PROJECT REPORT

on

Defaulters Tracking Model

Submitted towards the partial fulfillment of the criteria for award of Genpact Data Science Prodegree by Imarticus

Submitted By:

Rishabh Surve

Prathamesh Pawaskar

Nikhil Bhosale

Jiipson Dsouza

Course and Batch: DSP Batch - 14 (July -Aug 2018)

Abstract

Data has the potential to transform business and drive the creation of business value. Data can be used for a range of simple tasks such as managing dashboards or visualizing relationships. However, the real power of data lies in the use of analytical tools that allow the user to extract useful knowledge and quantify the factors that impact events. Some examples include: Customer sentiment analysis, customer churn, geo-spatial analysis of key operation centres, workforce planning, recruiting, or risk- sensing.

Analytical tools are not the discovery of the last decade. Statistical regressions and classification models have been around for the best part of the 20th century. It is, however, the explosive growth of data in our times combined with the advanced computational power that renders data analytics a key tool across all businesses and industries.

In the Financial Industry some examples of using data analytics to create business value include fraud detection, customer segmentation, employee or client retention.

In order for data analytics to reveal its potential to add value to business, a certain number of ingredients need to be in place. This is particularly true in recent times with the explosion of big data (big implying data volume, velocity and variety).

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 **Miss. Nikita Tandel** as our mentor. She has readily shared his 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: September 11, 2018
Place: Mumbai

Rishabh Surve
Prathamesh Pawaskar
Nikhil Bhosale
Jiipson Dsouza

Certificate of Completion

I hereby certify that the project titled "**DEFAULTERS TRACKING MODEL**" was undertaken and completed under my supervision by Rishabh Surve , Prathamesh Pawaskar, Nikhil Bhosale, Jipson Dsouza from the batch of DSP (14).

Mentor: **Miss. Nikita Tandel**

Date: September 11, 2018

Place – Mumbai

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

1.1 Title & Objective of the study-

To explore qualitatively and quantitatively the risk associated with giving out credit for personal and commercial purposes and to model risk factor using a widely used machine learning classification method : Logistic Regression.

1.2 Need of the Study-

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.

1.3 Business or Enterprise under study

Credit risk refers to the probability of loss due to a borrower’s failure to make payments on any type of debt. Credit risk management is the practice of mitigating losses by understanding the adequacy of a bank’s capital and loan loss reserves at any given time – a process that has long been a challenge for financial institutions.

The global financial crisis – and the credit crunch that followed – put credit risk management into the regulatory spotlight. As a result, regulators began to demand more transparency. They wanted to know that a bank has thorough knowledge of customers and their associated credit risk. And new Basel III regulations will create an even bigger regulatory burden for banks.

To comply with the more stringent regulatory requirements and absorb the higher capital costs for credit risk, many banks are overhauling their approaches to credit risk. But banks who view this as strictly a compliance exercise are being short-sighted. Better credit risk management also presents an opportunity to greatly improve overall performance and secure a competitive advantage.

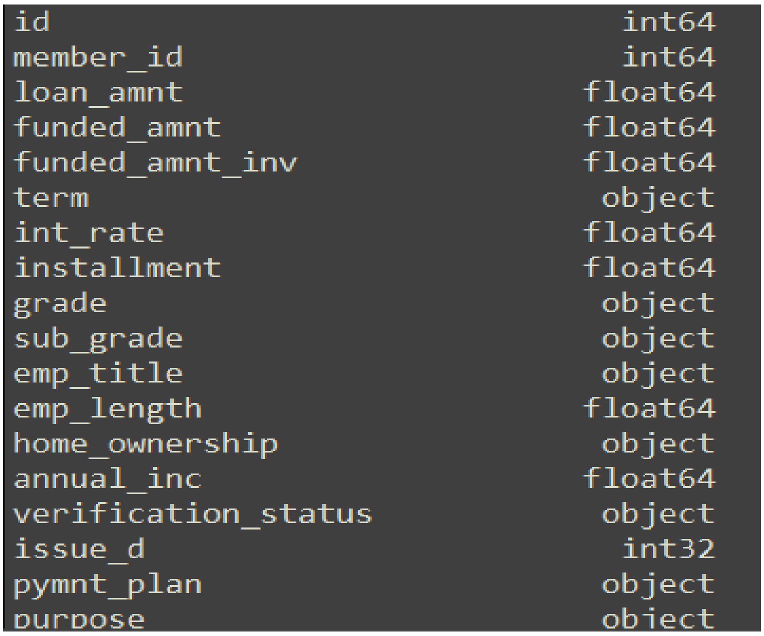
1.4 Data Sources

a) Dataset:

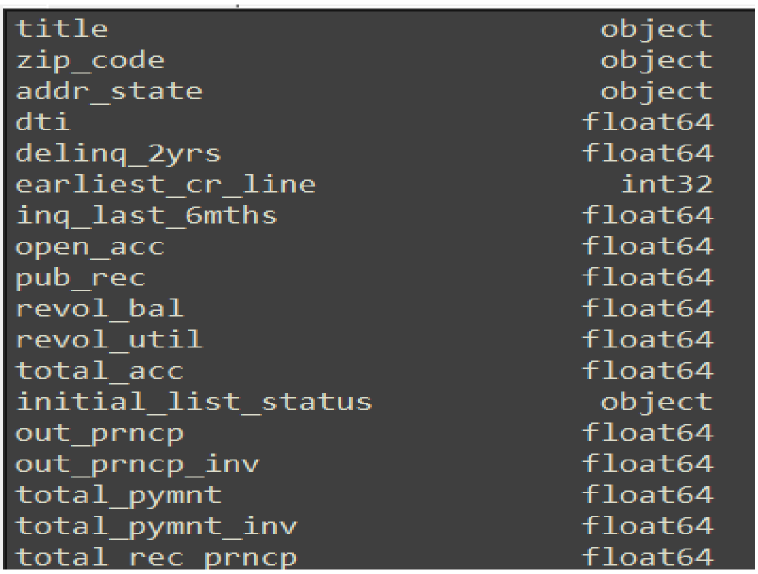
The dataset is taken as the banks record about the status of loan defaults and profile of customers. The dataset contains information like age, annual income, home ownership , grade of employee affects the loan paying capacity of the customers.

b) Dataset Description:

Contains 855969 rows and 73 columns



```
id                                int64
member_id                         int64
loan_amnt                       float64
funded_amnt                     float64
funded_amnt_inv                 float64
term                             object
int_rate                        float64
installment                     float64
grade                            object
sub_grade                        object
emp_title                        object
emp_length                      float64
home_ownership                   object
annual_inc                      float64
verification_status              object
issue_d                           int32
pymnt_plan                       object
purpose                          object
```



```
title                            object
zip_code                         object
addr_state                       object
dti                             float64
delinq_2yrs                     float64
earliest_cr_line                  int32
inq_last_6mths                  float64
open_acc                        float64
pub_rec                         float64
revol_bal                       float64
revol_util                      float64
total_acc                       float64
initial_list_status              object
out_prncp                       float64
out_prncp_inv                   float64
total_pymnt                     float64
total_pymnt_inv                 float64
total_rec_prncp                 float64
```

1.5 Tools & Techniques

Tools:

Annaconda Navigator, spyder(python 3.6)IDE for windows.

Spyder is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It features a unique combination of the advanced editing, analysis, debugging and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection and beautiful visualization capabilities of a scientific package.

Furthermore, Spyder offers built-in integration with many popular scientific packages, including NumPy, SciPy, Pandas, IPython, QtConsole, Matplotlib, SymPy, and more. Beyond its many built-in features, Spyder's abilities can be extended even further via first- and third-party plugins. Spyder can also be used as a PyQt5 extension library, allowing you to build upon its functionality and embed its components, such as the interactive console or advanced editor, in your own software.

Techniques:

Machine Learning Techniques:

Logistic Regressions :

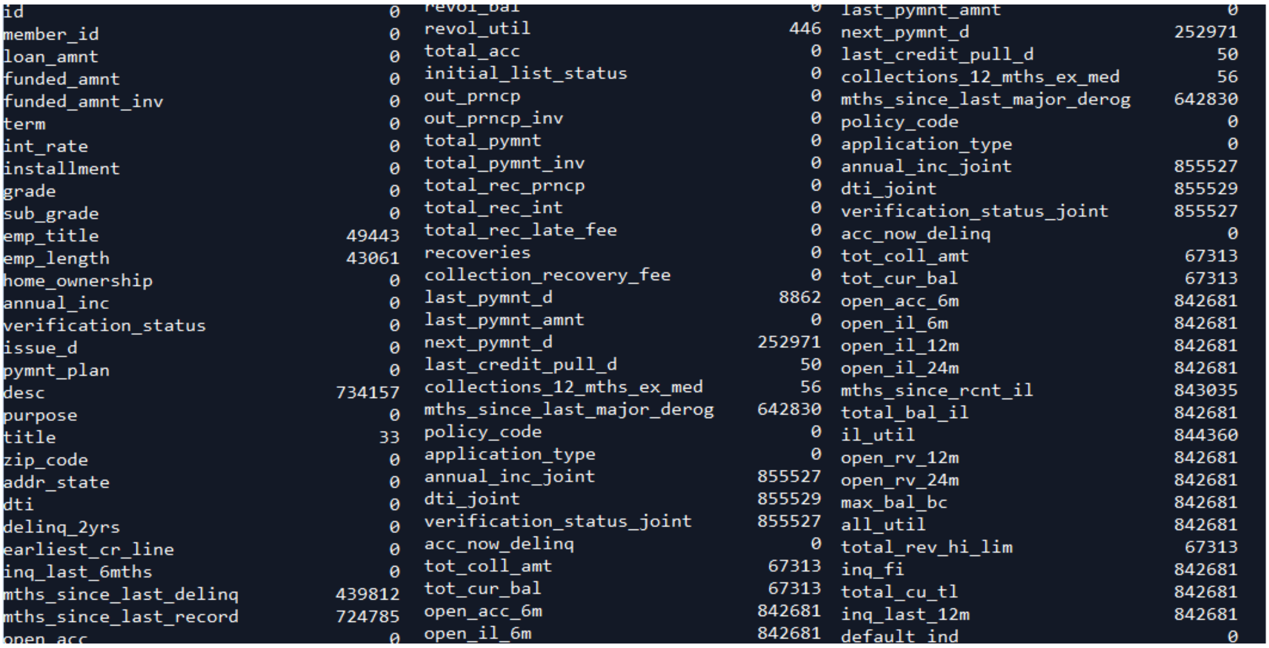
One of the most common, successful and transparent ways to do the required binary classification to “good” and “bad” is via a logistic function. This is a function that takes as input the client characteristics and outputs the probability of default.

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:

- Data source available in text i,e(.txt) format.
- To extract the data with spyder IDE we used pd.read_ csv function from pandas library delimiters " /t "
- The data is available in following datatypes:
    1) Integers
    2) Floats
    3) Object
- Where 49 out of 73 variables are floats , 21 variables are objects and remaining variables are integers type



```
id                                  0
member_id                           0
loan_amnt                           0
funded_amnt                         0
funded_amnt_inv                     0
term                                0
int_rate                            0
installment                         0
grade                               0
sub_grade                           0
emp_title                       49443
emp_length                      43061
home_ownership                      0
annual_inc                          0
verification_status                 0
issue_d                             0
pymnt_plan                          0
desc                           734157
purpose                             0
title                              33
zip_code                            0
addr_state                          0
dti                                 0
delinq_2yrs                         0
earliest_cr_line                    0
inq_last_6mths                      0
mths_since_last_delinq         439812
mths_since_last_record         724785
open_acc                            0
```

```
revol_bal                           0
revol_util                        446
total_acc                           0
initial_list_status                 0
out_prncp                           0
out_prncp_inv                       0
total_pymnt                         0
total_pymnt_inv                     0
total_rec_prncp                     0
total_rec_int                       0
total_rec_late_fee                  0
recoveries                          0
collection_recovery_fee             0
last_pymnt_d                     8862
last_pymnt_amnt                     0
next_pymnt_d                   252971
last_credit_pull_d                 50
collections_12_mths_ex_med         56
mths_since_last_major_derog    642830
policy_code                         0
application_type                    0
annual_inc_joint               855527
dti_joint                      855529
verification_status_joint      855527
acc_now_delinq                      0
tot_coll_amt                    67313
tot_cur_bal                     67313
open_acc_6m                    842681
open_il_6m                     842681
```

```
last_pymnt_amnt                     0
next_pymnt_d                   252971
last_credit_pull_d                 50
collections_12_mths_ex_med         56
mths_since_last_major_derog    642830
policy_code                         0
application_type                    0
annual_inc_joint               855527
dti_joint                      855529
verification_status_joint      855527
acc_now_delinq                      0
tot_coll_amt                    67313
tot_cur_bal                     67313
open_acc_6m                    842681
open_il_6m                     842681
open_il_12m                    842681
open_il_24m                    842681
mths_since_rcnt_il             843035
total_bal_il                   842681
il_util                        844360
open_rv_12m                    842681
open_rv_24m                    842681
max_bal_bc                     842681
all_util                       842681
total_rev_hi_lim                67313
inq_fi                         842681
total_cu_tl                    842681
inq_last_12m                   842681
default_ind                         0
```

Cleaning of data:

- Using isnull.sum function in pandas we conclude many of the variables having more than 50% of observations in data are missing.
- The data frame includes variables which contains more than 50% of missing Information.
- As it is not practically possible to build the model using those variables, hence we removed those variables.

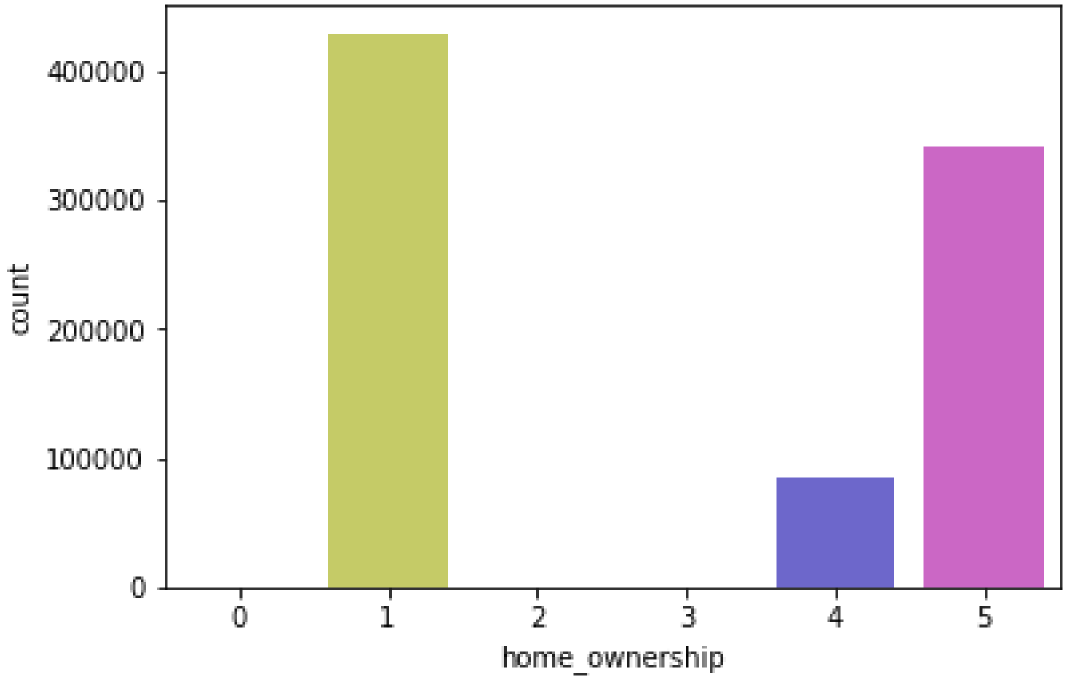
Missing Value Analysis and Treatment

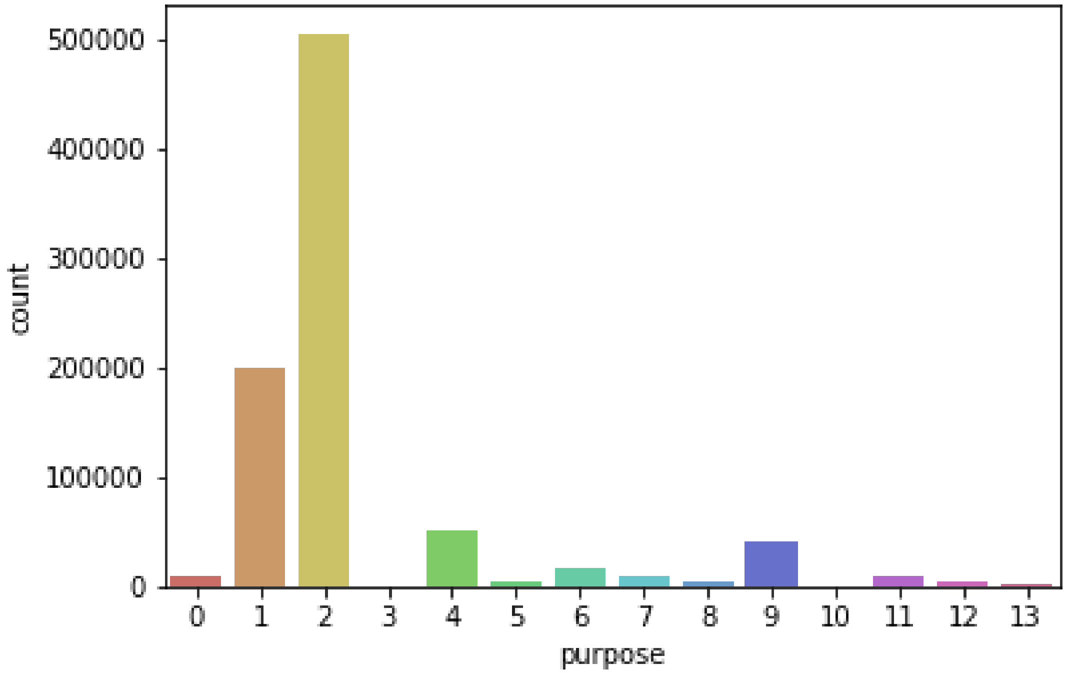
- In the data frame, variable called "emp_length", the values are in years. We are replacing < 1 with 0 years and 10 + years with 10 years according to the data dictionary.
- The missing values in the emp_length are replaced with 0 as the mode value will create the biasness in the model
- The other variables which contains missing values are backfilled with mean.
- The missing values in dates are backfilled with mode

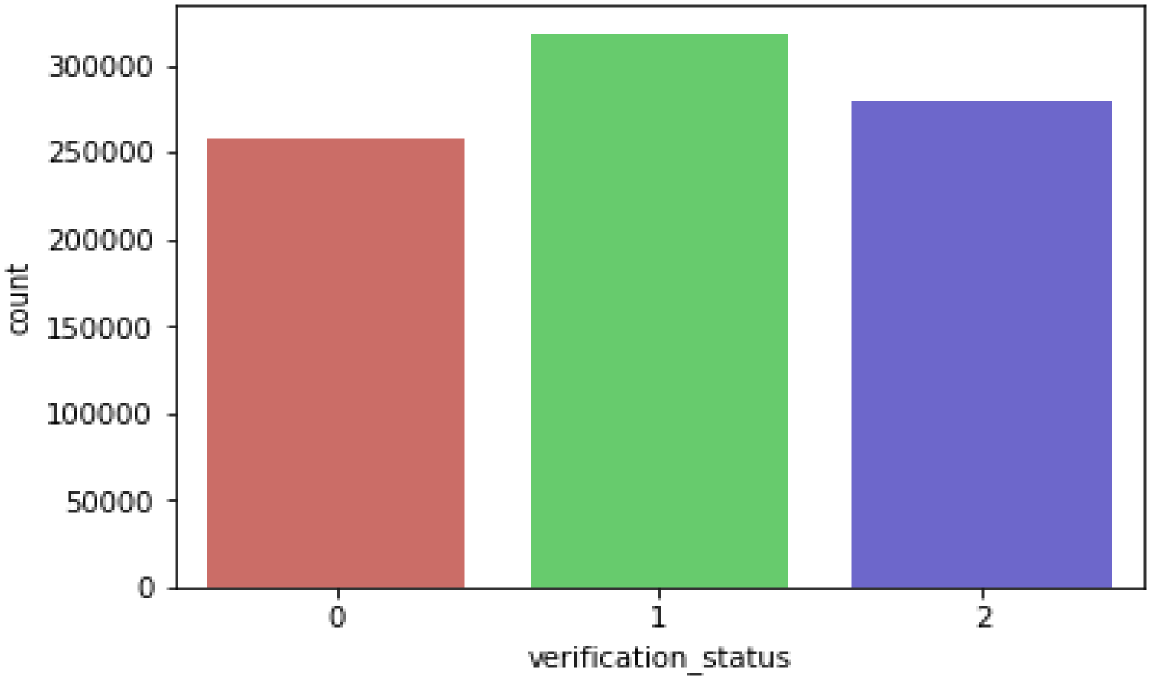
2.2 Data Dictionary

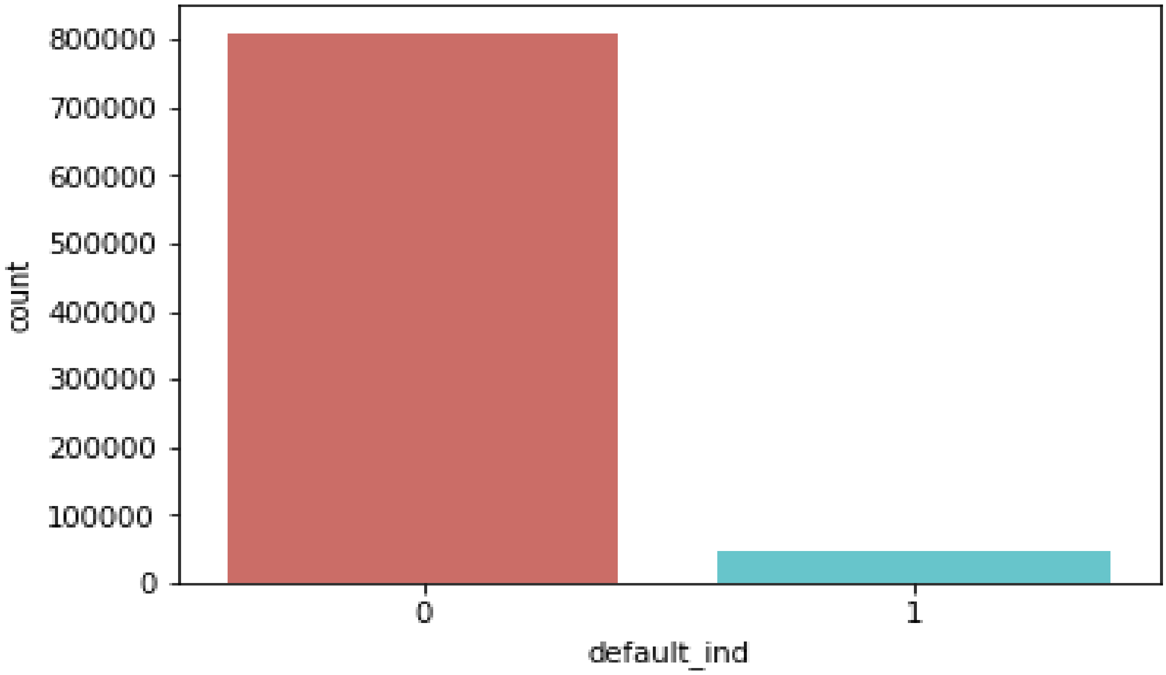
- <class 'pandas.core.frame.DataFrame'>
- Range Index: 855969 entries, 0 to 855968
- Data columns = 36 out of 73 columns are taken for the best accuracy of the model.
- In these, 36 variables (28) variables are float64 type, 4 variables are int32 type And remaining variables are int64 type.
- dtypes: float64(28), int32(4), int64(12).

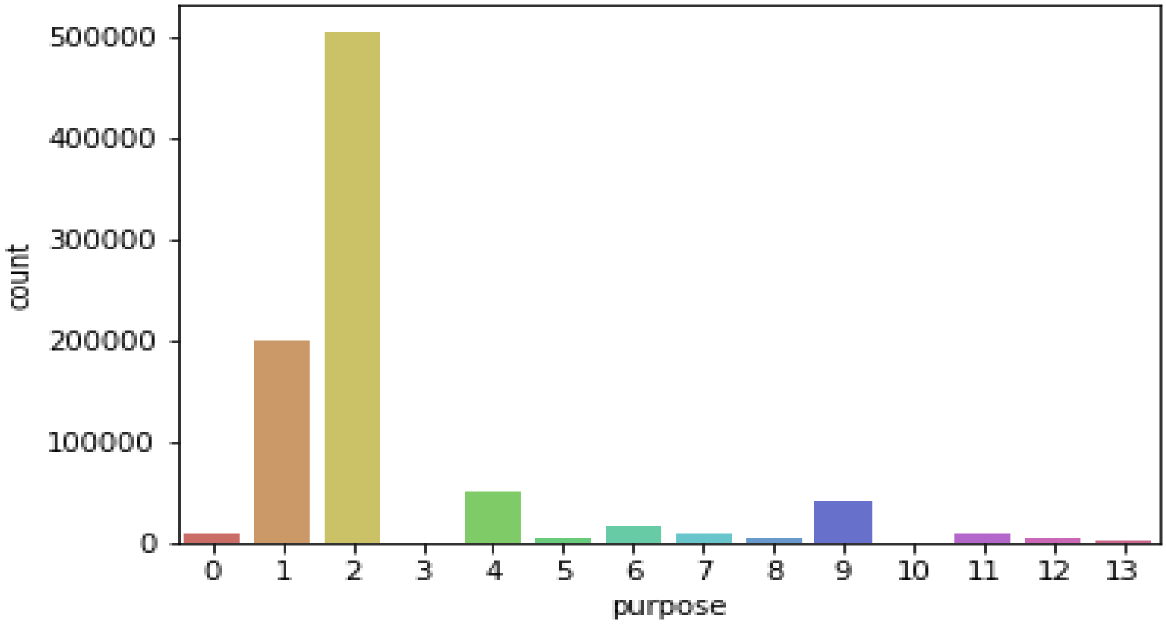
2.3 Data visualization :






300000
250000
200000
150000
100000
50000
0
count
0
1
2
verification_status


800000
700000
600000
500000
400000
300000
200000
100000
0
count
0
1
default_ind



2.4 Data Standardization

Data standardization is a process in which data attributes within a data model are organized to increase the cohesion of entity types. In other words, the goal of data standardization is to reduce and even eliminate data redundancy, an important consideration for application developers because it is incredibly difficult to store objects in a database that maintains the same information in several places.

A standardized data results are:

- Mean= 0
- Standard deviation=1
- Bell shaped curve

CHAPTER 3 : FITTING MODELS TO DATA

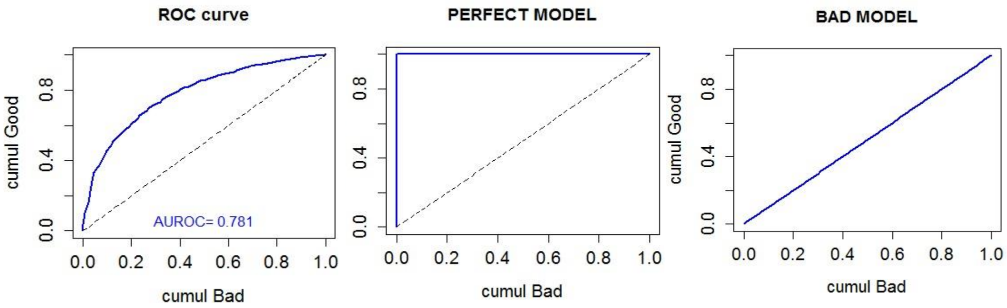
Before the analysis begins it is important to clearly state out what defines a default. This definition lies at the heart of the model. Different choices will have an impact on what the model predicts. Some typical choices for this definition include the cases that the client misses three payments in a row, or, that the sum of missed payments exceeds a certain threshold.

Classification

The aim of the credit scoring model is to perform a classification: To distinguish the “good” applicants from the “bad” ones. In practice this means the statistical models is required to find the separating line distinguishing the two categories, in the space of the explanatory variables (age, salary, education, etc.). The difficulty in the doing so is that the data is only a sample from the true population (e.g. the bank has records only from the last 10 years, or the data describes clients of that particular bank) and (ii) the data is noisy which means that some of significant explanatory variables may not have been recorded or that the default occurred by accident rather than due to the explanatory factors.

Typically we concentrate all information of the ROC curve into one number which is chosen to be the area under the ROC curve (the perfect model has an area equal to 1). Based on experience we givebelow a table of ROC values and their interpretation with respect to the model appropriateness

Acceptable	>70%
Good	>80%
VeryGood	>85%



Confusion Matrix

An additional measure of predictive power is the so-called confusion matrix. It has the form of the table below (which is a hypothetical example). We test the model's classification results against the actual observed classification. Of particular interest in this table is the "True Positive Rate" that corresponds to the fraction of Goods that are correctly classified [in the example below 7014/(7014+3171)) and the "True Negative Rate" that corresponds to the fraction of Bad that are correctly classified (in the example) below .

	Predicted Bad	Predicted Good
Observed Bad	357	178
Observed Good	3171	7014

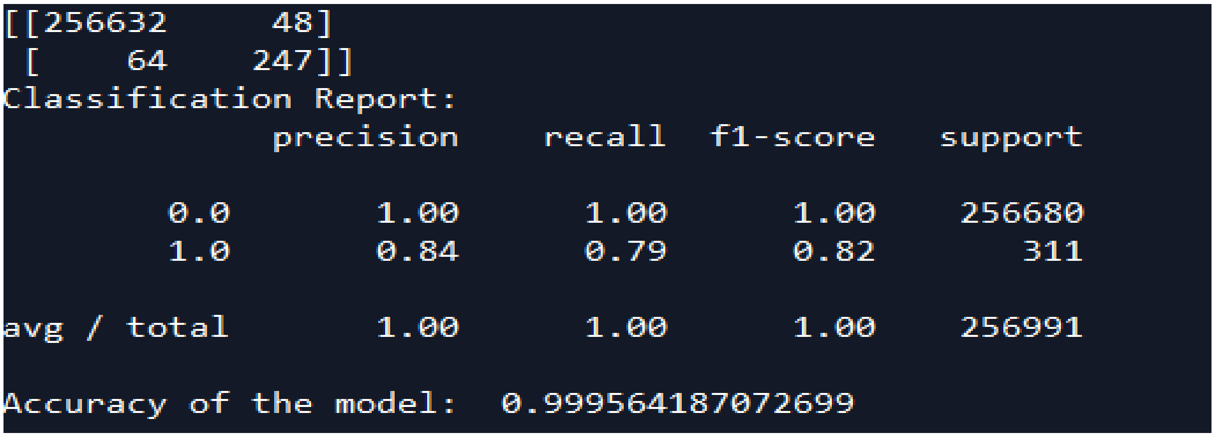
Based on experience we give below a table of figures that will allow us to interpret the results of the confusion matrix:

Acceptable	>60%
Good	>70%
Very Good	>85%

LOGISTIC REGRESSION MODEL

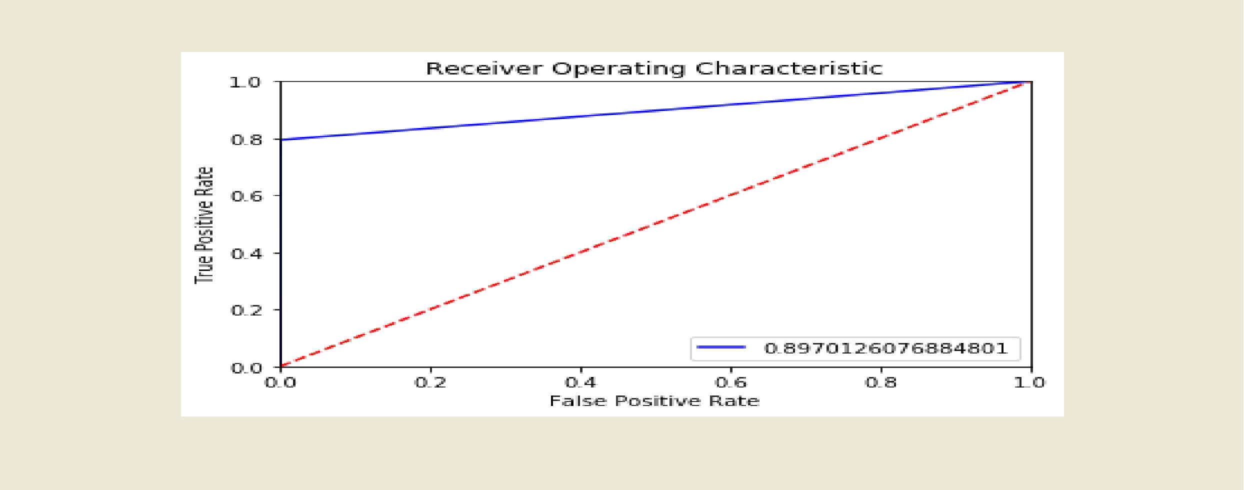
4.1.1 First Logistic Regression Model:

- From the data sets we selected 44 variables on our choice to run the model.
- Accuracy, precision and confusion matrix is shown below:



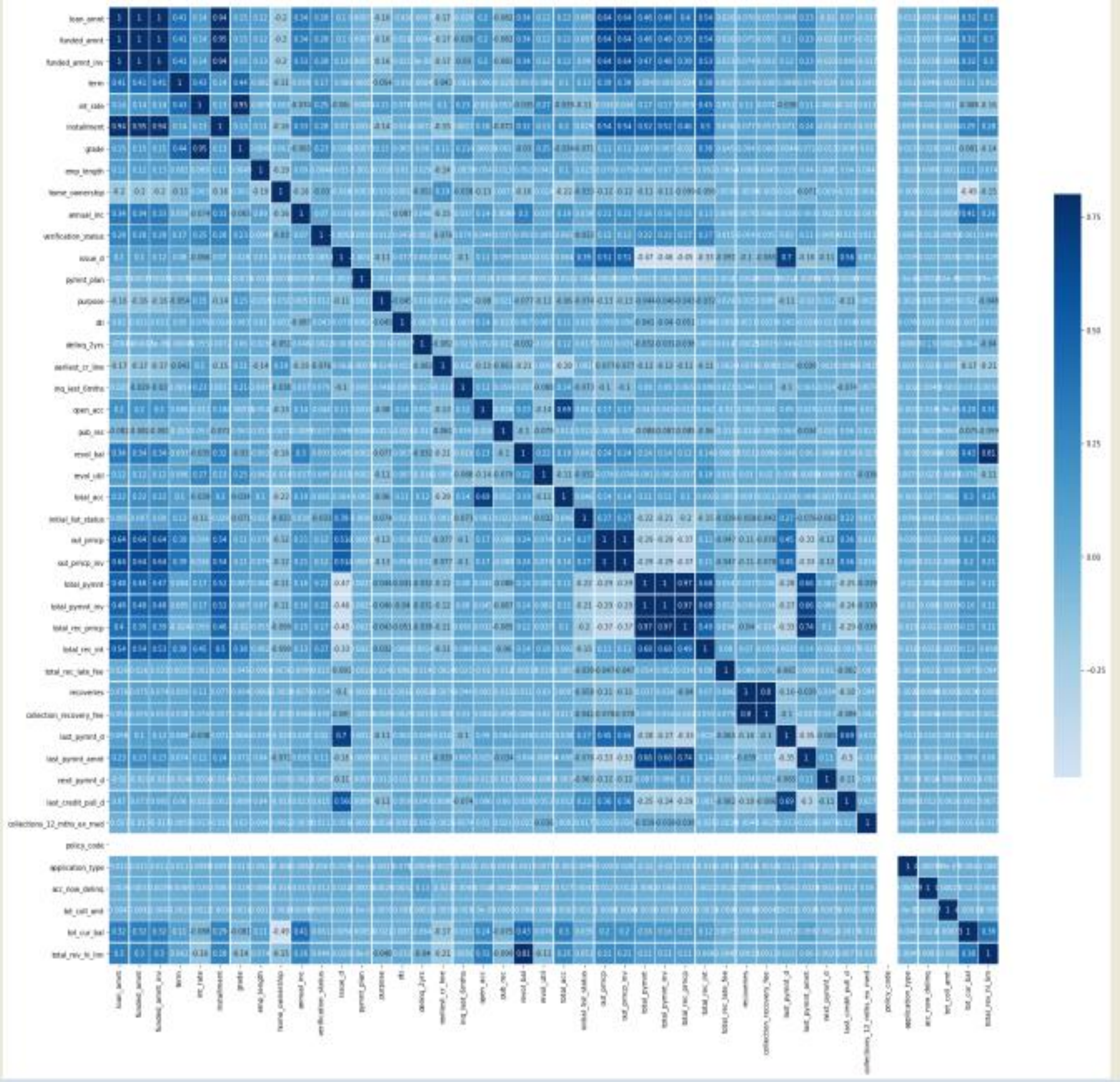
```
[[256632     48]
 [    64    247]]
Classification Report:
             precision    recall  f1-score   support

        0.0       1.00      1.00      1.00    256680
        1.0       0.84      0.79      0.82       311

avg / total       1.00      1.00      1.00    256991

Accuracy of the model:  0.999564187072699
```

- Receiver operating characteristics of model 1

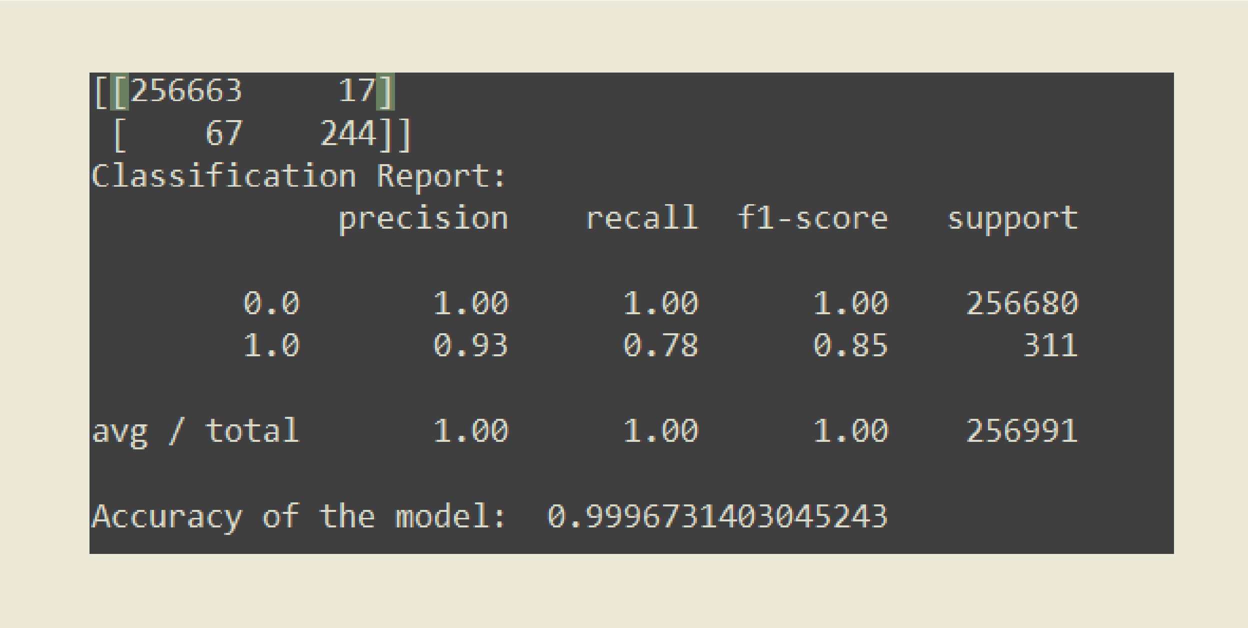


4.1.2 Second Logistic Regression model:

- From the dataset, we removed some variables on the basis of their multicollinearity and selected 36 variables.

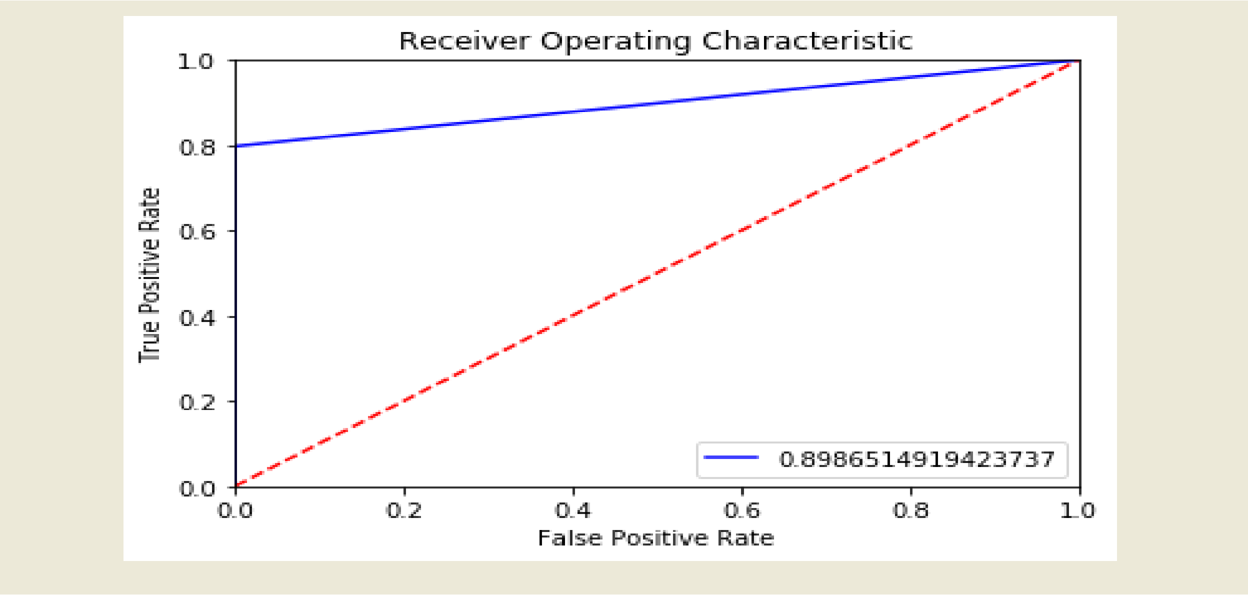


- Accuracy, precision and confusion matrix is shown below:



```
[[256663     17]
 [    67    244]]
Classification Report:
             precision    recall  f1-score   support

        0.0       1.00      1.00      1.00    256680
        1.0       0.93      0.78      0.85       311

avg / total       1.00      1.00      1.00    256991

Accuracy of the model:  0.9996731403045243
```

- Receiver operating characteristics of model 2:



4.1.3 Model Validation

K-Fold Cross Validation:

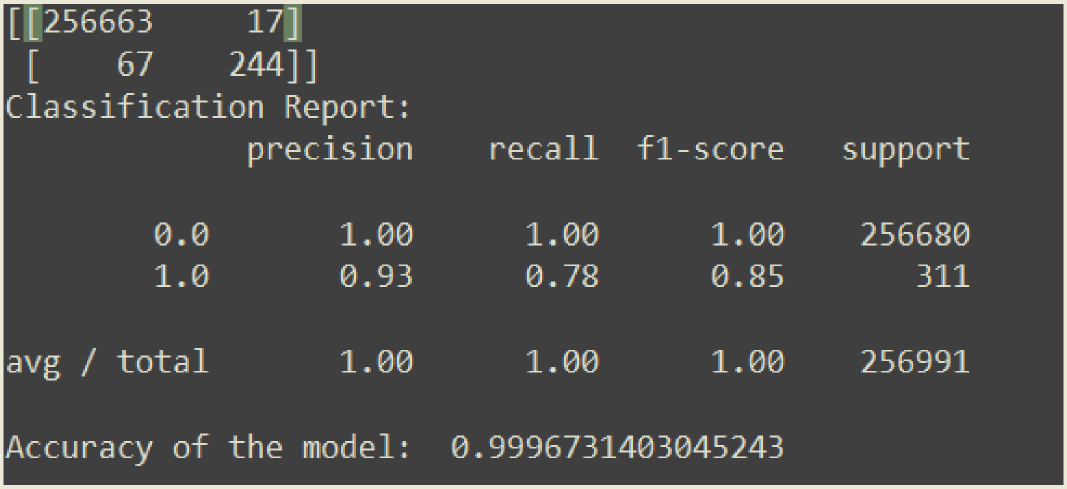
It’s easy to follow and implement. It is used to validate our model prediction. Below are the steps for it:

1. Randomly split your entire dataset into k”folds”
2. For each k-fold in your dataset, build your model on k – 1 folds of the dataset. Then test the model to check the effectiveness for *kth* fold
3. Record the error you see on each of the predictions
4. Repeat this until each of the k-folds has served as the test set
5. The average of your k recorded errors is called the cross-validation error and will serve as your performance metric for the model

We used the number of folds as 10 and after validating our model using K-fold ,we received the K fold mean as ().

4.1.4 Predicting New Data

On the basis of our model we predicted the test data which gave us the following confusion matrix.



```
[[256663     17]
 [    67    244]]
Classification Report:
             precision    recall  f1-score   support

        0.0       1.00      1.00      1.00    256680
        1.0       0.93      0.78      0.85       311

avg / total       1.00      1.00      1.00    256991

Accuracy of the model:  0.9996731403045243
```

4.1.5 Comparison of Logistic Regression model.

MODELS	TYPE 1 ERROR	TYPE 2 ERROR	ACCURACY	PRECISIONS
LOGISTIC REGRESSION MODEL 1	48	64	99.95	0.84
LOGISTIC REGRESSION MODEL 2	17	67	99.96	0.93

RECOMMENDATIONS AND CONCLUSION

- The data given by the bank is inherently biased towards the accepting loans.
- Two logistic models were used to predict the defaulters. The model with the more precision was selected. Different cut off gave different precision. The first model gave 0.88 as precision while the second model has 0.93 as precision. Hence, the most precise model is selected.
- The area under curve gives a measure of accuracy, which came out to be 90% approx.

REFERENCES

- https://datascienceplus.com
- https://www.investopedia.com/terms/c/creditrisk.asp
- https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
- http://analyticsvidya.com
- https://machinelearningmastery.com