Micro Credit Defaulter Project

FLIP ROBO TECHNOLOGIES

Submitted By:-

Sujay Nimbalkar

ACKNOWLDGEMENT

Here all the data set was been provided to me and on that bases the EDA, data visualization, analysis has been taking place

we have been taken much more references while surfing on the websites too

Below in the report everything is mentioned from introduction of the project till the conclusion how it worked and finally how we got the best accuracy score for the given data set

INTRODUCTION

Business Problem Framing

The company is 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. The company know the importance of communication so they have also focused on providing product and services to low-income families. To do this, they have a collaboration with MFI to provide micro credit on mobile balances to be paid back in 5 days. The customer is considered as defaulter if he fails to pay the sum of money within the stipulated period of time. 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). Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders. In order to make sure this underserved population has a positive loan experience; company makes use of a variety of alternative data include transactional information--to predict their clients' repayment abilities.

• Conceptual Background of the Domain Problem

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders. In order to make sure this underserved population has a positive loan experience, company makes use of a variety of alternative data include transactional information--to predict their clients' repayment abilities.

• Review of Literature

After going through the dataset, things which came to my mind was the dataset was unbalanced and it cannot be determined by only accuracy score, we have to consider precision and recall or confusion matrix and AUC ROC curve and its score. The data cleaning was the next part which I have gone through. First deleted the column which are adding value to our target column and then tried to replace the outliers with various methods by reviewing each column properly. I have done the visualization of every column and target to understand the data distribution. Next checked the skewness and tried to remove the skewness of every column distinctively using log transformation and square root transformation. Lastly, scaled the data before splitting and training. Used statistical model for our dataset and have done hyper parameter tuning, tried to find the confusion matrix for each model, accuracy score, AUC ROC score and its curve and finally chose the model which was performing best for our dataset.

• Motivation for the Problem Undertaken

The initiative of the company to provide Micro credit is very noble to help the low-income group of people but there are certain people who take advantage of this noble idea and don't bother to repay the money and become defaulter. So, it is necessary to stop this type of practice. Sometime people with good intension remain deprived of getting loan from the financial institution due to some dishonest people. Hence Machine Learning can be used to predict the defaulter and non-defaulter by using different parameters.

Analytical Problem Framing • Mathematical/ Analytical Modeling of the Problem

Starting with the dataset, when I looked through the statistical description, we come to see that most of the data are unbalanced. There is high standard deviation from the mean value. The difference between the third quantile and maximum value was huge in many cases which was quite abnormal and hence I decided to replace them with Q3+1.5(IQR) if it is more than Q3+1.5(IQR). In some places the minimum values were negative which also seem to be abnormal in that case. Hence, it was replaced by Q1-1.5(IQR) if it is below the minimum value. It was found in some variables that, the maximum value was abnormally high which was replaced by a normal high number of that variable. The visualization also helped to identify the skewness present in the data. Those skewness were also corrected using

Log transformation and square root transformation. At last, after data pre-processing we come the model building section, were I used Logistic Regression, Gaussian NB and Random Forest Classifier.

• Data Sources and their formats

This data is been provided by a Telecom company. 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. The company shared around 2 lakh data of their customer with different transaction behaviour to understand and to predict their future behaviour. The data is been provided in CSV format with 37 different variables in different columns and 209593 rows.

• Data Preprocessing Done

Most of the data in the dataset was full of outliers. Those outliers were corrected by replacing them with Q3+1.5(IQR) if it is more than Q3+1.5(IQR). The data was also skewed. Some of them were negatively whereas some are positively skewed. All the skewed data was corrected using square root transformation and log transformation where ever applicable.

• Data Inputs- Logic- Output Relationships

The input data provided, helps to understand the behavior of the customer, their various transaction records, their frequency of transaction during a period of time etc, all these helps to predict the customer's intension toward the repayment of loan.

• State the set of assumptions (if any) related to the problem under consideration

No as such assumption been done related to the circumstances.

• Hardware and Software Requirements and Tools Used

Data Science task should be done with sophisticated machine with high end machine configuration. But unfortunately, the machine which I'm currently using is powered by intel core i3 processor with 4GB of RAM. With this above-mentioned configuration, I managed to work with the data set in Jupyter Notebook which help us to write Python codes. As I'm using low configuration machine so it took more time then usual to execute codes. The library used for the assignment are Numpy, Pandas, Matplotlib, Seaborn, Scikit learn

Models Development and Evaluation

• Identification of possible problem-solving approaches (methods)

The data set contain more than 2 lakh data with no null values related to the customer. The dataset is imbalanced. Label 1 has 87.5% of data whereas label 0 has approximately 12.5%. As I went through the dataset, I found lot of outliers and skewness are present in the dataset. The outliers were corrected by replacing them with Q3+1.5(IQR) if it is more than Q3+1.5(IQR). The skewness was also reduced using Log transformation and square root transformation wherever applicable. There were certain columns which had least importance with our target variable, hence those were dropped. After data cleaning and data transformation, data visualization was done to represent data graphically. At last, the most important part was to build model for the data set.

• Testing of Identified Approaches (Algorithms)

Following are the algorithms used for the training and testing: - a. Logistic Regression b. Gaussian NB c. Random Forest Classifier.

• Run and Evaluate selected models

The algorithm used are a. Logistic Regression b. Gaussian NB c. Random Forest Classifier

• Key Metrics for success in solving problem under consideration

As mentioned earlier, the dataset is unbalanced with 87.5% of label 1 and 12.5% of label 0, which made it clear that, we cannot blindly rely on accuracy score for the prediction as it can lead to biasness. Hence, I have used confusion matrix and AUC ROC curve to determine the accuracy of the model.

• Visualizations

The plots used to visualize the data are :- a. Pie Chart b. Count plot c. Dist plot d. Hist plot

Observations: -

The Pie chart and the count plot are used to represent the label 0 and 1. From the visualization we can see the imbalance distribution of data in label 1 and 0. The number of data in label 1 and 0 can also be seen. The Dist plot and the hist plot on the other hand represent the distribution of data i.e., we can determine the skewness in the data.

• Interpretation of the Results

From the dataset, it was clear that most of the customers are inclined to pay the loan as 87.5% of the customer repaid it and only 12.5% of the customers are defaulter.

CONCLUSION

• Key Findings and Conclusions of the Study

Mostly, the customers have the intension of repaying. There are certain cases, when the customers have no intension of repayment but the number of such customers are few. With the model built, we can certainly determine customers having intension of repayment or not.

• Learning Outcomes of the Study in respect of Data Science

The dataset was full of outliers, skewness and unbalanced data which was the biggest challenge to overcome. Hence data cleaning was very important to get proper prediction. I have used Logistic Regression, Gaussian NB and Random Forest Classifier. Among the three algorithms Random Forest Classifier gave the best outcome. As the dataset was unbalanced, the other algorithm may overfit and can come out with wrong prediction whereas Random forest can control overfitting and give best prediction.

• Limitations of this work and Scope for Future Work

The solution can be applied to the customer having a transaction history but the model may not perform well with customer having new profile and no transaction history. Nevertheless, the model will perform well with customer having transaction history and can predict whether a person will be a defaulter or non-defaulter. Hence, we can say that this statistical model will be helpful in future for the prediction of micro credit defaulter and non-defaulter customer.

Thank You