CREDIT RISK MODELLING

Contents

Table of Figures

CREDIT RISK MODELLING

Ankur Wadhwa
wadhwa5@purdue.edu

Jeet Krishnan
krishnaj@purdue.edu

Swapnil Rai
rai16@purdue.edu

1.Executive Summary

The concept of lending money to the right person at the right time in exchange of an accrued interest component has been in practice from ages. However, identifying the right person for this transaction has been a challenge. Organizations in this sector must account for the risks of a customer defaulting as well as the opportunity cost of losing a good customer due to excessively strict vetting policies. In this project, we aim to build a model that helps organizations in the financial sector make this delicate decision.

For this novel purpose, we have taken the data provided by a peer to peer non-banking financial company (NBFC) named Lending Club. The dataset in its rawest form had over a million observations with over a hundred predicting variables. After thorough research on the subject matter, company specific details, and statistical parameters, we narrowed our input predictor space to twenty-seven relevant predictors. Our findings from the existing research in this area and expert opinions from industry experts helped us identify important variables that were missing in the original dataset. A key aspect that we have considered in our project is the inclusion of macroeconomic indicators. We collected important macroeconomic variables from government sources and mapped them with our existing dataset.

The problem in hand was a typical case of binary classification of an unbalanced data. We call the data as unbalanced because one class (Fully Paid) of the predictor variable significantly outweighs the other class (Default) in terms of number of observations. We identified many technical inaccuracies stemming out of building predictive models using this unbalanced dataset. Hence, we decided to make the training dataset balanced by using an up-sampling technique. On this dataset, we have tried to build the major models used for classification problems as identified by the existing research. We fitted a total of seven models namely Classification trees, Bagged trees, Random Forest, Logistic regression, Bayesian additive regression trees, Linear discriminant analysis and Quadratic discriminant analysis. The performance of these models was analyzed based on a combination of metrics and not a single metric (accuracy) as done in the existing research papers. This multi metric analysis helped us select the model generated by Random forest algorithm as the optimum model.

On analyzing the key predictors identified by our models, we found that the state in which loan was issued was an important predictor. On digging further into this, we noticed that Republican party was governing the top five defaulting states. The macroeconomic factors we added made it to the list of top predictors indicating an underlying relationship with the demographics of the location with interest rate. Our dataset highlights housing loan as being one of the most defaulting loan categories. This corroborates the conclusion of the analysis down in the aftermath of the great recession. The return on investment for the company on loans issued in these years was negative indicating a loss. The company bounced back since then and is now relatively stable. Overall the prediction rate of our model is good although there is room for improvement. We plan to run more advanced classification models and do better parameter tuning and feature engineering in the near future.

2.Introduction

Credit and default risks have been an area of concern for banking and non-banking financial companies (NBFC) ever since the introduction of the concept of loaning money in exchange for a market driven time value of money in form of interest. This problem has come to the forefront since the subprime mortgage crisis that culminated to into a global recession in 2008. In the aftermath of the crisis, people realized that one of the main causes of that crisis was that financial institutions granted loans freely to people without properly examining customers risk profile. An inferior risk modelling technique meant that financial institutes got their risk appetite calculations completely wrong leading to losses or in some cases bankruptcy. To prevent this from happening again, financial institutions turned in to the concept of Credit risk modelling to assess the credit risk of individuals and corporation more accurately.

Our research aims to study the existing work in this field and find ways to improve the modelling algorithm so that businesses can make better decisions. Our goal is to estimate the probability that a potential customer applying for a loan will experience delinquency beyond 30 days or default on his/her financial commitments to the company. Based on this probability, we plan to classify customers in different risk buckets so that a company is aware of the risk of acquiring a customer and can relate the overall risk potential to its risk appetite. While our research uses analysis from publicly available data of a company named Lending Club, we believe that the designed model can be used in other sectors and companies where some form of financial risk grading is involved.

3.Literature review

The techniques utilized in building credit scoring models rely mostly on classification methods[1]. Conventionally, the most widely applied method for credit scoring is logistic regression[2] followed by other linear methods, such as Linear discriminant analysis. However, based on our research we believe that this preference is not without a good reason since linear models provide in practice a very good compromise between classification accuracy and simplicity and interpretability[3]. Especially financial institutions are more reluctant to adopt less intuitive, "black box" approaches[4] since their legislative and operational framework imposes constraints on data availability, transparency, verifiability and interpretability of their risk evaluation methods and processes. We have used a Hybrid approach which is based on combining two (or more) different machine learning techniques, on a single predictor [5] i.e probability of default in our case. Traditional credit scoring models have focused mainly on individual level data or factors pertaining only to the customer/applicant. However, recent researches show the existence of a relation between loss given default rate (LGDR) and macroeconomic conditions [6]. Hence, we have added macroeconomic factors like Median Income and Poverty index at county levels to our dataset.

Reviewing the literature reveals that error rates were often used as the measurement of classification accuracy of models [7]. However, most records in the data set of credit card customers are non-risky (75-80%); therefore, the error rate is insensitive to classification accuracy of models. For the binary classification problem, area ratio in the lift chart can offer better solution for comparing the performance of different models than the one did by the error rate. Therefore, our study explores area ratio, instead of the error rate, to examine the classification accuracy among various data mining techniques.

4.Data Overview

4.1 Source

Our main data source is the dataset floated by a peer to peer lending company called as LendingClub (Dataset). The base dataset contained 1,321,864 observations with 111 variables. On this dataset, we have added variables like Median Income and Poverty which would give us an indication of the effects of macroeconomic factors on the default rate.

4.2 Data cleaning

- While doing exploratory data analysis, we noticed that one of the key variables had the following distribution of factors:

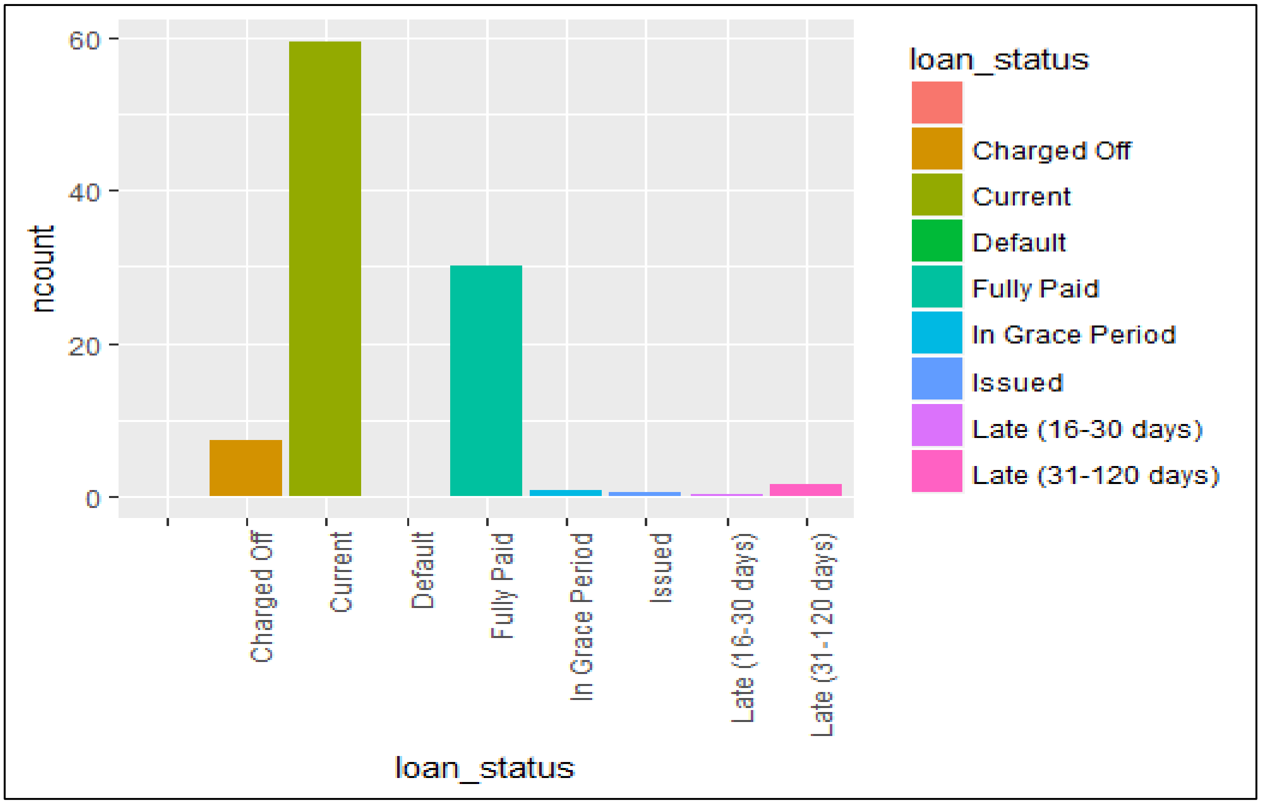


Figure 1: Distribution of response variable

Since the aim of our research is to predict the default probabilities and categorize the observations, we realized that loan status such as Current, In Grace Period, Issued, late (16-30 days) did not help us in achieving our objective. We needed only those observations which have a decisive outcome. The factors mentioned above (marked with grey color in the graph) did not result in a decisive outcome and would only obscure our training dataset. Hence, we decided to eliminate all observations which had the above-mentioned loan status. The other factors were merged based on the findings of our research in the finance market and lending club parameters. This leaves us with two main categories: Fully Paid, Default.

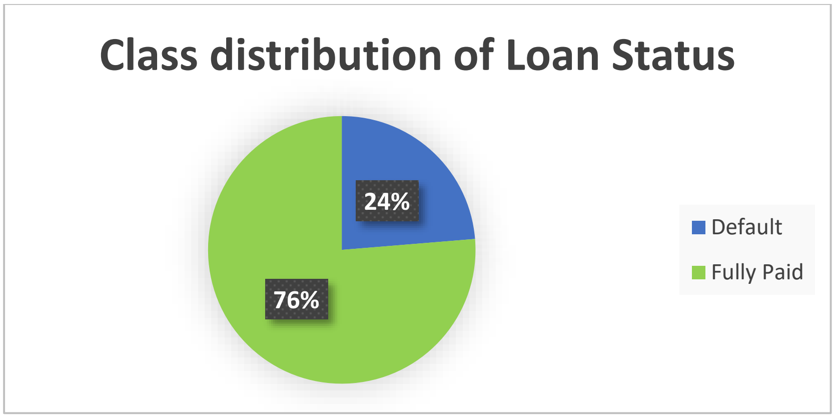


Figure 2: Distribution of response variable after initial cleaning

- We removed all the observations for which the predictor variable was missing
- The factor levels for many variables were not consistent and had bad data in them like '?' ,'*', '%' etc. We had to replace these values to 'NA' so that they did not create any complications during our analysis
- Majority of the variables had incorrectly coded class. Many numerical variables were stored as factors, many factors were stored as characters etc. We cross checked the data and made necessary corrections
- On analysis of the missing values in our dataset, we noticed that certain variables had more than 75% of the data missing. We decided to drop these columns given such high level of missing values

We also went through each and every predictor and researched on the company's website and internet to make a call as to whether a particular predictor should be included in our analysis or not. Few variables eliminated as a result of this scrutiny were 'total_pymnt'(total payment received till date) & 'out_principal'(outstanding principal amount). These variables provide metrics that are collected during an individual's journey through the loan. Since we need to make this prediction before acquisition of a client, these variables are irrelevant or misleading for our analysis.
We also carried out variable selection while building the model based on the significance of parameters and variable importance plot to improve predictive accuracy on the test set. This involves removing linearly dependent variables in linear models and low importance variables in tree based models and bagged models

4.3 Data merging

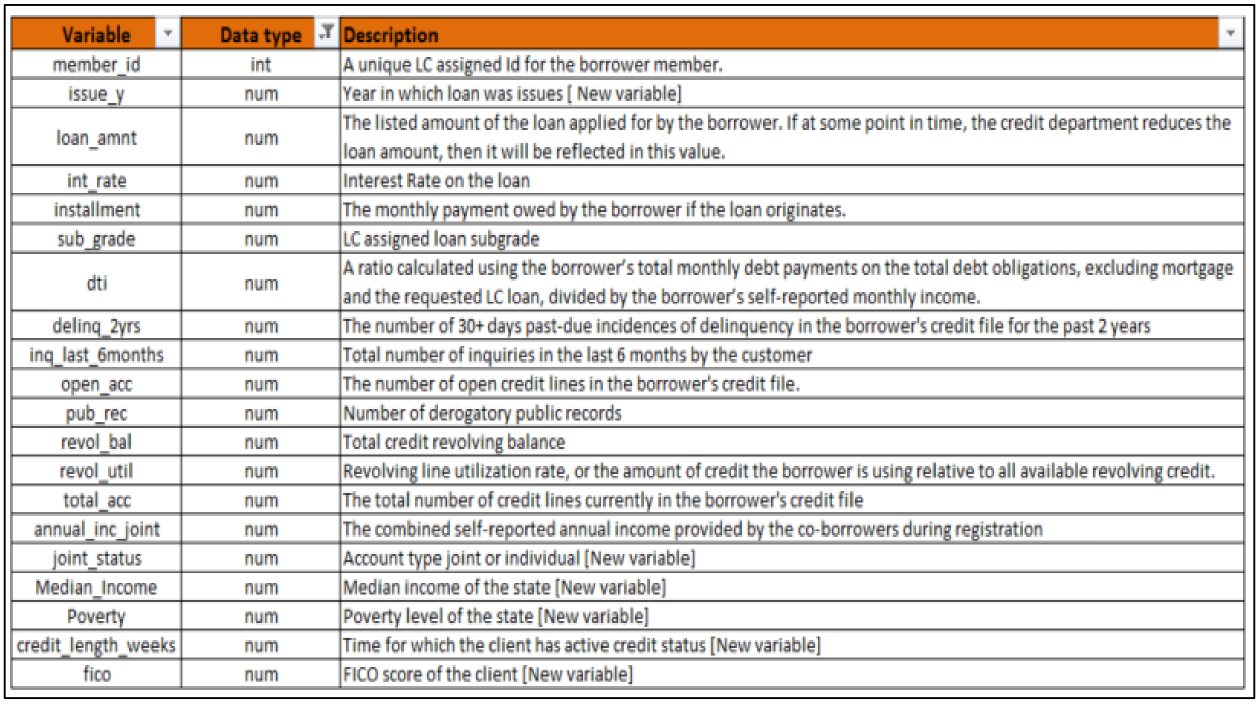
We collected the data for Median Income and Poverty for different years and counties from the website of census.gov Once we obtained this data, we performed basic sanity checks and modification so that it would be easier for us to merge with the original dataset.

We also collected FICO score variable from the dataset obtained from data world and mapped the FICO scores into our dataset by merging with the member_id field which is a unique identifier
Note: We excluded all the observations for which FICO score was not present

4.4 Data description

Response variable: ‘loan_status’: - This is a factor variable with two levels 0 and 1 corresponding to ‘Fully Paid’ and ‘Default’ respectively

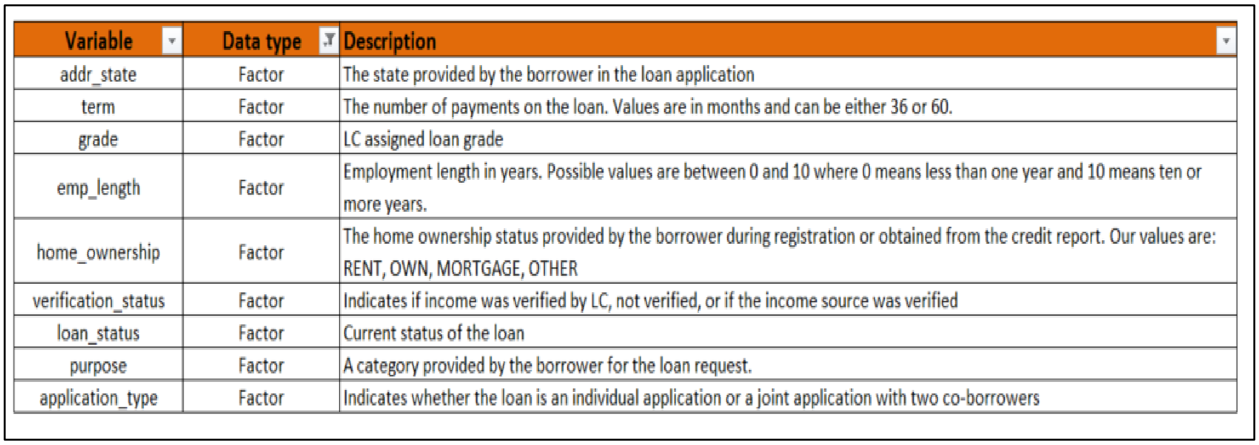
Numerical variables:



Variable	Data type	Description
member_id	int	A unique LC assigned Id for the borrower member.
issue_y	num	Year in which loan was issues [New variable]
loan_amnt	num	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
int_rate	num	Interest Rate on the loan
installment	num	The monthly payment owed by the borrower if the loan originates.
sub_grade	num	LC assigned loan subgrade
dti	num	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
delinq_2yrs	num	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
inq_last_6months	num	Total number of inquiries in the last 6 months by the customer
open_acc	num	The number of open credit lines in the borrower's credit file.
pub_rec	num	Number of derogatory public records
revol_bal	num	Total credit revolving balance
revol_util	num	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_acc	num	The total number of credit lines currently in the borrower's credit file
annual_inc_joint	num	The combined self-reported annual income provided by the co-borrowers during registration
joint_status	num	Account type joint or individual [New variable]
Median_Income	num	Median income of the state [New variable]
Poverty	num	Poverty level of the state [New variable]
credit_length_weeks	num	Time for which the client has active credit status [New variable]
fico	num	FICO score of the client [New variable]

Figure 3: Description of numerical variables

Factor variables:



Variable	Data type	Description
addr_state	Factor	The state provided by the borrower in the loan application
term	Factor	The number of payments on the loan. Values are in months and can be either 36 or 60.
grade	Factor	LC assigned loan grade
emp_length	Factor	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
home_ownership	Factor	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
verification_status	Factor	Indicates if income was verified by LC, not verified, or if the income source was verified
loan_status	Factor	Current status of the loan
purpose	Factor	A category provided by the borrower for the loan request.
application_type	Factor	Indicates whether the loan is an individual application or a joint application with two co-borrowers

Figure 4: Description of factor variables

4.5 Sampling

Imbalanced classification is a supervised learning problem where one class outnumbers other class by a large proportion. This problem is faced more frequently in binary classification problems than multi-level classification problems. In other words, a data set that exhibits an unequal distribution between its classes is considered to be imbalanced (Analytics Vidhya). Many research papers [8] have highlighted the problems of using machine learning algorithms to train unbalanced datasets. The main challenge in imbalanced datasets are that the small classes are often more useful, but standard classifiers tend to be weighed down by the huge classes and ignore the smaller ones. For a problem like ours, where default prediction is significant, we cannot afford to be losing out on accuracy of predicting defaults. We first tried to develop models using the original unbalanced data but as expected and corroborated by the papers, our accuracy was poor and heavily weighted towards predicting "Fully Paid" since it outnumbered "Defaults" 3:1.

We used different sampling methods like ROSE (by Nicola Lunardon), SMOTE (by Nitesh V. Chawla), up sampling and down sampling techniques. After running models on dataset generated by thee techniques and doing some sanity checks on our own, we decided to go ahead with the up-sampled dataset using SMOTE.

The up-sampling method works with class having fewer observations. It selects samples from the smaller class with replacement such that the total number of data points of the smaller class in the up-sampled dataset is equal to the number of data points in the majority class. An advantage of using this method is that it leads to no information loss.

4.6 Final dataset

After all the cleaning and up-sampling efforts, we were left with a final dataset having the following distribution

	Fully Paid	**Default**	**Total**
Up-train	28113	28113	56226
Test	7051	1216	8267
Total	**35164**	**29329**	**64493**

5.Exploratory Data Analysis

We performed various exploratory data analysis to understand our dataset better and identify the relationships that exist between predictors and between predictors and response variable. The anaysis and insights drawn from them are mentioned below.

- The distribution of the response variable in our dataset is shown in the above graph. As we can see the dataset is heavily unbalanced (as is the case with most real-life classification datasets)

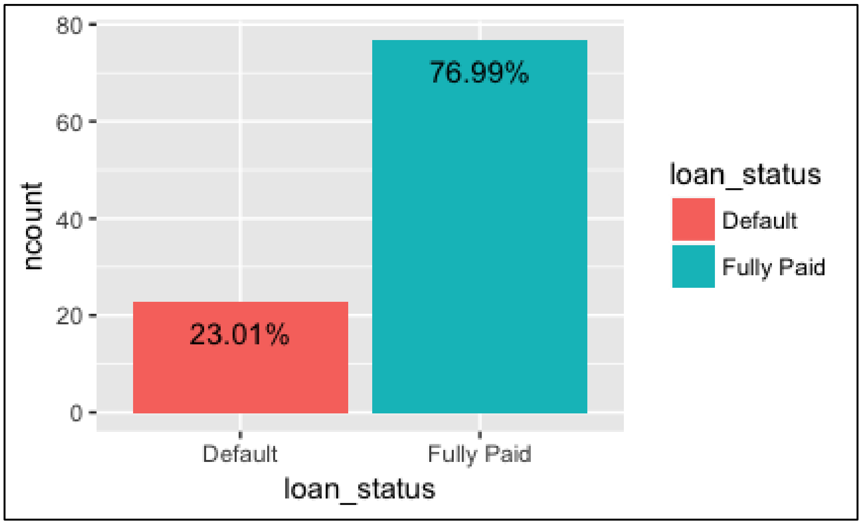


Figure 5: Final distribution of response variable

- This map shows the distribution of loans across the US. We can see that most of the loans have been issued in the bay area and the east coast

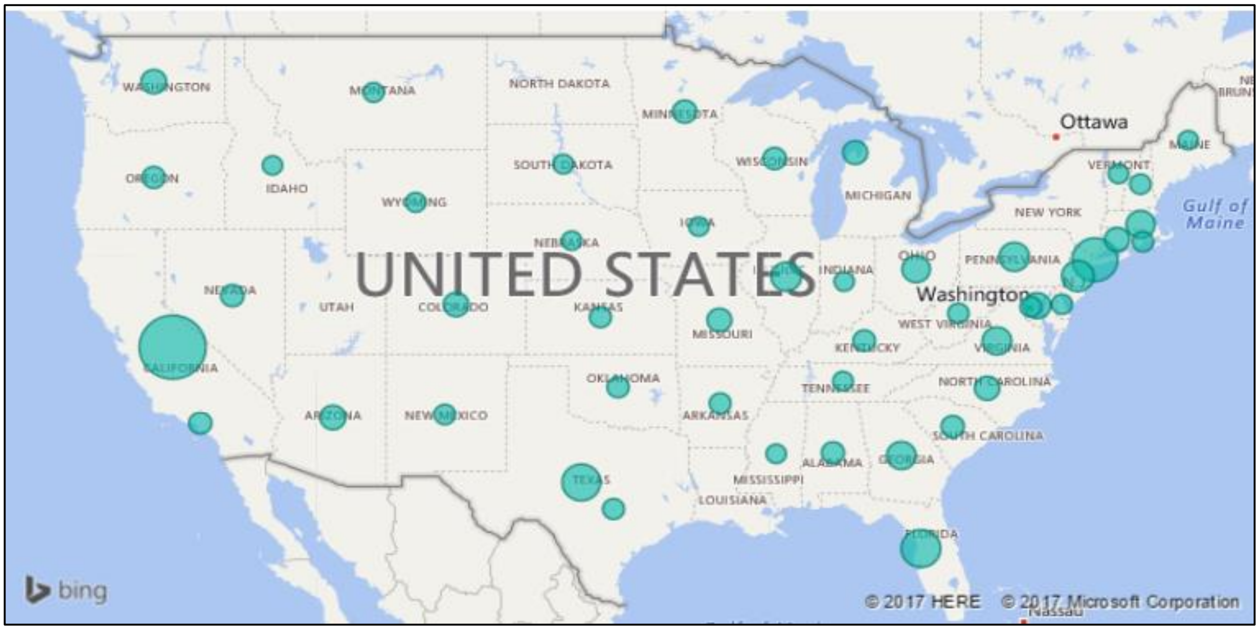


Figure 6: Distribution of loans across US

- We can see in the graph below that the default rate increases with grade. This is understandable because the different grades represent the different categories of loan in increasing order of risk profile. The interesting thing to note here is that the default rate for all the categories of loans is high in 2007 as compared to any other year. This is because of the sub-prime crisis that hit the US in 2007-2008.

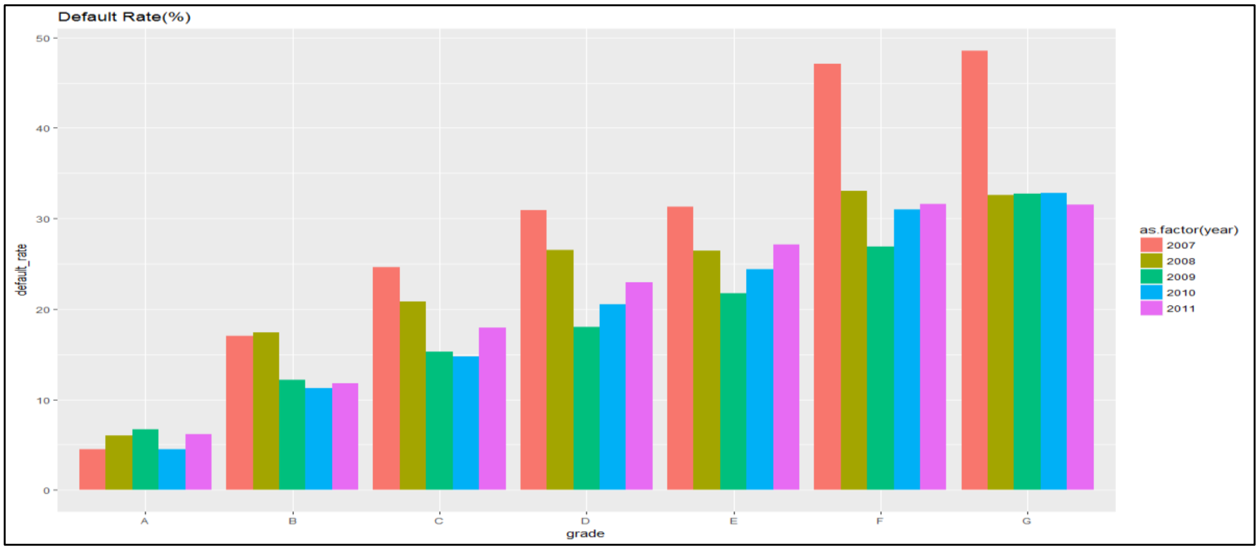


Figure 7: Default rate by grade over the years

- The graph comparing the interest rates for different grades of loan over the years tells us that there is an increasing trend of interest rate as loan grade or issue year increases

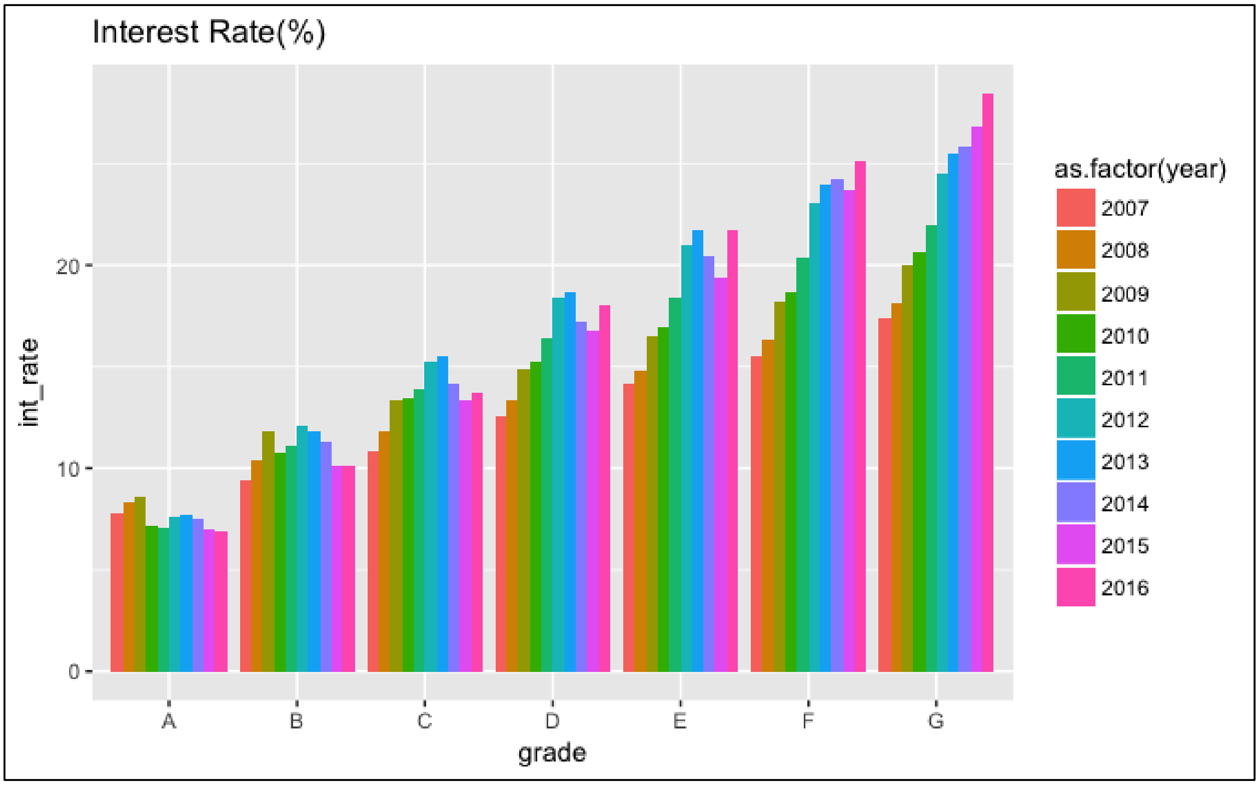


Figure 8: Interest rate by grade over the years

- The scatter plot of interest rate, loan amount, and term of payment tell us that most of the short term loans have a lower interest rate

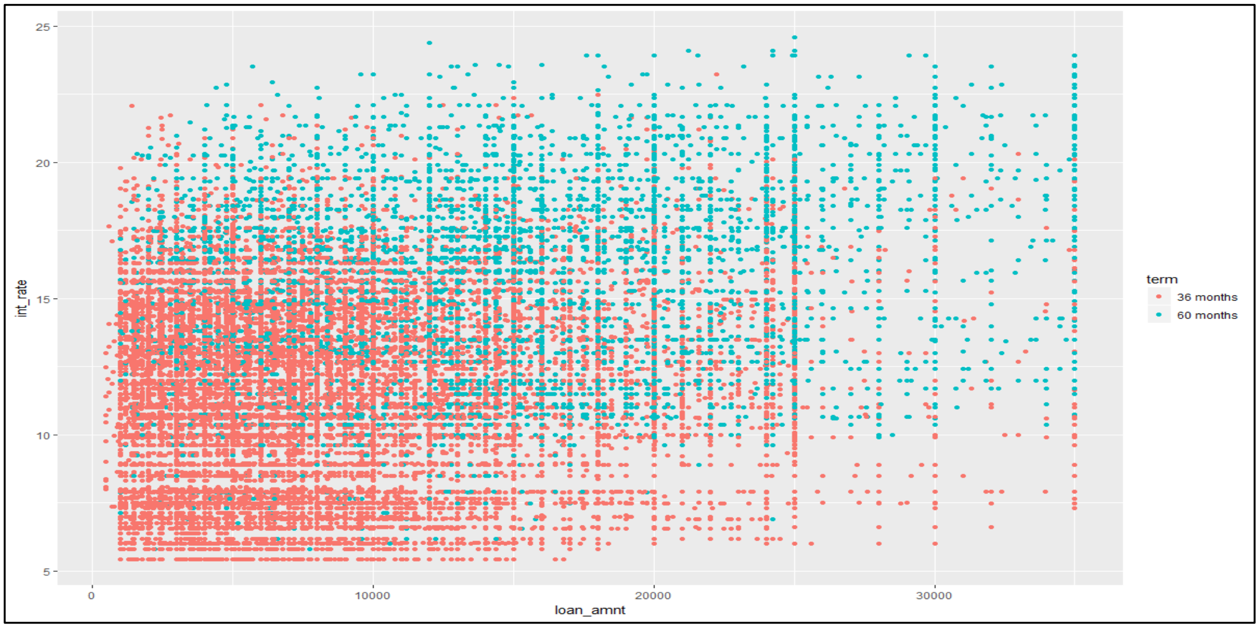


Figure 9: Scatter plot of interest rate, loan amount and term of payment

- This graph shows us that the most of 'Fully Paid' loans belonged to grades A and B whereas most of the defaulting loans belonged to grades C and D

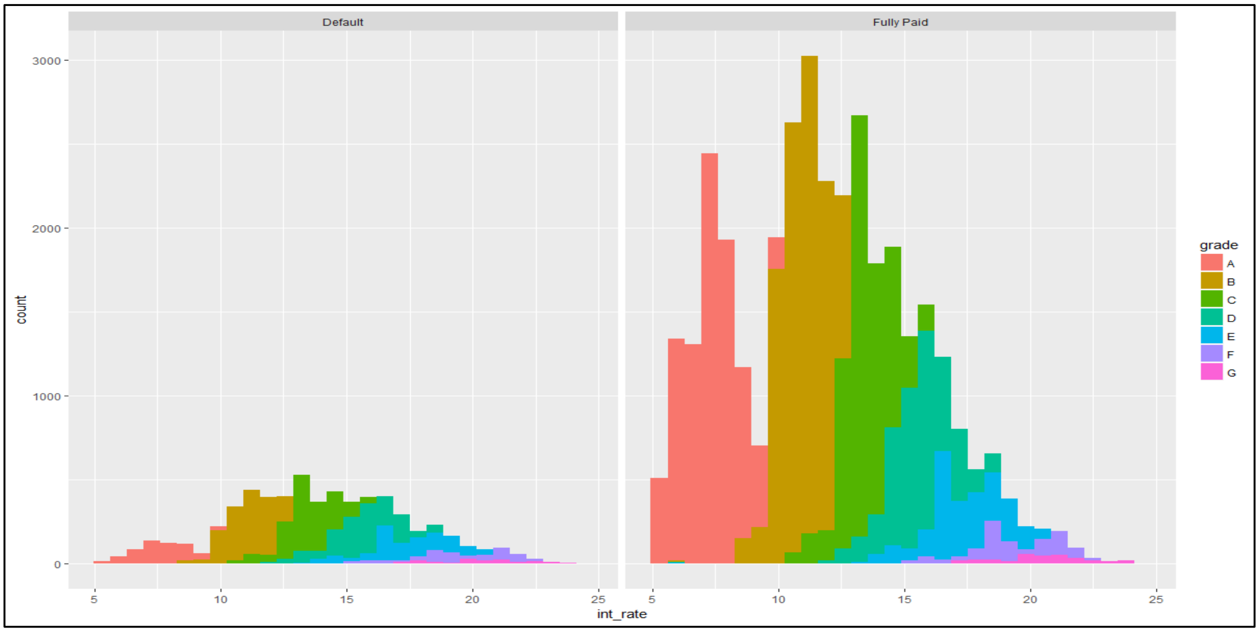


Figure 10: Distribution of response variable by grade

- The boxplot shows how interest rate varies with respect to purpose of loan. The highest interest rates are charges for housing and small businesses. Car loan is usually the cheapest with median rate of 11%

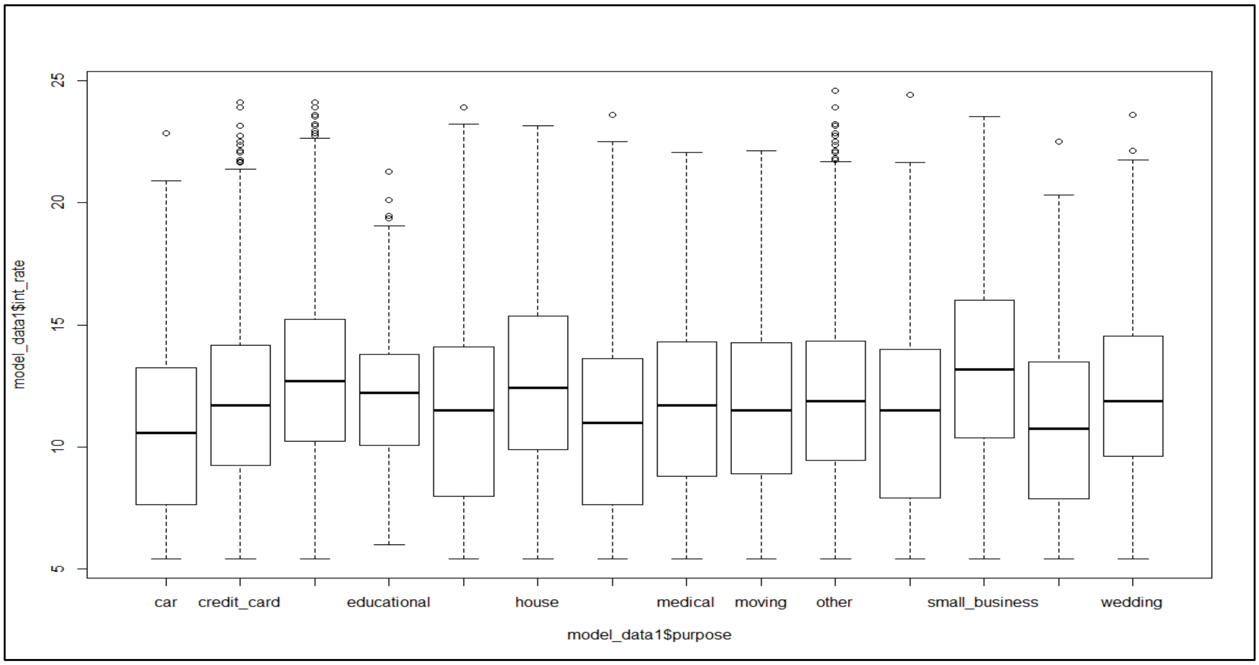


Figure 11: Boxplot of interest rate with loan type

- In continuation with the previous analysis, we tried to dig in to see if there is any home ownership type that is influencing the high interest rates. However, there doesn't seem to be any trend

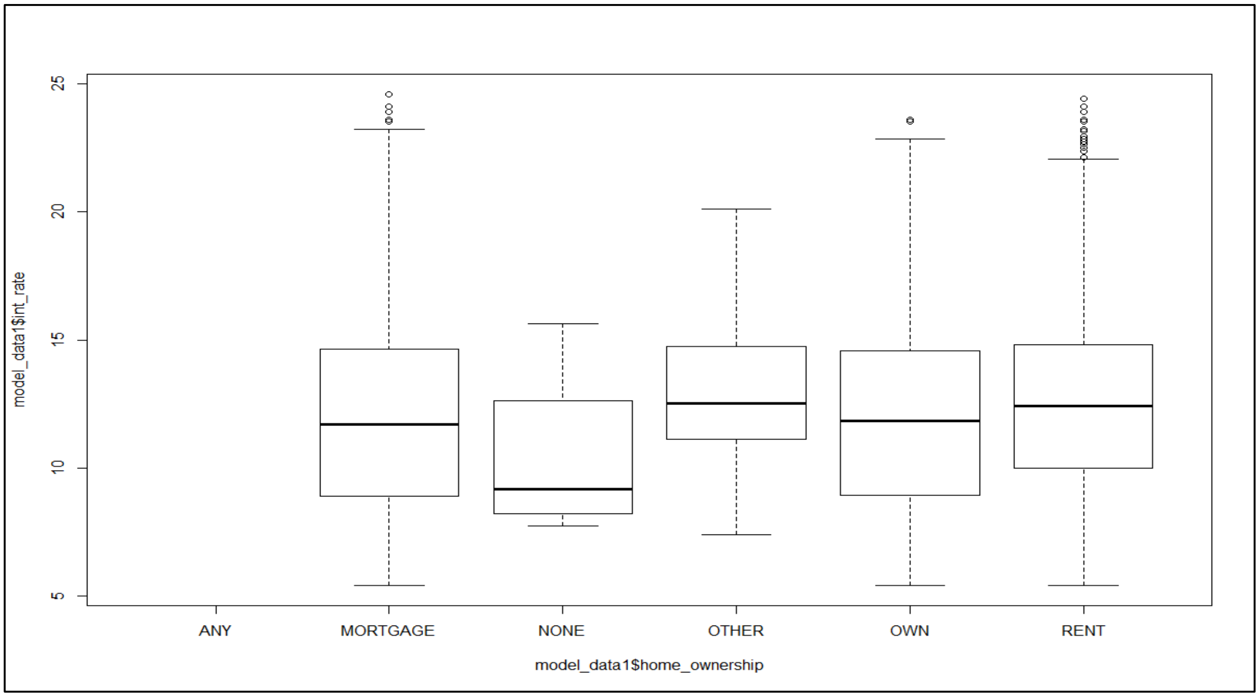


Figure 12: Boxplot of interest rate by type of home ownership

6.Methodology

Since our objective is binary classification of an up-sampled dataset, we have incorporated the well-known classification algorithms identified in various papers [9][10]. The classification techniques used have an underlying algorithm that calculates probabilities of falling in a particular class and classifies based on a cutoff value. The various techniques used in our research are described below along with the modifications and tuning used to get the optimum results.

6.1 Classification and regression trees (CART) [14]

The CART algorithm uses the concept of a decision tree which has a binary recursive partitioning procedure capable of processing continuous and nominal variables both as predictor and response variable. The classifier uses the Gini-coefficient to guide tree growing [15]. Trees are grown to a maximal size without the use of a stopping rule and then pruned back to the root via cost-complexity pruning. The results obtained by CART are easy to interpret as they are in the form of trees. For regression trees, the following equation is solved $\min(j,s)[\min_{c1} \sum(y_i - c_1)^2 + \min_{c2} \sum(y_i - c_2)^2]$

<u>Limitations:</u> It uses a greedy algorithm while making the splits i.e it does not try to minimize the overall error but only the error at the next node. The resulting trees are highly unstable and have a very high variability due to which the results may not be reliable always. It uses a binary partitioning algorithm on each node; this approach may not yield best results for all the models. It is known to overfit the data at times

6.2 Bagged Classification and regression trees [16]

This algorithm makes use of the bootstrap aggregation technique alternatively known as bagging. It is an ensemble method that creates individuals for its ensemble by training each classifier on a random redistribution of the training set. This random sampling is done with replacement and multiple times so that each individual is selected at least once in the ensemble. The final aggregate classifier can be obtained by averaging (regression) or majority voting (classification). When bagging with decision trees, we are not very concerned about individual tree overfitting the training data. Hence, the individual decision trees are grown deep and the trees are not pruned. The process of averaging and voting reduces the variance to a great extent

<u>Limitations:</u> Since the result obtained as a result of averaging or voting is no longer a tree, we lose the interpretability that we had in the original CART. The variable importance of predictors is not as accurate as predicted by other ensemble models like Random Forest

6.3 Random Forest [17]

Random forest is an ensemble method for predicting (classifying) observations by growing decision trees and randomly selecting subsets of predictor variables at each node. It first divides the 'd' dimensional space into discrete regions such that each region is homogeneous. This division is done based on least squares method and decision trees. Based on this decision tree, a new response observation is allocated a region. The value assigned to this new observation will be the mean of all the observations in that region. This procedure is repeated several times with different decision trees. The final value allocated to the new observation will be the average of predicted values by different decision trees. This averaging logic helps in reducing the variability in the dataset and leads to creation of stable trees.

Formally, a random forest is a predictor consisting of a collection of randomized base regression trees $\{r_n\ (x, \theta m, Dn), M \geq 1\}$, where $\theta_1 \ldots \theta_m$ are i.i.d outputs of a randomizing variable.

The other salient part of this model is the random subset selection. Due to this, effects of multicollinearity in the dataset are greatly reduced. This is because with random selection the relative importance of each variable can be distinctly calculated. The random forest algorithm beautifully balances the bias-variance and multicollinearity-Rsq tradeoffs and does the best variable selection while ensuring that the Rsq value is not affected greatly. Also, the random forest model does not have assumptions about distribution of data like in case of regression models. In other words, it is a non-parametric model.

Limitations: Like in case of the Bagged CART algorithm, the results are not easily interpretable. The interpretability is better than Bagged CART because this model gives us an accurate variable importance plot however, it is still lesser than models like linear regression and decision trees

6.4 Logistic Regression [18]

This model is an extension of the regression models and used mainly for classification type problems. Unlike actual regression, logistic regression does not try to predict the value of a numeric variable. It predicts a probability that the given input point belongs to a certain class (binary classes in our case). The key assumption in this algorithm is that our input space can be separated into distinct regions (2 in our case), by a linear boundary. This linear boundary is referred to as the linear discriminant. They key statistical metric used in this case is the logit variable. The logit variable is defined as the logarithm of ratio of probability of an event happening over probability of the event not happening.

Logit = $\log(\frac{\text{probability of an event happening}}{1-\text{ probability of an event happening}})$ = log (odds)

Instead of finding the best fitting line by minimizing the squared residuals like in the case of linear regression, in logistic regression, the algorithm tries to find the smallest possible deviance between the observed and predicted values. This I known as the maximum likelihood. For a binary classification problem with two predictors, the regression equation that the model will come up with is log(odds) = $\beta_0 + \beta_1 x_1 + \beta_2 x_2$. = a. Once the regression equation is developed, we can calculate the odds ratio by doing e^a. After getting the individual odds ratios, we can calculate the individual probabilities as follows P = $\frac{e^a}{1+e^a}$

Limitations: It can only predict categorical outcomes and not continuous outcomes. It tends to overfit the data

6.5 Bayesian Additive Regression trees (BART) [20]

Bayesian Additive Regression Trees (BART) is a statistical sum of trees model. It can be considered a Bayesian version of machine learning tree ensemble methods where the individual trees are the base learners. It is a combination of trees model where each tree is constrained by a regularization prior to be a weak learner (like in boosting models). It uses an iterative Bayesian backfitting approach to generate samples from the posterior. It has a robust algorithm as the trees are made using conditional probability

Y = $(\sum_{j=1}^{m} g(x; Tj, Mj)$ + error (error is Normally distributed with mean of 0 and variance of σ^2)

Where Tj is a binary regression tree.

Limitations: It is computationally very intensive and takes a lot of time to give results

6.6 Linear Discriminant Analysis (LDA) [21][22]

The general LDA approach is like a Principal Component Analysis just that in addition to finding the component axes that maximize the variance of our data (PCA), we also try to maximize the separation between multiple classes. LDA projects feature spaces into smaller subspaces while maintaining the class discriminatory information. LDA maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability.

6.7 Quadratic Discriminant Analysis (QDA) [23]

Quadratic discriminant analysis is similar to LDA. The main difference is that we assume different covariance matrix for each class and therefore we will have to estimate the covariance matrix for each separately for each class

$\delta_k(x) = -0.5 * \log|\Sigma_k| - 0.5 * (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k$

The decision boundaries are quadratic in x and because QDA allows more flexibility in the in the model for covariance matrix, it tends to perform better than the QDA.

7.Model fitting and results

7.1 Classification and regression trees

We used 10-fold cross validation to give us the best combination of arguments with the goal of minimizing our misclassification error. The optimum parameters obtained were complexity parameter of 0.01. Max depth of trees (stopping rule) as 4 and minimum split of 10. Using these parameters, the output obtained were as follows:

	Misclass. Error	Test accuracy	Sensitivity - Train	Specificity - Train	AUC - Train	Sensitivity - Test	Specificity - Test	AUC - Test
RPART	0.39	0.5	0.78	0.45	0.61	0.77	0.45	0.61

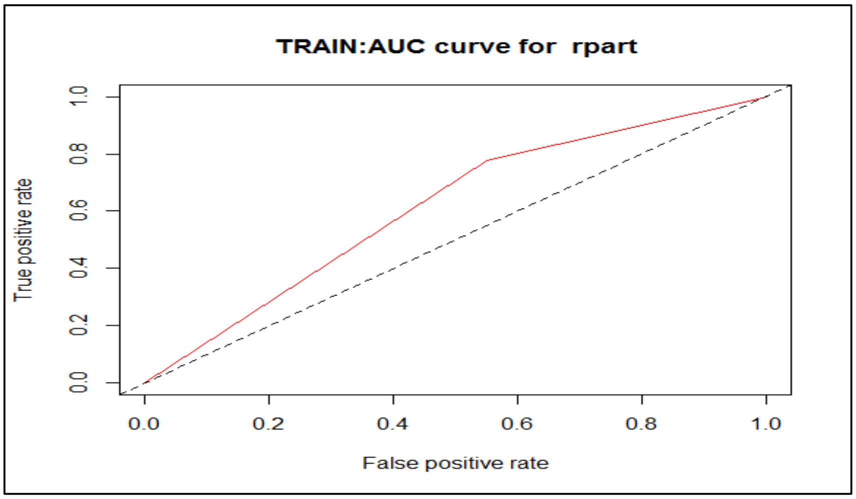


Figure 13: AUC curve for rpart - Train

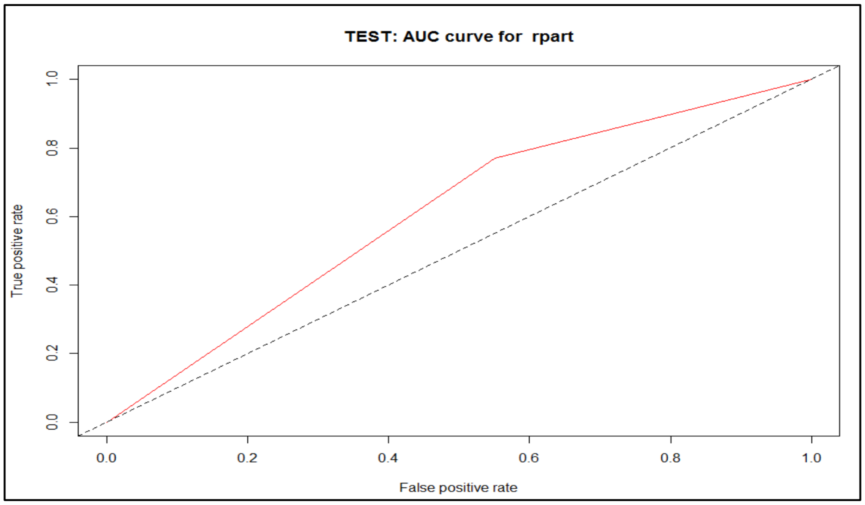


Figure 14: AUC curve for rpart - Test

7.2 Bagged classification and regression trees

We used cross validation function and obtained the optimum parameters for our model. The parameters obtained were as follows:

	Misclass. Error	Test accuracy	Sensitivity - Train	Specificity - Train	AUC - Train	Sensitivity - Test	Specificity - Test	AUC - Test
Bagging	0.00	0.84	1.00	1.00	1.00	0.08	0.97	0.69

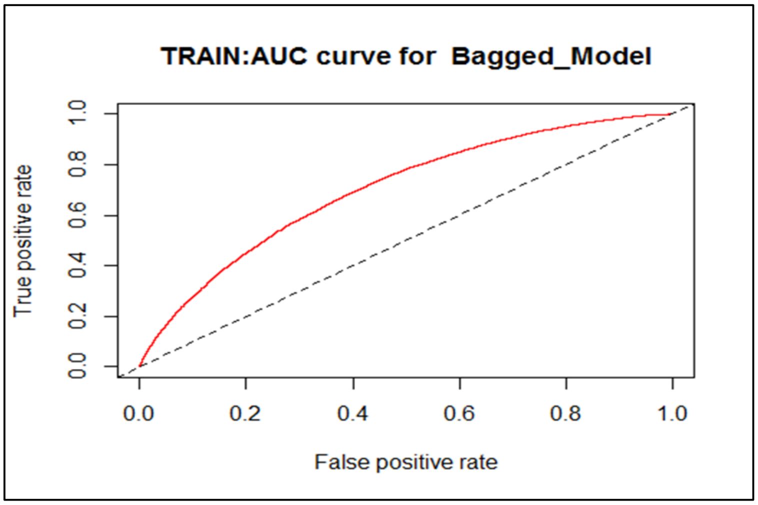


Figure 15: AUC curve for bagged CART - Train

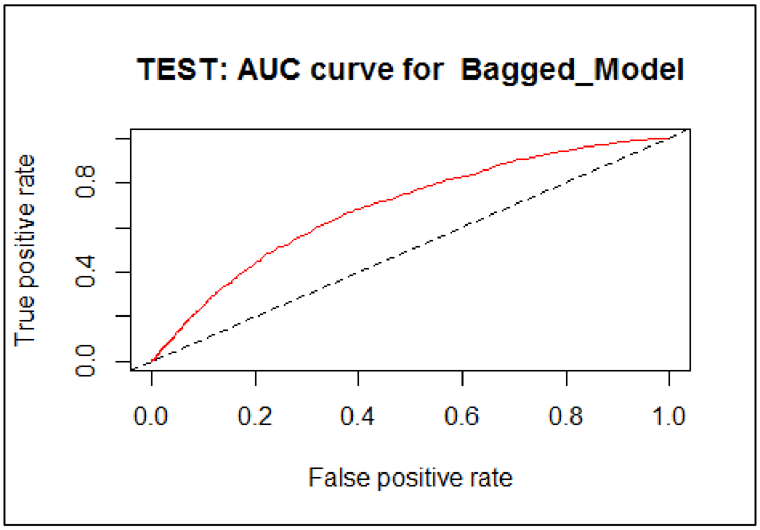


Figure 16: AUC curve for bagged CART - Test

7.3 Random Forest

We designed a function and ran the random forest model multiple times for different combinations of argument. Using this iterative approach of parameter selection, we identified our best parameters. We selected the number of trees as 800 with 5 variable splits at each node. The results we obtained were as followed:

	Misclass. Error	Test accuracy	Sensitivity - Train	Specificity - Train	AUC - Train	Sensitivity - Test	Specificity - Test	AUC - Test
Random forest	0.10	0.80	0.93	0.90	0.95	0.24	0.90	0.67

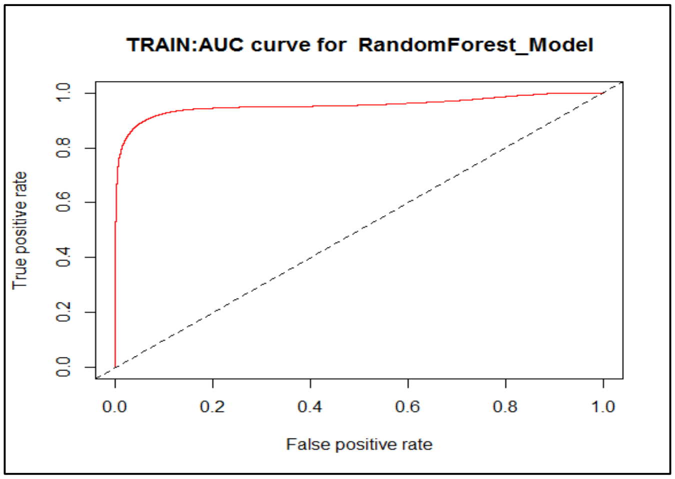


Figure 17: AUC curve for Random forest - Train

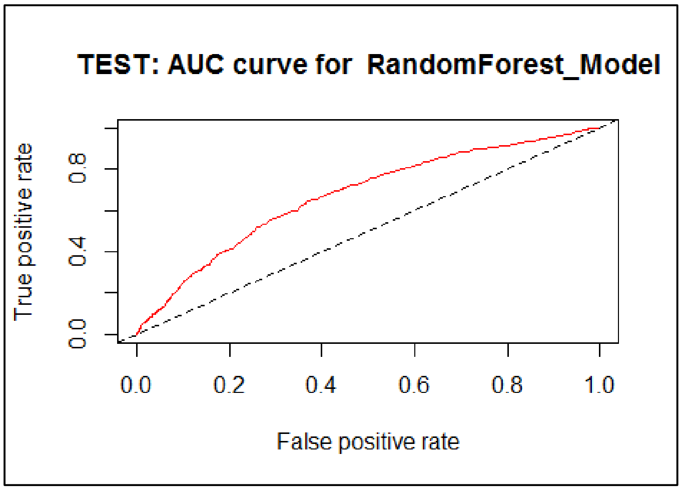


Figure 18: AUC curve for Random forest - Test

7.4 Logistic regression

The underlying assumption in all regression models is that the predictor variables should not be correlated with each other. It was imperative that we check this assumption before fitting our model. We generated a correlation matrix of our predictor variables and removed the collinear variables (r≥0.9) before fitting the model. We first built the logistic model with default parameter and later we tuned the different values of logistic. Using 10-fold CV, we optimized the Elastic-net mixing parameter(α) and regularization parameter (λ). The optimized values were 0.5 and 0.00502 respectively.

	Misclass. Error	Test accuracy	Sensitivity - Train	Specificity - Train	AUC - Train	Sensitivity - Test	Specificity - Test	AUC - Test
Logistic regression	0.34	0.65	0.62	0.64	0.69	0.66	0.64	0.71

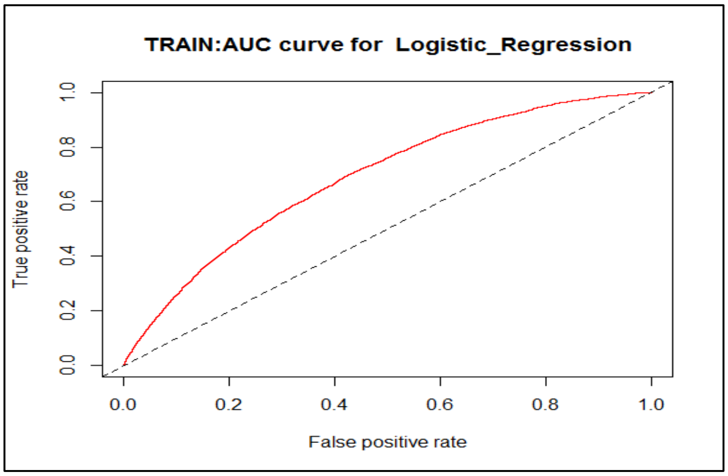


Figure 19: AUC curve for Logistic regression - Train

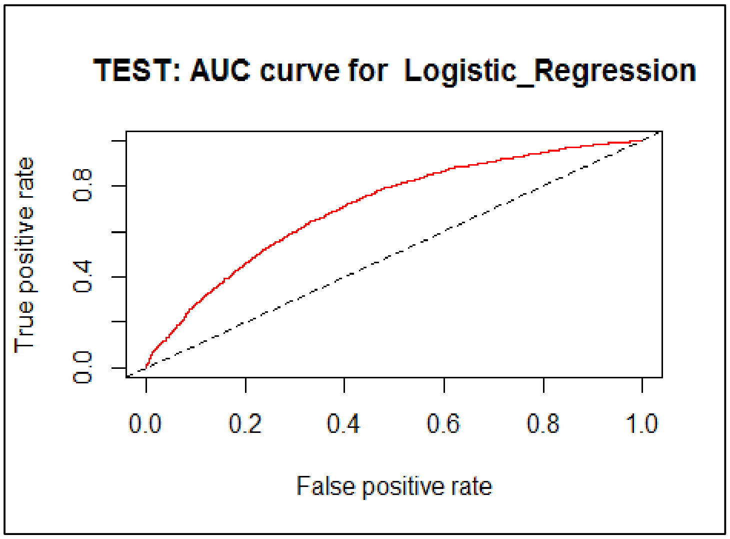


Figure 20: AUC curve for Logistic regression - Test

7.5 Bayesian Additive Regression trees

Since the BART algorithm is computationally very expensive, we were not able to use cross validation or tuning to generate the best set of input parameters. However, we did not want to compromise on the efficiency of our model so we did some research and also read the original paper[20] on BART written by Hugh Chipman. The conclusion was that the default parameters tend to perform very well for this algorithm and the only tuning parameter that could improve our predictive performance was number of trees. Hence, we generated three BART models with 50, 100 and 130 trees respectively. Our findings were in line with the research and only observed small incremental change in the third decimal. The parameter that we selected for our final BART model was 100 trees. The results obtained are as follows:

	Misclass. Error	Test accuracy	Sensitivity - Train	Specificity - Train	AUC - Train	Sensitivity - Test	Specificity - Test	AUC - Test
BART	0.28	0.65	0.75	0.70	0.80	0.63	0.68	0.71

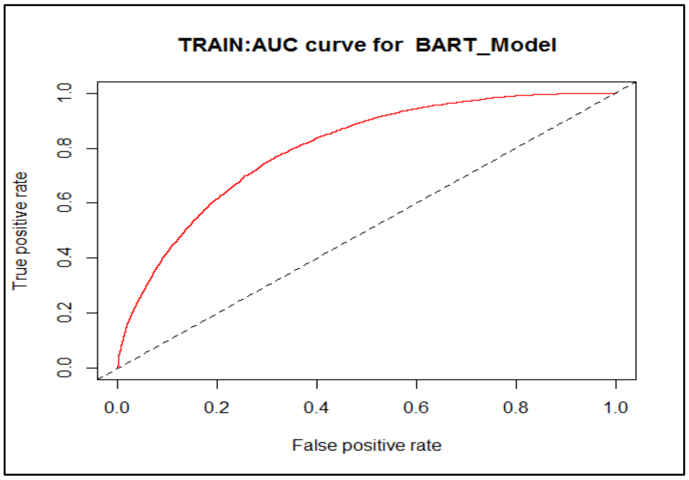


Figure 21: AUC curve for BART - Train

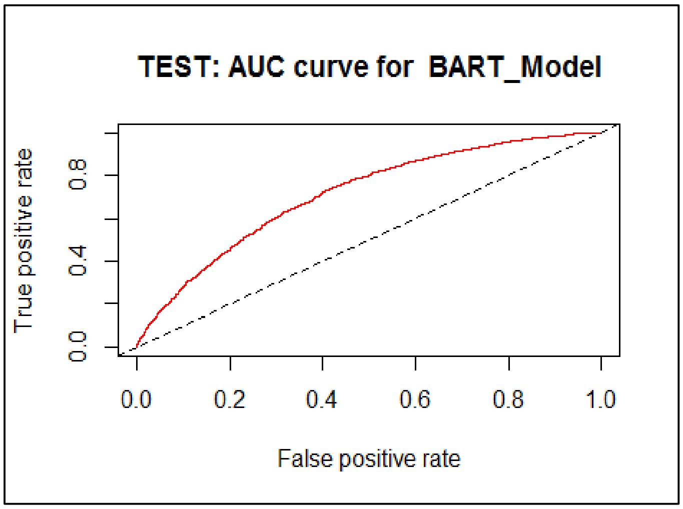


Figure 22: AUC curve for BART - Test

7.6 Linear Discriminant Analysis

We used 10-fold cross validation approach to fit this model and obtained the following results.

	Misclass. Error	Test accuracy	Sensitivity - Train	Specificity - Train	AUC - Train	Sensitivity - Test	Specificity - Test	AUC - Test
LDA	0.64	0.65	0.65	0.64	0.70	0.67	0.64	0.70

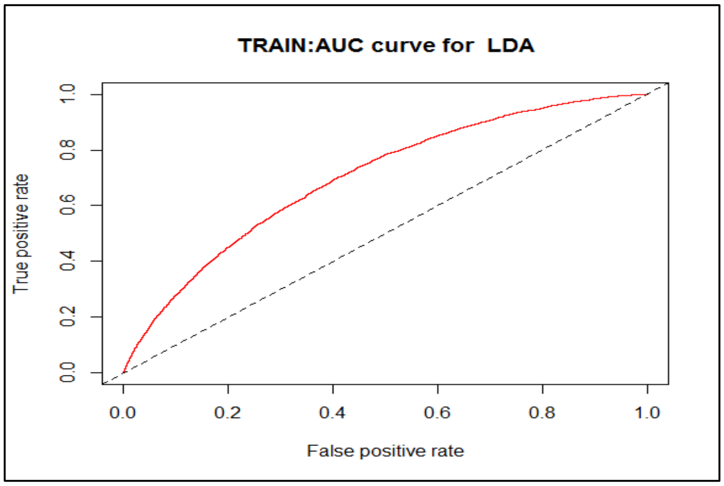


Figure 23: AUC curve for LDA - Train

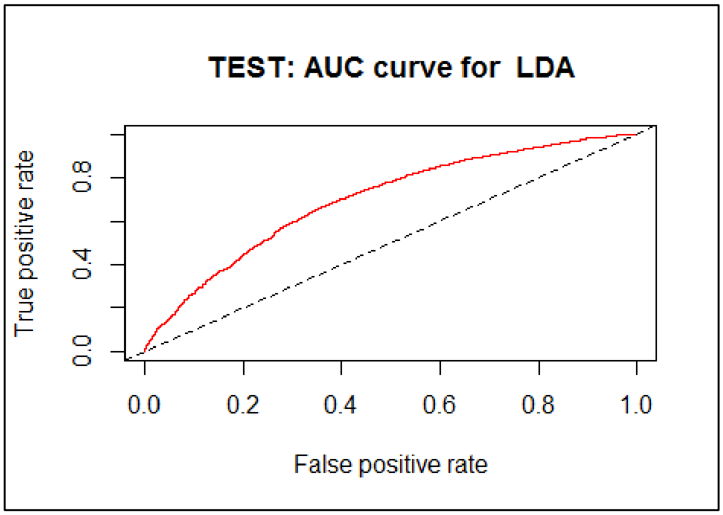


Figure 24: AUC curve for LDA - Test

As we can see from the table and graph that our performance is not good in this model. We then went on to check the assumption of this model and found that the basic assumption was that our dataset should be linearly separable. Unfortunately, this was not the case and that's the reason this model did not work well.

7.7 Quadratic Discriminant Analysis

We used 10-fold cross validation approach to fit this model and obtained the following results.

	Misclass. Error	Test accuracy	Sensitivity - Train	Specificity - Train	AUC - Train	Sensitivity - Test	Specificity - Test	AUC - Test
QDA	0.65	0.63	0.46	0.78	0.68	0.48	0.77	0.68

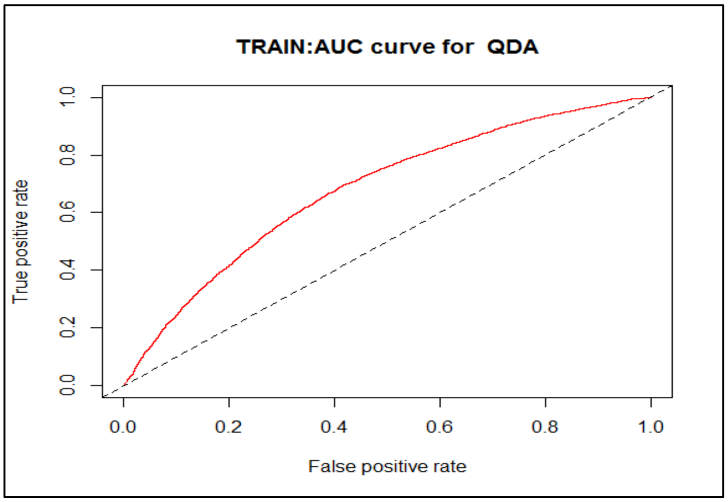


Figure 25: AUC curve for QDA - Train

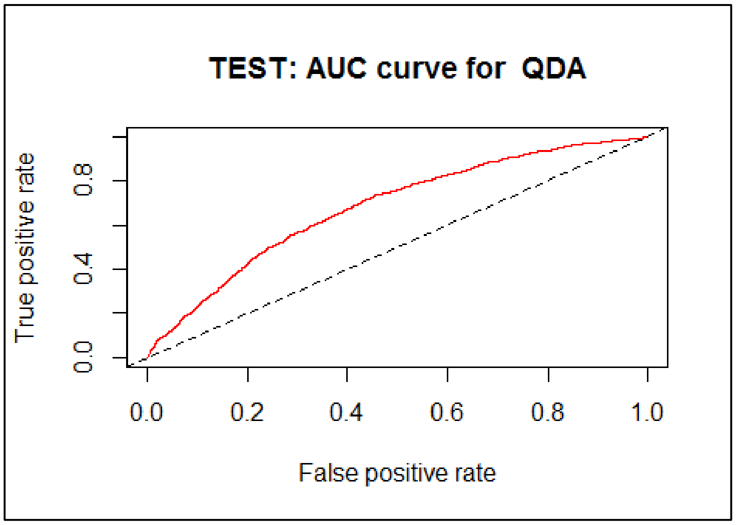


Figure 26: AUC curve for QDA - Test

8.Model Selection

Traditionally test accuracy is considered as a key metric for measuring the accuracy of the dataset. However, for unbalanced datasets, this approach is considered erroneous. This is because we are orienting our learners in the models to be oriented towards predicting the majority class. As a result, the resulting figure would be misleading. This was one of the gaps in the existing research that we set out to rectify. The first step that we took towards addressing this issue was to make our training dataset balanced. This eliminated the training bias that was created while building models using an unbalanced dataset. Based on our research, we decided that the best model for our dataset should be selected on a variety of statistical parameters like AUC, sensitivity, specificity, accuracy, and G-mean score [25].

Based on the results obtained after running all the models, we selected Random Forest as the best model because it performed well on majority of the key metrics identified by our research. We also looked into the variable importance chart to identify out key predictors.

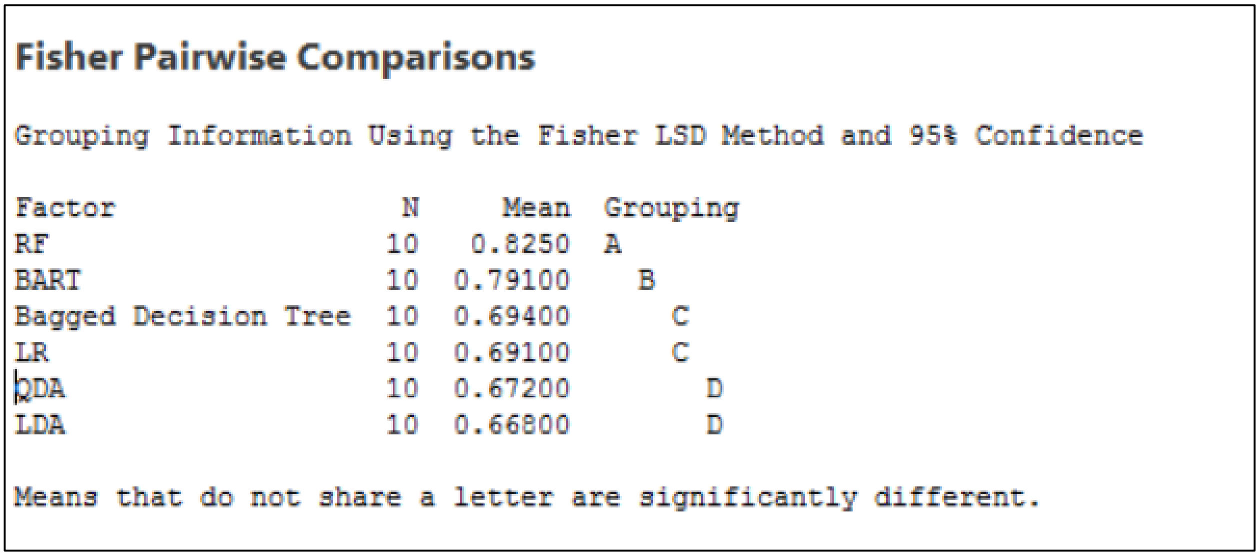
	Train				
Models	Accuracy	Sensitivity	Specificity	AUC	G-mean
Mean only	0.5	0.5	0.5	0.5	0.5
BART	0.72	0.75	0.70	0.80	0.72
Bagging	0.81	0.82	0.60	0.72	0.70
Random Forest	0.91	0.93	0.90	0.95	0.91
Logistic Regression	0.63	0.62	0.64	0.69	0.63
QDA	0.62	0.46	0.78	0.68	0.59
LDA	0.64	0.65	0.64	0.70	0.64

	Test				
Models	Accuracy	Sensitivity	Specificity	AUC	G-mean
BART	0.67	0.63	0.68	0.71	0.66
Bagging	0.84	0.08	0.97	0.69	0.27
Random Forest	0.81	0.54	0.90	0.67	0.69
Logistic Regression	0.65	0.66	0.64	0.71	0.65
QDA	0.73	0.48	0.77	0.68	0.61
LDA	0.64	0.67	0.64	0.70	0.65

8.1 Fisher Pairwise Comparisons

We used the Fisher test to compare the cross validated AUC and G-mean values for different models. The results are as follows:

AUC:



```
Fisher Pairwise Comparisons

Grouping Information Using the Fisher LSD Method and 95% Confidence

Factor                 N     Mean  Grouping
RF                    10   0.8250  A
BART                  10  0.79100    B
Bagged Decision Tree  10  0.69400      C
LR                    10  0.69100      C
QDA                   10  0.67200        D
LDA                   10  0.66800        D

Means that do not share a letter are significantly different.
```

Figure 27: Fisher test for AUC

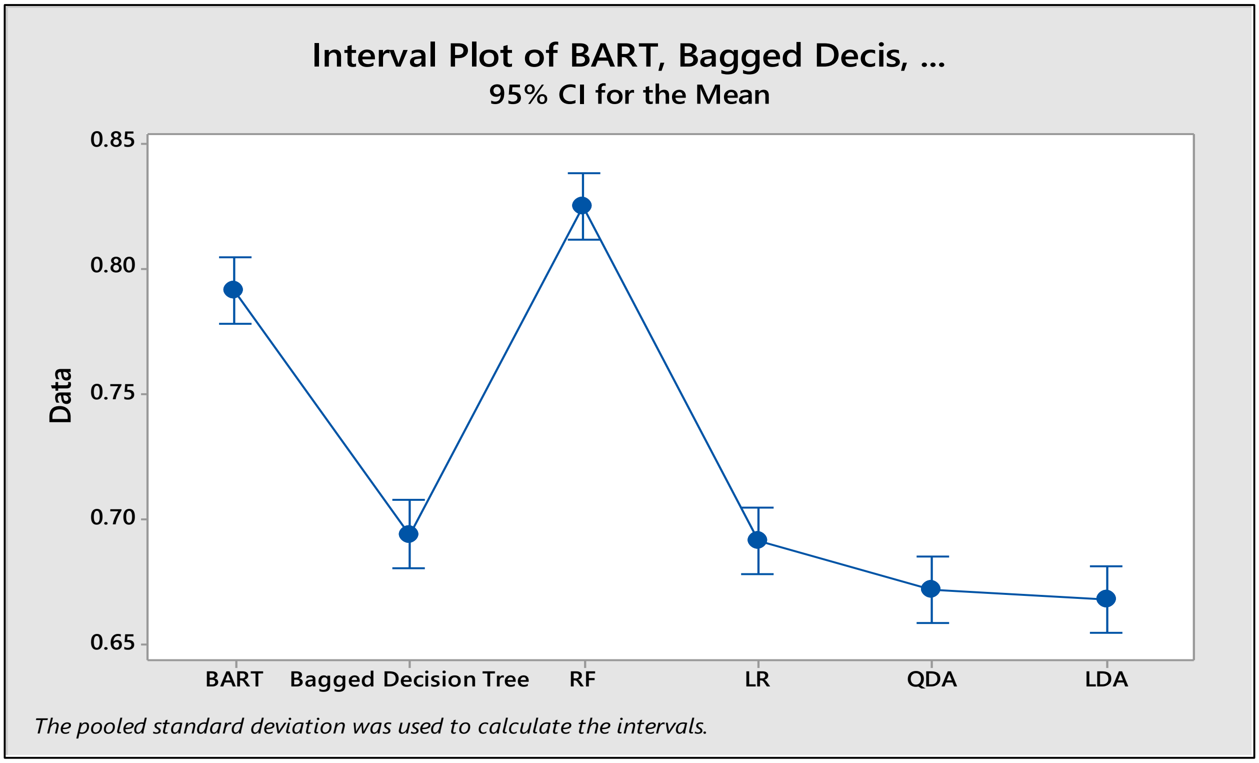
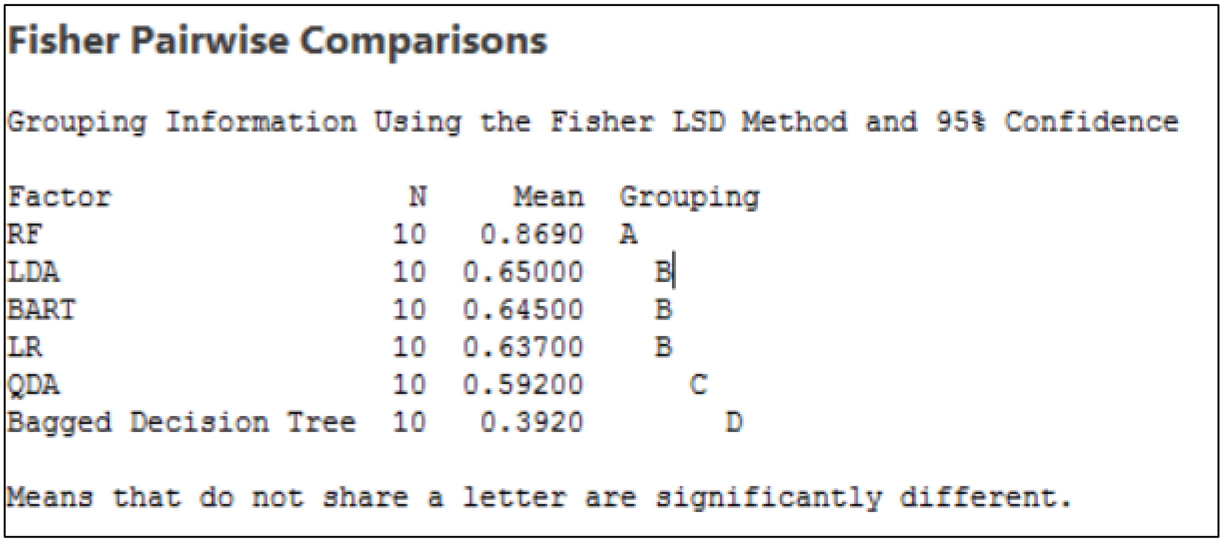


Figure 28: Interval plot for AUC

G-Mean:



```
Fisher Pairwise Comparisons

Grouping Information Using the Fisher LSD Method and 95% Confidence

Factor               N     Mean  Grouping
RF                  10   0.8690  A
LDA                 10  0.65000    B
BART                10  0.64500    B
LR                  10  0.63700    B
QDA                 10  0.59200      C
Bagged Decision Tree 10   0.3920        D

Means that do not share a letter are significantly different.
```

Figure 29: Fisher test for G-mean

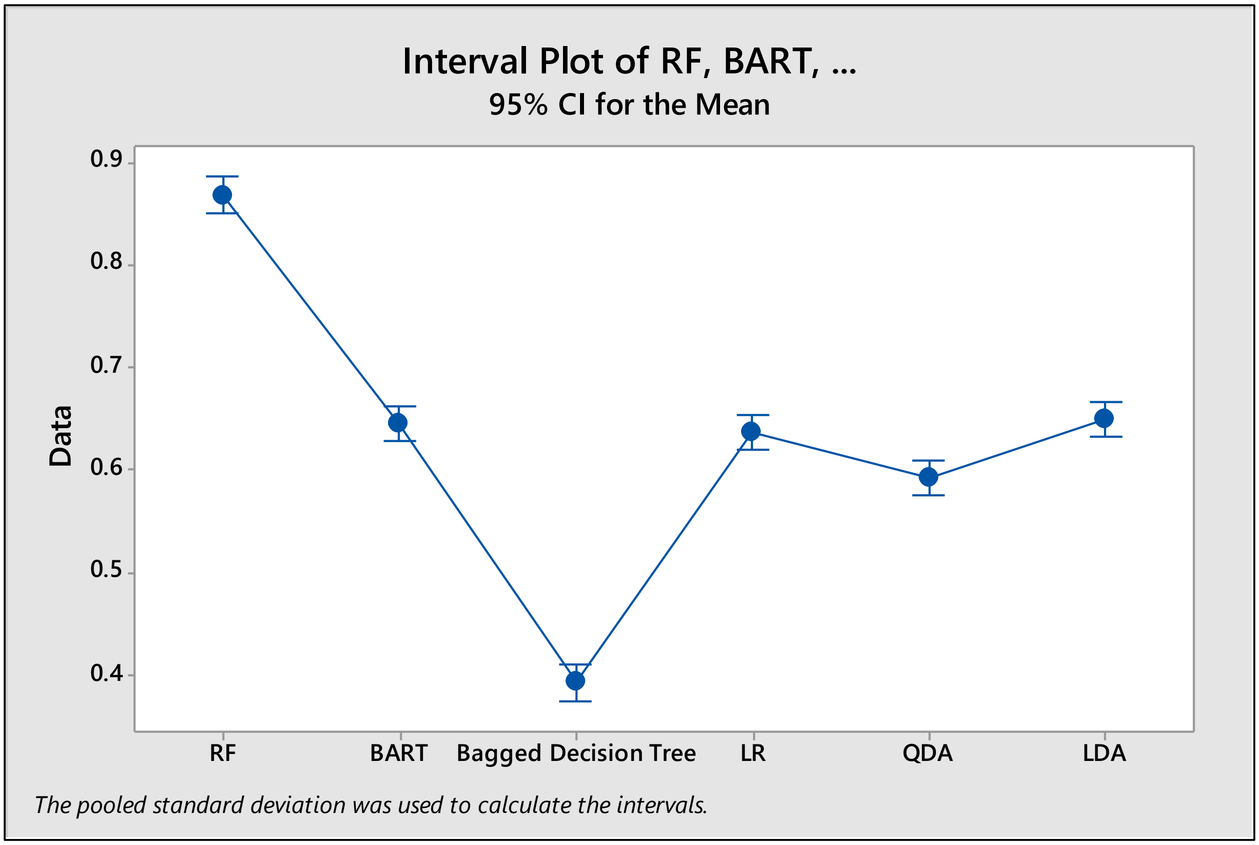


Figure 30: Interval plot for G-mean

These results clearly suggest that Random forest significantly performs better than all other models in terms of the two key evaluation metrics AUC and G-mean. Thus, we selected Random forest as the best model to explain our dataset.

9.Inferences

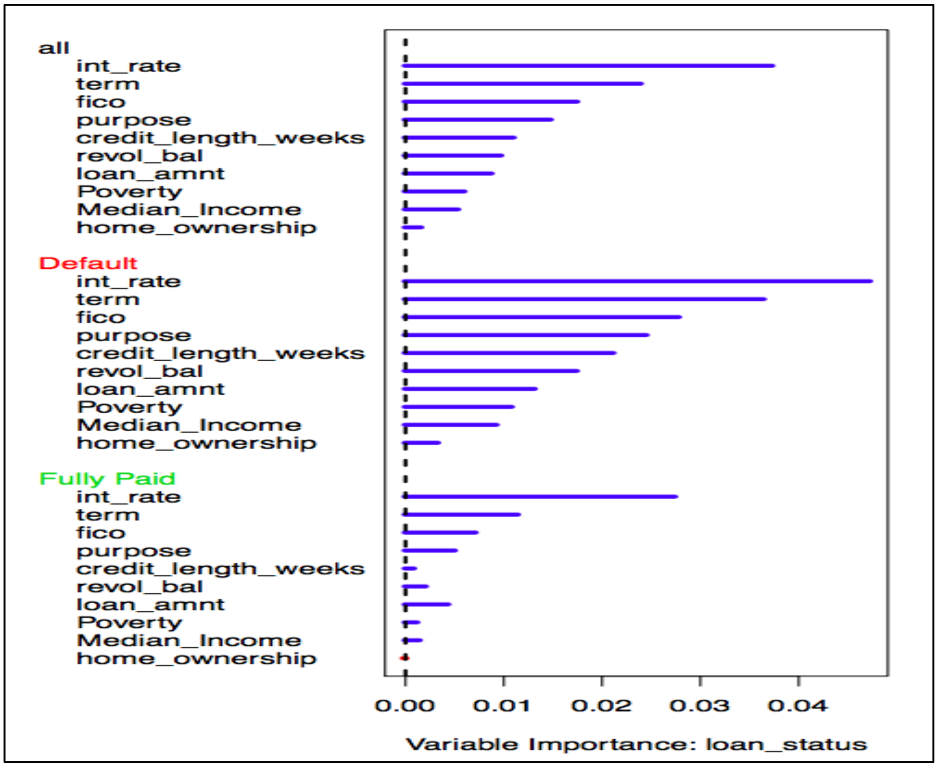


Figure 31: Variable importance plot

As we can see from the plot above, 'int_rate' (interest rate) is our top predictor. Followed by 'term' (loan term – short vs long). At the third position is the credit score of a person which traditionally has been the main if not the only predictor.

9.1 Variable importance

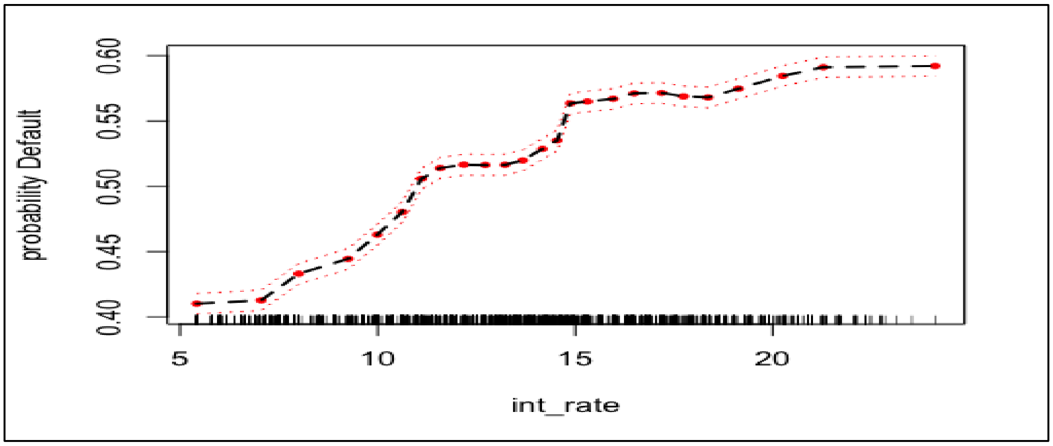


Figure 32: Partial plot of interest rate

The plot above suggests that as interest rate increases, the probability of default also increases. This is probably because the way lending club calculates interest rates is based on the perceived risk of a customer. Since our analysis is also on similar lines, it’s natural to get interest rate as an important predictor.

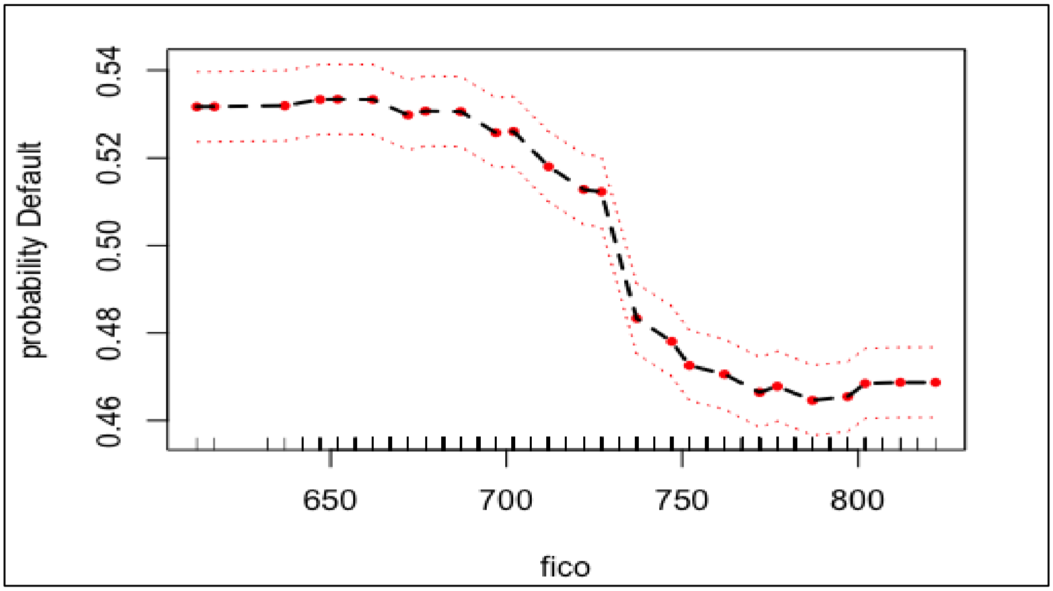


Figure 33: Partial plot of fico score

The plot above suggests that upto a fico score of about 720, the probability of default is high. Post that, there is a sharp decrease in the probability of default. It becomes fairly constant after a score of 760.

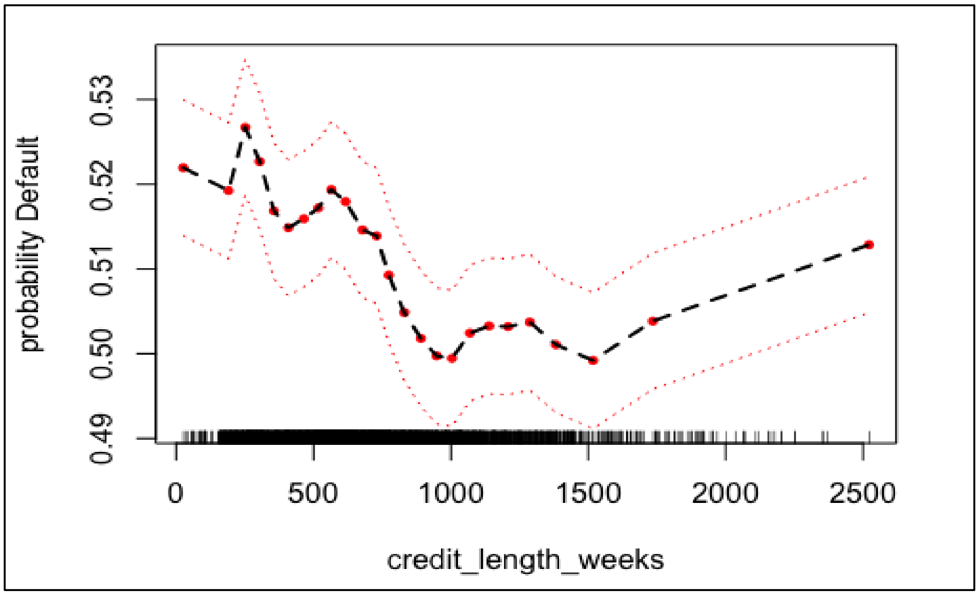


Figure 34: Partial plot of credit length

The above plot suggests that the probability of default is pretty high for customers in the age group of 18-38 (assuming average credit line starts at the age of 18). The probability reduces between the age group of 38-60 and increases again post 60. We can probably hypothesize that this is because younger people tend to be taking more risk and making investments for buying a house starting a business etc. Middle aged people tend to have a stable income and trying to settle down in life. Post the age of 60, there is no steady income and no incentive to build a positive credit score due to which the probability increases again.

9.2 Other inferences

- Our research highlighted that the default rate was highest for the loan purpose ‘house’. This result corroborates the existing work done in this field[24]
- On analyzing the results of our best model, we noticed that the state in which loan was issued was an important predictor. On digging deeper into the causes of this, we noticed that the Republican party was in power in the top 5 states in terms of default rates. This result is somewhat counter-intuitive because one would expect the Republican states to do well when the Republicans govern the congress as well.

State	Default Rate	Party
Nebraska	0.55	Republican
Indiana	0.37	Republican
Mississippi	0.23	Republican
Nevada	0.23	Republican
Alaska	0.20	Republican

- The macroeconomic factors added in our analysis like Median Income and Poverty level have made it to the list of top predictors suggesting that these factors play an important role in prediction
- The interest rate of the customer is also a key predictor because it is function of the type of loan the person is taking and the credit score (fico score) of the person
- The credit length of a person (created via feature engineering) i.e the amount of time a person has been receiving credit also determines the default probability of a person. Customers who have received credit previously or at a young age, tend to be better customers in terms of risk aversion
- The company made heavy losses in the years of the recession and came back to normal operations only after 2008. This is reflected in the year on year return on investment metric

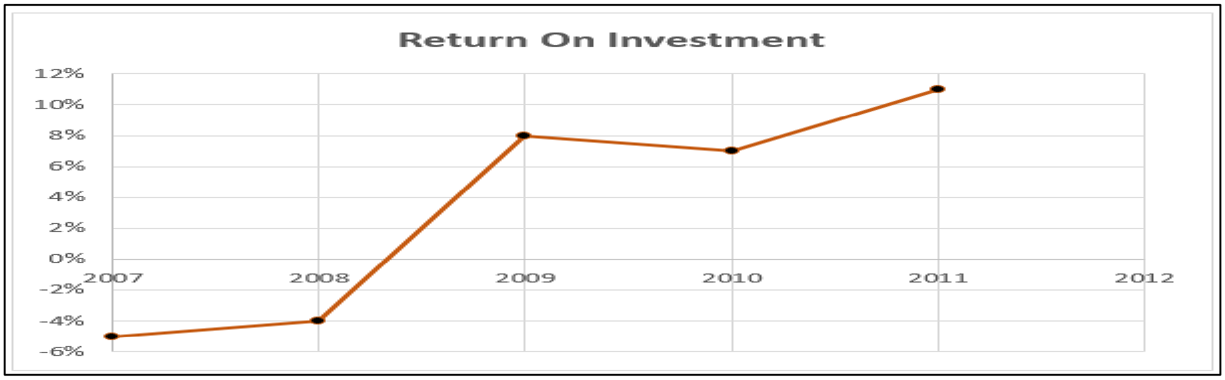


Figure 35: Return on investment over the year

10. Response to reviewers

We would like to thank professor Nateghi and all the reviewers for taking time out and reviewing our project. We got the much needed “outsiders” point of view and it really helped us in understanding the positives of our project and the areas of improvement. We sincerely appreciate the feedback and have tried our best to address the comments.

- ➢ Comment 1: Figure captions and table of figures is missing
- ✓ Response 1: Added captions for all the figures and tables and also included the table of contents

- ➢ Comment 2: Data labels are not clear in Figure 1.
- ✓ Response 2: We modified the graph and made sure that the data labels are more legible to the reader.

- ➢ Comment 3: Support vector machine should have been applied to the dataset as it is known to do very well for classification problems
- ✓ Response 3: We used this algorithm on our dataset however, the results were not as good as the best model that we selected, hence we decided to not include the additional algorithm in our report

- ➢ Comment 4: The term LDA was used before describing the term
- ✗ Response 4: We respectfully disagree with this comment. We have described all the abbreviations in the contents and methodology sections. Only after describing them we have used the abbreviations

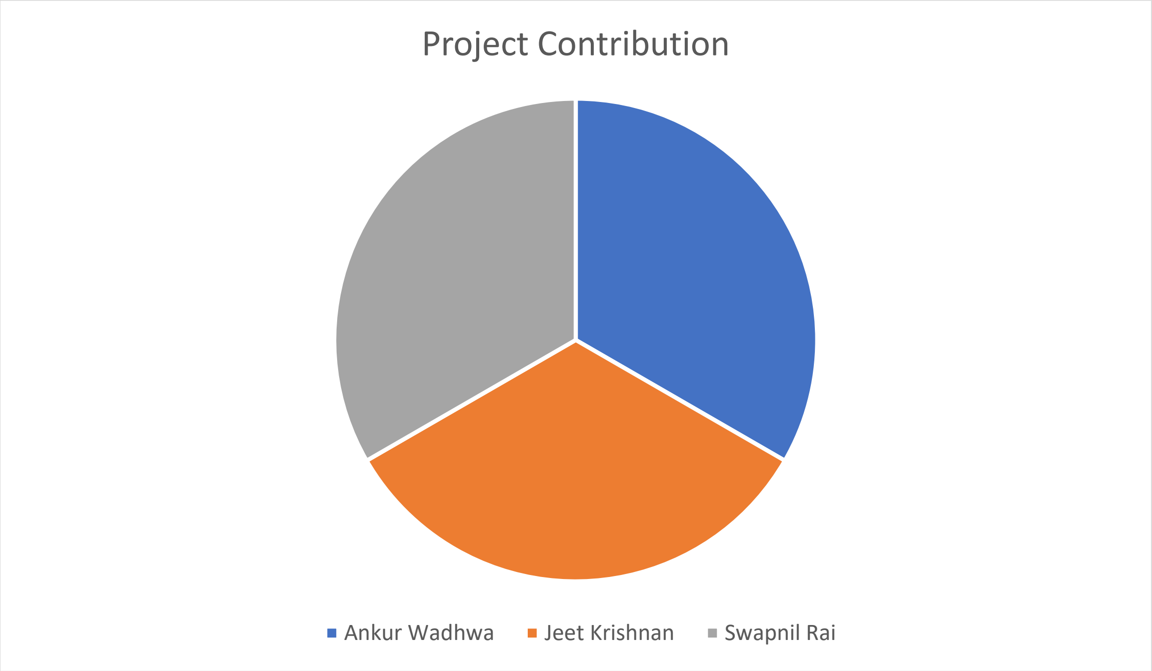
- ➢ Comment 5: G-mean was mentioned for model selection but not explained
- ✓ Response 5: We have explained G-mean in literature review section

- ➢ Comment 6: Lift area mentioned but not explained.
- ✓ Response 6: Earlier we planned to use that but later we moved to better methods so it has been removed

11.Future scope of work

- We would like to allocate a penalty in the model for incorrectly predicting defaults so that our specificity improves
- We plan to expand our analysis and fit a regression based algorithm to predict the interest rate that a potential customer should receive
- To calculate Bias, Variance, and noise of each model from the error term so that we can work individually on these parts

12.Contribution



Roles and Responsibility

- Ankur Wadhwa:
- Jeet Krishnan:
- Swapnil Rai:

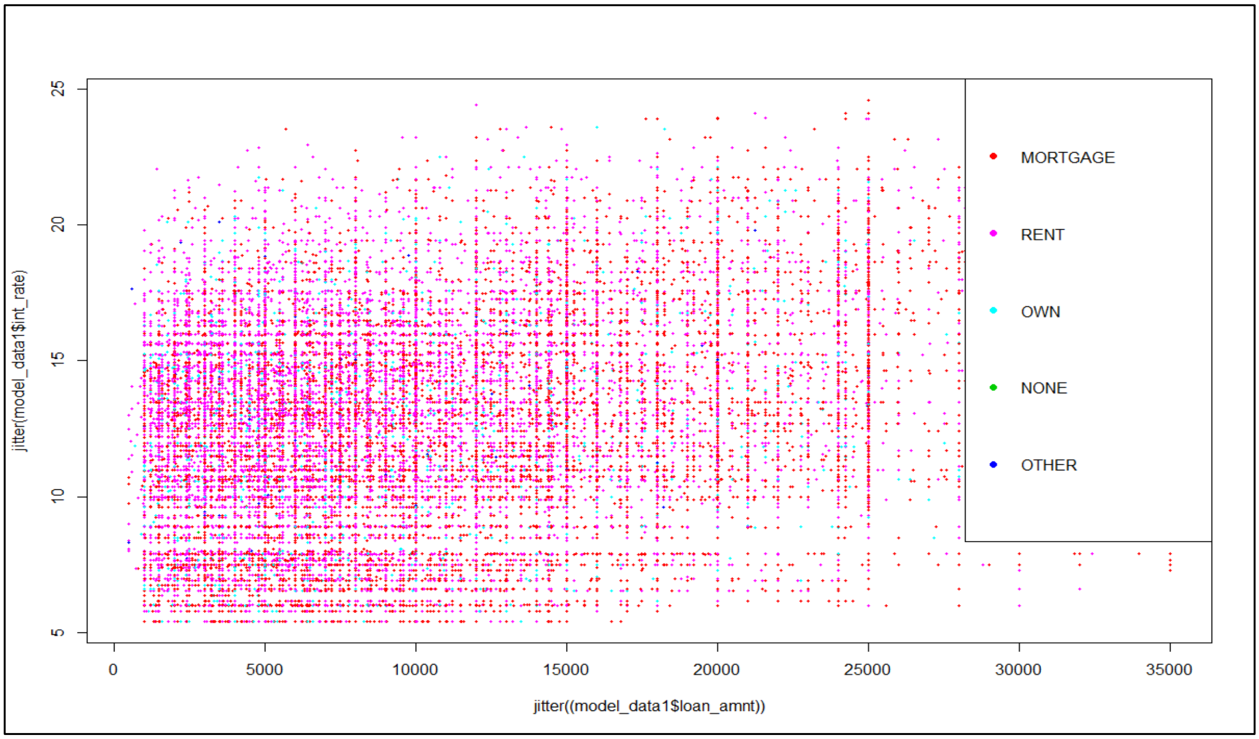
13.References

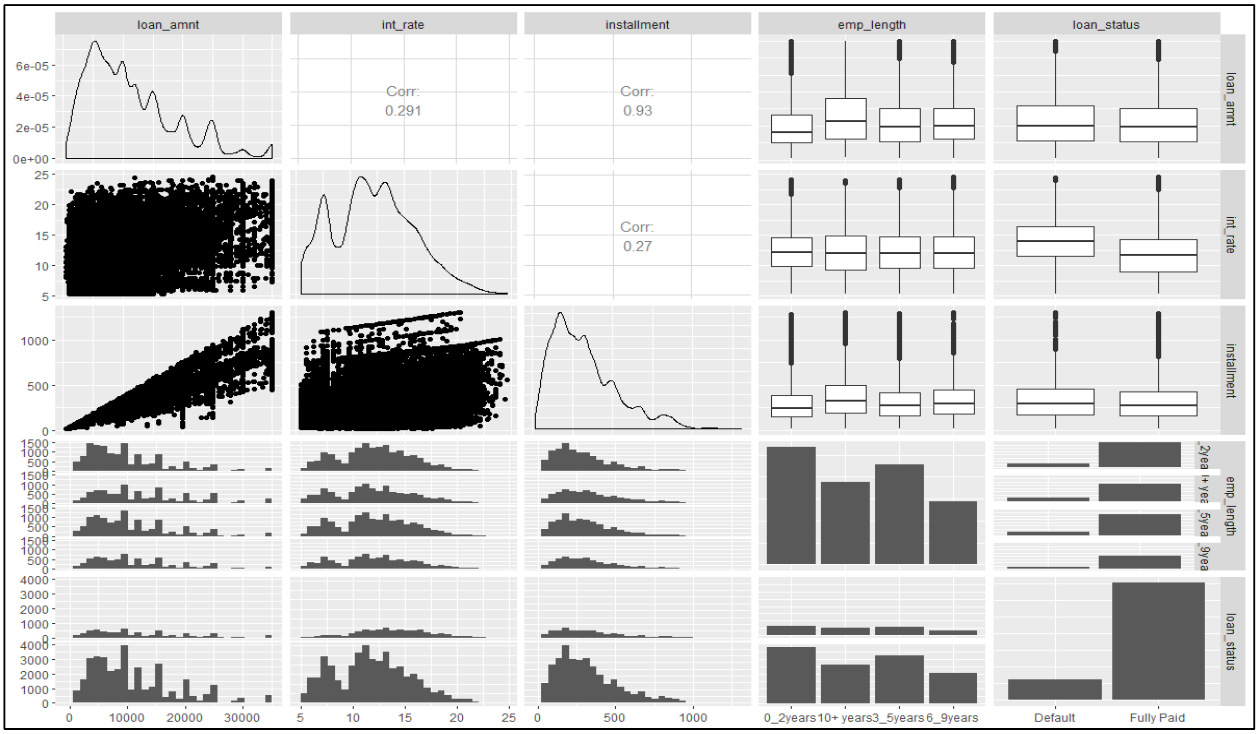
1. Nikolaidis D., Doumpos M., Zopounidis C. (2017) Exploring Population Drift on Consumer Credit Behavioral Scoring. In: Grigoroudis E., Doumpos M. (eds) Operational Research in Business and Economics. Springer Proceedings in Business and Economics. Springer, Cham
2. Thomas et al (2003)
3. Yu et al. 2008a, b
4. Sousa et al. 2013
5. Verikas et al. 2010
6. Caselli, S., Gatti, S. & Querci, F. J Finan Serv Res (2008) 34: 1. doi:10.1007/s10693-008-0033-8
7. Jain, Duin, & Mao, 2000; Nelson, Runger, & Si, 2003
8. IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW; Dr. D Ramyachitra et.al
9. A Comparative Study of Classification Techniques for Intrusion Detection' (H.Chauhan et.al)
10. Benchmarking state-of-the-art classification algorithms for creditscoring: A ten-year update (Stefan Lessmann et.al)
11. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.

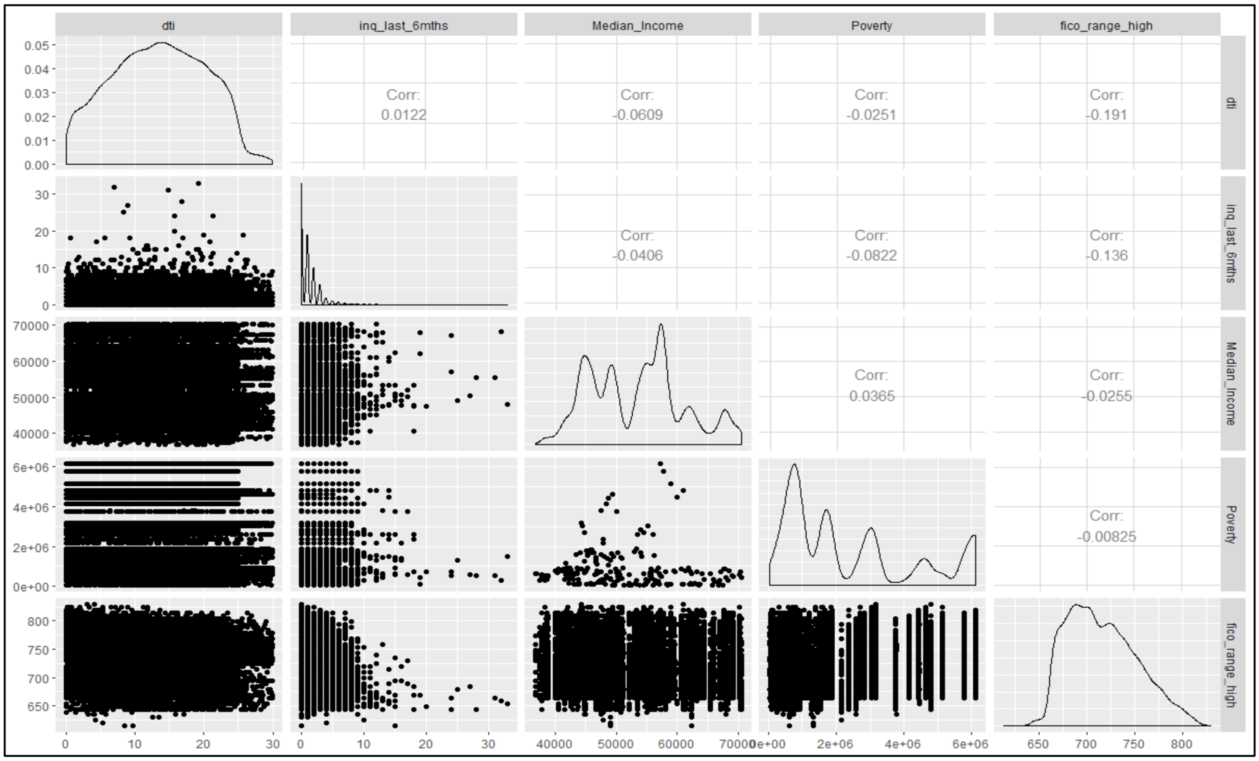
12. Maria Stepanova, Lyn Thomas, (2002) Survival Analysis Methods for Personal Loan Data. Operations Research 50(2):277-289
13. Statistical Classi®cation Methods in Consumer Credit Scoring: a Review by D.J Hand and W.E Henley
14. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont
15. Hastie et al., 2009
16. Breiman 1996 Machine Learning, 24(2), 123-140.
17. Hastie, Tibshirani, & Jerome, 2009
18. David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 128 &
19. Cox, DR (1958). "The regression analysis of binary sequences (with discussion)". J Roy Stat Soc B. 20: 215–242
20. Hugh Chipman - The Annals of Applied Statistics 2010, Vol. 4, No. 1, 266–29
21. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS – R.A Fischer
22. Two-Dimensional Linear Discriminant Analysis – Jieping Ye
23. Bayesian Quadratic Discriminant Analysis – Santosh Srivastava
24. A Summary of the Primary Causes of the Housing Bubble and the Resulting Credit Crisis: A Non-Technical Paper – Jeff Holt
25. M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in Proceedings of the Fourteenth International Conference on Machine Learning. Morgan Kaufmann, 1997, pp. 179- 186.

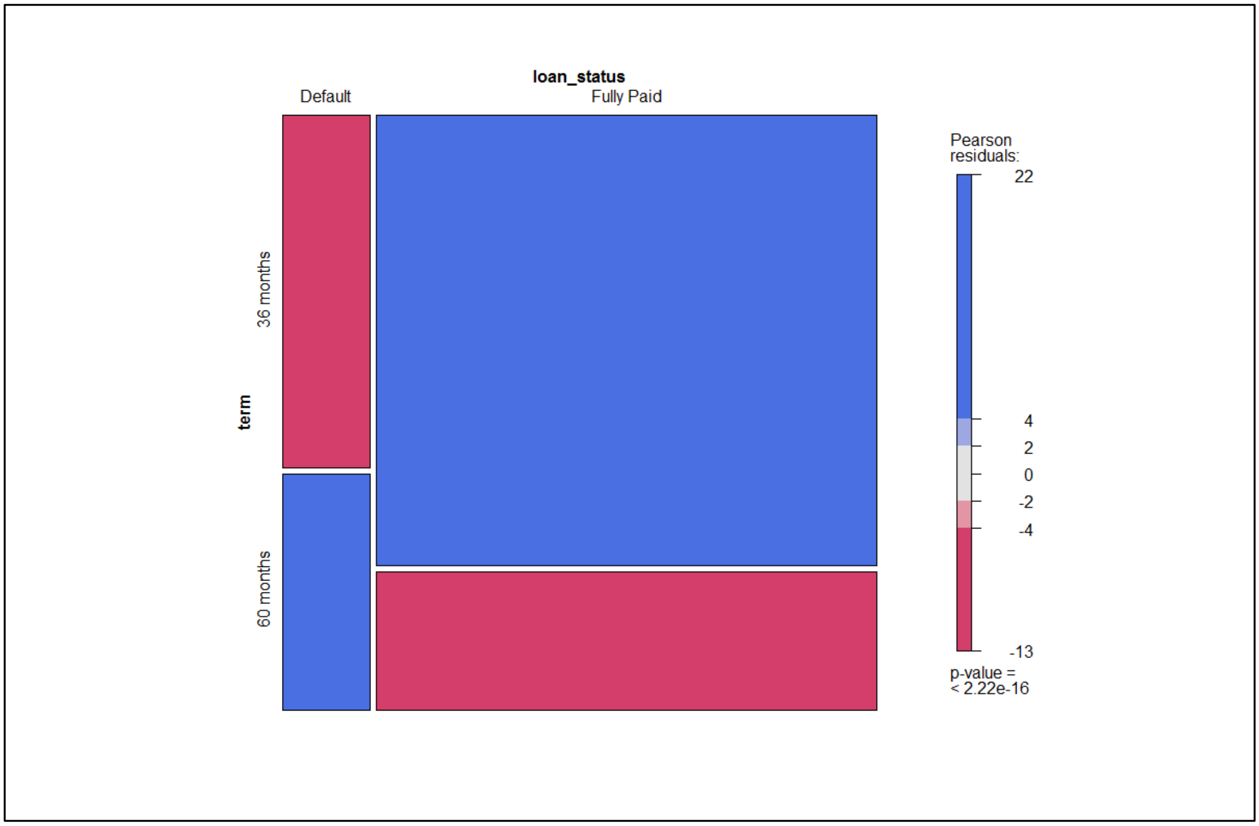
13.Appendix

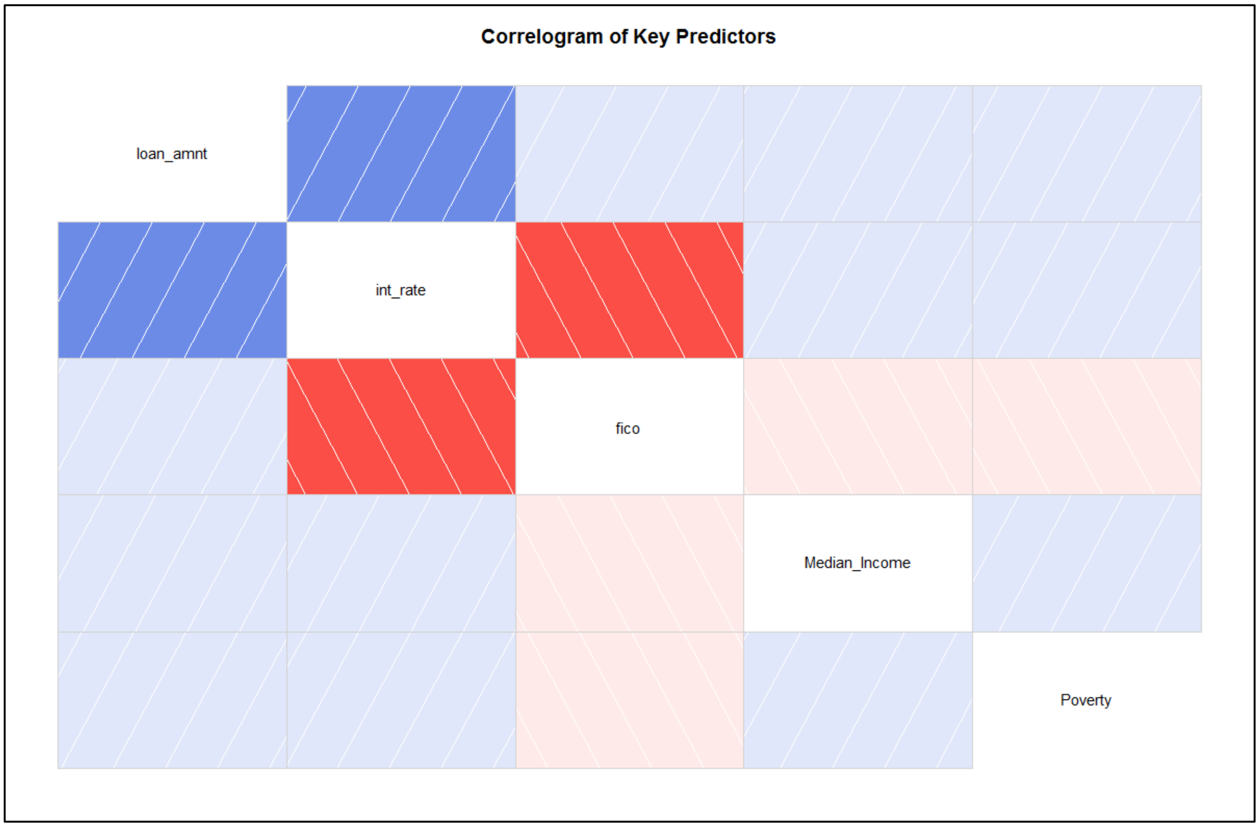
EDA

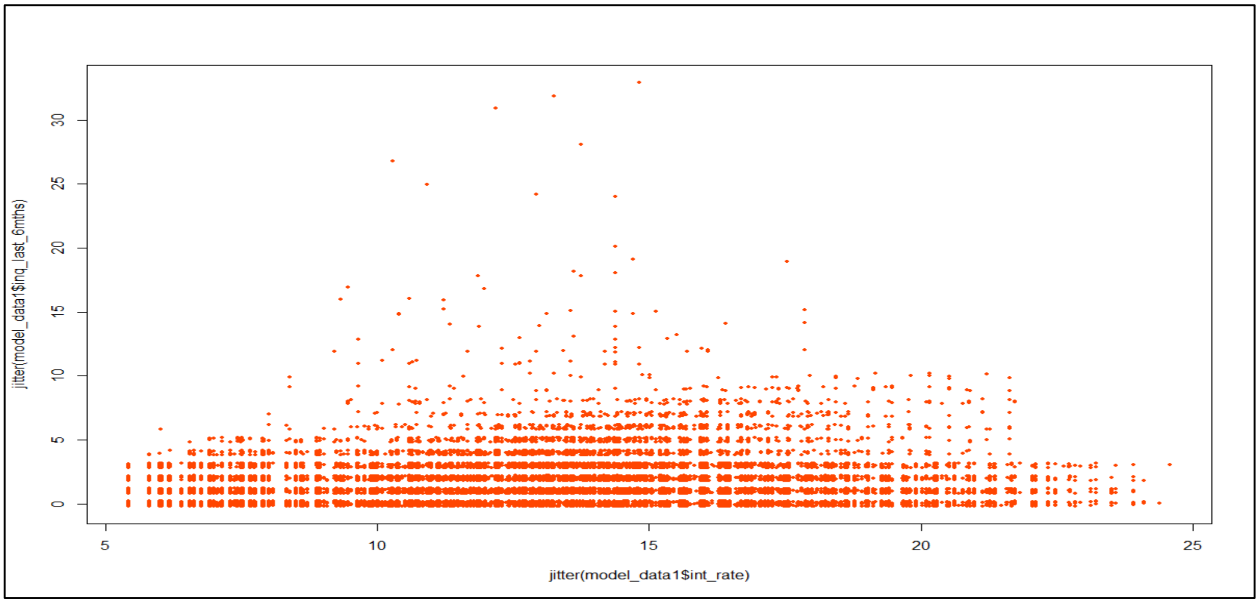


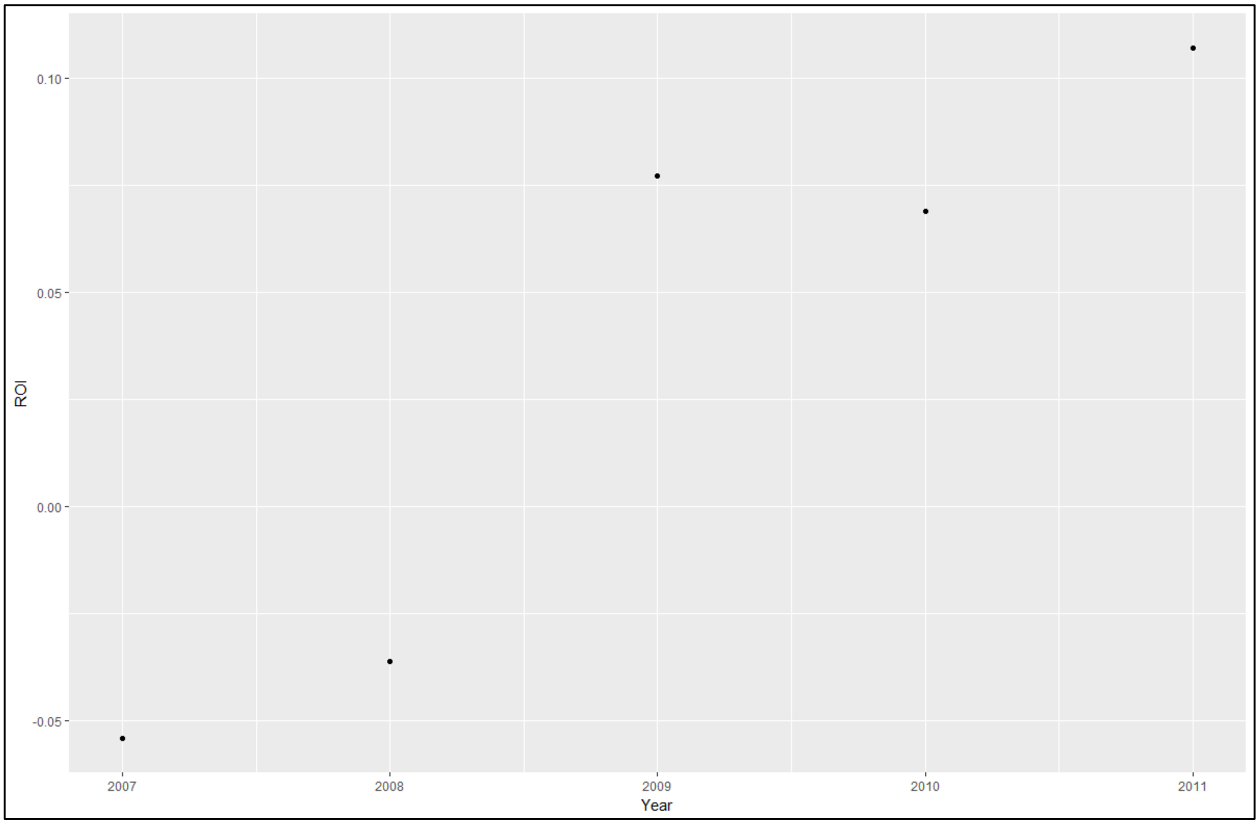


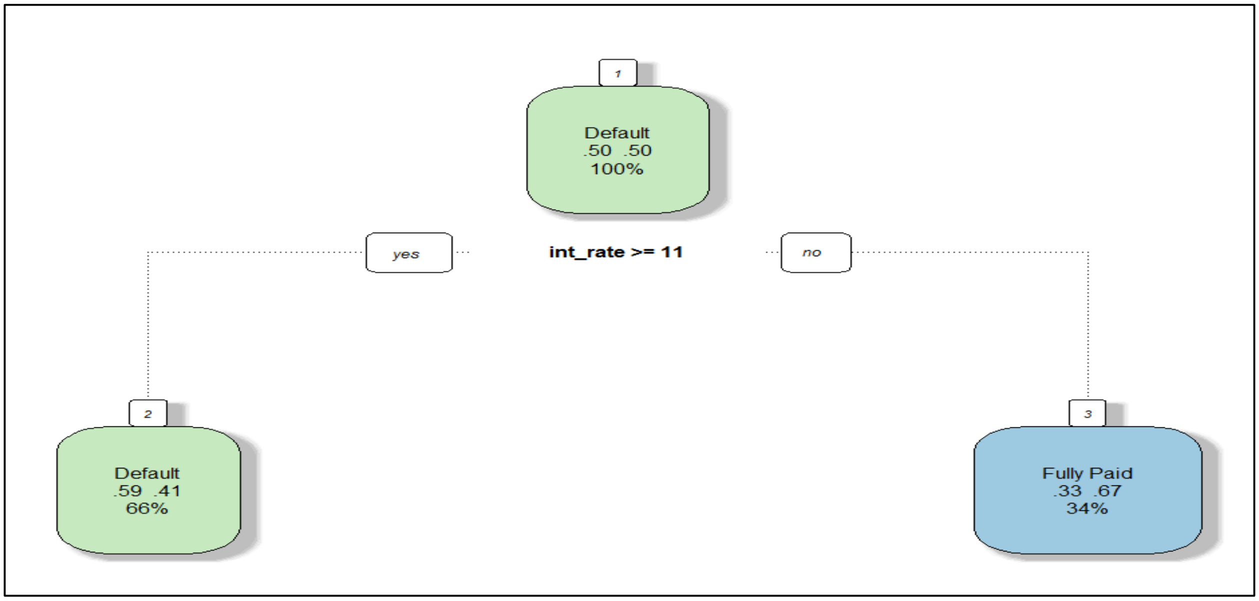

dti
inq_last_6mths
Median_Income
Poverty
fico_range_high
Corr:
0.0122
Corr:
-0.0609
Corr:
-0.0251
Corr:
-0.191
Corr:
-0.0406
Corr:
-0.0822
Corr:
-0.136
Corr:
0.0365
Corr:
-0.0255
Corr:
-0.00825
0.05
0.04
0.03
0.02
0.01
0.00
30
20
10
0
70000
60000
50000
40000
6e+06
4e+06
2e+06
0e+00
800
750
700
650
0
10
20
30
0
10
20
30
40000
50000
60000
70000
0e+00
2e+06
4e+06
6e+06
650
700
750
800


loan_status
Default
Fully Paid
term
36 months
60 months
Pearson
residuals:
22
4
2
0
-2
-4
-13
p-value =
< 2.22e-16


Correlogram of Key Predictors
loan_amnt
int_rate
fico
Median_Income
Poverty


jitter(model_data1$inq_last_6mths)
0
5
10
15
20
25
30
5
10
15
20
25
jitter(model_data1$int_rate)


0.10
0.05
0.00
-0.05
ROI
2007
2008
2009
2010
2011
Year


1
Default
.50 .50
100%
yes
int_rate >= 11
no
2
Default
.59 .41
66%
3
Fully Paid
.33 .67
34%