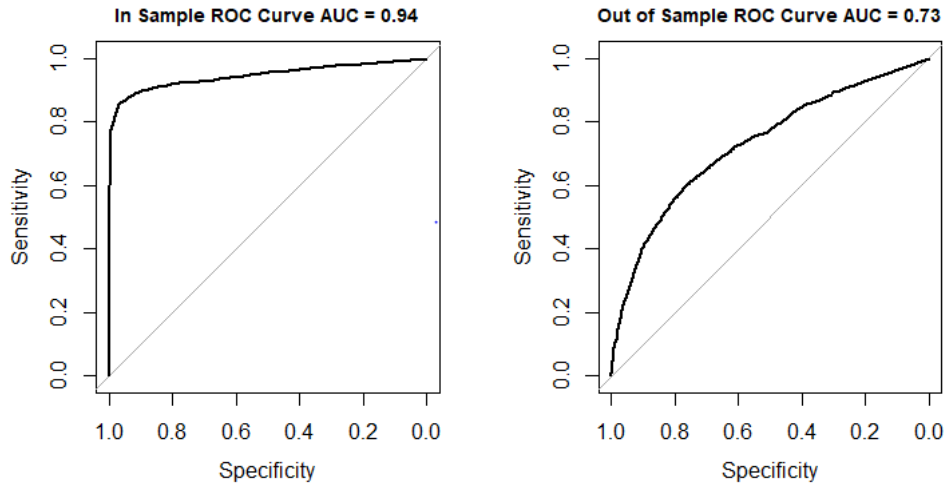

Figure 14 - ROC curve for Random Forest model training and test sets.

5.2 Gradient Boosted Model

The second model that was implemented was a Gradient Boosted model that used the same predictor variables as the Random Forest model. The boosting method has three parameters that can be tuned. They are the number of trees, the shrinkage parameter lambda, which controls the rate at which boosting learns, and the interaction depth which determines the number of splits in each tree. (James, G., Witten, D., Hastie, T., & Tibshirani, R. 2017).

The ‘caret’ package in R was used to conduct a grid search to find the optimal tuning parameters. Using 5-fold cross validation a grid search with interaction depth values of [0,1,2,3,4,5] and number of trees values of [50,100,150,200,250] was implemented. The shrinkage parameter was held constant at a value of 0.1. The parameters that resulted in the best model were number of trees equal to 50 and interaction depth equal to two.

A variable importance plot of the predictor variables was generated and is shown in Figure 15. The most important variable was PAY_R1_G1 again. However, the second and third most important variables from the Random Forest model, SEX_M and AGE_26_40, have very little importance in the Gradient Boosted model.

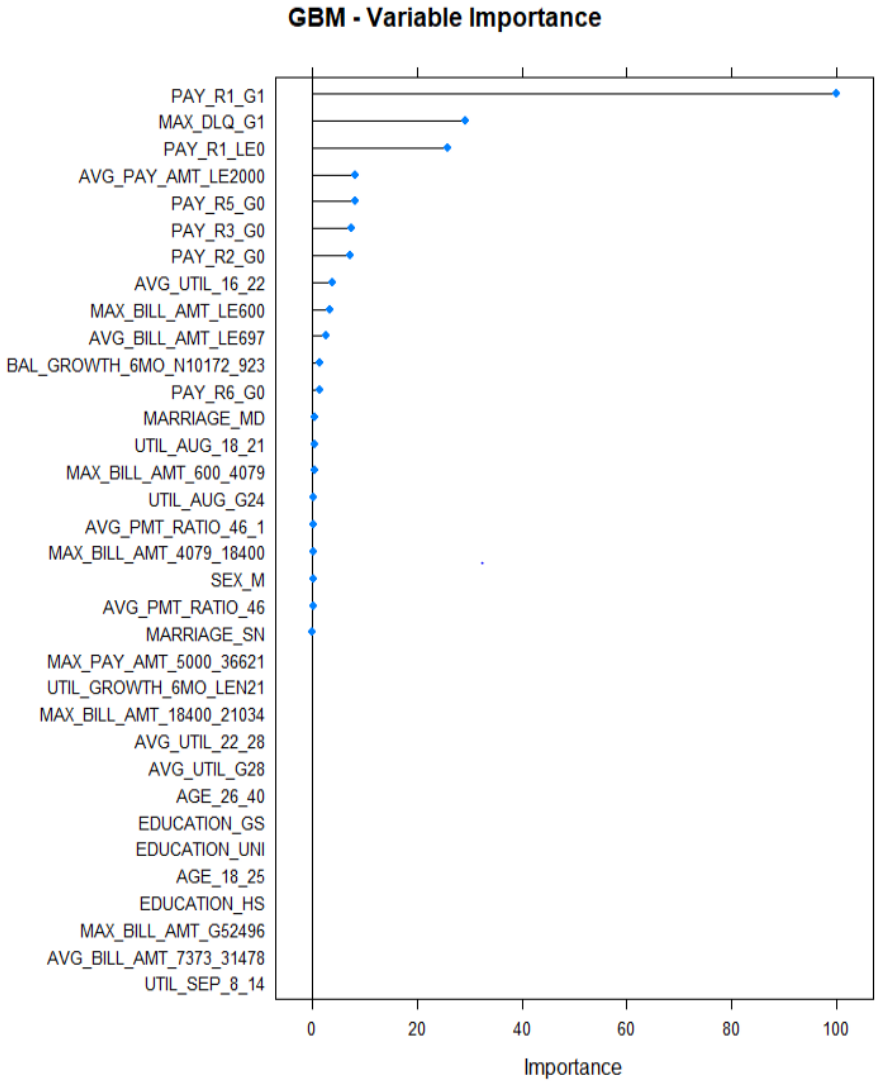

Figure 15 - Gradient Boosted model variable importance plot.

The in sample and out of sample model performance are shown in Table 13. It does not appear as though the training model was overfit as the predictive performance of the model is very similar between training and test sets. The ROC curves shown in Figure 16 appear to be almost identical.

Table 13 - Gradient boosted model performance metrics.

	In Sample	Out of Sample
Cut off	0.254	0.234
TPR	0.602	0.630
FPR	0.182	0.206
Accuracy	0.769	0.759
AUC	0.776	0.776

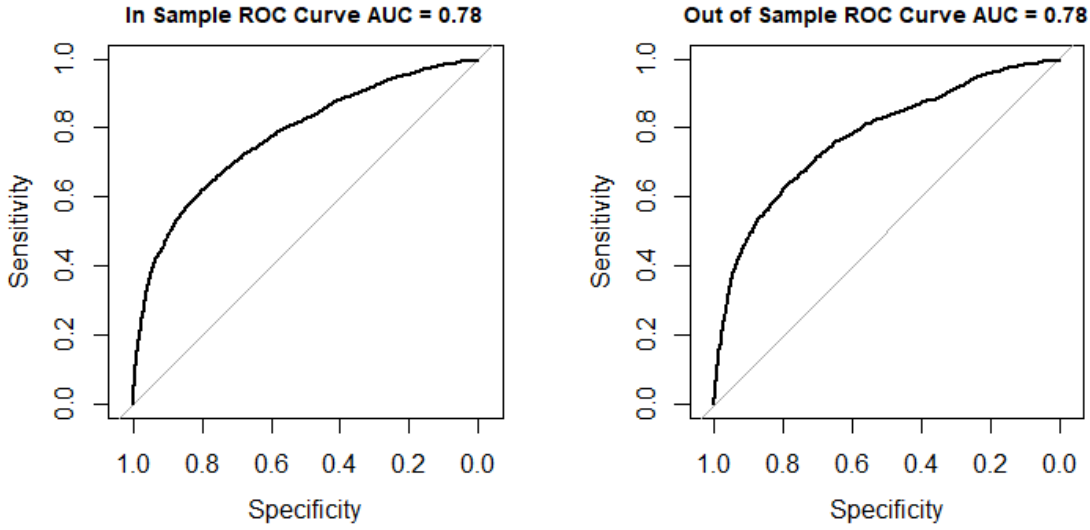

Figure 16 - ROC curve for Gradient Boosted model training and test sets.

5.3 Logistic Regression

A baseline Logistic Regression model was fitted using the top three variables of importance from the Gradient Boosted model. They were PAY_R1, MAX_DLQ_G1 and PAY_R1_LE0. The logit regression coefficients are shown in Table 14. By taking the exponential of the regression coefficients, the odds ratio can be calculated. The probability can then be calculated using the odds ratio. The two formulas below show how this is done.

(1) Odds_Y = exp(Logit_Y)

(2) Prob_Y = Odds_Y/(1 + Odds_Y)

The positive associate of the PAY_R1_G1 coefficient makes logical sense. Delayed payments increase the probability of defaulting. The negative association of the PAY_R1_LEO also makes sense. No delayed payments decrease the probability of defaulting.

Table 14 – Logit coefficient estimates.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.21534    0.06034  -20.14   <2e-16 ***
PAY_R1_G1    1.18736    0.07750   15.32   <2e-16 ***
MAX_DLQ_G1   0.90961    0.05418   16.79   <2e-16 ***
PAY_R1_LE0  -0.73350    0.06061  -12.10   <2e-16 ***
---
```

The in sample and out of sample model performance are shown in Table 15. The ROC curves shown in Figure 17 show that the in sample and out of sample performance was very similar.

Table 15 - Logistic Regression model performance metrics.

	In Sample	Out of Sample
Cut off	0.177	0.177
TPR	0.639	0.663
FPR	0.246	0.256
Accuracy	0.728	0.727
AUC	0.728	0.737

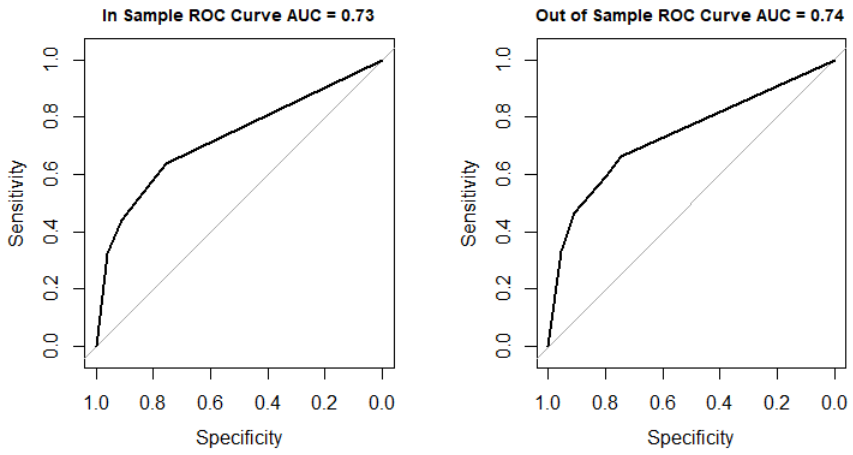

Figure 17 - ROC curve for Logistic Regression model training and test sets.

5.4 Logistic Regression with Backward Variable Selection

A second Logistic Regression model was fitted using a backward variable selection algorithm. The backward selection technique first starts with a model with all the predictors (full model) and iteratively removes variables based on a selection criterion. The stepAIC function in R uses Akaike's Information Criterion (AIC) in its iterative algorithm to decide if a variable should be included in the model. AIC considers the number of parameters used in the model as well as the goodness-of-fit. The full model contained the 21 variables that had importance in the Gradient Boosted model. The final model contained 15 predictor variables. The variables in the final model and their logit coefficients are shown in Table 16.

Table 16 - Logit coefficient estimates for the backward variable selection model.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.25251	0.09481	-23.758	< 0.0000000000000002
PAY_R1_G1	1.35710	0.08183	16.585	< 0.0000000000000002
MAX_DLQ_G1	0.63702	0.07075	9.004	< 0.0000000000000002
PAY_R1_LE0	-0.34898	0.07392	-4.721	0.0000023446534336
AVG_PAY_AMT_LE2000	0.31176	0.05242	5.948	0.0000000027192309
PAY_R5_G0	0.45818	0.07480	6.125	0.0000000009067502
PAY_R3_G0	0.31164	0.07411	4.205	0.0002607426237333
PAY_R2_G0	0.11886	0.08328	1.427	0.15354
MAX_BILL_AMT_LE600	1.11345	0.10837	10.274	< 0.0000000000000002
BAL_GROWTH_6MO_N10172_923	0.11074	0.05161	2.146	0.03191
MARRIAGE_MD	0.15834	0.04420	3.583	0.00034
MAX_BILL_AMT_600_4079	0.63606	0.08743	7.276	0.0000000000003451
UTIL_AUG_G24	0.44047	0.05701	7.726	0.0000000000000111
AVG_PMT_RATIO_46_1	-0.10205	0.05348	-1.908	0.05636
MAX_BILL_AMT_4079_18400	0.31606	0.06670	4.738	0.0000021562288779
SEX_M	0.14318	0.04478	3.197	0.00139

The in sample and out of sample performance metrics are shown in Table 17. The ROC curves shown in Figure 18 reveal that the in sample and out of sample performance are similar.

Table 17 - Logistic Regression model with backward selection performance metrics.

	In Sample	Out of Sample
Cut off	0.201	0.241
TPR	0.654	0.618
FPR	0.230	0.196
Accuracy	0.744	0.764
AUC	0.778	0.778

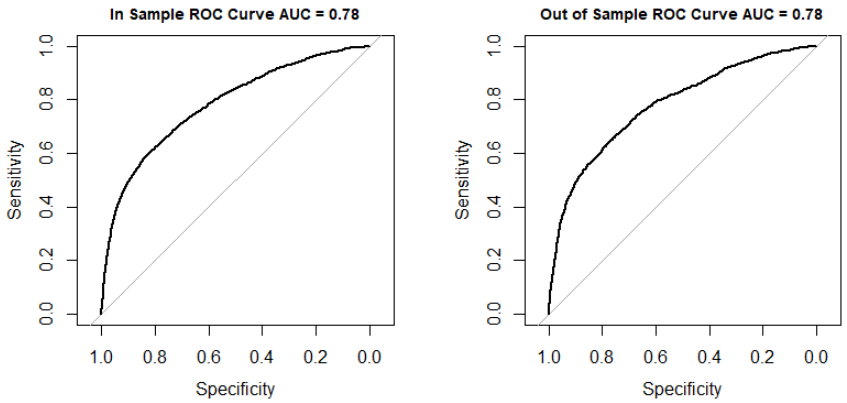

Figure 18 - ROC curve for Logistic Regression with backward selection training and test sets.

5.5 Naïve Bayes

A Naïve Bayes classifier was fitted using the 21 variables that had importance in the Gradient Boosted model. The Naïve Bayes classifier is based on Bayes theorem and calculates conditional and joint probabilities. It assumes that the predictors are independent and the predictors have an equal effect on the outcome. The classifier assigns the observation to the class with the highest conditional probability. The model was fast to train taking only 0.15 seconds. The performance metrics are shown in Table 18. The ROC curves in Figure 19 show that the in sample and out of sample performance are similar.

Table 18 - Naive Bayes model performance metrics.

	In Sample	Out of Sample
Cut off	0.046	0.275
TPR	0.644	0.629
FPR	0.231	0.211
Accuracy	0.741	0.755
AUC	0.768	0.770

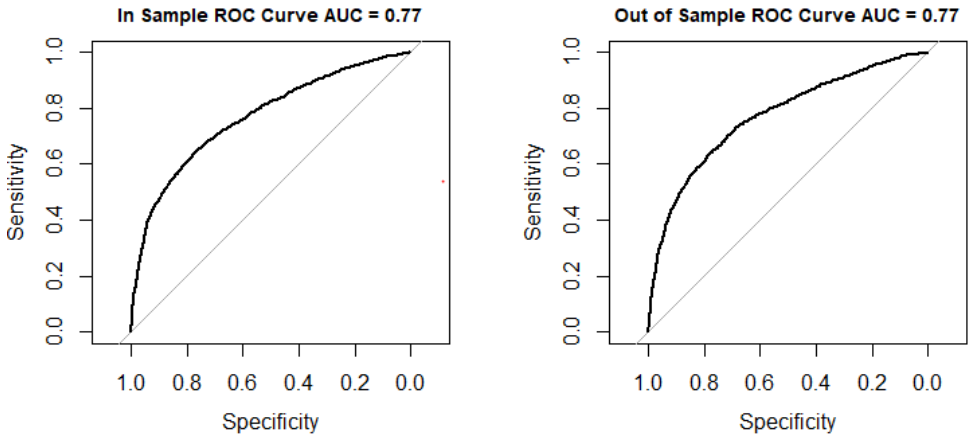

Figure 19 - ROC curves for Naive Bayes training and test sets.

5.6 Support Vector Machine

The final modelling algorithm that was implemented on the data set was Support Vector Machine (SVM). The implementation of SVM was broken into two different sections. Modelling using the discrete variable pool and modelling using the continuous variable pool.

5.6.1 Discrete Variables

The models fitted in this section used the discrete variables that were determined to have importance in the Gradient Boosted model. There are several parameters that can be tuned when fitting SVM models. A SMV model can be fitted with linear or non-linear kernels which influence the shape of the decision boundary between the classes. To determine which type of decision boundary best suited the data set a linear kernel and a radial kernel model were fitted while keeping cost and gamma constant (1 and 0.5 respectively).

Table 19 provides a performance comparison of the two models. Using out of sample AUC as the scoring metric the model with a non-linear decision boundary performed best.

Table 19 - Performance comparison of linear and radial kernels.

Model	In Sample AUC	Out of Sample AUC
Linear Kernel	0.62	0.62
Radial Kernel	0.75	0.69

The next model fitted used a radial kernel while using a grid search technique and 10-fold cross validation to find the optimal parameters for cost and gamma. The results of the grid search are shown in Table 20. The model with the lowest error used values of 10 and 0.5 for cost and gamma respectively.

Table 20 - Performance results for SVM model fitted using parameter tuning.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma
   10   0.5

- best performance: 0.2059947

- Detailed performance results:
   cost gamma     error  dispersion
1   0.1   0.5 0.2179183 0.010023379
2   1.0   0.5 0.2076416 0.007354040
3  10.0   0.5 0.2059947 0.007040843
4   0.1   1.0 0.2204216 0.009742348
5   1.0   1.0 0.2092885 0.008080390
6  10.0   1.0 0.2092885 0.008080390
7   0.1   2.0 0.2208169 0.009289342
8   1.0   2.0 0.2092885 0.008080390
9  10.0   2.0 0.2092885 0.008080390
10  0.1   4.0 0.2208169 0.009289342
11  1.0   4.0 0.2092885 0.008080390
12 10.0   4.0 0.2092885 0.008080390
```

The in sample and out of sample performance metrics for the model that used the optimal parameters are shown in Table 21. The ROC curves are shown in Figure 20.

Table 21 - Discrete variable SVM model performance metrics

	In Sample	Out of Sample
Cut off	0.163	0.262
TPR	0.554	0.478
FPR	0.045	0.134
Accuracy	0.865	0.783
AUC	0.749	0.679

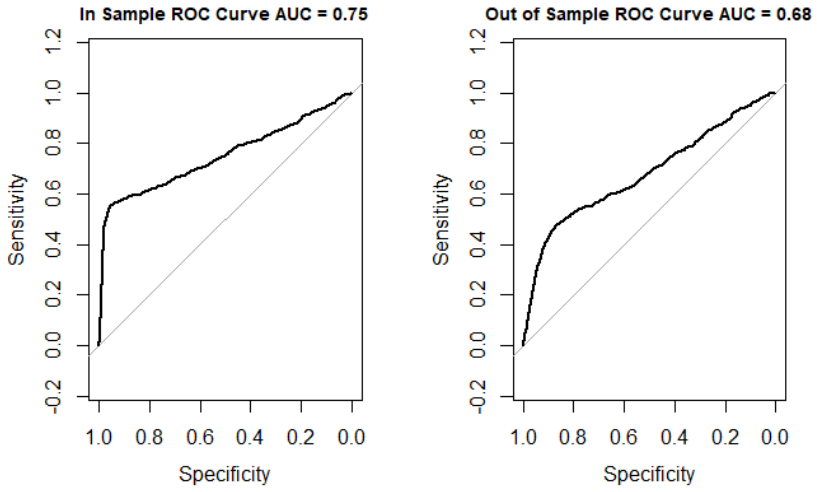

Figure 20 - ROC curves for discrete variable SVM training and test sets.

5.6.2 Continuous Variables

Due to the poor performance of the SVM model using the discrete variables compared to the other modelling algorithms it was decided to try an SVM model using the predictor variables from the 'Discrete and Continuous' column in Table 10, Section 4.4 of the report. The model was fitted using a radial kernel and the best model used values of 1 and 0.5 for cost and gamma respectively as shown in Table 22.

Table 22 - Performance results for SVM model fitted using parameter tuning.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma
    1   0.5

- best performance: 0.212253

- Detailed performance results:
   cost gamma     error  dispersion
1   0.1   0.5 0.2254941 0.008600700
2   1.0   0.5 0.2122530 0.010382550
3  10.0   0.5 0.2337286 0.012863386
4   0.1   1.0 0.2254941 0.008600700
5   1.0   1.0 0.2195652 0.008465079
6  10.0   1.0 0.2444005 0.009913090
7   0.1   2.0 0.2254941 0.008600700
8   1.0   2.0 0.2220685 0.006620107
9  10.0   2.0 0.2494071 0.004587218
10  0.1   4.0 0.2254941 0.008600700
11  1.0   4.0 0.2270092 0.006470933
12 10.0   4.0 0.2484190 0.006988601
```

The in sample and out of sample performance of the SVM model that used the continuous variables is shown in Table 23. There was a 0.027 increase in AUC compared to the SVM model that used discrete variables. The ROC curve shown in Figure 21 reveals that the training model was most likely overfit.

Table 23 - Continuous variable SVM model performance metrics

	In Sample	Out of Sample
Cut off	0.149	0.303
TPR	0.902	0.473
FPR	0.087	0.141
Accuracy	0.910	0.777
AUC	0.926	0.706

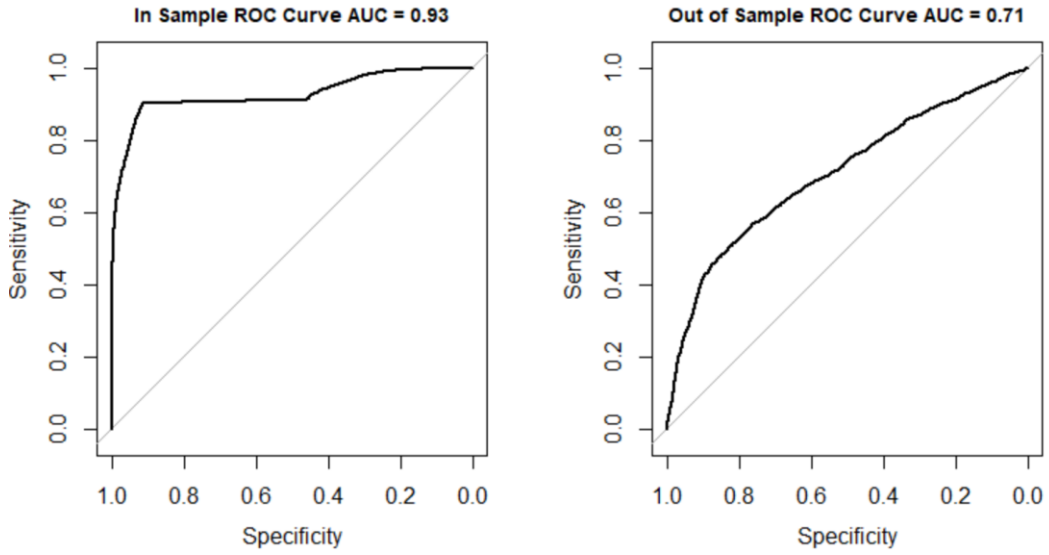

Figure 21 - ROC curves for continuous variable SVM training and test sets.

6.0 Comparison of Results

A summary of the in sample performance of the models fitted in Section 5 is provided in Table 24. The SVM models training times of greater than 12 hours is significantly larger than any of the other models. The fastest model to train was the Naïve Bayes followed closely by the Logistic Regression models. The Random Forest and SVM model that used continuous variables over fit the data.

Table 24 - Comparison of in sample model performance.

Model	Training Time	Accuracy	TPR	FPR	AUC
Random Forest	2min 20s	0.94	0.86	0.03	0.94
Gradient Boosted	10min 30s	0.77	0.60	0.18	0.78
Logistic Regression	0.2s	0.73	0.64	0.25	0.73
Backward Selection	14s	0.74	0.65	0.23	0.78
Naïve Bayes	0.15s	0.74	0.64	0.23	0.77
SVM Discrete Vars	16 hrs	0.87	0.55	0.05	0.75
SVM Contin. Vars	13hrs	0.91	0.90	0.09	0.93

A summary of the out of sample performance of the models fitted in Section 5 is provided in Table 25. Using AUC as the metric to rank the models, the two best performing models were the Gradient Boosted model and the Logistic Regression model with backward variable selection. Both these models achieved an AUC of 0.78. The Naïve Bayes classifier followed closely with an AUC value of 0.77. The Logistic Regression model that was fitted with only

three variables achieved an AUC value of 0.74. The Random Forest model, which used all of the predictor variables only achieved an AUC value of 0.73.

Table 25 - Comparison of out of sample model performance.

Model	Accuracy	TPR	FPR	AUC
Random Forest	0.73	0.60	0.24	0.73
Gradient Boosted	0.76	0.63	0.21	0.78
Logistic Regression	0.73	0.66	0.26	0.74
Backward Selection	0.76	0.62	0.20	0.78
Naïve Bayes	0.76	0.63	0.21	0.77
SVM Discrete Vars	0.78	0.48	0.13	0.68
SVM Contin. Vars	0.78	0.47	0.14	0.71

Figure 22 combines the out of sample ROC curves for all the models. From this plot it can be clearly seen that the SVM model that used the discrete variable pool was the worst performing model. The SVM model that used continuous variables performed better than the SVM with discrete variables however it was still the second worst performing.

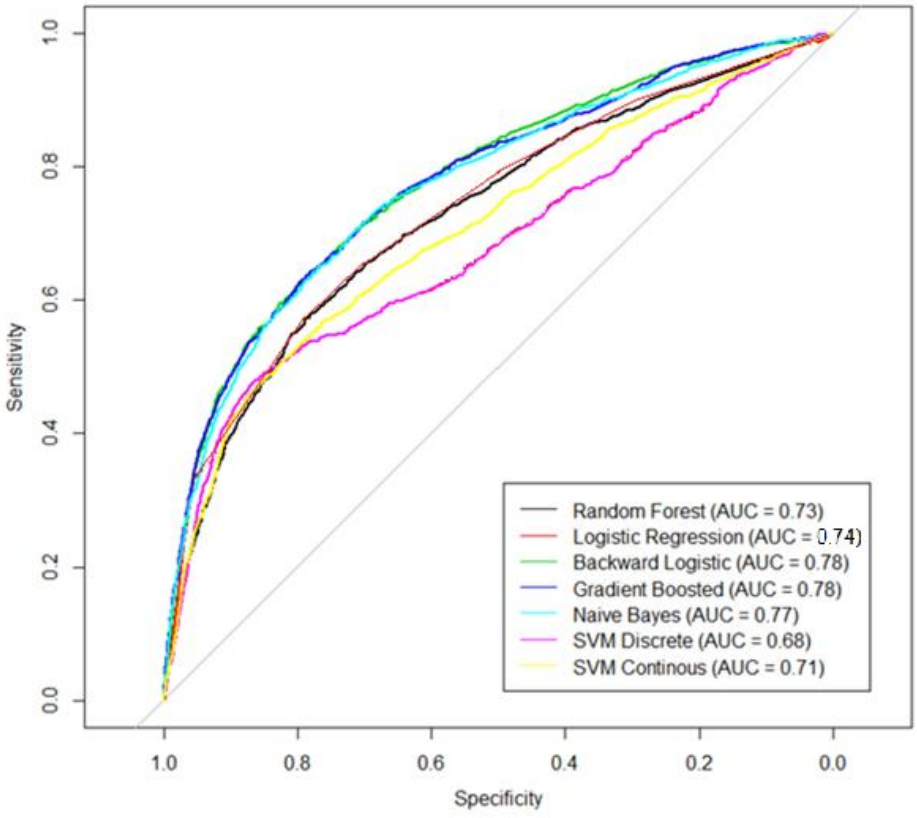

Figure 22 - ROC curves for out of sample model performance.

7.0 Conclusion

This report detailed the process of implementing five different machine learning algorithms to predict the probability of a credit card client defaulting on payment. The data set contained 30,000 observations and 23 predictor variables consisting of customer demographics and billing and payment history. Feature engineering was conducted on the billing and payment history to refine the raw variables into predictors that would be useful for training the model. The weight of evidence supervised binning algorithm was used to discretise continuous variables into optimal bin sizes.

The five machine learning algorithms that were implemented were Random Forest, Gradient Boosted, Logistic Regression, Naïve Bayes and Support Vector Machine. The Random Forest and Gradient Boosted models identified important variables. The most important variable for predicting a default payment was the variable related to if the client had delayed a payment by one or more months. The out of sample performance of the models were ranked using AUC as the scoring metric.

The Gradient Boosted model and the Logistic Regression model that used backward variable selection were the best performing models and were closely followed by the Naïve Bayes model. The Support Vector Machine models were the most computationally intensive and resulted in the worst performing models. Alternative machine learning algorithms that could be fitted in future research include Neural Network or K-Nearest Neighbour.

References

James, G., Witten, D., Hastie, T. J., & Tibshirani, R. J. (2017). *An introduction to statistical learning: With applications in R*. New York: Springer.

NCSS. (n.d.). One ROC Curve and Cutoff Analysis. Retrieved April 25, 2019, from https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/One_ROC_Curve_and_Cutoff_Analysis.pdf

UCI. (2016). Default of Credit Card Clients Data Set. Retrieved April 10, 2019, from https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Appendix 1 – Discretising Continuous Variables

1.1 PAY_R2

Table 26 - Intervals and WOE scores for PAY_R2 variable.

Interval	WOE Score	Indicator Variable
[-Inf, 0]	37.9	Base
[0, Inf]	-149	PAY_R2_G0

1.2 PAY_R3

Table 27 - Intervals and WOE scores for PAY_R3 variable.

Interval	WOE Score	Indicator Variable
[-Inf, 0]	13.3	Base
[0, Inf]	-134.9	PAY_R3_G0

1.3 PAY_R4

Table 28 - Intervals and WOE scores for PAY_R4 variable.

Interval	WOE Score	Indicator Variable
[-Inf, 0]	26.1	Base
[0, Inf]	-140	PAY_R4_G0

1.4 PAY_R5

Table 29 - Intervals and WOE scores for PAY_R5 variable.

Interval	WOE Score	Indicator Variable
[-Inf, 0]	22.8	Base
[0, Inf]	-148.2	PAY_R5_G0

1.5 PAY_R6

Table 30 - Intervals and WOE scores for PAY_R6 variable.

Interval	WOE Score	Indicator Variable
[-Inf, 0]	21.3	Base
[0, Inf]	-135.1	PAY_R6_G0

1.6 MAX_DLQ

Table 31 - Intervals and WOE scores for MAX_DLQ variable.

Interval	WOE Score	Indicator Variable
[-Inf, 1]	66.5	Base
[1, Inf]	-111.0	MAX_DLQ_G1

1.7 AVG_PMT_RATIO

Table 32 - Intervals and WOE scores for AVG_PMT_RATIO variable.

Interval	WOE Score	Indicator Variable
[-Inf, 46.06399026]	-104.7	AVG_PMT_RATIO_LE46
[46.06399026, 46.06627761]	-18.2	AVG_PMT_RATIO_46
[46.06627761, 46.12973924]	43.7	AVG_PMT_RATIO_46_1
[46.12973924, Inf]	3.9	Base

1.8 AVG_BILL_AMT

Table 33 - Intervals and WOE scores for AVG_BILL_AMT variable.

Interval	WOE Score	Indicator Variable
[-Inf,697.28]	-34.8	AVG_BILL_AMT_LE697
[697.28,2861.17]	-0.1	Base
[2861.17, 7373.52]	31.5	AVG_BILL_AMT_2861_7373
[7373.52, 31478.83]	-6.4	AVG_BILL_AMT_7373_31478
[31478.83, Inf]	7	AVG_BILL_AMT_G31478

1.9 AVG_PAY_AMT

Table 34 - Intervals and WOE scores for AVG_PAY_AMT variable.

Interval	WOE Score	Indicator Variable
[-Inf, 2000]	-36.9	AVG_PAY_AMT_LE2000
[2000, 11842.2]	26.7	Base
[11842.2, Inf]	98.0	AVG_PAY_AMT_G11842

1.10 MAX_BILL_AMT

Table 35 - Intervals and WOE scores for MAX_BILL_AMT variable.

Interval	WOE Score	Indicator Variable
[-Inf, 600]	-49.2	MAX_BILL_AMT_LE600
[600, 4079]	-23.9	MAX_BILL_AMT_600_4079
[4079,18400.65]	3.4	MAX_BILL_AMT_4079_18400
[18400.65, 21034]	-25.9	MAX_BILL_AMT_18400_21034
[21034, 52496.15]	-1.4	Base
[52496.15, Inf]	19.3	MAX_BILL_AMT_G52496

1.11 MAX_PAY_AMT

Table 36 - Intervals and WOE scores for MAX_PAY_AMT variable.

Interval	WOE Score	Indicator Variable
[-Inf, 168]	-75.9	MAX_PAY_AMT_LE168
[168, 5000]	-22.9	Base
[5000, 36621.4]	26.8	MAX_PAY_AMT_5000_36621
[36621.4, Inf]	82.5	MAX_PAY_AMT_G36621

1.12 UTIL_AUG

Table 37 - Intervals and WOE scores for UTIL_AUG variable.

Interval	WOE Score	Indicator Variable
[-Inf,18.2]	7.3	UTIL_AUG_LE18
[18.2,21.8]	45.4	UTIL_AUG_18_21
[21.8,24.4]	6.5	Base
[24.4, Inf]	-28.6	UTIL_AUG_G24

1.13 UTIL_SEP

Table 38 - Intervals and WOE scores for UTIL_SEP variable.

Interval	WOE Score	Indicator Variable
[-Inf, 8.8]	-14.7	Base
[8.8, 14.7]	29.1	UTIL_SEP_8_14
[14.7, Inf]	-23.1	UTIL_SEP_G14

1.14 AVG_UTIL

Table 39 -Intervals and WOE scores for AVG_UTIL variable.

Interval	WOE Score	Indicator Variable
[-Inf, 16.61]	-41.1	AVG_UTIL_LE16
[16.61, 16.73]	2	Base
[16.73, 22.94]	39.8	AVG_UTIL_16_22
[22.94, 28.31]	-17.4	AVG_UTIL_22_28
[28.31, Inf]	-54.1	AVG_UTIL_G28

1.15 BAL_GROWTH_6MO

Table 40 - Intervals and WOE scores for BAL_GROWTH_6MO variable.

Interval	WOE Score	Indicator Variable
[-Inf, -21881.5]	66.6	BAL_GROWTH_6MO_LEN21881
[-21881.5, -10172.8]	14.8	BAL_GROWTH_6MO_N21881_N10172
[-10172.8, 923]	-36.3	BAL_GROWTH_6MO_N10172_923
[923, Inf]	27.2	Base

1.16 UTIL_GROWTH_6MO

Table 41 - Intervals and WOE scores for UTIL_GROWTH_6MO variable.

Interval	WOE Score	Indicator Variable
[-Inf,-21.5]	-47.5	UTIL_GROWTH_6MO_LEN21
[-21.5, Inf]	18.8	Base