FLIP ROBO

Micro Credit Defaulter

Submitted by:

Ravinder Singh

ACKNOWLEDGMENT

1. Abhishek Chib Airtel (Operation Team Lead) Africa

2. https://en.wikipedia.org/wiki/Microcredit/

3. https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/

INTRODUCTION

- Business Problem Framing

  In this Problem Micro financing Institutions want to know which customer can be a defaulter in the future to avoid losses,

  Real world : Some telecome companies in India do offer micro credit on there prepaid connection to the customer in case of emergency to build more trust on the network. It's done to make a customer loyal to them

- Conceptual Background of the Domain Problem

  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

  Research done by me shows that a customer who are on the network for a longer duration tends to pay there loan.

- Motivation for the Problem Undertaken

  I have been using the same telecom company as my telecome operator(Vodafone) since 2011 and had taken micro credits back in 2012-2014 . it made me trust the Network more and later i switched from prepaid to a postpaid connection .

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Data looks Perfect for the direct modelling. But has high skewness in the target variable and Multicollinearity.

- Data Sources and their formats
  Most columns were either Integer or float.
  Except(Msisdn,pcircle,pdate

```
---  ------                 --------------  -----
 0   label                  209593 non-null  int64
 1   msisdn                 209593 non-null  object
 2   aon                    209593 non-null  float64
 3   daily_decr30           209593 non-null  float64
 4   daily_decr90           209593 non-null  float64
 5   rental30               209593 non-null  float64
 6   rental90               209593 non-null  float64
 7   last_rech_date_ma      209593 non-null  float64
 8   last_rech_date_da      209593 non-null  float64
 9   last_rech_amt_ma       209593 non-null  int64
 10  cnt_ma_rech30          209593 non-null  int64
 11  fr_ma_rech30           209593 non-null  float64
 12  sumamnt_ma_rech30      209593 non-null  float64
 13  medianamnt_ma_rech30   209593 non-null  float64
 14  medianmarechprebal30   209593 non-null  float64
 15  cnt_ma_rech90          209593 non-null  int64
 16  fr_ma_rech90           209593 non-null  int64
 17  sumamnt_ma_rech90      209593 non-null  int64
 18  medianamnt_ma_rech90   209593 non-null  float64
 19  medianmarechprebal90   209593 non-null  float64
 20  cnt_da_rech30          209593 non-null  float64
 21  fr_da_rech30           209593 non-null  float64
 22  cnt_da_rech90          209593 non-null  int64
 23  fr_da_rech90           209593 non-null  int64
 24  cnt_loans30            209593 non-null  int64
 25  amnt_loans30           209593 non-null  int64
 26  maxamnt_loans30        209593 non-null  float64
 27  medianamnt_loans30     209593 non-null  float64
 28  cnt_loans90            209593 non-null  float64
 29  amnt_loans90           209593 non-null  int64
 30  maxamnt_loans90        209593 non-null  int64
 31  medianamnt_loans90     209593 non-null  float64
 32  payback30              209593 non-null  float64
 33  payback90              209593 non-null  float64
```

Data Preprocessing Done

Remove columns where number of unique value is only 1. Let's look at no of unique values for each column. We will remove all columns where number of unique value is only 1 because that will not make any sense in the analysis

- Data Inputs- Logic- Output Relationships
  * 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- colinearity 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 more than 20% of our data will be removed. So the approach will be capping and flooring the variables using INTER QUARTILE RANGE

  After pre processing the data and normalazation it should be in a linear relation and we obtained a 100% accuracy for the same

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

  No assumptions taken

- Hardware and Software Requirements and Tools Used
  Processor – Intel i3 7th gen
  RAM – 4GB
  Hard Disk -512 Gb
  Software used
  Jupyter Notebook.
  Google colab

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Statistical:the data has a lot skewness and duplicacy is also very high with a high Multi-colinearity

  Analytical: As the data is very large its not possible to use Jupyter notebook /used google colab and tableau (free version) doesn't consider columns more than 30k

- Testing of Identified Approaches (Algorithms)

  Logistic Regression

  Guassian Naive Bayes

  Decision Tree Classifier

  Random Forest Classifier

  XgBoost Classifier

- Run and Evaluate selected models

  Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

- Key Metrics for success in solving problem under consideration

  Classification Report:A Classification report is used to measure the quality of predictions from a classification algorithm.

  Confusion Matrix: A confusion matrix is a table that is used to define the performance of a classification algorithm.

Auc-Roc Curve: The ROC curve shows the trade off between the true positive fraction (TPF) and false positive fraction (FPF) as one change the criterion for positivity

- ## Visualizations

  * We plot correlation matrix via heat-map 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 defaluter /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.

Interpretation of the Results

Was able to achieve a 100% result in detecting Defaulters

CONCLUSION

- Key Findings and Conclusions of the Study

  So here every model acted as the best model after the Random sampling(SMOTE) was done we were able to get a 100% result f results for Label ‘0’ indicates that the loan has not been payed i.e. defaulter.

- Learning Outcomes of the Study in respect of Data Science

  Random sampling can improve the results drastically.Multi-collinearity in the data should be removed.

- Limitations of this work and Scope for Future Work

  Inter quartile range for the Capping and flooring of outliers

  Removing the columns with high multicolinearity

  Feature Engineering(date)

  Normalization(standard scaler)

  Random sampling(Smote)