**Dataset** – <https://www.kaggle.com/wendykan/lending-club-loan-data>

**Reference links** – <https://rpubs.com/gabrielmartos/discriminantR>

http://rstudio-pubs/static.s3.amazonaws.com/5346\_a009a0dc6db04fb8858a20467da5636a.html

**Reason for selecting this dataset:**

Predicting how likely a borrower will default on a loan is a dilemma faced by banks and lenders. Typically, a borrower is classified into grades based on which it is determined whether the loan should be sanctioned to him/her. Our dataset contains the complete loan data for all the loans issued through 2007 – 2015 by the Lending Club in USA which is a peer-to-peer lending company. The primary product of the organization is Loans and investments.

There were other data sets available on Kaggle for Loan data like – loan data, credit card data, credit fraud data, etc. but we picked this dataset as we wanted the dataset to contain a large number of observations and features so that we can predict the probability of defaulting on a loan more accurately. This dataset has 887k observations and 74 features which fit the bill of our requirement. The dataset includes features like current loan status, latest payment information, credit scores, region etc, all of which give a clear picture about the financial stability of each client based on which an informed decision can be made about sanctioning a loan.

We begin exploring the data by importing the data as a data frame into R and using the structure function to get an understanding of the number of observations and features.

**Objective:**

The objective of this project is to model the data of ongoing loans and to predict whether a client’s loan will be approved or rejected based on the credit risk. We have created a new feature in the data called loan\_approval based on credit grade. The data has 7 grades (A to G), A and B are considered to be least risky and have a loan\_approval status of “Approved” while the remaining grades C to G having a higher risk have a status of “Rejected”.

**Data Exploration:**

One of the important steps in classifying the grades of the prospective clients is Data exploration. Data exploration helps us the understand the structure of the data better. It exposes the outliers and anomalies in the data.

We used the R function structure to view the structure of the data frame.

'data.frame': 887379 obs. of 74 variables:

$ id : int 1077501 1077430 1077175 1076863 1075358 1075269 1069639 1072053 1071795 1071570 ...

$ member\_id : int 1296599 1314167 1313524 1277178 1311748 1311441 1304742 1288686 1306957 1306721 ...

$ loan\_amnt : num 5000 2500 2400 10000 3000 ...

$ funded\_amnt : num 5000 2500 2400 10000 3000 ...

$ funded\_amnt\_inv : num 4975 2500 2400 10000 3000 ...

$ term : Factor w/ 2 levels " 36 months"," 60 months": 1 2 1 1 2 1 2 1 2 2 ...

$ int\_rate : num 10.7 15.3 16 13.5 12.7 ...

$ installment : num 162.9 59.8 84.3 339.3 67.8 ...

$ grade : Factor w/ 7 levels "A","B","C","D",..: 2 3 3 3 2 1 3 5 6 2 ...

$ sub\_grade : Factor w/ 35 levels "A1","A2","A3",..: 7 14 15 11 10 4 15 21 27 10 ...

$ emp\_title : Factor w/ 299273 levels ""," \tAdv Mtr Proj Fld Rep",..: 1 224800 1 9368 282199 285977 246848 171062 1 256905 ...

$ emp\_length : Factor w/ 12 levels "< 1 year","1 year",..: 3 1 3 3 2 5 10 11 6 1 ...

$ home\_ownership : Factor w/ 6 levels "ANY","MORTGAGE",..: 6 6 6 6 6 6 6 6 5 6 ...

$ annual\_inc : num 24000 30000 12252 49200 80000 ...

$ verification\_status : Factor w/ 3 levels "Not Verified",..: 3 2 1 2 2 2 1 2 2 3 ...

$ issue\_d : Factor w/ 103 levels "Apr-2008","Apr-2009",..: 22 22 22 22 22 22 22 22 22 22 ...

$ loan\_status : Factor w/ 10 levels "Charged Off",..: 6 1 6 6 2 6 2 6 1 1 ...

$ pymnt\_plan : Factor w/ 2 levels "n","y": 1 1 1 1 1 1 1 1 1 1 ...

$ url : Factor w/ 887379 levels "https://www.lendingclub.com/browse/loanDetail.action?loan\_id=1000007",..: 21292 21256 21242 21220 20692 20684 19191 19811 19796 19657 ...

$ desc : Factor w/ 124471 levels "","\t Loan for purchase of grand piano. Piano will further diversify an already profitable business. Monthly budge"| \_\_truncated\_\_,..: 113402 113407 1 113258 113232 1 112347 111631 113230 111646 ...

$ purpose : Factor w/ 14 levels "car","credit\_card",..: 2 1 12 10 10 14 3 1 12 10 ...

$ title : Factor w/ 63146 levels "","\tcredit\_card",..: 10496 4975 52500 50874 50267 42595 36948 7263 24371 6112 ...

$ zip\_code : Factor w/ 935 levels "007xx","008xx",..: 810 296 572 856 909 803 267 839 897 729 ...

$ addr\_state : Factor w/ 51 levels "AK","AL","AR",..: 4 11 15 5 38 4 28 5 5 44 ...

$ dti : num 27.65 1 8.72 20 17.94 ...

$ delinq\_2yrs : num 0 0 0 0 0 0 0 0 0 0 ...

$ earliest\_cr\_line : Factor w/ 698 levels "","Apr-1955",..: 265 43 572 210 276 575 342 287 48 690 ...

$ inq\_last\_6mths : num 1 5 2 1 0 3 1 2 2 0 ...

$ mths\_since\_last\_delinq : num NA NA NA 35 38 NA NA NA NA NA ...

$ mths\_since\_last\_record : num NA NA NA NA NA NA NA NA NA NA ...

$ open\_acc : num 3 3 2 10 15 9 7 4 11 2 ...

$ pub\_rec : num 0 0 0 0 0 0 0 0 0 0 ...

$ revol\_bal : num 13648 1687 2956 5598 27783 ...

$ revol\_util : num 83.7 9.4 98.5 21 53.9 28.3 85.6 87.5 32.6 36.5 ...

$ total\_acc : num 9 4 10 37 38 12 11 4 13 3 ...

$ initial\_list\_status : Factor w/ 2 levels "f","w": 1 1 1 1 1 1 1 1 1 1 ...

$ out\_prncp : num 0 0 0 0 767 ...

$ out\_prncp\_inv : num 0 0 0 0 767 ...

$ total\_pymnt : num 5861 1009 3004 12226 3242 ...

$ total\_pymnt\_inv : num 5832 1009 3004 12226 3242 ...

$ total\_rec\_prncp : num 5000 456 2400 10000 2233 ...

$ total\_rec\_int : num 861 435 604 2209 1009 ...

$ total\_rec\_late\_fee : num 0 0 0 17 0 ...

$ recoveries : num 0 117 0 0 0 ...

$ collection\_recovery\_fee : num 0 1.11 0 0 0 0 0 0 2.09 2.52 ...

$ last\_pymnt\_d : Factor w/ 99 levels "","Apr-2008",..: 42 7 58 42 43 42 43 42 6 80 ...

$ last\_pymnt\_amnt : num 171.6 119.7 649.9 357.5 67.8 ...

$ next\_pymnt\_d : Factor w/ 101 levels "","Apr-2008",..: 1 1 1 1 35 1 35 1 1 1 ...

$ last\_credit\_pull\_d : Factor w/ 104 levels "","Apr-2009",..: 43 102 43 42 43 104 43 25 14 67 ...

$ collections\_12\_mths\_ex\_med : num 0 0 0 0 0 0 0 0 0 0 ...

$ mths\_since\_last\_major\_derog: num NA NA NA NA NA NA NA NA NA NA ...

$ policy\_code : num 1 1 1 1 1 1 1 1 1 1 ...

$ application\_type : Factor w/ 2 levels "INDIVIDUAL","JOINT": 1 1 1 1 1 1 1 1 1 1 ...

$ annual\_inc\_joint : num NA NA NA NA NA NA NA NA NA NA ...

$ dti\_joint : num NA NA NA NA NA NA NA NA NA NA ...

$ verification\_status\_joint : Factor w/ 4 levels "","Not Verified",..: 1 1 1 1 1 1 1 1 1 1 ...

$ acc\_now\_delinq : num 0 0 0 0 0 0 0 0 0 0 ...

$ tot\_coll\_amt : num NA NA NA NA NA NA NA NA NA NA ...

$ tot\_cur\_bal : num NA NA NA NA NA NA NA NA NA NA ...

$ open\_acc\_6m : num NA NA NA NA NA NA NA NA NA NA ...

$ open\_il\_6m : num NA NA NA NA NA NA NA NA NA NA ...

$ open\_il\_12m : num NA NA NA NA NA NA NA NA NA NA ...

$ open\_il\_24m : num NA NA NA NA NA NA NA NA NA NA ...

$ mths\_since\_rcnt\_il : num NA NA NA NA NA NA NA NA NA NA ...

$ total\_bal\_il : num NA NA NA NA NA NA NA NA NA NA ...

$ il\_util : num NA NA NA NA NA NA NA NA NA NA ...

$ open\_rv\_12m : num NA NA NA NA NA NA NA NA NA NA ...

$ open\_rv\_24m : num NA NA NA NA NA NA NA NA NA NA ...

$ max\_bal\_bc : num NA NA NA NA NA NA NA NA NA NA ...

$ all\_util : num NA NA NA NA NA NA NA NA NA NA ...

$ total\_rev\_hi\_lim : num NA NA NA NA NA NA NA NA NA NA ...

$ inq\_fi : num NA NA NA NA NA NA NA NA NA NA ...

$ total\_cu\_tl : num NA NA NA NA NA NA NA NA NA NA ...

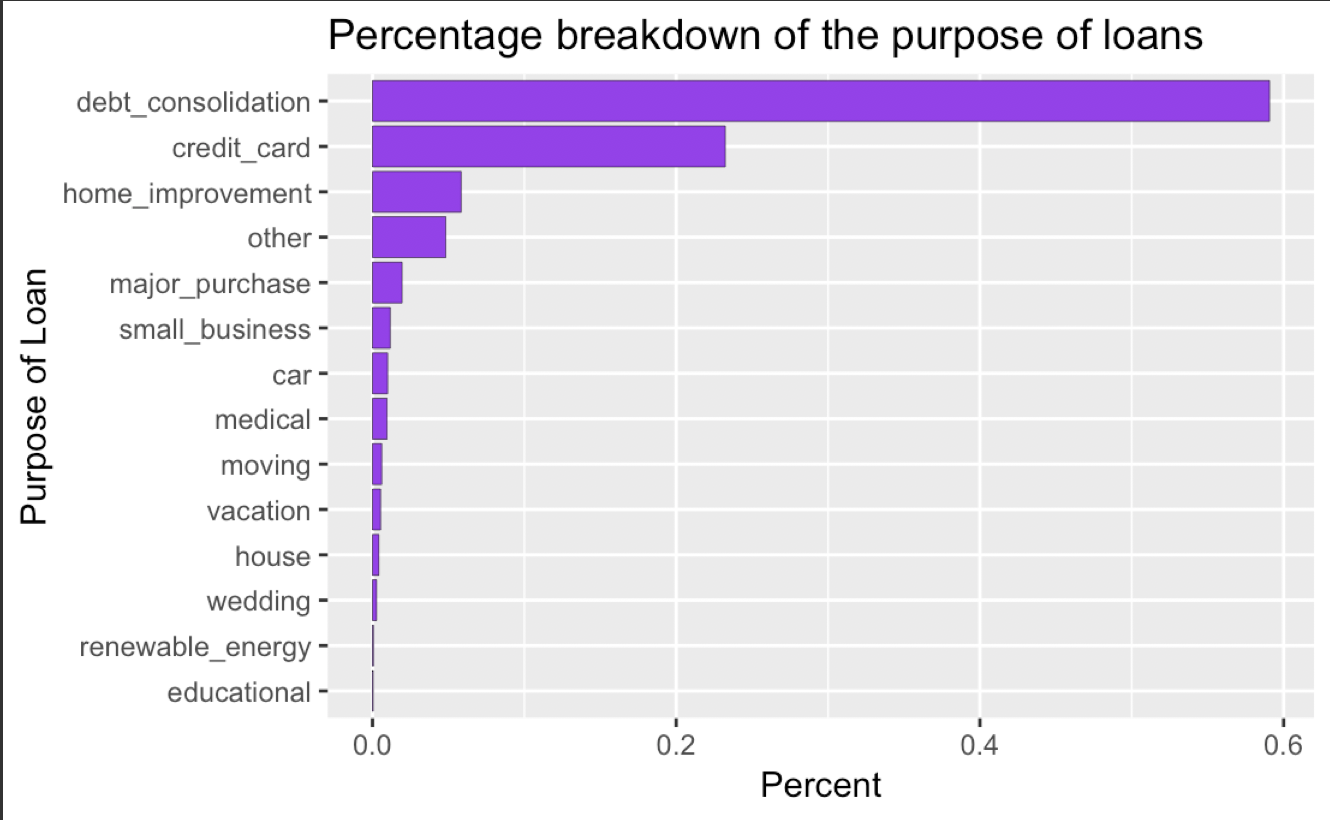
$ inq\_last\_12m : num NA NA NA NA NA NA NA NA NA NA ...

The data we have chosen consists of 74 features (23 factor variables and 51 numerical variables) and 887k observations.

**Data Visualization:**

We use data visualization techniques to significance of data by placing it in a visual context. When the data is in the form of text, certain patterns or trends may go undetected. These trends can be exposed through graphical representation.

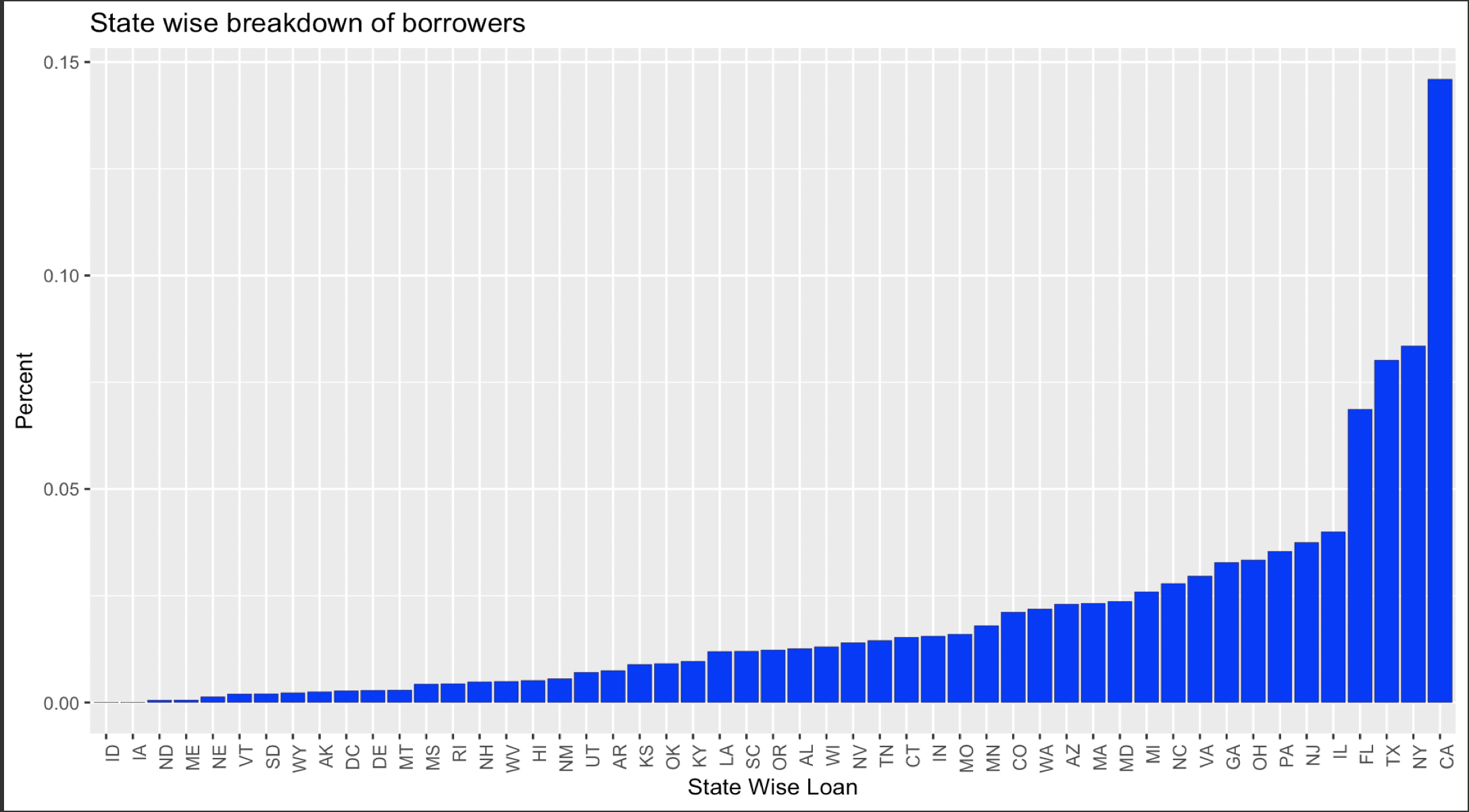
We begin by visualizing the breakdown of the purpose of the loan from the data.

****

This graph indicates shows us the various reasons for which people have taken loans form Loan Club. The most common purpose of borrowing a loan is debt\_consolidation. Personal reasons such as car, medical, moving, vacation and housing totally account to less than 10% of purpose of loan.

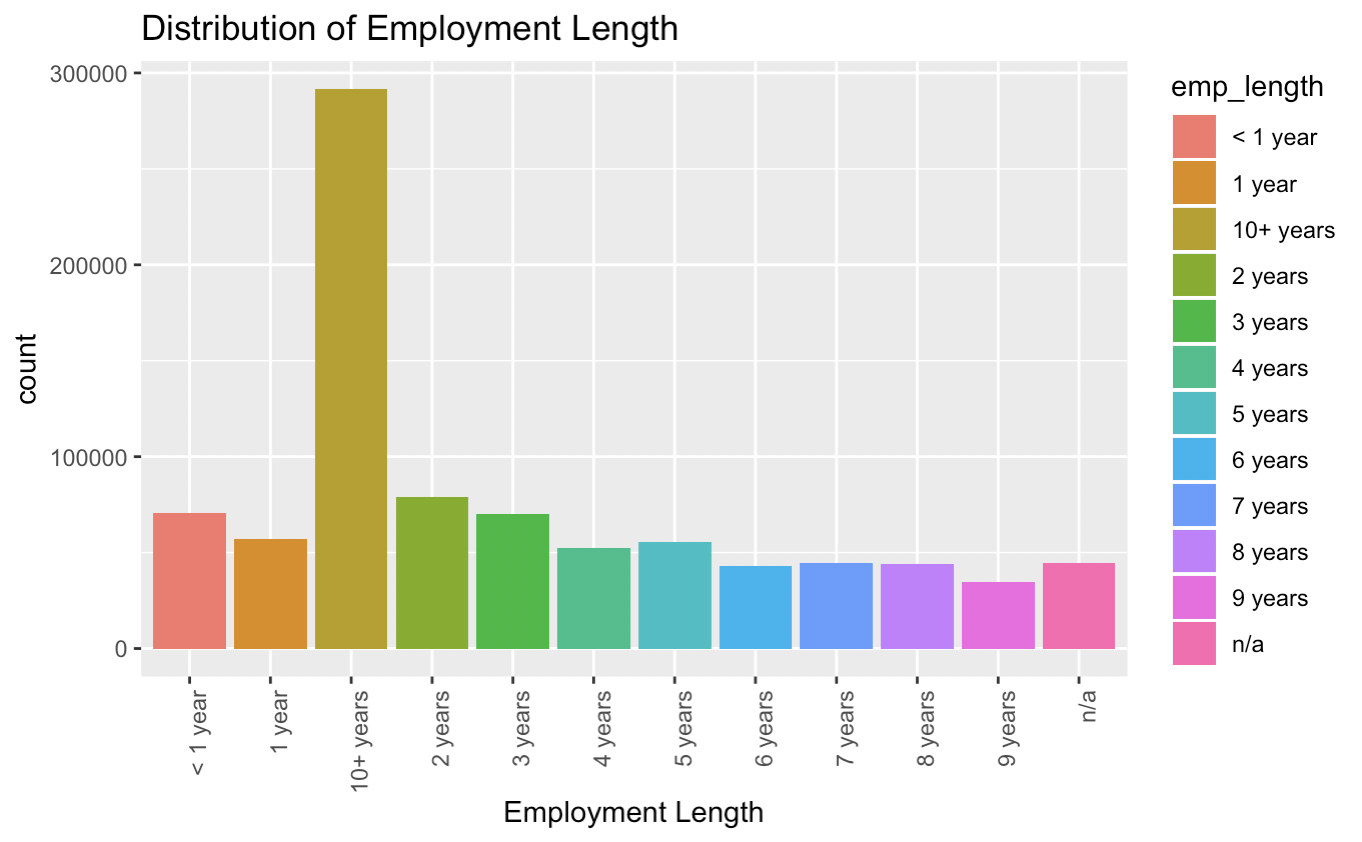
Debt consolidation means taking out a new loan to pay off a number of liabilities and consumer debts, generally unsecured ones. Debt consolidation is the process of combining all of our unsecured loans into single monthly payment.

People tend to borrow less loan for personal reasons such as house, wedding, car and education and are willing to take risk to borrow more to pay their credit card bills and to consolidate their debts.

****

From the above graph, we infer that people from California have borrowed nearly 15% of the total volume of loans across United states. Then followed by people of New York who have borrowed nearly 7.5% of the total volume of loan.

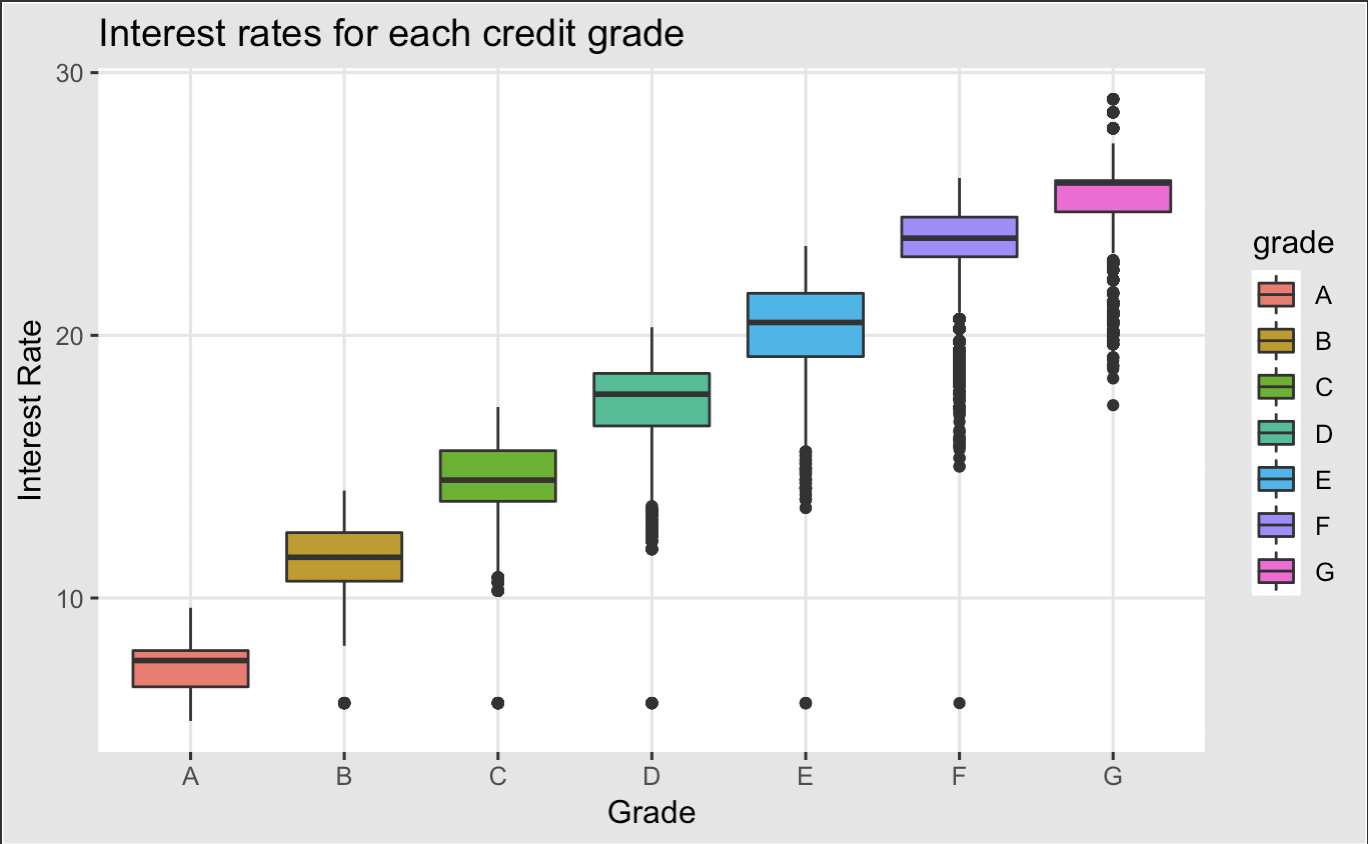
Business opportunities in states like California, New York and Texas are more when compared to states like Nevada, Utah and Idaho. Huge population and high cost of living could also be vital factors for high volume of loan amount borrowed by people of California and New York.



From this plot, it would be reasonable to assume that people with more years of experience earn more and less dependent on loans. But the data tells us a different story. For instance, people with more than 10 years of experience have borrowed more than people with relatively less years of experience.

We can understand that people who have just started their career are less likely to apply for loans. As they progress in their career and as their income increases, they are willing to borrow more loans. Work experience is an important factor in assessing the credit risk history.

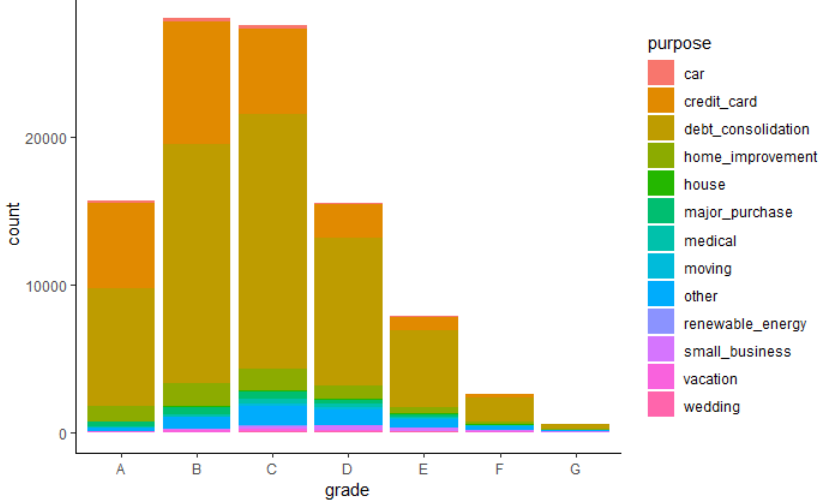
It is coherent to assume that people with more years of experience have good credit history when compared to people who have just started their career.



Grades represent the quality of credit history of the borrowers. Better the credit history, higher is the grade assigned to them. For example, people with the best credit history are grouped in Grade A and they sum to 150k of 887k population. We can infer from the plot that a poor grade F & G have a higher interest rate when compared with A grade which is the highest credit rating.

In this dataset, the credit history scores of clients are grouped as Grade A, B, C, D, E, F and G. The interest rate plays as an important feature in assessing the credit history grade of the clients. People of grade A pay the lowest interest rate because of very good credit history.

Whereas people of grade G pay an high interest rate because of poor credit history.



The above chart explains the distribution of various purposes across all grades. People with good credit history scores are grouped in group A and group B. The major reason for borrowing loan from these group are credit\_card and debt\_consolidation.

Grade B and Grade C are the most common categories of all the grades listed in the dataset. Even people with good credit history tend to borrow less loan on personal reasons such as vacation, wedding, car, house and education

**Data Preprocessing**

Out of 74 features, 22 features have more than 20% of null values. These features are not suitable to be selected as predictors for our model as we would be sub-setting a sample form the data which can lead to sampling errors.

As the objective of modelling is to predict the grade of the proposed loan, modelling any information related to principal payment and recoveries for ongoing loans will not be suitable; also, there were many variables like description of the loan, member id, etc which won’t contribute to the model. Thus, we removed all such columns from the scope of modelling.

The elimination of these features brought down the scope of the analysis to 27 variables. After this we conducted a qualitative assessment and research based on which we concluded that following 11 predictors had the highest impact on estimating the risk associated with a loan:

|  |  |
| --- | --- |
| **Features** | **Description** |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| 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. |
| emp\_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. |
| installment | The monthly payment owed by the borrower if the loan originates. |
| int\_rate | Interest Rate on the loan |
| loan\_amnt | 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. |
| revol\_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| verification\_status | Indicates if the borrower’s income was verified by LC, not verified, or if the income source was verified |
| tot\_cur\_bal | Total current balance of all accounts |

**Feature Scaling**

Since the features that we considered are of different scale, this will lead to a bias. In order to avoid the bias, we scale the features so that the corresponding units will be eliminated. We scale the data by a process called normalization.

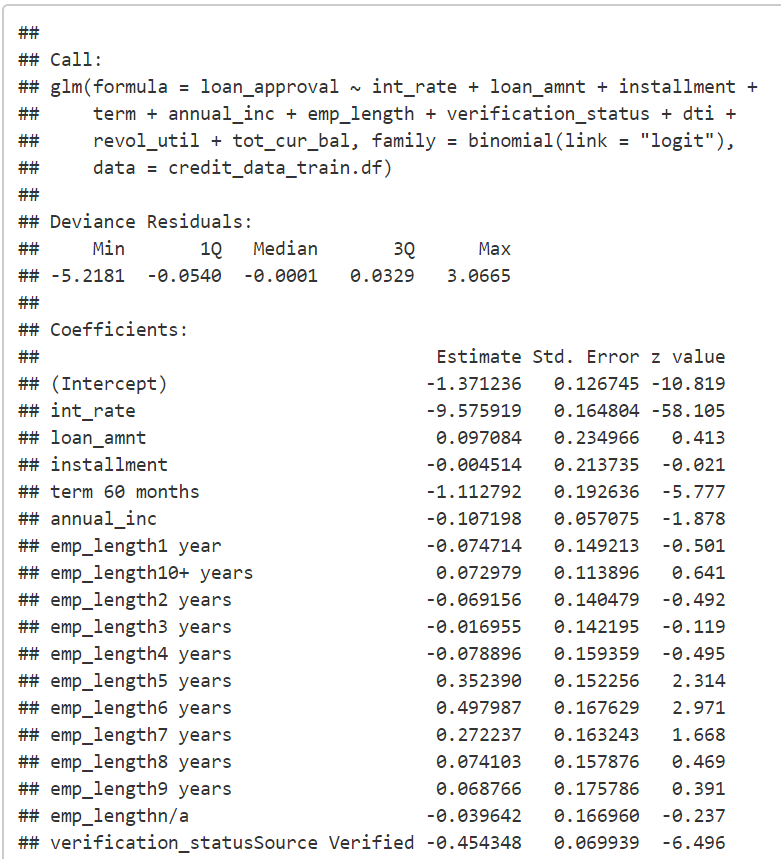
Normalization is a process of subtracting a record or an observation from it’s mean and dividing this value by the Standard deviation.

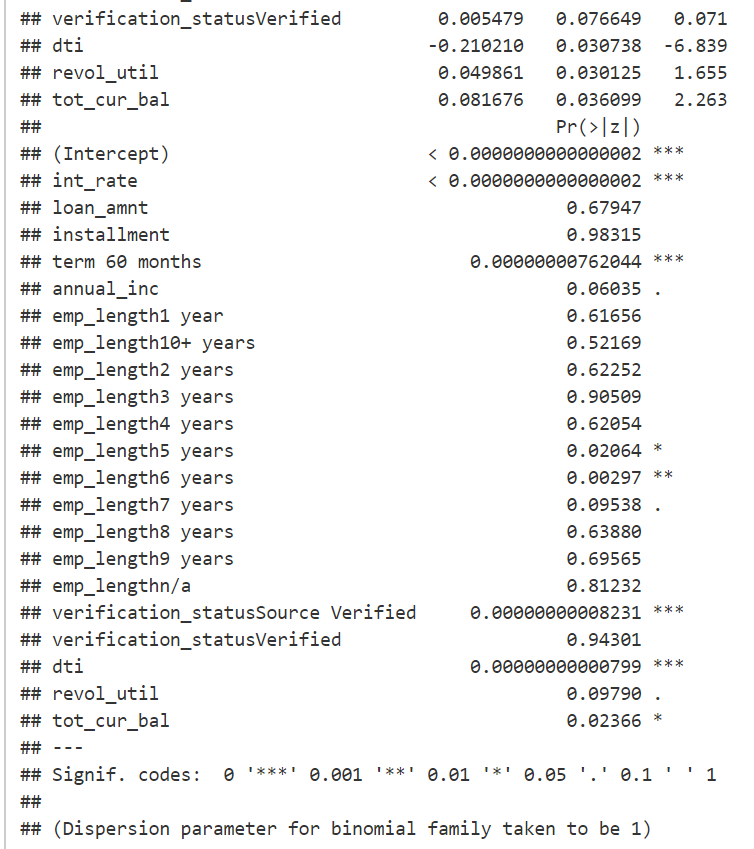
**Data Modelling**

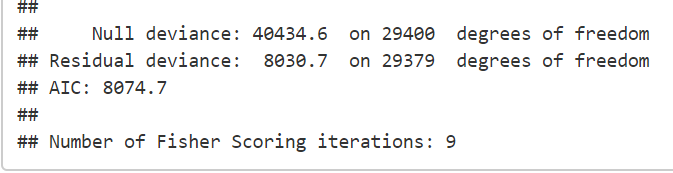
1. Logistic Regression

Logistic Regression is a regression model which is used when the dependent variable Y is categorical. For example, if Y denotes buying/selling/holding a house, we have three categories present in the dependent variable. In this scenario logistic regression can be used to classify the record/observation into one of the classes based on the values of the predictor variables. This regression model helps to establish relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Summary

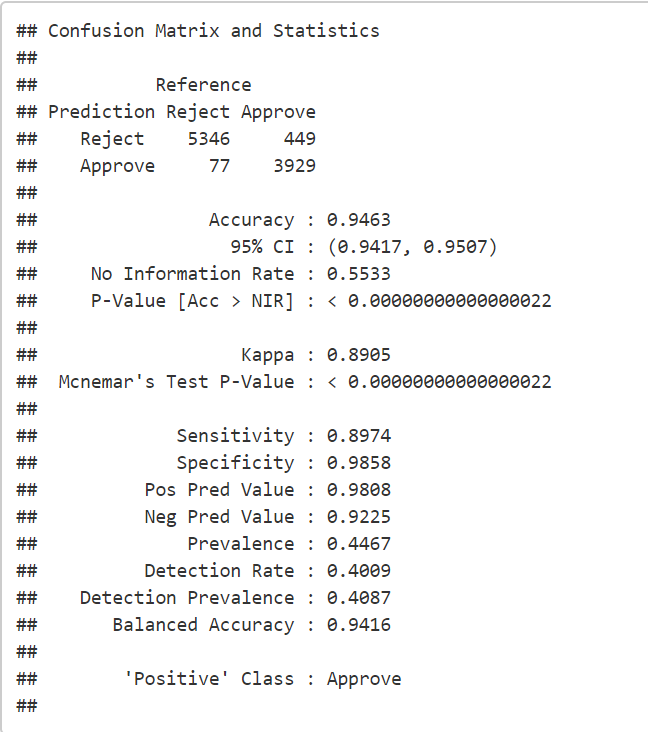






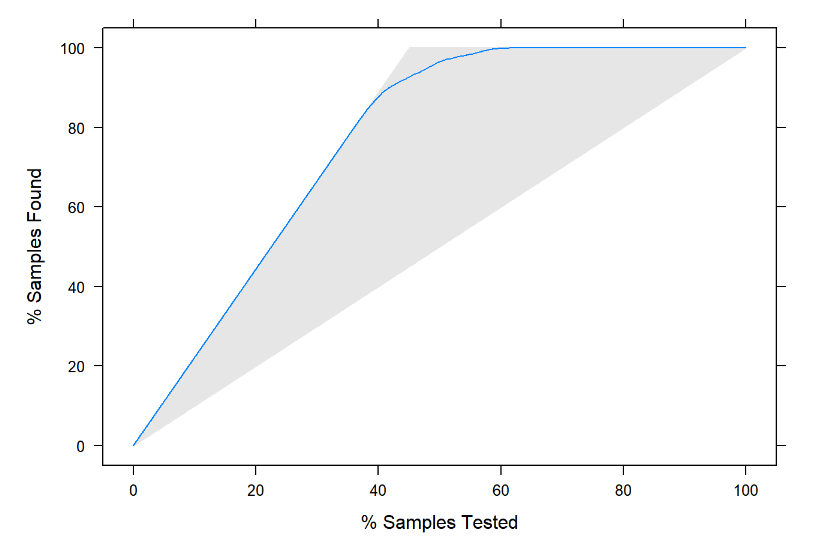
From the summary, we can infer that Interest Rate and the term significantly affect the risk in a negative way, whereas employment length and whether the financial status of the borrower was verified significantly affect the risk in a positive way.

Confusion Matrix



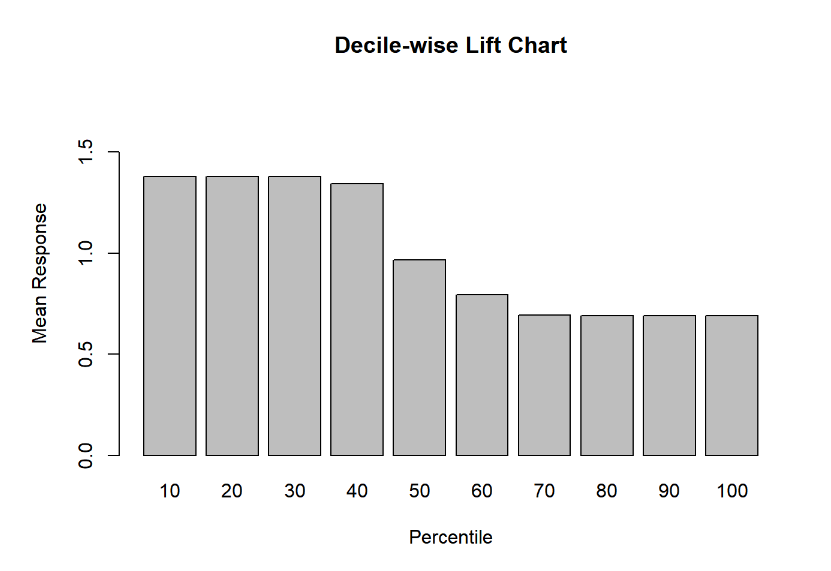
Running logistic regression on the validation data with a threshold of 0.7 for approval probability gave an accuracy of 94.63 %, with a specificity of 98.58 %; however, sensitivity was low at 89.74%.

Lift chart



It can be clearly seen that the model has a very good lift compared to the naïve rule; for 40 % of the sample tested, naïve rule is able to correctly classify 40 % of the data, whereas logistic regression is able to correctly classify about 85 % of the data.

Decile chart

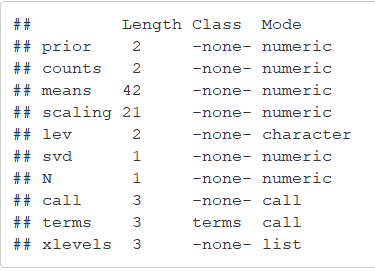


For the first decile, logistic regression model has performed about 1.4 times better than the naïve model; at the fifth decile, it starts performing like the naïve model and then drops further.

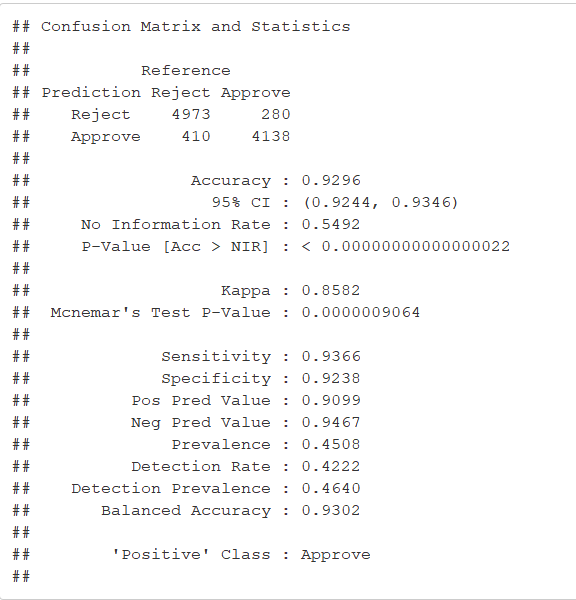
1. Linear Discriminant Analysis

LDA is a classification method like Logistic regression and can be used for classification and profiling. It a method to find linear combination that will best classify the observations into different classes/groups with increased separation between them. Some of the popular use cases involve classifying customers in potential loan buyers and non-buyers, credit cards.

Summary

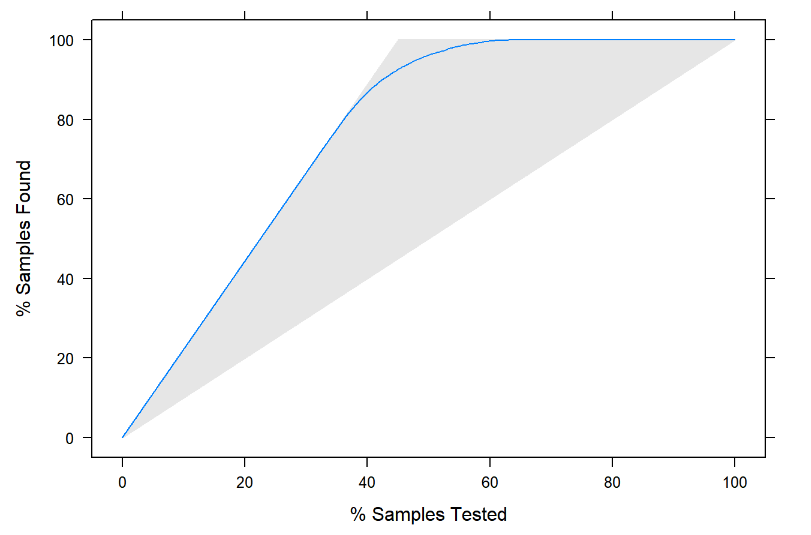


Confusion Matrix



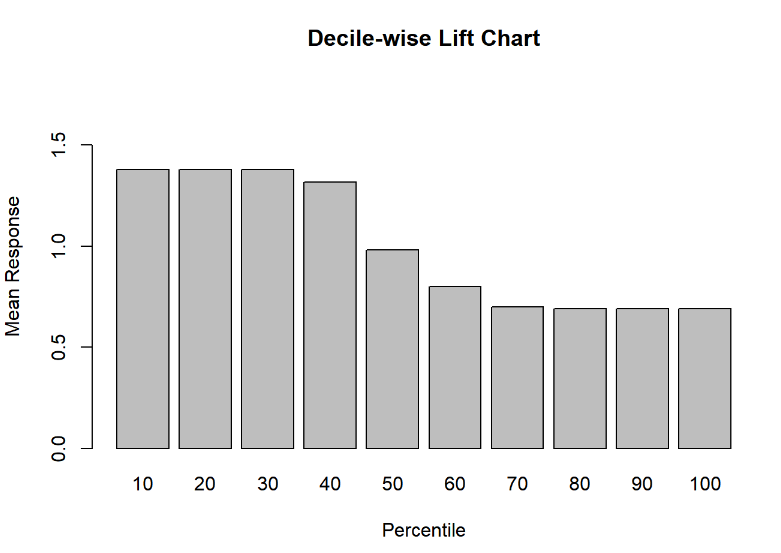
Running linear discriminant analysis on the validation data gave an accuracy of 92.96 %, with a specificity of 92.38 % and improving the sensitivity to 93.66% compared to the logistic regression model.

Lift chart



It can be clearly seen that the model has a very good lift compared to the naïve rule; for 40 % of the sample tested, naïve rule is able to correctly classify 40 % of the data, whereas logistic regression is able to correctly classify about 82 % of the data.

Decile chart

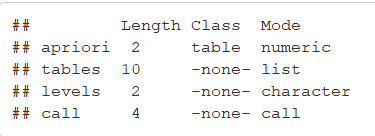


For the first decile, linear discriminant analysis model has performed about 1.4 times better than the naïve model; at the fifth decile, it starts performing like the naïve model and then drops further.

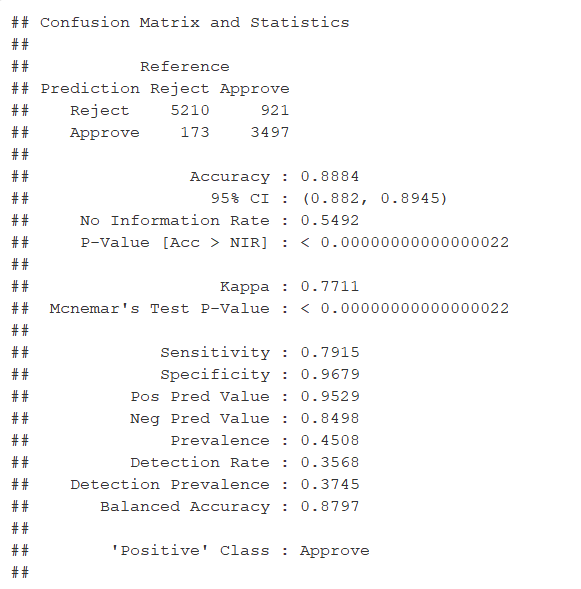
1. Naïve Bayes classifier

Naïve Bayes classifier algorithm is used for classification problems. It is primarily used for text classification datasets. A few examples are spam filtration, sentimental analysis, and classifying new articles. This algorithm learns the probability of object belonging to a grouping class. It’s named naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.

Summary

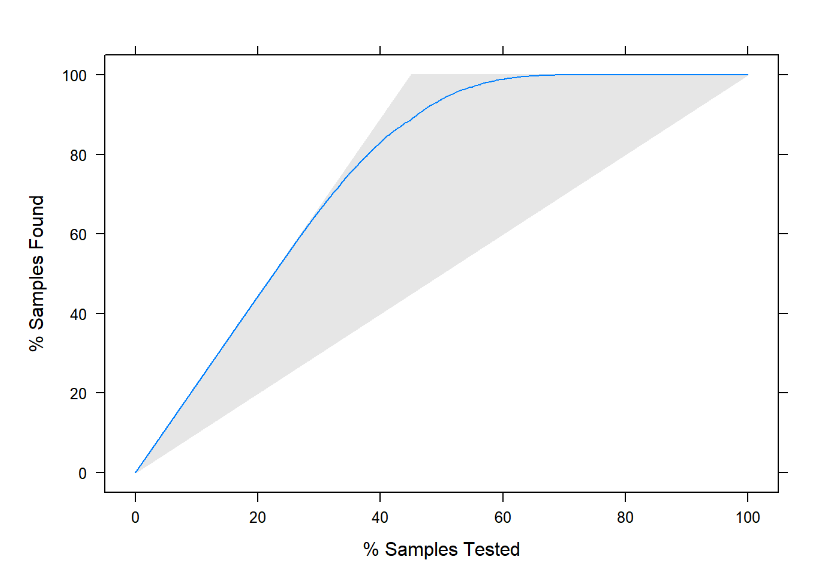


Confusion Matrix



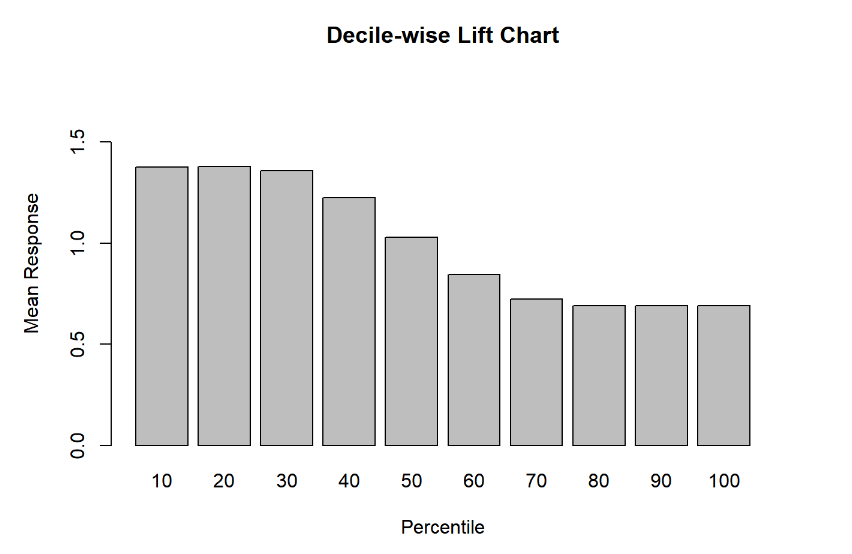
Running naïve Bayes classifier on the validation data gave lower accuracy of 88.84 %, with a specificity of 96.79 %; however, the sensitivity of the model was bad at about 79.15%.

Lift chart



It can be clearly seen that the model has a very good lift compared to the naïve rule; for 40 % of the sample tested, naïve rule is able to correctly classify 40 % of the data, whereas logistic regression is able to correctly classify about 80 % of the data.

Decile chart

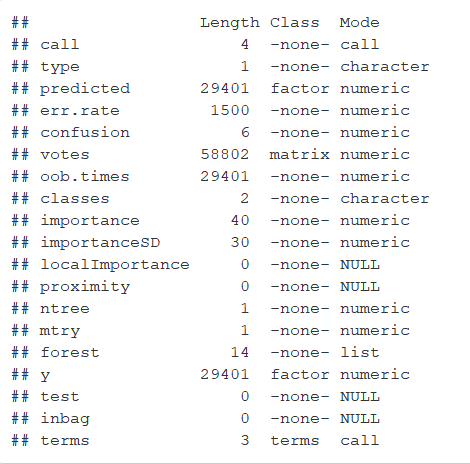


For the first decile, naïve Bayes model has performed about 1.4 times better than the naïve model; it then follows a step-wise pattern and at the fifth decile, it starts performing like the naïve model and then drops further.

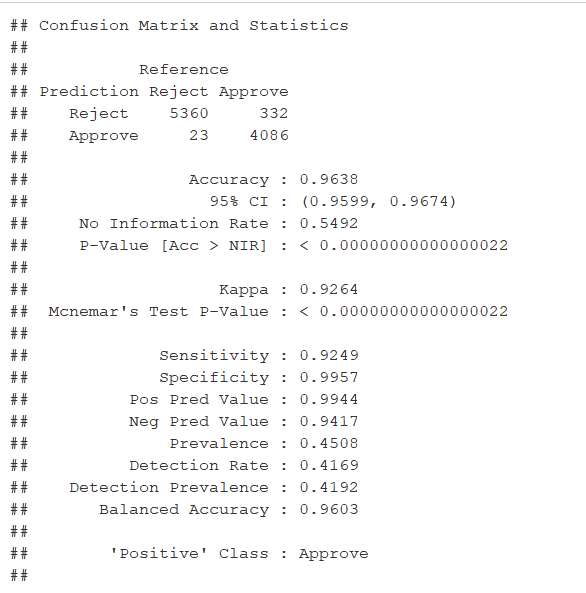
1. Random Forest classifier

It builds multiple decision trees from the different subsets of the same data set and merges them together to get a more accurate prediction. Each decision tree in the forest considers a random subject of features when forming questions and only must access to a random set of training data points. This deals to increase in the diversity of the forest which in turn produces robust predictions. When it comes time to make a prediction, the random forest takes an average of all the individual decision tree estimates. One of the main advantages of using Random Forest is that it’s used for both classification and regression problems. Some of the use cases of this algorithm is found in remote sensing & object detection.

Summary

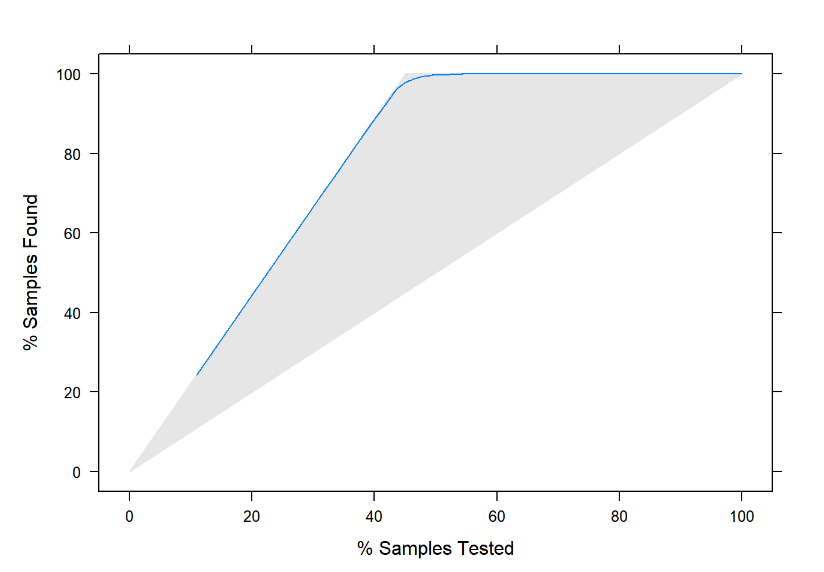


Confusion Matrix



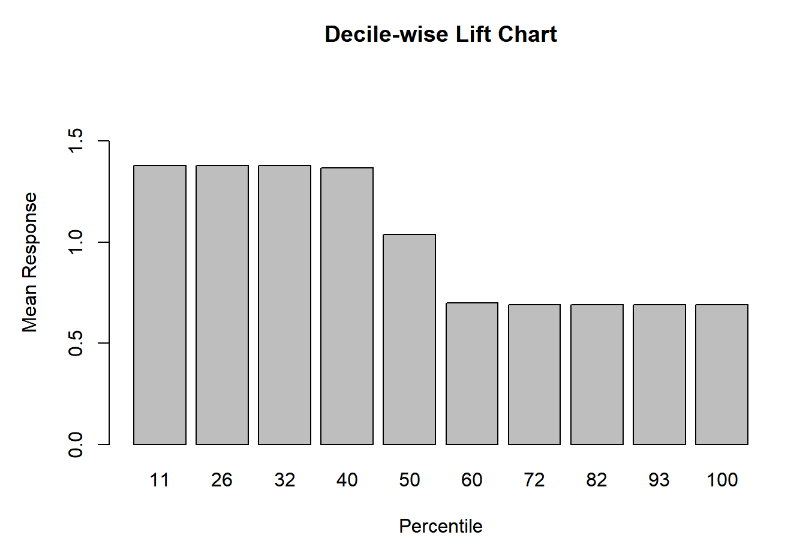
Running random forest classifier on the validation data gave the highest accuracy of 96.38 %, with a specificity of 99.79 %; also, the sensitivity achieved was decent at about 92.49 %.

Lift chart



It can be clearly seen that the model has a very good lift compared to the naïve rule; for 40 % of the sample tested, naïve rule is able to correctly classify 40 % of the data, whereas logistic regression is able to correctly classify about 90 % of the data.

Decile chart



For the first 4 deciles, random forest model has performed about 1.4 times better than the naïve model; it then follows a step-wise pattern and at the fifth decile, it starts performing like the naïve model and then drops further.

**Conclusion**

Based on the analysis of the confusion matrices, decile and lift charts of all 4 models that we ran, we can conclude that the Random Forest Classifier has performed better than all the other models that we ran as it gives the best combination accuracy, specificity and sensitivity along with the best lift.

The accuracy of the Random Forest classifier is 96.38%.