Model #101: Credit Card Default Model

Model Development Guide

YANINA STRYLETS

1. Introduction

Credit scoring models are widely used by banks and other financial institutions, in order to assess the risk of default of applicants for loans. Credit scoring is a classification problem. Credit scoring models take a vector of attributes for a loan applicant, and given these attributes, attempt to discriminate between Good and Bad cases; that is, to discriminate between those that are not likely to default or be in arrears with their payments, and those that are.

The credit scoring models have been built based on a statistical analysis of past borrowers' characteristics such as demographics and historical financial performance data. The data have been obtained from UCI Machine Learning Repository and contains 30000 observations and 24 raw variables. As it is often the case in credit scoring, the target data is imbalanced. The data have been preprocessed before using in the analysis. Derived variables have engineered by transforming raw variables to obtain a set of more predictive features. The purpose of the models is optimizing credit decisions and make lending more cost-efficient. The present analysis implies cost– efficient as realizing the balance between two conflicting objectives of the lending practice, increasing the expected profit and minimizing the expected loss due to lending to risk default customers. The credit scoring models have been developed using proven supervised classification models, binary Logistic regression, Random forest, Gradient boosting, Stacking model of both Random forest and XGBoost models, and Naive Bayes. The models have been evaluated based on their performances in cross-validation, on train and test data sets according to two discrimination measures Koglomorov-Smirnov test (KS) and AUC (The area under Receiver Operating Curve), F1 score and the sum of True positive and True negative values as a measure of accuracy of a model. The goal was to build the model which performs best in terms of detecting as much as possible credit defaulters while preserving the largest volume of customers. The models' hyperparameters (where applicable) as well as thresholds are fine-tuned using F1 score to realize the tradeoff objective. XGboost was found to be the best performing classifier, outperforming a number of other classifiers.

2. The Data

2.1 Data Description

The credit scoring models are built using the credit card default data set obtained from the UCI Machine Learning Repository. The data consist of 30000 observations and 24 variables which are mix of categorical and numeric variables. The response variable is binary and indicates two possible outcomes, default denoted as 1 and non- default denoted as 0. The data dictionary of the default variables is provided in Table 1.

Table 1. The Data Dictionary

VARIABLE NAME	**DEFINITION**
DEFAULT	Default payment (Yes = 1, No = 0)
LIMIT_BAL	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
SEX	Gender (1 = male; 2 = female)
EDUCATION	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
MARRIAGE	Marital status (1 = married; 2 = single; 3 = others)
AGE	Age (year)
PAY_1 – PAY_6 **(X6 – X11)**	History of past payment. The past monthly payment records (from April to September, 2005) have been tracked as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .; X11 = 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.
BILL_AMT1-BILL_AMT6 **(X12 – X17)**	Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
PAY_AMT1-PAY_AMT6 **(X18 – X23)**	Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

2.2 Data Quality Check

The data quality check reveals inconsistency with the data dictionary which has been treated by remapping invalid values to valid observations. The data have no missing value. The data have been transformed to alleviate the effect of outliers by discretizing, winsorizing or applying power transformation.

Discrepancies are found between the data and what is supposed to be in the data based on the data dictionary. The variables EDUCATION and MARRIAGE contain values which do not align with the data dictionary. The minimum and maximum statistics for EDUCATION variable shown in Table 2. points out that the variable have invalid records, some of observations have 0, 5 and 6 values for the variable which are not legitimate data points.

Table 2: Summary Statistics for Credit Card Default Data

Statistic	*N*	*Mean*	*St. Dev.*	*Min*	*Pctl(25)*	*Median*	*Pctl(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

Similarly, MARRIAGE variable is supposed to have 3 levels, but the data quality check reveals that some observations have 0 values which are not defined in the dictionary. To remedy these data issues the invalid data points have been replaced with NA and predicted using Multivariate Imputation by Chained Equation leveraged in MICE imputation package and classification and regression trees model (CART) as an imputation engine. Other methods are also tested to perform imputations, but CART produced the most plausible values, the distribution of CART imputations resembles distribution of the observed data the most. Thus, the invalid values for EDUCATION and MARRIAGE variables have been mapped to valid values. The inspection of PAY variable also revealed that the variable has values which are not defined in the data dictionary or the values are different from the levels defined in the data dictionary. The records of the variable have -2, and 0 values which are not in data dictionary and the values span between -2 and 8 while they are supposed to lie within the range between -1 and 9. The calculations of PAY variable described in data dictionary have been used to find the meaning of the undefined values. It has been observed that -2 and -1 values indicate the balance repayment in full, and observations with 0 values are the records with partially repaid balance statements. According to the description of the PAY variable in the dictionary, -1 values indicate that the previous month balance has been currently repaid either partially or fully. For example, if the outstanding balance for August is $500 has been repaid in

September, the repayment status in August should be -1. The same underlying characteristics of repayment status have records with – 2 values that assumes BILL_AMT _n <= PAY_AMT(n-1) and PAY_AMT(n-1) is not equal to zero unless BILL_AMT _n is equal to zero (some payment has been made), thus -2 values have been remapped to -1. The Table 3. demonstrates the distribution of the PAY__n variable {n = 2,3,4,5,6} conditioned by BILL_AMT _n <= PAY_AMT(n-1) which illustrates that the vast majority of observations that met the above-mentioned condition have either -2 and -1 values.

Table 3. The frequency table of Pay variable under condition that the balance has been paid in full

	-2	-1	0	1	2	3
PAY_2	36.6	59.39	3.08	0.21	0.71	0.01
PAY_3	38.55	56.73	3.38	NA	1.34	NA
PAY_4	41.29	54.53	3.14	NA	1.03	0.01
PAY_5	42.74	52.3	4.4	NA	0.57	NA
PAY_6	43.05	50.34	6.39	NA	0.22	NA

Note: PAY_1 variable is not included since the data don't have the payment records for October.

Payment status of 0 indicates that BILL_AMT_n > PAY_AMT(n-1) but PAY_AMT(n-1) is not necessarily ZERO. In other words, the payments have been made but partially. For example, out of 15730 observations belonging to class 0 in the variable PAY_2, 15418 have made some sort of payment of previous balance, but not necessarily full amount and the rest 312 paid 0 but the previous balance was either 0 or negative. All of this suggests that the 0 values in the variable PAY_2 represent a duly payment but not in full amount. The Table 4. shows the distribution of the observations that met the above-mentioned condition across PAY variable levels verifying that the 0 values in the variable means that the balance has been paid but not in full amount.

Table 4. The frequency table of Pay variable under condition that the balance has been partially repaid

	-2	-1	0	1	2	3	4	5	6	7
PAY_2	0.64	0.6	86.71	0.04	11.37	0.49	0.15	0.01	0.01	NA
PAY_3	0.47	0.34	87.48	0.02	11.28	0.28	0.11	0.03	NA	NA
PAY_4	0.44	0.29	89.75	0.01	9.13	0.26	0.09	0.03	0.01	0.01
PAY_5	0.47	0.41	91	NA	7.81	0.21	0.07	0.02	NA	0.01
PAY_6	0.44	0.64	90.16	NA	8.45	0.15	0.09	0.02	0.04	NA

Note: PAY_1 variable is not included since the data don't have the payment records for October.

As a result of data quality check, the values -2 in PAY variable have been remapped to -1 and the dictionary of the variable has been overwritten as follows: -1 = pay duly in full amount; 0 = partial balance repayment; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months. Finally, the data exploration reveled that the data is purely numeric while according to data dictionary some of the variables such as GENDER or SEX are supposed to be of categorical type. Thus, the variables have been converted to appropriate type.

2.3 Train, test and validate data sets

The data has been split into train, test, and validate data sets as follows:

Table 5: The composition of train, test and validate data sets

	# of observations	% distribution
Train	15,180	50.60
Test	7,323	24.41
Validate	7,497	24.99

The training and testing data sets are used to build and validate the performance of the credit card models in sample and out sample. The validation data set is used for performance validation at the monitoring stage.

3. Feature Engineering

The raw variables except for SEX, MARRIAGE, EDUCATION, LIMT_BAL and the target variable have been used to engineer more informative predictors of the credit default event and excluded from the further analysis and modeling process. The Table 6. contains a new set of engineered predictors and their definition:

Table 6. The dictionary of engineered predictors

VARIABLE	DEFINITION	DESCRIPTION
AGE	Age (Age_18_25, Age_26_40, Age_41_100) – The variable has been discretized using Weight Of Evidence (WOE) binning.	According to WOE results, the customers within the age group of 26 – 40 are the most financially stable while younger and older customers have a higher likelihood to default.
Avg_Bill_Amt	Average Bill Amount Average Bill Amount is created by averaging the monthly bill amount (expenditure) over the six months.	In theory, people who spend more default more.
Avg_Pmt_Amt	Average Payment Amount Average Payment Amount is created by averaging the monthly pay amount (repayment history) over the six months.	Average payment amount can be used as a proxy for income or ability to pay.
Pmt_Ratio_2 - Pmt_Ratio_6	Payment Ratio Payment Ratio is created by calculating ratio of payment amount to bill amount as follows: Pmt_Ratio_n = PAY_AMT(n-1)/ BILL_AMT_n to account for a time delay of payments in the data. For the observations with payment amount exceeding the amount owed or balance amount is less or equal to zero, value of 1 have been assigned indicating that the balance has been repaid fully.	The variable can be used as an indicator of how much of each bill does the customer pay each month and either he/she makes partial or full balance repayments.
Avg_Pmt_Ratio	Average Payment Ratio Average Payment Ratio is created by averaging the monthly payment ratio over the six months.	The variable can be used as an indicator of a customer payment behavior.

Util_1- Util_6	Utilization Current Balance / Credit Limit Util_1 = fraction of credit used in September 2005; Util_2 = fraction of credit used in August, 2005; . . .; Util_6 = fraction of credit used in April, 2005. For the instances where current balance exceeds Credit Limit, values of 1 have been assigned to indicate that the limit has been reached, and 0 values have been assigned for the observations with negative Current Balance to indicate 0 % usage of credit card on the current month. The above- mentioned manipulations have been made to make sure that the variable is bounded between 0 and 1 and represent the percentages and do not have extreme outliers. If it is not done the range of the variable spans between -1.40 and 10 with the vast majority still lying between 0 and 1. Additional variable has been created to flag the customers who exceeds their credit limit occasionally in order to not lose this piece of information as a result of truncating Utilization variable.	The variable indicates how much of the credit line is the customer using.
Avg_Util	Average Utilization Average Payment Ratio is created by averaging the monthly credit utilization over the six months.	The variable tells how much of the credit line has been used on average for the period of 6 months. The variable can reveal the credit habit of a customer.
OVER_LIMIT	The number of months closed with over limit. The variable is calculated by summing the months with balance amount exceeding credit limit.	The variable identifies the number of months with balance amount higher than credit limit. Customer who tends to go over their limit might have a greater chance to default.
Bal_Growth_6mo	Balance Growth Over 6 Months The variable is created by subtracting the beginning (April) amount owed to the bank from ending (September) balance and dividing it by original (April) balance. Thus, the variable measures the growth of the balance over six months compared to the original balance. The original balances of 0 have been replaced with 1 value to avoid dividing by zero.	Is the balance growing due to continued spending with only partial payments each month?
Util_Growth_6mo	Utilization Growth Over 6 Months	Is the balance getting close to the credit limit?

	The variable is created by calculating the difference (change) in credit usage between September and April, beginning and ending months of the 6-month period (Util_1 – Util_6).	
Max_Bill_Amt	Maximum billed amount over the 6 months.	If credit card limit aspect is disregarded, the variable could be used to identify the consumers who spend more, and thus might have a greater likelihood to go into debt.
Max_Pmt_Amt	Max Payment Amount Maximum payment amount over the 6 months.	If credit card limit aspect is disregarded, the variable could be used to identify the consumers who pay more, and thus are more financially stable.
Max_Util	Maximum Utilization Maximum credit utilization over the 6 months.	Average utilization alone could not represent the real credit utilization behavior of the customers. Max utilization might help to capture the truth more accurately especially in the situations where customers average utilization is low because they previously reached their credit card limit and that is why using their credit cards significantly less.
Min_Util	Minimum Utilization Minimum credit utilization over the 6 months.	
OVER_PMT	Overpayment over the 6 months. The binary variable indicates whether the customer has ever paid more than it is due causing the balance to become negative.	The customers who overpay what is due might be more financially responsible and less likely to default.
MAX_Util_ratio	Ratio of current utilization to maximum credit utilization. Is a customer reaching his/her limit.	
Freq_PAY	The most frequent repayment status for a customer assigned by the bank over 6 months. The variable is created by finding the mode of PAY variables. The woe. Binning suggested to bin the variable into two categories: customers with most frequent previous payment status 0 and -1 into one category (Z_DULY_PAY), and the observations with any number of months of delay in another category - DELAY_PAY.	If a customer makes most of the time just partial payments, he/she is less reliable customer than the one who pays off the bill every month but less risky than the customers who constantly late to repay the balance.

REPAY_PATTERN	The variable categorizes the bank customers according to their balance repayment patterns as follows: 0 - no obvious repayment pattern with maximum two occasions of balance growth but no sign of continuous non-repayment or partial repayment pattern; 1 - the balance grew for two consecutive months; 2 – the balance grew for four but not consecutive months; 3 - the balance grew for more than two consecutive months; 4 –the balance grew every other month in a altering sequence manner.	The variable might reveal the customer credit repayment pattern that will be a sign of the risky behavior. For example, category 1 might identify the bank clients who did not repay the balance or paid non-significant amount compared to the amount owed which caused the balance to steadily grow for two months. Category 3 suggests consistent partial repayment of the balance which let the balance grow continuously for 3 or more months which might be a sign of financial distress or irresponsible credit behavior. Category 4 might indicate inconsistent income or the customer financial irresponsible behavior.
UTIL_PATTERN	Utilization pattern variable encodes different customer' credit utilization patterns. 0 - the utilization increased twice in non- continuous manner over the 6 months period; 1= the utilization increased for two consecutive months; 2 = the utilization increased two times for two consecutive months over the 6 months; 3- the utilization has been steadily growing for more than 3 consecutive months.	The customer behavior in terms of credit card utilization might contain information on financial stability and financial responsibility of the credit consumers.

4. Exploratory Data Analysis

4a. Traditional EDA

The summary statistics table reveals no serious issues with the data. Some of the numeric variables such as LIMIT_BAL, Avg_Bill_Amt, Avg_Pmt_Amt, Max_Bill_Amt , Bal_Growth_6mo have a large difference between their 75th percentiles and maximum values which suggest the presence of outliers and right skewness of the variables.

Table 7: Summary Statistics for train data with engineered features

Statistic	*N*	*Mean*	*St. Dev.*	*Min*	*Pctl(25)*	*Median*	*Pctl(75)*	*Max*
DEFAULT	15,180	0.23	0.42	0	0	0	0	1
LIMIT_BAL	15,180	168,064.70	130,547.60	10,000	50,000	140,000	240,000	800,000
Avg_Bill_Amt	15,180	44,934.91	63,150.01	-56,043	4,789.1	21,198.4	56,880.5	592,432
Avg_Pmt_Amt	15,180	5,255.11	10,150.14	0.00	1,111.75	2,389.42	5,554.42	627,344.30
Pmt_Ratio_2	15,180	0.40	0.44	0	0.04	0.1	1	1
Pmt_Ratio_3	15,180	0.41	0.44	0	0.04	0.1	1	1
Pmt_Ratio_4	15,180	0.40	0.45	0	0.04	0.1	1	1
Pmt_Ratio_5	15,180	0.40	0.45	0	0.04	0.1	1	1
Pmt_Ratio_6	15,180	0.43	0.45	0	0.04	0.1	1	1
Avg_Pmt_Ratio	15,180	0.41	0.40	0.00	0.05	0.19	0.82	1.00
Util_1	15,180	0.41	0.39	0.00	0.02	0.32	0.83	1.00
Util_2	15,180	0.40	0.38	0.00	0.02	0.30	0.81	1.00
Util_3	15,180	0.39	0.37	0.00	0.02	0.28	0.75	1.00
Util_4	15,180	0.36	0.36	0	0.02	0.2	0.7	1
Util_5	15,180	0.33	0.34	0	0.01	0.2	0.6	1
Util_6	15,180	0.32	0.34	0	0.01	0.2	0.6	1
Avg_Util	15,180	0.37	0.34	0.00	0.03	0.29	0.68	1.00
OVER_LIMIT	15,180	0.28	0.87	0	0	0	0	6
OVER_PMT	15,180	0.50	1.17	0	0	0	0	5
Bal_Growth_6mo	15,180	2,187.36	14,720.96	-70,532	-0.2	0.1	1.4	388,897
Util_Growth_6mo	15,180	0.10	0.27	-1.00	-0.03	0.01	0.17	1.00
Max_Util	15,180	0.48	0.39	0.00	0.07	0.43	0.92	1.00
Min_Util	15,180	0.26	0.31	0	0	0.1	0.5	1
MAX_Util_ratio	15,180	0.70	0.38	0	0.4	0.9	1	1
Max_Bill_Amt	15,180	60,425.45	77,746.88	-2,900	10,050.8	31,587.5	79,119.5	823,540

Max_Pmt_Amt	15,180	15,620.73	35,279.24	0	2,195.8	5,000	12,200.8	1,215,471
Max_DLQ	15,180	0.68	1.07	0	0	0	2	8
Freq_PAY	15,180	-0.18	0.88	-1	-1	0	0	8
REPAY_PATTERN	15,180	1.23	1.38	0	0	1	3	4

4a.a. Univariate EDA and Transformations:

CONTINUOUS VARIABLES

Since one of the methods used to predict the risk of customer default is logistic regression, which is parametric and sensitive to outliers, some of the predictor variables with abnormal distribution have been transformed to correct for influential observations. While logistic regression does not expect assumption of the multivariate normality of the data to be hold multivariate normality yields a more stable solution. The Table 8. summarizes the reasons, transformations applied to the certain continuous variables and new variables obtained as a result of corrective actions. To see the distributions of the continuous variables before and after transformation refer to Figure 1. through Figure 12, Appendix A 1.

Table 8. Transformation of continuous predictors

Variable	Reason for Transformation	Transformation applied	New Variable
LIMIT_BAL	LIMIT_BAL variable has 83 outliers which is reflected in the highly right skewed histogram of the variable (see Figure 1.). Very few records have credit limit greater than 600000.	To correct for the abnormal distribution of the variable and prevent the negative effect of outliers on the regression analysis, two new variables have been constructed by binning using weight of evidence algorithm and taking the square root of the original variable.	LIMIT_BAL_bin has three levels: low – (10000 – 30000], customers with low credit limit; average (30000 – 160000] – customers with average credit limit and greater than 160000 are clients with high credit limits. Sqrt_LIMIT_BAL created by taking the square root of the original variable to correct for right skewness and downweigh the effect of outliers (see Figure 2.).
Avg_Bill_Amt	The long right tail of the variable and 1298 outliers including 430 e xtreme outliers of the variables (s ee Figure 3.).	The issue has been handled b y winsorizing the variable to 1st and 99 percentiles and app lying Yeo-Johnson transform	Avg_Bill_Amt_tfm The variable might be more suitable to use in logistic regression since the statistical method is parametric and assumes the normal

		ation[1] with lambda 0.2. Woe binning suggested no differen ce in performance of a target variable for negative values a nd 0s of Avg_Bill_Amt, and f or values greater than 302000 and values of 99th percentile. The new variable.	distribution Figure 4. shows the significant improvement of the variable' distribution after corrective actions have been taken.
Max_Bill_Amt	The large number of outliers (1137 and 373 extreme outliers), and the right skewness of the Max_Bill_Amt variable advised to transform the variable (see Figure 5.).	The variable has been truncated to 1st and 99th percentiles and transformed using Yeo-Johnson transformation with lambda 0.2.	Max_Bill_Amt_tfm Figure 3. depicts the distribution of the variable before and after transformation (See Figure 6.).
Avg_Pmt_Amt	The variable is highly right skewed (see Figure 7.).	Avg_Pmt_Amt variable has been transformed using 0 lambda Yeo-Johnson transformation.	Avg_Pmt_Amt_tfm The transformation considerably improved the right skewed distribution of the variable (see Figure 8.)

Bal_Growth_6mo	The significance difference in mean (2,187.36) and median (0.01) values and large standard deviation (14,720.96) of Bal_Growth_6mo variable (see Table 7.) and the presence of 2810 outliers where 2390 of them are extreme outliers strongly suggest that the variable should be transformed. The histogram of the predictor variable and the boxplot of Bal_Growth_6mo vs DEFAULT clearly demonstrates the serious issue with the distribution of the variable (see Figure 9. And Figure 10.).	The reason of such abnormal distribution of the engineered variable is that some records had 0 beginning balance and large positive balance on September which resulted in extremely large percentage of change. Since the percentage increase could not be legitimately calculated against zero starting point, the variable has been binned using woe.tree.binning .	Bal_Growth_6mo_bin has three categories:
			Large Negative Growth - (-Inf,-0.37];
			Small Growth (-0.37,0.13];
			Large Positive Growth - (0.13, Inf].

Max_Pmt_Amt	Max_Pmt_Amt variable has 1950 outliers and highly right skewed distribution (see Figure 11.).	The variable has been transformed using Yeo-Johnson transformation with lambda 0.	Max_Pmt_tfm exhibits the improved distribution after transformation (see Figure 12.).
Avg_Util	Since some of the modeling techniques to be used in the analysis generally preferred discretized variables, additional binned variable has been created for Avg_Util.	Avg_Util has been binned using WOE technique.	Avg_Util_bin has the following levels: No_Util – the customers with utilization 0; Small_Util – the customers with credit utilization <= 0.37; Mod_Util – the customers with credit utilization <= 0.82;

[1] This is the Box-Cox transformation of *U+1* for nonnegative values, and of *|U|+1* with parameter *2-lambda* for *U* negative.

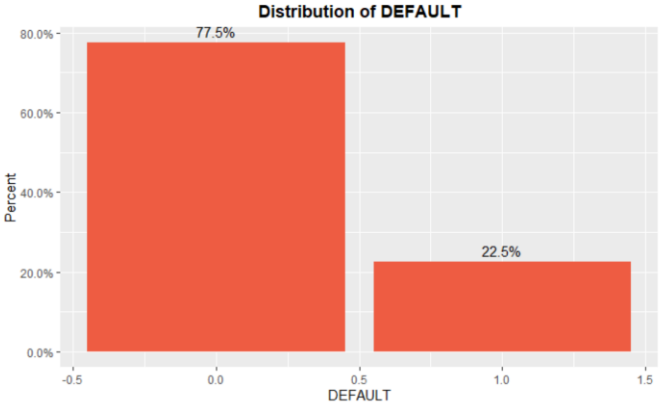
			High_Util – the customers with credit utilization >.82 <= 1.
Util_Growth_6mo	A large number of outliers is detected (see Figure 12.).	The variable has been binned using WOE algorithm.	Util_Growth_6mo_bin variable has the following levels: Large Negative Growth - (-1,-0.2]; Small Negative Growth - (-0.2,-0.03]; Mod Negative Growth - (-0.03,0]; Mod Positive Growth - (0,0.26]; Large Positive Growth – (0.26 , 1].

The variables transformed as a result of EDA observations and the original variables have been preserved in the data to examine which ones would perform best in various classification algorithms.

DESCRETE/ CATEGORICAL VARIABLES

The response variable has unbalanced classes with the majority of the observations belonging to 0 class which is a normal phenomenon for credit scoring response variable (see Figure 13.).

Figure 13.



Most of the categorical predictor variables have relatively even distributions across their classes. However, the predictors with heavily unbalanced distributions have been transformed using binning. The Table 9. summarizes the reasons and transformations undergone by categorical predictors. To see the distributions of the categorical variables before and after transformation refer to Appendix A.2, Figure 14. through Figure 19.

Table 9. Transformation of the categorical predictors

Variable	Reason for Transformation	Transformation applied	New Variable
OVER_LIMIT	The uneven distribution of the observations across classes with almost 87% of records belonging to 0 class (see Figure 14.).	Recategorized to a binary variable. Weight of evidence algorithm suggested that as far as a response variable is concerned the categories 1 and greater perform similarly.	The new binary predictor, OVER_LIMIT indicates whether a customer has ever gone over his/her credit limit during 6-month period (see Figure 15.).
Max_DLQ	The variable has been binned due to uneven distribution of the records across its levels (see Figure 16.).	WOE binning revealed no difference in risk of default for delinquency status 1 and greater than 1 suggesting combining all the levels greater than 1.	As a result of transformation Max_DLQ has two classes, '1' and '2' , 1- duly paid no more than one month delay in balance payment, 2 – delayed payment for more than 1 month (see Figure 17.).
REPAY_PATTERN	The variable needs to be corrected for the uneven distribution of the records across its categories with only 1% of observations belonging to class 2 (see Figure 18.).	WOE binning applied. 1, 2 and 3 categories are combined	REPAY_PATTERN_bin has three levels, 0; 1-3 (1, 2 and 3 categories combined) and category 4.
REPAY_PATTERN	The variable needs to be corrected for the uneven distribution of the records across its categories with only 1% of observations belonging to class 2 (see Figure 19.).	WOE binning applied. 1, 2 and 3 categories are combined	UTIL_PATTERN_bin has category 0 and 1-3 which is a result of 1, 2 and 3 classes being combined.
MARRIAGE		The variable has been releveled to make OTHER class to be a reference category for convenience of interpreting the regression coefficients.	
EDUCATION		The variable has been releveled to make OTHER class to be a reference category for convenience of interpreting the regression coefficients.	

The original and binned variables are used in the modeling to see which variables would perform best in terms of predicting the outcome.

4a.b. Bivariate analysis. Correlation with the response variable and Multicolliniarity

Based on the examination of correlation matrix of continuous predictor variables (see Figure 20.) the folowing observations have been made:

1. Pearson coefficient does not exceeds 0.17 in any direction of linear relationship indicating that none of the continuous variables explain more than 17 % of variance in the response variable (see Figure 10.). Similar results produced correlation plot based on spearman correlation coefficient (see Appendix B) which does not assume parametric distribution of the variables and measure not only linear relationship but monotonic relationship.
2. Transformed variables have similar Spearman (see Appendix B) and Person coeffcients which suggest that the relationship between the reponse variable and predictors have been straightened after transformation.
3. Transformed variables have a higher degree of association with a response variable according to Pearson which suggests to use transformed variables in the further analysis at least in case of logistic regression which assumes the linear relationship of reaponse variable and logits of predictor variables.
4. The multicolliarity is present in the data.

The correlation matrix with discrete variables reveal the same pattern of multicollinearity and relatively weak association with a response variable (See Figure 21.).

Figure 20. Correlation matrix of continuous variables (Pearson)

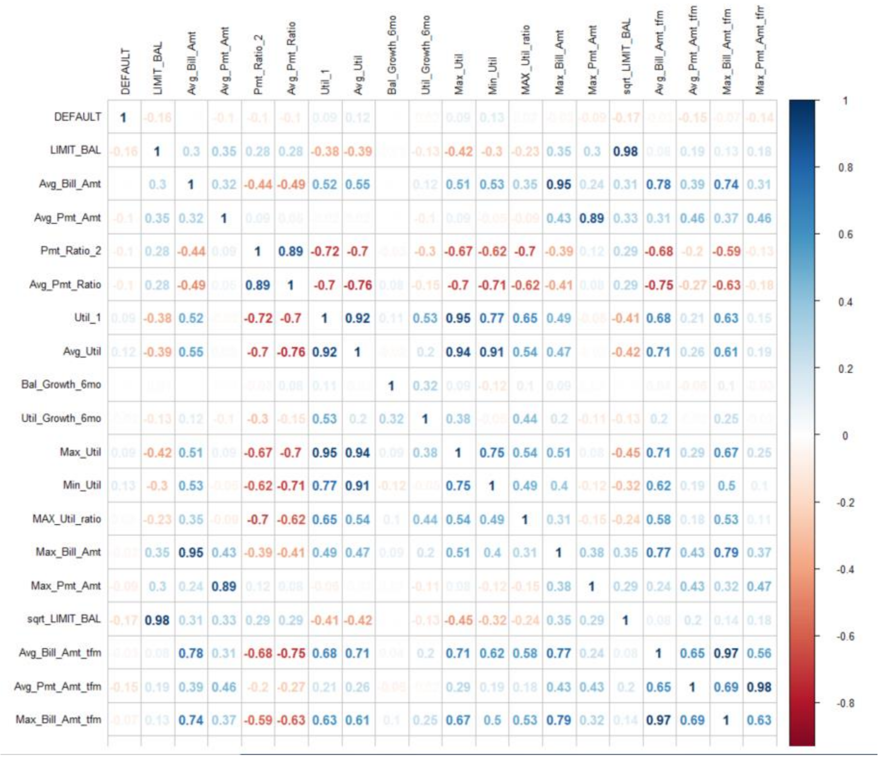
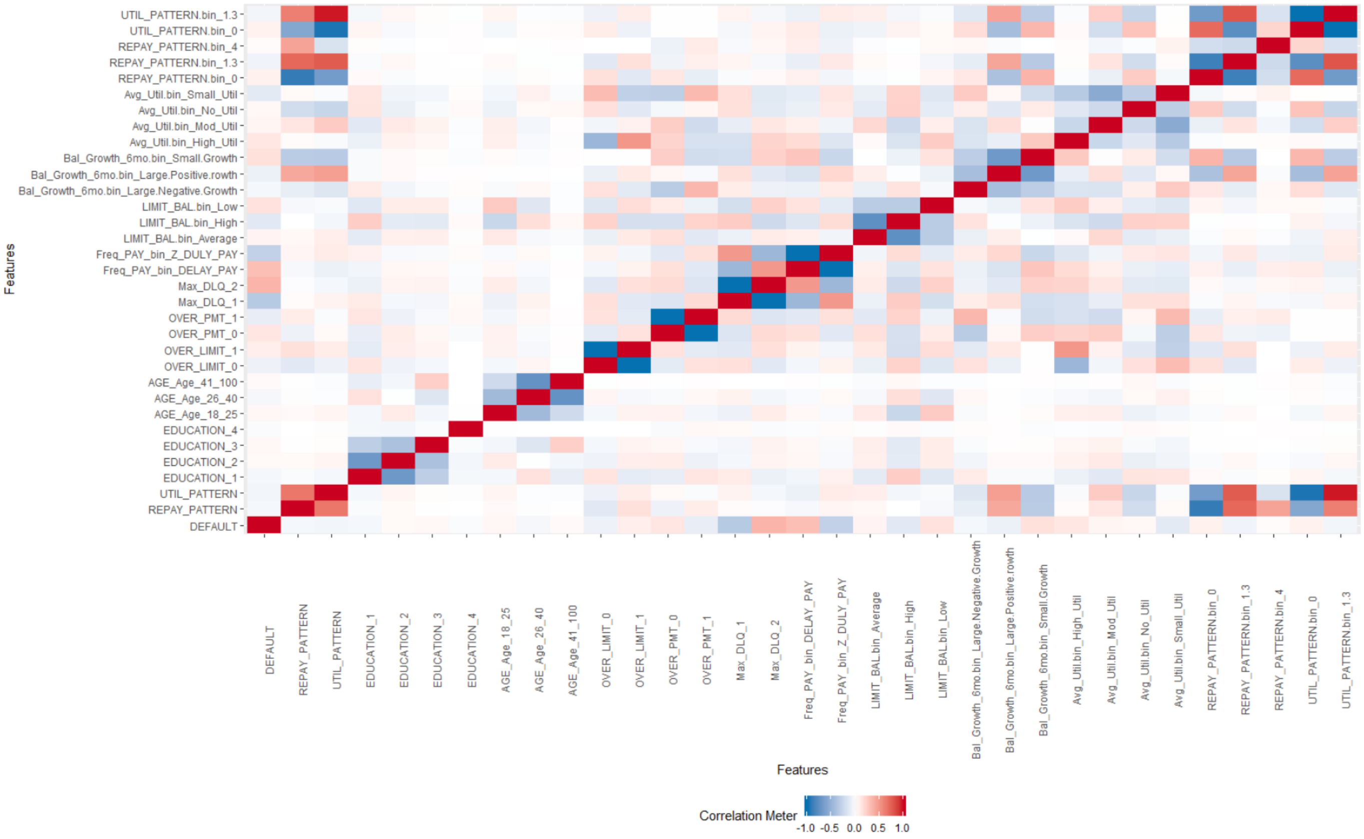


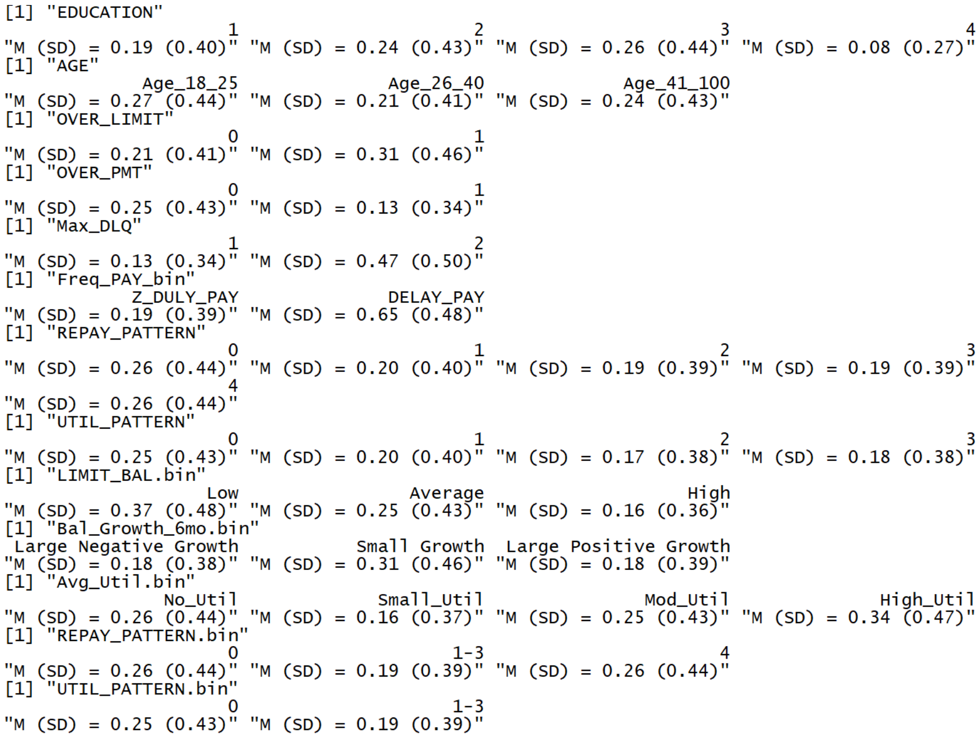
Figure 21. Correlation matrix of categorical variables. (Pearson)



4.a.c. Bivariate Analysis. Categorical Variables

The Table 10. displays the mean and standard deviation of the response variable for each class of every categorical variable. Most variables have the differences in mean values of the response variable for each class which indicates that the variables have predictive power. EDUCATION variable has least predictive power based on the insignificant difference in means for the response variable between its categories. The categories 2 and 3 in REPAY_PATTERN have the same mean which suggest using binned variable for modeling which combined these two categories. Similarly, the binned counterpart of UTIL_PATTERN has more significant difference between the mean values of the response variable across its categories. Bal_Growth_6mo.bin variable indicates that large negative growth and large growth have the same impact on the likelihood of a customer to default.

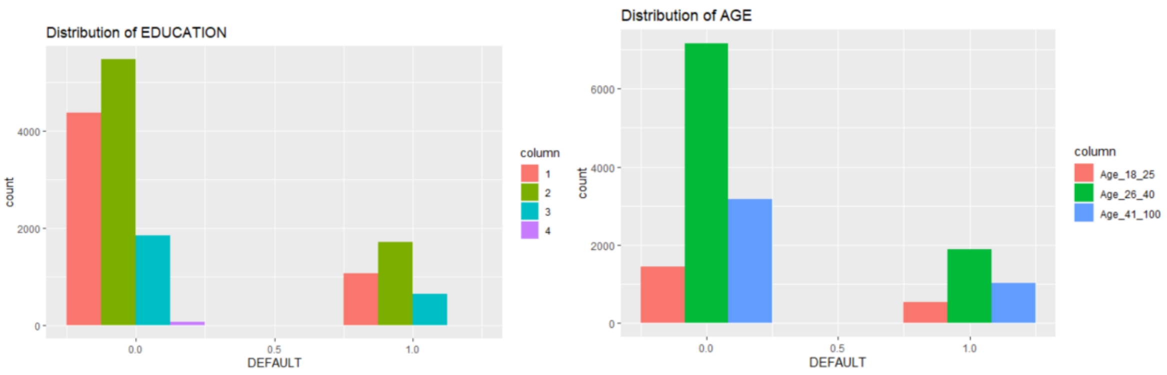
Table 10: Mean and Standard deviation of DEFAULT variable for each class

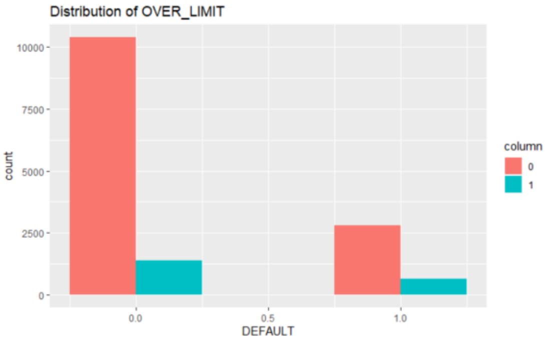


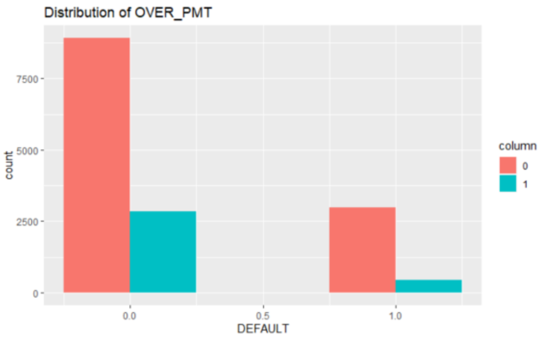
```
[1] "EDUCATION"
                     1                      2                      3                      4 
"M (SD) = 0.19 (0.40)" "M (SD) = 0.24 (0.43)" "M (SD) = 0.26 (0.44)" "M (SD) = 0.08 (0.27)" 
[1] "AGE"
             Age_18_25              Age_26_40             Age_41_100 
"M (SD) = 0.27 (0.44)" "M (SD) = 0.21 (0.41)" "M (SD) = 0.24 (0.43)" 
[1] "OVER_LIMIT"
                     0                      1 
"M (SD) = 0.21 (0.41)" "M (SD) = 0.31 (0.46)" 
[1] "OVER_PMT"
                     0                      1 
"M (SD) = 0.25 (0.43)" "M (SD) = 0.13 (0.34)" 
[1] "Max_DLQ"
                     1                      2 
"M (SD) = 0.13 (0.34)" "M (SD) = 0.47 (0.50)" 
[1] "Freq_PAY_bin"
            Z_DULY_PAY              DELAY_PAY 
"M (SD) = 0.19 (0.39)" "M (SD) = 0.65 (0.48)" 
[1] "REPAY_PATTERN"
                     0                      1                      2                      3 
"M (SD) = 0.26 (0.44)" "M (SD) = 0.20 (0.40)" "M (SD) = 0.19 (0.39)" "M (SD) = 0.19 (0.39)" 
                     4 
"M (SD) = 0.26 (0.44)" 
[1] "UTIL_PATTERN"
                     0                      1                      2                      3 
"M (SD) = 0.25 (0.43)" "M (SD) = 0.20 (0.40)" "M (SD) = 0.17 (0.38)" "M (SD) = 0.18 (0.38)" 
[1] "LIMIT_BAL.bin"
                   Low                Average                   High 
"M (SD) = 0.37 (0.48)" "M (SD) = 0.25 (0.43)" "M (SD) = 0.16 (0.36)" 
[1] "Bal_Growth_6mo.bin"
 Large Negative Growth           Small Growth  Large Positive Growth 
"M (SD) = 0.18 (0.38)" "M (SD) = 0.31 (0.46)" "M (SD) = 0.18 (0.39)" 
[1] "Avg_Util.bin"
               No_Util             Small_Util               Mod_Util              High_Util 
"M (SD) = 0.26 (0.44)" "M (SD) = 0.16 (0.37)" "M (SD) = 0.25 (0.43)" "M (SD) = 0.34 (0.47)" 
[1] "REPAY_PATTERN.bin"
                     0                    1-3                      4 
"M (SD) = 0.26 (0.44)" "M (SD) = 0.19 (0.39)" "M (SD) = 0.26 (0.44)" 
[1] "UTIL_PATTERN.bin"
                     0                    1-3 
"M (SD) = 0.25 (0.43)" "M (SD) = 0.19 (0.39)" 
```

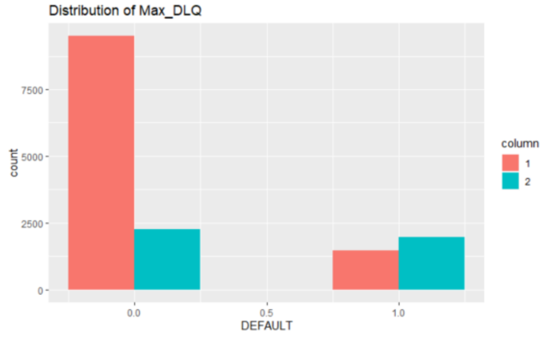
The same conclusion is drawn from the examination of conditioned by the response variable histograms of predictor variables (See Figure 22.). It is worth to notice that the observations of OTHER category of Education variable exclusively belong to non-defaulters. The variables Max_DLQ, LIMIT_BAL_bin, Util_Growth_6_mo, Balance_Growth_6mo_bin, Average_Util, and REPAY_PATTERN have different composition of classes for non-defaulters and defaulters which is evident of the variables' strong association with the response variable. For example, the plot of Max_DLQ conditioned by the response variable reveals that the vast majority of non- defaulters overly delinquency status 1 (- duly paid no more than one month delay in balance payment) while opposite is true for defaulters. LIMIT_BAL_bin' distribution suggests that average credit limit is more indicator of a default event than high or low credit limits. Since the observations with small growth in balance prevail among defaulters and are not prevalent among non- defaulters, the belonging to the category might be an indicator of risk to go in debt. In Avg_Util variable the records with small utilization has considerably larger presence among non-defaulters while customers with high average credit utilization have larger presence in the default category. Moderate on average utilization of the credit limit is prevalent among defaulters and not prevalent among non – defaulters. The categories 0 and 1-3 in REPAY_PATTERN are good indicators of safe and risky credit behaviors respectively. 0 status is prevalent among non – defaulters and not prevalent among defaulters and 1-3 repayment status in opposite is prevalent among defaulters and not prevalent among non – defaulters. The same pattern is true for UTIL_PATTERN variable. The customers in 1 category for Freq_PAY_bin and OVER_LIMIT variables largely belong to defaulters which suggests that customers who frequently delayed their balance payments and went over their limit at least once over 6-months period are risky customers. OVER_PMT plot indicates that the customers less likely to default if they overpaid their balance at least once over 6-months period.

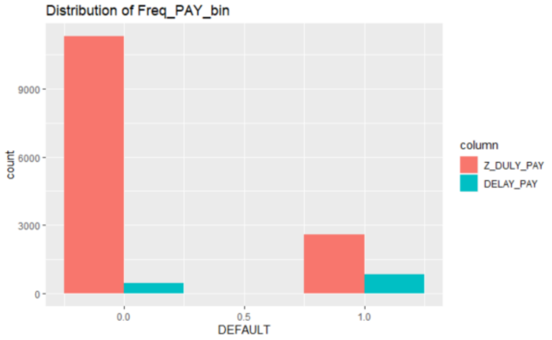
Figure 22. The conditioned by the response variable histograms of the categorical predictor variables

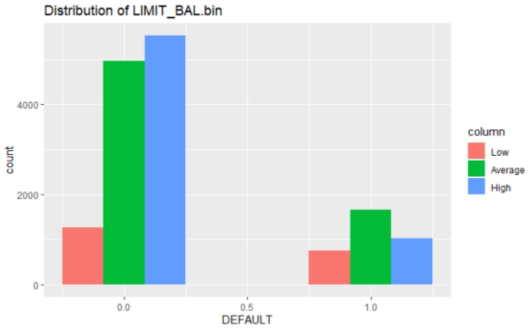


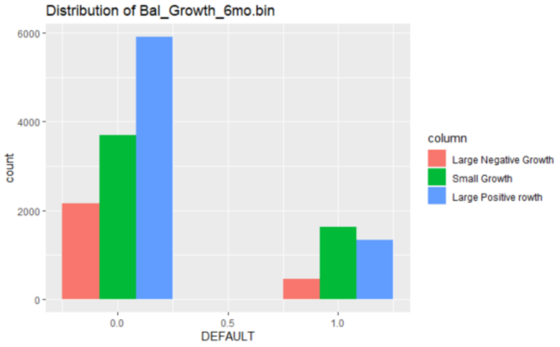

Distribution of OVER_LIMIT
count
10000
7500
5000
2500
0
0.0
0.5
1.0
DEFAULT
column
0
1

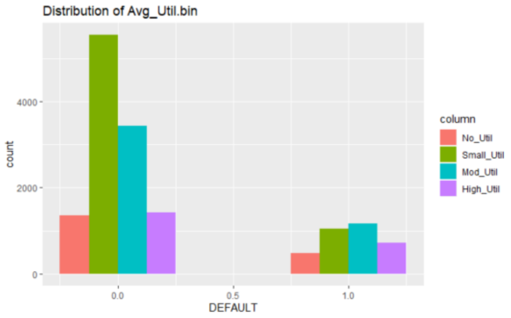

Distribution of OVER_PMT
count
7500
5000
2500
0
0.0
0.5
1.0
DEFAULT
column
0
1

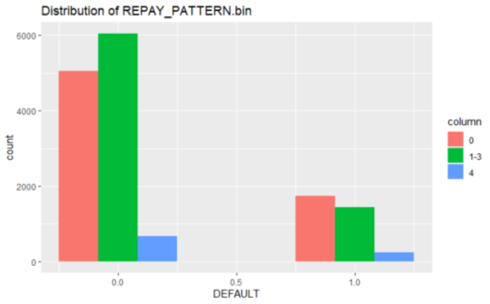

Distribution of Max_DLQ
count
7500
5000
2500
0
0.0
0.5
1.0
DEFAULT
column
1
2

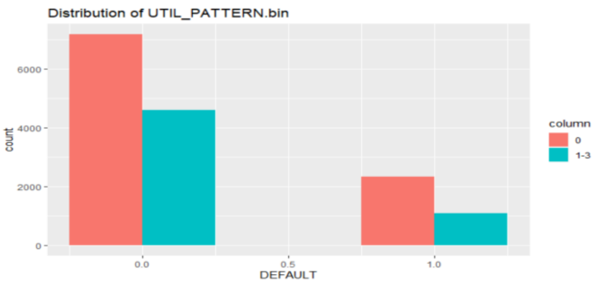

Distribution of Freq_PAY_bin
count
9000
6000
3000
0
0.0
0.5
1.0
DEFAULT
column
Z_DULY_PAY
DELAY_PAY

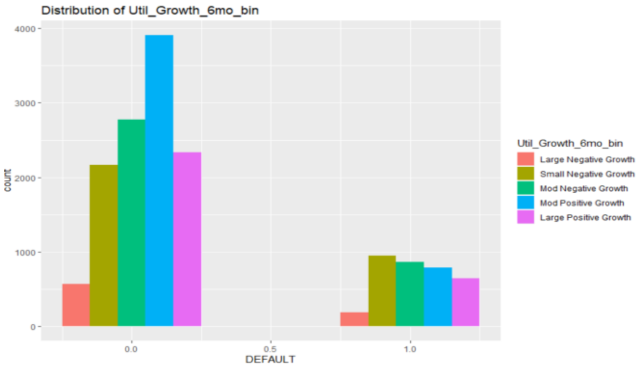

Distribution of LIMIT_BAL.bin
count
4000
2000
0
0.0
0.5
1.0
DEFAULT
column
Low
Average
High


Distribution of Bal_Growth_6mo.bin
count
6000
4000
2000
0
0.0
0.5
1.0
DEFAULT
column
Large Negative Growth
Small Growth
Large Positive rowth


Distribution of Avg_Util.bin
count
4000
2000
0
0.0
0.5
1.0
DEFAULT
column
No_Util
Small_Util
Mod_Util
High_Util


Distribution of REPAY_PATTERN.bin
count
6000
4000
2000
0
0.0
0.5
1.0
DEFAULT
column
0
1-3
4

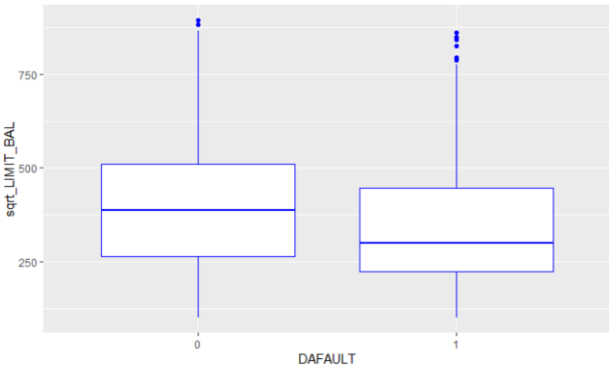




4.a.d. Bivariate Analysis. Continuous Variables

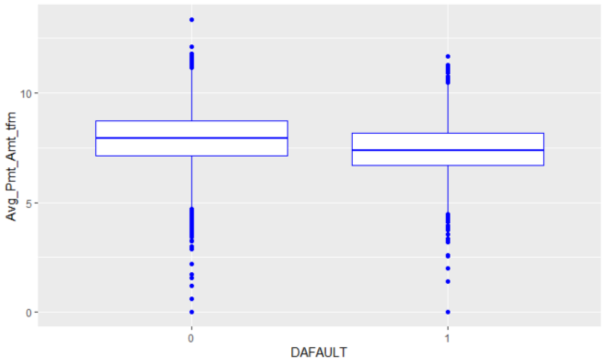
The boxplots of continuous predictors also exhibit that the variables are useful in terms of predicting the risk of default except for Avg_Bill_Amt_tfm, Max_Bill_amt_tfm, Max_Pamt_amt_tfm, Max_Bill_Amt_tfm, Max_Util_Ratio which have weak relationships with the response. sqrt_LIMIT_BALANCE variable demonstrates the difference in median values of the predictor variable for defaulters and non- defaulters (See Figure 23.).

Figure 23.



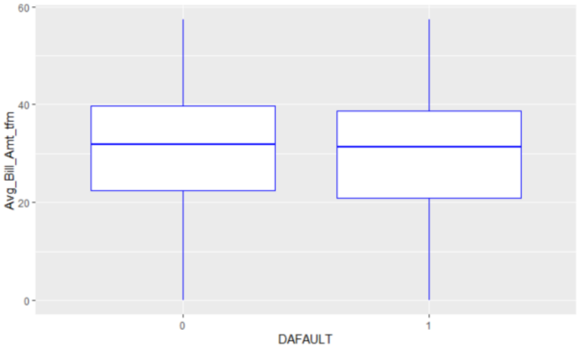
The boxplot of Avg_Pmt_Amt_tfm exhibits the different medians for two groups which suggests the variable's importance in predicting the outcome variable (see Figure 24).

Figure 24.



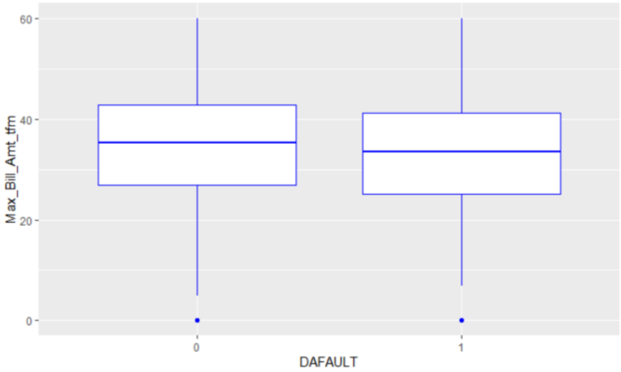
The significant difference in the medians of Avg_Bill_Amt_tfm variable is not observed for defaulters and non- defaulters which suggets that the variable is a weak predictor of a default risk (See Figure 25.).

Figure 25.



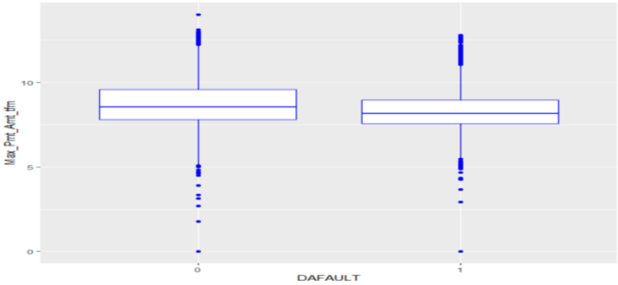
Max_Bill_Amt_tfm variable has only week association with the response variable (see Figure 26.).

Figure 26.



The boxplot if Max_Pmt_Amt_tfm variable does not indicate a strong relationship between variables (See Figure 27).

Figure 27.



The Table 11, which summarizes how the median of Pmt_Ratio variable changes over 6-month period**,** reveals that at the beginning of the 6-month period, in April and May, the customers in both groups have being paying less portion of their bills than in August or May. Both groups have been steadily and at the same rate incresing their balance payments over the 6-month period. Pmt_Ratio_5 deviates form the overall trend by exhibiting the decline in payments. The large differences in medians for Pmt_Ratio variable and Avg_Pmt_Ratio indicate that the less fraction of the bill is paid by a customer the more chances fo him/her to default.

Table 11: The change in Median of Pmt_Ratio variable over 6-month period

	DEFAULT	Pmt_Ratio_2	Pmt_Ratio_3	Pmt_Ratio_4	Pmt_Ratio_5	Pmt_Ratio_6	Avg_Pmt_Ratio
1	0	0.12	0.12	0.10	0.09	0.11	0.24
2	1	0.07	0.07	0.06	0.05	0.06	0.08

Based on the examination of the table of tracing the change in median utilization over time it could be concluded that the utilization steadily grown from April through September and as well as the difference in medians between two groups. The credit utilization of the default group has grown more on average compared to non- defaulters. So, the utilization growth might be a strong predictor of the default. Overall, Util and Avg_Util variables are strong predictors of the default since they have significantly median values for default and non-default groups. The increase in utilization increases the risk of default.

Table 12: The change in Median of Util variable over 6-month period

	DEFAULT	Util_1	Util_2	Util_3	Util_4	Util_5	Util_6	Avg_Util
1	0	0.26	0.25	0.22	0.20	0.18	0.15	0.24
2	1	0.49	0.50	0.47	0.41	0.38	0.35	0.46

MAX_UTIL and MIN_UTIL variables have a significance difference in their means for risky and non-risky customers which is a sign of a strong predictive power (See Figures 28, 29). The greater the value for maximum and minimum of credit utilization for a customer the greater the likelihood for this customer to default. MAX_Util_Ratio does not display a strong relationship with the outcome variable (See Figure 30.).

Figure 28.

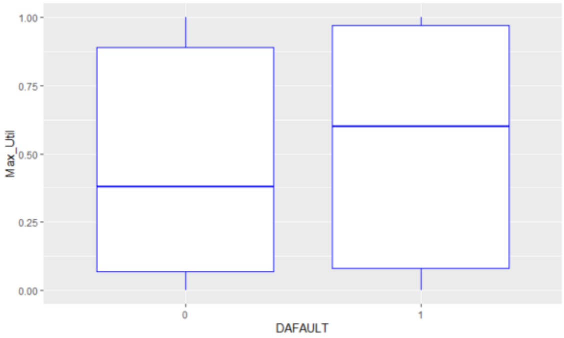


Figure 29.

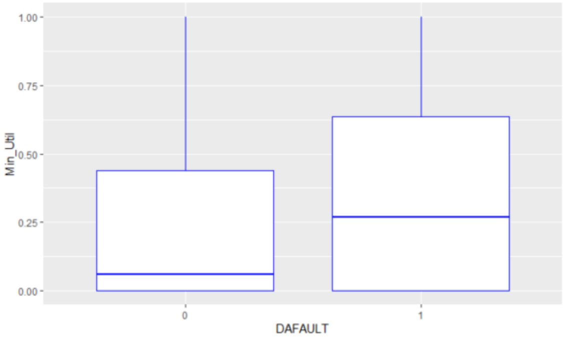
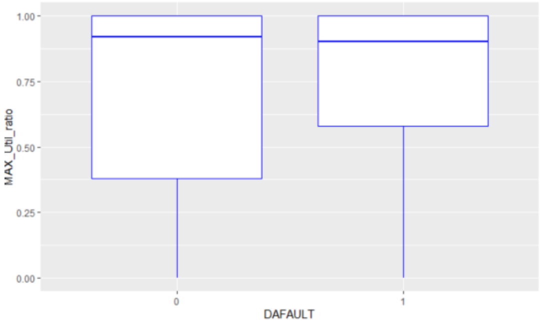


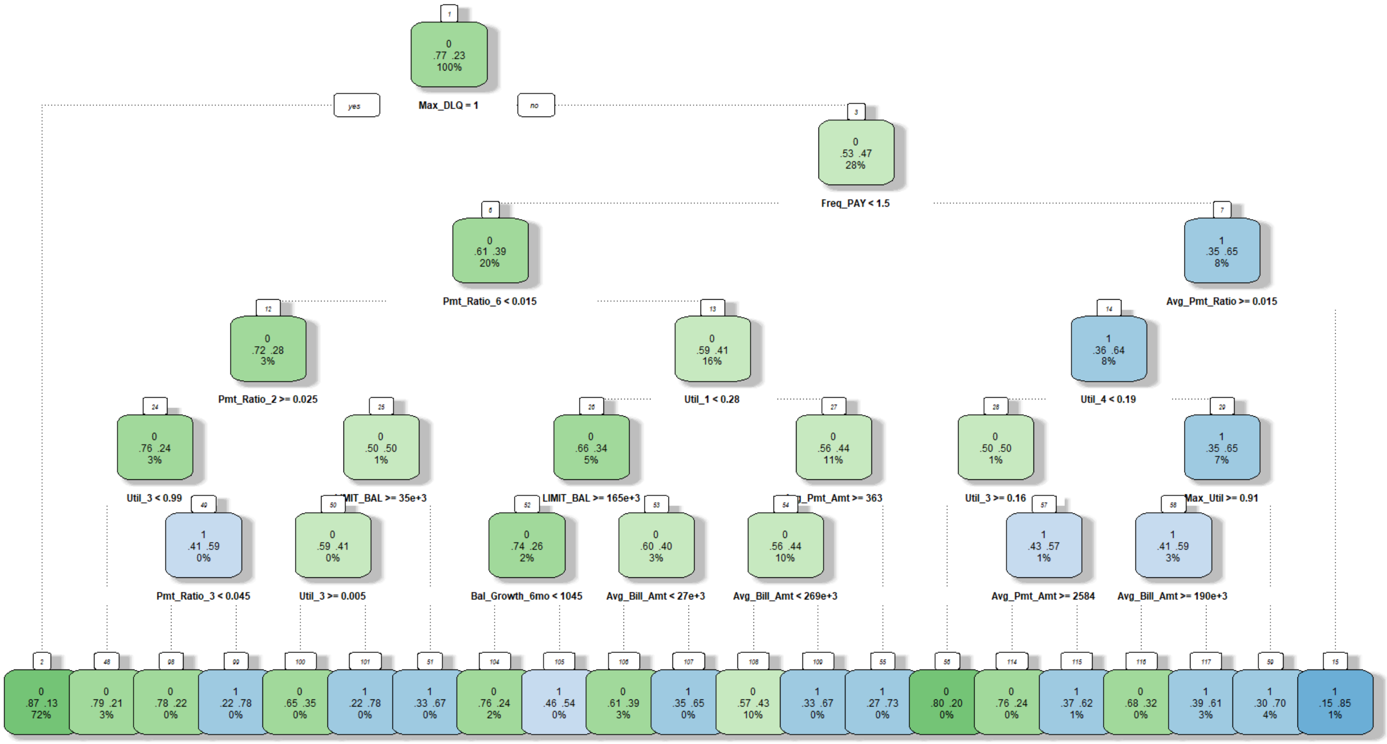
Figure 30.



4b. Model – based EDA

Recursive Partitioning and Regression Trees algorithm has been used with original (Figure 30.) and transformed data sets (Figure 31.) to identify the important variables to use in the classification analysis. Rpart algorithm used original data produced lower cross validation error (0.8892784) than Rpart algorithm which used transformed data (0.8898627). The variables' importance has been ranked by both algorithms slightly different. It is important to notice that variables with highly skewed distributions and large number of extreme outliers such as Avg_Bill_Amt, Max_Bill_Amt, Avg_pmt_Amt, Bal_Growth_6mo have been weighted by the algorithm much less after transformation and Sqrt_Limit_BAL is more important than LIMIT_BAL_bin. Nevertheless, in both cases MAX_DLQ and FREQ_PAY variables are ranked as the strongest predictors of the outcome variable.

Figure 30. Recursive Partitioning and Regression Tree. Original Data.



Rattle 2019-Oct-30 14:21:20 yanin

Figure 31. Variable importance plot. Original data.

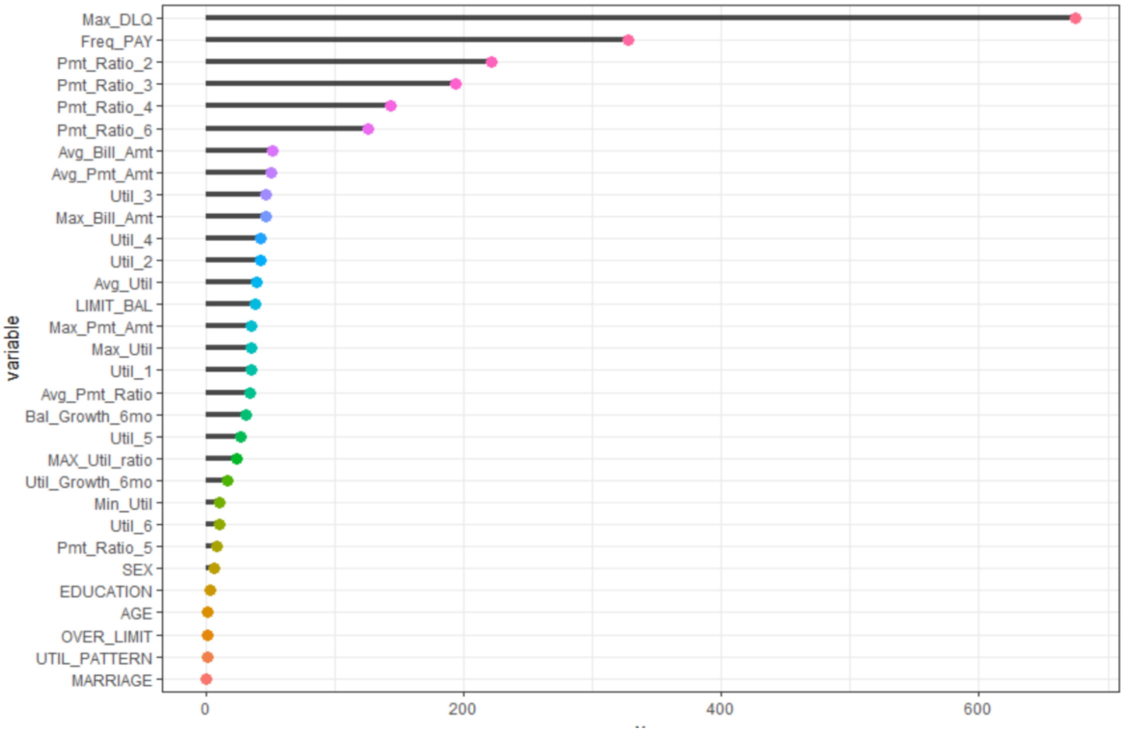


Figure 32. Recursive Partitioning and Regression Tree. Transformed Data.

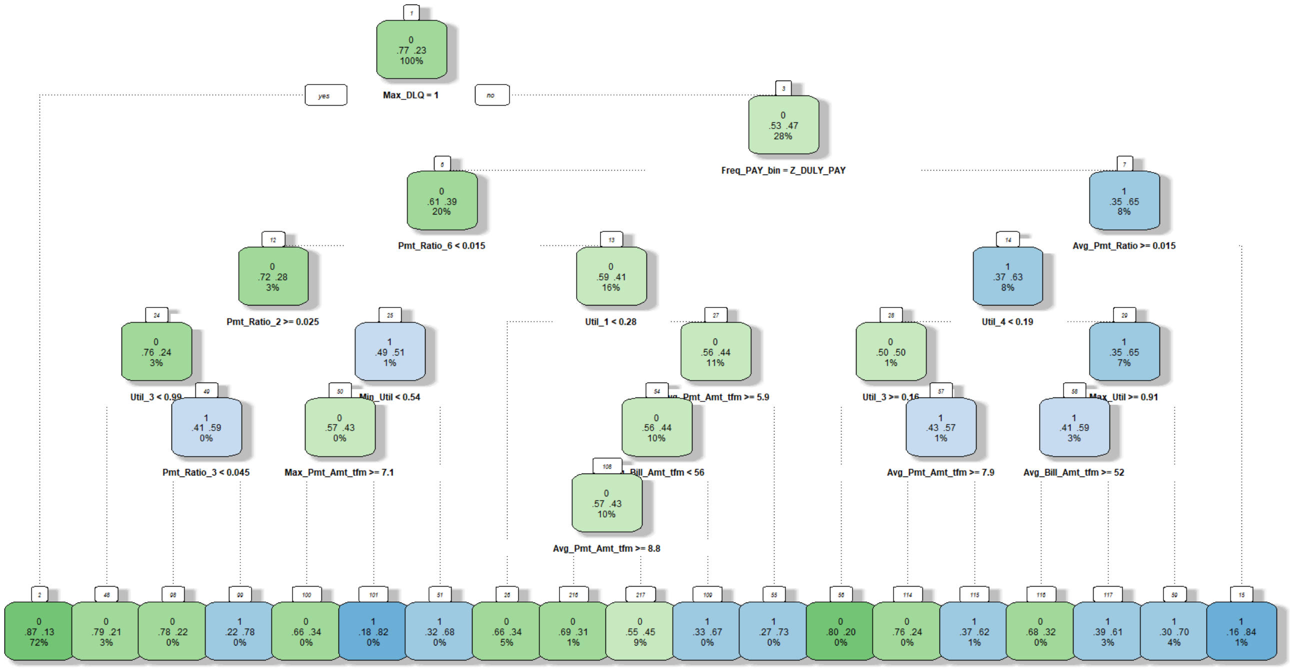
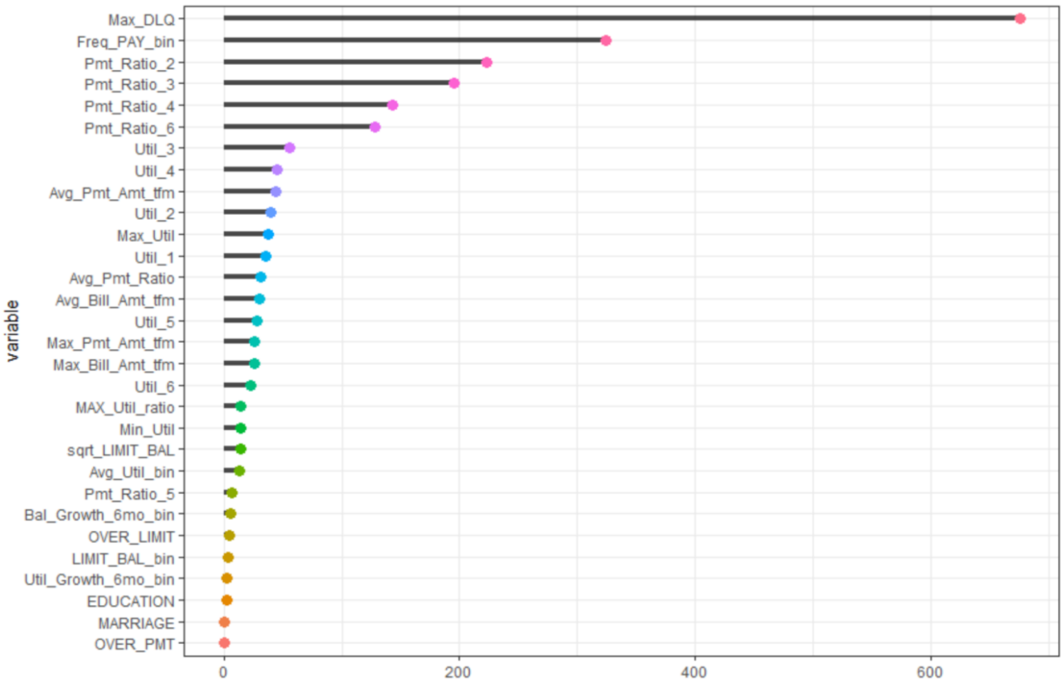


Figure 33. Variable importance plot



One Rule algorithm has been used to build baseline model, the best simple model, for models' performance evaluation and to retrieve importance ranking of variables[2]. One Rule algorithm suggests that essentially any model should reach the level of 80.57% accuracy on the test data or sum of TP and TN of 1.20[3]. This is achieved with just one simple rule when customers have the frequent delinquency status of Z_DULY_PAY, they are risk free customers with 79.93% accuracy and when customers have DELAY_PAY (delay in repayment more than 1 month) as their most frequent delinquency status they are most likely candidates for default on credit payment. The models might have difficulties to detect the defaulters since the data is unbalanced. The evaluation of the OneRule model' performance is presented in the table below.

[2] The transformed data set has been used since One Rule requires binned predictors and by default continuous variables are binned automatically using equal binning approach which might be less robust than woe binning.

[3] It is more reasonable to look at TP+TN since the data is unbalanced and the high accuracy could be obtained by j predicting the prevalent negative class with no consideration of Recall.

Table 13. OneRule model performance on test data.

Model #0: OneRule Model													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.24	TP+TN	1.20	AUC	
Class	0	1			Class	0	1	TN	0.96	Precision	0.61	Sensitivity	0.24
0	5,528	238	5,766		0	0.96	0.04	Type I Error	0.04	Recall	0.24	Specificity	0.96
1	1,185	372	1,557		1	0.76	0.24	Type II Error	0.76	F1	0.37		

One rule identified FREQ PAY variable as most significant for classifying customers as credit risky or not. Interesting to note that FREQ_PAY has been ranked second important by Rpart algorithm after MAX_DLQ. All other predictors produce 77.45 % accuracy values (see Appendix C).

5. Predictive Modeling: Methods and Results

At the modeling stage, transformed and original data have been used to build classifiers. The evaluation of the classifiers' performances on transformed and original data revealed that models fitted to the binned data produce better results in – sample and out-of- sample[4]. Thus, all the learners presented in the section are built using transformed data. The choice of performance metrics for assessing the models' predictive capabilities and establishing cut-off values for categorical forecasts have been based on the imbalanced characteristic of the response variable[5] and the goal to optimize tradeoff between multiple objectives, minimizing the risk of monetary loss due to borrower defaulting and maximizing the market share. The cost of TYPE II error is assumed to be high and the goal of classifiers is to catch as much risky customers as possible. The expected loss due to credit defaulting could be realized by maximizing Recall or minimizing Type II error. The increase in Recall inevitably leads to the decrease in Precision which translates into the loss of profit by shrinking the market share. Hence two objectives are inversely correlated F1 score[6] was used to tune the models' hyperparameters (where applicable) and obtain the

[4] Appendix D contains the evaluation of Random Forest model built with original data to compare the results with the model trained on the transformed data.

[5] A-priori probabilities of the response: 0 - 0.7745059; 1 - 0.2254941. Since the response variable is imbalanced, the model could disregard the minority class without compromising accuracy.

[6] F1 score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. The F1 score conveys the balance between Precision and Recall. F1 is usually more useful than accuracy, especially in case of an uneven class distribution because it is sensitive to imbalances in data

optimal cutoff values to realize trade-off between expected loss and profit. The models' ability to discriminate good and bad customers have been measured using Kolmogorov–Smirnov test and Area under the ROC Curve. The models have been evaluated and compared using the same performance measures in – sample and out- of -sample. Confusion matrix has been used to shed light on the population odds and reveal the truth on the model ability catch the risk defaulters which is not implicitly observed using any discrimination measures. The models have been cross-validated and tested on transformed test data to detect overfitting and measure the models' powers to predict unseen instances.

5.a Random Forest

Three Random Forest models have been built using different tuning techniques and metric to optimize. The random search and grid search tuning strategies have been used to fine tune the models' hyperparameters and cut-off values based on the objective to either maximize the True positive rate (Model #1) or F1 performance measure (Model #2 and Model #3). The models have been cross - validated and tested on test data. Stratification has been applied for resampling in cross- validation to account for imbalances in the response variable. The confusion matrices and the performance results measured on train and test data are presented below. The threshold is tuned to maximize F1 score.

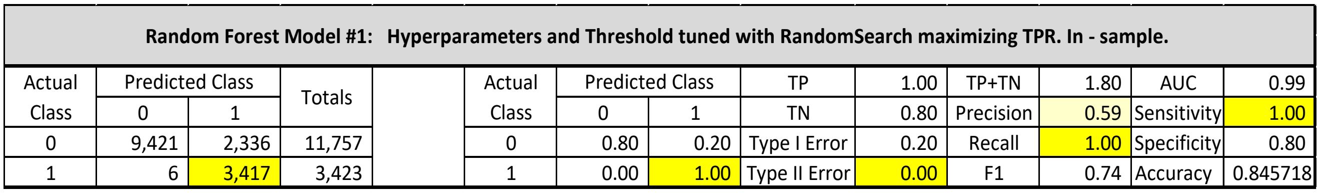
5.a.a. Random Forest Model #1

The first Random Forest model has been built using Random Search tuning technique to find the best random combinations of hyperparameters that gives the highest TPR solution. The set of optimal hyperparameters is confined to the predefined parameter space and Random Search implies that not all the parameters are tested to find the optimal solution only some number of random configurations in the parameter space are tested[7]. The model has produced very optimistic results on train data with Recall of 1, F1 of 0.74, and AUC of 0.99 (see Table 14.) . However, cross – validation results and performance on the

[7] However, the parameter optimization strategy is less computation expensive and sometimes work best for relatively low-dimensional data

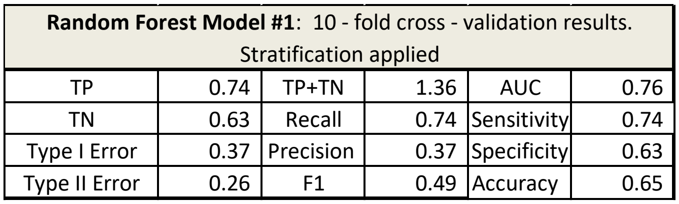
test data are much less promising which suggests that the model overfitted the train data (see Table 15 and Table 16.). Nevertheless, the main drawback of the model is high Type I error which is the result of high Recall value. If the extra 877 true positives (compared to baseline One – Rule model) in the portfolio compensate for the extra 2,523 false positives, the model is optimal. Kolmogorov–Smirnov test revealed a relatively strong discriminatory power of the model, 0.39 as well as AUC value of 0.75 (see Table 16) .

Table 14.



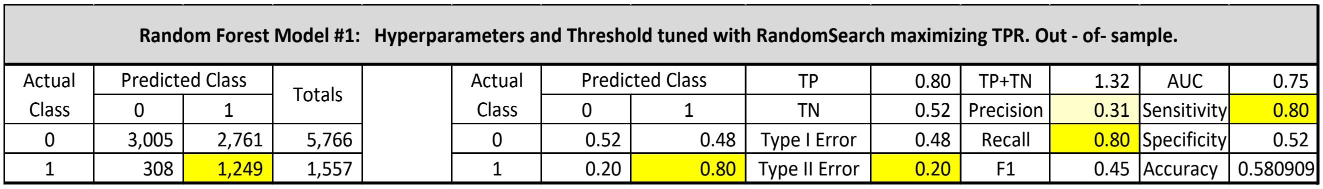
Random Forest Model #1: Hyperparameters and Threshold tuned with RandomSearch maximizing TPR. In - sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	1.00	TP+TN	1.80	AUC	0.99
Class	0	1			Class	0	1	TN	0.80	Precision	0.59	Sensitivity	1.00
0	9,421	2,336	11,757		0	0.80	0.20	Type I Error	0.20	Recall	1.00	Specificity	0.80
1	6	3,417	3,423		1	0.00	1.00	Type II Error	0.00	F1	0.74	Accuracy	0.845718

Table 15.



Random Forest Model #1: 10 - fold cross - validation results. Stratification applied					
TP	0.74	TP+TN	1.36	AUC	0.76
TN	0.63	Recall	0.74	Sensitivity	0.74
Type I Error	0.37	Precision	0.37	Specificity	0.63
Type II Error	0.26	F1	0.49	Accuracy	0.65

Table 16.

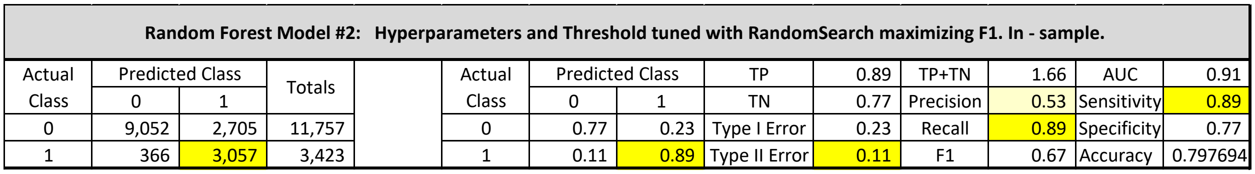


Random Forest Model #1: Hyperparameters and Threshold tuned with RandomSearch maximizing TPR. Out - of- sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.80	TP+TN	1.32	AUC	0.75
Class	0	1			Class	0	1	TN	0.52	Precision	0.31	Sensitivity	0.80
0	3,005	2,761	5,766		0	0.52	0.48	Type I Error	0.48	Recall	0.80	Specificity	0.52
1	308	1,249	1,557		1	0.20	0.80	Type II Error	0.20	F1	0.45	Accuracy	0.580909

5.a.b.Random Forest Model # 2

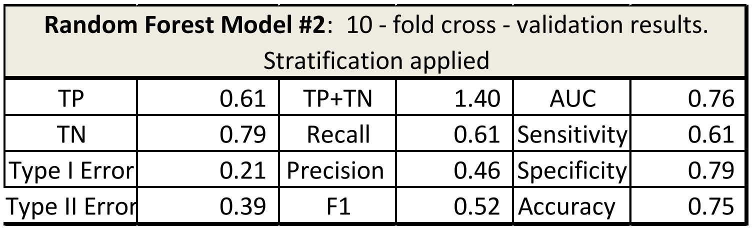
To improve Precision of the model while preserving as highest as possible Recall, Random forest model has been tuned to maximize F1 value. Again, the performance results on the training data is very optimistic suggesting the model's strong power to discriminate the classes and accurately predict the default. The model performed worse than the Model #1 on train data with lower Precision and Recall (see Table 17.) However, the model performed considerably better both in cross- validation and on test data (see Table 18 and 19). Accuracy and Precision of the model has increased producing more false negatives. Nevertheless, the model is more balanced in terms of its capacity to detect the customers with a high credit default risk and accurately identify them as such. KS value has just slightly improved to 0.40407 while AUC stayed the same suggesting more or less the same capability of the model to differentiate good customers from bad. Since the goal is to build a classifier that realizes tradeoff between Precision and Recall, the considerably improved F1 measure indicates that Model #2 outperforms Model #1. The difference in performance on train and test data sets suggests that the model overfitted the train data but the gap in performance is less than in case of Model #1. The similar results obtained from 10-fold cross- validation and on the test data indicates that the model performs well.

Table 17 .



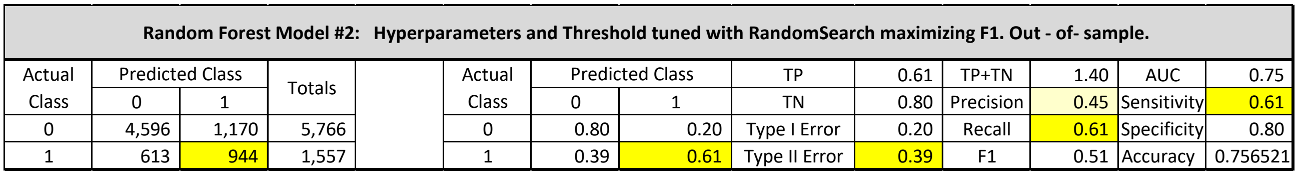
Random Forest Model #2: Hyperparameters and Threshold tuned with RandomSearch maximizing F1. In - sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.89	TP+TN	1.66	AUC	0.91
Class	0	1			Class	0	1	TN	0.77	Precision	0.53	Sensitivity	0.89
0	9,052	2,705	11,757		0	0.77	0.23	Type I Error	0.23	Recall	0.89	Specificity	0.77
1	366	3,057	3,423		1	0.11	0.89	Type II Error	0.11	F1	0.67	Accuracy	0.797694

Table 18.



Random Forest Model #2: 10 - fold cross - validation results. Stratification applied					
TP	0.61	TP+TN	1.40	AUC	0.76
TN	0.79	Recall	0.61	Sensitivity	0.61
Type I Error	0.21	Precision	0.46	Specificity	0.79
Type II Error	0.39	F1	0.52	Accuracy	0.75

Table 19.



Random Forest Model #2: Hyperparameters and Threshold tuned with RandomSearch maximizing F1. Out - of- sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.61	TP+TN	1.40	AUC	0.75
Class	0	1			Class	0	1	TN	0.80	Precision	0.45	Sensitivity	0.61
0	4,596	1,170	5,766		0	0.80	0.20	Type I Error	0.20	Recall	0.61	Specificity	0.80
1	613	944	1,557		1	0.39	0.61	Type II Error	0.39	F1	0.51	Accuracy	0.756521

5.a.c. Random Forest Model #3

Since the desirable performance level has not been obtained, the third Random Forest model has been built using Grid Search hyperparameter optimization strategy assuming that the tuning technique would find a better set of hyperparameters to further optimize model performance based on F1 score[8]. The results from cross-validation and performance on the test data presented in the tables below demonstrate that accuracy, Recall and F1 score of the model increased as well as the discriminatory power of the classifier, KS value has increased from 0.4 to 0.42. Moreover, the model demonstrated the least difference in performance on train and test data suggesting to be least overfitted out of three random forest models.

Table 20

Random Forest Model #3: Hyperparameters and Threshold tuned with GridSearch maximizing F1. In-sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.87	TP+TN	1.63	AUC	0.89
Class	0	1			Class	0	1	TN	0.76	Precision	0.52	Sensitivity	0.87
0	8,986	2,771	11,757		0	0.76	0.24	Type I Error	0.24	Recall	0.87	Specificity	0.76
1	461	2,962	3,423		1	0.13	0.87	Type II Error	0.13	F1	0.65	Accuracy	0.787088

Table 21.

Random Forest Model #3: 10 - fold cross - validation results. Stratification applied					
TP	0.58	TP+TN	1.40	AUC	0.76
TN	0.82	Recall	0.58	Sensitivity	0.58
Type I Error	0.18	Precision	0.48	Specificity	0.82
Type II Error	0.42	F1	0.52	Accuracy	0.76

Table 22.

Random Forest Model #3: Hyperparameters and Threshold tuned with GridSearch maximizing F1. Out-of-sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.62	TP+TN	1.42	AUC	0.75
Class	0	1			Class	0	1	TN	0.80	Precision	0.45	Sensitivity	0.62
0	4,599	1,167	5,766		0	0.80	0.20	Type I Error	0.20	Recall	0.62	Specificity	0.80
1	592	965	1,557		1	0.38	0.62	Type II Error	0.38	F1	0.52	Accuracy	0.759798

[8] The Grid Search approach allows to exhaust all the possible hyperparameters' combinations within the predefined hyperparameter space.

5.a.d. Champion Random Forest. Feature Importance plot and ROC and Precision Recall Curves

Based on the business objective to build the balanced model, Random Forest Model # 3 has been selected as a champion model which produced highest KS value, sum of TP and TN, Recall, and F1 score compared to other models. While 0.15 threshold has been used to meet tradeoff requirement of the model, the ROC(see Figure 34.) and Precision– Recall (see Figure 35.) curves presented below could be used to further fine tune the threshold if the business requirements change and it has been decided to increase or reduce threshold to either increase the market share or decrease the expected loss. The ROC and Precision- Recall curves suggest that the chosen 0.15 threshold for the test data is optimal to produce the best possible Recall value without accruing a lot of FPR or drastically reducing precision.

Figure 34.

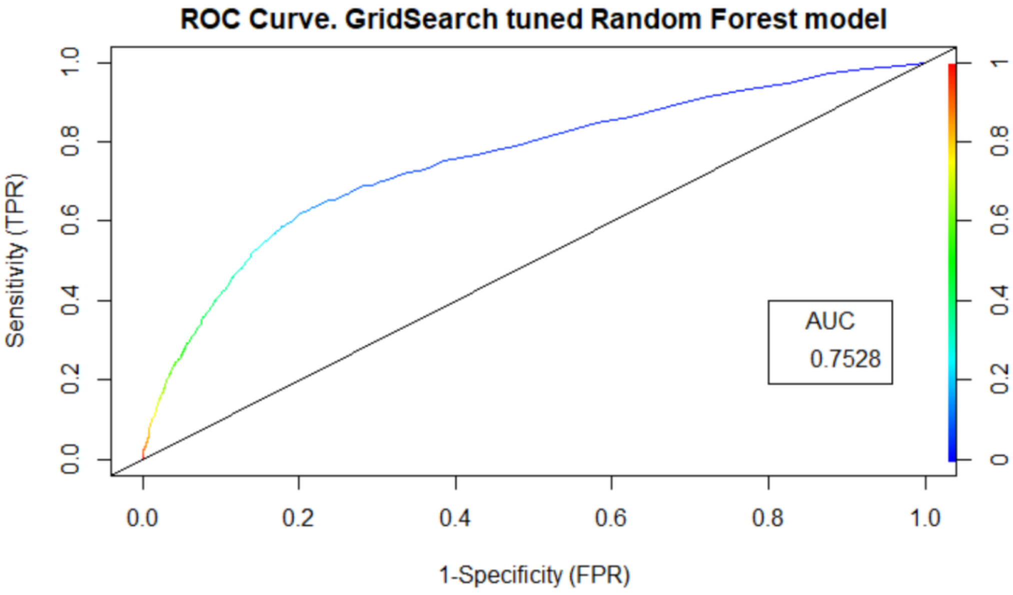
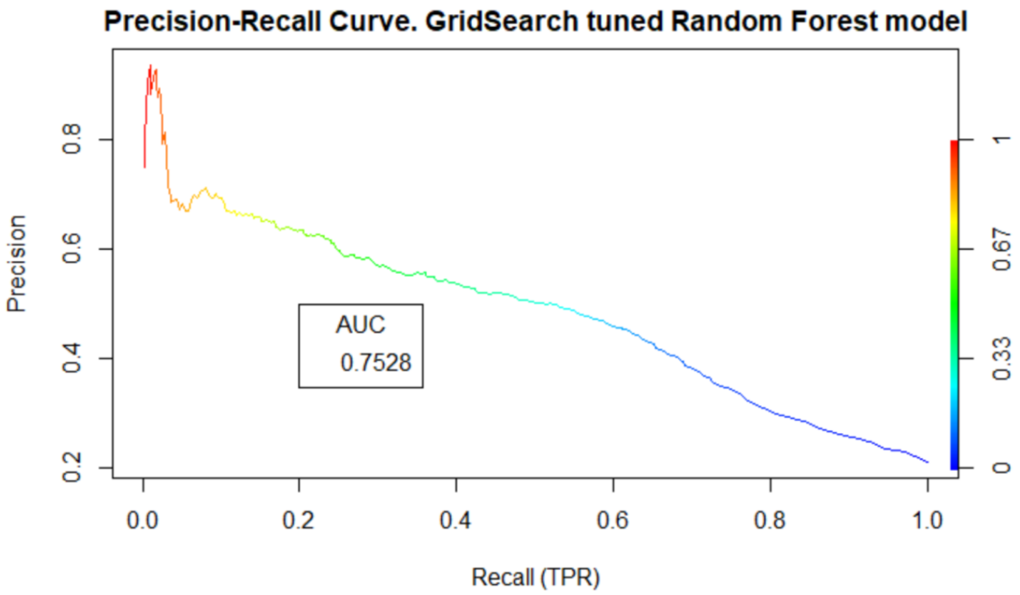


Figure 35.



The importance of features measured by contribution to decrease in Gini impurity (Figure 36.) and F1 criteria (Figure 37.) are presented below. Interesting to note that Freq_Pay_ bin variable has been ranked differently according to F1 and Gini Index performance criteria.

Figure 36. Importance ranking of variables (Gini Index).

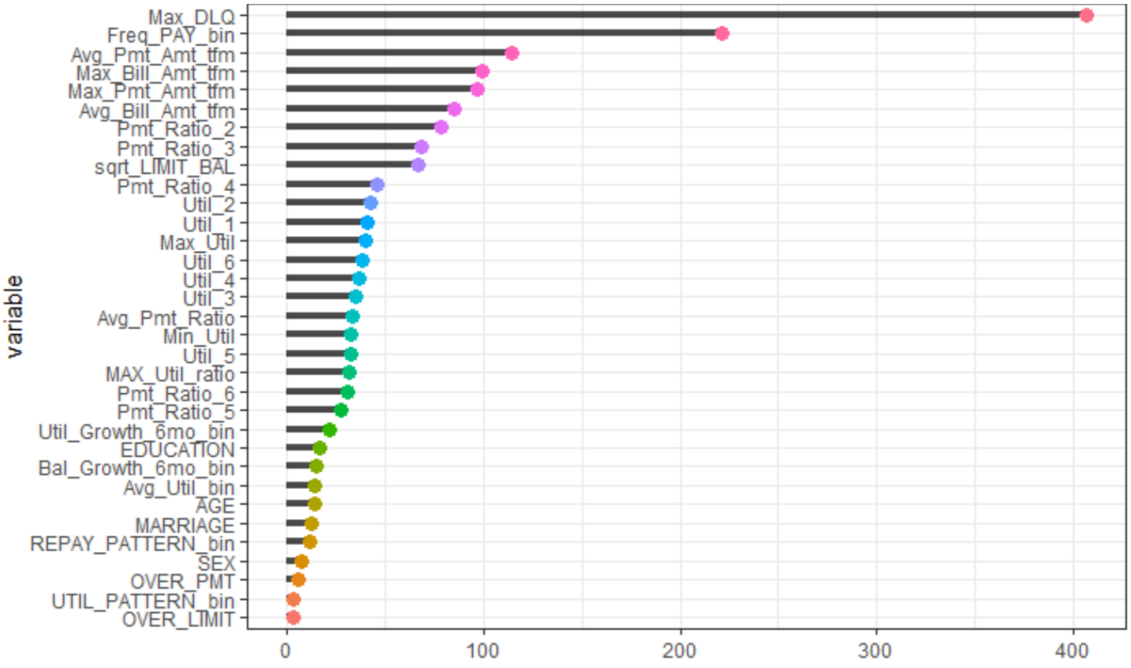
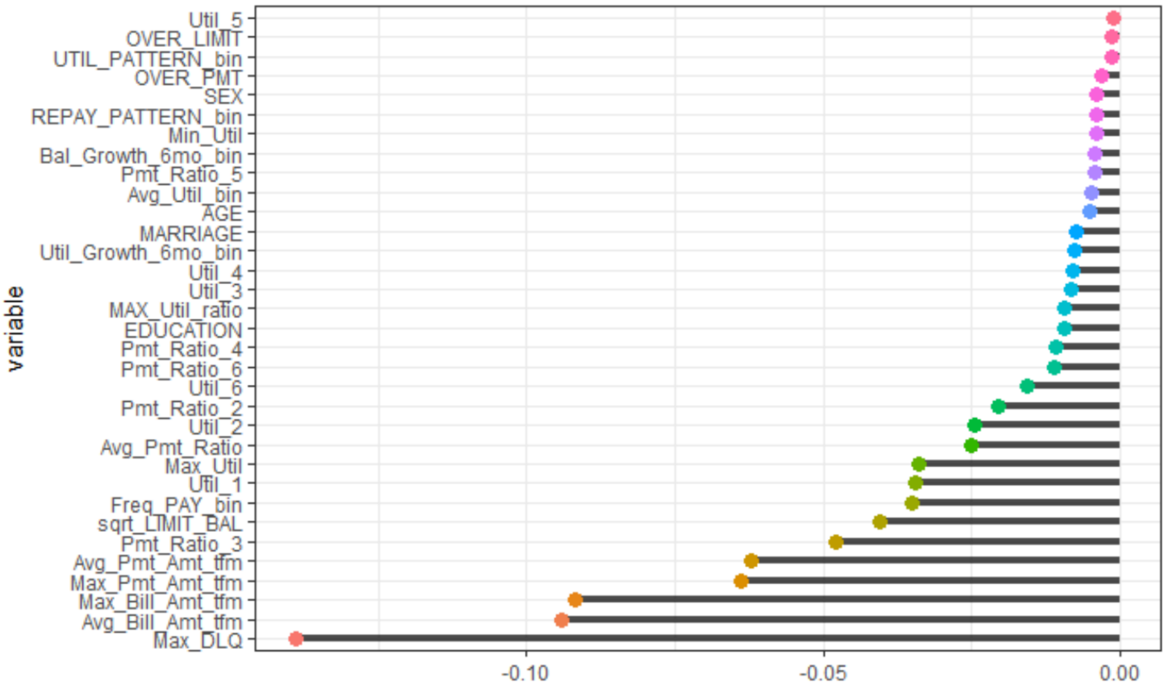


Figure 37. Variable importance plot (F1 measure).



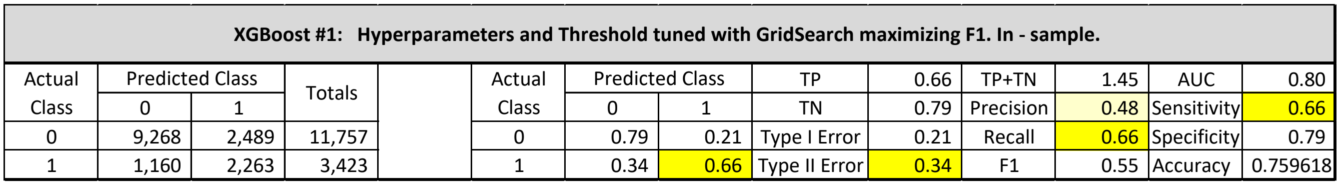
5.b Gradient Boosting

Extreme Gradient Boosting model has been built next. The method applies the principle of boosting weak learners using the gradient descent architecture. The algorithm is efficient in terms its training time and the use of hardware but most importantly it proved to produce decent result and handles overfitting through regularization[9]**.** The XGBoost algorithm has been fitted to the transformed train data and validated using cross- validation and test data. GridSearch is used to find the most optimal hyperparameters to obtain the highest F1 score. According to two discriminatory measures, AUC (0.76) and KS (0.42), the resulted model is able to discriminate the good customers from bad customers which

[99] XGboost penalizes more complex models through both LASSO (L1) and Ridge (L2) regularization to prevent overfitting.

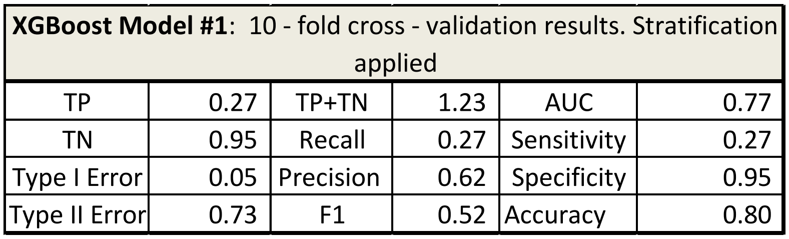
indicates that the model works well. Moreover, the performance of the model in-sample and out- of-sample are similar with less indication of overfitting on the training data than in case of Random forest models. The overall predictive capability of the model is decent with 61% default customers' detection rate and F1 score of 52.

Table 23.



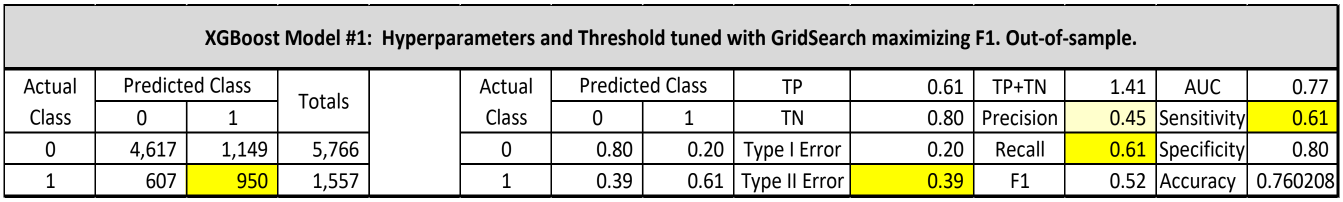
XGBoost #1: Hyperparameters and Threshold tuned with GridSearch maximizing F1. In - sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.66	TP+TN	1.45	AUC	0.80
Class	0	1			Class	0	1	TN	0.79	Precision	0.48	Sensitivity	0.66
0	9,268	2,489	11,757		0	0.79	0.21	Type I Error	0.21	Recall	0.66	Specificity	0.79
1	1,160	2,263	3,423		1	0.34	0.66	Type II Error	0.34	F1	0.55	Accuracy	0.759618

Table 24.



XGBoost Model #1: 10 - fold cross - validation results. Stratification applied					
TP	0.27	TP+TN	1.23	AUC	0.77
TN	0.95	Recall	0.27	Sensitivity	0.27
Type I Error	0.05	Precision	0.62	Specificity	0.95
Type II Error	0.73	F1	0.52	Accuracy	0.80

Table 25.

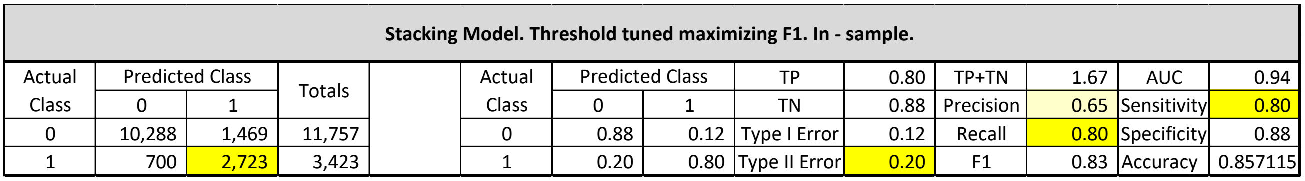


XGBoost Model #1: Hyperparameters and Threshold tuned with GridSearch maximizing F1. Out-of-sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.61	TP+TN	1.41	AUC	0.77
Class	0	1			Class	0	1	TN	0.80	Precision	0.45	Sensitivity	0.61
0	4,617	1,149	5,766		0	0.80	0.20	Type I Error	0.20	Recall	0.61	Specificity	0.80
1	607	950	1,557		1	0.39	0.61	Type II Error	0.39	F1	0.52	Accuracy	0.760208

5.c. Stacking model

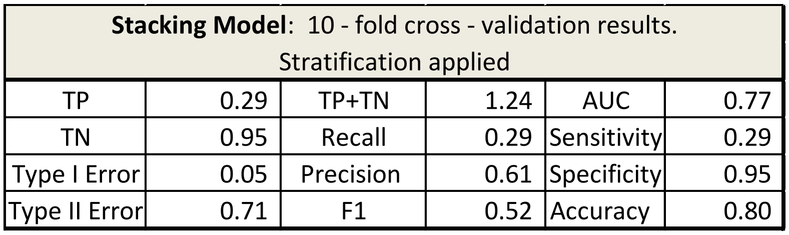
Stacking model has been created by combining predictive powers of Random Forest model tuned with Random Search strategy to maximize TPR and XGboost model tuned with Grid Search to maximize F1 score. The predictions of base learners without weights have been averaged to obtain a final set of predictions. The resulted model was expected to achieve two goals: maximize TPR and keep FPR to a minimum. The model met the established goals based on the results obtained from the training data with a strong AUC score of 0.94, F1 score of 0.83, and TPR of 0.80 and TNR of 0.88 (see Table 26.). The evaluation of the classifier's quality in cross-validation and on the test data produced less optimistic results with considerably higher Type II error but the model is still competitive. The obtained KS value of 0.4142111 and AUC of 0.77 on the test data set suggest that the model discriminates the classes well.

Table 26.



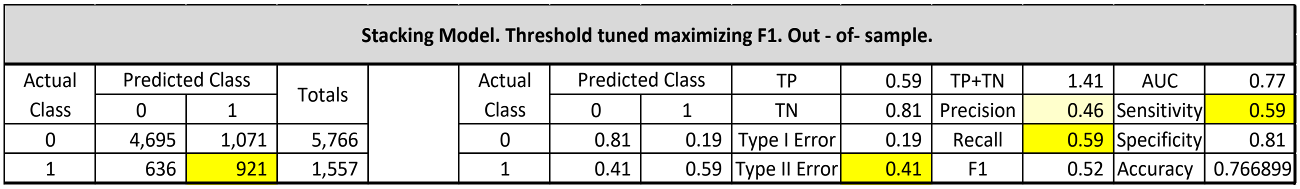
Stacking Model. Threshold tuned maximizing F1. In - sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.80	TP+TN	1.67	AUC	0.94
Class	0	1			Class	0	1	TN	0.88	Precision	0.65	Sensitivity	0.80
0	10,288	1,469	11,757		0	0.88	0.12	Type I Error	0.12	Recall	0.80	Specificity	0.88
1	700	2,723	3,423		1	0.20	0.80	Type II Error	0.20	F1	0.83	Accuracy	0.857115

Table 27.



Stacking Model: 10 - fold cross - validation results. Stratification applied					
TP	0.29	TP+TN	1.24	AUC	0.77
TN	0.95	Recall	0.29	Sensitivity	0.29
Type I Error	0.05	Precision	0.61	Specificity	0.95
Type II Error	0.71	F1	0.52	Accuracy	0.80

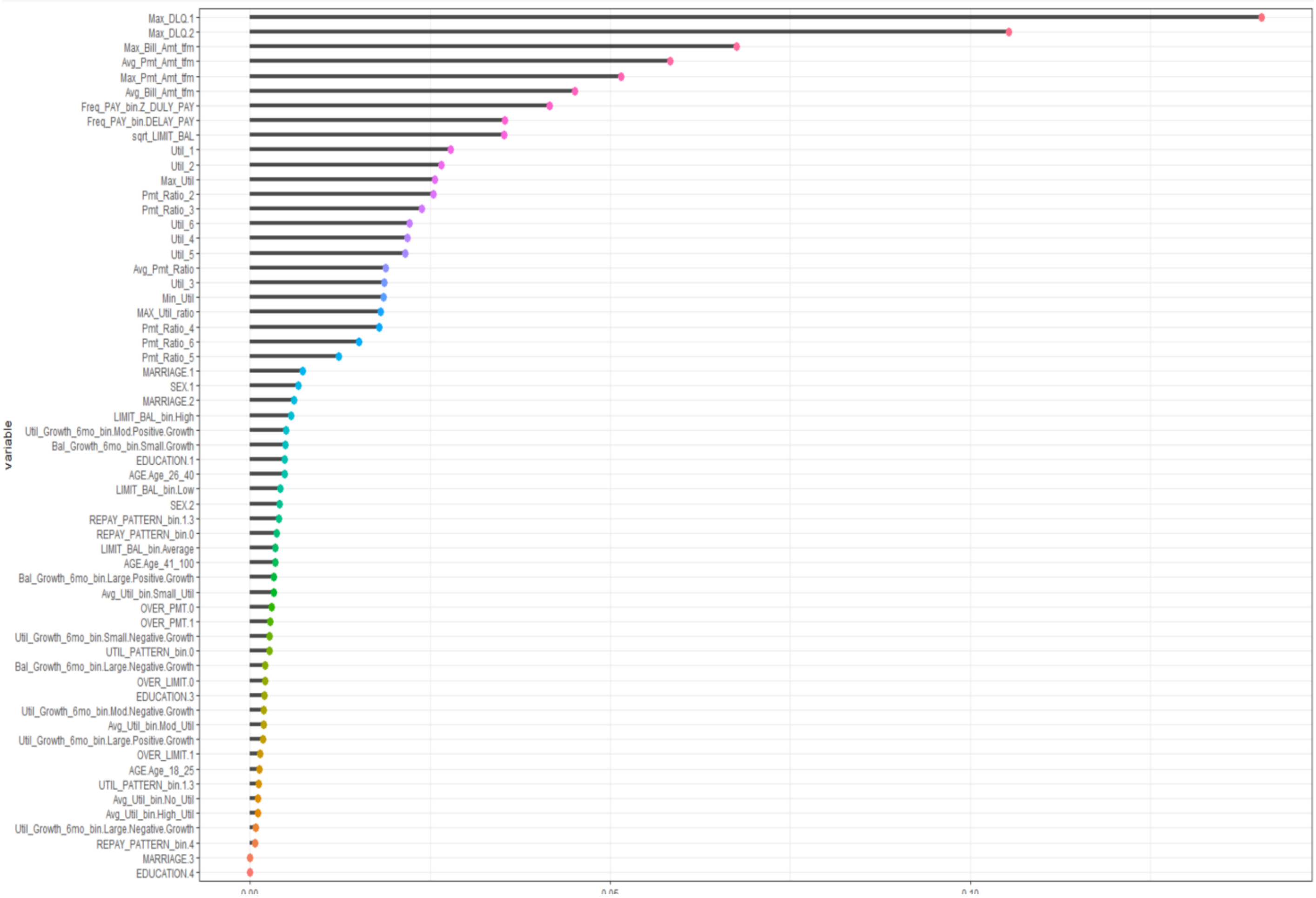
Table 28.



Stacking Model. Threshold tuned maximizing F1. Out - of- sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.59	TP+TN	1.41	AUC	0.77
Class	0	1			Class	0	1	TN	0.81	Precision	0.46	Sensitivity	0.59
0	4,695	1,071	5,766		0	0.81	0.19	Type I Error	0.19	Recall	0.59	Specificity	0.81
1	636	921	1,557		1	0.41	0.59	Type II Error	0.41	F1	0.52	Accuracy	0.766899

The importance of features measured by contribution to decrease in Gini impurity is illustrated in Figure 38.

Figure 38. Variable importance plot (Gini Index).



The model's out-of-sample performance in terms of FPR and TPR and Precision and Recall for various cutoff values are illustrated in ROC (see Figure 39) and with Precision – Recall (see Figure 40.) curves respectively.

Figure 39.

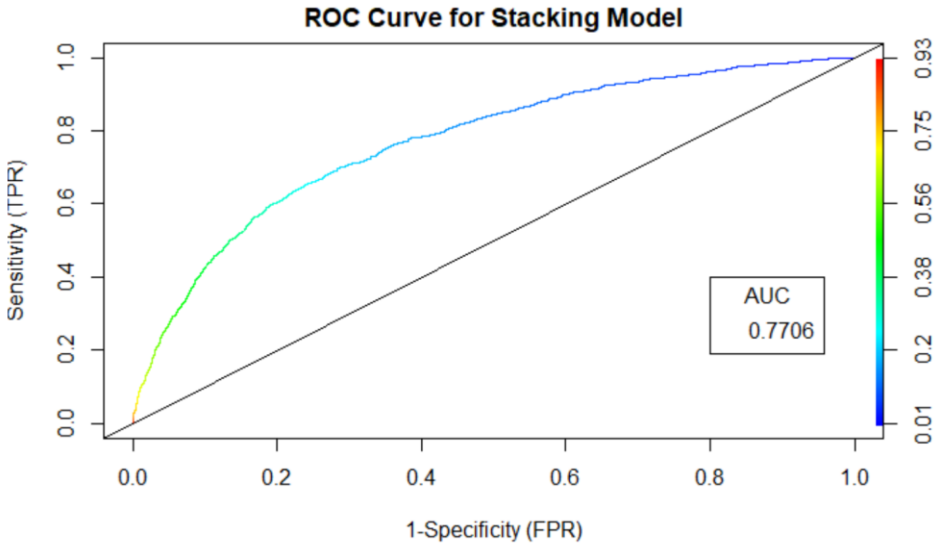
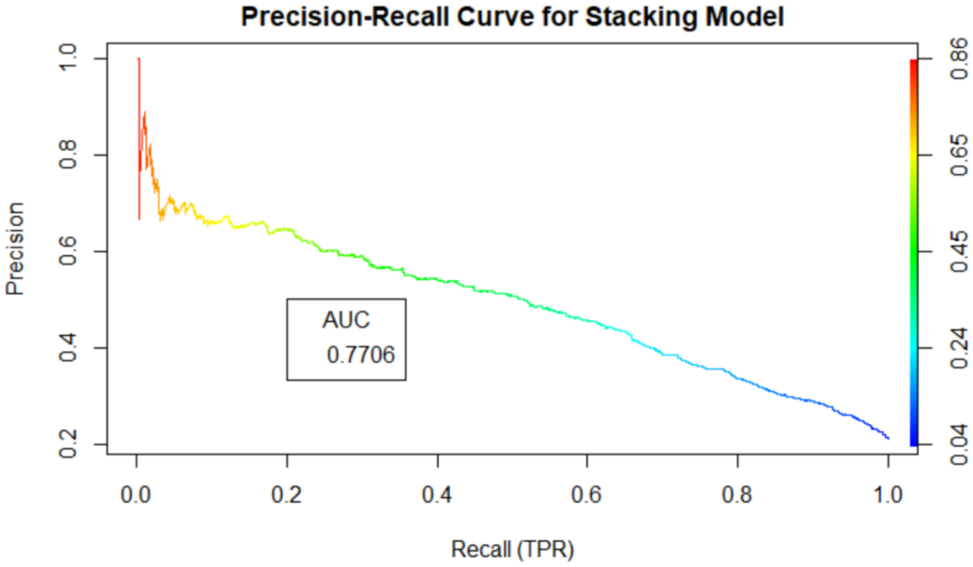


Figure 40.



5.d Logistic Regression with Variable Selection

Three regression models from a set of candidate predictor variables have been built using automated forward, stepwise and backward variable selection methods, that remove or enter, or remove and enter predictors based on AIC values in a stepwise manner until there is no variable left to remove (enter) any more. Another three models have been created using the same Automated Variable Selection approaches but including interaction terms identified using ANOVA and statistical significance test (see Appendix E). Finally, another model has been built using only variables that have been identified as important by Random forest according to various performance metrics such as F1, Gini Index and TPR as well as Rpart and Boruta algorithms. The scoring method has been used to eliminate the useless features as follows, one point is granted to a feature if it is identified as strong predictor, -1 if it is rejected by any algorithm and -0.5 if it is condemned as a tentative predictor (see Appendix F). As a result SEX, EDUCATION and UTIL_PATTERN_bin have been excluded from consideration and the stepwise approach used on the rest of predictors to obtain the final model. According to evaluation results from all six models based on their in – sample and out-of- sample performances using AIC, KAPPA, Precision and Recall metrics, it was concluded that most of the models have similar predictive qualities and the analysis has proceeded with the model created using the predefined set of predictors and stepwise variable selection algorithm. The coefficient table revealed that MARRIAGE variable is not statistically significant under 0.05 statistical significance level, and thus has been excluded from the model without compromising the predictive performance of the model (see Appendix G). VIF values for Util_1 and Util_2 are larger than 10 suggesting that the variables are correlated but since it does not affect predictive capability of the model and values are in a tolerable range, they were preserved in the model. The final model' coefficients are presented in the Table 29 below.

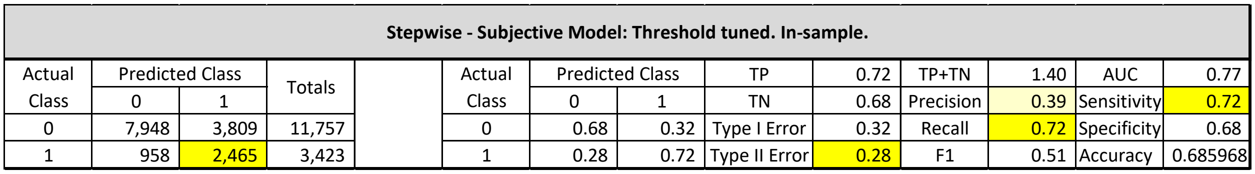
Table 29: The Coefficient table. VIF values included

	Est.	2.5%	97.5%	z val.	p	VIF
(Intercept)	-0.24	-0.45	-0.02	-2.17	0.03	
Util_1	0.50	0.08	0.91	2.34	0.02	15.26
Util_2	0.45	0.05	0.84	2.21	0.03	13.79
Util_5	-0.39	-0.67	-0.10	-2.67	0.01	5.76
MAX_Util_ratio	-0.41	-0.59	-0.23	-4.45	0.0000	2.73

Avg_Pmt_Amt_tfm	-0.10	-0.13	-0.07	-6.18	0	3.08
Max_Bill_Amt_tfm	-0.01	-0.02	-0.001	-2.24	0.02	7.31
WOE_Freq_PAY	-0.005	-0.01	-0.004	-14.21	0	1.30
WOE_Avg_Util	-0.003	-0.01	-0.0001	-2.02	0.04	4.17
WOE_LIMIT_BAL	-0.002	-0.003	-0.0003	-2.37	0.02	2.18
WOE_MAX_DLQ	-0.01	-0.01	-0.01	-27.90	0	1.41
WOE_OVER_PMT	-0.003	-0.005	-0.001	-3.40	0.001	1.41
WOE_REPAY_PATTERN	-0.004	-0.01	-0.002	-3.28	0.001	1.33
Pmt_Ratio_6	0.22	0.08	0.37	2.97	0.003	2.53

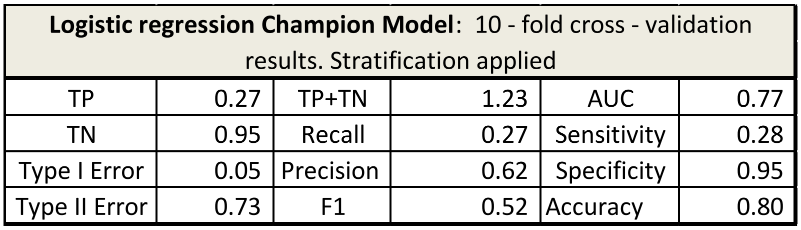
The model produced good results in – sample with KS value equal to 0.4143922 and AUC value of 0.7653802. The results on the test data set does not differ significantly from the results obtained on the training data set with KS equal to 0.4128808 and AUC of 0.7632956 which suggest that the model does not overfit the training data and could produce equally good results on seen and unseen instances. Overall, the quality of the model is satisfactory with tolerable Type II error on the test data of 0.28 and F1 value of 0.49 (see Table 32).

Table 30.



Stepwise - Subjective Model: Threshold tuned. In-sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.72	TP+TN	1.40	AUC	0.77
Class	0	1			Class	0	1	TN	0.68	Precision	0.39	Sensitivity	0.72
0	7,948	3,809	11,757		0	0.68	0.32	Type I Error	0.32	Recall	0.72	Specificity	0.68
1	958	2,465	3,423		1	0.28	0.72	Type II Error	0.28	F1	0.51	Accuracy	0.685968

Table 31.



Logistic regression Champion Model: 10 - fold cross - validation results. Stratification applied					
TP	0.27	TP+TN	1.23	AUC	0.77
TN	0.95	Recall	0.27	Sensitivity	0.28
Type I Error	0.05	Precision	0.62	Specificity	0.95
Type II Error	0.73	F1	0.52	Accuracy	0.80

Table 32.

Stepwise - Subjective Model: Threshold tuned. Out-of-sample.													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.72	TP+TN	1.39	AUC	0.76
Class	0	1			Class	0	1	TN	0.67	Precision	0.37	Sensitivity	0.72
0	3,840	1,926	5,766		0	0.67	0.33	Type I Error	0.33	Recall	0.72	Specificity	0.67
1	433	1,124	1,557		1	0.28	0.72	Type II Error	0.28	F1	0.49	Accuracy	0.677864

The model's out-of-sample performance in terms of precision and recall and FPR and TPR for various cutoff values are illustrated with Precision – Recall curve (see Figure 42.) and ROC curve respectively (see Figure 41.).

Figure 41.

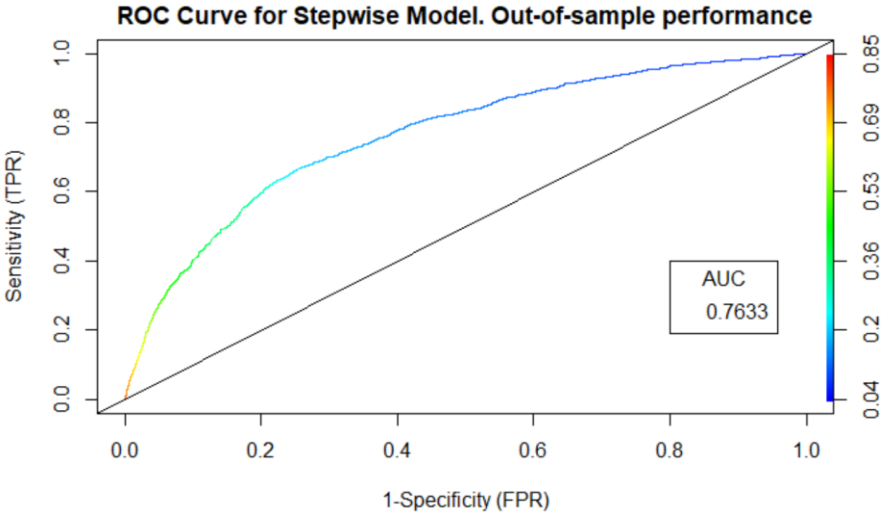
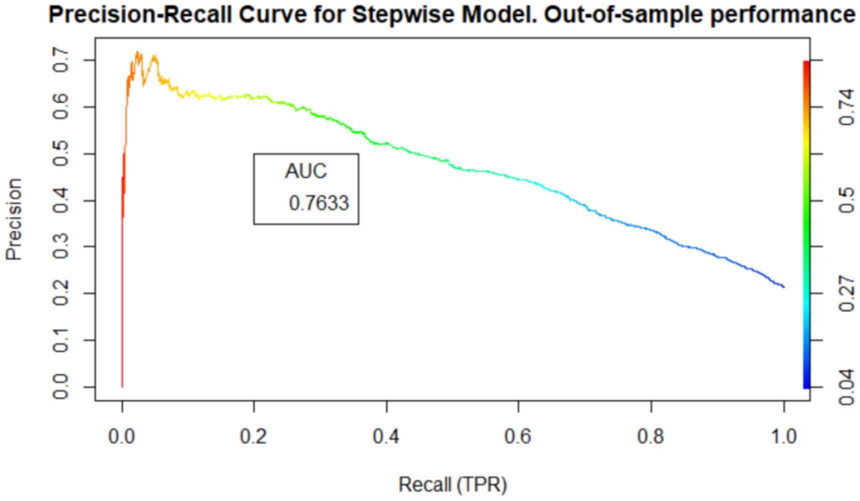


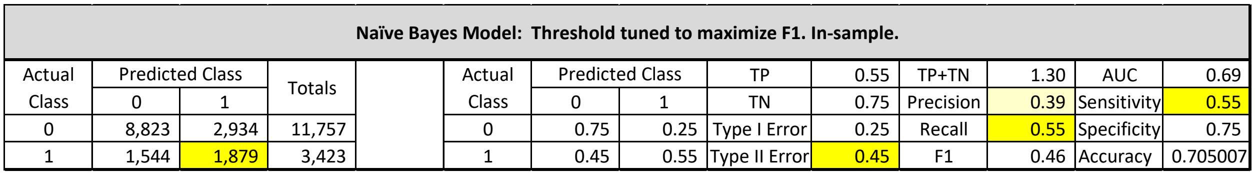
Figure 42.



5.e. Naïve Bayes

The Naive Bayes model produced relatively good results being able to accurately predict around 70% of no-defaulters and 55% defaulters in-sample and out-of-sample. KS values of 0.280 and 0.299 produced on test and train data respectively signal the model's poor discriminatory power. Also, relatively low AUC values have been obtained from predictions on both train and test data sets as well as in cross-validation, not exceeding 0.69.

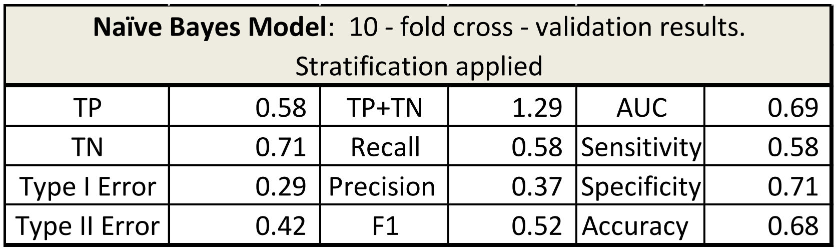
Table 33.



Naïve Bayes Model: Threshold tuned to maximize F1. In-sample.

Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.55	TP+TN	1.30	AUC	0.69
Class	0	1			Class	0	1	TN	0.75	Precision	0.39	Sensitivity	0.55
0	8,823	2,934	11,757		0	0.75	0.25	Type I Error	0.25	Recall	0.55	Specificity	0.75
1	1,544	1,879	3,423		1	0.45	0.55	Type II Error	0.45	F1	0.46	Accuracy	0.705007

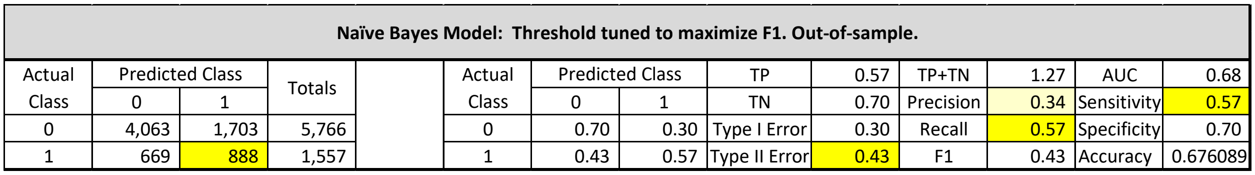
Table 34.



Naïve Bayes Model: 10 - fold cross - validation results. Stratification applied

TP	0.58	TP+TN	1.29	AUC	0.69
TN	0.71	Recall	0.58	Sensitivity	0.58
Type I Error	0.29	Precision	0.37	Specificity	0.71
Type II Error	0.42	F1	0.52	Accuracy	0.68

Table 35.



Naïve Bayes Model: Threshold tuned to maximize F1. Out-of-sample.

Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.57	TP+TN	1.27	AUC	0.68
Class	0	1			Class	0	1	TN	0.70	Precision	0.34	Sensitivity	0.57
0	4,063	1,703	5,766		0	0.70	0.30	Type I Error	0.30	Recall	0.57	Specificity	0.70
1	669	888	1,557		1	0.43	0.57	Type II Error	0.43	F1	0.43	Accuracy	0.676089

The classifier's out-of-sample performance in terms of precision and recall and FPR and TPR for various cutoff values are illustrated with Precision – Recall curve (see Figure 44.) and ROC curve respectively (see Figure 43.).

Figure 43.

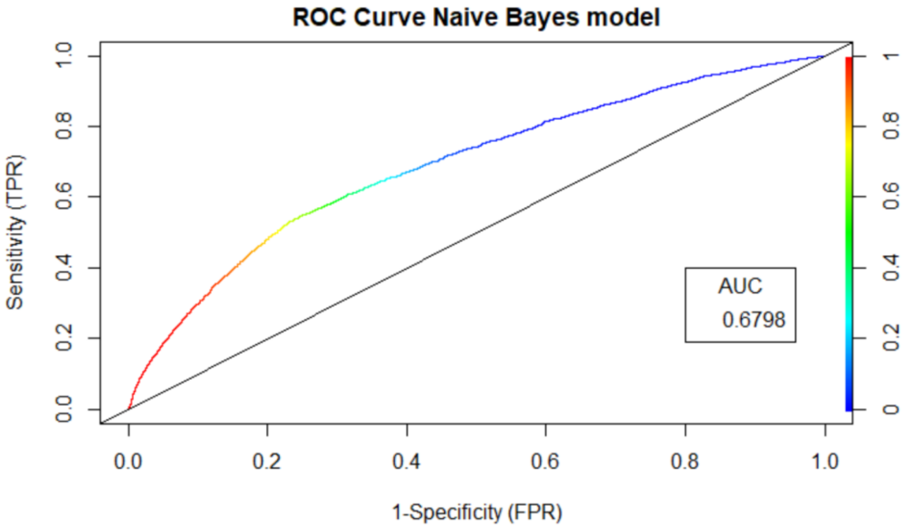
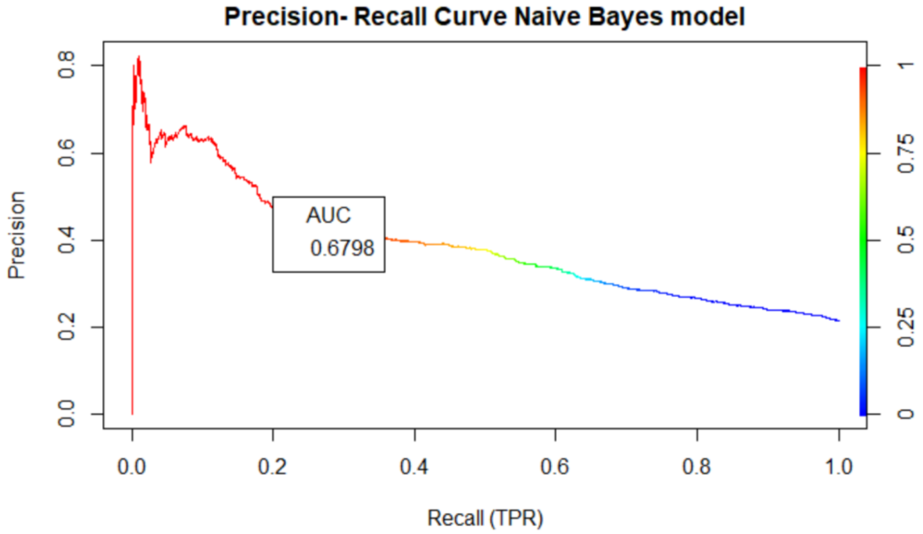


Figure 44.



6. Comparison of Results

The classifiers have been assessed and compared in terms their ability to produce accurate predictions (TP+TN), their ability to discriminate defaulters from non-defaulters (KS and AUC), realize trade-off between Recall and Precision, and their capacity to catch the minority class which is the class of interest. The comparison table suggests that Naïve Bayes classifier performed the worst out of all built models.

The logistic regression built with predefined pool of important predictors and stepwise variable selection strategy produced the highest Recall value by being able to catch more risky customers than any other models. If the cost of committing Type II error is significantly higher than Type I error, the model might be the most optimal to use. However, logistic regression model is not as accurate as other models with the sum of rates of true negatives and true positives equal to 1.39 which is 0.3 less than the most accurate model. The reason is low Precision. The best performance results have been produced by Champion Random Forest, XGboost and Stacking models with not significant difference in their performances. Nevertheless, Stacking model achieved the highest Precision but the lowest out of three Recall. While two other models have Precision values which are by 0.1 lower than Precision obtained by Stacking model, they achieved Recall which is higher by at least 0.2 (Random Forest 0.3) than that obtained by Stacking model. Assuming that the cost of Type II error is higher than Type I, the increase in Recall considerably offset the 0.1 reduction in Precision. Thus, two models, Random Forest and XGboost have more satisfactory performance results than Stacking model. Random Forest outperformed XGBoost on several important metrics, KS (very insignificant difference though), Recall by 0.1, and TP+TN by 0.1. The models have the same values for Precision and F1 test. XGBoost has the highest AUC value which is 0.2 higher than Random Forest' AUC. However, the fact that XGboost generated similar results in-sample and out-of -sample which is not true for Random forest indicates that the model might produce better results on unseen data than Random Forest classifier since the latter is overfitted.

Table 36. Comparison table.

	AUC train	AUC test	KS train	KS test	Recall train	Recall test	Precision train	Precision test	F1 train	F1 test	TP+TN train	TP+TN test	Accuracy train	Accuracy test
Naïve Bayes Classifier	0.690582	0.6798	0.29977	0.28048	0.55	0.57	0.39	0.34	0.4563	0.43	1.3	1.27	0.70501	0.67689
Champion Random Forest	0.89	0.7528	0.648	0.417388	0.87	0.62	0.52	0.45	0.65	0.52	1.63	1.42	0.787088	0.759798
XGBoost	0.801927	0.77326	0.46116	0.416858	0.66	0.61	0.48	0.45	0.55	0.52	1.45	1.41	0.759618	0.76028
Stacking	0.939229	0.77063	0.727401	0.414211	0.8	0.59	0.65	0.46	0.83	0.52	1.67	1.41	0.857115	0.76698
Logistic Regression	0.76538	0.7633	0.414392	0.412881	0.72	0.72	0.39	0.37	0.51	0.49	1.4	1.39	0.685968	0.677864

7. Conclusions

As a result of the extensive assessment of the classifiers' performances, XGBoost has been proved to have the strongest predictive power (TP+TN of 1.41 is obtained on the test data) as well as produced the greatest separation between Good Credit and Bad Credit cases observed in the test data, and the best ability to identify customers who failed their financial responsibilities to a bank. Since in credit scoring context there is a high cost associated with False Negative, the Recall value has been monitored and was taken into equation while evaluating the models' qualities. XGBoost produced the second highest Recall value detecting 66% of customers in default on the training data and 61% on the test data. The classifier achieved F1 score of 0.52 on the test data which is a good indicator of the model ability to find balance between Precision and Recall. The threshold could be further adjusted based on a company's growth goals and risk appetite.

The prediction results obtained by the supervised algorithms are dependent on the quality of data and features derived. Thus, one of the ways to improve is to have more features or more data. Perhaps, if more features such as income, number of accounts, the length of credit history, any charge offs, debt settlements, lawsuits, wage garnishments or public judgments against a customer, and etc.. are obtained, there are the chances to arrive at a better model using the same supervised modeling approaches. Moreover, a larger dataset might expose a different and perhaps more balanced perspective on the classes. The model's performance could be also boosted by evening-up the classes using various methods such as oversampling, undersampling or generating synthetic samples using SMOTE algorithm (the Synthetic Minority Over-sampling Technique)[10]. Unsupervised modeling techniques such as Neural network may be considered which proved to produce the state-of-the-art results.

[10] SMOTE is an oversampling method. It works by creating synthetic samples from the minor class instead of creating copies. The algorithm selects two or more similar instances (using a distance measure) and perturbing an instance one attribute at a time by a random amount within the difference to the neighboring instances (Brownlee, 2015).

References

Brownlee, J. (2015). Oversampling methods to improve the predictability of the minority class. Retrieved from https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes- in-your-machine-learning-dataset/

Thomas, L.C. (2009). Consumer Credit Models: Pricing, Profit, and Portfolios. New York, NY: Oxford University Press.

Appendix A.1.

Figure 1.

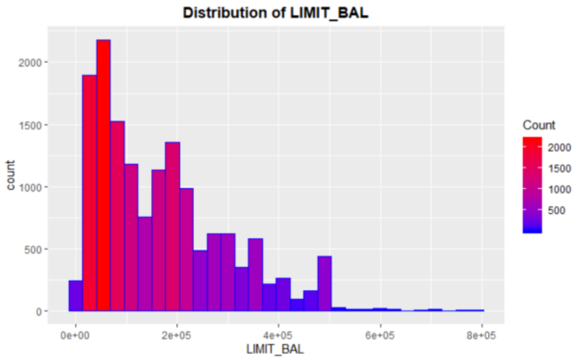


Figure 2.

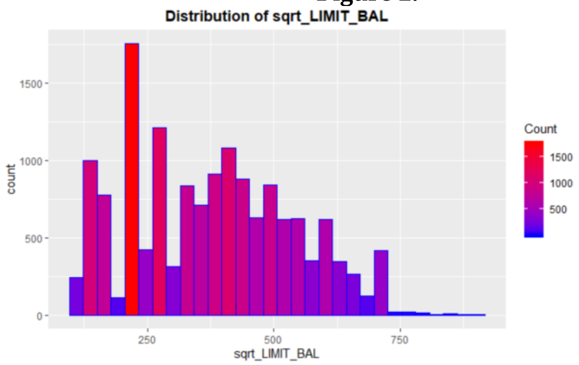


Figure 3.

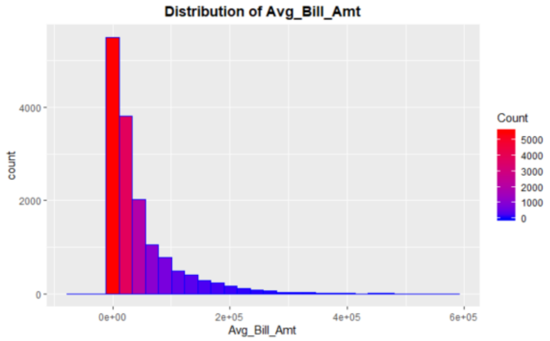


Figure 4.

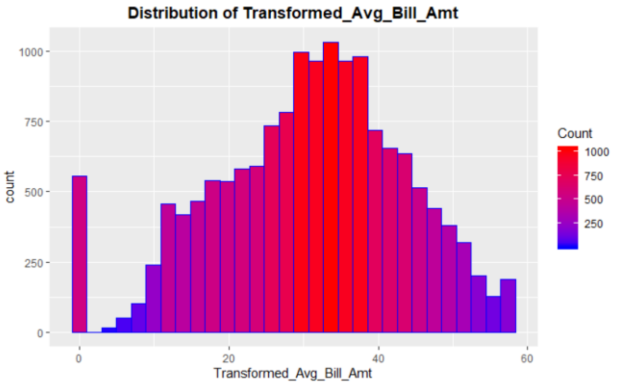


Figure 5.

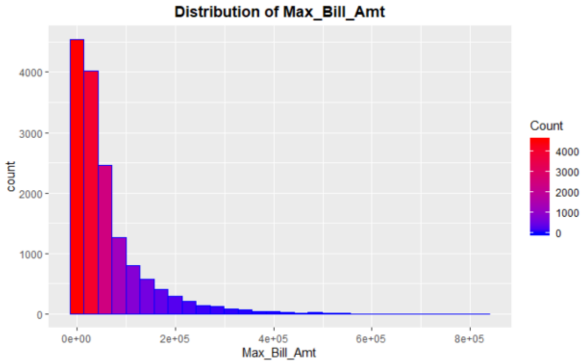


Figure 6.

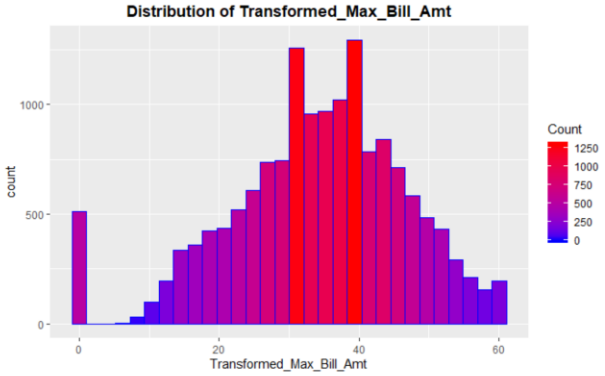


Figure 7.

Figure 8.

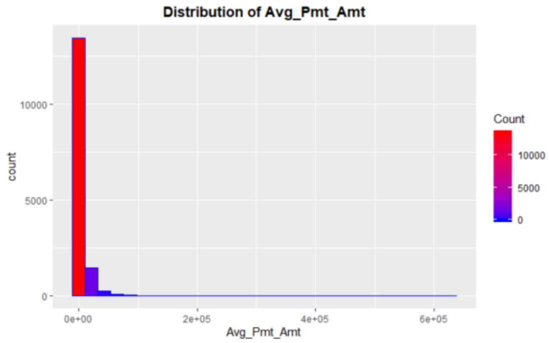


Figure 9.

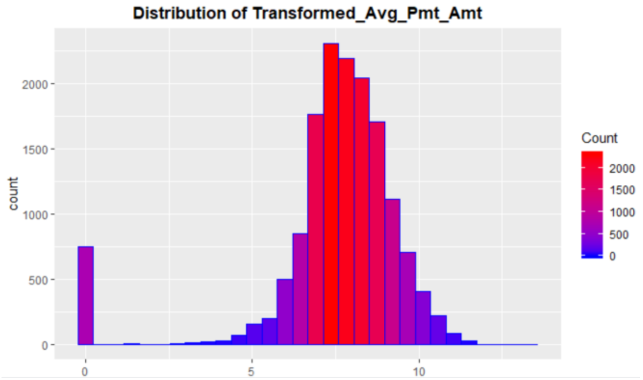


Figure 10.

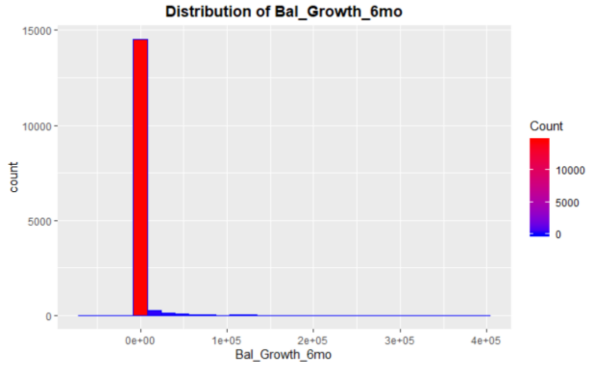


Figure 11.

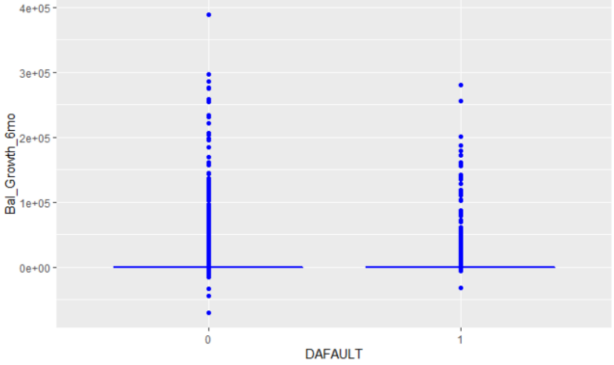
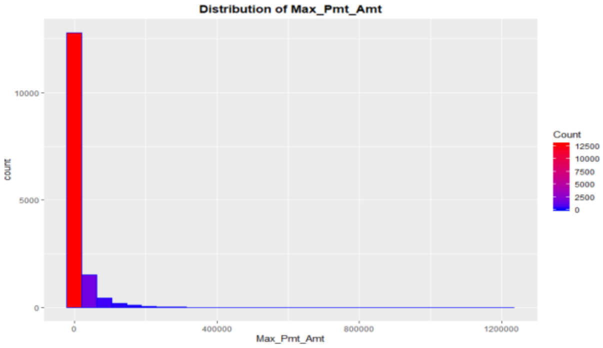


Figure 12.



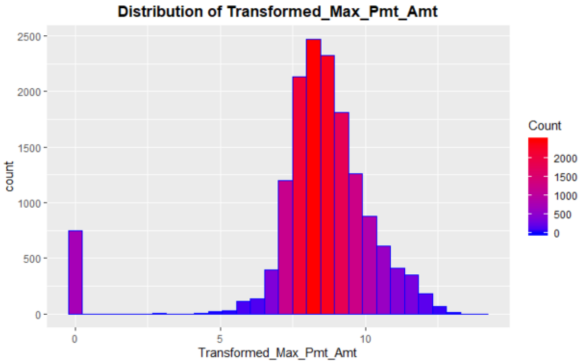
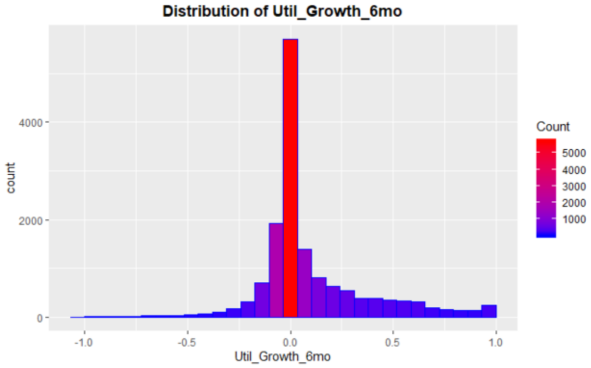


Figure 12.


Distribution of Util_Growth_6mo
4000
2000
0
count
-1.0
-0.5
0.0
0.5
1.0
Util_Growth_6mo
Count
5000
4000
3000
2000
1000

Appendix. A.2.

Figure 14.

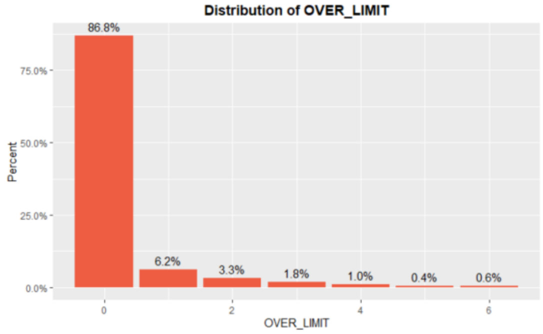


Figure 15.

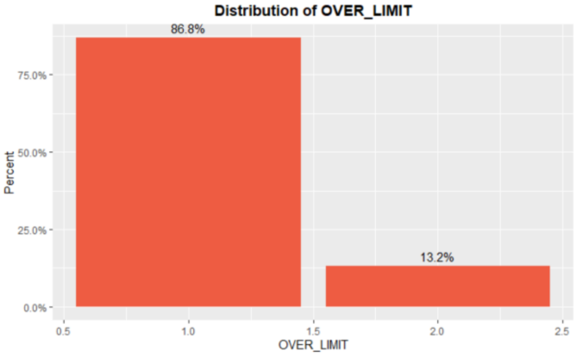


Figure 16.

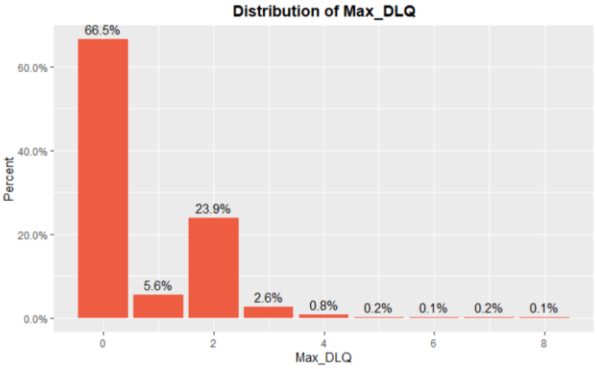


Figure 17.

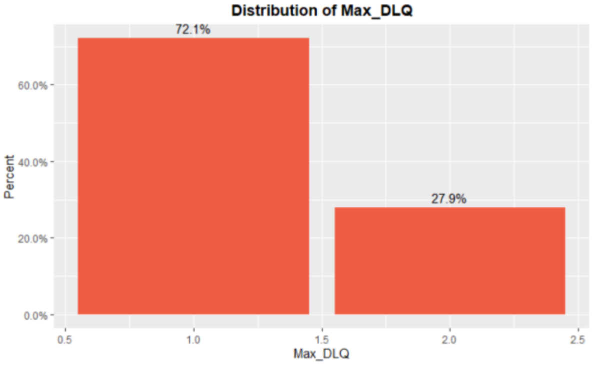


Figure 18.

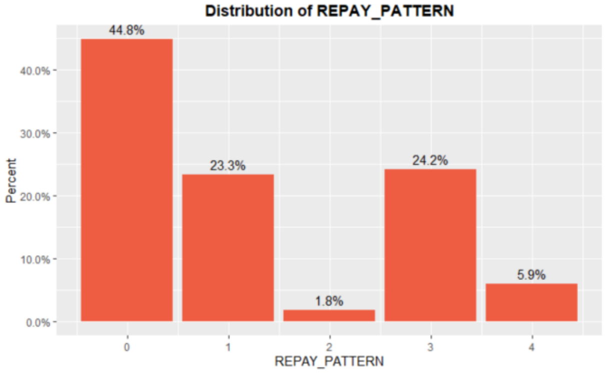
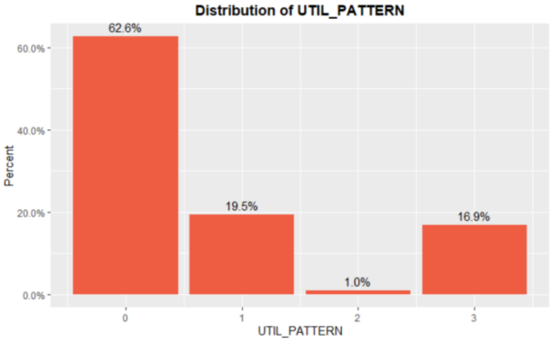
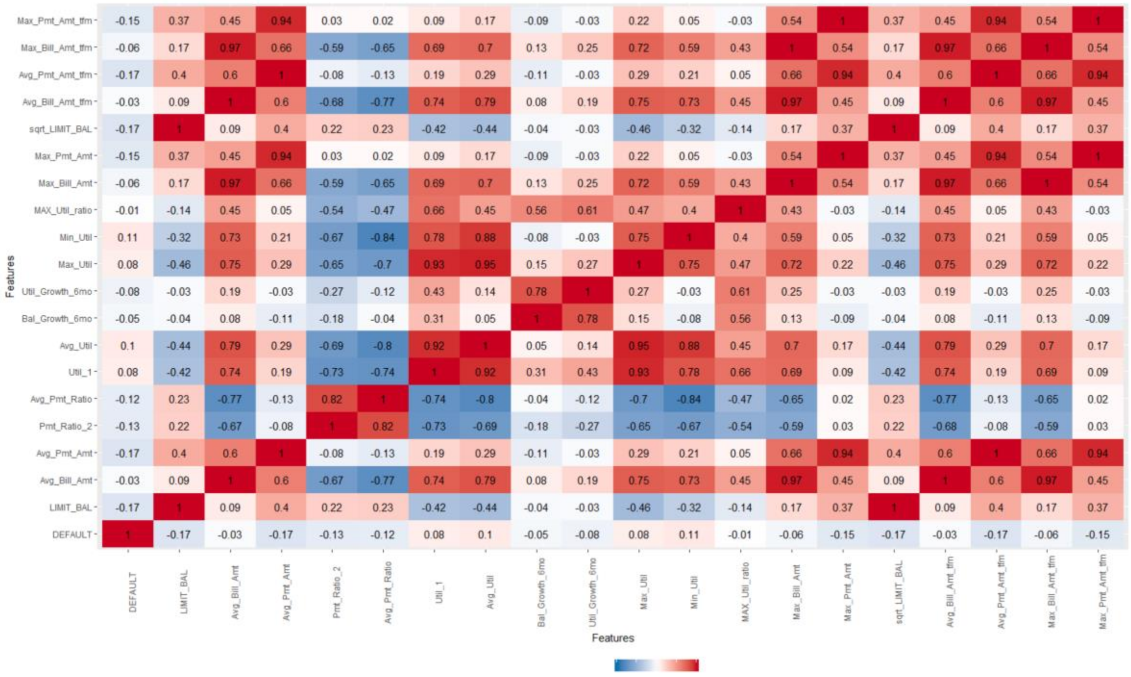


Figure 19.



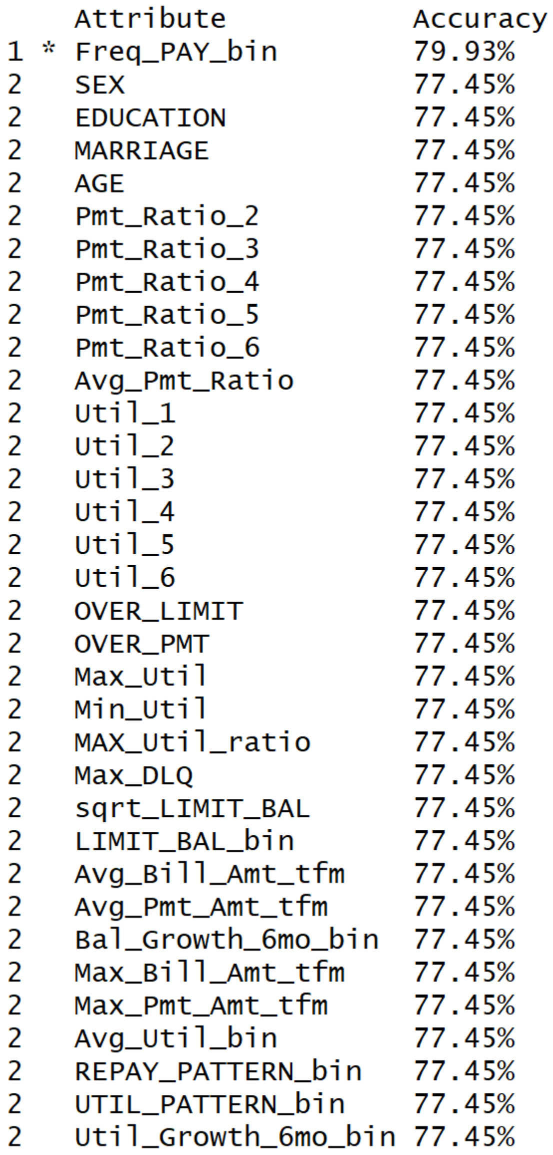
Appendix B.

Correlation matrix of continuoous variables. (Spearman)



	DEFAULT	LIMIT_BAL	Avg_Bill_Amt	Avg_Pmt_Amt	Pmt_Ratio_2	Avg_Pmt_Ratio	Util_1	Avg_Util	Bal_Growth_6mo	Util_Growth_6mo	Max_Util	Min_Util	MAX_Util_ratio	Max_Bill_Amt	Max_Pmt_Amt	sqrt_LIMIT_BAL	Avg_Bill_Amt_tfm	Avg_Pmt_Amt_tfm	Max_Bill_Amt_tfm	Max_Pmt_Amt_tfm
Max_Pmt_Amt_tfm	-0.15	0.37	0.45	0.94	0.03	0.02	0.09	0.17	-0.09	-0.03	0.22	0.05	-0.03	0.54	1	0.37	0.45	0.94	0.54	1
Max_Bill_Amt_tfm	-0.06	0.17	0.97	0.66	-0.59	-0.65	0.69	0.7	0.13	0.25	0.72	0.59	0.43	1	0.54	0.17	0.97	0.66	1	0.54
Avg_Pmt_Amt_tfm	-0.17	0.4	0.6	1	-0.08	-0.13	0.19	0.29	-0.11	-0.03	0.29	0.21	0.05	0.66	0.94	0.4	0.6	1	0.66	0.94
Avg_Bill_Amt_tfm	-0.03	0.09	1	0.6	-0.68	-0.77	0.74	0.79	0.08	0.19	0.75	0.73	0.45	0.97	0.45	0.09	1	0.6	0.97	0.45
sqrt_LIMIT_BAL	-0.17	1	0.09	0.4	0.22	0.23	-0.42	-0.44	-0.04	-0.03	-0.46	-0.32	-0.14	0.17	0.37	1	0.09	0.4	0.17	0.37
Max_Pmt_Amt	-0.15	0.37	0.45	0.94	0.03	0.02	0.09	0.17	-0.09	-0.03	0.22	0.05	-0.03	0.54	1	0.37	0.45	0.94	0.54	1
Max_Bill_Amt	-0.06	0.17	0.97	0.66	-0.59	-0.65	0.69	0.7	0.13	0.25	0.72	0.59	0.43	1	0.54	0.17	0.97	0.66	1	0.54
MAX_Util_ratio	-0.01	-0.14	0.45	0.05	-0.54	-0.47	0.66	0.45	0.56	0.61	0.47	0.4	1	0.43	-0.03	-0.14	0.45	0.05	0.43	-0.03
Min_Util	0.11	-0.32	0.73	0.21	-0.67	-0.84	0.78	0.88	-0.08	-0.03	0.75	1	0.4	0.59	0.05	-0.32	0.73	0.21	0.59	0.05
Max_Util	0.08	-0.46	0.75	0.29	-0.65	-0.7	0.93	0.95	0.15	0.27	1	0.75	0.47	0.72	0.22	-0.46	0.75	0.29	0.72	0.22
Util_Growth_6mo	-0.08	-0.03	0.19	-0.03	-0.27	-0.12	0.43	0.14	0.78	1	0.27	-0.03	0.61	0.25	-0.03	-0.03	0.19	-0.03	0.25	-0.03
Bal_Growth_6mo	-0.05	-0.04	0.08	-0.11	-0.18	-0.04	0.31	0.05	1	0.78	0.15	-0.08	0.56	0.13	-0.09	-0.04	0.08	-0.11	0.13	-0.09
Avg_Util	0.1	-0.44	0.79	0.29	-0.69	-0.8	0.92	1	0.05	0.14	0.95	0.88	0.45	0.7	0.17	-0.44	0.79	0.29	0.7	0.17
Util_1	0.08	-0.42	0.74	0.19	-0.73	-0.74	1	0.92	0.31	0.43	0.93	0.78	0.66	0.69	0.09	-0.42	0.74	0.19	0.69	0.09
Avg_Pmt_Ratio	-0.12	0.23	-0.77	-0.13	0.82	1	-0.74	-0.8	-0.04	-0.12	-0.7	-0.84	-0.47	-0.65	0.02	0.23	-0.77	-0.13	-0.65	0.02
Pmt_Ratio_2	-0.13	0.22	-0.67	-0.08	1	0.82	-0.73	-0.69	-0.18	-0.27	-0.65	-0.67	-0.54	-0.59	0.03	0.22	-0.68	-0.08	-0.59	0.03
Avg_Pmt_Amt	-0.17	0.4	0.6	1	-0.08	-0.13	0.19	0.29	-0.11	-0.03	0.29	0.21	0.05	0.66	0.94	0.4	0.6	1	0.66	0.94
Avg_Bill_Amt	-0.03	0.09	1	0.6	-0.67	-0.77	0.74	0.79	0.08	0.19	0.75	0.73	0.45	0.97	0.45	0.09	1	0.6	0.97	0.45
LIMIT_BAL	-0.17	1	0.09	0.4	0.22	0.23	-0.42	-0.44	-0.04	-0.03	-0.46	-0.32	-0.14	0.17	0.37	1	0.09	0.4	0.17	0.37
DEFAULT	1	-0.17	-0.03	-0.17	-0.13	-0.12	0.08	0.1	-0.05	-0.08	0.08	0.11	-0.01	-0.06	-0.15	-0.17	-0.03	-0.17	-0.06	-0.15

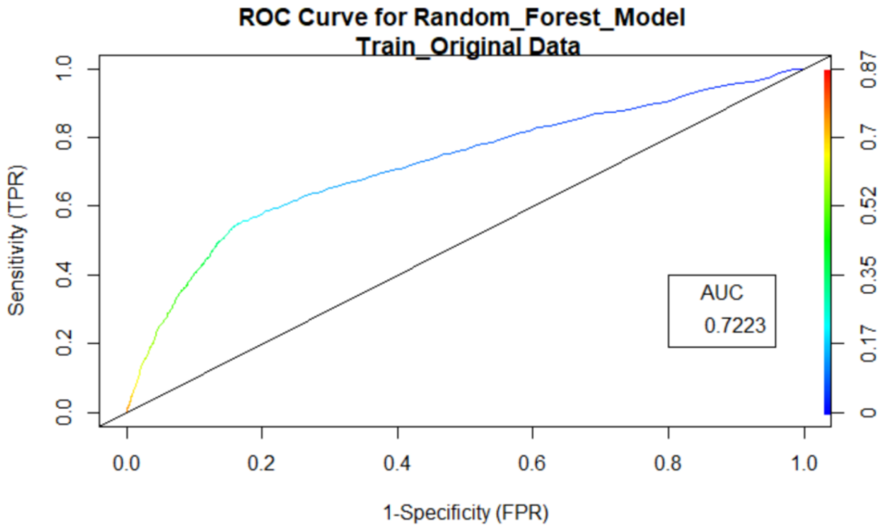
Appendix C.

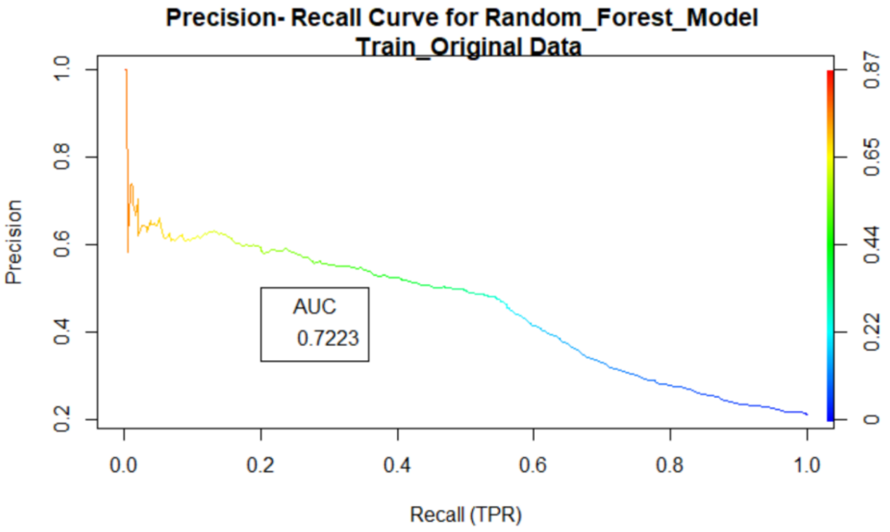


```
    Attribute           Accuracy
1 * Freq_PAY_bin        79.93%
2   SEX                 77.45%
2   EDUCATION           77.45%
2   MARRIAGE            77.45%
2   AGE                 77.45%
2   Pmt_Ratio_2         77.45%
2   Pmt_Ratio_3         77.45%
2   Pmt_Ratio_4         77.45%
2   Pmt_Ratio_5         77.45%
2   Pmt_Ratio_6         77.45%
2   Avg_Pmt_Ratio       77.45%
2   Util_1              77.45%
2   Util_2              77.45%
2   Util_3              77.45%
2   Util_4              77.45%
2   Util_5              77.45%
2   Util_6              77.45%
2   OVER_LIMIT          77.45%
2   OVER_PMT            77.45%
2   Max_Util            77.45%
2   Min_Util            77.45%
2   MAX_Util_ratio      77.45%
2   Max_DLQ             77.45%
2   sqrt_LIMIT_BAL      77.45%
2   LIMIT_BAL_bin       77.45%
2   Avg_Bill_Amt_tfm    77.45%
2   Avg_Pmt_Amt_tfm     77.45%
2   Bal_Growth_6mo_bin  77.45%
2   Max_Bill_Amt_tfm    77.45%
2   Max_Pmt_Amt_tfm     77.45%
2   Avg_Util_bin        77.45%
2   REPAY_PATTERN_bin   77.45%
2   UTIL_PATTERN_bin    77.45%
2   Util_Growth_6mo_bin 77.45%
```

Appendix D.

Model #1: Random Forest Model. Train_original Data used. RadomSearch F1 and Threshold tuned													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.61	TP+TN	1.37	AUC	0.72
Class	0	1			Class	0	1	TN	0.76	Precision	0.40	Sensitivity	0.61
0	4,355	1,411	5,766		0	0.76	0.24	Type I Error	0.24	Recall	0.61	Specificity	0.76
1	604	953	1,557		1	0.39	0.61	Type II Error	0.39	F1	0.66	Accuracy	0.72484





Appendix E.

Interaction terms table.

```
 SEX2:WOE_MAX_DLQ                         -1.601e-03  6.927e-04  -2.311  0.02085 *
EDUCATION1:WOE_Freq_PAY                  -3.851e-03  1.264e-03  -3.047  0.00231 **
 MARRIAGE1:WOE_Bal_Growth_6mo             -2.817e-02  1.323e-02  -2.129  0.03321 *
 MARRIAGE2:WOE_Bal_Growth_6mo             -2.708e-02  1.322e-02  -2.048  0.04055 *
 MARRIAGE1:WOE_OVER_LIMIT                  2.802e-02  1.473e-02   1.902  0.05716 .
 MARRIAGE2:WOE_OVER_LIMIT                  2.607e-02  1.468e-02   1.775  0.07585 .
 Pmt_Ratio_2:WOE_Avg_Util                  3.354e-01  1.494e-01   2.246  0.02474 *
 Pmt_Ratio_3:WOE_Avg_Util                  3.407e-01  1.492e-01   2.284  0.02237 *
 Pmt_Ratio_3:WOE_Util_Growth_6mo          -1.890e-01  9.117e-02  -2.073  0.03815 *
 Pmt_Ratio_4:WOE_Avg_Util                  3.308e-01  1.493e-01   2.215  0.02675 *
 Pmt_Ratio_4:WOE_Util_Growth_6mo          -1.863e-01  9.126e-02  -2.041  0.04121 *
 Pmt_Ratio_5:WOE_Avg_Util                  3.274e-01  1.493e-01   2.194  0.02827 *
 Pmt_Ratio_5:WOE_Util_Growth_6mo          -1.798e-01  9.114e-02  -1.973  0.04853 *
 Pmt_Ratio_6:WOE_Avg_Util                  3.426e-01  1.495e-01   2.292  0.02193 *
 Pmt_Ratio_6:WOE_Util_Growth_6mo          -1.789e-01  9.126e-02  -1.961  0.04991 *
 Avg_Pmt_Ratio:WOE_Avg_Util               -1.663e+00  7.458e-01  -2.230  0.02574 *
 Avg_Pmt_Ratio:WOE_Util_Growth_6mo         9.055e-01  4.550e-01   1.990  0.04656 *
Util_2:MAX_Util_ratio                     9.924e+00  3.686e+00   2.693  0.00709 **
 Util_2:sqrt_LIMIT_BAL                    -1.923e-02  9.359e-03  -2.055  0.03992 *
 Util_2:WOE_Avg_Util                      -6.060e-02  2.487e-02  -2.436  0.01483 *
 Util_3:MAX_Util_ratio                    -8.962e+00  3.793e+00  -2.363  0.01814 *
 Util_3:Avg_Pmt_Amt_tfm                    2.479e+00  9.843e-01   2.518  0.01179 *
Util_3:Max_Pmt_Amt_tfm                   -2.273e+00  8.465e-01  -2.685  0.00725 **
              Util_4:sqrt_LIMIT_BAL                     -1.612e-02  7.859e-03  -2.
 MAX_Util_ratio:WOE_Freq_PAY              -8.883e-03  4.055e-03  -2.191  0.02847 *
 MAX_Util_ratio:WOE_Avg_Util               1.469e-02  7.398e-03   1.986  0.04705 *
 Util_4:WOE_MAX_DLQ                       -8.214e-03  3.244e-03  -2.532  0.01134 *
 Max_Util:sqrt_LIMIT_BAL                   3.116e-02  1.347e-02   2.313  0.02074 *
 Max_Util:WOE_Avg_Util                     8.417e-02  3.664e-02   2.297  0.02163 *
 sqrt_LIMIT_BAL:WOE_Freq_PAY               2.385e-05  9.434e-06   2.528  0.01146 *
 sqrt_LIMIT_BAL:WOE_OVER_LIMIT             1.505e-04  6.690e-05   2.249  0.02449 *
 Avg_Bill_Amt_tfm:WOE_OVER_PMT             1.650e-03  8.027e-04   2.055  0.03984 *
Avg_Pmt_Amt_tfm:Max_Pmt_Amt_tfm          -2.412e-02  9.188e-03  -2.625  0.00866 **
Avg_Pmt_Amt_tfm:WOE_OVER_PMT             -9.349e-03  3.448e-03  -2.712  0.00670 **
 Max_Bill_Amt_tfm:WOE_Avg_Util            -1.908e-03  9.678e-04  -1.972  0.04864 *
 Max_Bill_Amt_tfm:WOE_OVER_PMT            -1.192e-03  6.383e-04  -1.867  0.06187 .
Max_Pmt_Amt_tfm:WOE_OVER_PMT              8.156e-03  2.956e-03   2.759  0.00579 **
 WOE_AGE:WOE_REPAY_PATTERN                -2.839e-04  1.613e-04  -1.761  0.07829 .
 WOE_Freq_PAY:WOE_LIMIT_BAL               -5.868e-05  2.419e-05  -2.426  0.01528 *
 WOE_Freq_PAY:WOE_OVER_LIMIT              -6.228e-05  2.726e-05  -2.285  0.02231 *
 WOE_LIMIT_BAL:WOE_UTIL_PATTERN            2.585e-04  1.317e-04   1.962  0.04973 *
 WOE_OVER_LIMIT:WOE_Util_Growth_6mo        1.957e-04  8.424e-05   2.322  0.02021 *
WOE_REPAY_PATTERN:WOE_Util_Growth_6mo     3.217e-04  1.019e-04   3.157  0.00159 **
WOE_Util_Growth_6mo:WOE_UTIL_PATTERN     -3.453e-04  1.137e-04  -3.035  0.00240 **
```

Appendix F.

	Rpart	Champion Random Forest F1	Champion Random Forest Gini Idx	Champion Random Forest TPR	Boruto	Negative score	Positive score	Total Score
sqrt_LIMIT_BAL	Confirmed	Confirmed	Confirmed	Rejected	Confirmed	1	1	0
SEX	Rejected	Tentative	Tentative	Confirmed	Rejected	3	0	-3
EDUCATION	Tentative	Rejected	Confirmed	Confirmed	Rejected	2	0	-2
MARRIAGE	Tentative	Confirmed	Confirmed	Confirmed	Tentative	1	0	-1
AGE	Rejected	Confirmed	Confirmed	Confirmed	Confirmed	1	0	-1
Avg_Bill_Amt_tfm	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	3	3
Avg_Pmt_Amt_tfm	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	3	3
Pmt_Ratio_2	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	1	1
Pmt_Ratio_3	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	2	2
Pmt_Ratio_4	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	1	1
Pmt_Ratio_5	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	1	1
Pmt_Ratio_6	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	1	1
Avg_Pmt_Ratio	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	1	1
Util_1	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	1	1
Util_2	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	0	0
Util_3	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	1	1
Util_4	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	0	0
Util_5	Confirmed	Rejected	Confirmed	Confirmed	Confirmed	1	0	-1
Util_6	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	0	0
Avg_Util_bin	Confirmed	Confirmed	Confirmed	Tentative	Confirmed	0.5	0	-0.5
OVER_LIMIT	Tentative	Rejected	Rejected	Tentative	Confirmed	3	0	-3

OVER_PMT	Tentative	Confirmed	Tentative	Tentative	Confirmed	1.5	0	-1.5
Bal_Growth_6mo_bin	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	0	0
Util_Growth_6mo_bin	Tentative	Confirmed	Confirmed	Tentative	Confirmed	1	0	-1
Max_Util	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	1	1
Min_Util	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	0	0
MAX_Util_ratio	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	0	0
Max_Bill_Amt_tfm	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	3	3
Max_Pmt_Amt_tfm	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	3	3
Max_DLQ	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	4	4
Freq_PAY_bin	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed	0	3	3
REPAY_PATTERN_bin	Rejected	Confirmed	Confirmed	Confirmed	Confirmed	1	0	-1
UTIL_PATTERN_bin	Tentative	Rejected	Rejected	Rejected	Confirmed	3.5	0	-3.5

Appendix G.

Table : The Coefficient table. VIF values included

	Est.	2.5%	97.5%	z val.	p	VIF
(Intercept)	-0.20	-0.63	0.23	-0.93	0.35	
MARRIAGE1	0.06	-0.32	0.44	0.31	0.75	1.03
MARRIAGE2	-0.12	-0.50	0.26	-0.64	0.52	1.03
Util_1	0.50	0.08	0.91	2.34	0.02	15.27
Util_2	0.45	0.05	0.85	2.22	0.03	13.80
Util_5	-0.38	-0.66	-0.10	-2.63	0.01	5.76
MAX_Util_ratio	-0.42	-0.60	-0.24	-4.52	0.0000	2.74
Avg_Pmt_Amt_tfm	-0.10	-0.13	-0.07	-6.14	0	3.08
Max_Bill_Amt_tfm	-0.01	-0.02	-0.001	-2.25	0.02	7.31
WOE_Freq_PAY	-0.005	-0.01	-0.004	-14.13	0	1.30
WOE_Avg_Util	-0.002	-0.005	0.0000	-1.96	0.05	4.16
WOE_LIMIT_BAL	-0.002	-0.004	-0.001	-2.69	0.01	2.20
WOE_MAX_DLQ	-0.01	-0.01	-0.01	-27.82	0	1.41
WOE_OVER_PMT	-0.003	-0.005	-0.001	-3.41	0.001	1.41
WOE_REPAY_PATTERN	-0.004	-0.01	-0.002	-3.19	0.001	1.33
Pmt_Ratio_6	0.22	0.07	0.37	2.90	0.004	2.53