PROJECT NAME:-

[Micro Credit Defaulter Project]

SUBMITTED BY:

SHIVAM SHARMA

INTRODUCTION

Problem Statement:

A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and are very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

We are working with one such client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have 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 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).

SO HERE WE HAVE TO-

Build a model 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 insurance 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.

Review of Literature

THE TOPIC IS ABOUT THE CREDIT WHICH IS GIVEN BY THE TELECOM INDUSTRY AND IT WANTS TO MAKE A MODEL SO THEY CAN EASILY PREDICT WHICH KIND OF CUSTOMERS ARE ABLE TO RETURN THEIR LOANS OR WHICH KIND OF CUSTOMERS DO FRAUDS OR WE CAN SAY NOT RETURNING THE LOANS ..

BY THE MEANS OF ANALYSING THEY CAN EASILY UNDERSTAND TO WHICH TYPES OF PEOPLE THEY SHOULD GIVE THE LOANS.. .

Motivation for the Problem Undertaken

We Build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back so that the the client will less suffer from the defaulters who are not going to give their loans back …

Data Sources and their format

- **data.csv file downloaded from PMT**

```
arnings.filterwarnings('ignore')
```

```
ata = pd.read_csv('Data file.csv')
ata.head()
```

Unnamed: 0	label	msisdn	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	...	maxamnt_loans30	medianamı
1	0	21408I70789	272.0	3055.050000	3065.150000	220.13	260.13	2.0	0.0	...	6.0	
2	1	76462I70374	712.0	12122.000000	12124.750000	3691.26	3691.26	20.0	0.0	...	12.0	
3	1	17943I70372	535.0	1398.000000	1398.000000	900.13	900.13	3.0	0.0	...	6.0	
4	1	55773I70781	241.0	21.228000	21.228000	159.42	159.42	41.0	0.0	...	6.0	
5	1	03813I82730	947.0	150.619333	150.619333	1098.90	1098.90	4.0	0.0	...	6.0	

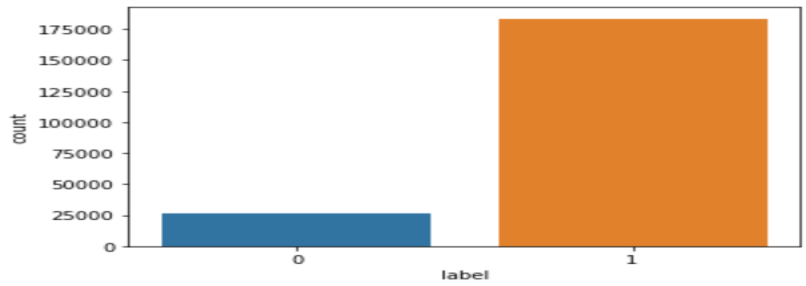
rows × 37 columns

Data Preprocessing and EDA

The steps followed for the cleaning and EDA of the data:-

1) FIRST WE ANALYSE THE DATA BY SIMPLY DATA.INFO() METHOD
2) Then we used Drop function for delete some columns from the dataset which having some IDS and (ex- Unnamed:0)
3) Then we manipulate the Dates and make it for machine usable
4) Then we analyse the data with Barplot visualization techniques with label on x axis by which it is easy to interpret the dat which types of people are defaulters or which paying the loans back.
5) After that we count the label values as we can clearly shows from the graph that it has imbalance data.

```
: sns.countplot(data['label'])
: <AxesSubplot:xlabel='label', ylabel='count'>
```



6) Then we replace some columns value which has object datatype so that they can be easily understandable by machine i.e- int/float datatypes..
7) Now we will check the correlation between the independent variables and between the independent variables and dependent variables.
8) Plot a Heatmap to check clearly the correlations between all
9) Then we check the skewness by plotting the Distribution plot and analyse the data after that we apply skew() and transform this skewness by the help of log , cbrt , sqrt transformation in many columns..

10) Then after removing the skewness , we check the outliers by plotting the Boxplot for data

11) After this we found some nans are present in the data so we fill these nans by the means of their mean as data is continous
12) We drop the columns which are highly correlated between the independents variables and zero related with the Output having greater skewness and large outliers..
13) After that methods we seprate the lables and features in different dataframe so that we can easily apply some methods on the labels and features differently..
14) After this we applied scaling on the features (x) by the help of Standard Scaler…
15) After scaling the data we use Smote techniques for oversampling the data i.e- balancing the data …
16) Now we are ready for the model building……

17) Seprate the output variable from the dataframe and train the data for the precition

Algorithm used in this project:-

For building machine learning models there are several models present inside the Sklearn module.

Sklearn provides two types of models i.e. regression and classification. Our dataset's target variable is to predict whether fraud is reported or not. So for this kind of problem we use classification models.

But before the model fitting we have to seprate the predictor and target variable, then we pass this variable to the train_test_split method to create the training set and testing set for the model training and prediction.

We can build as many models as we want to compare the accuracy given by these models and to select the best model among them.

I have selected 5 models:

1. Logistic Regression
2. Decision Tree Classifier
3. Random forest Classifier
4. KNN Classifier

5. Gradient boosting Classifier

BEST Algo. From these And why?

Random Forest Classifeir is the best algo from all of these algo which is used in this data to predict because the difference between the cross val score and the accuracy score is min for the RF algo AND IT ALSO GIVE BETTER ACCURACY(approx. 95%)that's why we USED THIS **Algo.**

- **Lets hypertune this algo with the help of Grid searchCV**
- **Save the model for later predictions**

 # Now my model is ready to predict

CONCLUSIONS

THE DATAFRAME CONTAINS THE DATA WHICH RELATED TO THE LOAN PREDICTION THAT THE LOAN WILL BE PAID BACK OR NOT BY THE CUSTOMERS SO THAT WE CAN PREDICT THE DEFAULTERS OF THAT LOAN.

We got our best model i.e RANDOM FOREST CLASSIFIER with the accuracy score of 95.2%. Here our model predicts 43877 true positive cases out of 46087 positive cases and 43416 true negative cases out of 45636 cases. It predicts 2213 false positive cases out of 46087 positive cases and 2210 false negative cases out of 45636 cases. It gives the f1 score of 95%.....

KEYFINDING:-

WE DON'T CHECK THE OUTLIERS IN THIS DATASET BECAUSE THERE IS TOO MUCH OUTLIERS PRESENT AND AS WE ARE GOING TO REMOVE THE OUTLIERS FROM THIS PROJECT WE FIND DIFFICUTLY AS IT REMOVE ALL THE ROWS AND WHEN WE SELECTING SOME COLUMNS THEN ALSO IT REMOVE WHOLE ROWS SO WE DECIDED NOT TO LOSE THE DATA SO MUCH AS THE DATA IS VERY CRUICAL THAT'S Y WE DID NOT GO TO REMOVE THE OUTLIERS FROM THIS DATA..