MICRO-CREDIT DEFAULTER MODEL

Submitted by:

Snehal Kumawat

INTRODUCTION

- ## Business Problem Framing

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients. They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. 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.

- ## Conceptual Background of the Domain Problem

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). The sample data is provided to us from our client database. It is hereby given to you for this exercise.

Review of Literature

The reviewed literature was divided in three separate parts. First, the studies that emphasize then housing price evaluation using machine learning techniques are reviewed. The second part includes the studies focusing on hedonic-based regression, and other stochastic approaches for the price prediction problem. The third part of the literature review concentrates on the studies related to price prediction model using specifically machine learning algorithms.

- ## Motivation for the Problem Undertaken

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..

- ## Data Sources and their formats

The dataset was in the form of a CSV file, so I used the read_CSV file function from the pandas module. The picture of the dataset I just have given below, you can observe that it consists of 209593 rows (records) and 37 columns (features).

```
In [2]: df=pd.read_csv("Data file.csv")
        df
```

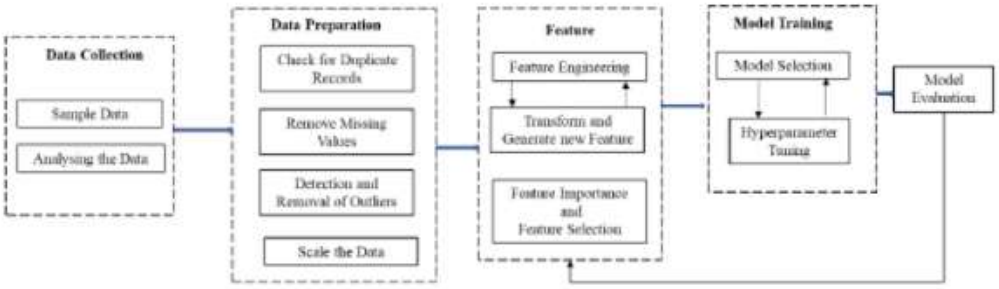
Out[2]:

	Unnamed: 0	label	msisdn	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	...	maxamnt_loans30	med
0	1	0	21408I70789	272.0	3055.050000	3065.150000	220.13	260.13	2.0	0.0	...	6.0	
1	2	1	76462I70374	712.0	12122.000000	12124.750000	3691.26	3691.26	20.0	0.0	...	12.0	
2	3	1	17943I70372	535.0	1398.000000	1398.000000	900.13	900.13	3.0	0.0	...	6.0	
3	4	1	55773I70781	241.0	21.228000	21.228000	159.42	159.42	41.0	0.0	...	6.0	
4	5	1	03813I82730	947.0	150.619333	150.619333	1098.90	1098.90	4.0	0.0	...	6.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
209588	209589	1	22758I85348	404.0	151.872333	151.872333	1089.19	1089.19	1.0	0.0	...	6.0	
209589	209590	1	95583I84455	1075.0	36.936000	36.936000	1728.36	1728.36	4.0	0.0	...	6.0	
209590	209591	1	28556I85350	1013.0	11843.111667	11904.350000	5861.83	8893.20	3.0	0.0	...	12.0	
209591	209592	1	59712I82733	1732.0	12488.228333	12574.370000	411.83	984.58	2.0	38.0	...	12.0	
209592	209593	1	65061I85339	1581.0	4489.362000	4534.820000	483.92	631.20	13.0	0.0	...	12.0	

209593 rows × 37 columns

1. Data Preparation and Cleaning

Fig. 1 captures the research framework for the housing price prediction problem. It includes five major blocks, namely data collection, data preparation, feature processing, mode training, and model evaluation. These blocks of the diagram are explained in detail in the next subsections.



In this model, we need to feed the available independent variables to the model will predict the possible price of houses. For designing the model the machine learning method I used is multiple regression model and the tool I used for coding is jupyter notebook.

About Project

Machine Learning Regression model is developed to predict the price of houses. Data contains 209593 entries each having 37 variables.

Packages used: Pandas, Numpy, Seaborn, Matplotlib, Scikit and Stats models.

2. Data Analysis

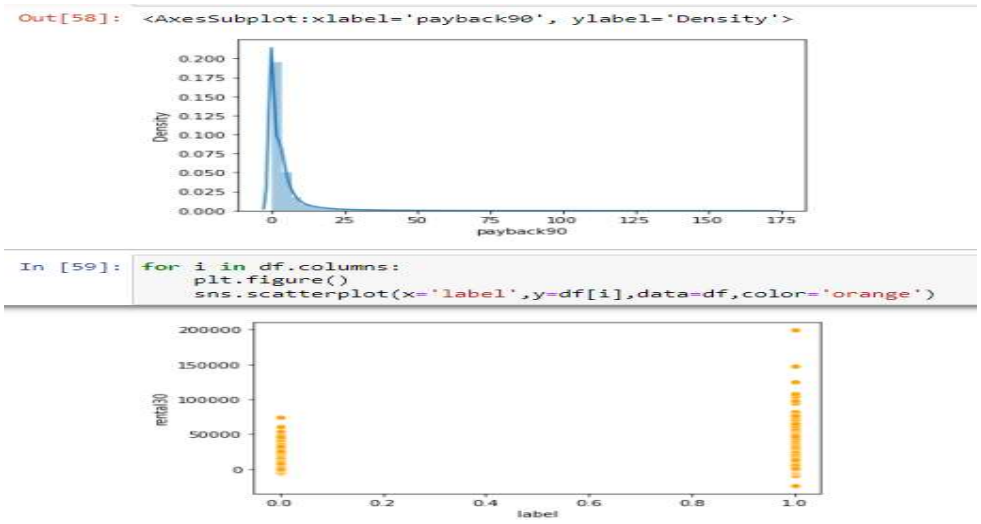
In this dataset, there are no null values. The dataset is imbalanced. Label ‘1’ has approximatel y 87.5% records, while, label ‘0’ has approximately 12.5% records. So we have balanced the dataset using SMOTE analysis. These are counts for target column “label”

1 183431
0 26162

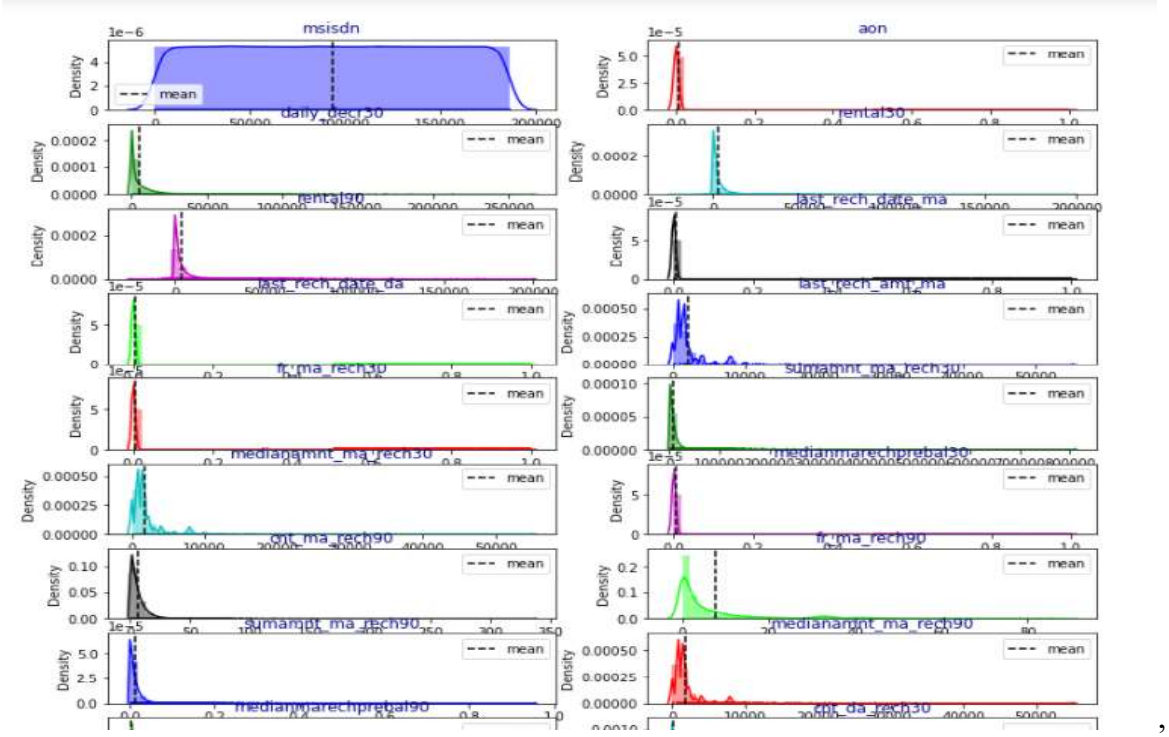
Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Let's now plot the correlation of some variables with target variable (Sale price).



After taking correlation multicollinearity is present hence we have to reduce it.

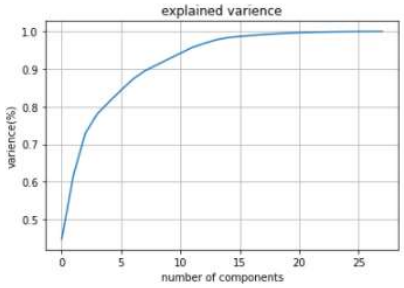


From the above plot it is observed that data is not normalized.

After that we checking the Outliers and we have not remove it as data loss is huge. So keep the data as it is. Then we scaled the data and we need to use Principal Component Analysis (PCA).

Code

```
from sklearn.decomposition import PCA

pca=PCA()

pc=pca.fit_transform(x_t)

plt.figure()

plt.plot(np.cumsum(pca.explained_variance_ratio_))

plt.xlabel("number of components")

plt.ylabel("varience(%)")

plt.title("explained varience")

plt.show()
```



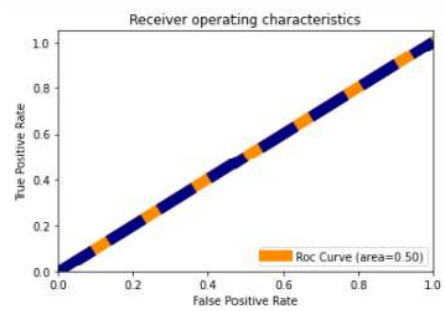
So now we selected 20 columns hence principal_x is having 20 columns and 209593 rows.

Now this train data is ready for classification.

3. Logistic Regression

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

Using this we will get output Cross validation score is:- 74.08
Accuracy_score is :- 74.085
And AUC ROC curve is as follows



4. Classification Model Building:

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

Splitting the Dataset

As usual for supervised machine learning problems, we need a training dataset to train our model and a test dataset to evaluate the model.

Code

```
x_train,x_test,y_train,y_test=train_test_split(trainx,trainy,random_state=53,test_size=0.20)
```

Decision Tree Classifier

A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learni ng where the data is continuously split according to a certain parameter.

Using Decision Tree Classifier and cross validation we got

Cross validation score is:- 81.854

Accuracy_score is :- 81.911

KNeighborsClassifier

The K in the name of this classifier **represents the k nearest neighbors**, where k is an integer value specified by the user.

Using KNeighborsClassifier and cross validation we got

Cross validation score is:- 83.391

Accuracy_score is :- 85.471

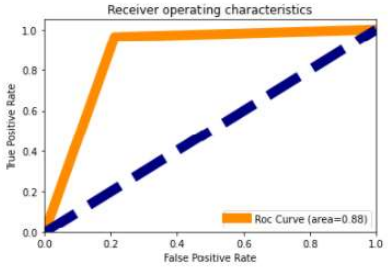
RandomForestClassifier

A random forest classifier. A random forest is **a meta estimator that fits a number of decision tree c lassifiers on various sub-samples of the dataset** and uses averaging to improve the predictive accura cy and control over-fitting. The number of trees in the forest.

We got Cross validation score is:- 87.795

Accuracy_score is :- 89.846

AUC ROC Curve



Ada Boost Classifier

We got Cross validation score is:- 87.795

Accuracy_score is :- 75.646

The best model is Random Forest Classifier.

Since the difference between the percentages score of cross validation and r2_score is optimum. After that we save the best model using pickle method.

5. Conclusion

It is seen that the most effective attribute in predicting in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of ins urance of loan and that Random Forest Classifier is the most effective model for our Dataset

And here is original and predicted values.

	original	predicted
0	0	0
1	0	0
2	1	1
3	1	1
4	0	1
...	...	...
73368	1	0
73369	0	0
73370	1	1
73371	0	0
73372	0	0

73373 rows × 2 columns

We built several classification models to predict a probability for each loan transaction. W e evaluated and compared each model to determine the one with highest performance. We als o looked at how some models rank the features according to their importance. We followed th e data science process starting with getting the data, then cleaning and pre-processing the dat a, followed by exploring the data and building models, then evaluating the results and commu nicating them with visualizations.

6. Limitations of this work and Scope for Future Work

This study, like any other, came not without limitations. Although our work was concentrated on employing real data from a micro-lending institution, we will base our experimental analysis on a more extensive data set in future works. While some broad qualitative conclusions about the importance of various features and the use of ensemble classifiers in micro-lending scenarios can be drawn from our results, the particular choice of features, etc., may not be universally applicable across other countries and other institutions. The use of an extensive data set might boost the model's performance and provide more accurate estimations. Similarly, we might control the number of outliers more efficiently while understanding machine learning algorithms' limits.