FRA: INDIAN CREDIT RISK DEFAULT MODEL

PGP-DSBA 18 FEB 2022

By: THILAK RAJ

Table of Contents

Table of figures

Table of Tables

1. Introduction of the business problem

Problem objective:

The objective of this assignment is to create an Indian credit risk(default) model, using the data provided in the spreadsheet.

Dependent variable - We need to create a default variable which should take the value of 1 when net worth next year is negative & 0 when net worth next year is positive.

Validation Dataset - We need to build the model on train dataset and check the model performance measures on validation dataset.

Problem Statement:

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

This exploration report will consists of the following:

- Importing the dataset in python
- Understanding the structure of dataset
- Graphical exploration
- Descriptive statistics
- Logistic regression
- Random forest
- Insights from the dataset

2. Data Report

Variable description:

Networth Next Year	Net worth of the customer in next year
Total assets	Total assets of customer

Net worth	Net worth of the customer of present year
Total income	Total income of the customer
Change in stock	difference between value of current stock and the value of stock in last trading day
Total expenses	Total expense done by customer
Profit after tax	Profit after tax deduction
PBDITA	Profit before depreciation, income tax and amortization
PBT	Profit before tax deduction
Cash profit	Total Cash profit
PBDITA as % of total income	PBDITA / Total income
PBT as % of total income	PBT / Total income
PAT as % of total income	PAT / Total income
Cash profit as % of total income	Cash Profit / Total income
PAT as % of net worth	PAT / Net worth
Sales	Sales done by customer
Income from financial services	Income from financial services
Other income	Income from other sources
Total capital	Total capital of the customer
Reserves and funds	Total reserves and funds of the customer
Deposits (accepted by commercial banks)	All blank values
Borrowings	Total amount borrowed by customer
Current liabilities & provisions	current liabilities of the customer
Deferred tax liability	Future income tax customer will pay because of the current transaction
Shareholders funds	Amount of equity in a company, which is belong to shareholder
Cumulative retained profits	Total cumulative profit retained by customer
Capital employed	Current asset minus current liabilities
TOL/TNW	Total liabilities of the customer divided by Total net worth
Total term liabilities / tangible net worth	Short + long term liabilities divided by tangible net worth
Contingent liabilities / Net worth (%)	Contingent liabilities / Net worth
Contingent liabilities	Liabilities because of uncertain events
Net fixed assets	purchase price of all fixed assets
Investments	Total invested amount
Current assets	Assets that are expected to be converted to cash within a year
Net working capital	Difference of current liabilities and current assets
Quick ratio (times)	Total cash divided by current liabilities
Current ratio (times)	Current assets divided by current liabilities
Debt to equity ratio (times)	Total liabilities divided by its shareholder equity
Cash to current liabilities (times)	Total liquid cash divided by current liabilities
Cash to average cost of sales per day	Total cash divided by average cost of the sales
Creditors turnover	Net credit purchase divided to average trade creditors
Debtors turnover	Net credit sales divided by average accounts receivable
Finished goods turnover	Annual sales divided by average inventory
WIP turnover	The cost of goods sold for a period divided by the average inventory

	for that period
Raw material turnover	Cost of goods sold is divided by the average inventory for the same period
Shares outstanding	Number of issued shares minus the number of share held in the company
Equity face value	cost of the equity at the time of issuing
EPS	Net income divided by total number of outstanding share
Adjusted EPS	Adjusted net earning divided by the weighted average number of common share outstanding on a diluted basis during the plan year
Total liabilities	Sum of all type of liabilities
PE on BSE	Company current stock price divided by its earning per share

The dataset contains 4256 unique rows and 52 columns.
All the data are numerical variables.

Data sample:

	Num	Networth Next Year	Total assets	Net worth	Total income	Change in stock	Total expenses	Profit after tax	PBDITA	PBT	...	Debtors turnover	Finished goods turnover	WIP turnover	Raw material turnover	Shares outstanding	Equity face value	EPS
0	1	395.3	827.6	336.5	534.1	13.5	508.7	38.9	124.4	64.6	...	5.65	3.99	3.37	14.87	8760056.0	10.0	4.44
1	2	36.2	67.7	24.3	137.9	-3.7	131.0	3.2	5.5	1.0	...	NaN	NaN	NaN	NaN	NaN	NaN	0.00
2	3	84.0	238.4	78.9	331.2	-18.1	309.2	3.9	25.8	10.5	...	2.51	17.67	8.76	8.35	NaN	NaN	0.00
3	4	2041.4	6883.5	1443.3	8448.5	212.2	8482.4	178.3	418.4	185.1	...	1.91	18.14	18.62	11.11	10000000.0	10.0	17.60
4	5	41.8	90.9	47.0	388.6	3.4	392.7	-0.7	7.2	-0.6	...	68.00	45.87	28.67	19.93	107315.0	100.0	-6.52

5 rows × 51 columns

Table 1: Sample data

3. Exploratory Data Analysis

Data preprocessing:

Duplicate variables:

- We found no duplicate variables in the dataset.

Null values in the data:

- We found null values in37 variables.
- We are imputing the null values using the median of the column variables as the data is skewed and having outliers.

PE on BSE	61.724624
Investments	40.296053
Other income	36.560150
Contingent liabilities	32.941729
Deferred tax liability	32.166353
Income from fincial services	26.104323
Finished goods turnover	20.535714
Equity face value	19.031955
Shares outstanding	19.031955
WIP turnover	17.951128
Change in stock	12.922932
Borrowings	10.126880
Raw material turnover	10.056391
Creditors turnover	9.187030
Debtors turnover	9.046053
Sales	7.166353
Total income	5.427632
Total expenses	3.876880
Cash profit	3.618421
PBT	3.618421
Profit after tax	3.618421
PBDITA	3.618421
Net fixed assets	3.101504
Current liabilities & provisions	2.584586
Quick ratio (times)	2.467105
Cash to current liabilities (times)	2.467105
Current ratio (times)	2.467105
Cash to average cost of sales per day	2.349624
Reserves and funds	2.302632
Current assets	1.879699
Cash profit as % of total income	1.856203
PAT as % of total income	1.856203
PBT as % of total income	1.856203
PBDITA as % of total income	1.856203
Cumulative retained profits	1.057331
Net working capital	0.869361
Total capital	0.117481

Table 2: Percentage of null values in the variables.

The missing values are completely handled using median value imputation.

Outlier treatment:

- We noticed both positive and negative skewness in the data so there are outliers which are with larger values.

Networth Next Year	36.375204
Total assets	26.422680
Net worth	31.851686
Total income	31.443117
Change in stock	18.024259
Total expenses	32.190391
Profit after tax	24.290606
PBDITA	24.124350
PBT	22.275883
Cash profit	27.667906
PBDITA as % of total income	-29.030769
PBT as % of total income	-37.936981
PAT as % of total income	-37.170128
Cash profit as % of total income	-36.017775
PAT as % of net worth	17.761978
Sales	31.233587
Income from fincial services	40.462142
Other income	35.591580
Total capital	31.492327
Reserves and funds	34.108966
Borrowings	20.891301
Current liabilities & provisions	26.506920
Deferred tax liability	23.739302
Shareholders funds	31.549033
Cumulative retained profits	27.824601
Capital employed	28.275799
TOL/TNW	8.893421
Total term liabilities / tangible net worth	9.033640
Contingent liabilities / Net worth (%)	24.542580
Contingent liabilities	37.762615
Net fixed assets	37.623727
Investments	19.442847
Current assets	21.325079
Net working capital	8.836809
Quick ratio (times)	27.431505
Current ratio (times)	33.284368
Debt to equity ratio (times)	16.330812
Cash to current liabilities (times)	26.456958
Cash to average cost of sales per day	38.840939
Creditors turnover	19.719291
Debtors turnover	22.907662
Finished goods turnover	20.844660
WIP turnover	25.686670
Raw material turnover	60.607761
Shares outstanding	11.034062
Equity face value	-29.192120
EPS	-63.287482
Adjusted EPS	-63.287529
Total liabilities	26.422680
PE on BSE	37.196834
dtype: float64	

Table 3: skewness in the data

- Outliers are treated by imputing lower and upper range of values according to the outliers.

Descriptive summary of the data looks better after the treatment of the data.

Target Variable:

We created a target variable that takes "1" when net worth next year is negative and "0" When net worth next year is positive. Basically it is a binary value and named it as ‘default’ variable.

	default	Networth Next Year
3245	0	118.4
1604	0	69.1
3503	0	76.3
3232	1	0.0
4221	1	-5.4
3482	0	633.9
3764	0	3.6
2030	1	0.0
2204	1	0.0
1455	0	31.5

Table 4: Sample default variable data.

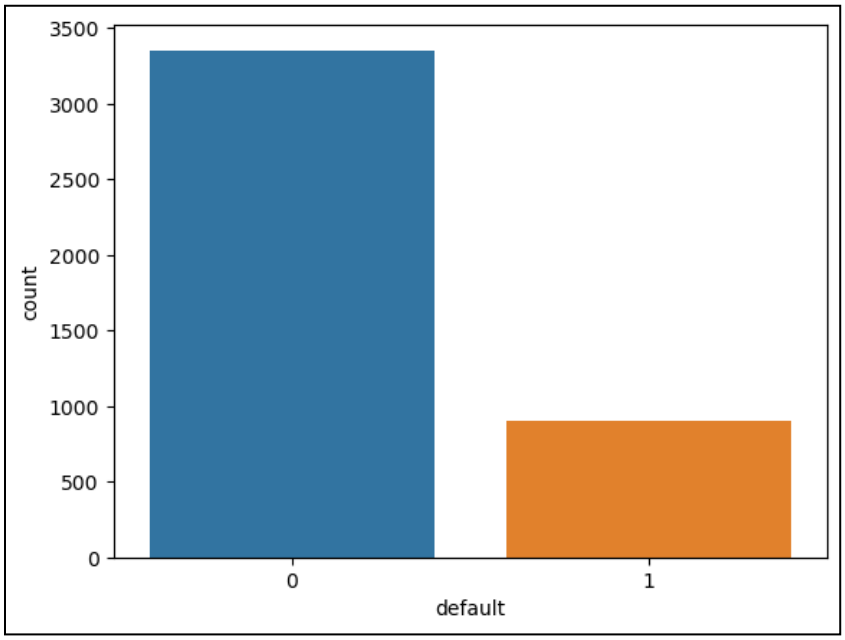

Figure 1: Population of default data

Value count:
0:3352
1:904

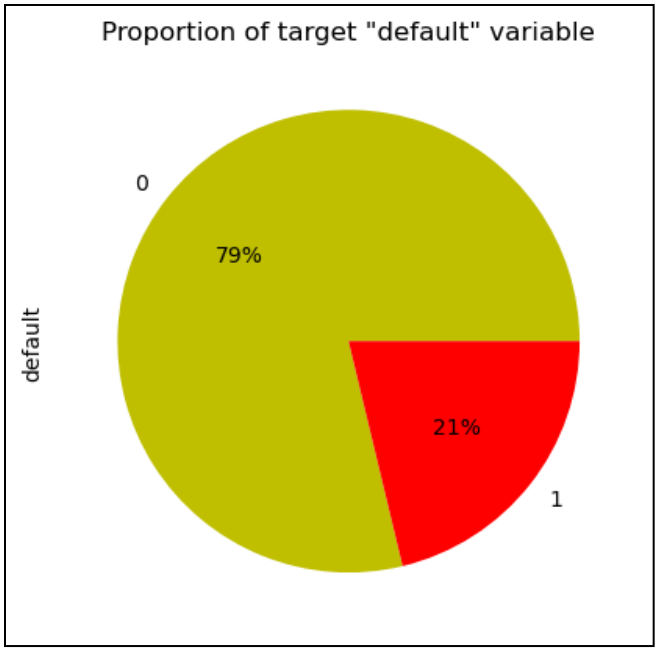

Figure 2: Proportion of default variable.

Univariate Analysis:

We used Boxplot, Distplot and Histogram to conduct the univariate Analysis and below are the findings.

- The skewness of the data has been reduced after null value imputation and the outlier treatment.
- Even if the variable data are highly skewed, we can find the noticeable skewness in the both positive and negative direction.
- There are not variable with equally distributed values in any variables but we can find some variables closer to normal distribution.
- We are not able to see any significant data distribution in Equity face value and PE on BSE values. So we can ignore these variables.
- Please refer code document for Boxplot, Distplot and Histogram for all the variables.

Bi-variate Analysis:

We plotted the boxplot of each variables against the target variable 'default'. Below are the findings.

- There are number of variables with outliers for both 0 and 1 default values.
- Networth next year (as target variable is completely rely on this variable),TOL/TNW, investments and Debt to equity ratio are showing the influence on default variable when we compare to all other variables. So we can consider as important variables when building the models.

- Equity face value and PE on BSE are not showing any significant details as the data is set to one value.

Multivariate analysis:

- We used correlation matrix and heatmap to check the coliniarity between the variables.
- We did not find any negative correlation between any variables.
- **Sales, Total income** and **Total expenses** are having positive correlation with each other.
- Below are the other few attributes with higher positive correlation.
    - **Total assets** with **capital employed** and **Total liabilities**
    - **Net worth** with **Shareholder funds**
    - **Profit after tax** with **PBT.**

We dropped PE on BSE and Equity face value attributes as there is no significant contribution to the model building process

Train and Test split:

- We used the standard Scalar method to standardize the data before we proceed.
- We split the data into train and test data with the ratio of 70:30

Value count of target '**default'** variables in train and test data.

Train data:

```
0    0.782315
1    0.217685
Name: default, dtype: float64
```

Test data:

```
0    0.808685
1    0.191315
Name: default, dtype: float64
```

4. Logistic Regression Model.

Parameters used for LR model:

- solver='newton-cg'
- max_iter=10000
- penalty='none'
- verbose=True

- n_jobs=2

Model Evaluation:

The accuracy of train data is 100%

The accuracy of test data is 99.53%

AUC and ROC for the training data:

AUC:1.00

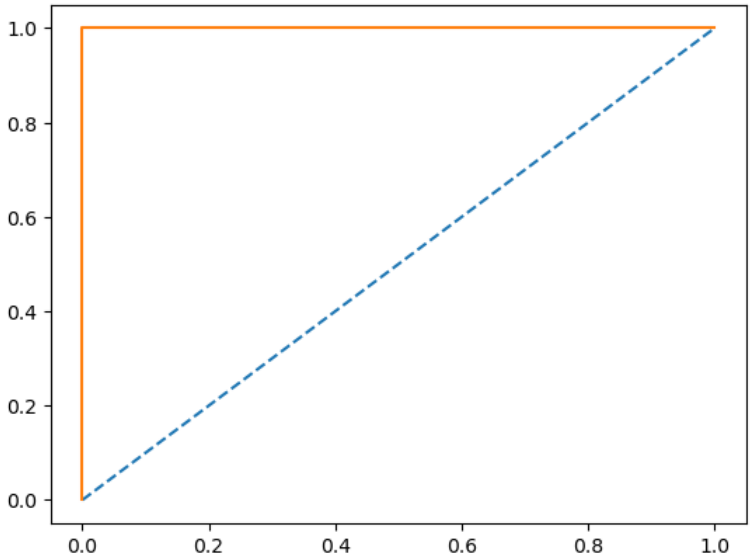

Figure 3: ROC train data- LR model

AUC and ROC for the testing data:

AUC:1.00

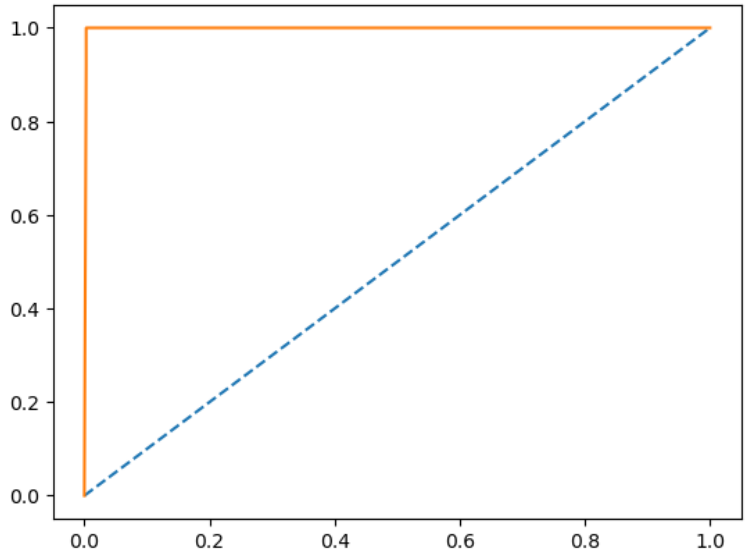

Figure 4: ROC test data- LR model

Confusion matrix and classification report of LR model:

Train Data:

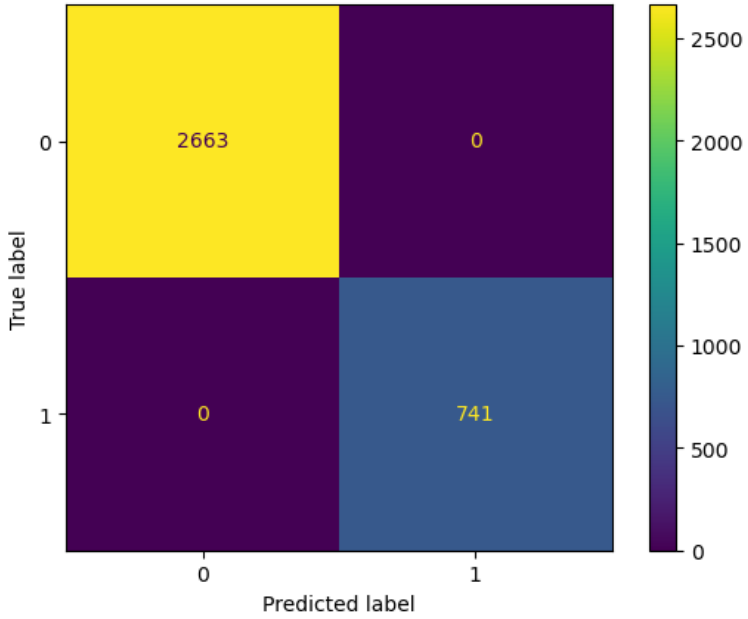

Figure 5: Confusion matrix-Train data- LR model

Classification report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2663
1	1.00	1.00	1.00	741
accuracy			1.00	3404
macro avg	1.00	1.00	1.00	3404
weighted avg	1.00	1.00	1.00	3404

Table 5: Classification report-Train data-LR model

Insights:

- The accuracy is 100%
- Recall and the f1-score also 100%

Test data:

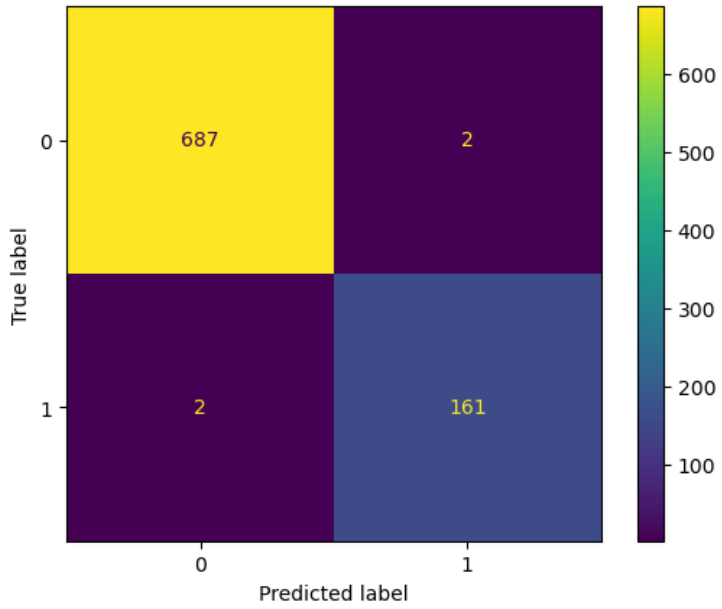

Figure 6: confusion matrix-Test data-LR model

Classification report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	689
1	0.99	0.99	0.99	163
accuracy			1.00	852
macro avg	0.99	0.99	0.99	852
weighted avg	1.00	1.00	1.00	852

Table 6: Classification report-Test data-LR model

Insights:

- The accuracy is 100%
- Precision, Recall and the f1-score is 99%

5.Random forest Model.

We used RandomForestClassifier to build the Random Forest model.

Train Accuracy of the Random Forest model: 1.0 (100%)
Test Accuracy of the Random Forest model: 1.0 (100%)

Confusion matrix and classification report of RF model:

Train Data:

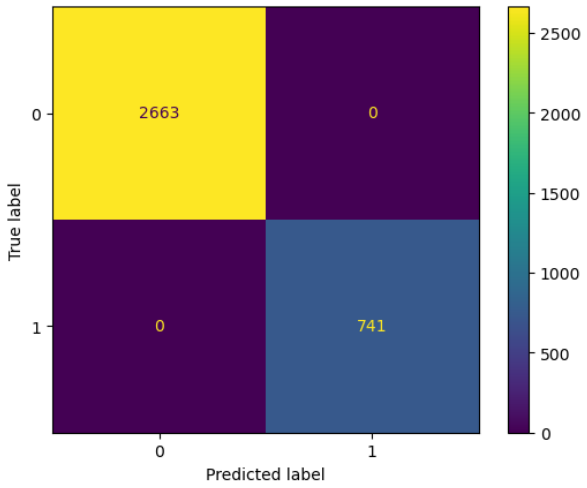

Figure 7: confusion matrix-Train data- RF model

Classification report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2663
1	1.00	1.00	1.00	741
accuracy			1.00	3404
macro avg	1.00	1.00	1.00	3404
weighted avg	1.00	1.00	1.00	3404

Table 7: Classification report-Train data-RF model

Insights:

- The accuracy is 100%
- Precision, Recall and the f1-score is 100%

Test data:

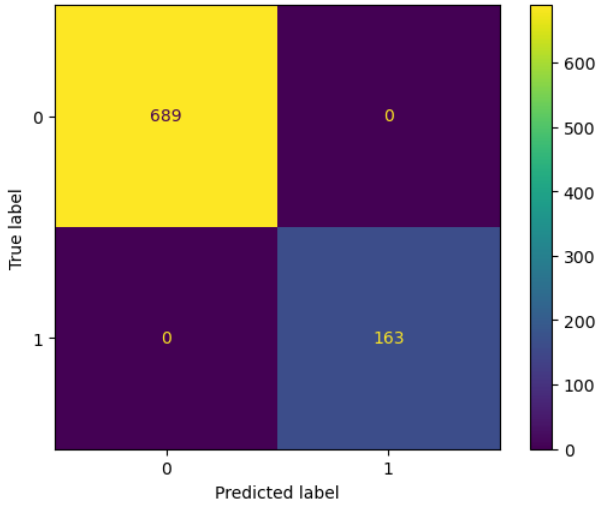

Figure 8: confusion matrix-Test data- RF model

Classification report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	689
1	1.00	1.00	1.00	163
accuracy			1.00	852
macro avg	1.00	1.00	1.00	852
weighted avg	1.00	1.00	1.00	852

Table 8: Classification report-Test data-RF model

Insights:

- The accuracy is 100%
- Precision, Recall and the f1-score is 100%

6. Final interpretation:

- Initially the Data was looking good with record counts of 51 variables and 4256 records.
- There were lots of missing values in the data.
- Shares outstanding variable having high in Std. Deviation and other parameters i.e. 3Q's and mean etc.
- In given data, not a single column is normally distributed, and data is skewed.
- There are proportion of default for "0" has 3352 and "1" has 904 counts from the total population. After defending variable that take the value of 1 when net worth next year is negative and 0 when net worth next year is positive.

- 21% of the population has showing as defaults.
- Top ten variables are highly correlated which has more than 98%.

7. Recommendations:

- As we noticed the data given to us some of the details are not completely populated mainly “PE on BSE” which has high missing values. And also there are many more missing values. Hence, if the data provided with missing values it will help us to predict with more accurately.
- Both the LR and RF models are having the accuracy of 99% and 100% respectively which meets the industrial standards which is 95%.
- Finally, we are able to achieve a descent recall value without over-fitting. Considering the opportunities such as outliers, missing values and correlated features this is a fairly good model.
- It can be improved if we get better quality data where the features explaining the default are not missing to this extent.
- Of course we can try other techniques which are not sensitive towards missing values and outliers.

---------- END----------