

**P2P LENDING RISK ANALYSIS**

**Submitted towards partial fulfillment of the criteria**

**For the award of PGPDSE by Great Lakes Institute of Management**

**Submitted By**

**Group No. 2 [Batch: Apr- 2019]**

**Group Members**

**K.S.V.Samanth**

**Abhishek Banerjee**

**Arpita Bose**

**Junaid Ahamad**

**Pramod Reddy Gaddam**

**Mentor**

**Vidhya K**



**Certificate of Completion**

I hereby certify that the project titled defaulter prediction and interest rate prediction for case resolution was undertaken and completed under my supervision by Arpita Bose, Abhishek Banerjee, Junaid Ahamad, Samanth Koduru and Pramod of Post Graduate Program in Data Science and Engineering (PGP – DSE).

**Vidhya K.**

**Date:02/10/2019**

**Place: Hyderabad**

**Declaration**

I declare that the project entitled **P2P LENDING RISK ANALYSIS** is a project work carried out by me under the supervision and guidance of Vidhya K for the award of degree PGP DSE, and this has not been previously submitted for the award of any Degree, Diploma or other similar title of any other University/ Institute.

Date: 02/10/2019

Place: Hyderabad Group No :- 2

K.S.V.Samanth

Abhishek Banerjee

Arpita Bose

Junaid Ahamad

Pramod Reddy Gaddam

**ACKNOWLEDGEMENT**

In any work there are contributions from all quarters. Since they cannot all be mentioned, I will name the few who have contributed the most.

I would like to acknowledge my supervisor Vidhya K for providing me the necessary guidance and valuable support throughout this research project. I value the assistance of Great Learning, Hyderabad campus. Learning from their knowledge helped me to become passionate about my research topic.

I will be failing in my duty if I don’t express my gratitude for my team members, for their valuable suggestions and cheerful assistance during the implementation phase of this project.

**ABSTRACT**

This documentation shows the application of Regression and Classification for predicting if a customer is likely to default; and given the risk how much interest rate should be levied. This will give fair amount of idea to the investors for decision making. The prediction model is built using historical data from Lending Club for period from 2007 until 2011. The original dataset has 144 variables and 45k+ observations.

The classification and regression models are built on 29 independent variables and 45k+ observations after data cleaning process. The model is tested on test data and results are analysed using metrics from accuracy score, recall and precision in case of classification and r-square and RMSE values for regression. Furthermore, classification models are improved using up-sampling method to overcome the problem of imbalanced classification. Results show that loan prediction models have fairly poor performance mainly because of the imbalanced classification problem, whereas regression models are performing pretty effectively in determining the interest rate. The best performance was shown by Random Forest classifier with hyper parameter tuning in case of predicting defaulters whereas Random Forest regressor performed the best in predicting how much interest rate has to be levied according to the risk associated with the loan to minimize losses.

* Techniques: Predictive Modelling
* Tools: Python, Tableau
* Domain: Financial and Risk Analytics

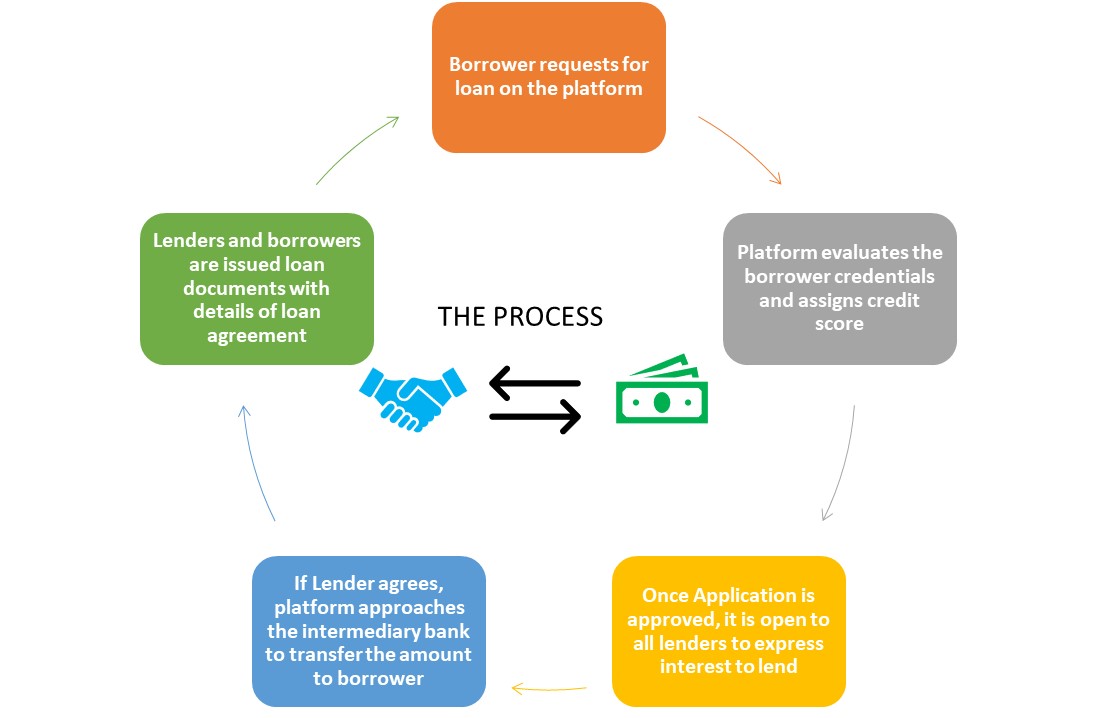
# **INTRODUCTION**

## 1.1 LENDING CLUB WORKING MODEL

LendingClub is a US [peer-to-peer lending](https://en.wikipedia.org/wiki/Peer-to-peer_lending) company, headquartered in [San Francisco, California](https://en.wikipedia.org/wiki/San_Francisco,_California). LendingClub enables borrowers to create loan listings on its website by supplying details about themselves and the loans that they would like to request. All loans are unsecured personal loans and can be between $1,000 - $40,000.

On the basis of the borrower’s [credit score](https://en.wikipedia.org/wiki/Credit_score), credit history, desired loan amount and the borrower’s [debt-to-income ratio](https://en.wikipedia.org/wiki/Debt-to-income_ratio), LendingClub determines whether the borrower is credit worthy and assigns to its approved loans a credit grade that determines payable interest rate and fees. The standard loan period is three years; a five-year period is available at a higher interest rate and additional fees.

Only investors in 39 US states are eligible to purchase notes on the LendingClub platform. Investors make money from interest. Rates vary from ~6% to ~24%, depending on the credit grade assigned to the loan request. The grades assigned to these requests range alphabetically from A to G, with A being the highest-grade, lowest-interest loan. Each of these letter grades has five finer-grain sub-grades, numbered 1 to 5, with 1 being the highest sub-grade.



## 1.2 PROBLEM STATEMENT:

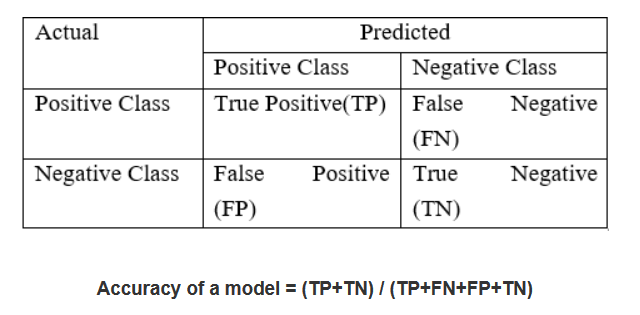
* To predict if the borrower is likely to fully pay the amount or charge off.
* To predict the potential interest rate that can be given to the customer according to the risk associated with the loan.

# **LITERATURE SURVEY**

We have to take a different approach here from a normal machine learning flow because of the nature of our data. The conventional model evaluation methods do not accurately measure model performance when faced with imbalanced datasets.

Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes which have number of instances. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class.

Evaluation of a classification algorithm performance is measured by the Confusion Matrix which contains information about the actual and the predicted class.

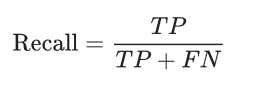
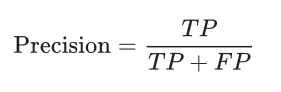


However, while working in an imbalanced domain accuracy is not an appropriate measure to evaluate model performance**.** For e.g.: A classifier which achieves an accuracy of 98 % with an event rate of 2% is not accurate, if it classifies all instances as the majority class. And eliminates the 2% minority class observations as noise.

To fully evaluate the effectiveness of our model, we must examine **precision** and **recall** as well. Unfortunately, precision and recall are often in tension. That is, improving precision typically reduces recall and vice versa.

**Precision:** What proportion of positive identifications was actually correct?

**Recall:** What proportion of actual positives was identified correctly?



Evaluation of a regression algorithm performance is measured by the RMSE and R square:

**R-squared** is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%:

* 0% indicates that the model explains none of the variability of the response data around its mean.
* 100% indicates that the model explains all the variability of the response data around its mean.

## 2.1 HANDLING IMBALANCED DATA USING SMOTE

Dealing with imbalanced datasets entails strategies such as improving classification algorithms or balancing classes in the training data (data preprocessing) before providing the data as input to the machine learning algorithm. The later technique is preferred as it has wider application.

The main objective of balancing classes is to either increasing the frequency of the minority class or decreasing the frequency of the majority class. This is done in order to obtain approximately the same number of instances for both the classes.

# **2.2 TENTATIVE LIST OF ALGORITHMS & INITIAL APPROACH**

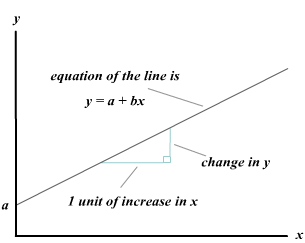
Since our problem statement deals with both regression and classification problems, we will be using the following algorithms in modelling:

* Linear Regression
* Logistic Regression
* Tree Based Classifiers / Regressor
  + Decision Tree
  + Random Forest
  + Bagging

## 2.3 LINEAR REGRESSION

When we explore the relationship between two quantitative variables graphically has a linear (or straight line) pattern, the correlation provides a numerical measure of the strength and direction of the relationship. We can analyse the data further by finding an equation for the straight line that best describes that pattern. That equation predicts the value of the response variable from the value of the explanatory variable. The straight line that best describes the linear pattern is called the regression line. The equation of the regression line predicts the value for the response variable y from the explanatory variable x. The equation for the regression line has the form:

Y = a + bx

Where ‘a’ denotes the y-intercept, and b denotes the slope.

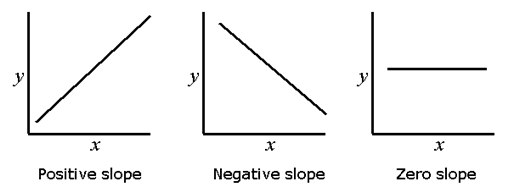
Regression line

Note: The response variable is also referred as the dependent variable; and the explanatory variable is also referred as the independent variable. The regression is referred as multiple regression when we have two or more independent variables.

A regression equation is often called a prediction equation, since it predicts the value of the response variable y at any value of x.

### **INTERPRETING THE REGRESSION LINE**

In descriptive statistics, the simple linear regression line describes the nature of the relationship between the dependent variable y and the independent variable x. The slope of the line reflects the degree to which the variable y changes *linearly* as a function of changes in the variable x. The sign of the slope indicates whether the relationship between x and y is positive or negative. A positive slope indicates that both variables changes in the same direction; and a negative slope indicates that x and y changes in opposite directions. If there is no linear relationship between x and y, then the regression line will be flat (zero slope).



Interpreting the slope of the regression line

**Assumptions of Linear Regression:**

* The regression model is linear in parameters.
* The mean of residuals is zero.
* Homoscedasticity of residuals or equal variance.
* No autocorrelation of residuals.
* The X variables and residuals are uncorrelated.
* The variability in X values is positive.
* The regression model is correctly specified.
* No perfect multi-collinearity.

## RIDGE REGRESSION

Ridge regression is a form of linear regression where the parameters of the regression equation are reduced to prevent multicollinearity. This reduces the model complexity and there by reducing variance. The regularization technique used to reduce the coefficients is L2 regularization technique. The cost function for ridge regression is:

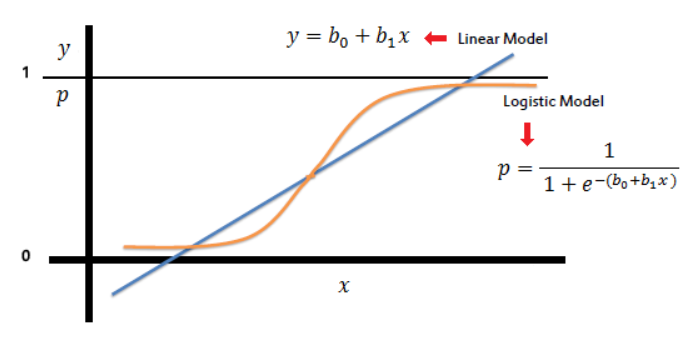


## LOGISTIC REGRESSION

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

* A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)
* Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.



**Assumptions or Requirements of Logistic Regression:**

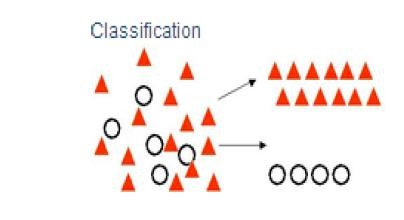
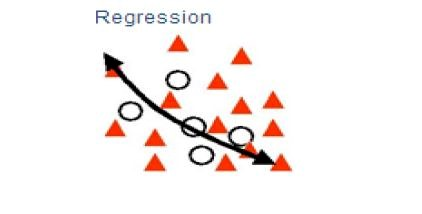
* First, binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
* Second, logistic regression requires the observations to be independent of each other.  In other words, the observations should not come from repeated measurements or matched data.
* Third, logistic regression requires there to be little or no multi-collinearity among the independent variables.  This means that the independent variables should not be too highly correlated with each other.
* Fourth, logistic regression assumes linearity of independent variables and log odds.  Although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.
* Finally, logistic regression typically requires a large sample size.

## 2.5 DECISION TREE (CART)

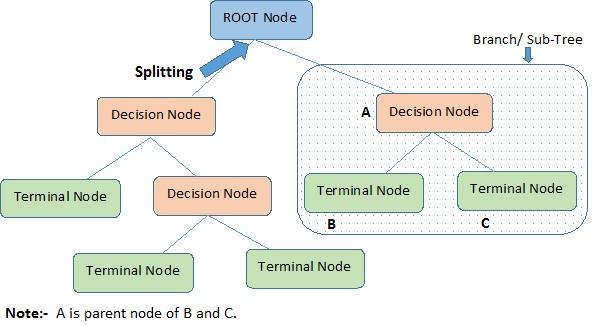
A Decision tree (CART) is a schematic, tree-shaped diagram used to determine a course of action or show a statistical probability. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

**Types of Decision Tree**

* **Classification Trees:** where the Dependent variable is categorical and the tree is used to identify the "class" within which a Dependent variable would likely fall into.
* **Regression Trees:** where the Dependent variable is continuous and tree is used to predict its value. (e.g. the price of a house, or a patient's length of stay in a hospital).



**Layout / flow of Decision Tree**



**Advantages of CART**

* Simple to understand, interpret, visualize.
* Decision trees implicitly perform variable screening or feature selection.
* Can handle both numerical and categorical data. Can also handle multi-output problems.
* Decision trees require relatively little effort from users for data preparation.
* Nonlinear relationships between parameters do not affect tree performance.

**Disadvantages of CART**

* Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting.
* Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This is called variance, which needs to be lowered by methods like bagging and boosting.
* Greedy algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees, where the features and samples are randomly sampled with replacement.
* Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the data set prior to fitting with the decision tree.

## 2.6 RANDOM FOREST

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is, that it can be used for both classification and regression problems.



Random Forest has nearly the same hyper parameters as a decision tree or a bagging classifier. Fortunately, we don’t have to combine a decision tree with a bagging classifier and can just easily use the classifier-class of Random Forest. Like I already said, with Random Forest, you can also deal with Regression tasks by using the Random Forest regressor.

Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

**Advantages of Random Forest:**

* There is no need for feature normalization
* Individual decision trees can be trained in parallel
* Reduced overfitting
* Require almost no input preparation
* Performs implicit feature selection
* It’s very quick to train

**Disadvantages of Random Forest:**

* No interpretability

## 2.7 BAGGING

Bagging is used typically when you want to reduce the variance while retaining the bias. This happens when you average the predictions in different spaces of the input feature space.

In bagging, first you will have to sample the input data (with replacement) to generate multiple sets of input data. For each of those sets, the same baseline predictor (such as a SVM, Neural Net, etc) is run to get a trained model for each of the training set.

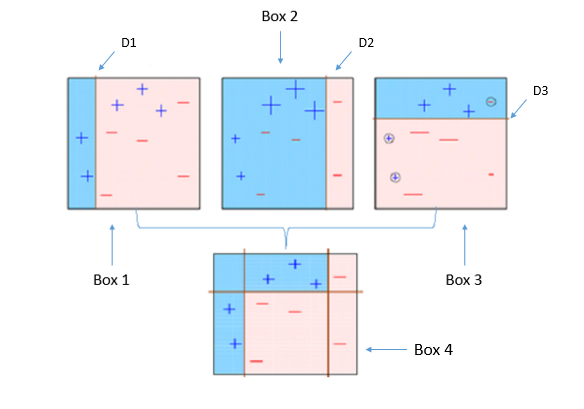
Now, to do the prediction on an unseen test sample, it is run through these individual models and the predictions are now averaged to get the final decision.

Bagging is effective because you are improving the accuracy of a single model by using multiple copies of it trained on different sets of data.

Bagging is not recommended on models that have a high bias. In such cases, boosting (Ada-boost) is used which goes a step ahead and eliminates the effect of a high bias present in the baseline model,

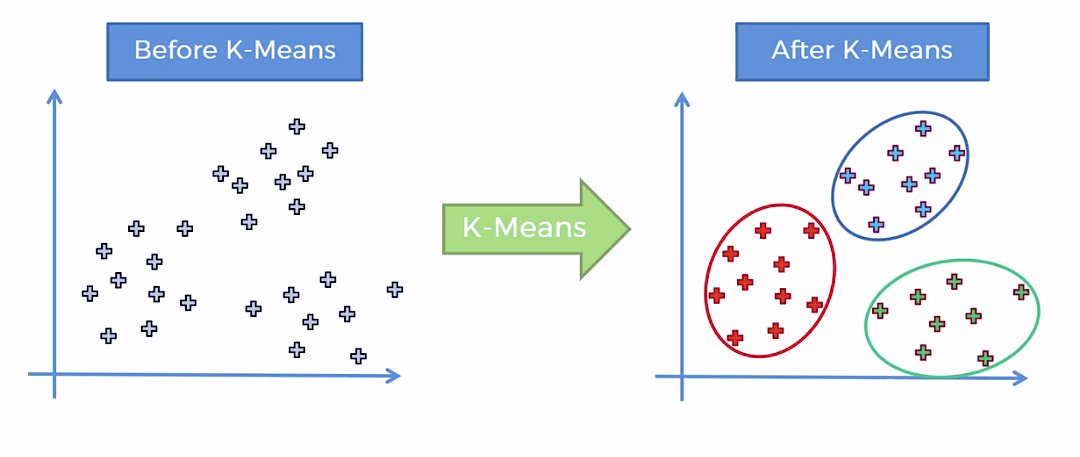
## 2.8 ADABOOST

AdaBoost is an ensemble algorithm where multiple weak learners are combined to form a strong learner. It works by learning on the original dataset in the first instance and adjusts weights of the parameters according to the errors and then re-learns the dataset. AdaBoost learns in a sequential manner on one base estimator. AdaBoost is used in both classification and regression.



## 2.9 KMEANS CLUSTERING

Kmeans is an iterative algorithm where the dataset is partitioned into pre-defined K number of clusters where each data point belongs to only one cluster. The algorithm tries to make the inter cluster data points as similar as possible and clusters as far as possible. Kmeans first initializes centroids by shuffling the dataset and then randomly selecting K data points as centroids without replacement. It then computes sum of the squared distances between data points and all centroids and assigns the data point to the centroid that is closest. The algorithm reiterates until there is no change to the centroids.



## 2.10 HANDLE IMBALACED CLASSIFICATION PROBLEMS IN MACHINE LEARNING

While performing the conventional machine learning on the data which is imbalanced the model will be inaccurate and biased. In this case number of observations in one class will be significantly lesser than other. This problem is predominant in scenarios where anomaly detection is crucial like electricity pilferage, fraudulent transactions in banks, identification of rare diseases, etc. Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes which have number of instances. This happens because Machine Learning Algorithms are usually designed to improve accuracy by reducing the error. Thus, they do not take into account the class distribution / proportion or balance of classes. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored.

# **DATA DESCRIPTION**

## 3.1 DATA SET

<https://www.lendingclub.com/info/download-data.action>

The data is obtained from the official Lending Club website. The analyzed period was for ~5 years from 2007 to 2011. The dataset has 42540 rows and 144 columns.

## 3.2 TARGET VARIABLE

The target variable is ‘potential loan amount to offer’ in case of regression analysis and ‘defaulter prediction’ which was derived from loan status column in order to perform classification analysis

# **4. DATA PRE-PROCESSING**

## 4.1 MISSING VALUE TREATMENT

* Out of 144 columns present in the dataset, 55 columns have been removed due to presence of more than 70% of missing values
* Out of ~46k rows, ~1k rows have been removed due to missing values in various features

## 4.2 FEATURE TRANSFORMATION

* Few columns with categorical data type have been converted into numeric data type (ordinal data type)
* Some of the date columns which are in object type are converted to date time object to identify the trend over the years
* Usage of standard scalar

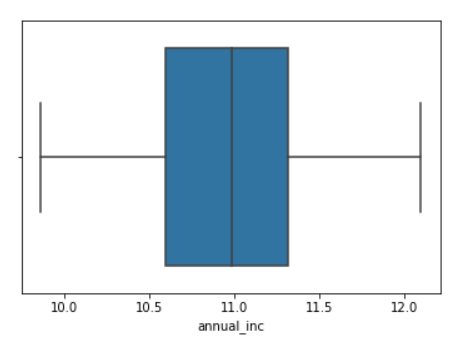
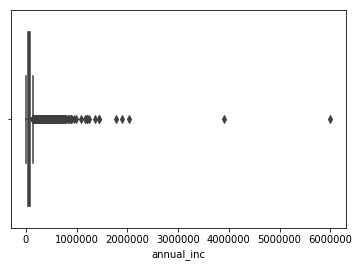
## 4.3 FEATURE ENGINEERING

* Loan status column has been split into two. One column tells if the borrower has met the credit policy, whereas, the other one tells if the borrower has fully paid or charged off.

## 4.4 OUTLIER TREATMENTS

We have identified few columns in the dataset have high number of outliers. In order to treat the outliers, Winsorization technique has been applied.

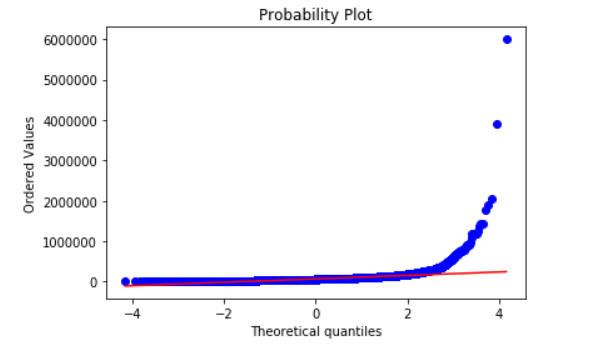
Below is an example of annual income column where a large number of outliers were found. The pictures represent the data before and after the treatment.



## 4.5 VARIABLE TRANSFORMATION AND NORMALIZATION

Log Transform and Square Root Transformation were used to get the data into normalized form. QQ-Plots and Anderson-Darling tests were used to test the normality of the data after normalization.

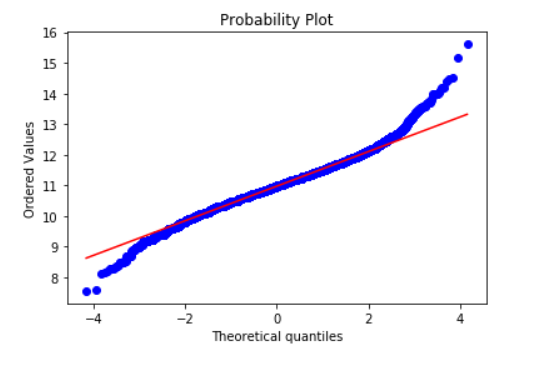
As shown in the QQ-Plot, below is an example of annual income column with and without normalized data. Anderson-Darling scores tells how close the data is to the normalized form.



Anderson Result = 3622

Critical Values = 0.787

Significance = 5%



Anderson Result = 50

Critical Values = 0.787

Significance = 5%

## 4.6 MULTI-COLLINEARITY TREATMENT

**Using correlation matrix for numerical data type**

We were able to trace some multi-collinearity between features because they are closely related such as loan amount, funded amount, installment amount etc. However, correlation matrix provided us an evidence and helped us in ruling out features to reduce multi-collinearity.

**ANOVA Test to check if there are any signification differences between the means**

ANOVA tells us if the means of various features are coming from the same larger population. It compares means of various features to see if they are significantly equal or unequal.

**The null hypothesis of the ANOVA** test is that the means of all the groups are equal to each other.

The p-value will tell us if our test results are significant or not. In order to perform an ANOVA test and get the p-value, you need two pieces of information:

* Degrees of freedom within the group.
* Degrees of freedom among the groups.
* The alpha level (α). The usual alpha or significance level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

*We reject the null hypothesis when the P-value is less than the set significance level.*

**Chi-Square Test for Independence**

A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.

In our case, we need to determine whether there is indeed a relationship between a predictor variable and any of the target variables to a significant degree. We only need to consider these features for our further analysis.

**The null hypothesis of the Chi-Square** test is that no relationship exists on the categorical variables being tested. I.e. they are independent.

The p-value will tell us if our test results are significant or not. In order to perform a chi square test and get the p-value, you need two pieces of information:

* Degrees of freedom. That’s just the number of categories minus 1.
* The alpha level (α). The usual alpha or significance level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

*We reject the null hypothesis when the P-value is less than the set significance level.*

**After applying all the above techniques, we were able to reduce the shape of our dataset to (72200, 34) for the classification problem. The SMOT technique was not used for the regression problem as class imbalance does not affect the regression. The classes in the categorical variables were encoded using get dummies to reduce bias in a feature. The shape of the dataset for regression problem was (42445,139).**

**5. EXPLORATORY DATA ANALYSIS**

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

The importance of EDA are as follows:

1. It is always a good idea to explore a data set with multiple exploratory techniques, especially when they can be done together for comparison. Every data scientist should compile a cookbook of techniques in exploratory data analysis. Once we fully understand our data set, it is quite possible that we may need to revisit one or more data munging tasks in order to refine or transform the data even further.

2. Another side benefit of EDA is to refine your selection of feature variables that will be used later for machine learning. Once we gain deep familiarity with your data set, we may need to revisit the feature engineering step – we may find the features we selected do not serve their intended purpose. Moreover, we may discover other features that add to the overall picture the data presents. Once we complete the EDA stage, we should have a firm feature set with which to use for supervised and unsupervised statistical learning.

So now we should proceed with the EDA so as to get better insights of Data and we can get better inferences of how loan is given in lending club, the types of grades decided by the Lending Club and the sum of fully paid or charged off loans.

We will use Data Visualization Software Tableau for getting inferences and insights of the Data Provided to us. This Inferences will be very helpful while making the model and it will also help us in classifying the people grade wise and help us to predict the interest rate.

A close up of a map

Description automatically generated

We can see that Average Delinquency rate in East Region is the highest of about 33.70% and it is followed West of 26.21%. We might wonder the reason behind such high Delinquency rate is because of the amount of Not Verified individuals in the in the East Region. The same goes for West region Where the amount of Non-Verified individuals are high. This can be seen from the next Figure.:

A screenshot of a cell phone

Description automatically generated

East

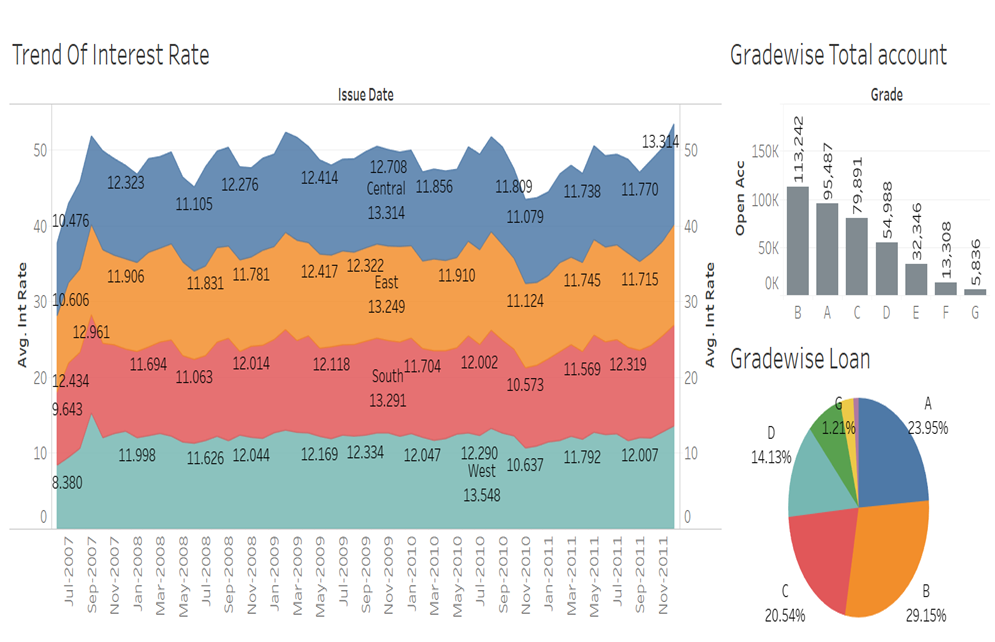
Central

South

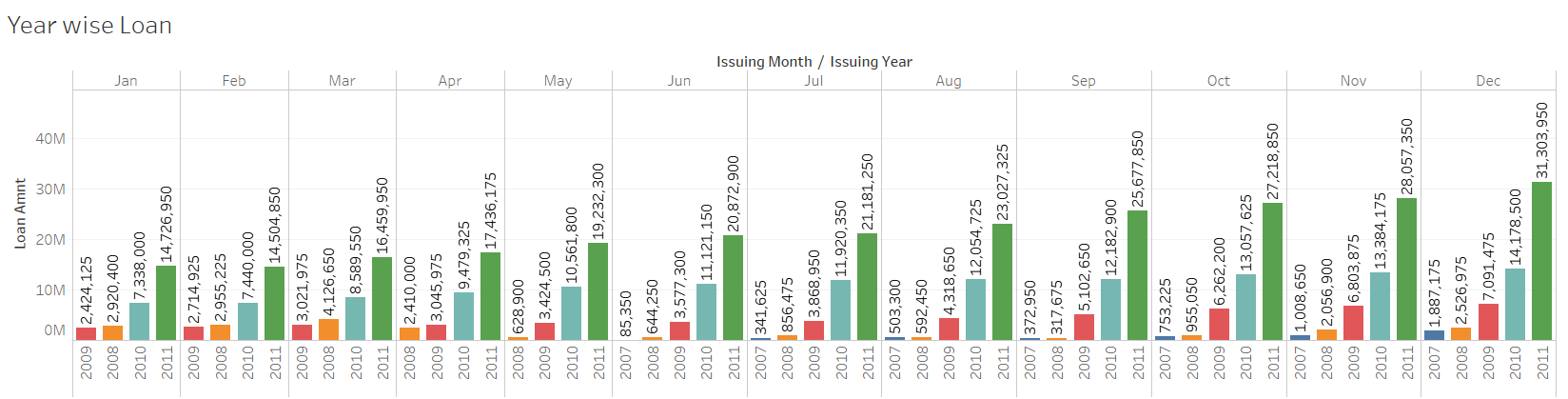
West

A screenshot of a cell phone

Description automatically generated



From these graphs we can see that Grade B has taken the highest amount of Loan and has the highest number of Account Opened with Interest Rate Ranging from 9.01% to 12.05%. After grade B comes Grade A and Grade C. We can also see that at the End of each Financial Year Lending Club has reduced the Interest rate so more loan is disbursed. We can see a dip in the Interest rate so that more amount of Loan can be disbursed at the end of Each Financial year.

****

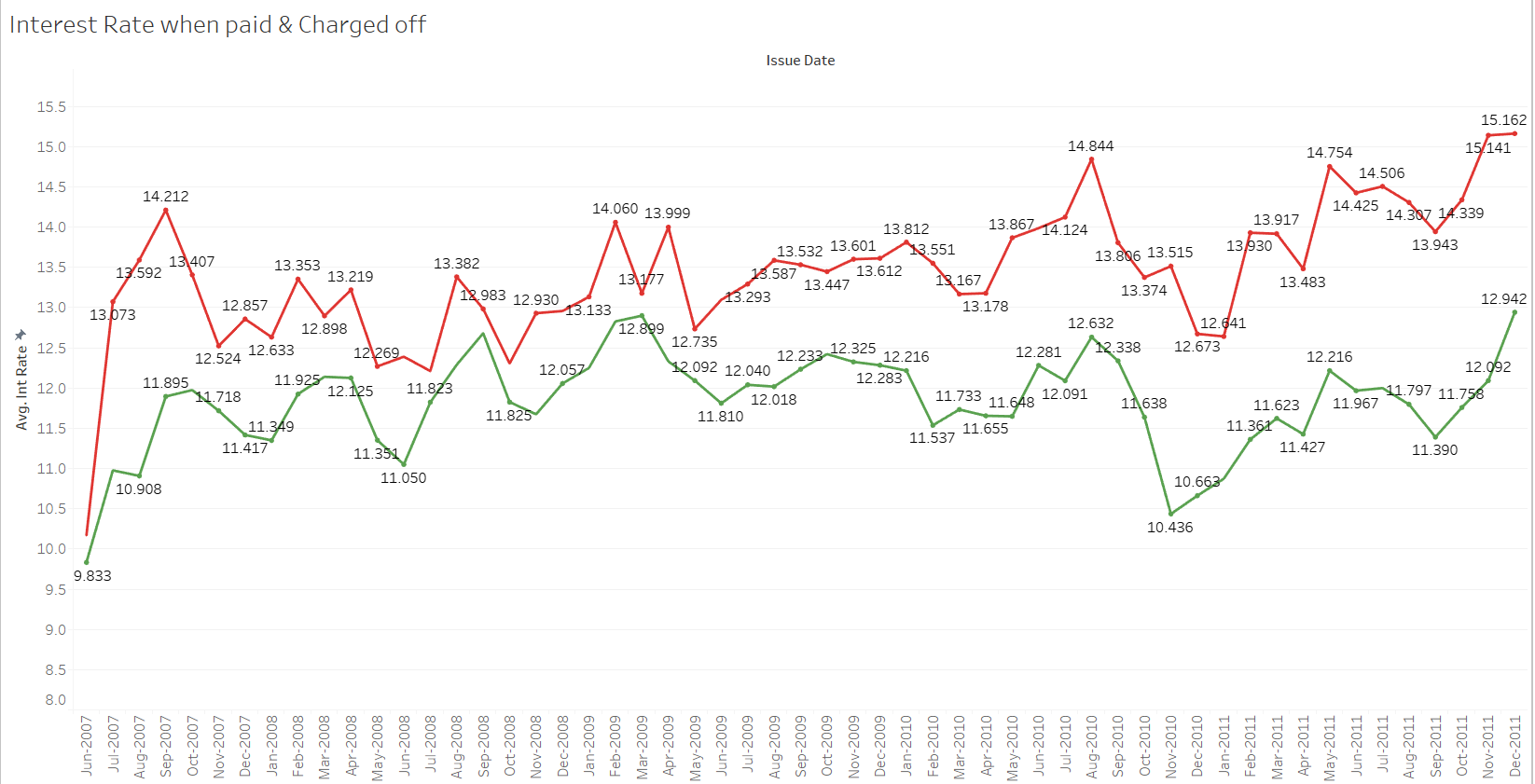
A close up of a logo

Description automatically generated

A picture containing screenshot

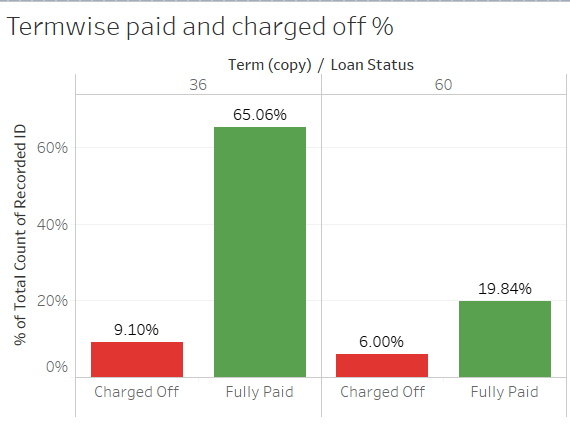
Description automatically generated

The Bubble Chart shows a quite clear picture that Debt consolidation has the maximum %. Debt Consolidation is a way where one can pay off debts quite easily. [Debt consolidation](https://www.thebalance.com/how-to-choose-a-credit-consolidation-loan-960821) is the process of combining all of our unsecured debts into a single monthly payment. Debt consolidation might be done with a [debt consolidation loan](https://www.thebalance.com/types-of-debt-consolidation-loans-960610). The loan is used to pay off our debts, then we pay off the new consolidation loan rather than dividing our payments to our creditors. we may be able to take out a debt consolidation on our own using a home equity loan or a debt consolidation loan from a bank. Box Plot shows not much extreme outliers for a Particular grade However it shows an increase in the interest rates when anyone moves from Grade A to Grade G. In the below diagram Green line indicates fully paid loans whereas the red line indicates the charged off loans.



A close up of a map

Description automatically generated



29.04% Loan Amount has been Disbursed to Grade B people and then 20.14% to Grade C people as shown in the below figure. The number of Delinquency in Grade C and Grade D people are more if we take Total Loan Amount (Credit) into Consideration. We prioritize our focus on Grade C and Grade D people more and then we should Focus towards Grade B people for giving Loans. We can also see when the term is more there is high chances of charged off loans.

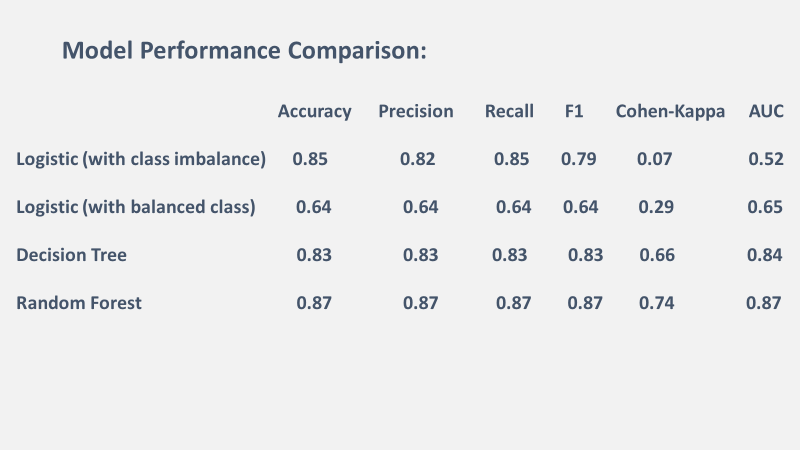
A close up of a map

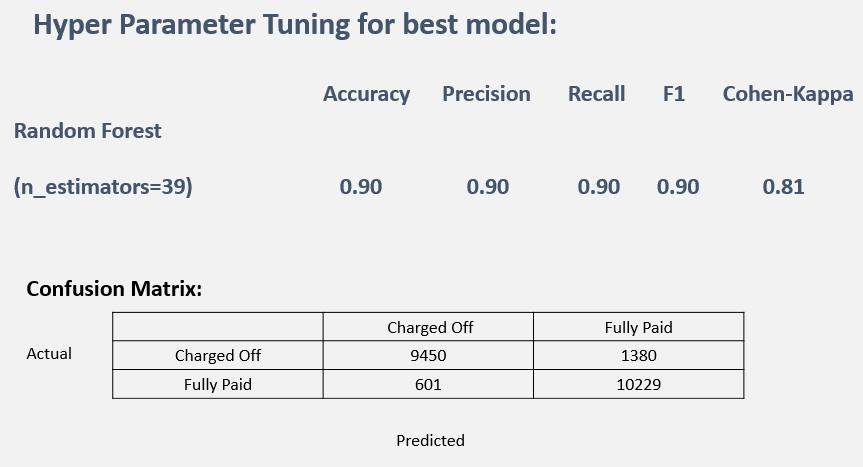
Description automatically generated

Grade B has the highest number of not Verified IDs so the club should focus on getting them verified however first focus should be on Grade C and Grade D people because they are the people who take less loan but their chances of Delinquency is very High. In credit card terms, a **Revolving balance** is the portion of credit card spending that goes unpaid at the end of a billing cycle. The amount can vary, going up or down depending on the amount borrowed and the amount repaid. Grade B People have the highest Amount of Credit Revolving Balance so the club should focus on decreasing the Revolving Balance. Focus should also be given on Grade F and Grade G people before giving the loans.

# **6. RESULTS AND COMPARISON STUDY**

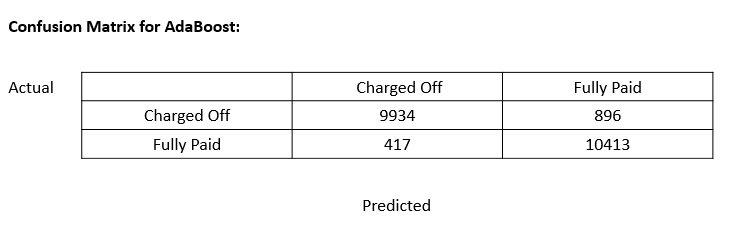
## 6.1 DEFAULTERS PREDICTION USING CLASSIFICATION:

****



**WITH ENSEMBLE**

****



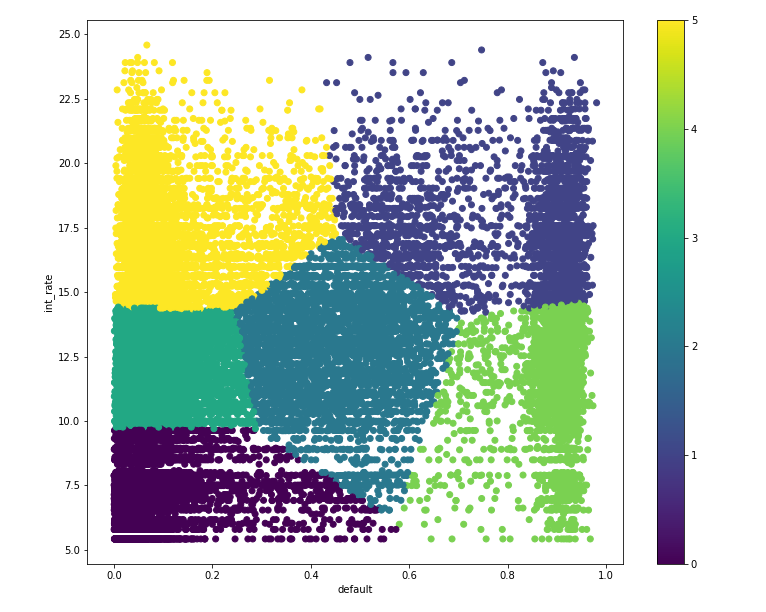
**INFERENCE:** In this case we have identified that AdaBoost with random forest as base estimator to be the best model because of higher precision and recall rates. With AdaBoost model we are able to identify charged off borrowers with greater accuracy.

## 6.2 INTEREST RATE PREDICTION USING REGRESSION:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **R-Squared** |
| Linear Regression | 0.58 | 0.61 | 0.78 | 0.95 |
| Ridge Regression | 0.58 | 0.61 | 0.78 | 0.95 |
| Random Forest Regressor | 0.49 | 0.50 | 0.70 | 0.96 |
| AdaBoost Regressor with Linear | 0.88 | 1.32 | 1.15 | 0.90 |
| AdaBoost Regressor with Ridge | 0.76 | 1.04 | 1.02 | 0.93 |

**INFERENCE**: From the above table, linear regression and ridge regression perform equally well and errors are slightly less when random forest regressor is applied. The ensemble model adaboost regressor with base estimators linear and ridge increases the errors.

## 6.3 RISK ASSESSMENT:



Cluster details:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Risk categories | Cluster Number | Interest Minimum | Interest Maximum | Default probability Minimum | Default Probability Maximum |
| low | 0 | 5.42 | 9.7 | 0 | 0.57 |
| HIgh | 1 | 14.22 | 24.4 | 0.43 | 0.98 |
| Moderate | 2 | 6.54 | 17.04 | 24.94 | 0.69 |
| Low-medium | 3 | 9.7 | 14.42 | 0 | 0.28 |
| Moderate-High | 4 | 5.42 | 14.57 | 0.57 | 0.97 |
| Low-high | 5 | 14.35 | 24.59 | 0 | 0.45 |

Within Cluster credit profiles:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Credit profile grade | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
| A | 9415 |  | 242 |  | 522 |  |
| B | 688 |  | 1422 | 9083 | 1196 |  |
| C | 4 | 210 | 1349 | 5369 | 981 | 827 |
| D | 4 | 1075 | 711 | 413 | 147 | 3666 |
| E | 5 | 941 | 140 | 68 | 48 | 2192 |
| F | - | 474 | 14 |  |  | 813 |
| G | - | 203 |  |  |  | 309 |

**INFERENCE:** From the above cluster picture and cluster details table and within cluster credit profiles table, the data can be divided into risk categories and then interest rate can be derived as the new borrower is classified in a cluster. The cluster 0 has low interest rates even though the maximum default probability is 0.57 because the credit profile of the borrower is of grade A. This can be observed throughout the data for other credit profile grades.

From the tables it can be inferred that the risk factor is highly dependent on probability of borrower defaulting on loan and interest rate assigned to the borrower is dependent on default probability and the credit profile grade of the borrower.

# **7. CONCLUSION**

This section will reflect on the results obtained from this project and also any further works that could be done. Extensive research has been done previously in this area with satisfactory results. However they primarily focused on the classification of borrowers based on their loan status, we also created credit report scorecard using a regression model to suggest interest rates to be charged to hedge the risk involved.

Overall, the accuracy rates show that the methods chosen were successful for the classification task. Further, the similar results for the different methods show that the algorithms are mutually competitive. Further, while there are many different types of features that can be extracted, the best results are those based on heavily pre-processed data. This shows how preprocessing data in such a way can later be beneficial for many other applications.

From the models built and the tests performed, Random Forest Regressor is the best method to predict the how much interest rate has to be levied on a particular loan application.

In case of defaulter’s prediction Random Forest Classifier with hyper parameter tuning was the best model and we recommend that model because

• Improves stability & accuracy of machine learning algorithms

• Reduces variance

• Overcomes over fitting

# **8. APPENDIX**

**REFERENCES:**

* <https://www.saedsayad.com/logistic_regression.htm>
* <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>
* <https://en.wikipedia.org/wiki/Random_forest>
* <https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>
* <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>
* <https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/>
* <https://en.wikipedia.org/wiki/LendingClub>
* <http://cs229.stanford.edu/proj2015/199_report.pdf>
* <https://en.wikipedia.org/wiki/Peer-to-peer_lending>
* <https://www.valuepenguin.com/loans/what-are-p2p-loans>
* <https://www.sofi.com/learn/content/understanding-p2p-lending-works/>