**Lending Club Loan Default Prediction**

**1. Overview**

* Lending Club is a peer-to-peer lending company, headquartered in San Francisco, California.
* It is platform which offers loan trading where investor lends the money to the borrower.
* Both Investor and Borrower need to register with Lending Club.
* Purpose of loan can be for variety of reasons such as education, wedding, mortgage and debt consolidation.
* Lending Club enables borrowers to create unsecured personal loans between $1,000 and $40,000.The standard loan period is three years.
* Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose.
* Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee.

**2. Project Background**

* Investor here is any financial firm which lends money.
* Borrower here can be any individual, a small business company, etc.
* Borrower has multiple data associated with his/her profile.

For example- Employment data (like Current Company, Employment length, Company grade, Annual earning), Loan Amount requirement, Purpose of loan, Address, Home Ownership (like Mortgage, Rent, Own).

* There is multiple loan status. Below are the status in chronological order-
  + Current
  + Issued
  + In-Grace Period
  + Late (16-30 days)
  + Late (31-120 days)
  + Default
  + Charged Off
  + Fully Paid
* The status of a loan will change occasionally as borrower makes payments or is unable to do so.
* The loans under categories of ‘In Grace Period’, ‘Late (16-30 days)’, and ‘Late (31-120) days’ are technically not NPAs and present an opportunity for the lending institution to levy late payment charges above and beyond the principal and interest. These relatively less serious categories of loans actually bring in additional revenue for the institution and the lending institution may tolerate these. In contrast, the categories ‘Default’ and ‘Charged Off’ are a money lenders nightmare because they directly translate into NPAs. The lending institution tries to avoid such loans. The reality of money lending business is that any ‘Current’ or ‘Issued’ loan can get delayed or defaulted.

**Problem Statement**

* **Objective 1**- We as an investor are predicting, if a new borrower will default the loan or not. It’s a classic binary classification problem.
* **Objective 2**- We as an investor are predicting the similarities of Current Status loan with risky loan status. It’s achieved through comparative technique.

**3. Analysis**

**Business Understanding**

* About Lending club
* Lending Club is a US peer-to-peer lending company.
* Platform of lending club
* Entities associated in business (Investor and borrower)
* Provides a more flexible platform compared to other individual bodies.
* Terms associated in loan business (like debt to income ratio, delinq, revolve balance, multiple loan status)

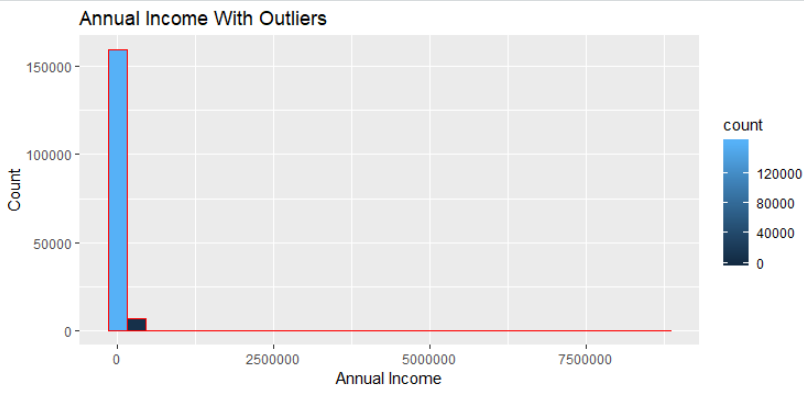
**Data Understanding**

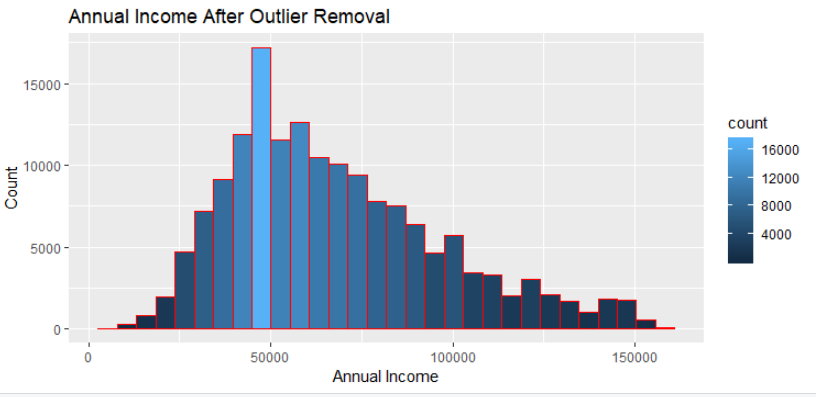
* **Loan Amount -** The listed amount of the loan applied for by the borrower.
* **Term -** The number of payments on the loan. Values are in months and can be either 36 or 60.
* **Interest Rate** **-** Interest Rate on the loan
* **Installment/EMI -** The monthly payment owed by the borrower if the loan originates.
* **Grade** - Lending Club assigned grade.
* **Employment Length -** 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 -** The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE.
* **Annual Income** - The self-reported annual income provided by the borrower during registration.
* **Loan Status -** Current status of the loan.
* **Purpose -** A category provided by the borrower for the loan request.
* **Address (state) -** The state provided by the borrower in the loan application
* **DTI -** 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 -** The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
* **Inq\_last\_6mths** - The number of inquiries in past 6 months (excluding auto and mortgage inquiries).
* **Open Account** - The number of open credit lines in the borrower's credit file.
* **Public Record** - Number of derogatory public records.
* **Revolve Balance -** Total credit revolving balance.
* **Revolve Utilization -** Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.

**Data Preparation**

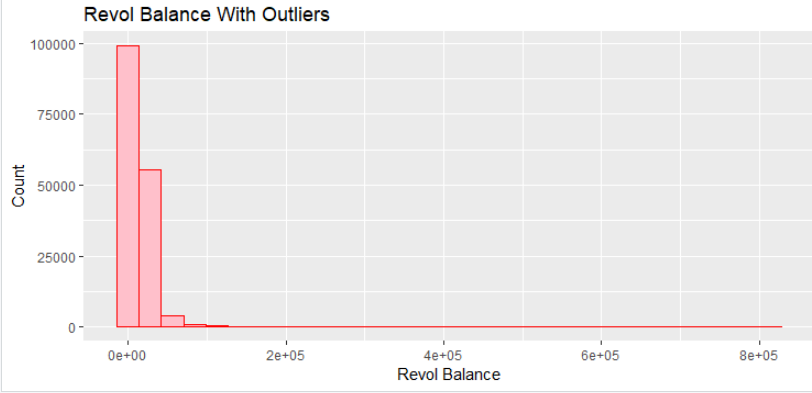
1. **Creating a new column Default**: From the possible loan status we considered only “Fully Paid”, “Default” and “Charged Off” status for imputing new column Default. We ignored other status since we didn’t want the model to learn data we aren’t sure if they should be classified as default or not.
2. **Removing redundant levels:** We could see that purpose column has 14 levels and not all of them were significant. Hence, we clubbed the rare levels into a generic level “other” and considered only 5 levels for purpose. This would also help us to reduce the number of columns generated during One Hot Encoding.
3. **Bucketing Employment Length:** We created bins of Employment Length for better visualization of data.
4. **Converting Grade to Numeric:** In Model building part, we converted Ordinal Data Grade to Numeric starting from grade A having the highest value to G being the lowest. This again helped us to reduce the number of columns imputed during One Hot Encoding.
5. **Imputing NAs:** We used mice to impute NA values in employment length and revolving utilization columns instead of using the Mean or the Median imputation technique.
6. **Feature Selection:** Out of 120 columns we used only 30 columns for our analysis based on domain knowledge. We went through a thorough understanding of the features and took about 3 days to select relevant features.
7. **Removing special characters:** We removed special characters >,+ from employment length.
8. **Outlier Removal**: We could clearly see how outliers were impacting our understanding process. Below are the graphs of before and after outlier removal.

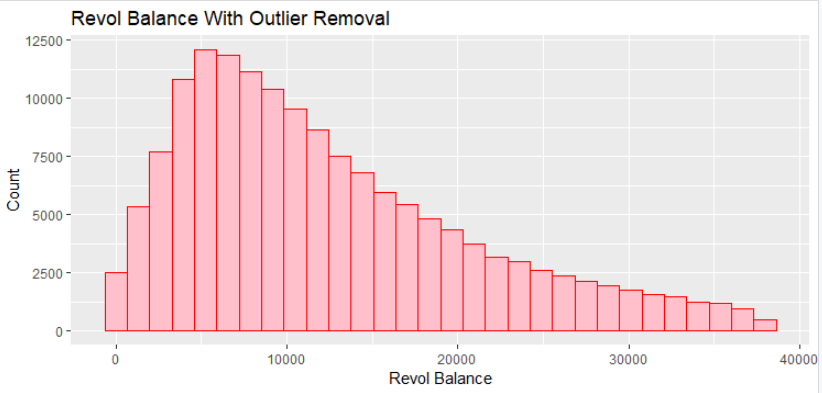
7a. Annual income



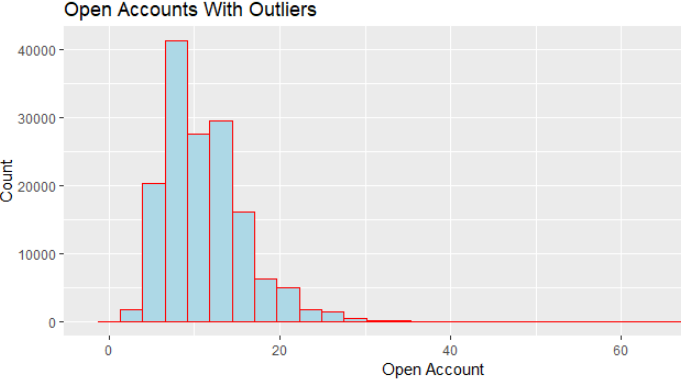


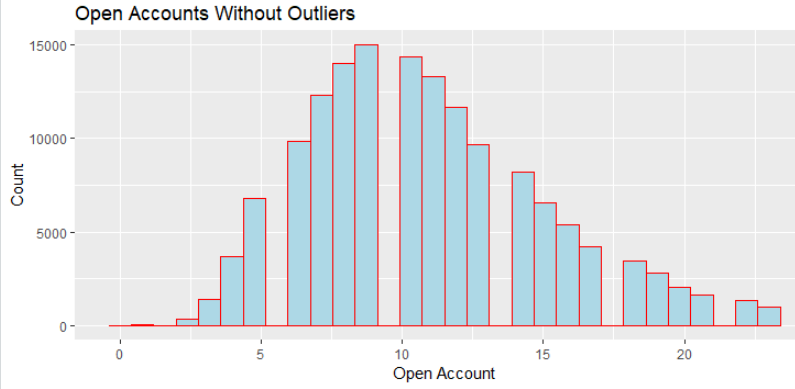
7b. Revolve Balance



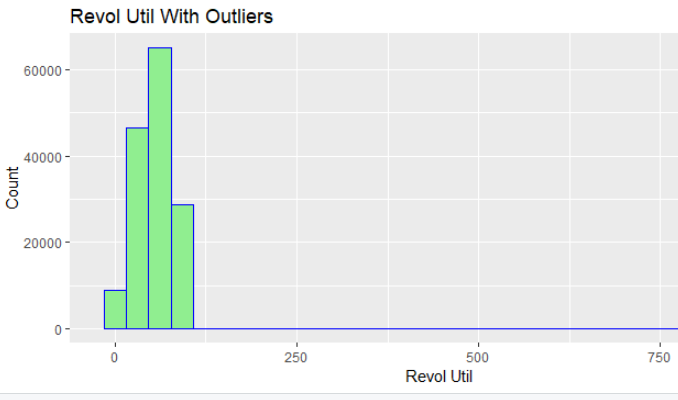


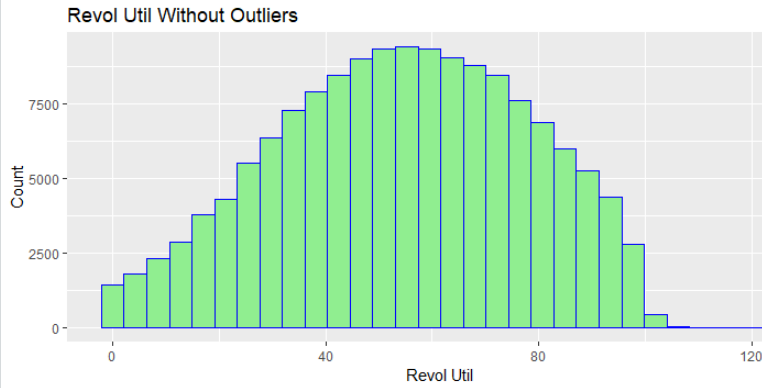
7c. Open Accounts





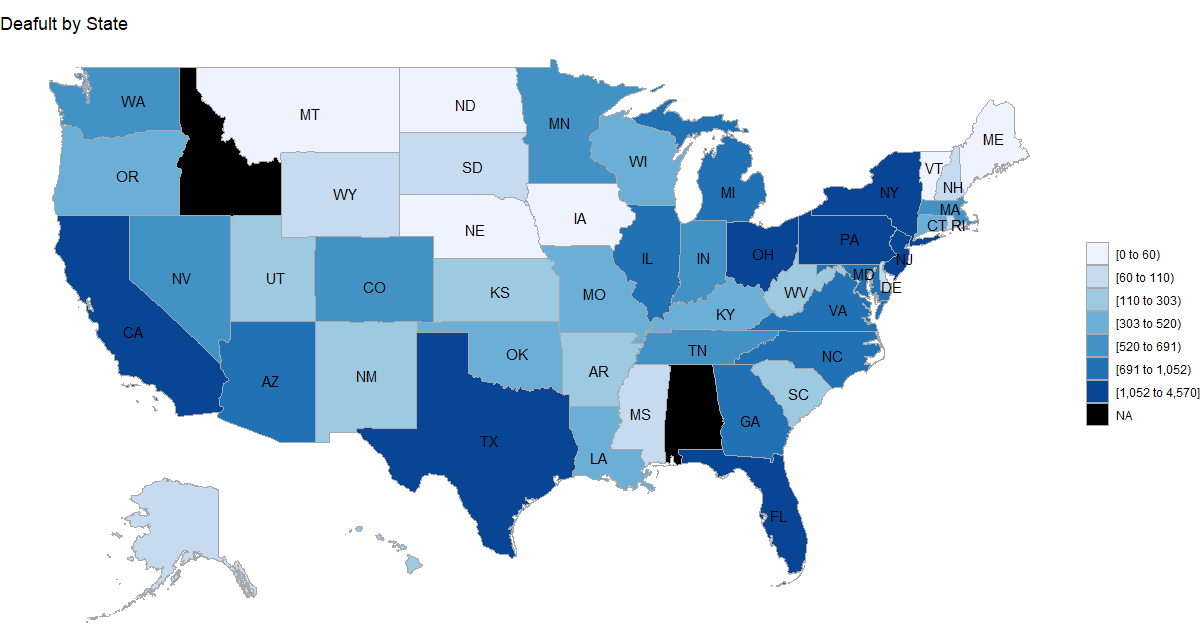
7d. Revolve Utilization





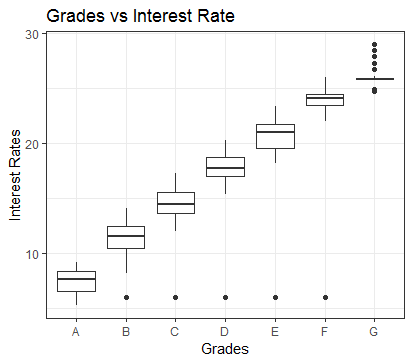
**Exploratory Data Analysis**

1. **State wise default**



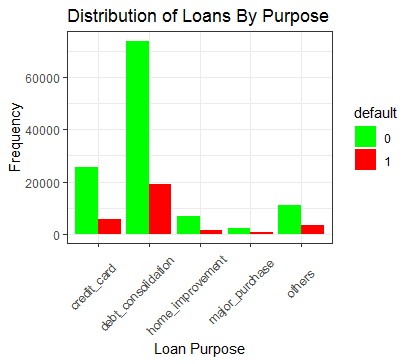
**Inference**: We could see that Texas, California, Pennsylvania, New York and Ohio states had the highest numbers of defaulters.

1. **Grade v/s Interest**



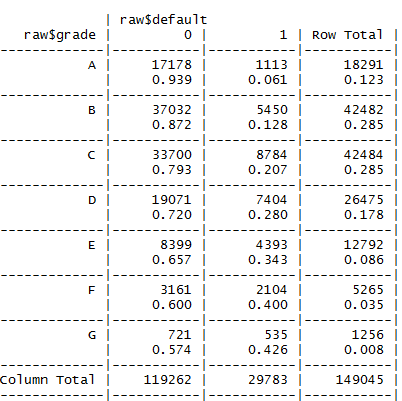
**Inference -** Another important observation to check validity of dataset is to examine relation between interest rates and grade scores. Better grades should have lower interest rates, while poorer grades should have higher interest rates. The best way to control this relation is by making a scatter diagram shows clearly that interest rates get higher as grade loan gets poorer.

1. **Purpose v/s Default**



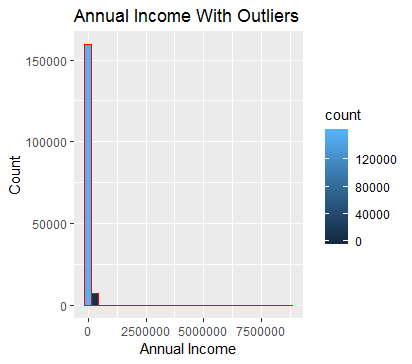
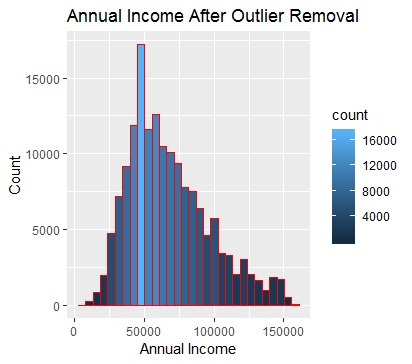
**Inference -** Debt consolidation means taking out a new loan to pay off a number of liabilities and consumer debts. Debt consolidation is the highest purpose of loan among the borrower.

1. **Grade v/s Default**



**Inference -** There are hardly any defaulters in grade ‘A’ whereas it’s a total different scenario in grade ‘G’. We conclude that loans grading is correct and as grade score quality decreases the percentage of defaults increases.

1. **Annual Income outlier handling**

**Inference-**

***With Outlier***

There are two main methods how we can handle these outlier. The first method is using the rule of thumb where upper limit is computed as 1.5 \* IRQ, where IRQ = 3rd Quartile – 1st Quartile. Therefore, upper limit for outliers would be $ 157,500 USD.

The second method is using expert judgment. According to the official statistics, the median income of US household is $ 55,775 USD. Therefore, in this case all annual incomes which are larger than $ 1,000,000 USD I will consider as outliers and will remove from dataset. There are 91 observations that meets this criteria and which were deleted.

***Without Outlier***

Distribution of annual income after upper outliers were removed. We can see that distribution is not normal and it is skewed to the right with most observations that fall within range between 50,000 and 100,000. This make sense since annual income usually follows this shape (only small percentage of population have extremely high income), while majority of population falls within range which is in line with official statistics. Therefore, we can conclude that there is no unusual behavior in this feature.

**4. Methodology**

**Handling Imbalanced Data**

Our data is highly skewed. We can see only approximately 20% default cases in our dataset. We used the following Resampling techniques for balancing our data.

1. **Over Sampling**

Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample.

* Advantages
  + Unlike under sampling this method leads to no information loss.
  + Outperforms under sampling
* Disadvantages
  + It increases the likelihood of overfitting since it replicates the minority class events.

1. **Random Under Sampling**

Random Under sampling aims to balance class distribution by randomly eliminating majority class examples.  This is done until the majority and minority class instances are balanced out.

* Advantages
* It can help improve run time and storage problems by reducing the number of training data samples when the training data set is huge.
* Disadvantages
* It can discard potentially useful information which could be important for building rule classifiers.
* The sample chosen by random under sampling may be a biased sample. And it will not be an accurate representative of the population. Thereby, resulting in inaccurate results with the actual test data set.

1. **Both (Using Under and Over Sampling)**

Using a combination of both under and over sampling giving 0.5 probability to each.

1. **Synthetic Sampling**

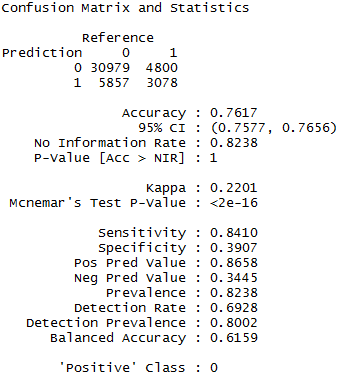
This technique is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created.

**Results of Resampling**

Following were the results for each of the resampled technique using logistic regression.

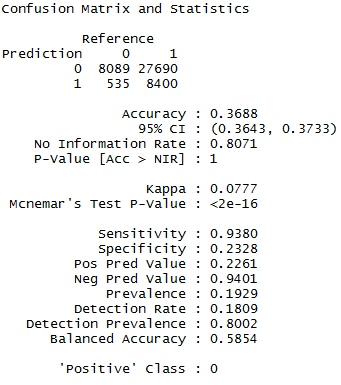
* **Logistic Regression (With Under Sampling)**

Below is the evaluation metric of logistic regression with under sampling technique: -



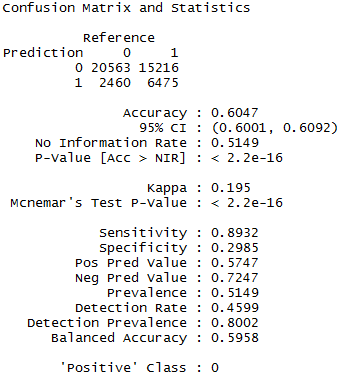
* **Logistic Regression (With Over Sampling)**

Below is the evaluation metric of logistic regression with over sampling technique: -



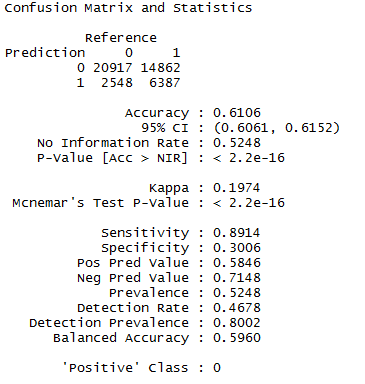
* **Logistic Regression (With Both (i.e. Under and Over) Sampling)**

Below is the evaluation metric of logistic regression with both (i.e. Under and Over) sampling technique: -

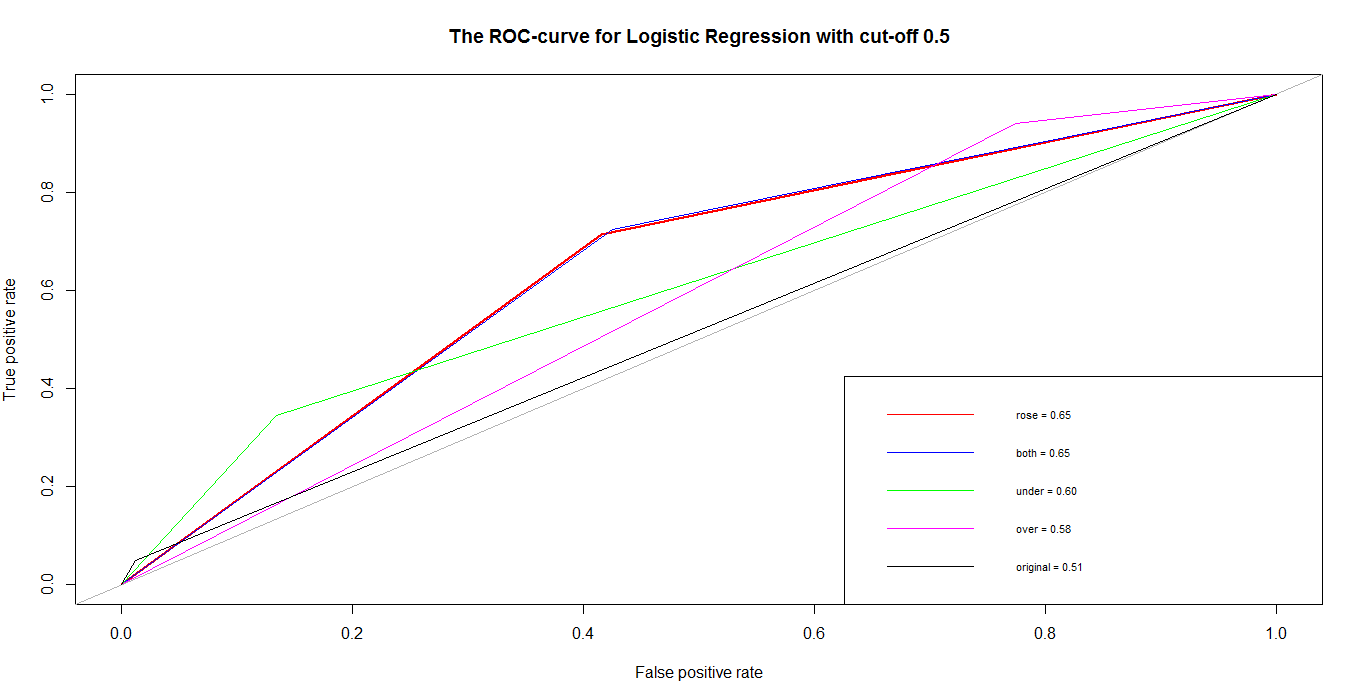


* **Logistic Regression (With Synthetic Sampling)**

Below is the evaluation metric of logistic regression with synthetic sampling technique: -



* **ROC Curve with multiple sampling technique**



We could see that the highest AUC we got was for sampling using ROSE and Both(under and over sampling) that is 0.65.

**Algorithms and Implementation**

**1. Logistic Regression.**

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. It is a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

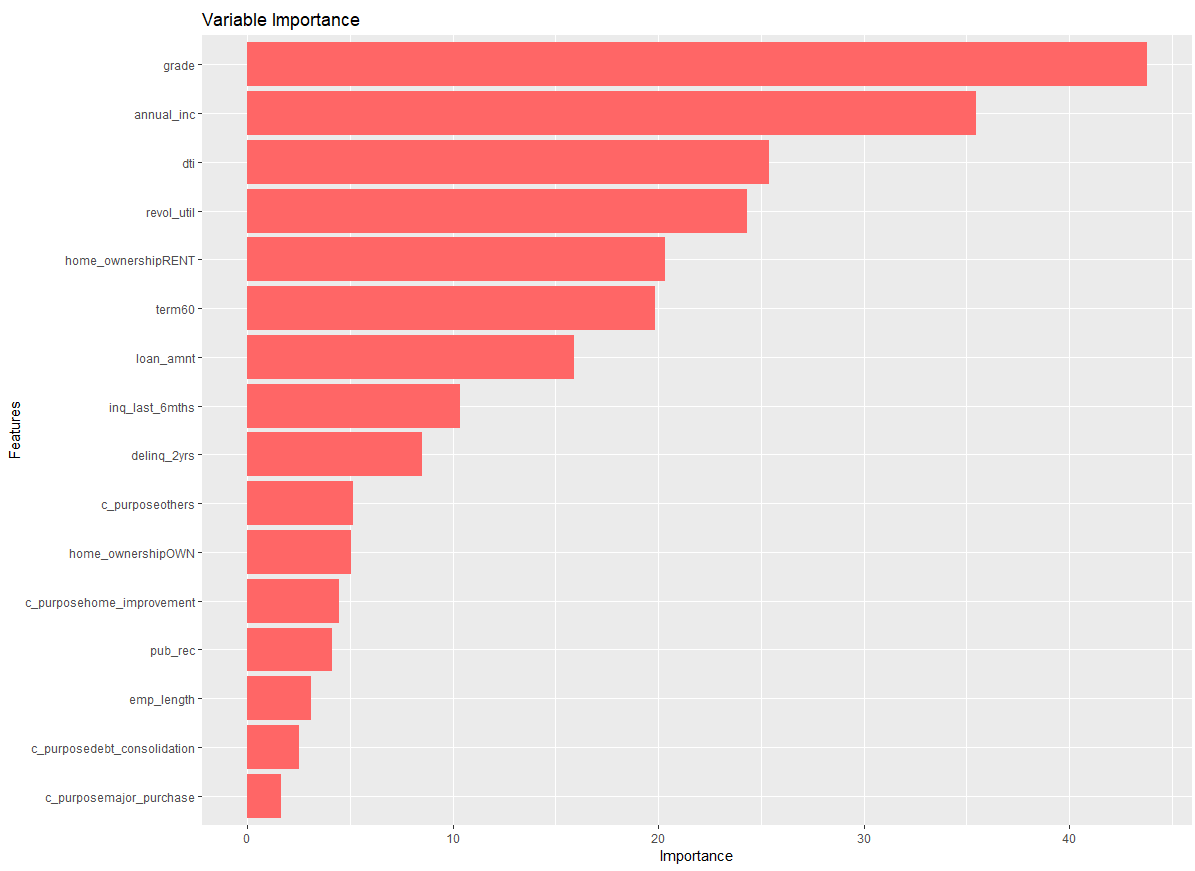
It does not use OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE). Errors need to be independent but not normally distributed.

Advantages:

* Gives the probability of a person to default.
* Allows to adjust threshold for classifying person as default
* Easier to interpret. Depending on the variable importance and the coefficients we can interpret the results. This is essential since Bank has to justify reasons for classifying a person as default.

After building a logistic regression model results were as follows.

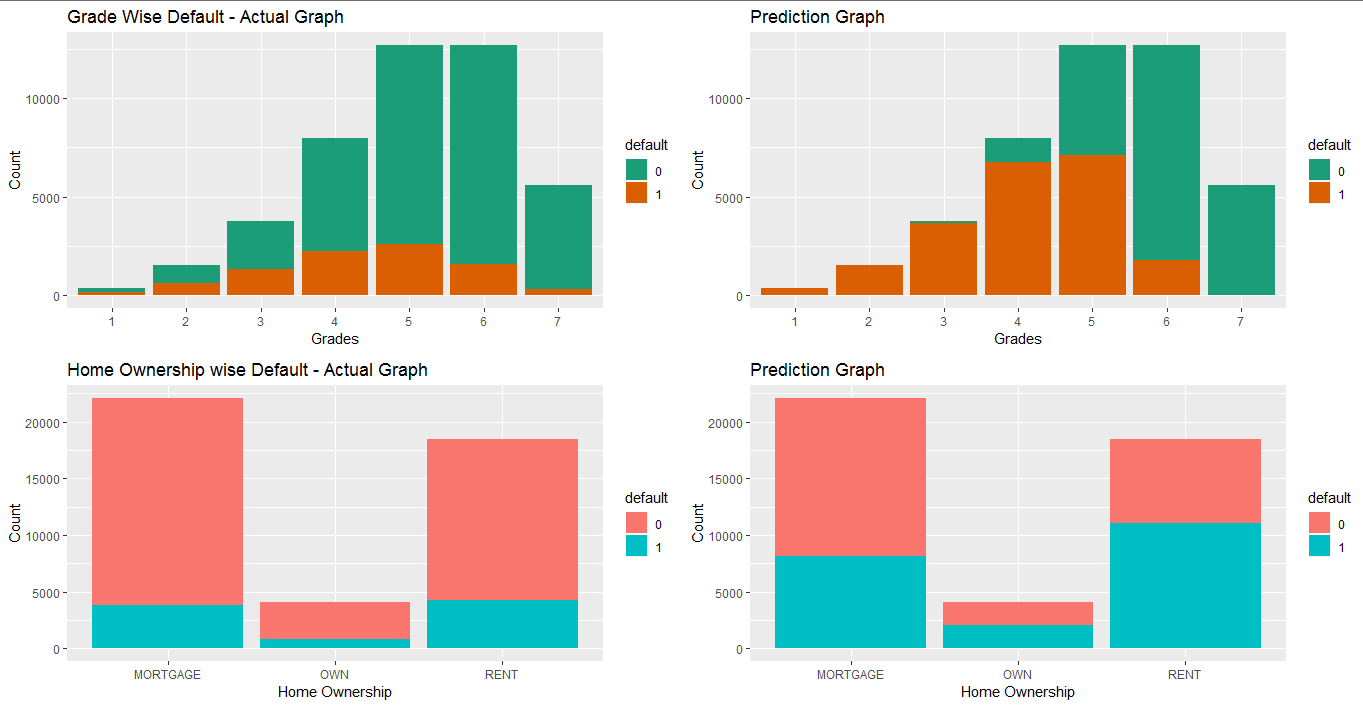
**Variable Importance Graph**



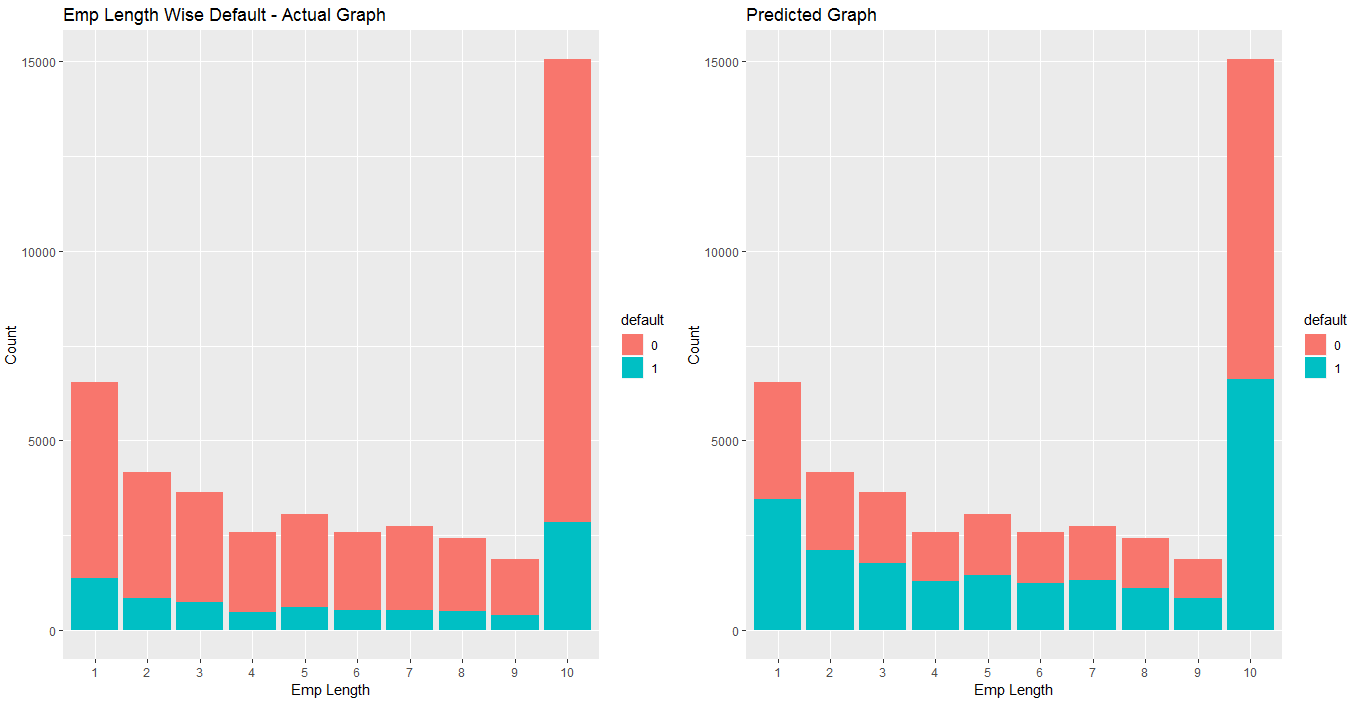
We can see that Grade, Annual Income, DTI, Revolving Utilization and Term were highly important in prediction of default.

**Comparing actual vs precited graph for this model.**

* **Grade and Homeownership Prediction Frequency**



* **Employment Length Prediction Frequency**



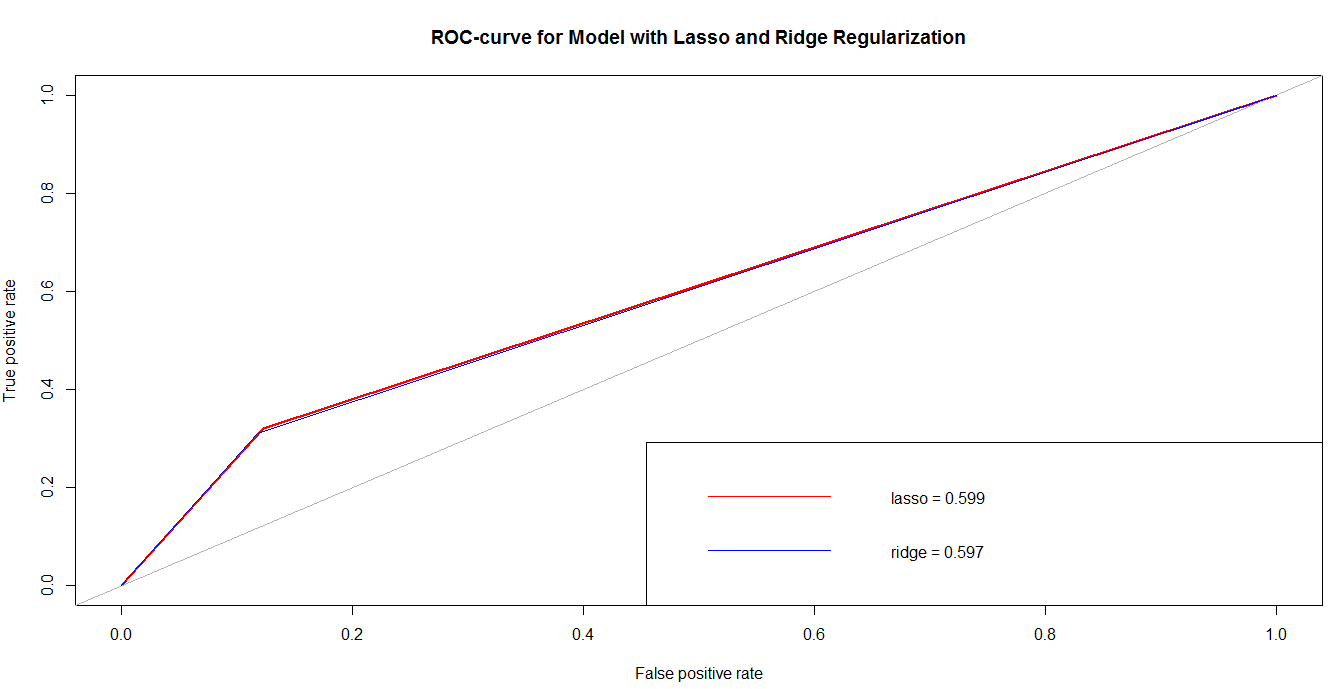
**Inferences-**

We have built a more conservative model as we do not want to provide loans to any risky customers who will eventually default and increase the non-performing asset (NPA) percent of a company.

**2. Lasso and Ridge Regularization**

We also tried L1 and L2 regularization techniques to see if it increased the AUC but we could see regularization didn’t help much as we had already very wisely chosen the features and there were no redundant features in the model.

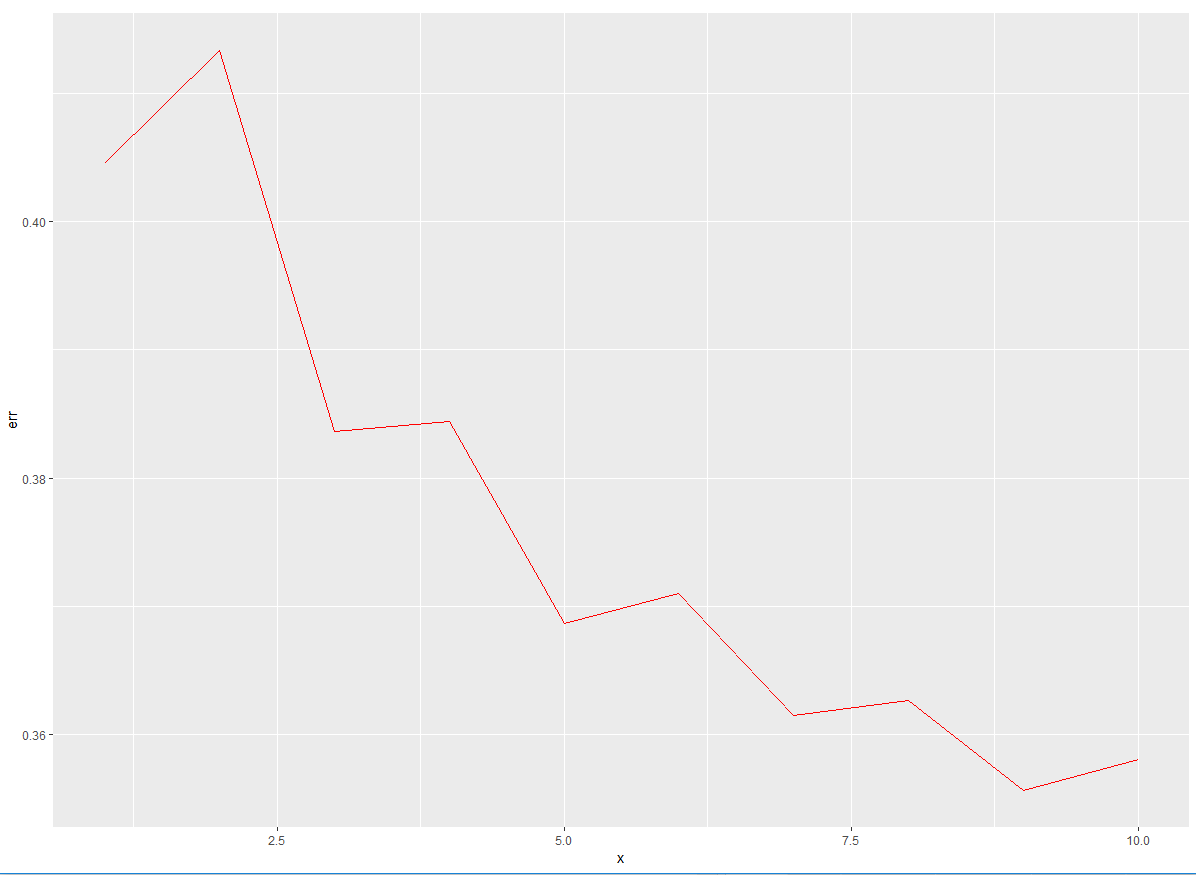
* **ROC Curve with Lasso and Ridge technique**



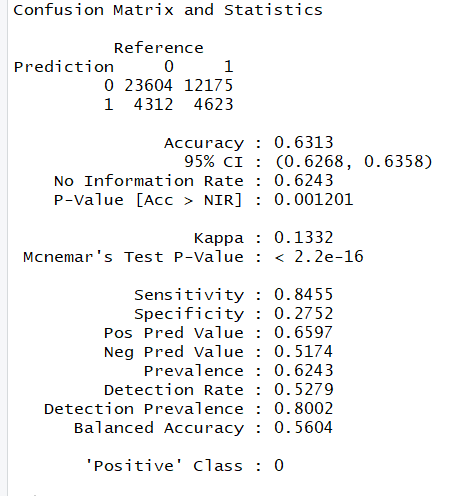
**3. K-Nearest Neighbor**

KNN is a non-parametric and instance-based algorithm. It is highly versatile and robust classifier. the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the K most similar instances to a given “unseen” observation. Similarity is defined according to a distance metric between two data points.

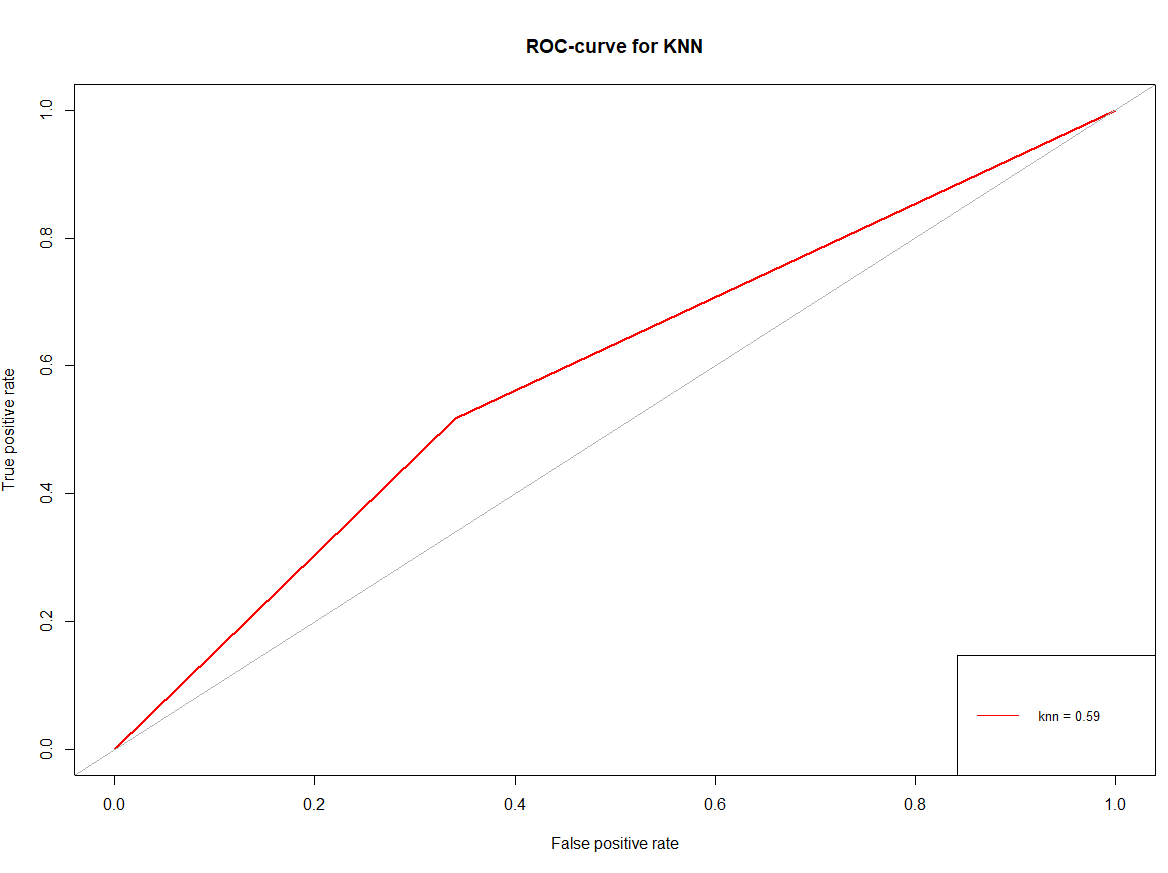
**Finding the optimal value of K**



**Results of KNN**

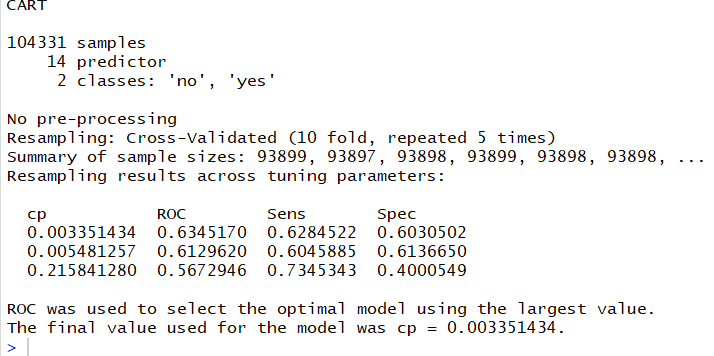


ROC Curve

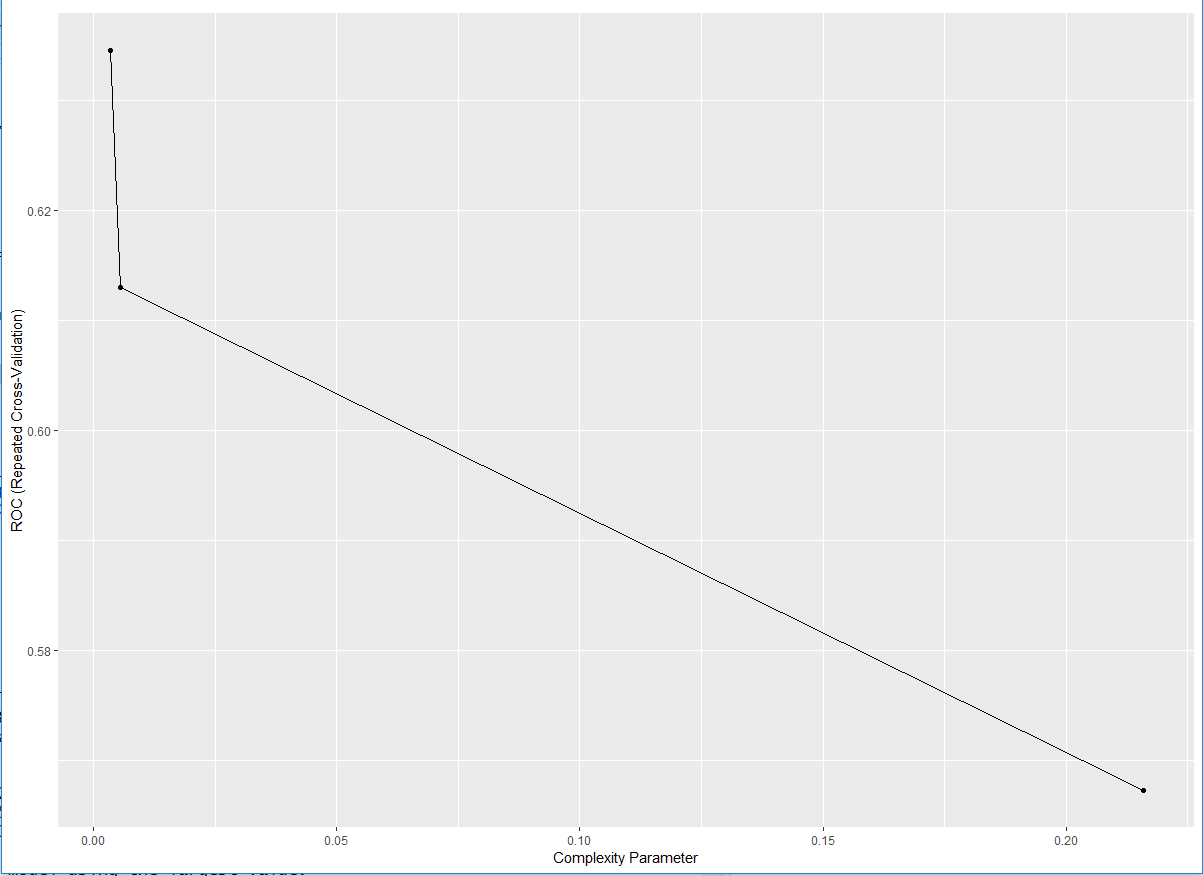


**4. Decision Trees**

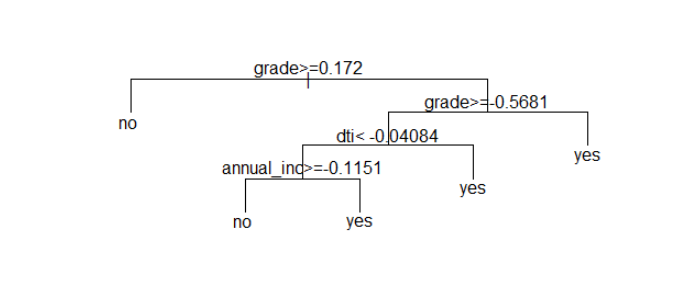
We used tree-based approach for our predictive analysis by using rpart from the caret library. The results were as follows.



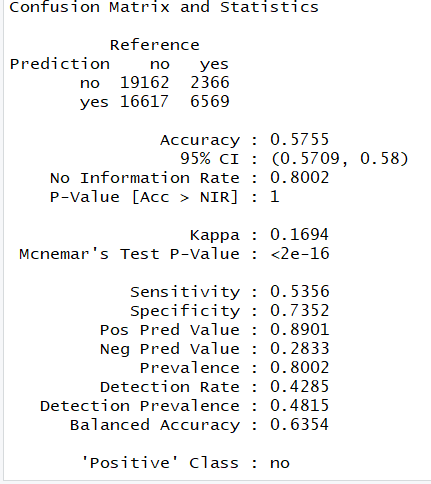
**Complexity parameter curve**



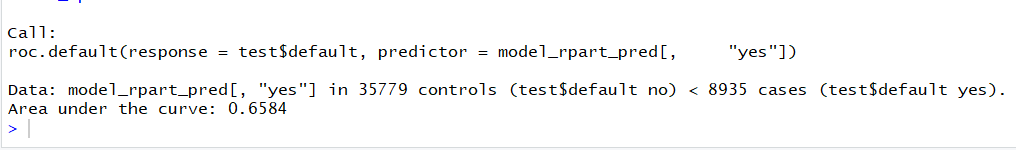
**Printing the Tree**



**Confusion Matrix**



**ROC**



**Evaluation Metric**

**ROC Curve:**

Receiver Operating Characteristic (ROC) summarizes the model’s performance by evaluating the trade-off between true positive rate (sensitivity) and false positive rate (1- specificity).  The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model.

**Sensitivity**

Sensitivity (also called the true positive rate, or the recall in some fields) measures the proportion of actual positives which are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition), and is complementary to the false negative rate. Sensitivity= true positives/ (true positive + false negative)

**Specificity**

Specificity (also called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate. Specificity=true negatives/ (true negative + false positives)

**Accuracy**

It’s the ratio of correctly classified records over the total records. But it can be biased if the ratio of one class is higher than the other.

**Model Selection – (objective 1)**

We could see the best results we got were from Logistic Regression Model. It is simple and easy to interpret. Hence, we selected Logistic regression model as our final model. We could see from the variable importance graph that Grade and Annual Income were highly influencing the prediction of a person being a default. We gave preference to ROC, Sensitivity and Specificity instead of accuracy. We could see that some models ended up giving accuracy of 0.80 but sensitivity, specificity and roc got worse.

**Challenges**

Understanding the domain (features)- This took a long time. We had to go through a lot of documentation for understanding the data.

Parameter tuning- We tried parameter tuning for decision trees by using grid search instead of the default search but the results were not impressive. We tried various variations of glm (by using L1, L2) Regularization techniques but they didn’t outstand the results of simple glm. Simpler the model, better the results.

**Model Building (Objective 2)**

***Identifying Potentially Bad Loans:***

The dataset from Lending Club contains 4,20,853 loans for year 2015. It is not time series data where payment history is recorded over the life-time of each loan. It is a snapshot of all loans on the book. About 2% loans are ‘Issued’, 89.7% are ‘Current’ and 5.5% are ‘Fully Paid’. About 0.7% are ‘In Grace Period’, 0.3% are ‘Late (16-30) days’ and 1% are ‘Late (31-120 days)’, 0.1% are ‘Default’ and 0.7% are ‘Charged Off’. The negative loan statuses make up 2.8% of the sample.

Bad loans can be identified from the pool of ‘Current’ or ‘Issued’ loans by seeing how similar they are to ‘Default’, ‘Charged Off’ or other late loans. There are numeric and categorical variables in the dataset and therefore numeric distance methods like euclidean method do not work. The similarity is computed using the ‘Gower’ distance method using the ‘kmed’ R-package. There are additional methods such as ‘huang’ and ‘wishart’ available in the kmed package. The gower distance ranges from 0 to 1. A dissimilarity (distance) matrix is computed which shows distance between any pair of loans.

The R function selectSimilar() select loans from the pool of ‘Current’ or ‘Issued’ loans that at most 15% dissimilar to, say, ‘Default’ loans.

> def = selectSimilar(dist.gower,target='Default', data=small, distance=0.15)

> def[order(def$distance, def$similar),]

curr\_issue similar distance

10 377064 397587 0.1302409

13 105260 397587 0.1307428

9 227034 397587 0.1308821

11 343664 366994 0.1311406

17 320323 366994 0.1324361

4 222569 397587 0.1325747

In the output, the curr\_issue column contains row labels of loans which are ‘Current’ or ‘Issued’. The similar column contains row labels of ‘Default’ loans. The distance column shows the distance between the loans.



***Selling-off Potentially Bad Loans:***

A portfolio manager can look at parameters of potentially bad loans such as its grade and decide that they are not worthy of being kept around. Such potentially bad loans can be sold off to other buyers who are interested. This kind of transactions often happen in the business of mortgage-backed securities (asset-backed securities, in general).