FLIP ROBO

Project Report
on
Micro Credit Loan Defaulters

Submitted by:

Smita More

INTRODUCTION

- A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. Microfinance services (MFS) becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The MFS provided by MFI are Group Loans, Agricultural Loans, and Individual Business Loans and so on.
- Now a days telecom industries launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.
- They understand the importance of communication and how it affects a person's life, thus focusing on providing their services and products to low income families and poor customers that can help them in the need of hour. They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in some days e.g. in 5/10/15 days.

Problem Statement

- MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days.

- For the loan amount of 5 (in Indonesian Rupiah), payback amount should be 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah).

- The sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

- So here we will build a model using classification technique which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of issuance of loan. In this case, Label '1' indicates that the loan has been payed i.e. Non- defaulter, while, Label '0' indicates that the loan has not been payed i.e. defaulter.

Exploratory Data Analysis and Data Cleaning

- First we check the information of the given dataset because it tells that how many rows and columns are present in our dataset and data type of the columns whether they are object, integer or float.

- Drop duplicates rows if present in dataset. Then we check for the null values present in our dataset.

- If null values are present then fill it via mean, median or mode. Or also you can remove that rows but kindly check it properly.

- After that we check the summary statistics of our dataset. This part tells about the statistics of our dataset i.e. mean, median, max value ,min values and also it tell whether outliers are present in our dataset or not.

- We also check the correlation of our dataset to check the correlation of the columns with each other. If columns are highly correlated with each other let's say 90% or above then remove those columns to avoid multi coli-nearity problem.

- We extract data from date column and make new columns like day, month and year to see the outcomes with our target column that is label.

- We delete the pcircle column because it has only one unique value that tells that collected data is only for one circle.

- We cannot remove outliers because data loss is more than 7%

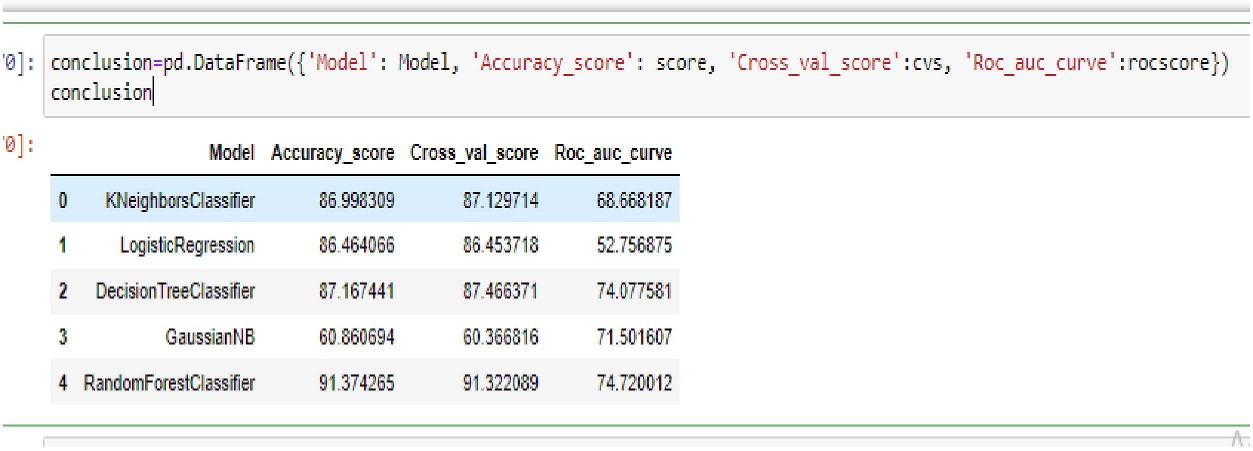
Visualization

- We plot correlation matrix via heatmap to see the correlation of the columns with other columns.
- We also visualize the correlation of columns with target column via bar graph to see which column is highly correlated with target column.
- We see the number of defaulter and non-defaulter customers with the help of count plot.
- We plot histogram to displays the shape and spread of continuous sample data.
- We also see the customers' labels i.e defaulter/Non-defaulter according to date and month with count plot.
- We also see the distribution of the data with the help of distribution plot whether it is left skewed or right skewed.

Model Building

- We know that this is classification problem so we use accuracy score, classification report and confusion matrix as our evaluation matrix. We also see the AUC score and also plot the AUC_ROC curve for our final model.
- As we know this dataset is imbalance so we don't too much focus on accuracy score. We see the precision and recall value along with f1_score.
- First we see the result without doing any sampling technique and for that I use Logistic Regression with K-Fold cross validation and hyper-parameter tuning.
- We also use Random Forest Classifier as our evaluation model without using hyper-parameter tuning because our dataset is too large and it takes more than hour to give the result.

Conclusion



```
conclusion=pd.DataFrame({'Model': Model, 'Accuracy_score': score, 'Cross_val_score':cvs, 'Roc_auc_curve':rocscore})
conclusion
```

	Model	Accuracy_score	Cross_val_score	Roc_auc_curve
0	KNeighborsClassifier	86.998309	87.129714	68.668187
1	LogisticRegression	86.464066	86.453718	52.756875
2	DecisionTreeClassifier	87.167441	87.466371	74.077581
3	GaussianNB	60.860694	60.366816	71.501607
4	RandomForestClassifier	91.374265	91.322089	74.720012

So here 'RandomForestClassifier Model' is the best model out of all model tested above and by looking this we can conclude that our model is predicting around 91% of correct results for Label '0' indicates that the loan has not been payed i.e. defaulter.