PROJECT: BANK FEARS LOANLINESS

" Predicting The Probability, That A Member Will Default ; Using Python"

By,

Chetan Jawale.

ABSTRACT

The purpose of the study is to predict the credit risk involved in granting loan. Many factors are at play in a lender's final decision on a mortgage loan. These factors are all analyzed during the underwriting process through specialized software programs.

In this project we had to will have to put yourselves in the shoes of a loan issuer and manage credit risk by using the past data and deciding whom to give the loan to in the future. The data set used for prediction is a structured data and we shall use a supervised learning technique for modelling. We shall use multivariate analysis such as Classification techniques to determine the status of Independent variables basis multiple predictor variables.

TABLE OF CONTENT

INTRODUCTION

OBJECTIVE OF STUDY

The purpose of the study is to predict the credit risk involved in granting loan. The underwriter has to devise a process that leads to a final loan approval or denial, which is determined by a professional underwriter

In general, whenever an individual/corporation applies for a loan from a bank (or any loan issuer), their credit history undergoes a rigorous check to ensure that whether they are capable enough to pay off the loan (in this industry it is referred to as credit-worthiness).

The issuers have a set of model/s and rule/s in place which take information regarding their current financial standing, previous credit history and some other variables as input and output a metric which gives a measure of the risk that the issuer will potentially take on issuing the loan. The measure is generally in the form of a probability and is the risk that the person will default on their loan (called the probability of default) in the future.

Based on the amount of risk that the issuer is willing to take (plus some other factors) they decide on a cutoff of that score and use it to take a decision regarding whether to pass the loan or not. This is a way of managing credit risk. The whole process collectively is referred to as underwriting.

SCOPE OF STUDY

The study has several implications and usage in Banking, Wealth management, Mortgage and loan Underwriting and Asset Management industries. The training model developed can be used by the underwriter to predict on the loan application. There are a few possible outcomes at this point. The loan can be approved outright or the lender may determine that conditions must be fulfilled before the application can be approved.

For example, you might be required to provide additional verification of income or conclude the sale of your current property. The loan might be denied if the borrowers do not meet underwriting requirements. If you are denied for a mortgage loan, the lender will send an explanation of the decision.

LIMITATIONS OF STUDY

The study is conducted on the dataset provided, hence

1. The insights drawn using model developed by training is only true to the subject of study.
2. The model is only true for the dependent variables used under study.
3. The feature selection is done basis our domain knowledge and understanding to increase the accuracy and keep Type I and Type II errors as optimal.
4. The dataset is of USA hence the model developed is true for that particular geography and economy.

LITERATURE REVIEW

Since the 2007 recession, the lending marketplace has changed a lot. Before that, the local bank was pretty much the only lender a small business owner could go to for a loan. Banks are still the most common lenders for small businesses, but the amount of capital going to small businesses fell significantly after the recession.

To fill in those gaps, alternative lenders have popped up to serve small businesses. With these new lenders, businesses have access to much more capital to finance their company's growth. However, with these new loan products come new application and underwriting requirements.

The term "underwriting" refers to the process that leads to a final loan approval or denial, which is determined by a professional underwriter. Many factors are at play in a lender's final decision on a mortgage loan. These factors are all analysed during the underwriting process through specialized software programs.

Application

Filing a formal application for the loan is the first step in the underwriting process. This generally includes submitting evidence of current income and current assets, along with estimates of existing debt obligations and a current credit score. Next, the property's value is determined by an appraiser and a title search is completed to ensure there are no liens against the property. After these steps, the loan can move to the underwriting phase.

Credit Review

Your credit score and history heavily affect whether you will be approved for a mortgage loan. Through underwriting, the complete credit report is analyzed. The type of credit you possess, the way you use it and any red flags are considered. The better your credit, the more likely you are to be approved. Every lender is different, but some are more lenient than others when it comes to a few late payments over the course of your credit history.

Income to Debt Ratio

Another factor analyzed in the underwriting process is your income-to-debt ratio. This is simply the amount of monthly expenses you have divided by the amount of monthly income. For example, your proposed mortgage payment is $1,200 and additional debts -- such as auto loans, student loans, and credit cards -- require monthly payments totaling

$500. If you make $5,000 a month, the ratio is determined by dividing $1,700 by $5,000, which equals 34 percent. The lower the ratio, the better. This shows the lender you have additional funds coming in each month and are not overextending yourself.

Income Verification

You will most likely be required to provide some type of income verification to the lender, such as an official pay stub showing your year-to-date earnings. This is generally enough proof if you work a typical job, receiving biweekly or weekly pay. If you have an unconventional job with varying income or you work on commission you may need other forms of verification. Accepted documents might include tax returns, bank statements and accounting records if you are self-employed.

Approval Decision

Once the underwriter has reviewed all the necessary information and documents, he will make a decision on the loan application. There are a few possible outcomes at this point. The loan can be approved outright or the lender may determine that conditions must be fulfilled before the application can be approved. For example, you might be required to provide additional verification of income or conclude the sale of your current property. The loan might be denied if the borrowers do not meet underwriting requirements. If you are denied for a mortgage loan, the lender will send an explanation of the decision.

METHOD OF STUDY

STEPS USED IN STUDY:

1. To understand the problem and objective
2. To understand the data and develop some business sense.
3. EDA, segmentation if required.
4. Data Cleaning and Preprocessing.
5. Feature engineering.
6. Splitting the data into training and testing datasets.
7. Model Building using several Algorithm analysis.
8. Training the model for train dataset and determining the accuracy.
9. Testing the model and predicting the values for independent variable.
10. Cross-validation and Accuracy score .
11. Final results, recommendations and plots/visualizations.
12. To draw insights or recommendations that you can give from the data which will help the business.

LANGUAGES AND TOOLS USED IN STUDY:

Parameters	Specifications	Version	Purpose
Language Used	Python	3.6	Open Source and has inbuilt packages for implementing machine learning algorithms
IDE used	Spyder	8.0	For Data modelling and Prediction of Accuracy score.
Tool Used	Tableau	10.1	For Data Visualization and graphs

PYTHON PACKAGES USED:

Pandas :-

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.

Numpy :-

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. Rather, it's an extra tool that provides a more streamlined way of working with numerical and tabular data in Python. You can use pandas data structures but freely draw on Numpy and Scipy functions to manipulate them

Sk-learn:-

Scikit-learn (formerly scikits.learn) is a free software machine learning library for the Python programming language.
It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Matplot lib:-

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.
It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

EXPLORATORY DATA ANALYSIS

PROBLEM UNDERSTANDING:-

In this project you will have to put yourself in the shoes of a loan issuer and manage credit risk by using the past data and deciding whom to give the loan to in the future. The text files contain complete loan data for all loans issued by XYZ Corp. through 2007-2015. The data contains the indicator of default, payment information, credit history, etc.

The data should be divided into train (June 2007 - May 2015) and out-of-time test (June 2015 - Dec 2015) data. You will have use the training data to build models/analytical solution and finally apply it to test data to measure the performance and robustness of the models.

ALGORITHM UNDERSTANDING:-

We have to build a data model to predict the probability of default, and choose a cut-off based on what you feel is suitable. Alternatively you can also use a modelling technique which gives binary output

Based on the data that is available during loan application, build a model to predict default in the future. This will help the company in deciding whether or not to pass the loan.

Since the outcome or independent variable is predicting the status of the loan being defaulted or not. It is a binary classification problem, hence the best suitable algorithm is Logistics Regression.

LOGISTIC REGRESSION:-

Definition:

Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.

Advantages:

Logistic regression is designed for this purpose (classification), and is most useful for understanding the influence of several independent variables on a single outcome variable.

Disadvantages: Works only when the predicted variable is binary, assumes all predictors are independent of each other, and assumes data is free of missing values.

```
from sklearn.linear_model import LogisticRegression
lr =  LogisticRegression()
lr.fit(x_train, y_train)
y_pred=lr.predict(x_test)
```

ABOUT THE DATASET:-

These files contain complete loan data for all loans issued through the 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. The data contains defaulters, successful payers and customers who were current during that time. To simplify the problem, customers under 'current' status have been considered as non-defaulters in the dataset.

Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others. The file is a matrix of about 855 thousand observations and 73 variables.

DATA DICTIONARY:-

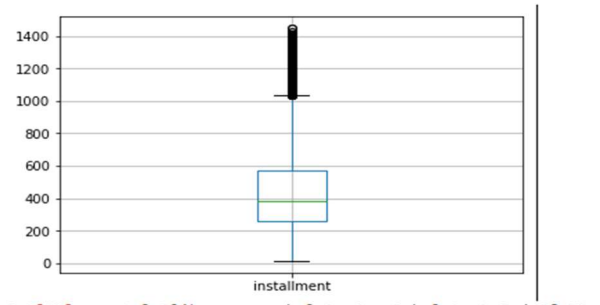
Variables	Description
addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower’s self-reported monthly income.
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested loan, divided by the co-borrowers' combined self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened

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.
emp_title	The job title supplied by the Borrower when applying for the loan.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	XYZ corp. assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
id	A unique assigned ID for the loan listing.
initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
last_credit_pull_d	The most recent month XYZ corp. pulled credit for this loan
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
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.
loan_status	Current status of the loan
member_id	A unique Id for the borrower member.
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record.
next_pymnt_d	Next scheduled payment date
open_acc	The number of open credit lines in the borrower's credit file.
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
policy_code	publicly available policy_code=1 new products not publicly available policy_code=2

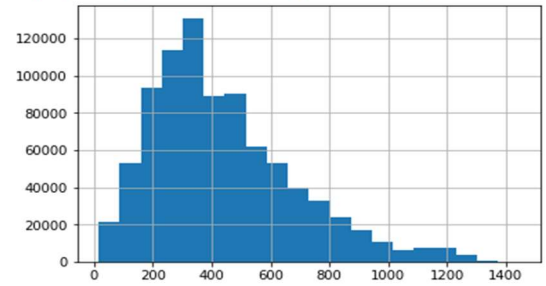
pub_rec	Number of derogatory public records
purpose	A category provided by the borrower for the loan request.
pymnt_plan	Indicates if a payment plan has been put in place for the loan
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	XYZ assigned assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower
total_acc	The total number of credit lines currently in the borrower's credit file
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
verified_status_joint	Indicates if the co-borrowers' joint income was verified by XYZ corp., not verified, or if the income source was verified
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
open_acc_6m	Number of open trades in last 6 months
open_il_6m	Number of currently active installment trades
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
mths_since_rcnt_il	Months since most recent installment accounts opened
total_bal_il	Total current balance of all installment accounts
il_util	Ratio of total current balance to high credit/credit limit on all install acct
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
max_bal_bc	Maximum current balance owed on all revolving accounts
all_util	Balance to credit limit on all trades
total_rev_hi_lim	Total revolving high credit/credit limit
inq_fi	Number of personal finance inquiries
total_cu_tl	Number of finance trades
inq_last_12m	Number of credit inquiries in past 12 months

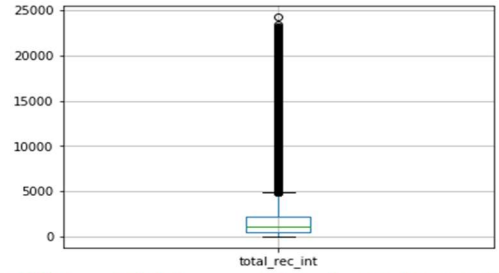
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
tot_coll_amt	Total collection amounts ever owed
tot_cur_bal	Total current balance of all accounts
verification_status	Was the income source verified

BOXPLOT AND HISTOGRAM:

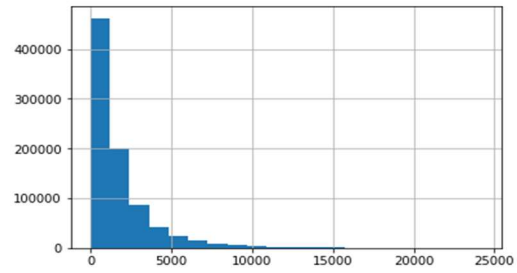


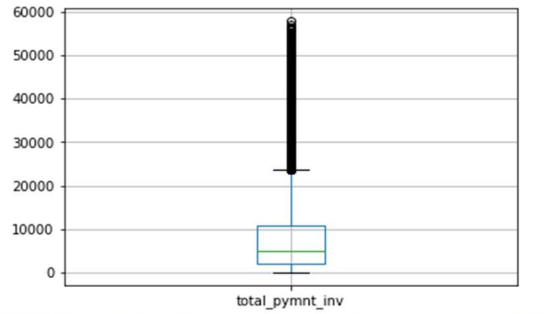

```
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0xda8fe10>
```



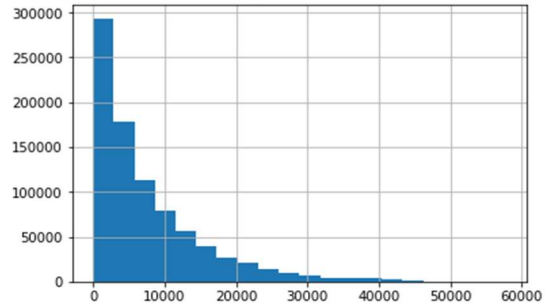


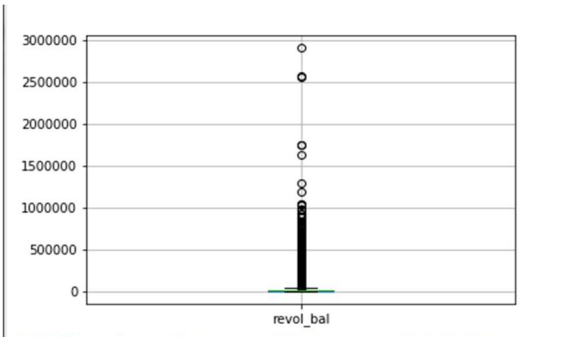

```
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0xddfb7f0>
```



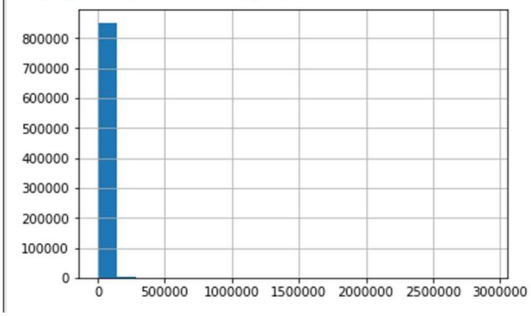



```
Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0xdf32c18>
```






```
Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0xdfc6278>
```



FEATURE SELECTION:

Attributes	Reason
id	Unique reference number which does not have any effect on the data set
member_id	Unique reference number which does not have any effect on the data set
funded_amnt_inv	It is a duplicate dataset most of the values are exactly the same to the attribute funded_amnt. The variance percentage is almost 11%
grade	It is a duplicate dataset to the variable sub grade
emp_title	This attribute is redundant to the study
pymnt_plan	The values in the attribute is biased as people opting for payment plan is quite less and it will be only part of test dataset.
desc	This attribute is redundant to the study
title	This attribute is redundant to the study
addr_state	This attribute is relevant to the study but we are eliminating the attribute because the attribute zip code will give more a clear picture.
inq_last_6mths	This attribute is redundant to the study
mths_since_last_record	Missing value more than 50%
initial_list_status	Does not have impact on the dataset
mths_since_last_major_derog	This dataset is not relevant to the study and missing value more than 50%
policy_code	This attribute is redundant to the study
application_Type	The values in the attribute is biased as people with individual application type is far more than joint application type.
annual_inc_joint	Missing values more than 50%
dti_joint	
verification_status_joint	
tot_coll_amt	
tot_cur_bal	
open_acc_6m	
open_il_6m	
open_il_12m	
open_il_24m	
mths_since_rcnt_il	
total_bal_il	
il_util	
open_rv_12m	
open_rv_24m	
max_bal_bc	
all_util	
inq_fi	
total_cu_tl	
inq_last_12m	

MISSING VALUES TREATMENT:

1. For all continuous variables which have missing values mean is used for treatment of missing values.
2. For all categorical variables mode or “frequency” is used for treatment of missing values.

```
#%%

# Imputing categorical missing data with mode value

colname1=['term','sub_grade','emp_length','home_ownership','verification_status',
        'issue_d','purpose','zip_code','earliest_cr_line','last_pymnt_d',
        'next_pymnt_d','last_credit_pull_d']
for x in colname1[:]:
   loan_df1[x].fillna(loan_df1[x].mode()[0],inplace=True)

# Imputing continuous or numerical missing data with mean value.

colname2=['mths_since_last_delinq','revol_util','collections_12_mths_ex_med',
        'total_rev_hi_lim']
for x in colname2[:]:
   loan_df1[x].fillna(loan_df1[x].mean(),inplace=True)
loan_df1.isnull().sum()
```

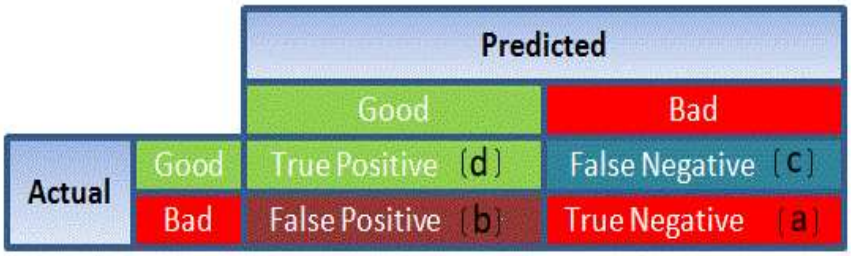
OUTLIER TREATMENT:-

Since the variables have r continuous numbers the outliers are not required to be treated in this dataset as the accuracy and confusion matrix is getting impacted.

DATA INTERPRETATION AND FINDINGS

PERFORMANCE OF LOGISTIC REGRESSION MODEL:

Confusion Matrix: It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. **False positive is Type I errors and False negative is Type II error**



You can calculate the accuracy of your model with:

$$\frac{\text{True Positive + True Negatives}}{\text{True Positive + True Negatives + False Positives + False Negatives}}$$

From confusion matrix, Specificity and Sensitivity can be derived as illustrated below:

$$\left.\begin{aligned}\text{True Negative Rate (TNR), specificity} &= \frac{A}{A+B}\\ \text{False Positve Rate (FPR), } 1-\text{specificity} &= \frac{B}{A+B}\end{aligned}\right\}\text{sum to 1}$$

$$\left.\begin{aligned}\text{True Positive Rate (TPR), sensitivity} &= \frac{D}{C+D}\\ \text{False Negative Rate (FNR)} &= \frac{C}{C+D}\end{aligned}\right\}\text{sum to 1}$$

SOME SPECIFIC ATTRIBUTES

The **precision** is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

The **recall** is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

The **F-beta** score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.

The F-beta score weights recall more than precision by a factor of beta. beta == 1.0 means recall and precision are equally important.

The **support** is the number of occurrences of each class in Y.

FINDINGS AND INTERPRETATION:-

Post running the Logistic Regression Classifier we get the below values as we get confusion matrix as

179632	77048
13	298

Classification Report

Parameters	Precision	Recall	f1-score	Support	Accuracy	Sensitivity	Specificity
0.0 (Fully Paid)	1.00	0.70	0.82	256680	70%	70%	96%
1.0 (Late or Charged Off)	0.00	0.96	0.01	311			
Avg/Total	1.00	0.70	0.82	256991			

Adjusting the threshold manually to 0.35 then we get confusion matrix as

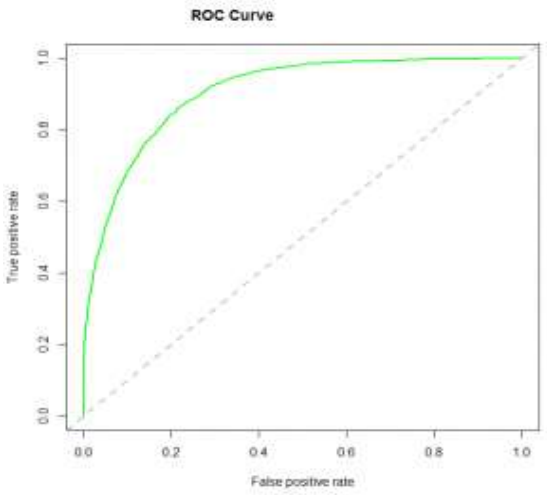
197652	59028
23	288

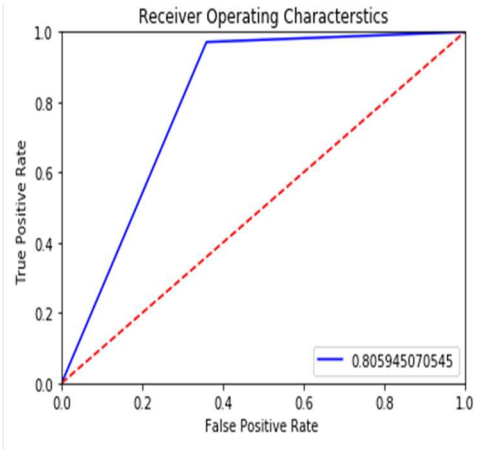
Classification Report

Parameters	Precision	Recall	f1-score	Support	Accuracy	Sensitivity	Specificity
0.0 (Fully Paid)	1.00	0.70	0.82	256680	77%	77%	93%
1.0 (Late or Charged Off)	0.00	0.96	0.01	311			
Avg/Total	1.00	0.70	0.82	256991			

Specificity and Sensitivity plays a crucial role in deriving ROC curve

ROC Curve: Receiver Operating Characteristic (ROC) summarizes the model's performance by evaluating the tradeoffs between true positive rate (sensitivity) and false positive rate(1- specificity). For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate. ROC summarizes the predictive power for all possible values of p > 0.5. 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. Below is a sample ROC curve. The ROC of a perfect predictive model has TP equals 1 and FP equals 0. This curve will touch the top left corner of the graph.





CONCLUSION

This dataset require lots of cleaning and feature engineering. The predictor variable with 0.35 threshold gives and accuracy of 77% for the model with optimal Type I and Type II errors.

FUTURE RECOMMENDATIONS

1. Improve data visualization so that fewer **good** loan data points are hidden under the **bad** loan data points.
2. Improve the prediction accuracy.
3. Developing a predictive tool for determining the date feature of the dataset data as data variables had missing values which cannot be replaced by mode.

REFERENCES

1. https://budgeting.thenest.com/mean-loan-goes-underwriting-23201.html
2. https://link.springer.com/content/pdf/bbm%3A978-1-4020-7898-9%2F1.pdf
3. https://www.fundera.com/business-loans/guides/business-loan-requirements
4. https://analyticsindiamag.com/7-types-classification-algorithms/
5. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/.
6. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8.
7. http://scikitlearn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html
8. https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/