NAME OF THE PROJECT

Micro-Credit Defaulter Project

Submitted by:

Chandan Kumar

FLIPROBO SME:

Ms. Khushboo Garg

ACKNOWLEDGMENT

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. And I want to express my huge gratitude to Ms. Khushboo Garg (SME Flip Robo), she is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to “Data trained” who are the reason behind my Internship at Fliprobo. Last but not least my parents who have been my backbone in every step of my life.

INTRODUCTION

- Business Problem Framing

  We have collaborated with a client from the Telecom Industry that provides micro-credit on mobile-balanced (main as well as data) to its customers. The company provides micro-credit to the customers which is to be paid within a period of 5 Days. We have to build a model that will help the client to Loan-out credit to those population 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 we are aware of the targeted customer of our client is the population with low income, we can say almost all of the customers would be mostly taking loan for calling and not for internet services. So, concentrating on the main balance should be the key here.

- ## Review of Literature

  Electronical Communication is the need of the day. For shopping for bread to hollering the Emergency Services there's no second to communication.

  The Companies belonging from the telecom industries are well aware of it, and they want to approach their customers furthermore than just providing telecom services. The Companies wants to lend 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.

  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 loan.

- ## Motivation for the Problem Undertaken

  Describe your objective behind to make this project, this domain and what is the motivation behind.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  The Problem is of Classification.

  The Data consists of 32 features in the dataset.

  There are more than 200K samples in the dataset.

  All the Data is of numerical datatype.

  Use various algorithms to build up the model.

  Almost all of the features are highly skewed, skewness is to be addresses.

  There are some samples with non-negotiable values to the respective features that must be addressed.

  The data is unscaled.

  There are no missing values in the dataset.

  The target classes are highly imbalance; imbalance must be addressed.

  The target has 2 classes only; it is a binary classification problem.

  As the target contains highly imbalance classes (85%-15%), we may use AUC_ROC as out primary scoring and evaluation metric.

- Data Sources and their formats

  The data was provided by the client to “FlipRobo Technologies”. The data is in the form of a comma separated file (CSV). The data i.e. the features and the target are in the single file.

```
mcd=pd.read_csv('Documents/Data file.csv')
```

```
mcd.shape
```

```
(209593, 37)
```

```
mcd.columns
```

```
Index(['Unnamed: 0', 'label', 'msisdn', 'aon', 'daily_decr30', 'daily_decr90',
       'rental30', 'rental90', 'last_rech_date_ma', 'last_rech_date_da',
       'last_rech_amt_ma', 'cnt_ma_rech30', 'fr_ma_rech30',
       'sumamnt_ma_rech30', 'medianamnt_ma_rech30', 'medianmarechprebal30',
       'cnt_ma_rech90', 'fr_ma_rech90', 'sumamnt_ma_rech90',
       'medianamnt_ma_rech90', 'medianmarechprebal90', 'cnt_da_rech30',
       'fr_da_rech30', 'cnt_da_rech90', 'fr_da_rech90', 'cnt_loans30',
       'amnt_loans30', 'maxamnt_loans30', 'medianamnt_loans30', 'cnt_loans90',
       'amnt_loans90', 'maxamnt_loans90', 'medianamnt_loans90', 'payback30',
       'payback90', 'pcircle', 'pdate'],
      dtype='object')
```

The different datatypes of these features are as shown in above figure. Out of all features only three features with object datatypes and rest are int64. We can note here ‘pdate’ has datatype of object instead of date time datatype.

- Data Pre-processing Done

  The Data pre-processing done is as follows:

  1. Skew transformation using Cube Root Transformation.
  2. Reducing Dimensions by removing the features with about 0 percent variance.
  3. Reducing Dimensions by removing the features that are independent of the target variable using ANOVA Test (classify)
  4. Over Sampling of the minority class using SMOTE. Increased minority class samples by 1.5% only.
  5. Last 0.5 Percent of Quantile stripped of some features, 2 to be exact (features had extremely high skew, about 1400 samples removed in total using quantile reduction)
  6. Standard Scaling the data

  (Note: All of the above processes are carried out strictly on the training dataset only and no test data is exposed to the model.)

- Data Inputs- Logic- Output Relationships

  Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

  Data is inputted in the form of a Pandas data frame to the model. The model is evaluated on a validation set using 3 Time Repeated, 5 fold Stratified cross validation set with AUC_ROC as the scoring parameter, And the model which preforms the best on the validation 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

  There are no such formal assumptions as we are Random-Forest to be the best model, producing better results than any other algorithms. Random Forests Algorithms are non-parametric and can thus handle skewed and multi-modal data as well as categorical data that are ordinal or non-ordinal.

- Hardware and Software Requirements and Tools Used

  Hardware Required:

  - A computer with a processor i3 or above.
  - More than 4 GiB of Ram.
  - GPU preferred.
  - Around 100 Mib of Storage Space.

Software Required:

- Python 3.6 or above
- Jupyter Notebook.
- Excel

Tools/Libraries Used:

1. Computing Tools:
   - Numpy
   - Pandas
   - Scipy
   - Sk-learn
2. Visualizing Tools:
   - Matplotlib
   - Seaborn
3. Saving Tools:
   - 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)

  Listing down all the algorithms used for the training and testing.

  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

- Naïve Bayes
- Random Forest
- Gradient Boosting

- Run and Evaluate selected models

```
Accuracy Score of LogisticRegression() is :
0.877205611544514

Confusion Matrix [[   517  14306]
 [   364 104281]]

Classification Report               precision    recall  f1-score   support

           0       0.59      0.03      0.07     14823
           1       0.88      1.00      0.93    104645

    accuracy                           0.88    119468
   macro avg       0.73      0.52      0.50    119468
weighted avg       0.84      0.88      0.83    119468
```

```
Accuracy Score of GaussianNB() is :
0.5529597883952188

Confusion Matrix [[13241  1582]
 [51825 52820]]

Classification Report               precision    recall  f1-score   support

           0       0.20      0.89      0.33     14823
           1       0.97      0.50      0.66    104645

    accuracy                           0.55    119468
   macro avg       0.59      0.70      0.50    119468
weighted avg       0.88      0.55      0.62    119468
```

```
Accuracy Score of DecisionTreeClassifier() is :
1.0

Confusion Matrix [[ 14823      0]
 [     0 104645]]

Classification Report               precision    recall  f1-score   support

           0       1.00      1.00      1.00     14823
           1       1.00      1.00      1.00    104645

    accuracy                           1.00    119468
   macro avg       1.00      1.00      1.00    119468
weighted avg       1.00      1.00      1.00    119468
```

```
Accuracy Score of KNeighborsClassifier(leaf_size=50) is :
0.8996886195466568

Confusion Matrix [[  5423   9400]
 [  2584 102061]]

Classification Report               precision    recall  f1-score   support

           0       0.68      0.37      0.48     14823
           1       0.92      0.98      0.94    104645

    accuracy                           0.90    119468
   macro avg       0.80      0.67      0.71    119468
weighted avg       0.89      0.90      0.89    119468
```

```
Accuracy Score of SVC() is :
0.8759249338735059

Confusion Matrix [[     0  14823]
 [     0 104645]]

Classification Report               precision    recall  f1-score   support

           0       0.00      0.00      0.00     14823
           1       0.88      1.00      0.93    104645

    accuracy                           0.88    119468
   macro avg       0.44      0.50      0.47    119468
weighted avg       0.77      0.88      0.82    119468
```

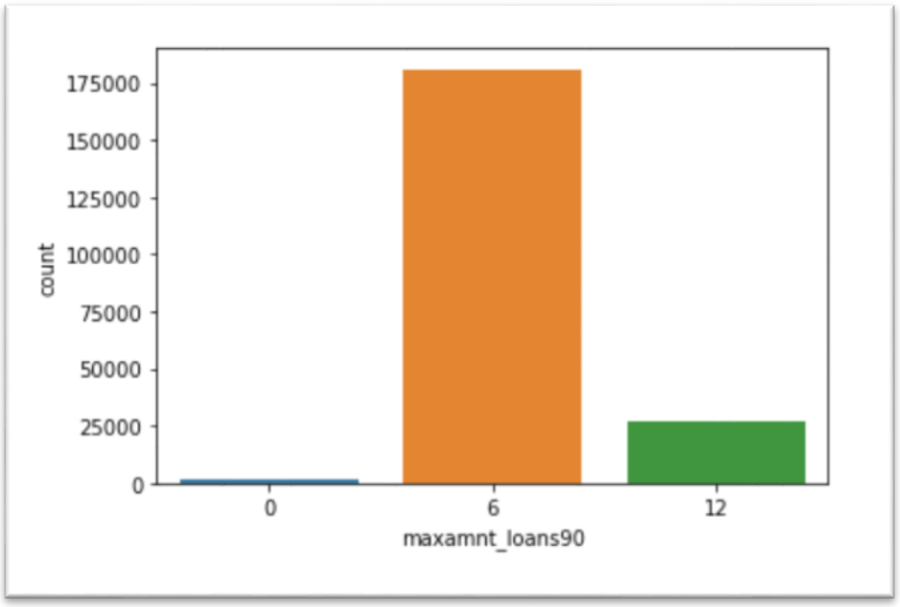
```
Accuracy Score of RandomForestClassifier(criterion='entropy', min_samples_leaf=4,
                       n_estimators=150) is :
0.954874945592125

Confusion Matrix [[ 10316   4507]
 [   884 103761]]

Classification Report               precision    recall  f1-score   support

           0       0.92      0.70      0.79     14823
           1       0.96      0.99      0.97    104645

    accuracy                           0.95    119468
   macro avg       0.94      0.84      0.88    119468
weighted avg       0.95      0.95      0.95    119468
```

- Key Metrics for success in solving problem under consideration

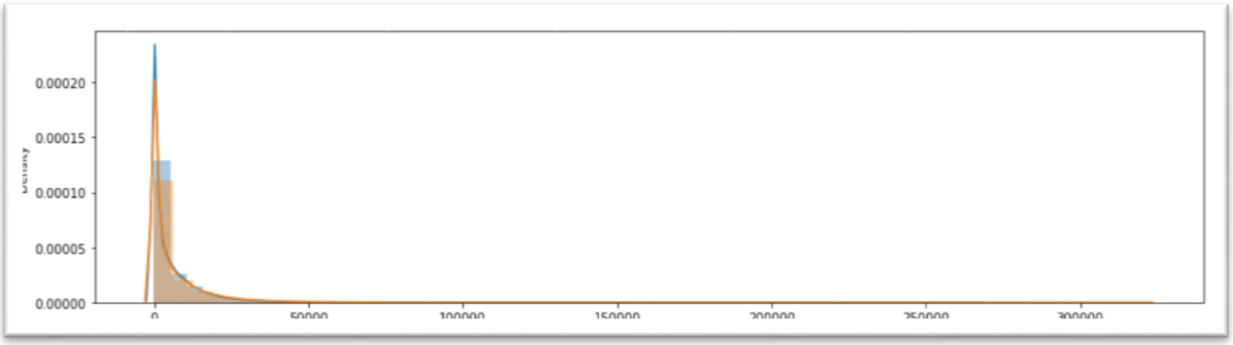
  The key-metric under considerations is AUC_ROC although the model was finalized on basis of other metrics as Matthew's Correlation Coefficient (MCC) as well as F1-score.

- Visualizations

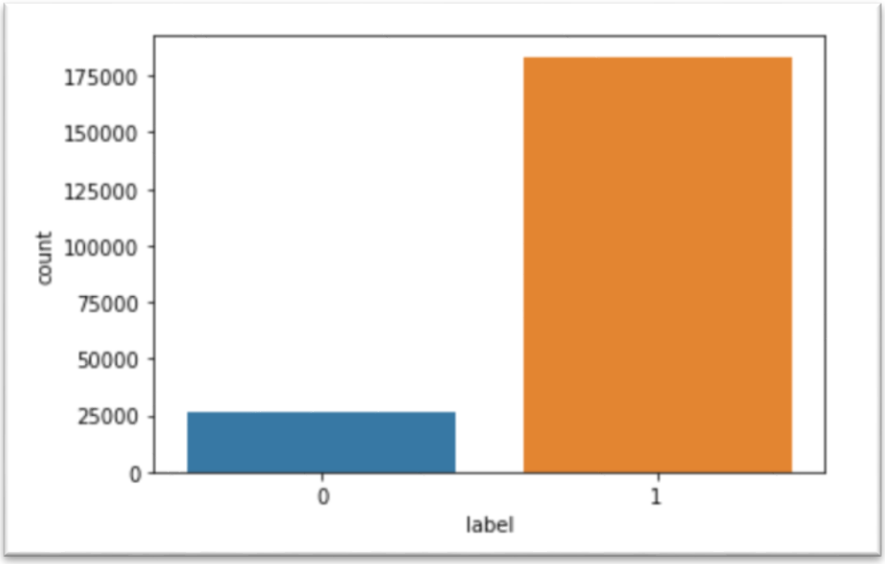
  Popular Loan that are preferred by the customers.



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



➢ Imbalance of the target classes



- Interpretation of the Results

```
lr_TS_auc=roc_auc_score(y_test,y_pred)
lr_TR_auc=roc_auc_score(y_train,train_pred)
```

```
lr_TS_auc,lr_TR_auc
```

```
(0.514993147774016, 0.5138069619192319)
```

```
fpr,tpr,thresholds=roc_curve(y_test,y_pred)
```

```
fpr
```

```
array([0.        , 0.96655199, 1.        ])
```

```
tpr
```

```
array([0.        , 0.99653828, 1.        ])
```

```
thresholds
```

```
array([2, 1, 0], dtype=int64)
```

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.

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

  List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

  Outcomes of the Study:

  - Almost 90 percent of the time is spent of data cleaning and data modelling.
  - Outliers are not to be removed casually as they may contribute to the model and our predictions.
  - You do not get a Gaussian distribution in real-word problem.
  - 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 converge on a Hugh dataset like this.
  - Naïve Bays is very quick as of converging rate.

- Limitations of this work and Scope for Future Work

  A data with statistics of beyond 90 days would be even better for data analysis.

  Some features such as network-speed received by the user, charges paid for roaming by the customer, may help in better detailed modelling.

  The model could be integrated with the analytics app used by the Data Analysts and Statistician of the respective telecom provider for easy decision.

  The model could be integrated with the mobile bot that lends loan to the customer to directly fetch the data from the database from the company servers of the particular customer, make predictions and allow or deny the customer loan bases on its predictions.

  Thank You