Micro-Credit Defaulter Model

Submitted by:

Dipta Bhattacharjee

ACKNOWLEDGMENT

I would like to express my earnest gratitude to “**FlipRobo Technologies**” that made this project possible. I would like thank my SME Mr. **Sajid Choudhary** for his guidance in building this project.

.

INTRODUCTION

- ## Business Problem Framing

We have been provided data from a Telecom client which provides micro-credit on mobile-balance (TT as well as data) to its customers. The company provides micro-credit to the customers which are to be paid within a period of 5 Days. The target is to build a model which will help the client to identify the defaulters and target customer who will be able to pay back the loan back within 5 Days based of the historical data.

- ## Conceptual Background of the Domain Problem

As mobile communication almost a basic necessity of life today, so it always important for a Telco to have a well defined strategy to increase revenue by identifying the targeted customer with low income and introduce micro loans for emergency balance top up. We can say almost all of the customers would be mostly taking loan for calling and not for internet services. So, the focus should be on the main balance .

- ## Review of Literature

As Electronic Communications is becoming a basic necessities of day to life chores. Telecom companies are making most of it to make more revenues by proving more custom solutions and they want to approach their customers for more than just providing telecom services. The Companies is lending a small amount of credit to those customers those who have their services expired and need to use the service right away without paying the complete bill at the point of time. This loan is to be paid back within a period of 5 days with a particular interest. If the customer does not pay the amount with interest within the given period of time, the consumer believed to be defaulter.

.

- ## Motivation for the Problem Undertaken

Our objective is 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 loan or not.

Analytical Problem Framing

- This a binary Classification problem with two label classes
- The target classes are highly imbalance, imbalance must be addressed
- It consists of 32 features in the dataset with are more than 200K samples in the dataset with no missing values.
- All the Data is of numerical data type.
- Almost all of the features are highly skewed, skewness is to be addresses.
- Some dirty values were there needs to cleansed
- The data is not scaled.
- As the target contains highly imbalance classes (85%-15%), we may use AUC_ROC as the primary scoring and evaluation metric.

.

• Data Sources and their formats

The data was accumulated client. This data was stacked in a spreadsheet with the values form the survey

• Data Preprocessing Done

Following data pre-processing was done on the training set:

1. Skew transformation using Cube root Transformation (cbrt).
2. Reducing Dimensions by removing the features with about 0 percent variance.
3. Over Sampling of the minority class using SMOTE. Increased minority class samples by 1.5% only.
4. Last 0.5 Percent of Quartile stripped for 2 features with highly skewed data, with loss was about 1300 records
5. Standard Scaling the data

- Data Inputs- Logic- Output Relationships

  Data is imported from the worksheets in the form of a Pandas Data frame to pre-process data and build the model. The model is evaluated on a training set using 5 fold cross validation set with **AUC_ROC** being the scoring parameter , the model which performs best on the training set is used for prediction of the classes and their respective probability per record on the test set. State the set of assumptions (if any) related to the problem under consideration

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

  N/A

- Hardware and Software Requirements and Tools Used

Hardware

- Intel Core i5 3rd Gen.
- Quad Core, 1.6 GHz Clock Speed.
- 8 GB DDR4 RAM.
- 1 TB Hard Disk.
- 13.3 inches, 1920 x 1080 pixels, Touch Screen.
- Windows 10 OS.

Software

- Junyper.
- MS WORD.
- MS Excel.

Libraries

- Numpy
- Pandas
- Scipy
- Sk-learn
- Matplotlib
- Seaborn
- Joblib

.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Describe the approaches you followed, both statistical and analytical, for solving of this problem.

- Testing of Identified Approaches (Algorithms).

  The Algorithms used for testing, training and Validating the models are as follows:

  - Logistic Regression
  - SVC (with an rbf kernel)
  - Decision Tree
  - K Nearest Neighbour
  - Random Forest
  - Gradient Boosting

  .

- Run and Evaluate selected models

```
cross_val_score(LogisticRegression(),X,y,scoring='roc_auc',n_jobs=-1)
array([0.8235133 , 0.82614822, 0.82627292, 0.82438163, 0.82576355])

cross_val_score(SVC(),X,y,scoring='roc_auc',n_jobs=-1)
array([0.74505401, 0.74409227, 0.74548028, 0.74845905, 0.74758005])

cross_val_score(DecisionTreeClassifier(),X,y,scoring='roc_auc',n_jobs=-1)
array([0.70255473, 0.70775761, 0.70902897, 0.70663978, 0.73262123])

cross_val_score(RandomForestClassifier(),X,y,scoring='roc_auc',n_jobs=-1)
array([0.8748551 , 0.8788937 , 0.87909887, 0.87893183, 0.8951725 ])
```

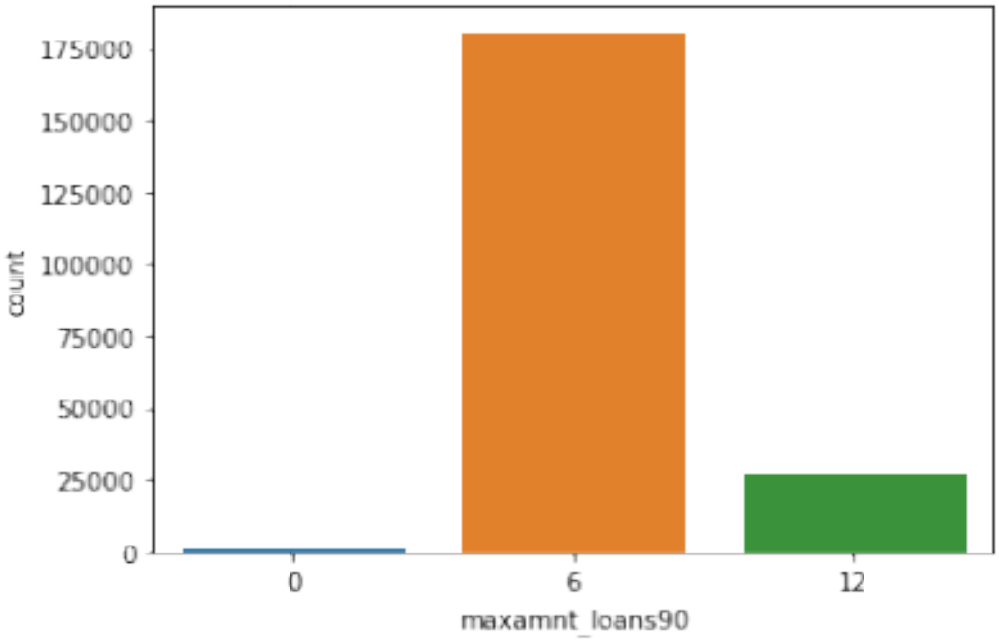
.

- Key Metrics for success in solving problem under consideration

  The key-metric under considerations is AUC_ROC although the model was finalized on basis of F1-score.

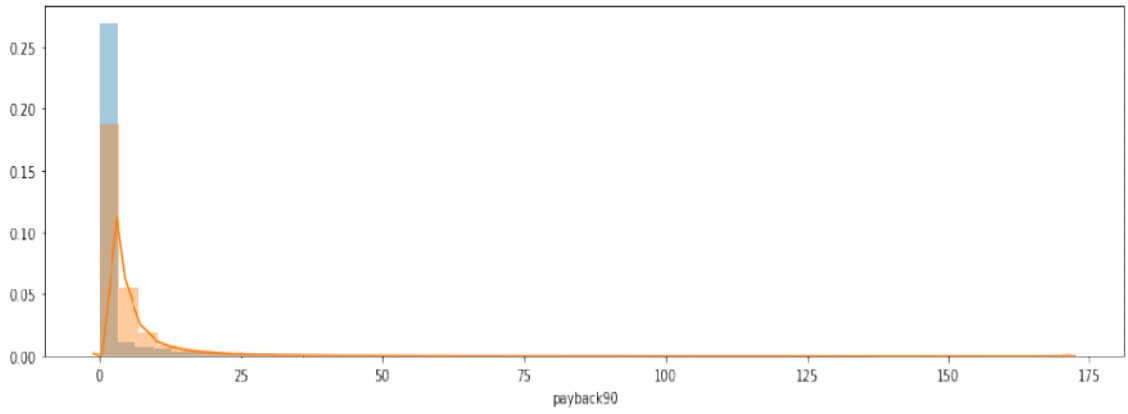
- Visualizations

Popular Loan that are preferred by the customers.



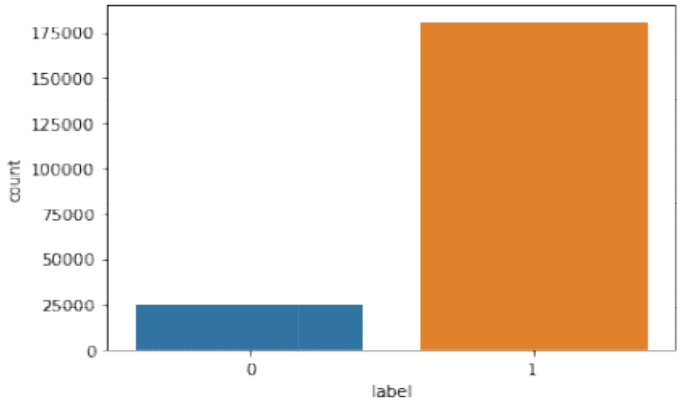
A large percentage of the population prefer loan of 6 Units than 12 Units.

plot of Defaulters vs Non Defaulters (orange: defaulters)



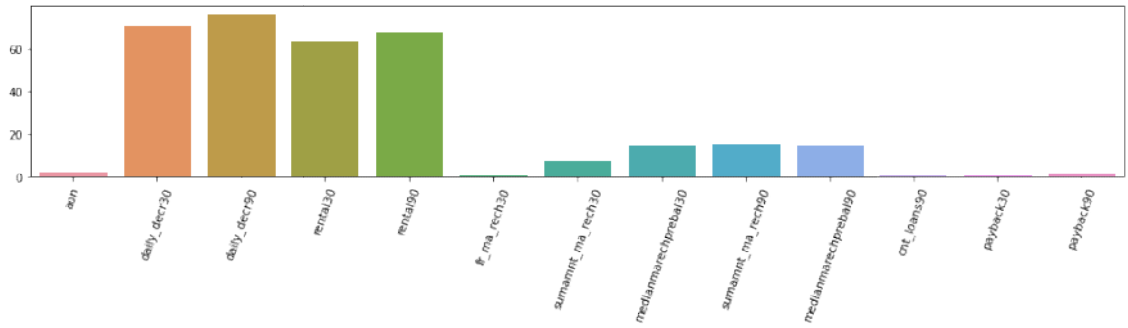
Density of the defaulter increase as the payback time increases. We can visualize that between the values 0 to 25 on X-axis.

Imbalance of the target classes



The target classes are highly imbalanced

Prominent Features chosen for data modelling



These are the selected features which contribute a weighted amount for prediction of the classes in the target .

- Interpretation of the Results

```
1 print(classification_report(y_test,y_preds))
```

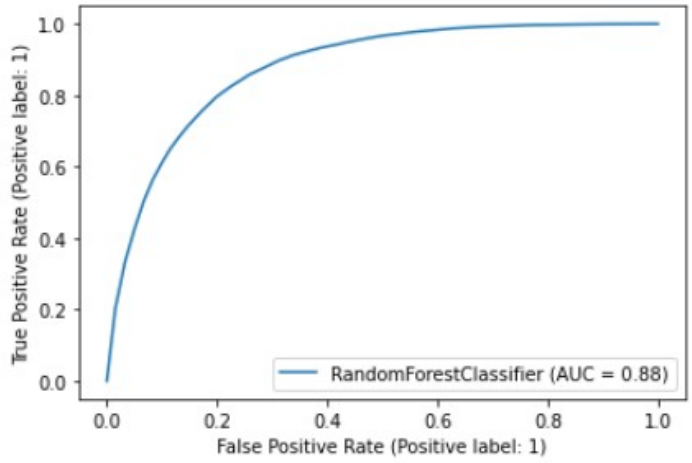
	precision	recall	f1-score	support
0	0.74	0.44	0.55	7648
1	0.93	0.98	0.95	54284
accuracy			0.91	61932
macro avg	0.83	0.71	0.75	61932
weighted avg	0.90	0.91	0.90	61932

This is the classifications report on the test set. Since we have high imbalance in our target classes we used AUC_ROC to evaluate the model.

➢ ROC curve on the test data

```
1 plot_roc_curve(best_model,X_test,y_test)  # --> ROC curve onthe test data
```

```
<sklearn.metrics._plot.roc_curve.RocCurveDisplay at 0x14f3aa71a90>
```



We received a AUC score of 88% on the test data using Random Forest.

CONCLUSION

- Key Findings and Conclusions of the Study
  - A very few of the customer take loan for Internet Services.
  - Most of the features pay their loan with interest on time.
  - Most of the population opt for the loan of 06 Units rather 12 Units.
  - Ensemble Techniques learn large data well without any extra-ordinary requirements.
- Learning Outcomes of the Study in respect of Data Science

Outcomes of the Study:

- Most efforts are in of data cleansing and EDA.
- Outliers are to be removed making a lot of data loss in not happening
- Every less than half a quantile of data may make the distributions highly skewed.
- Algorithms like Support Vector Machines and K nearest neighbours may take a long time to process large datasets.

.

- ## Limitations of this work and Scope for Future Work

  More historical data will result in a better model. Some features such as connection type 3G/4G preferred by the user may help in better detailed modelling.