**PROJECT REPORT**

**Loan Default Prediction using Machine Learning Techniques**

*Submitted towards the partial fulfillment of the criteria for award of PGA by Imarticus*

*Submitted By:*

*Nipsy James*

*Manjusha Gerish*

*Batch: November 2019*

# Abstract

**Keywords**

*Disclaimer: \*Data shared by the customer is confidential and sensitive, it should not be used for any purposes apart from capstone project submission for PGA. The Name and demographic details of the enterprise is kept confidential as per their owners’ request and binding.*

# Acknowledgements

We are using this opportunity to express my gratitude to everyone who supported us throughout the course of this group project. We are thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, we were fortunate to have **Mr.** **Kiran** as our mentor. They have readily shared their immense knowledge in data analytics and guide us in a manner that the outcome resulted in enhancing our data skills.

We wish to thank, all the faculties, as this project utilized knowledge gained from every course that formed the DSP program.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: Nov 15th, 2019 Member: Divyashree B G

Place: Bangalore

# Certificate of Completion

I hereby certify that the project titled “**XYZ bank lending and** **Loan Prediction for Default using Machine Learning Techniques**” was undertaken and completed by myself from the batch of DSP 18 (NOV 2019)

Mentor: Kiran

Date: Nov 10th, 2019

Place – Bangalore.

Table of Contents [Sample purpose only]

[Abstract 2](#_Toc9090473)

[Acknowledgements 2](#_Toc9090474)

[Certificate of Completion 3](#_Toc9090475)

[CHAPTER 1: INTRODUCTION 5](#_Toc9090476)

[1.1 Title & Objective of the study 5](#_Toc9090477)

[1.2 Need of the Study 5](#_Toc9090478)

[1.4 Business Model of Enterprise 5](#_Toc9090479)

[1.4 Data Sources 5](#_Toc9090480)

[1.5 Tools & Techniques 6](#_Toc9090481)

[1.6 Infrastructure Challenges 6](#_Toc9090482)

[CHAPTER 2: DATA PREPARATION AND UNDERSTANDING 6](#_Toc9090483)

[2.1 Phase I – Data Extraction and Cleaning: 6](#_Toc9090484)

[Decide On A Target Column 8](#_Toc9090485)

[2.2 Phase II - Feature Engineering 9](#_Toc9090486)

[Finding the correlation between variables 9](#_Toc9090487)

[2.3 Data Dictionary: 12](#_Toc9090488)

[2.4 Exploratory Data Analysis: 14](#_Toc9090489)

[Categorical Variables: 15](#_Toc9090490)

[CHAPTER 3: FITTING MODELS TO DATA 17](#_Toc9090491)

[4.1 Logistic REGRESSION MODEL 18](#_Toc9090492)

[4.1.1 First Logistic Regression Model: 18](#_Toc9090493)

[4.1.2 Second Logistic Regression model: 18](#_Toc9090494)

[4.1.3 Third Logistic Regression model: 19](#_Toc9090495)

[4.2 RANDOM FOREST 20](#_Toc9090496)

[CHAPTER 5: KEY FINDINGS 21](#_Toc9090497)

[CHAPTER 6: RECOMMENDATIONS AND CONCLUSION: 22](#_Toc9090498)

[CHAPTER 7: REFERENCES 23](#_Toc9090499)

# CHAPTER 1: INTRODUCTION

## Title & Objective of the study

The objective of our project is to predict whether a loan will default or not based on objective financial data only and whether investors should lend to a customer or not. Data from 2007-2015 will be used because most of the loans from that period have already been repaid or defaulted on.

## Need of the Study

In today’s world, obtaining loans from financial institutions has become a very common phenomenon. Every day many people apply for loans, for a variety of purposes. But not all the applicants are reliable, and not everyone can be approved. Every year, there are cases where people do not repay the bulk of the loan amount to the bank which results in huge financial loss. The risk associated with making a decision on a loan approval is immense. Hence, the idea of this project is to gather loan data from the Lending Club website and use machine learning techniques on this data to extract important information and predict if a customer would be able to repay the loan or not. In other words, the goal is to predict if the customer would be a defaulter or not.

## 1.4 Business Model of Enterprise

Financial lending is a way to borrow without using a traditional bank or credit union. For applicants with a good credit score (often a FICO credit score higher than 720), P2P loan rates can be surprisingly low. With less-than-perfect credit, an applicant still has a decent shot at being approved for an affordable loan with online lenders like XYZ corporation..

Financial loans are loans made by individuals and investors – as opposed to loans that come from a bank. People with extra funds offer to lend that money to others (individuals and businesses) in need of cash. A P2P service (such as a website) matches lenders and borrowers so that the process is relatively easy for all involved.

## 1.4 Data Sources

The provided dataset corresponds to all loans issued to individuals in the past from 2007-2015. The dataset has 855969 observations and 73 features. The data contains the indicator of default, payment information, credit history, etc. Customers under 'current' status have been considered as non-defaulters in the dataset. We have also been provided with a Data dictionary that best describes the features.

The dataset has quite a lot of missing values and the figures can be considered as ground truth, but lots of columns are either irrelevant, very sparse or non informative. Moreover, the dataset is unbalanced, with approximately 6% of loans considered as defaulted.

## 1.5 Tools & Techniques

Tools: Python 3.7.2, Jupyter Notebook, Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, Scipy

Techniques: Logistic regression, Random Forest Classifier, Gradient Boosting Classifier

# CHAPTER 2: DATA PREPARATION AND UNDERSTANDING

One of the first steps we engaged in was to outline the sequence of steps that we will be following for our project. Each of these steps are elaborated below

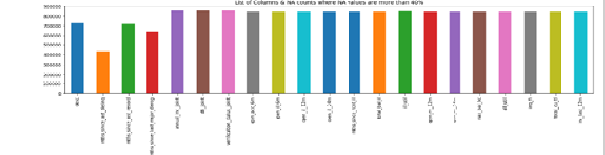
## 2.1 Phase I – Data Extraction and Cleaning:

* **Missing Value Analysis and Treatment**

In our dataset our target shows that 94% have not defaulted and 6% are defaulters or charged off. So this is clearly an unbalanced dataset.

The first issue was to know if the columns were filled with useful information or were mostly empty. Data exploration uncovered many empty or almost empty columns which were removed from the dataset because it would prove a difficult task to go back and try to answer for each data point that did not seem necessary at the time of the loan application.

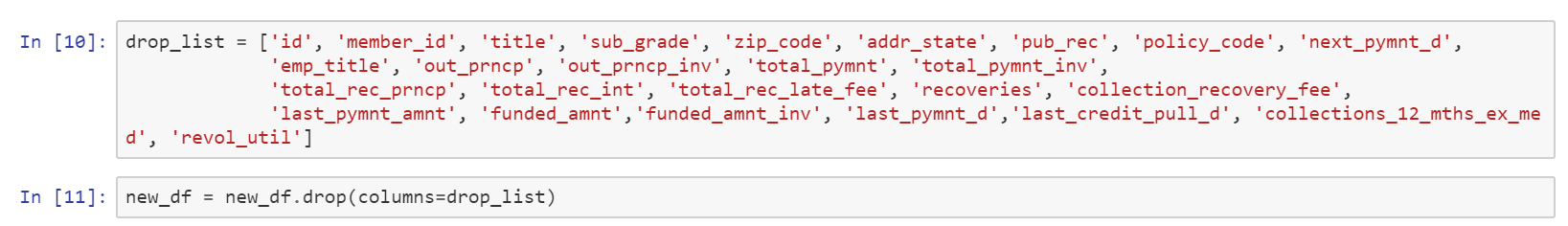
Our dataset has 855969 rows × 73 features including the target out of which 32 have missing values or NAN. Below we will look at a plot and get some insights.



Insights: So, we can see from the above plot that there are 20+ columns in the dataset where all the values are NA.

As we can see there are 855969 observations & 73 columns in the dataset, it will be very difficult to look at each column one by one & find the NA or missing values. So let's find out all columns where missing values are more than certain percentage, let's say 40%. We will remove those columns as it is not feasible to impute missing values for those columns.





Out of 73 features we only kept 53. So we removed about 21 features that had more than 40% missing values since it will not make any sense in further exploration.

We checked the variance and are removing low variance variables(excluding out target). However, since this is a sensitive problem we will remove based on our own discretion and business knowledge. Some irrelevant columns Unique ID's such as "id","member\_id" because they did not provide any useful information about the customer. As last 2 digits of zip code is masked 'xx', we can remove that as well.

* **Handling Outliers**

****

* **Feature Extraction**

### Decide On A Target Column

Now, let’s decide on the appropriate column to use as a target column for modeling – keep in mind the main goal is predict who will pay off a loan and who will default. We learned from the description of columns in the preview DataFrame that default\_ind is the only field in the main dataset that describe a loan status, so let’s use this column as the target column.

94% have not defaulted (fully paid) and 6% are defaulters or charged off.

**Purpose of loan** : Drop records where values are less than 0.75%

A screenshot of a cell phone

Description generated with very high confidence

We will analyze only those categories which contain more than 0.75% of records. Also, we are not aware what comes under 'Other' we will remove this category as well.

We are dropping columns that has more than 7 levels (categories) since it is not feasible to one hot encode them. The variable we removed are 'earliest\_cr\_line','addr\_state'. We're not removing issue\_d because the variable need to be used for splitting

## 2.2 Phase II - Feature Engineering

**Casting continuos variables to numeric:**

We have Cast all continuos variables that are necessary for our analysis to numeric so that we can find a correlation between them.

**Mapping:**

We are Mapping the issue date from "Jun-2015" to "Dec-2015" as Test for the ease of splitting our data to test set using Dictionary.

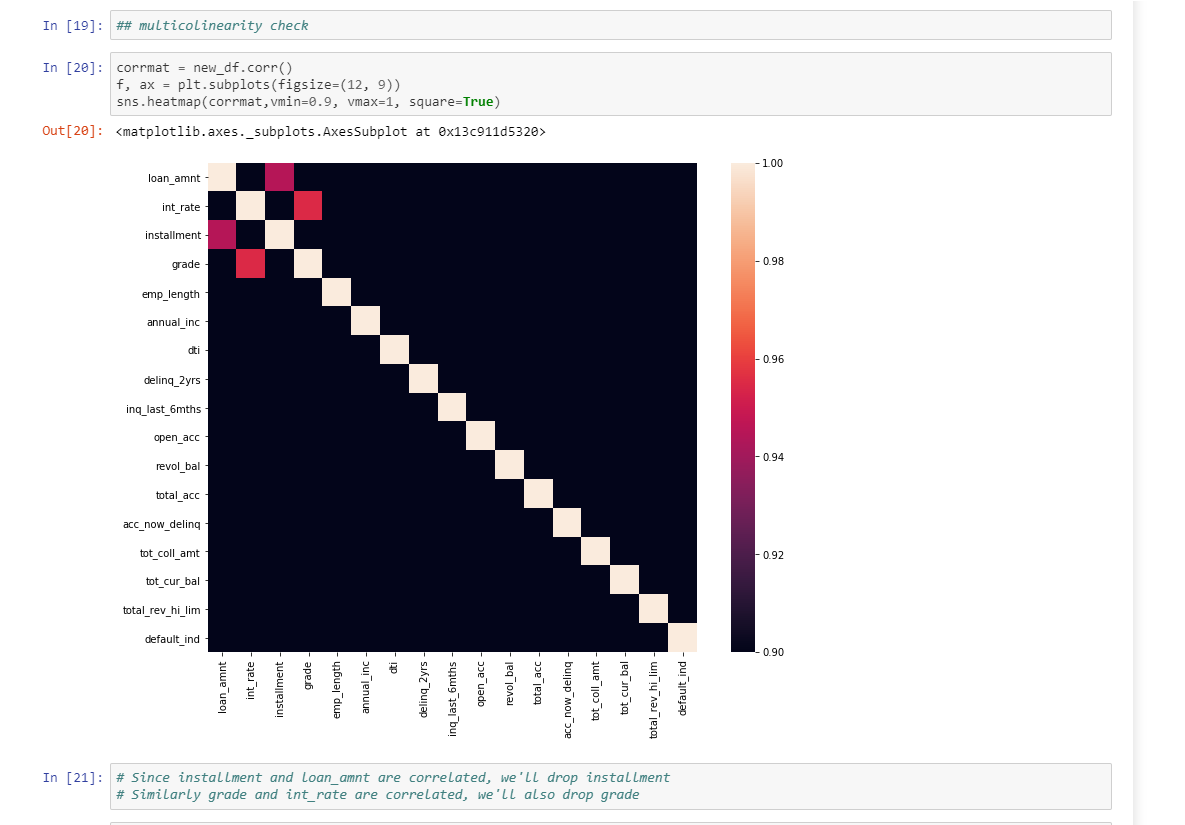
**Correlation:**

### Finding the correlation between variables

We will now look at the correlation structure between our variables that we selected above. This will tell us about any dependencies between different variables and help us reduce the dimensionality a little bit more

The variables checked for correlation are:

[loan\_amnt','funded\_amnt','funded\_amnt\_inv','installment','int\_rate','annual\_inc','dti','total\_pymnt','total\_pymnt\_inv','total\_rec\_prncp']

****

**Insights**: It is clear from the Heatmap that how **'loan\_amnt','funded\_amnt' & 'funded\_amnt\_inv'** are closely **interrelated**. So we can take any one column out of them for our analysis. Also ,**'total\_pymnt','total\_pymnt\_inv'** are highly correlated.

**Transformation:** Box-cox

Before training the data, we would first transform the data to account for any skewness in the variable distribution. Various transformation techniques ranging from log transform to power transformation are available. For our analysis, we'll be using Box-cox transformation. It is used to modify the distributional shape of a set of data to be more normally distributed so that tests and confidence limits that require normality can be appropriately used.

**Feature Scaling:**

We've have scaled the data so that each column has a mean of zero and unit standard deviation. We have scaled the training set and test set as well so as to reproduce the same results.

**Label Encoding**

Since we have some categorical variables for the analysis and the machine learning algorithms doesn't take categorical and string variables directly, we have to create dummy variables for them or We can either encode them using label encoder available for python, Here we are using Labels instead of one hot . Here Emp\_len , and Grades are two categorical variables, we cannot use them as it in our data if we are going to run any model on it. So converting this text data into model understandable numerical data, so here we Label Encoder class from sklearn library, here we replace the existing text data with new encoded numeric data.



## 2.3 Data Dictionary:

|  |  |
| --- | --- |
|  | **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 |

## 2.4 Exploratory Data Analysis:

**Derived Features For plotting:**

1. Loan amount to Annual Income ratio
2. Extract Year & Month from Issue date
3. Change order of months from Jan to Dec, currently it's in alphabetical order(A-Z)

**Univariate Analysis:**

#### **Loan Amount**

**A screenshot of a video game

Description generated with high confidence**

Insights: Most of the loan amounts are distributed between 8000 to 20000 USD.

1. **Interest Rate**

**A picture containing screenshot

Description generated with high confidence**

Insights: Most of the loans interest rates are distributed between 10% to 16%.

#### **3) Annual Income**

A close up of a map

Description generated with very high confidence

Max value is 9500000 which is approx. 150 times more than mean value, so we will remove the outliers from Annual Income.

Insights: Most of the applicants earns between 40000 to 90000 USD annually.

Remove Outliers (values from 99 to 100%)

### Categorical Variables:

#### **4) Default Ind**

A screenshot of a cell phone

Description generated with very high confidence

**About 6% of loans are charged off.**

#### **5. Purpose of loan**

A screenshot of a social media post

Description generated with very high confidence

**Insights**: Approx. 60% of the applicants applied loan for paying their other loans (Debt Consolidation).

#### **6. Home Ownership wise Loan**

A screenshot of a cell phone

Description generated with very high confidence

Insights: 40% of applicants are living in rented home whereas 52% applicants were mortgaged their home.

#### **7. Year wise Loan:**

A close up of a logo

Description generated with high confidence

Insights: loan applicants are increasing year on year, approx. 47% of loan applicants received loans in 2011.

# CHAPTER 3: FITTING MODELS TO DATA

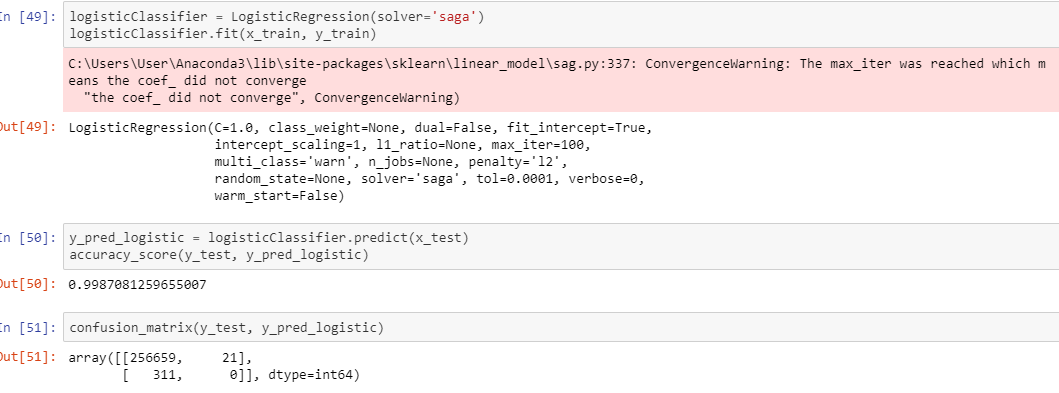
We have used the below models for our classification:

**Logistic regression**:

Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a [logistic function](https://en.wikipedia.org/wiki/Logistic_function).

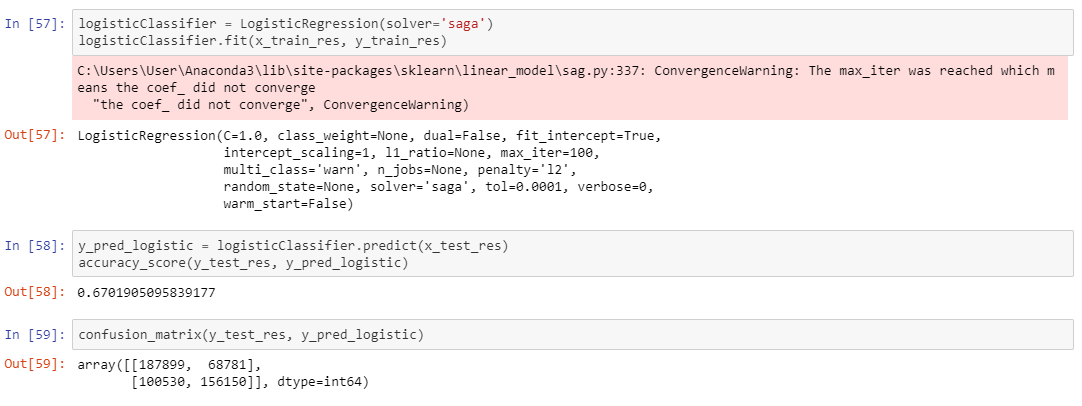
Logistic regression is implemented in [LogisticRegression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html" \l "sklearn.linear_model.LogisticRegression" \o "sklearn.linear_model.LogisticRegression). This implementation can fit binary, One-vs-Rest, or multinomial logistic regression with optional ℓ1, ℓ2 or Elastic-Net regularization. Note that regularization is applied by default.

### Logistic Regression Classifier without applying SMOTE



Problem: As observed above in the confusion matrix our data is clearly an imbalanced data so unable to derive any True Positive Predictions for the dataset. Solution: To mitigate this problem, below we have used SMOTE(Synthetic minority over sampling method), a technique for increasing the number of cases in your dataset in a balanced way. The module works by generating new instances from existing minority cases that you supply as input without changing the number of majority cases

### Logistic Regression Classifer after applying SMOTE



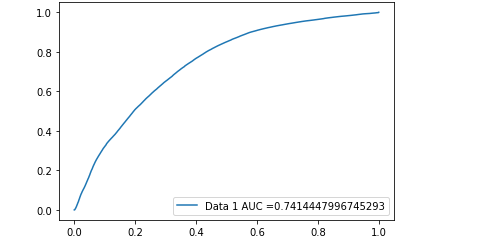
SMOTE is an over sampling method. What it does is, it creates synthetic(not duplicate) samples of the minority class equal to the majority class. SMOTE does this by selecting similar records and altering that record one column at a time by a random amount within the difference to the neighboring records .

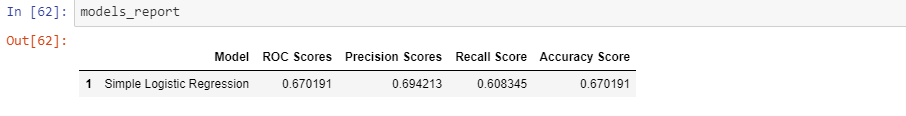
After applying SMOTE the accuracy is good , and looking at the confusion matrix we can see that Recall and precision are good . This result is achieved due to Balancing the data , through SMOTE .

Finally Got the better result in logistic regression module . here comes the codes and its results. For further better understanding we checked the ROC curve

**Putting the ROC Curve for Logistic Regression :**

**Creating the ROC curve on Y\_test and Y\_pred:**

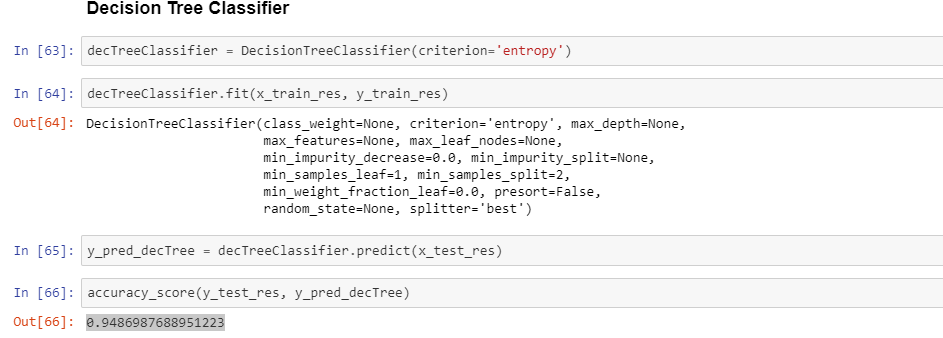




**Decision tree Classifier:**

We applied Decision tree estimator on the Training data set to validate if any further improvement of the model can be performed post Logistic regression.

Our Model gave us a 94% accuracy but the precision was very low. Below is the classification report. Below are our result



**Decision Tree Roc curve :**

****

**Random forest Classifier:**

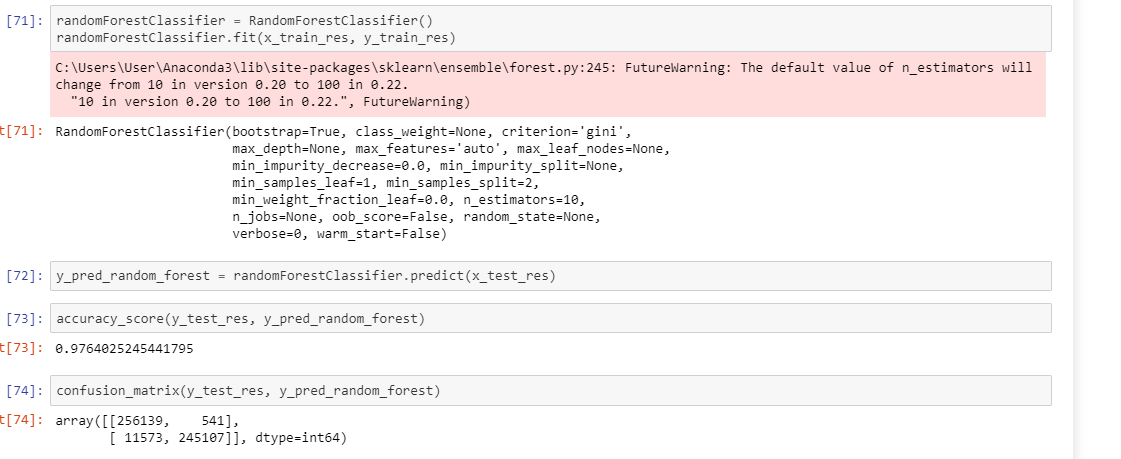
A random forest classifier.

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default).

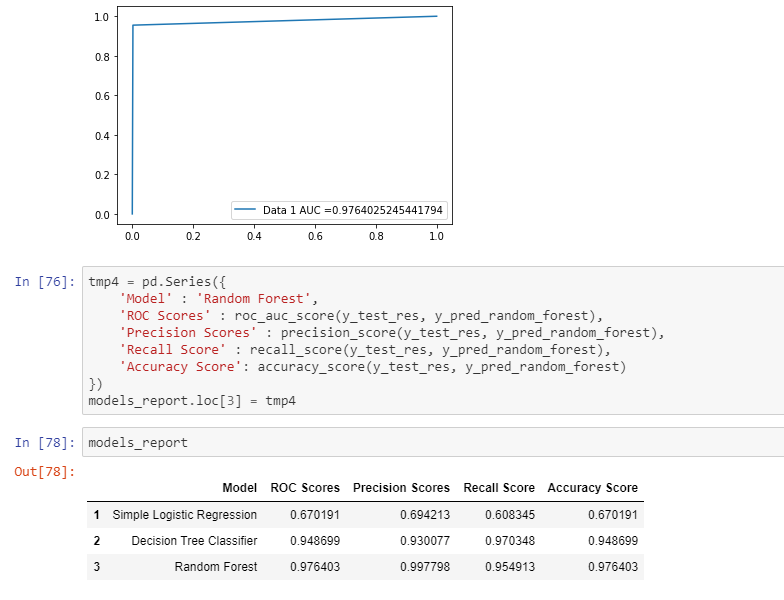
# 4.2 RANDOM FOREST

We applied Random Forest on the Training data set to validate if any further improvement of the model can be performed post Logistic regression.

Our Model gave us a 97% accuracy but the precision was very low. Below is the classification report. Below are our result



**ROC curve for Random forest :**



**Cross Validation of all the modules:**

Cross-validation is a statistical method used to estimate the skill of machine learning models.

It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

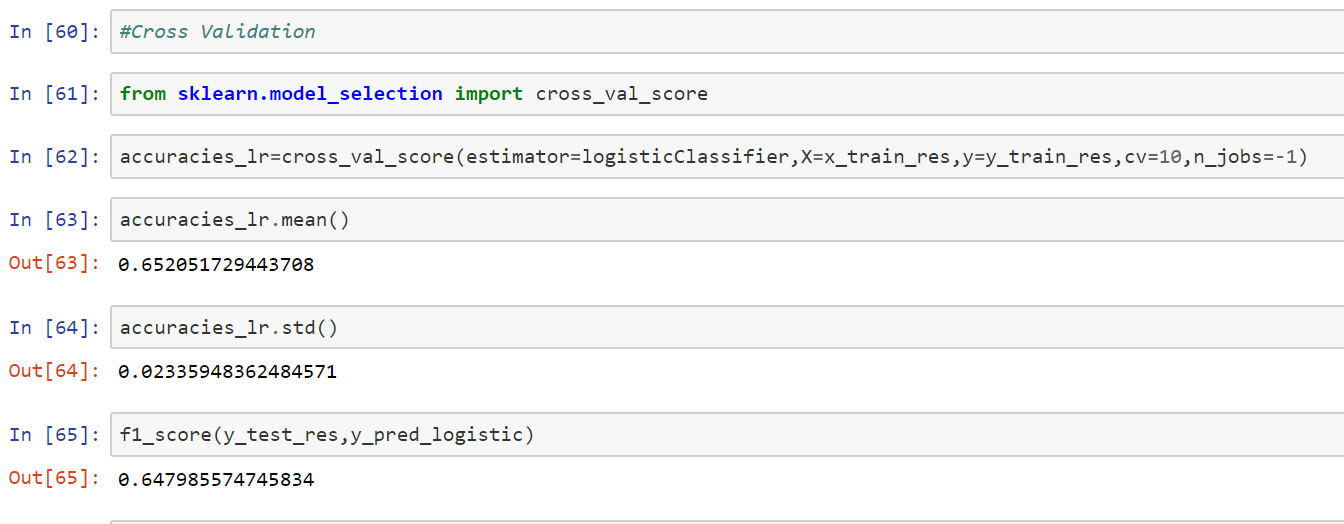
**cross\_val\_score**" splits the data into say 5 folds. Then for each fold it fits the data on 4

folds and scores the 5th fold. Then it gives you the 5 scores from which you **can** calcu

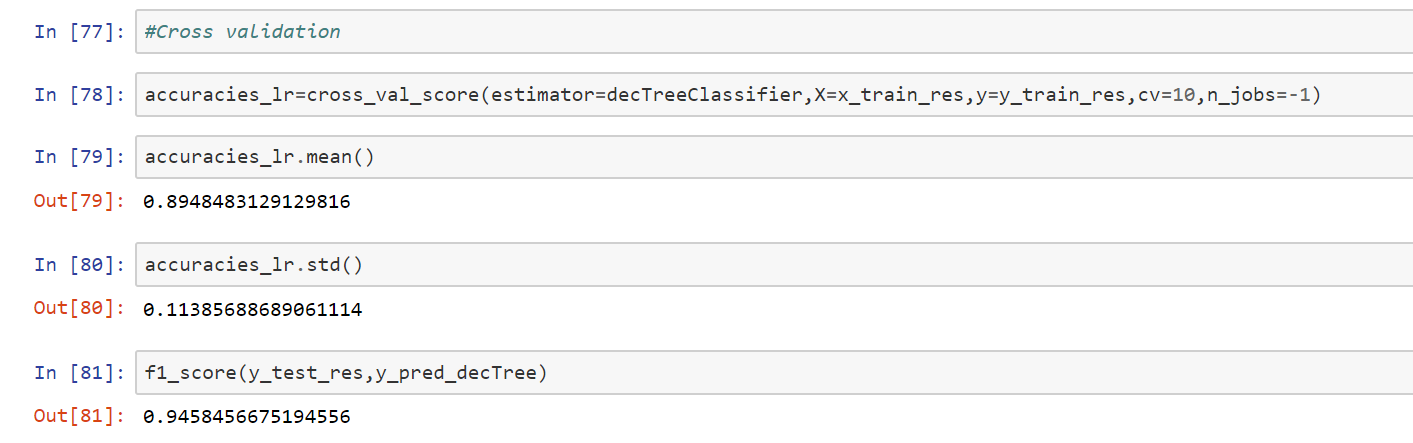
late a mean and variance for the score. ... The high score could mean more features **is** a

better model..

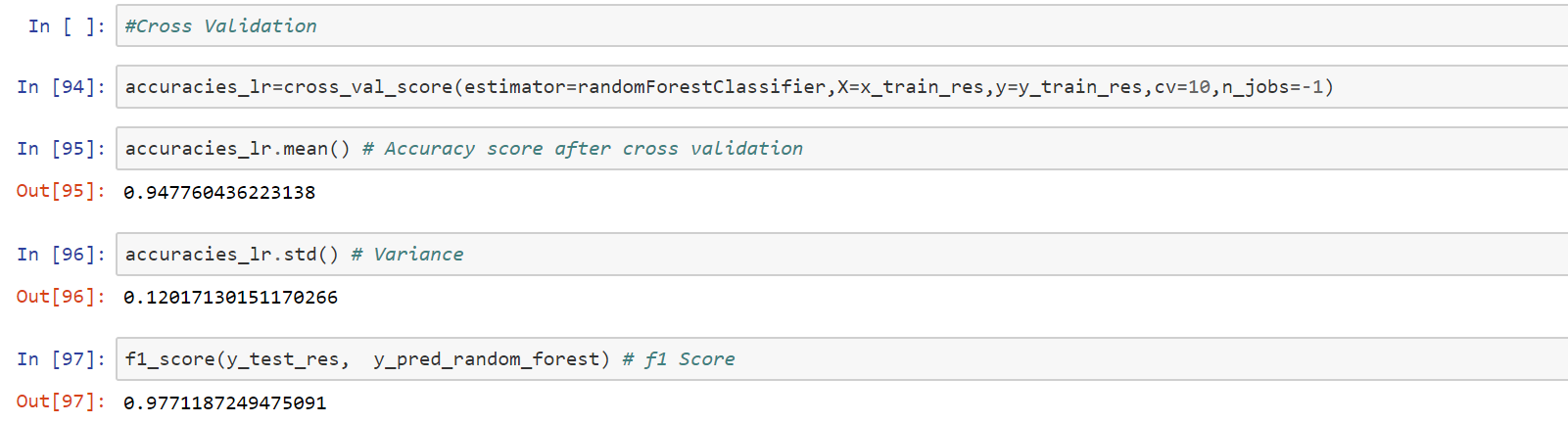
**Cross vaidation of Logistic Regression module:**

****

**Cross validation of Decision Tree:**

****

**Cross validation of Randon forest:**

****

CHAPTER 5: KEY FINDINGS

Significant Variables identified in linear models are also used in Random forest

Below table provides a snapshot of the various models which the business can choose from based on the pros and cons of each model.

**Below are some of the key findings:**

* The imbalance in the target category of loan repayment in the dataset, was due to the fact that 82 out of 100 loans were repaid. This indicates money could be lent continuously (always predicting that the borrower would repay) and be correct about 82.07% of the time that the loan was repaid.
* If we look at the confusion matrix, though, we see a big problem. The model can predict who are going to pay off the loan with a good accuracy of 99% but cannot predict who are going to default. The true positive rate of default (0 predicting 0) is almost 0. Since our main goal is to predict defaulter's, we have to do something about this. The reason this is happening could be because of high imbalance in our dataset and the algorithm is putting everything into 1.
* So to make the data balanced We used SMOTE.
* The cross-validation scores and ROC curves suggest the Random forest is the best model with 97% accuracy.
* We cannot perform grid search as the running time is very high.

CHAPTER 6: RECOMMENDATIONS AND CONCLUSION:

We have successfully built an machine learning algorithm to predict the people who might default on their loans.

Since test data is the data from June 2015 - Dec 2015, we can now estimate the monitory gain we'll have, if we've used our model at the time of lending the loan. The amount which we would have saved, if we have used our model during the period of Jun-Dec (2015) will be,

# the number of defaulters during this period \* mean of loan amount \* model accuracy

**Business Insights and Recommendations:**

The facts from our analysis shows that Applicants who has taken the Loan for 'small business' has the highest probability of charge off of 14%. Hence, bank should take extra caution like take some asset or guarantee while approving the loan for purpose of 'small business'

Banks should consider "Grade" as a major variable while providing loans.

Also, As the annual income is decreasing the probability that person will default is increasing with highest of 7% at (0 to 25000) salary bracket. The banks should either start with less principal loan amount and check the credibility.

Finally, As the interest rate is increasing the probability that person will default with highest of 9% at 15% & above bracket. Banks should consider minimizing their interest range for Applicants who are self-employed & less than 1 year of experience as they are more probable of charged off

CHAPTER 7: REFERENCES

1. <https://www.kaggle.com/deepanshu08/prediction-of-lendingclub-loan-defaulters>
2. <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html>
3. <https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score>
4. <https://scikit-learn.org/stable/model_selection.html#model-selection>
5. [https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier)