PREDICTING LOAN OUTCOMES USING MACHINE LEARNING

A project to predict loan status for the Lending Club dataset which will help enhancing the loan-approval pre-check system.

Submitted By
Mayanka Jha

Supervised By
Prof. Ravi Pandey

Introduction

Technology holds great promises for financial companies who have interests in financing personal loans but often face risks from customers who default. With the explosion of machine learning and predictive analytics, the outcome of these loans can be predicted using the information collected from applicants during the preliminary loan application. This will help lenders in making informed decision while investing their money and increase profitability. It will also help in faster processing of loans without much manual intervention and ease the entire process for both the customers and the lenders.

This paper applies machine learning algorithms to predict loan outcome status of a loan for one such organizations, Lending Club. Lending Club provides a platform for the US citizens to ethically loan and borrow money. A person can invest their money in the pool and the system then links their money with various borrowers' applications. Being an investor often comes with its own challenges, "How to ensure if the loan will not default?".

Different machine learning algorithms are trained using the Lending Club 2016 Quarter 1 data and predicted the 2016 Quarter 2 loan status. I have tried

Lending Club

Lending club is the world's largest peer-to-peer lending platform which uses technology to create a credit marketplace at a lower cost than traditional bank loan programs. By connecting borrowers and investors directly and allowing them to invest in and borrow from each other, it avoids the cost and complexity of the banking system and passes the savings on to borrowers in the form of lower rates and to investors in the form of solid returns. The entire process is online and relies on technology to promote affordability of the credit over availability of the credit.

How does Lending Club works?

1. Customers interested in a loan complete a simple application at LendingClub.com
2. Lending Club leverage online data and technology to quickly assess risk, determine a credit rating and assign appropriate interest rates. Qualified applicants receive offers in just minutes and can evaluate loan options with no impact to their credit score
3. Investors ranging from individuals to institutions select loans in which to invest and can earn monthly returns

Borrowers apply for loans.
Investors open an account.
Borrowers get funded.
Investors build a portfolio.
Borrowers repay automatically.
Investors earn & reinvest.

Data Cleaning

The dataset has 145 columns and 133889 rows. After going through the data dictionary, 12 most significant columns have been chosen for data exploration. Rows with any null values have been dropped from the dataset.

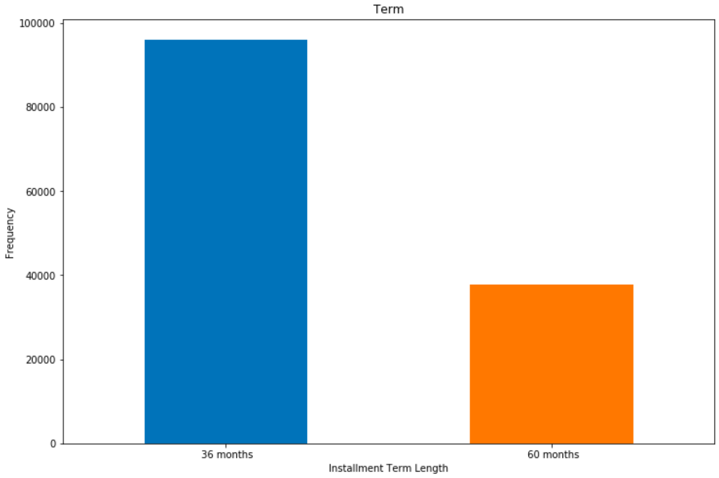
Single Variable Data Exploration

All 12 features have been individually explored to understand the dataset better.

Term

Term indicates the period for which loan is extended to the borrower, i.e. no of payments on the loan. The values of this field are in months and it’s either 36 months or 60 months.

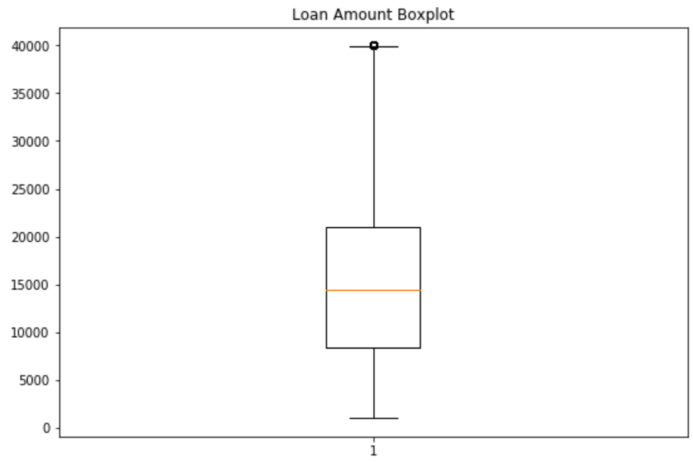
Term	Number of Loans
36 months	96120
60 months	37767

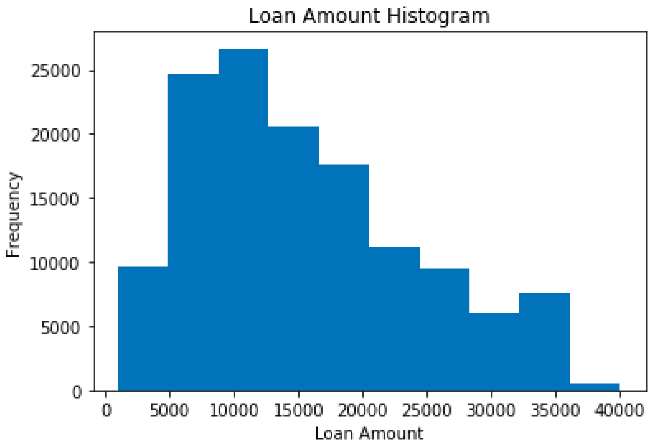


Loan Amount

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. In our dataset, the amount ranged from $1000 - $40,000 and the maximum number of loans were borrowed in the $7,500-$20,000 range.

```
count    133887.000000
mean      15589.394041
std        8858.198349
min        1000.000000
25%        8400.000000
50%       14400.000000
75%       21000.000000
max       40000.000000
```

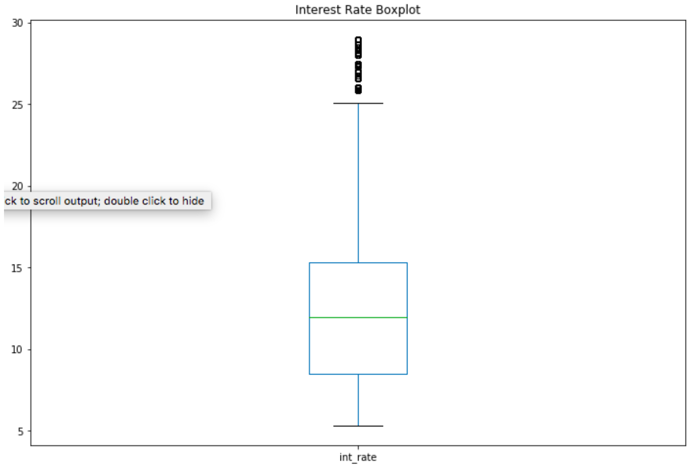


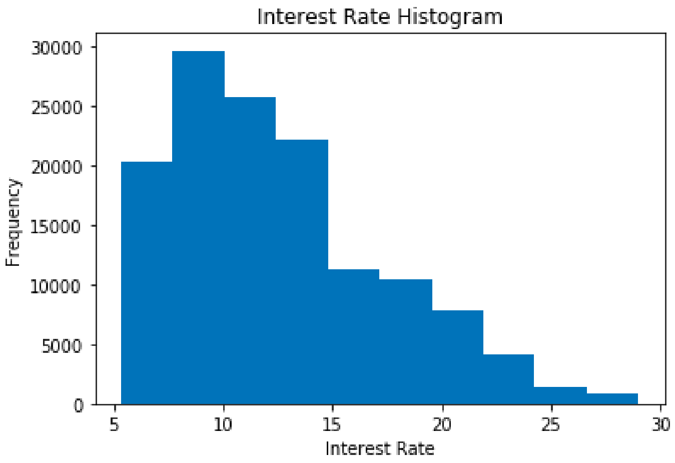


Interest Rate

Interest Rate is the rate of interest on the loan. As can be seen from the histogram, it is somewhat right skewed and even though the interest rate ranges from 5.32-28.9, a big part of the loans were extended for an interest rate up to 20%, with the average rate being 13.25%.

```
count    133887.000000
mean         12.476342
std           4.829203
min           5.320000
25%           8.490000
50%          11.990000
75%          15.310000
max          28.990000
```

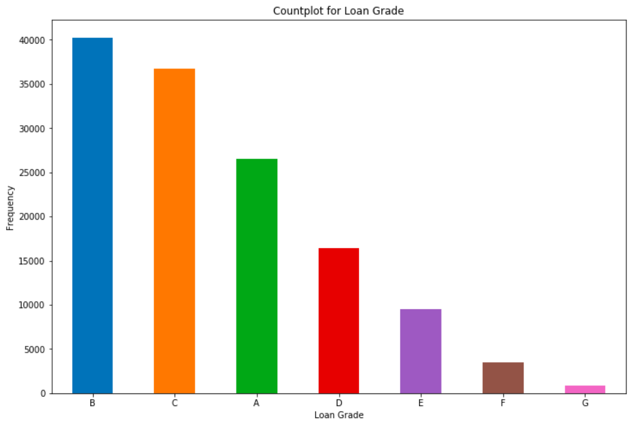




Grade

Grade field is the grade assigned to each loan by Lending Club. It reflects how likely the loan is to be paid off and is determined based on the creditworthiness of the borrower. The grades in the dataset range from A-G, B and C being the most assigned and very few G grade loans extended.

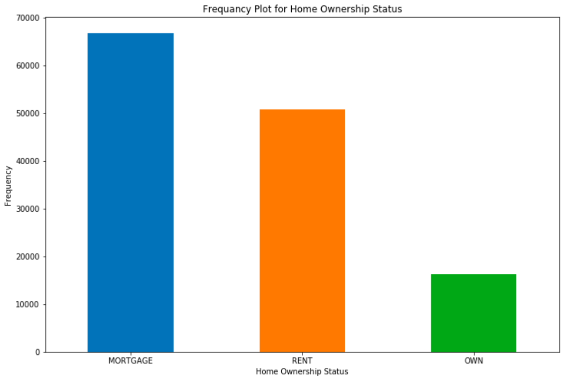
B	40267
C	36777
A	26482
D	16454
E	9540
F	3482
G	885



Home Ownership

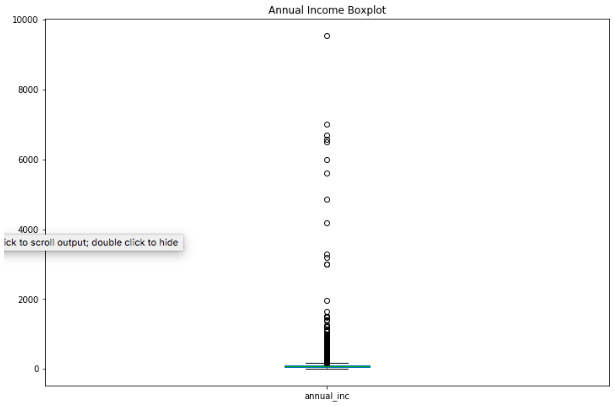
The home ownership status is provided by the borrower during initial application process and it can take be either RENT, OWN, MORTGAGE or OTHER. However, our data shows that none of the applicants selected ‘Other’. Also, approximately 50% of the loans were extended to people having an existing mortgage and those owning a house belong to the lowest 10%.

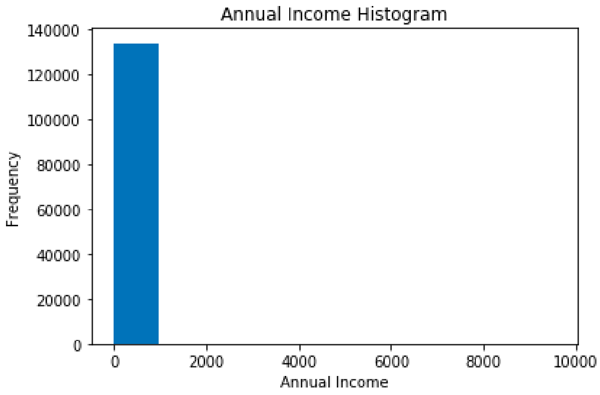
MORTGAGE	66829
RENT	50864
OWN	16194



Annual Income

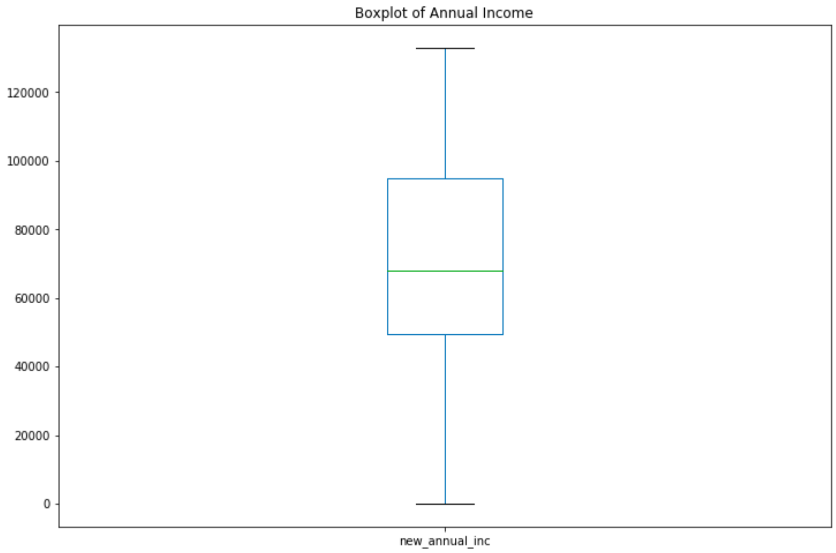
Applicants enter their annual income during the preliminary application process. In our dataset, the annual income ranges between $0- $1330,000. It can be observed from the histogram that the annual income is not normally distributed.





After removing outliers, we see the following distribution:

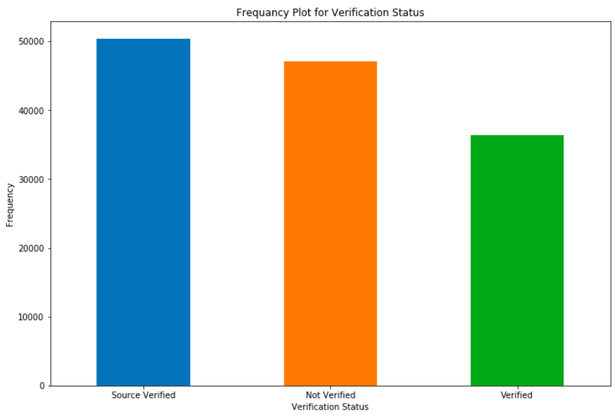
```
count   133887.000
mean     73676.920
std      32196.940
min          0.000
25%      49500.000
50%      68000.000
75%      95008.000
max     133000.000
```



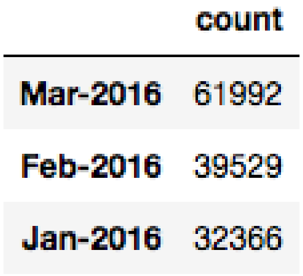
Verification Status

Indicates whether the co-borrower’s joint income has been verified by Lending Club. It is classified into ‘Verified’, ‘Not verified’ or ‘Source verified’. The frequency plot indicates that the number of loan status is uniformly distributed among the three categories with 37% being source verified and 30% not verified.

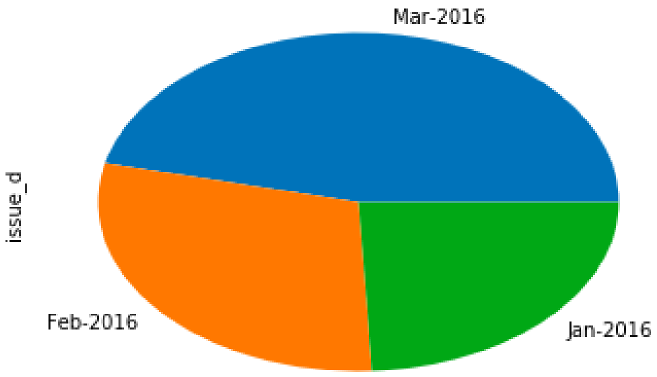
```
Source Verified    50425
Not Verified       47100
Verified           36362
```

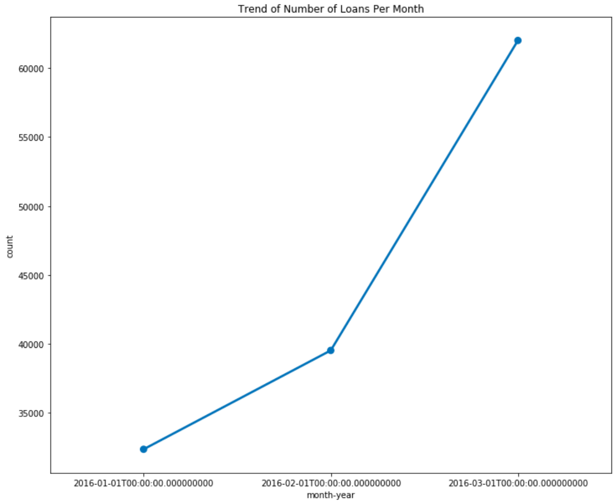


Issue Date



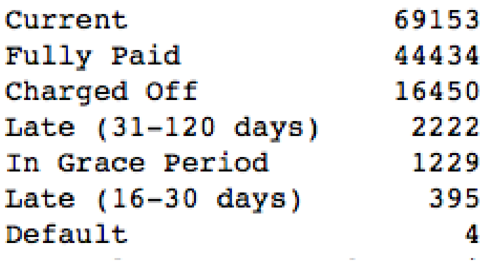
	count
Mar-2016	61992
Feb-2016	39529
Jan-2016	32366



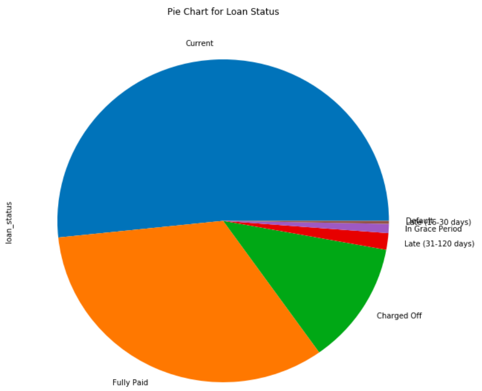


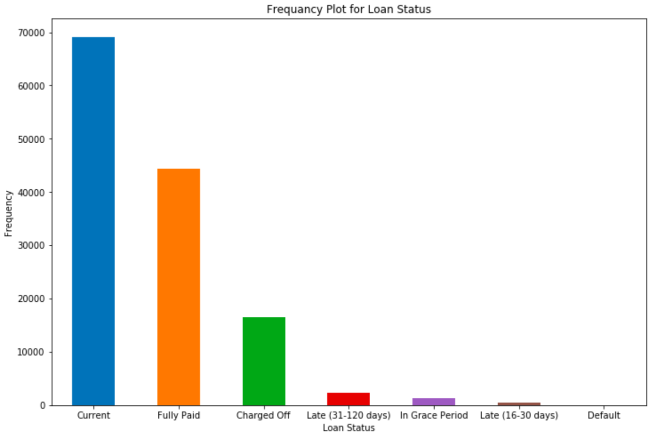
Loan Status

The term loan status means the current status of the borrower. There are 8 different categories a borrower can be placed under which are the following: charged off, current, default, fully paid, in grace period, issued, last (16-30 days), late (31-120 days). The largest number of borrowers are placed in the category “Current” followed by “Fully Paid”. As you can see in the diagram labelled “Frequency Plot for Loan Status”, our data is skewed to the right, 70 percent of borrowers are labelled as current status. “Charged off” status implies that loans for which are no longer a reasonable expectation of further payments. Generally, Charge Off occurs no later than 30 days after the Default status is reached.

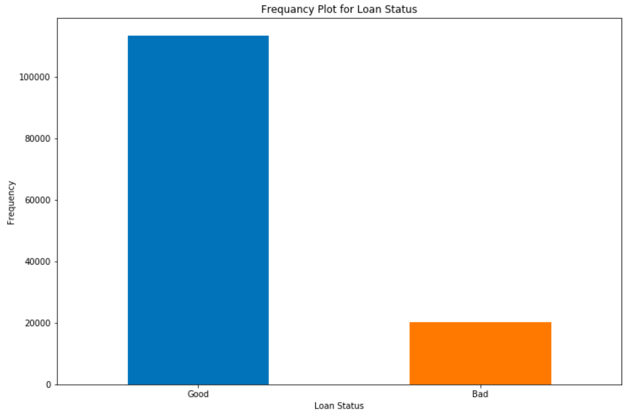


```
Current               69153
Fully Paid            44434
Charged Off           16450
Late (31-120 days)     2222
In Grace Period        1229
Late (16-30 days)       395
Default                   4
```





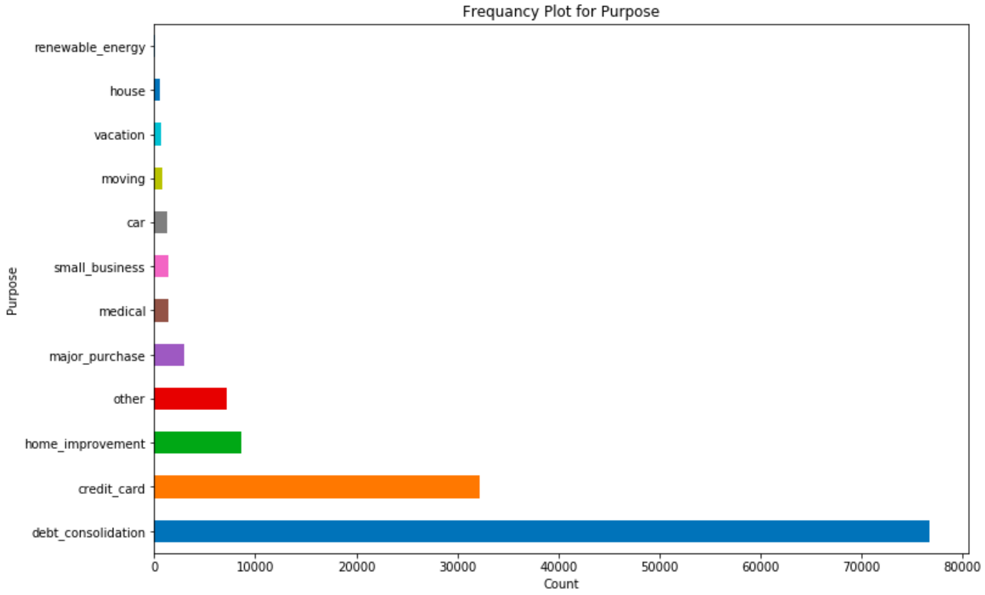
The status “Current” and “Fully Paid” are categorized as good loan whereas the rest of the categories are classified as bad loans. Accordingly, we will change the categories of loan status to “Good” and “Bad”.



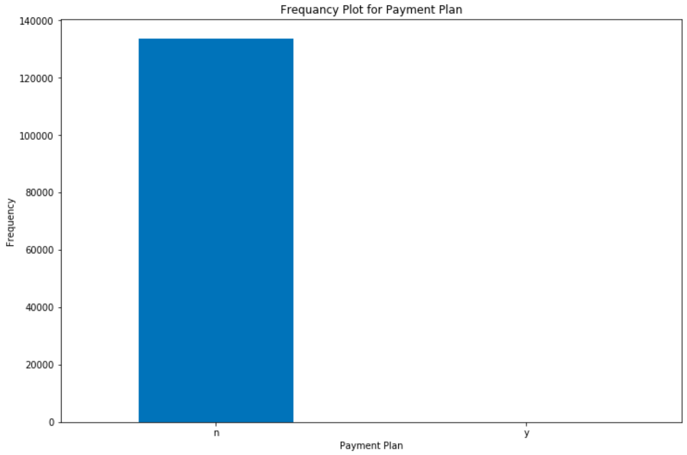
Purpose

Purpose is the reason for which applicant wants the loan and can take one of the four pre-specified categories: credit card, debt consolidation, home improvement, and others. Most of the applicants have chosen Debt consolidation as their purpose. 60 percent of people borrow money for debt consolidation.

23 percent of borrowers seek for a loan from Lending Club for credit card. 25 percent of borrowers seek loan for other type of reasons from Lending Club.

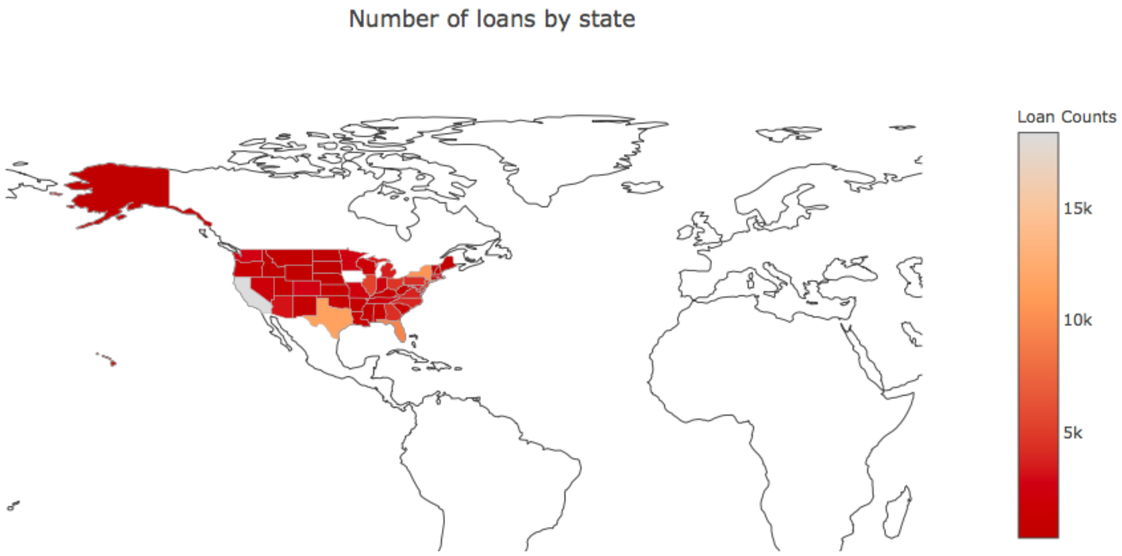


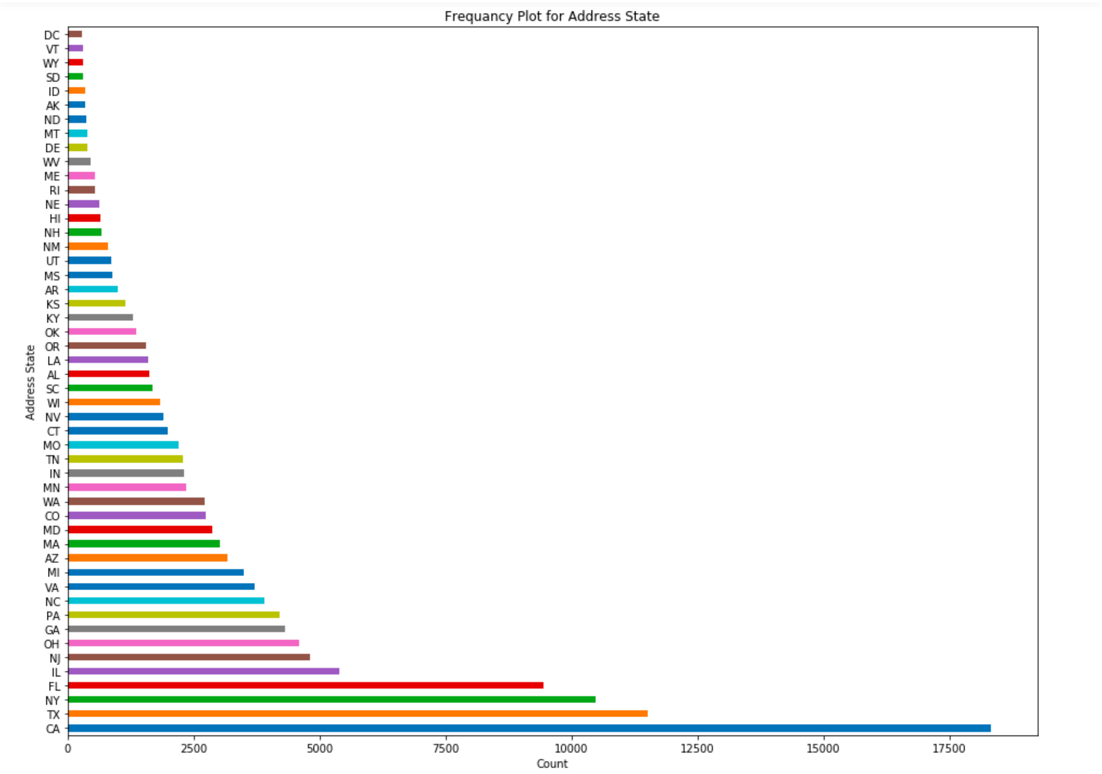
Payment Plan



State

The state information will help us in observing which state has the highest number of applicants. From the frequency plot, we can conclude that We can draw a conclusion that for the top four states California, New York, Texas, and Florida are metropolitan's states with more jobs to payback the loans they borrowed from Lending Club.

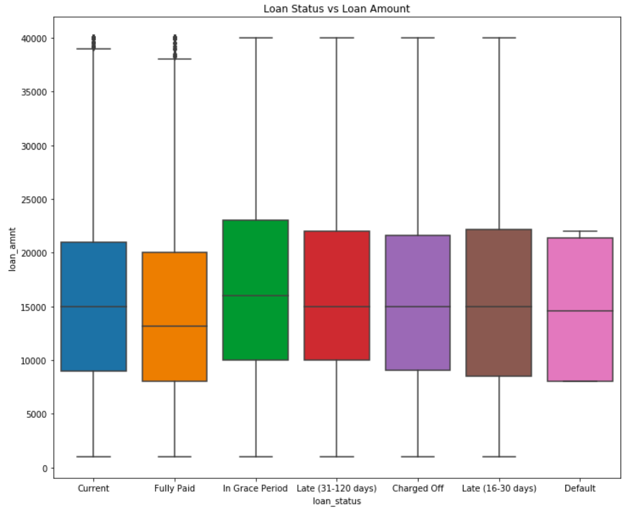

Number of loans by state
Loan Counts
15k
10k
5k


Frequancy Plot for Address State
Address State
DC
VT
WY
SD
ID
AK
ND
MT
DE
WV
ME
RI
NE
HI
NH
NM
UT
MS
AR
KS
KY
OK
OR
LA
AL
SC
WI
NV
CT
MO
TN
IN
MN
WA
CO
MD
MA
AZ
MI
VA
NC
PA
GA
OH
NJ
IL
FL
NY
TX
CA
0
2500
5000
7500
10000
12500
15000
17500
Count

Exploring Relationships: Multiple Variable Data Exploration

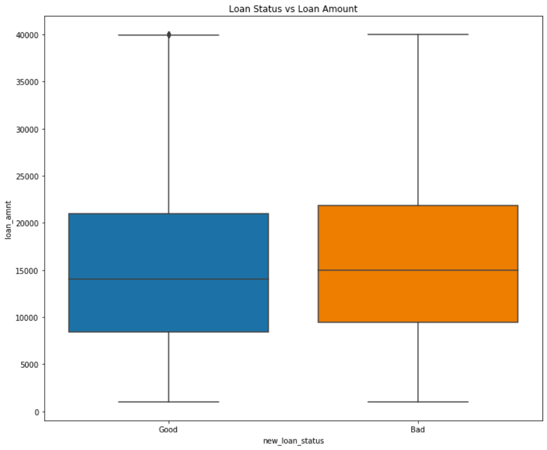
Loan Status vs Loan Amount

The relationship shown in this graph is between Loan Status and Loan amount. The lowest average loan amount has the loan status of Fully paid, which suggests that people are most likely to pay off the smaller loan amount. Similarly, the highest average loan amount has the status of loan as Grace period, so the loan amounts which are higher are in grace period and have not been fully paid.



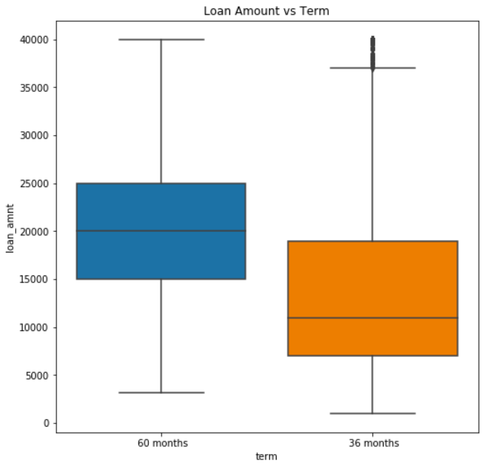
Loan Amount vs Loan Status (Good and Bad categories)

The relationship between Loan amount and Loan status shows that the loan amount does not affect whether that loan is good or bad, as the average loan amount is the same for both.



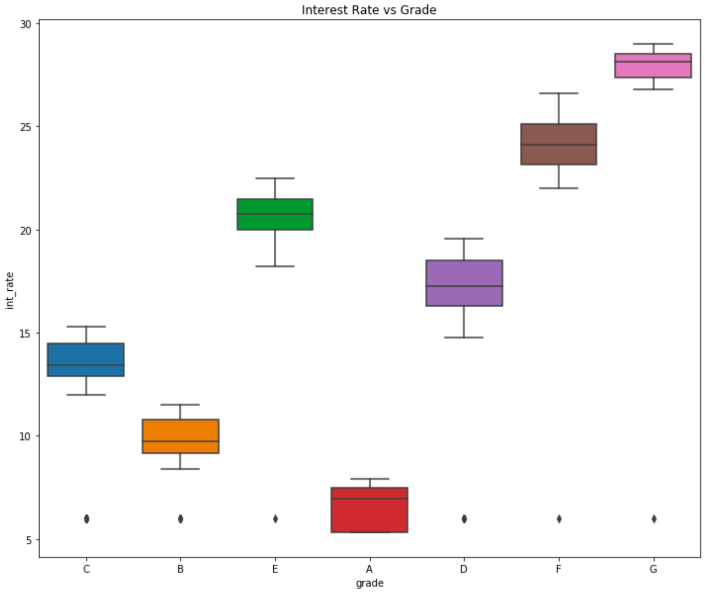
Loan amount vs Term

The relationship between Loan term and Loan amount suggests that if the term of loan is 36 months, the average loan amount will be less than the average loan amount if the term of loan is more, i.e. 60 months.



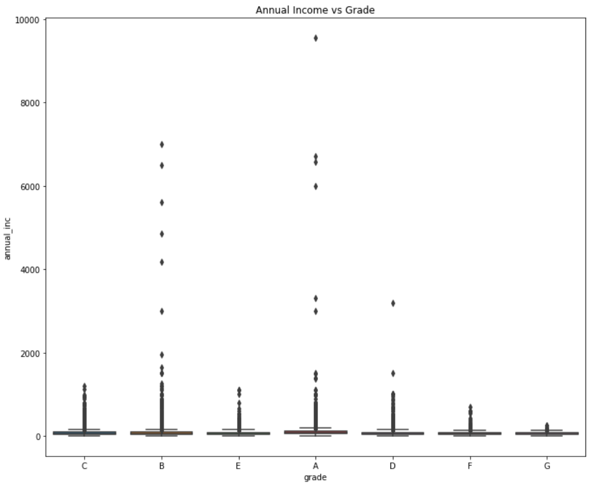
Interest Rate vs Grade

The relationship between Interest rate and Grade suggests that the Grade is the highest for loans with a lower rate of interest and lowest for the highest interest rate loans, grade G being assigned to loans with interest between 27-30 % and Grade A being assigned to loans with interest less than 7 %.

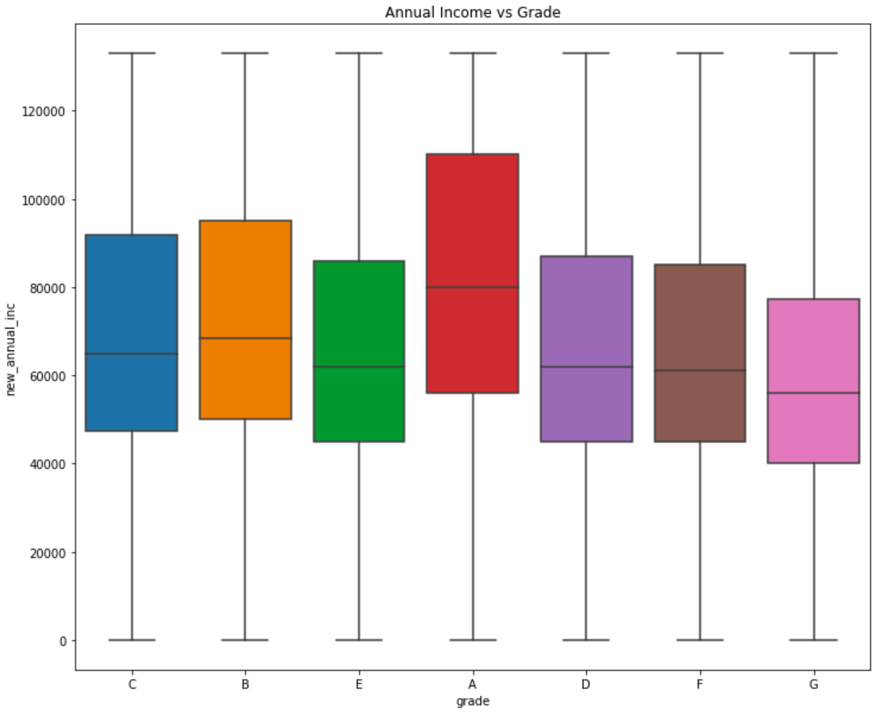


Annual Income vs Grade

Since, the annual income column has lot of outliers, it is difficult to make any observation about the relationship between annual income and grade.



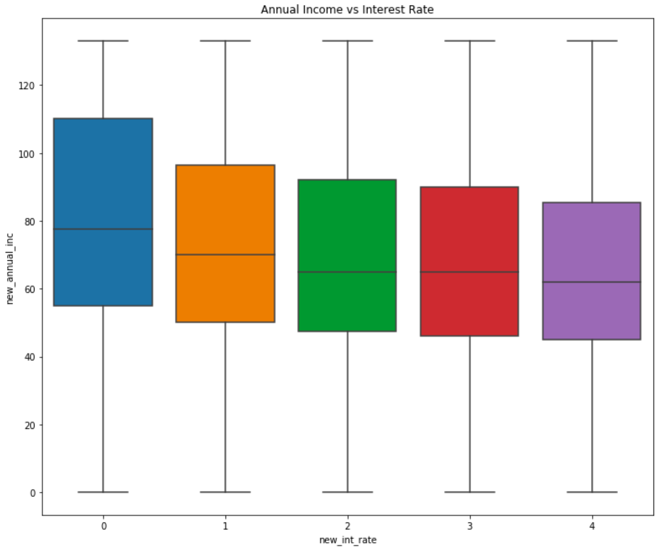
I have plotted the graph again after removing the outliers. The relationship between annual income and Grade suggests a high income corresponds to grade A, as a borrower with higher income is more likely to repay the loan. Grade D is assigned to a borrower with a median average income.



Annual Income vs Interest Rate

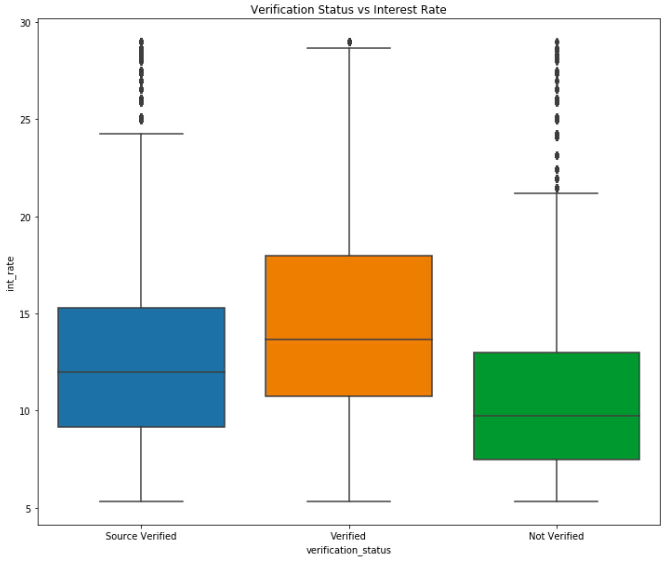
I have divided the interest rate in 5 buckets represented as follows:
0→ [0 to 5], 1→ [5 to 10], 2→ [10 to 15], 3→ [15 to 20] and 4 → [20 to 25].

The relationship between annual income and interest rate suggests that the lowest interest rate of 5- 10 % is given to the borrowers with highest average annual income.



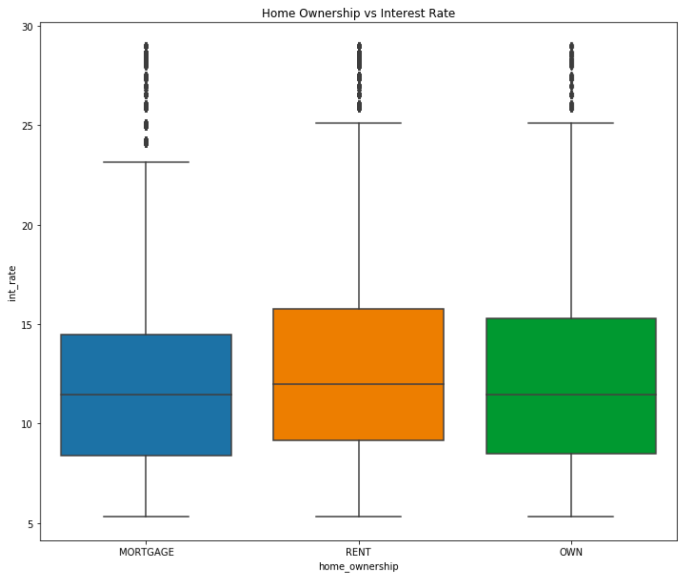
Verification Status vs Interest Rate

The relationship between verification state and interest rate suggests that the borrowers that have a verified income source are more likely to a get a loan on high interest rate. However, the lowest interest rate belongs the borrowers with not verified income source because their loan is most likely to be declined or they are borrowing small sums.



Home Ownership vs Interest Rate

The relationship between Home ownership status and interest rate suggests that the borrowers with a status of Mortgage are offered loan at the lowest interest rate and the ones with rent are offered the highest interest rate.



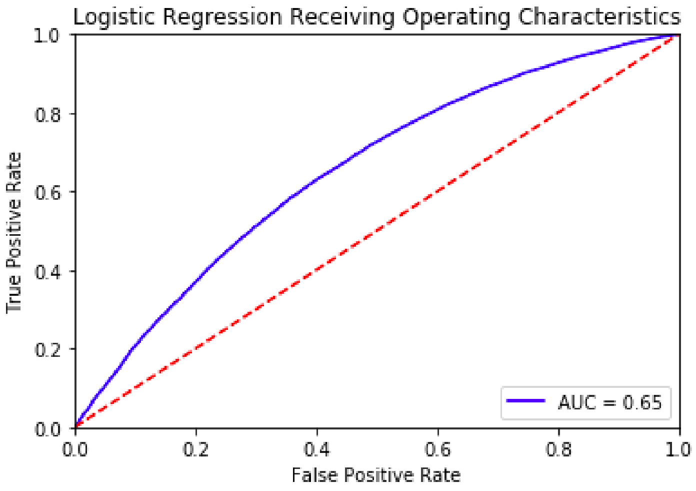
Predictive Modelling

Under Sampling

The data available to us is highly unbalanced with 113587 good loans and only 20300 bad loans. In such scenario, our model may not give us correct results. One of the methods to solve this problem is to under sample the good loans. I have used "NearMiss" algorithm for under sampling the data.

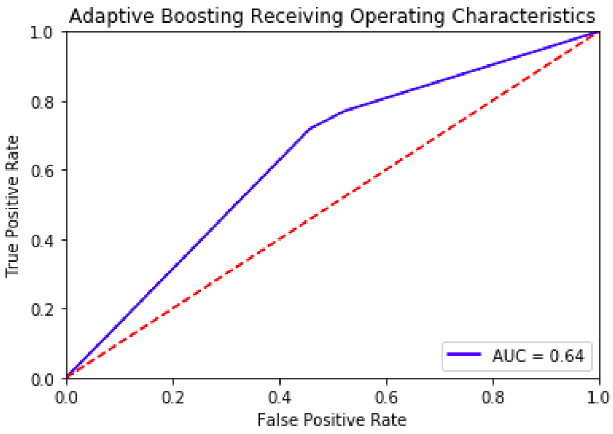
Logistic Regression

Accuracy Score: 0.50



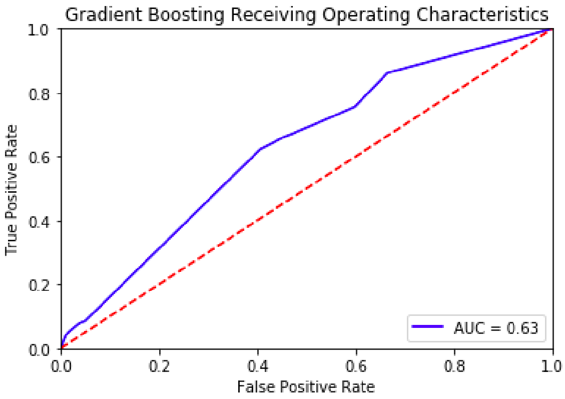
Adaptive Boosting

Accuracy Score: 0.73



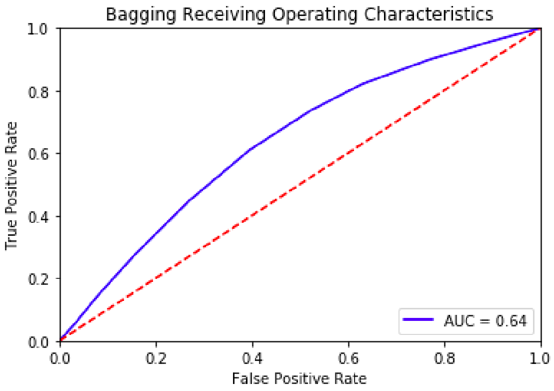
Gradient Boosting

Accuracy Score: 0.65



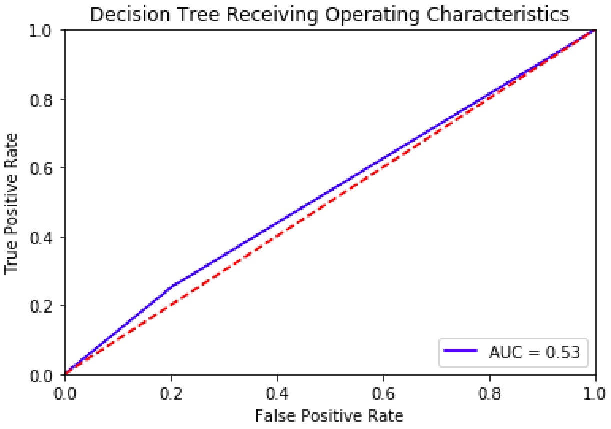
Bagging

Accuracy Score: 0.49



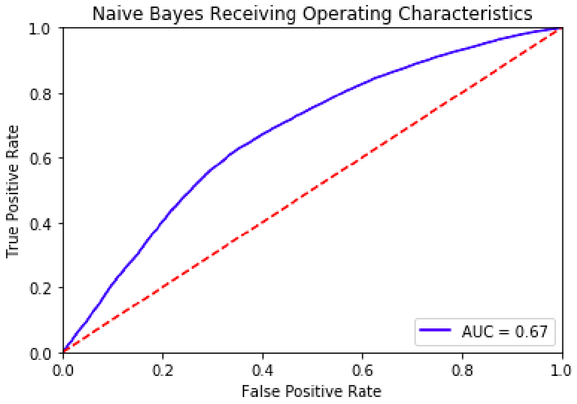
Decision Tree

Accuracy Score: 0.33



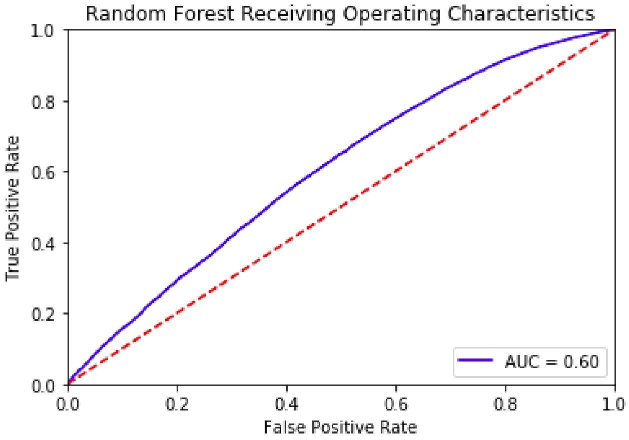
Naïve Bayes

Accuracy Score: 0.69



Random Forest

Accuracy Score: 0.4



Comparison of Models

Model Name	Accuracy Score	AUC (Area Under Curve)
Logistic Regression	0.50	0.65
Adaptive Boosting	0.72	0.64
Gradient Boosting	0.65	0.63
Bagging	0.48	0.64
Decision Tree	0.34	0.53
Naïve Bayes	0.68	0.67
Random Forest	0.39	0.6

Performance Summary

In credit risk analysis, accuracy does not play a major role in analysing performance. Predicting a good loan as defaulted loan (bad loan) will have less impact on the business than predicting a defaulted loan as good.

The Area under curve (AUC) ROC plots the sensitivity and specificity at different cutoff points. It is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative example. It measures the classifiers skill in ranking a set of patterns according to the degree to which they belong to the positive class, but without actually assigning patterns to classes.

Hence, we are choosing AUC as the performance measures. We can observe from the table above that Naïve Bayes has the highest AUC of 0.67 followed by Logistic Regression, AUC = 0.65.

Future Work and Business Recommendation

Before making any business recommendation, I need to increase the performance of my machine learning model. In the next stage, I am going to understand the concept of interaction variable and also play with different parameter of the ML algorithms to enhance the AUC of models.