STAT-551 Predictive Analytics I: Final project

Real Loans by Real People: What Should We Do to Keep Customers?

Kenneth Annan

05/05/2021

TABLE OF CONTENTS

Section I: Introduction

I. Background

In this project, I analyze real loans by real people data that come from a website *www.lendingclub.com.* My goal is to explore the data set and fit two models namely logistic regression and MARS (Multi-Adaptive Regression Splines) model for this analysis. The dataset contains 26,562 observations and 31 variables. For this project, only the *Education* and the M*onthsSinceLastRec* variables have more than 80 percent missing records. It is a good practice to delete those variables but for the sake of the project's instructions, I did not delete them but used a missing value imputation to handle it. For the education variable, I categorized it as Yes and NO. Yes, for all complete cases and NO for the missing cases. The MonthsSinceDQ variable has an unusual observation of 999, I converted those observations to NAs and used an imputation technique to handle it. I created another variable year to the dataset making a total of 32 variables. For the new variable year, I extracted the years of *when the date Loan was issued* and the *date of the first known credit line* and found the difference. The year variable measures the number of years a person has been using a credit line until he/she applies for the loan. I ranked the *appl_fico_band(FICO score band)* variable into good, very good, and excellent following the credit score rankings. The response variable 'Target' takes a value of *one* for a customer that did not pay his or her loan and is now seriously delinquent or in default and a value of *zero* for a customer that did not default (good customers). Among the 32 variables, I deleted 6 variables that have approximately zero variance (MOB, AccountsDQ, DelinquentAmount, DQ2yrs, PublicRec, current policy)- see data dictionary in the appendix section. I also deleted loan ID, duplicate term variable, issue date, state, earliest date. The reason for deleting the issue date and earliest date is that I created a year variable for it. After the data cleaning, I am left with all the 26,562 observations and 20 variables. I explored most of the categorical predictors, continuous predictors, and finally the response Versus most of the categorical variables. I randomly split the dataset into training data with 15,937 observations and validation data with 10,625 observations. For the MARS model, I used all the 20 variables except for *INFO* variables to fit the model and the model selected the important variables to be used for the prediction. However, for the logistic regression, I used the regsubset feature selection method to select four important variables. I binned two of the continuous variables(RevolvingLineUtilization and inquries6M) using the *WOE_custom*

command in R from the *Rprofet* library before fitting the logistic regression. The correlation plot shows that none of the continuous predictors used for fitting the logistic model are highly correlated (see Figure 20 in the appendix section). I then compared these two models using their Area Under the curve (AUC) and KS Value and the comparison went in favor of the MARS model as the best model.

II. Objective

The goal of this project is to make an appropriate model comparison, what we should do to retain customers, create gains and lift plot, compare how errors are measured within the context of logistic and MARS models when the dependent variable is categorical, and further use the Receiver Operating Characteristic (ROC) and Kolmogorov-Smirnov(KS) test as a model fit to detect the appropriate model. It is mostly best to aim at generating a model to predict the dependent variable target by selecting significant predictor variables, therefore, I used significant predictors for the logistic model and the MARS used significant variables by default. I will first create a MARS model and a logistic regression model, then decide on which model is better based on their ROC and the KS test statistic. This process is to fit a logistic and a MARS model based on the randomly selected training model, then apply each model to the validation data to evaluate performance. I made a visual display analysis of the variables used to build the two models and some of the variables in the dataset as seen in section two of this report.

Section 2: Findings, Conclusions, and Recommendations

I. Summary Statistics and Plots

The variables selected that best explain my chosen dataset and establish a relationship are explored below. The plot of the two predictors that I binned is in the appendix section of this report. Figures 1, 2, 3,4,5,6,7,8,9, and 10 show the histogram of variables DTI, amount of loan requested, monthly income, open credit line, total credit line, RevolvingCREDITBalance, RevolvingLineUtilization, years, interest rate and, target, respectively. I found that.

- Figures 1 3,5, and 8 show that the number of observations is somewhat close to normality in general, however, DTI and monthly income in logs are normally distributed with observations than the others.

- Figures 2 and 9 show that the interest rate and amount requested in logs are bimodal.
- Figure 10 shows that the number of customers who default(customers who do not pay their loans on time) is more than those who do not default.
- The boxplots in Figures 11a and 11b show that a customer will default or not depending on the interest rate, RevolvingLineUtilization, and Inquiries6M. The other variables explored in Figures 11a and 11b do not clearly determine whether a person will default or not default in the payment of loans.
- The plots in Figure 12 show that most loans for rent purposes and loans on a mortgage have the lowest probability of default. Also, most loans are for debt consolidation purposes and loans for small business has a higher chance of defaulting. Lastly, most loans are for people who have worked for 10 or more years in a certain company and people with 8 years working in a certain company have a lower chance of defaulting.

Figure 1: Histogram of DTI

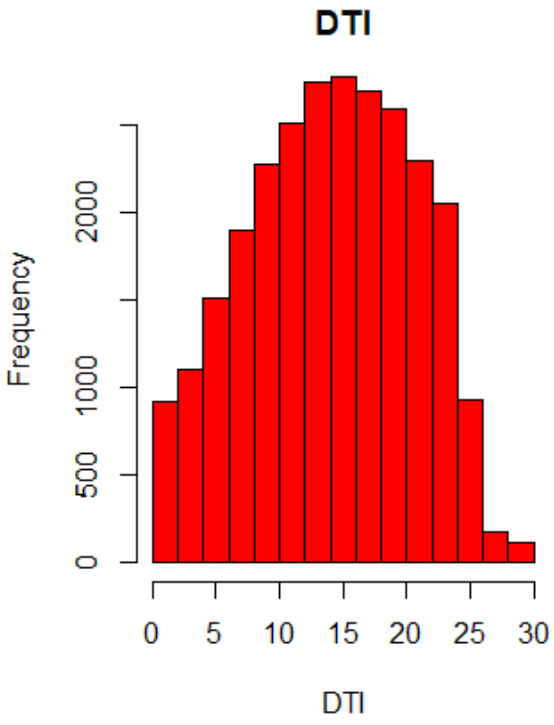


Figure 2: Histogram of Amount Requested

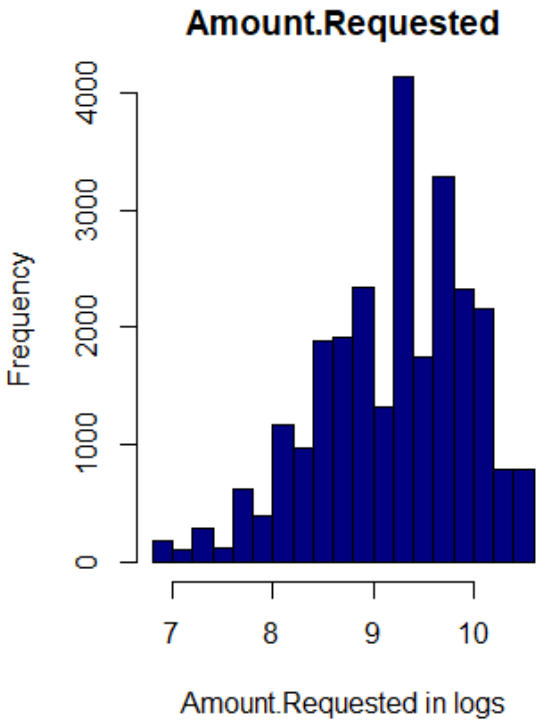


Figure 3: Histogram of Monthly Income

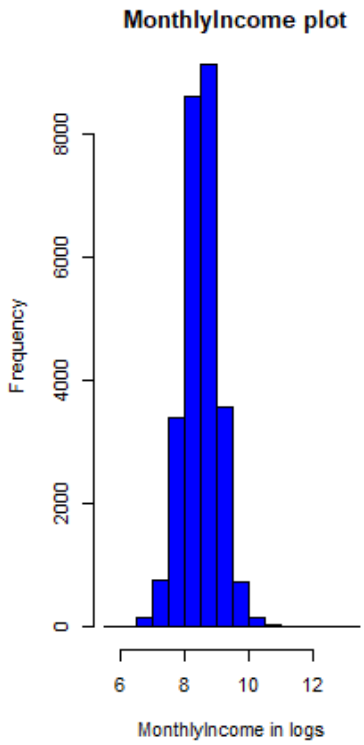


Figure 4: Histogram of open credit

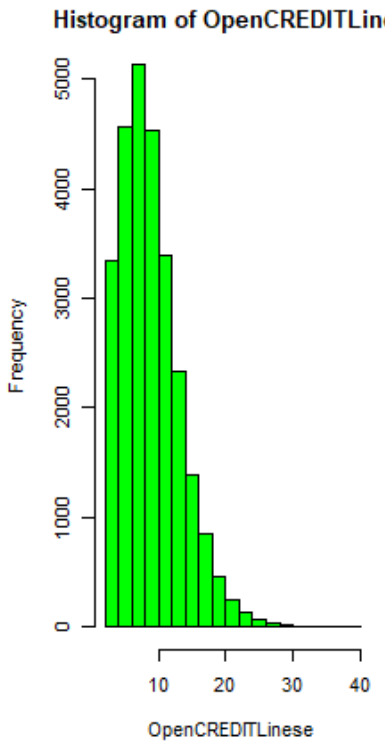


Figure 5: Histogram of Total credit

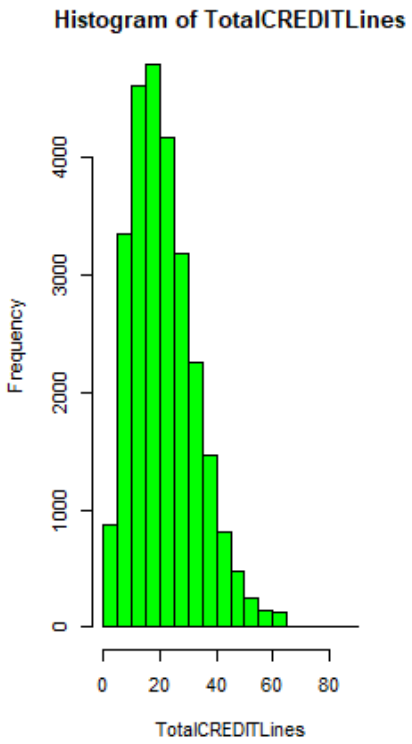


Figure 6: Histogram of credit balance

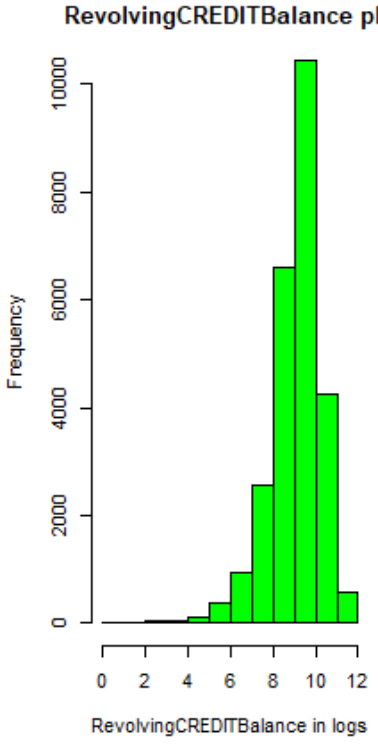


Figure 7: Histogram of line Utilization

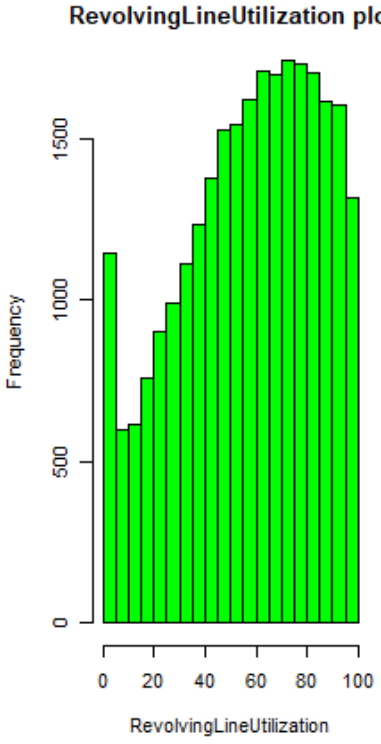


Figure 8: Histogram of years

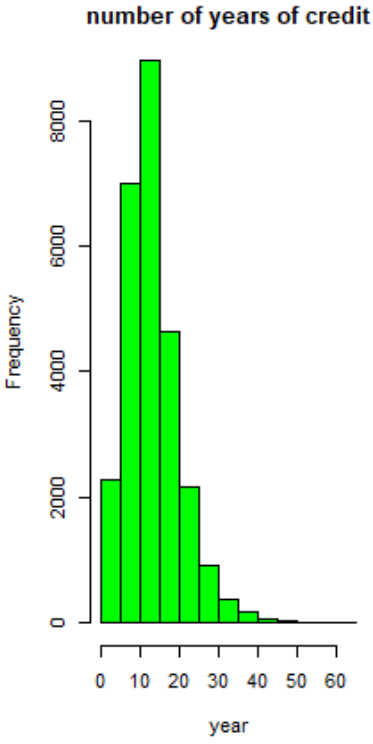


Figure 9: Histogram of interest rate

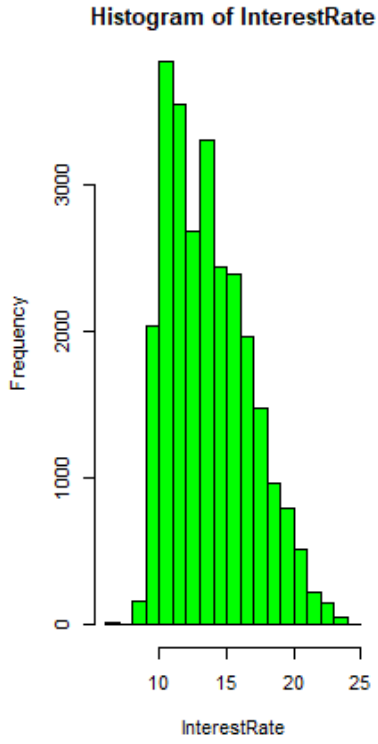


Figure 10: Histogram of target

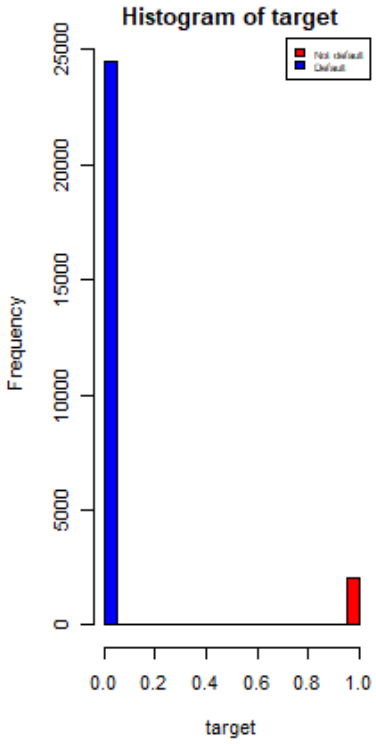


Figure 11a: Boxplot of target Vs the independent variables

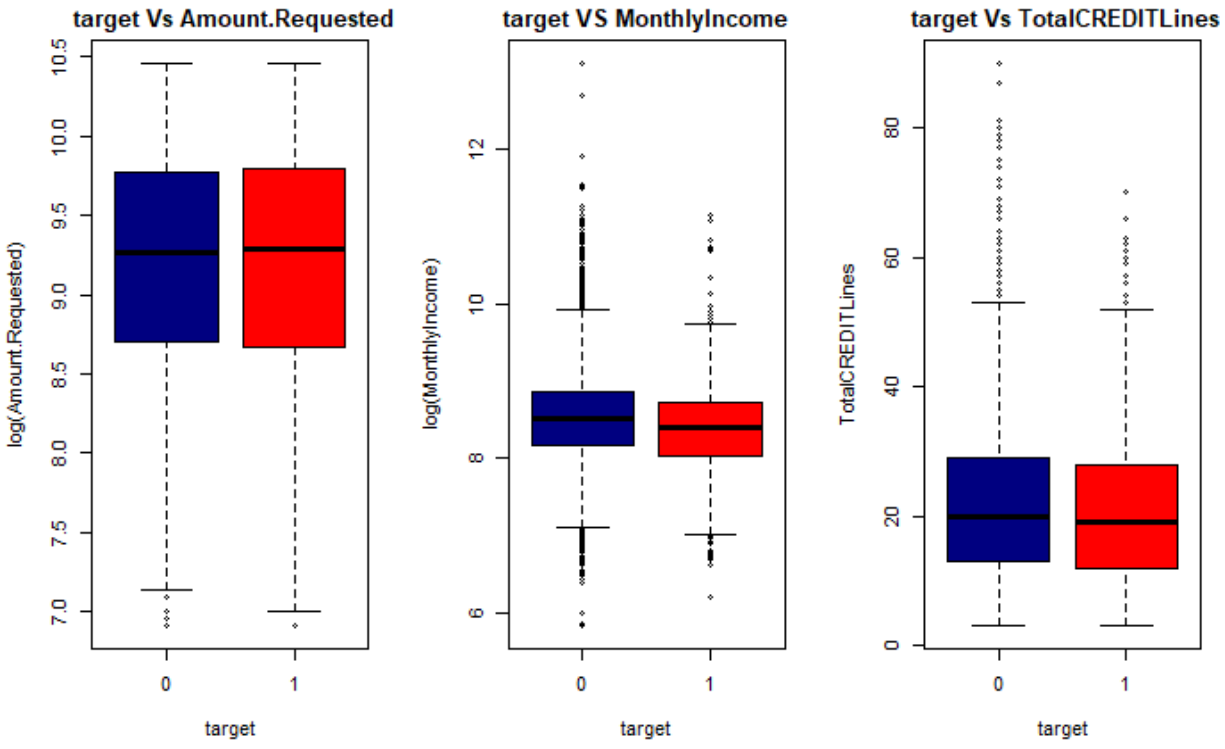


Figure 11b: Boxplot of target Vs the independent variables

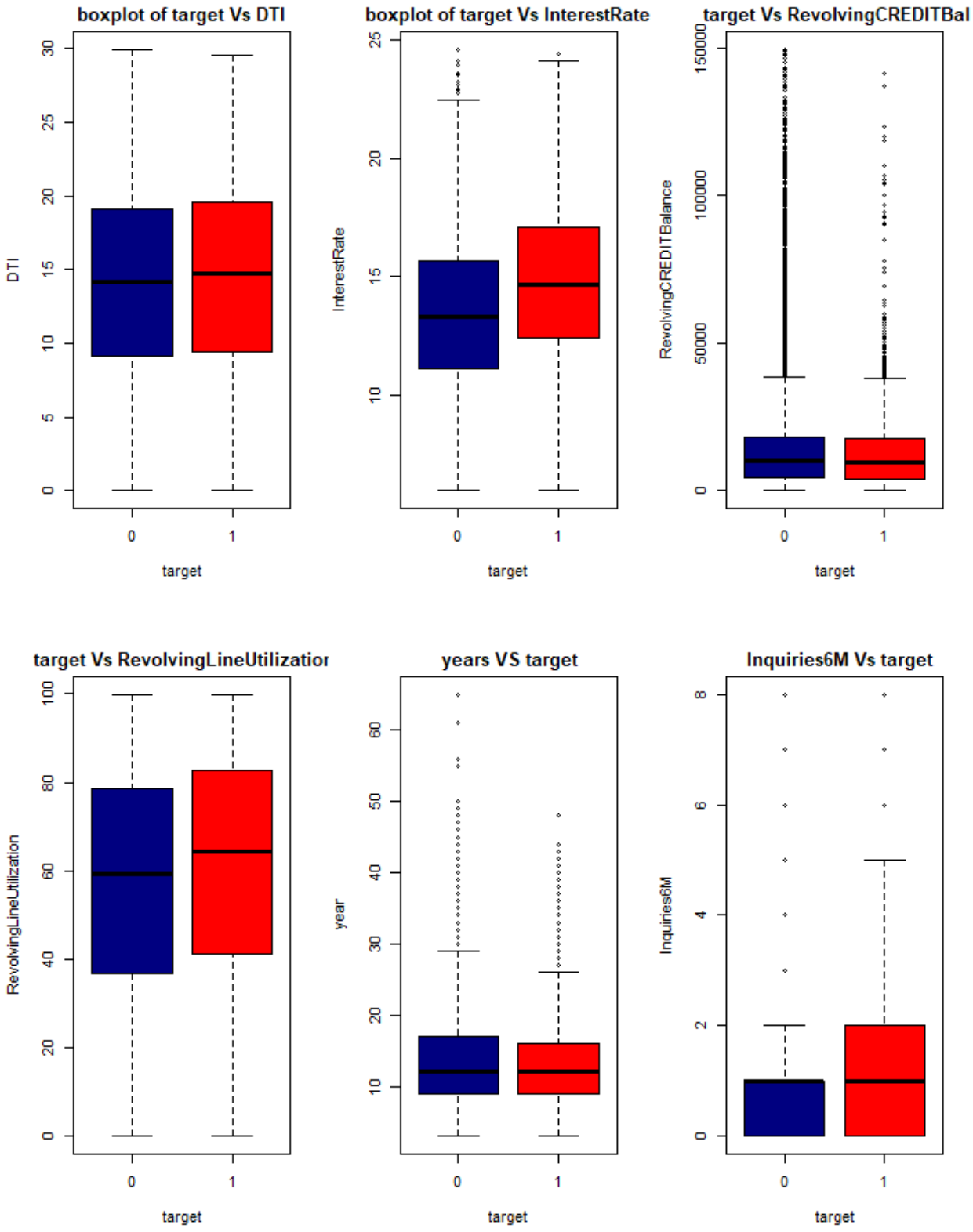
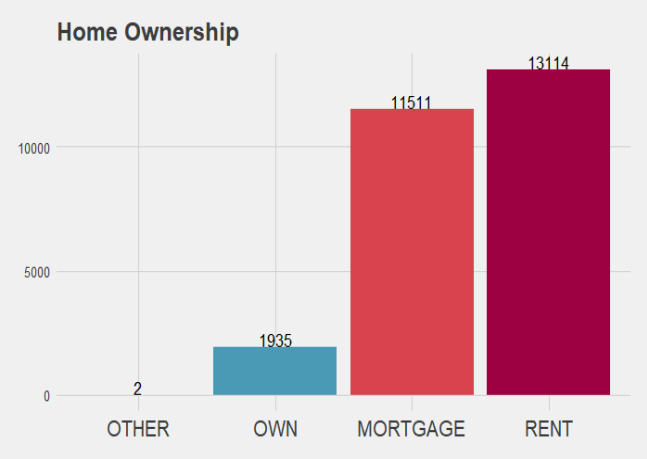
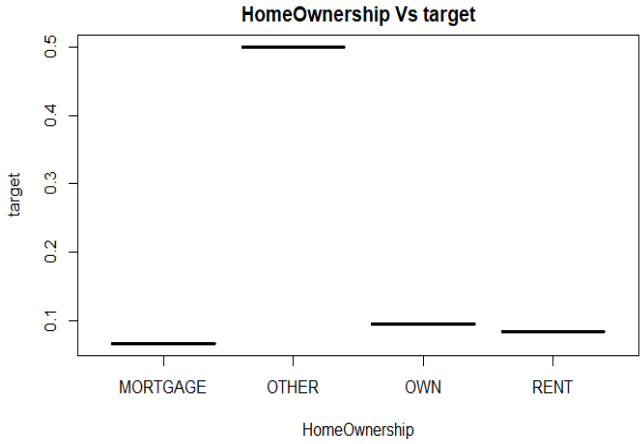
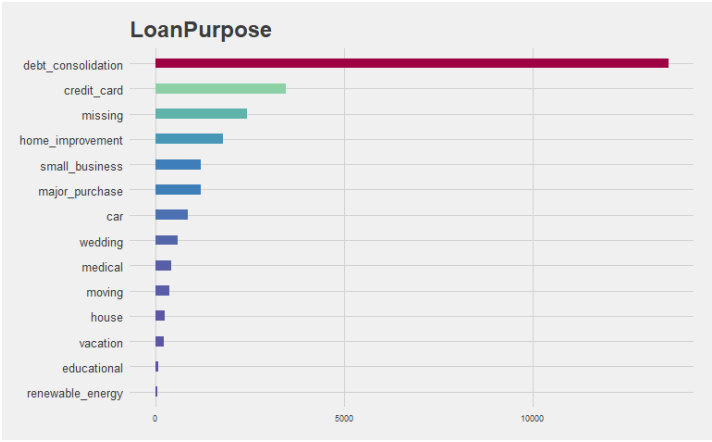
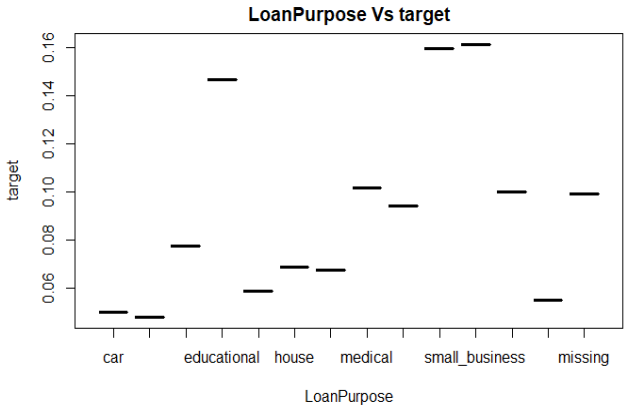


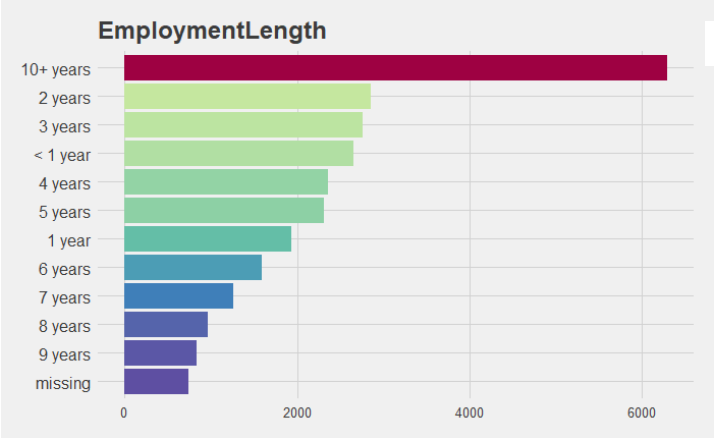
Figure 12: Plot of target Vs independent variables











II. Logistic Regression Model

After observing the plots. I generated a Logistic regression model based on the training data. My goal is to generate a model to predict the dependent variable by using the two predictor variables that I binned using the WOE_Custom command and the other predictors. I log-transformed the monthly income variable before fitting the logistic regression model.

Table 1: Summary of Logistic regression results

Dependent variable: Target

Coefficients	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	0.67023	0.50137	1.337	0.181289
log(MonthlyIncome)	-0.46531	0.05515	-8.436	2e-16***
Inquiries6M_custom[3,5)	0.45722	0.09317	4.907	9.23E-07***
Inquiries6M_custom[5, Inf)	1.07966	0.28548	3.782	0.000156***
RevolvingLineUtilization_custom[30,60)	0.08776	0.09355	0.938	0.348199
RevolvingLineUtilization_custom[60,80)	0.30283	0.0965	3.138	0.001701***
RevolvingLineUtilization_custom[80,Inf)	0.531	0.09465	5.61	2.02E-08***
LoanPurposecredit_card	-0.14413	0.23715	-0.608	0.543344
LoanPurposedebt_consolidation	0.51487	0.21503	2.394	0.016645**
LoanPurposeeducational	1.14762	0.4668	2.458	0.013952**
LoanPurposehome_improvement	0.32264	0.24758	1.303	0.192507
LoanPurposehouse	0.14426	0.42154	0.342	0.732187
LoanPurposemajor_purchase	0.38741	0.25415	1.524	0.127427
LoanPurposemedical	0.81795	0.28844	2.836	0.004572***
LoanPurposemoving	0.53608	0.32017	1.674	0.094058*
LoanPurposerenewable_energy	1.43438	0.44748	3.205	0.001348***
LoanPurposesmall_business	1.34846	0.23422	5.757	8.55E-09***
LoanPurposevacation	0.27233	0.40487	0.673	0.501174
LoanPurposewedding	-0.09981	0.33792	-0.295	0.767719
LoanPurposemissing	0.69331	0.22817	3.039	0.002377***

Notes:***p<0.01; **p<0.05; *p<0.1

In logistic regression, we use the logistic function as shown in equation 1 below.

$$p(X) = \frac{e^{\beta 0+\beta 1X}}{1+e^{\beta 0+\beta 1X}} \quad (1)$$

To fit the model in equation 1, we use a method called maximum likelihood. We use the log odds and equation 1 above is expressed as a linear function of X as shown in equation 2.

$$\log\left(\frac{P(X)}{1+P(X)}\right) = \beta 0 + \beta 1X \quad (2)$$

Thus, the above results can be expressed as

$$\widehat{\textbf{Target}} = In\left(\frac{p}{1-p}\right) = \beta 0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (3)$$

In equation (3), $\beta 0$ is the intercept and corresponds to a value of 0.67023 as shown in Table 1. Also, X_{Is} are the predictors used in the model with β_{Is} as their associated coefficients.

- Table 1 shows the summary results of the Logistic regression model with Target as the dependent variable and the predictors. At the 5% significant level, as monthly income increases, the probability that a customer will default in the loan payment decreases. The sign of the monthly income being negative is valid and makes sense because as customers' income increases, I would expect the probability of default to decrease.
- Also, at the 5% significant level, inquiries made about customers in the last 6 months(Inquiries6M) of 5 and above on loan are more likely to default than customers with Inquiries6M of 0 to 3. Thus, as inquiries made about customers in the last 6 months increase, customers are more likely to default. This positive sign of Inquiries6M also makes sense but not always the case because some inquiries about customers are for different reasons.
- Again, at the 5% significant level, customers with RevolvingLineUtilization of 60 and above on loan are more likely to default than customers with RevolvingLineUtilization of 0 to 30. However, customers with RevolvingLineUtilization of below 60 are not statistically different from customers with RevolvingLineUtilization of 0 to 30 in predicting whether a person is a bad or a good customer. This positive sign also makes sense because as the revolving credit balance to total revolving credit balance increases, I will expect customers to have a higher chance of defaulting.
- Lastly, loans directed purposely for debt consolidation, educational, medical, renewable energy, small business, and other purposes are more likely to default in their loan payment than loans for car purposes. However, loans directed purposely for a credit card, house, major purchase, home improvement**,** moving, and vacation purposes are not statistically different from customers with loan purpose on a car in predicting whether a person is bad or a good customer at the 5% significant level. The signs of loan purposes are a bit subjective, depend on the individual customers, and will be very difficult to judge. I would expect this relationship to be mixed.

III. MARS Model

I generated the MARS model based on the training data. My goal is to generate a model to predict the dependent variable target by using all the predictor variables chosen by the MARS model as the important features.

Table 2: Summary of the MARS model results.

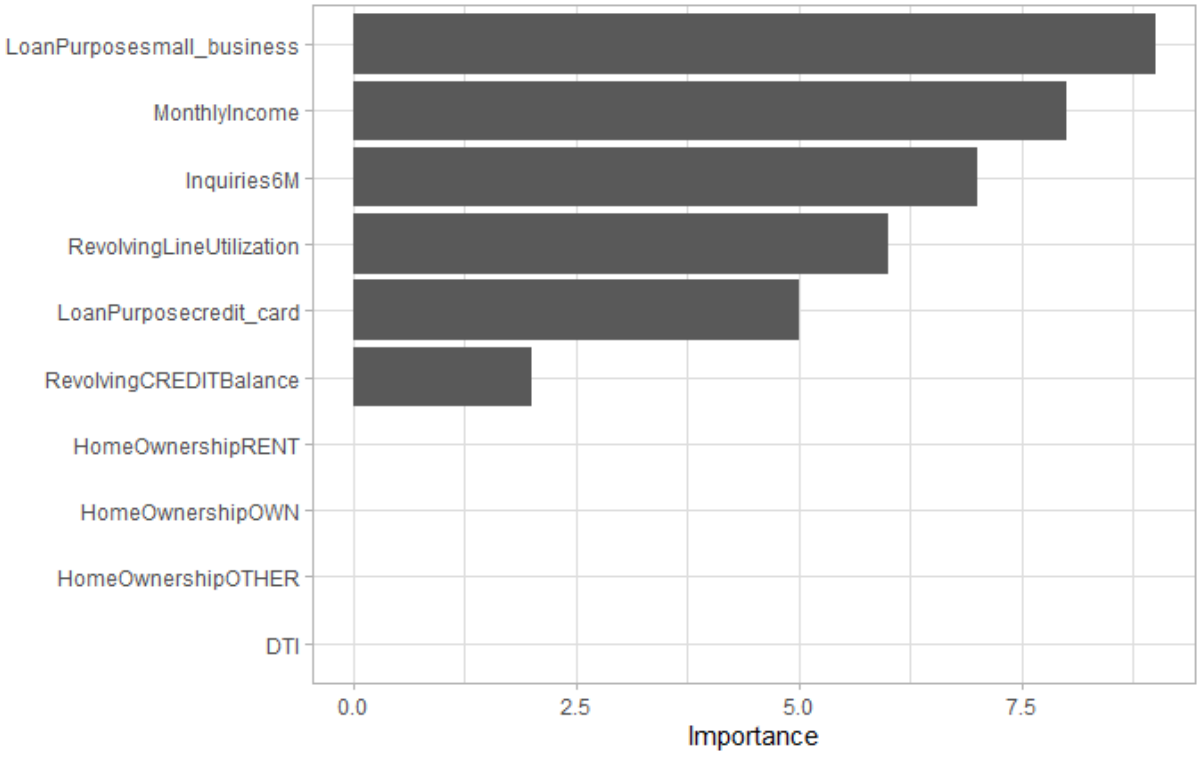
Dependent variable: Target

Coefficients	**Estimate**
(Intercept)	-2.83883
LoanPurposecredit_card	-0.61959
LoanPurposesmall_business	0.858454
h(8750-MonthlyIncome)	0.000146
h(6139-RevolvingCREDITBalance)	4.81E-05
h(RevolvingCREDITBalance-6139)	9.59E-06

h(92.6-RevolvingLineUtilization)	-0.00764
h(RevolvingLineUtilization-92.6)	0.084702
h(1-Inquiries6M)	-0.38224
h(Inquiries6M-1)	0.166671

The first four important variables chosen by the model are monthly income, loan purpose, inquires made in the last 6 months, and RevolvingLineUtilization in ascending order of magnitude as the most significant factors in determining whether a customer will default(bad customer) or not default(good customer)-see **Figure 13** plot below for the rankings of the variables in order of Importance by the MARS model. One drawback of the MARS model is that it does not generate a particular p-value to judge the significance of the variables. Thus, the more sophisticated the model is, the more it loses its interpretation.

Figure 13: Summary of Variables in order of Importance by the MARS model



IV. Evaluating the Models and Conclusion

Table 3: Gains Table for the Logistic regression model

Depth of file	N	Cume N	Mean Resp	Cume Mean Resp	Cume Pct Mean Resp	Lift index	Cume Lift	Mean modal score
10	1062	1062	0.17	0.17	21.60%	216	216	0.15
20	1065	2127	0.09	0.13	34.00%	124	170	0.11
30	1060	3187	0.08	0.12	45.00%	111	150	0.09
40	1063	4250	0.08	0.11	56.10%	110	140	0.08
50	1062	5312	0.07	0.1	65.20%	91	130	0.07
60	1063	6375	0.07	0.1	74.50%	93	124	0.07
70	1062	7437	0.06	0.09	81.80%	74	117	0.06
80	1072	8509	0.06	0.09	89.20%	73	111	0.05
90	1054	9563	0.05	0.08	95.10%	59	106	0.05
100	1062	10625	0.04	0.08	100.00%	49	100	0.03

Figure 14: Gains plot for the Logistic regression model.

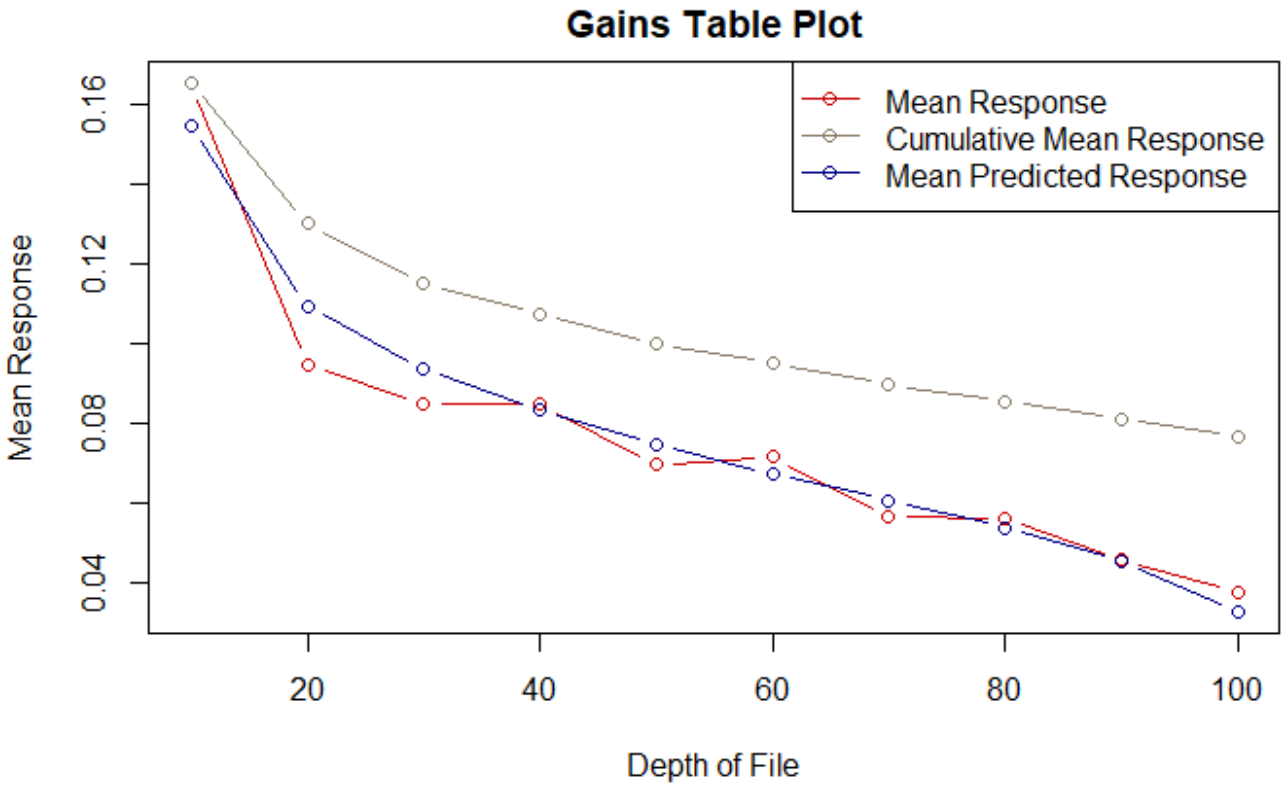


Table 3 and Figure 14 show the gains Table and plot for the Logistic regression model, respectively. From the gains table, it is observed that the Depth of file of 30 which has a cumulative observation of 3187 out of the 10625 has a 45 cumulative percentage of the total response. This means that in the case of predicting the customers with the highest probability of response if I choose 3187 of such customers from the predictions made by the model, there will be 45 percent of good responses. That is, the company should retain customers if the target of the company is not above the Depth of file of 30 which has a cumulative observation of 3187 out of the 10625and has a 45 cumulative percentage of total response with a cumulative lift curve of 150. In that case, at least 55(100-45) percent of the customers will not default their payment of loans (will be good customers), apart from this threshold, I suggest that the company should not retain customers. Lastly, the company must target the right customers to grant credit loans.

Table 4: Gains Table for the MARS model

Depth of file	N	Cume N	Mean Reply	Cume Mean Resp	Cume Pct Mean Resp	Lift index	Cume Lift	Mean modal score
10	1062	1062	0.16	0.16	21.10%	211	211	0.16
20	1063	2125	0.11	0.14	35.80%	147	179	0.11
30	1062	3187	0.08	0.12	46.00%	102	153	0.1
40	1063	4250	0.08	0.11	56.90%	109	142	0.08
50	1062	5312	0.07	0.1	66.50%	96	133	0.07
60	1063	6375	0.06	0.1	74.80%	83	125	0.07
70	1062	7437	0.06	0.09	82.80%	80	118	0.06
80	1063	8500	0.05	0.09	89.00%	61	111	0.05
90	1062	9562	0.05	0.08	95.70%	68	106	0.04
100	1063	10625	0.03	0.08	100.00%	43	100	0.03

Figure 15: Gains plot for the MARS model.

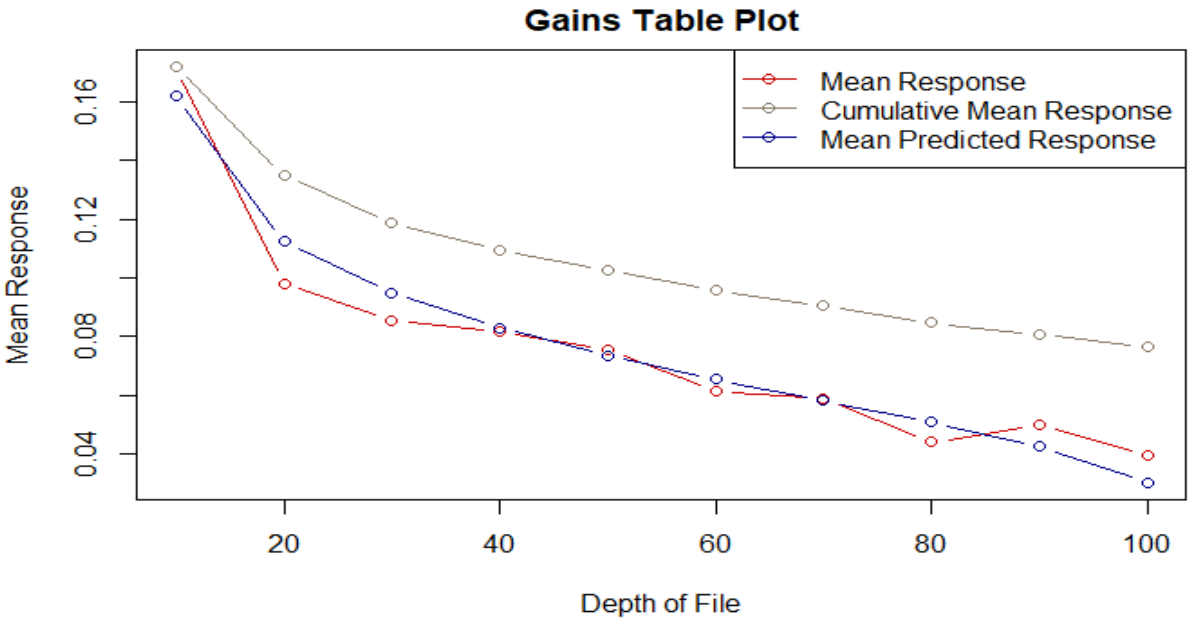


Table 4 and Figure 8 show the gains Table and plot for the MARS model, respectively. From the gains table, it is observed that the Depth of file of 30 which has a cumulative observation of 3187 out of the 10625 has 46 cumulative percentage of the total response. This means that in the case of predicting the customers with the highest probability of response if I choose 3187 of such customers from the predictions made by the model, there will be 46 percent of positive responses. That is, the company should retain customers if the target of the company is not above the Depth of file of 30 which has a cumulative observation of 153 out of the 10625 and has 46 cumulative percentage of the total response. In that case, at least 54 (100-46) percent of the customers will not default in the payment of their loans (will be good customers), apart from this threshold, I suggest that the company should not retain customers. Lastly, the company must target the right customers to grant credit loans.

V. Models Fit and Conclusion

Comparing the Two Models Using the ROC(AUC) and KS
Table 5: Compare logistic and MARS models.

Model	ROC(AUC)	KS value
Logistic regression	**0.623**	**0.17876**
MARS model	0.632	0.19148

Table 5 shows the calculated values of the ROC(AUC) and KS for both the logistic and the MARS model. Table 5 shows that the MARS model is marginally better than logistic regression because the MARS model has a slightly higher AUC value as compared to the logistic regression model.

Again, to compare models, I used the validation dataset for each model type since the validation dataset has a higher prediction accuracy. The ROC(AUC) value for the logistic regression **(MARS model)** is approximately 62.3**(63.2)** percent, respectively. This means that the MARS fits the data better than the logistic regression model because it has a higher AUC value. Therefore, the MARS model is a better 'fit with an AUC of 63.2% in predicting whether a customer will default or not default to pay his or her loans on time given the predictor variables in the data set.

Figure 15: Cumulative Gains and life chart plot for the logistic and MARS models

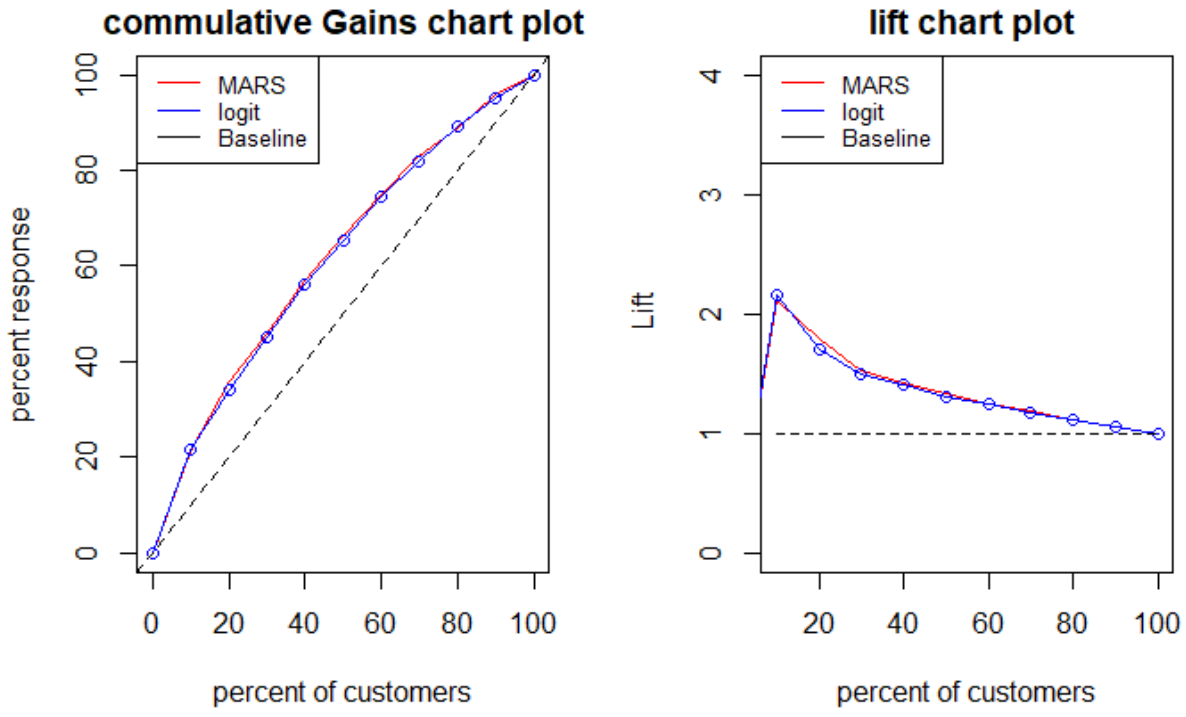


Figure 16: ROC plot for the logistic and MARS models on the same diagram

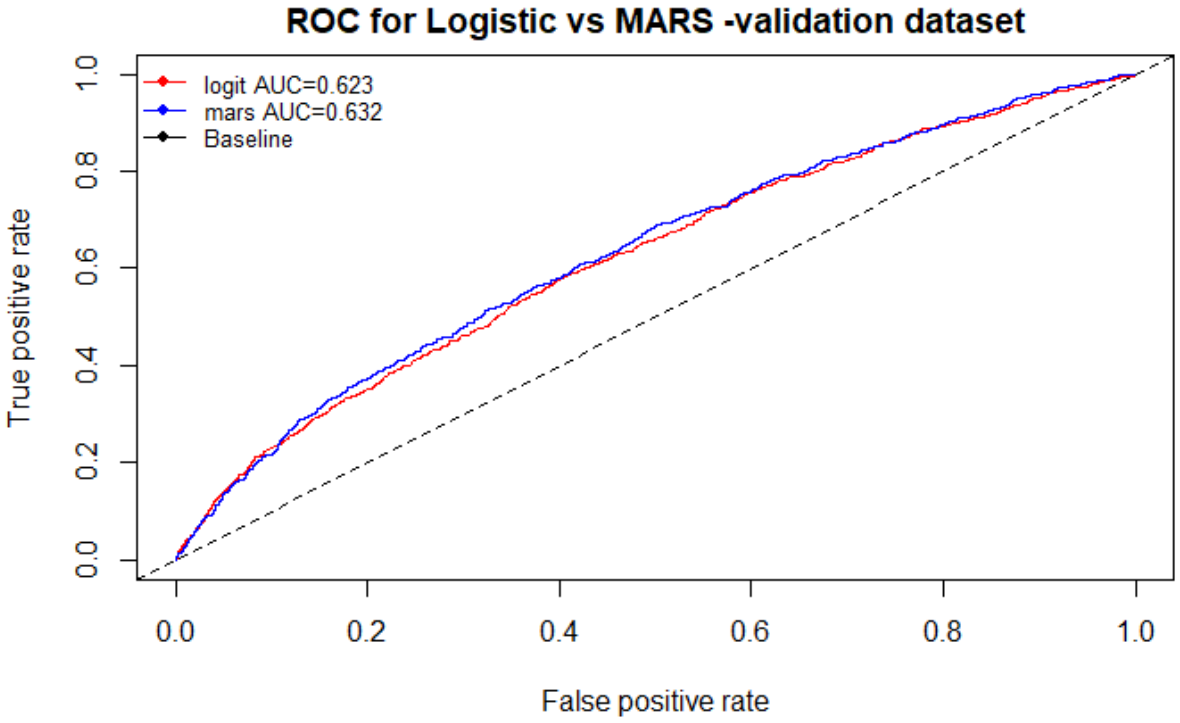
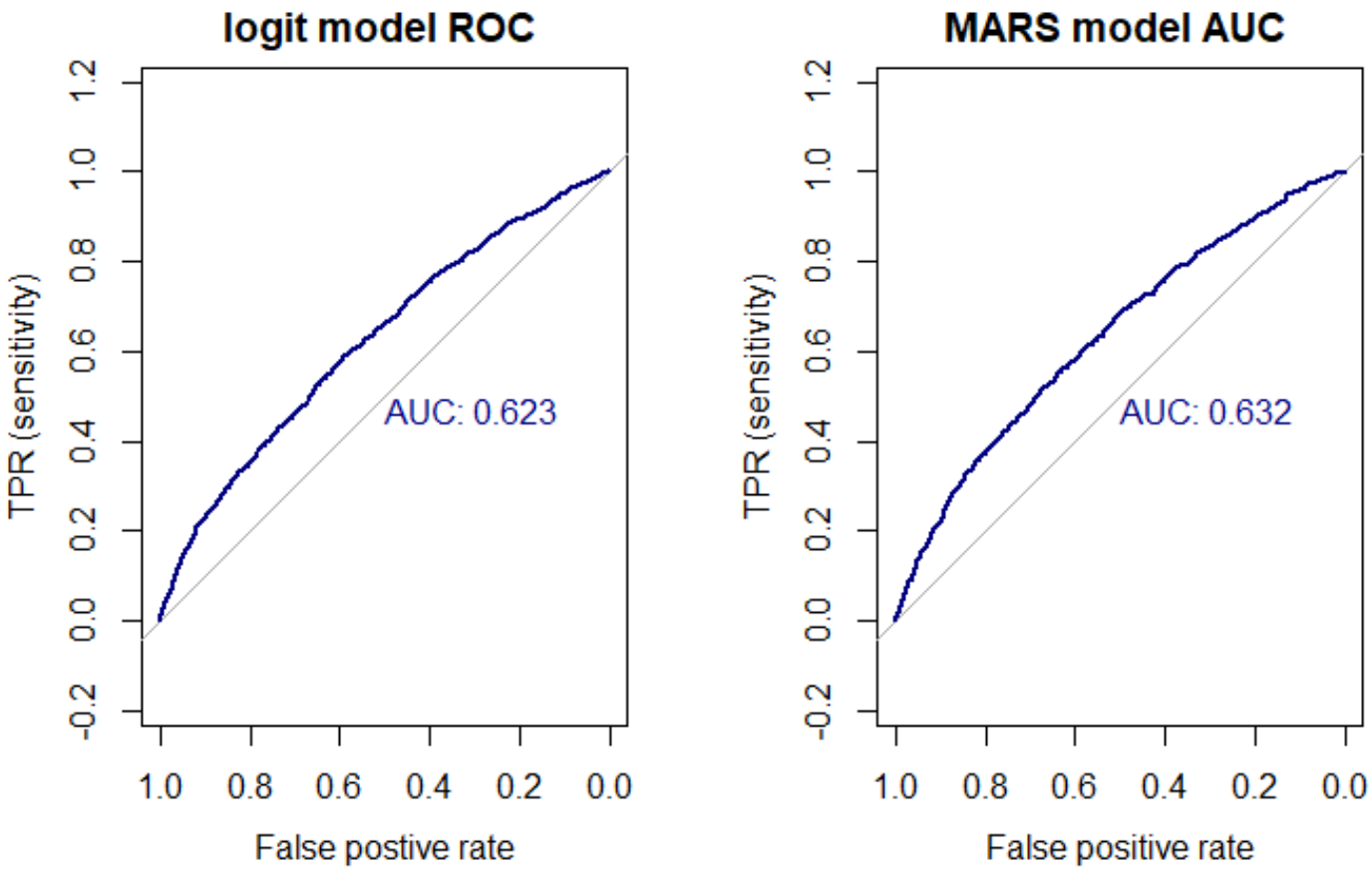
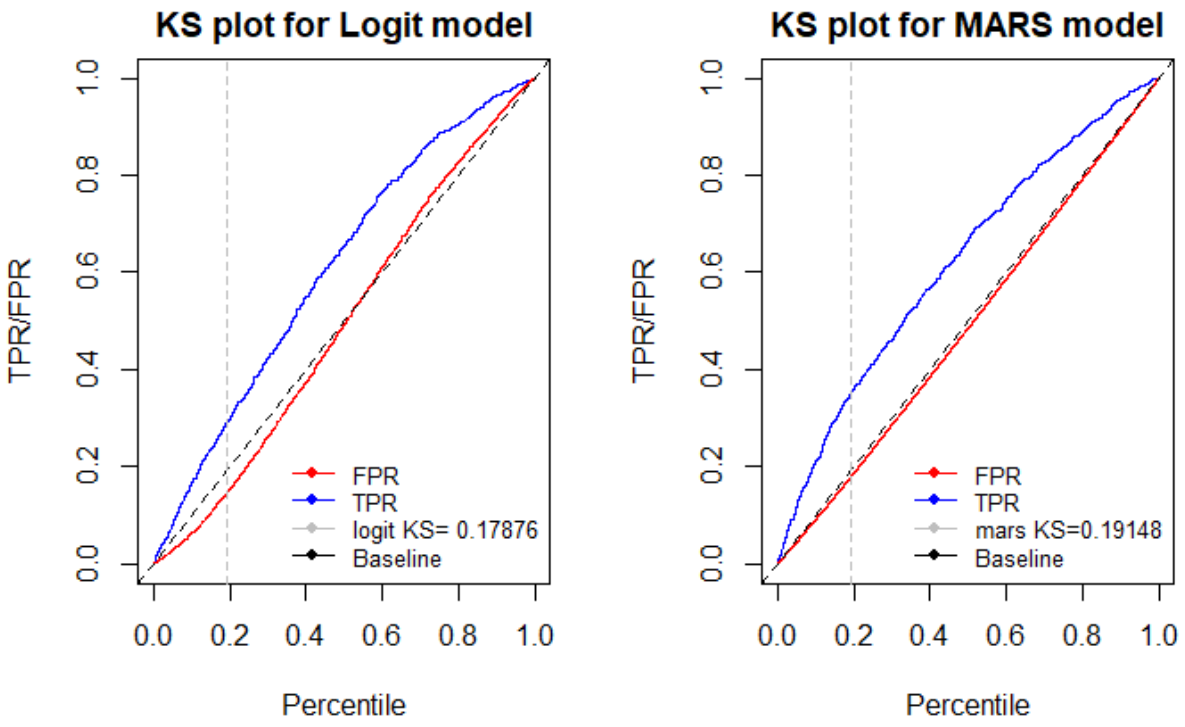


Figure 17: ROC plot for the logistic and MARS models



Figures 15, 16, and 17 plots confirm the results in Table 5 that using the ROC, the MARS regression performs better than the logistic model. This is because the True Positive rates are slightly higher for the MARS model than the logistic regression.- see Figure 16.

Figure 18: KS plot for the logistic and the MARS model



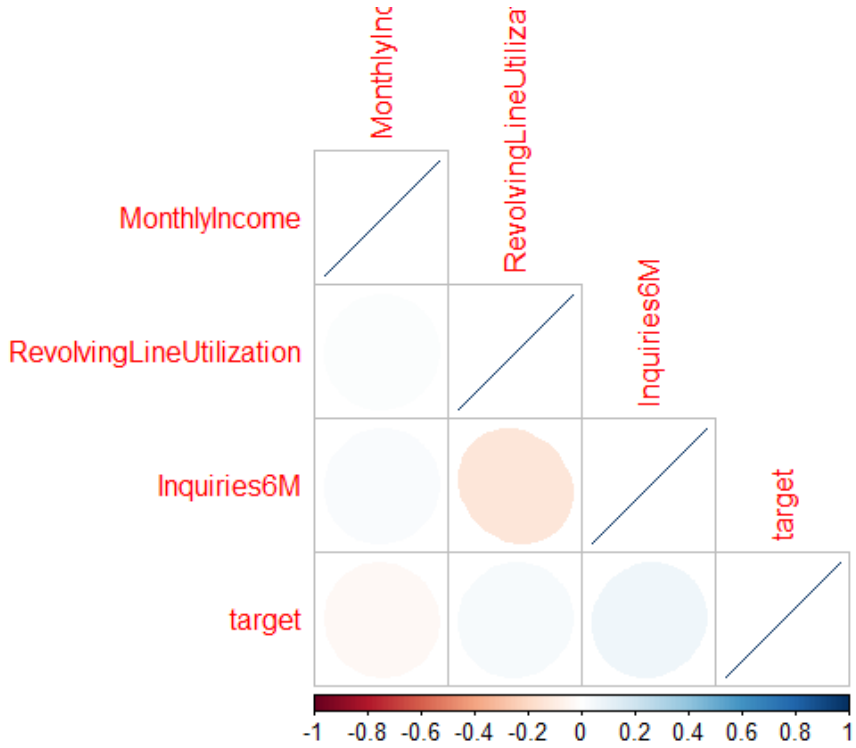
The Figures above also confirm the results in Table 5 that using the KS, the MARS model performs better than the logistic regression. This is because the True Positive rates are marginally higher than the False Positive rates for the MARS model as compared with the logistic regression -see Figure 19.

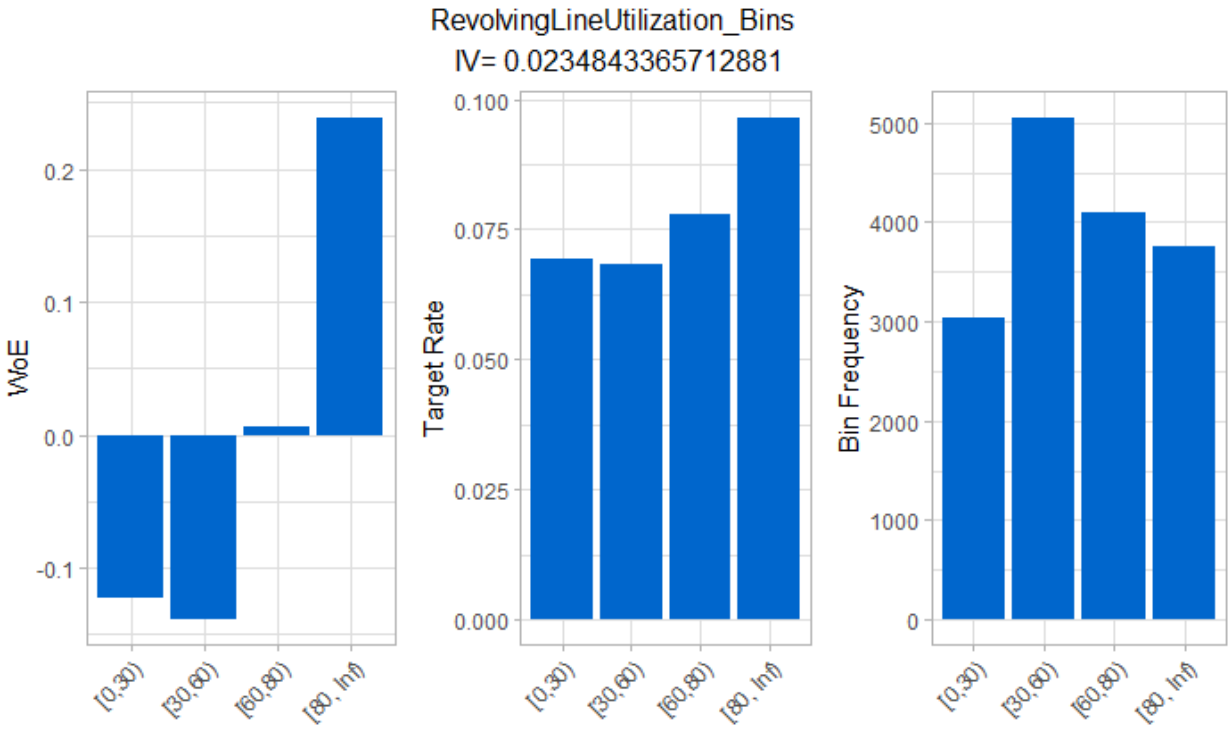
Conclusion

The analysis sheds light that using the ROC(AUC), and KS for both the logistic and the MARS model. The MARS model performs better than the logistic regression in predicting whether a customer will be default or not default in the payment of loans. Therefore, the MARS model is a better 'fit with a KS of 19.1%, and an AUC of 63.2% in predicting whether a customer will default his or her payment of loans given the predictor variables in the data set. This is also evidence in the ROC and KS plots and the plot of the logistic regression is a little bit above the MARS model. The company should therefore check the interest rate, Monthly income, purpose of loan, and inquiries in the last six months in ascending order of magnitude if possible before granting loans to customers.

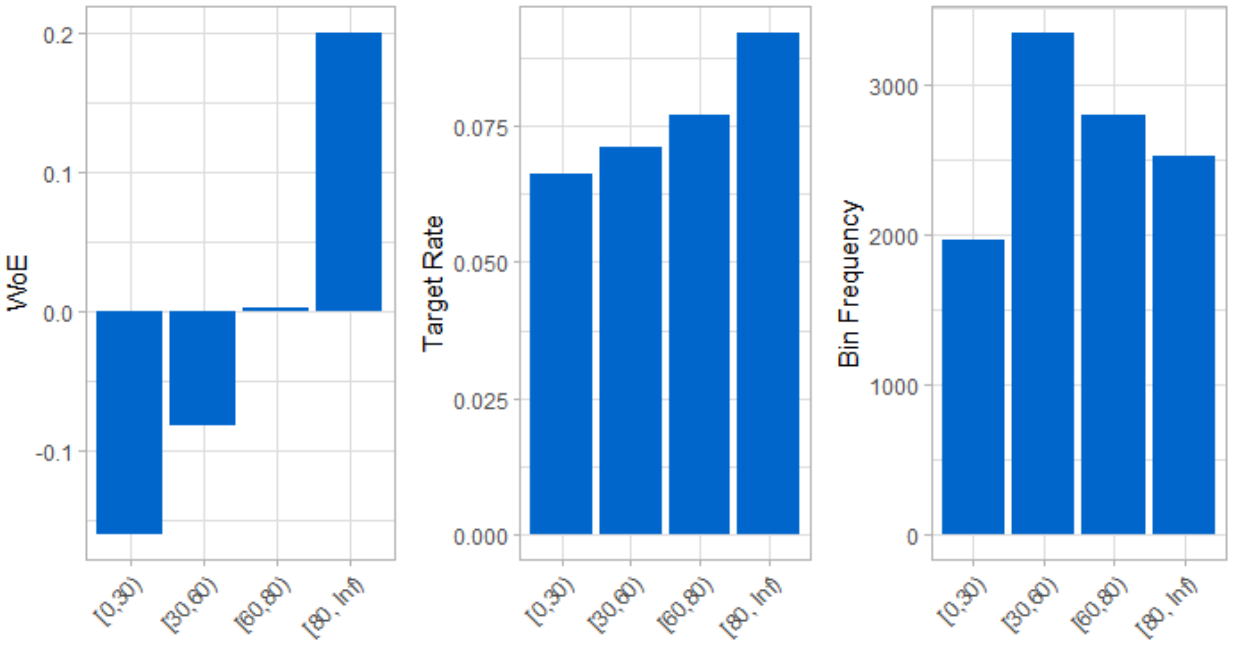
VI. Appendix and Reference

Figure 20: correlation plot of the continuous predictors in the logit model

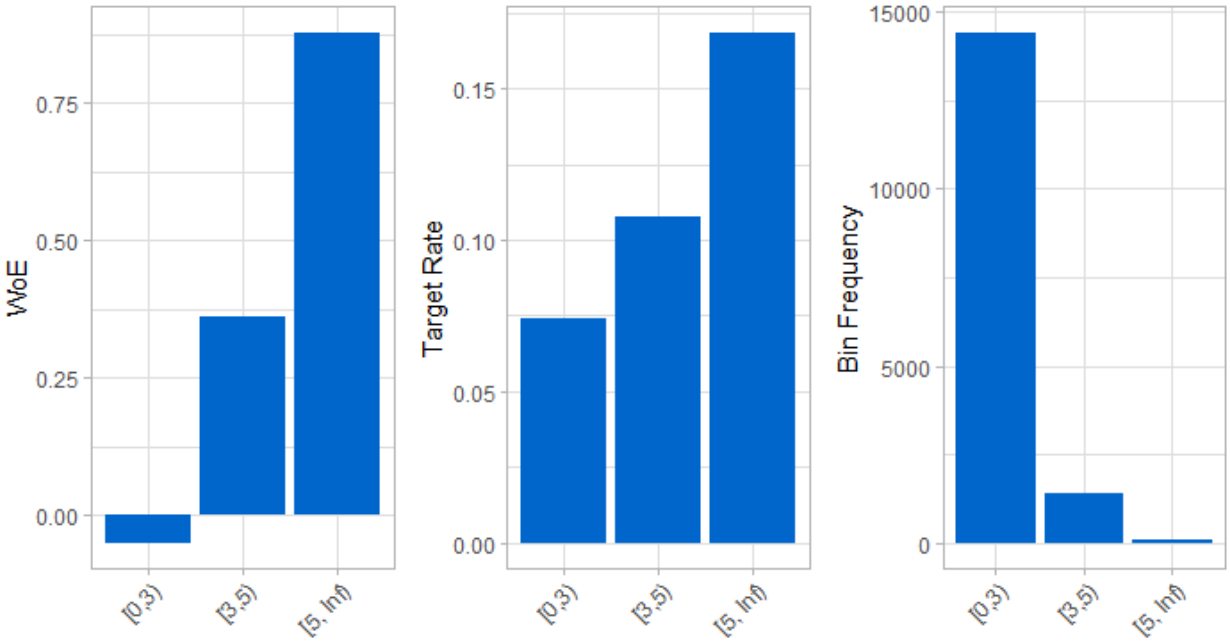


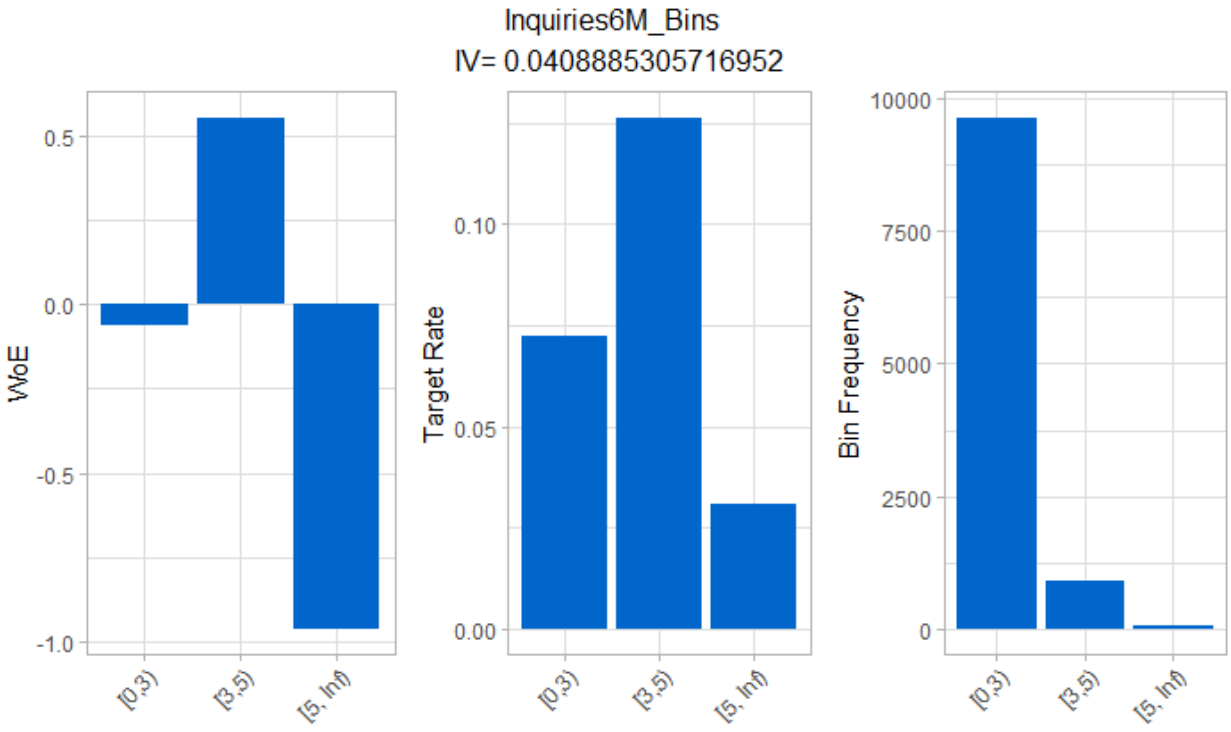


RevolvingLineUtilization_Bins
IV= 0.0167600212310547



Inquiries6M_Bins
IV= 0.0224144384360572





Data Dictionary

Variable	Usage	Definition
LOAN_ID	ID	Unique ID
MOB	Input	Months on Books--how long the customer has been a customer
InterestRate	Info Only	Interest Rate of the Loan
IssuedDate	Input	Date Loan was issued
DTI	Input	Debt to Income Ratio---monthly debt payments divided by monthly income
State	Input	State of residence
HomeOwnership	Input	Home Ownership status
MonthlyIncome	Input	Monthly income
EarliestCREDITLine	Input	Date of first known credit line

OpenCREDITLines	Input	Number of Open Credit Lines
TotalCREDITLines	Input	Total number of credit lines ever
RevolvingCREDITBalance	Input	Dollar amount of revolving credit (e.g. credit card balance total)
RevolvingLineUtilization	Input	Revolving Credit Balance divided by Total Revolving Credit Limit
Inquiries6M	Input	Number of inquiries customer has made for new credit in the last 6 months
AccountsDQ	Input	Number of credit accounts that are currently delinquent/past due
DelinquentAmount	Input	Dollars of credit accounts that are currently delinquent/past due
DQ2yrs	Input	Delinquencies in the last 2 years
MonthsSinceDQ	Input	Months since last delinquency
PublicRec	Input	Number of Public Records (bankruptcy, liens, judgements, etc….bad public records)
MonthsSinceLastRec	Input	Months since last public record
Education	Input	Education
EmploymentLength	Input	Length of time employed at current job
currentpolicy	Input	Does this customer meet current credit policy?
term	Info Only	Number of months of loan payments for this loan
appl_fico_band	Input	FICO Score Band (FICO is the credit score)
vintage	Info Only	Calendar Quarter the loan was issued in
TERM	Info Only	Same as term
MFMonth	Info Only	Master File Month
target	target	Bad customer default flag
Amount.Requested	Info Only	Amount of Loan Requested
LoanPurpose	Input	Stated purpose of loan
Year*	New input	Difference between Earliest date and date when loan was issued.
FICO*	New input	Good(appl_fico_band<=729);very good(729<appl_fico_band<=799; excellent(appl_fico_band>=800)
Education*	New input	Yes=complete cases, NO=missing records
Notes*	Notes*	I did not use all INFO variables when building the two models

Notes**	Notes**	Deleted all observations with zero variance and loan ID
Notes***	Notes***	Categorical variable NAs are replaced with missing Continuous variable NAS are replaced with Zeros, some cases with mean and binning where necessary.
Notes**	Notes**	
logMonthlyoncome	New input	I log transformed monthly income when fitting the logistic model.
Notes**	Notes**	I binned RevolvingLineUtilization and Inquires6M when fitting the logit model.
Employment years**	Notes**	I left it as factor, but feature selection methods did not see its importance. I will convert it to numeric to see its relevant for future research.
State**	Notes**	I deleted it but I will group them by regions to know its relevant for future research. Another thing to look at is to categorize it as Yes for states with more than 1200 observations and No otherwise.

References

- "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)", by Hastie, Tibshirani, and Friedman.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p.
- https://stackoverflow.com/questions/8161836/how-do-i-replace-na-values-with-zeros-in-an-r-dataframe
- https://stackoverflow.com/questions/11036989/replace-all-0-values-to-na
- https://stackoverflow.com/questions/36568070/extract-year-from-date/53340717
- https://www.experian.com/blogs/ask-experian/credit-education/score-basics/what-is-a-good-credit-score/