Project Report:

Micro Credit Loan Defaulter

Submitted by:

RAJ VAGHASIA

INTRODUCTION

- Business Problem Framing

A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and are very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

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.

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)

- Conceptual Background of the Domain Problem

  A telecom industry giving the loans for a period of 5 days which has to be repaid within 5 days. The loan provided are amount of 5 Indonesian Rupiah or 10 Indonesian Rupiah and the payback is 6 and 12 Indonesian Rupiah correspondingly. If the person is not able to clear the due, he is set as “Defaulter”.

- Motivation for the Problem Undertaken

  The motive to take on the project differs from person to person and one can invent the solution at bettering the assessment to provide micro loans while other person can take on the challenge to outreach the service to unrecognized remote area and compensate the outstanding loans.

Analytical Problem Framing

- Data Preprocessing

  - On the first sight of data, we checked for null values and the datatypes of features.
  - Separated object columns from numerical/continuous columns.
  - Checked for the duplicate features and rows.
  - Attributes with constant value are dropped which are irrelevant for algorithms.
  - Assesses the correlation within the features and highly correlated features are dropped.
  - Separated the date format into individual bundle.
  - Take a look at the correlation of independent variables with target variables by Pearson Coefficient Correlation matrix.

- Visualise the outliers via box-plot and remove the ones that are extreme of IQR to avoid loss of data.
- Check for the skewness and apply log transformation.
- Apply label encoding to categorical values.
- Lastly, for the preprocessing data execute standard scaling to unify the value range.

- 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.

  The data inputs describes the relative information of customer behaviour for taking loans during the course of 3 months i.e. Jun-Jul-Aug of the year 2016. The data also includes the relative information of amount spend from main account and its frequency.

  The output is in the form of flag (1: successful, 0: failure) indicating whether user paid back the credit amount within 5 days of issuing the loan.

  Ideally, the payback amount should be cleared or be equal to loan amount to be successful but the varying feature like frequency of data account getting recharged, amount of loan taken, account balance and age of cellular network affect the behaviour of customer willingness to payback the amount.

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

  - The main assumption considered to this problem is that the dataset is balanced.
  - Data of customer mobile number repeating are removed.
  - Since the data is believed to be collected from 2016, feature representing year is dropped.
  - Final model is considered based on ROC_AUC_Curve.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Statistically, we have put more influence on avoiding multicollinearity by building Pearson Matrix and dropping the column having high relation. There are other methods to build hypothesis around the same matter and can be checked using p-value significance.

  Analytically, we checked about the features having same value but there were none. Then we proceed to detect the outliers subjected to limit the data loss of no more than 8% - 10%.

- Testing of Identified Approaches (Algorithms)

  Since the project is based on binary classification algorithms used are

  The base model – Logistic Regression

  Decision Tree Classifier

  Random Forest Classifier

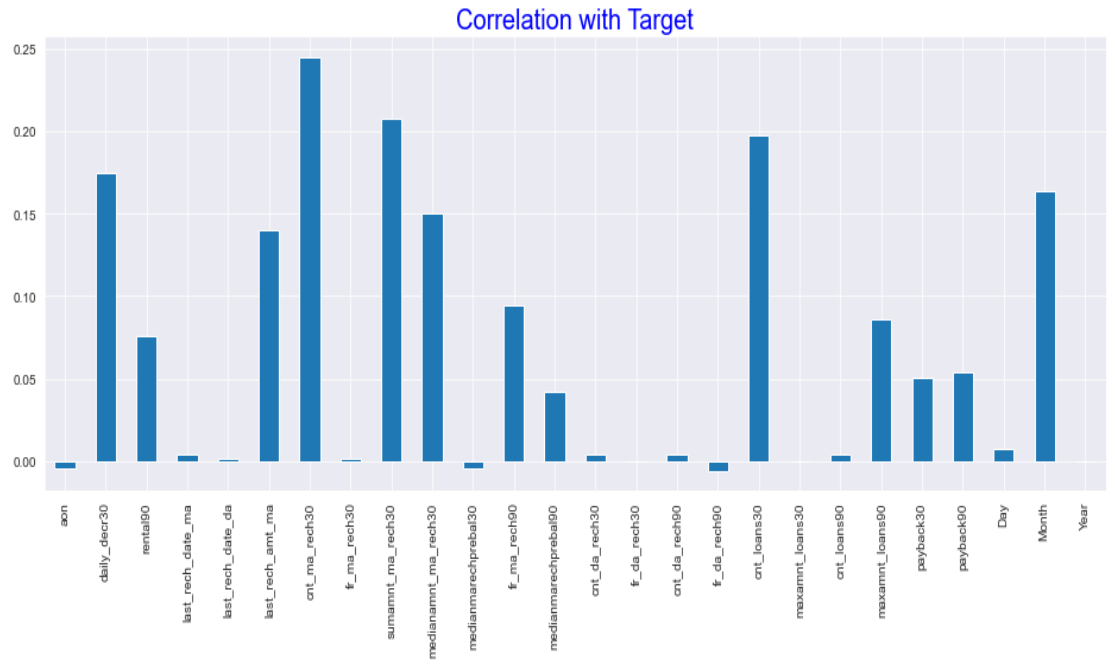
  SGD Classifier

  AdaBoost Classifier

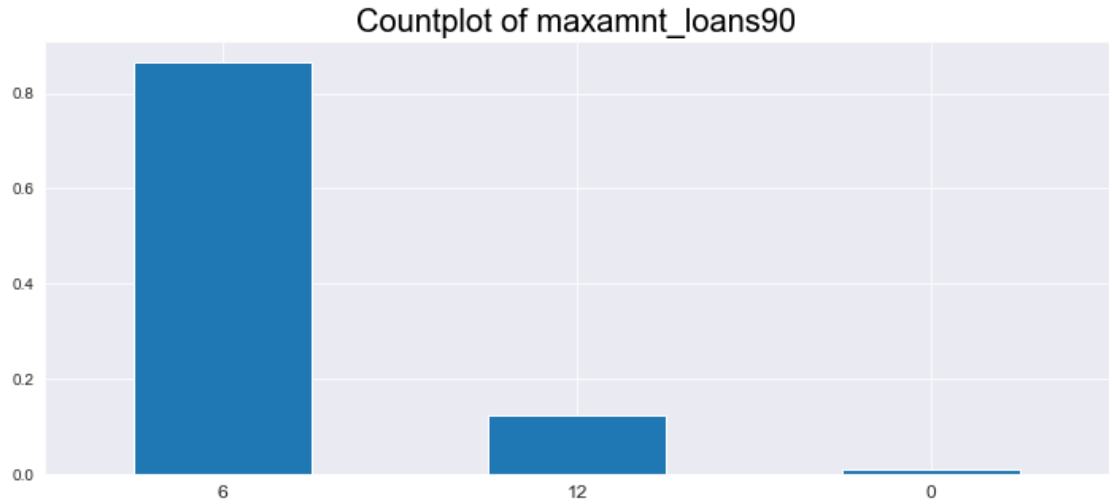
- Key Metrics for success in solving problem under consideration

  The key metrics used to evaluate the accuracy is highly relied on AUC_ROC_curve and cross validation. Based on these metrics, we are able to decide whether the score obtained on testing dataset is considered as final.

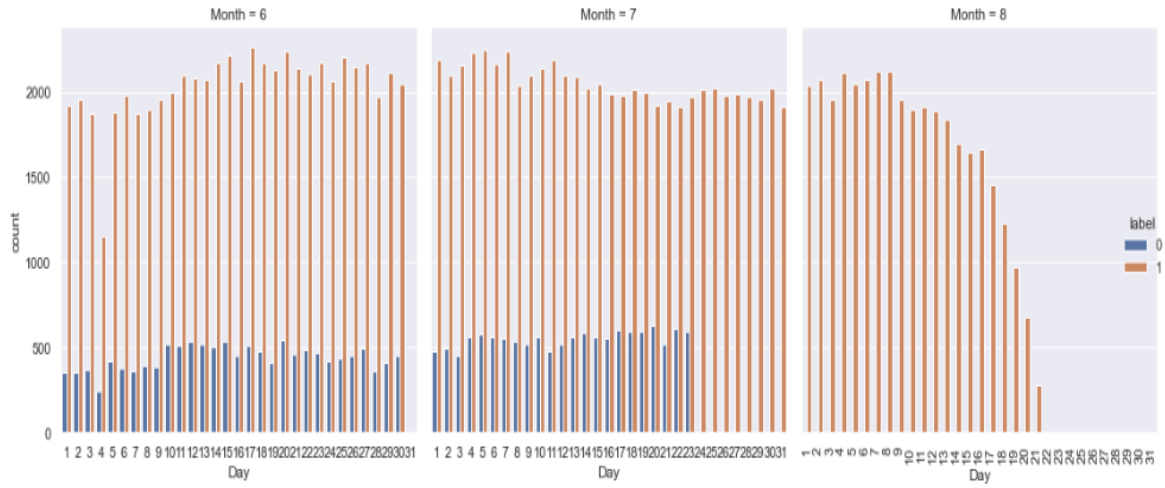
- Visualizations



The above bar plot shows the correlation of independent variables with the target variable. The feature having stronger relation is names 'cnt_ma_rech30' which shows the data of users indicating number of times main account got recharged within last 30 days.

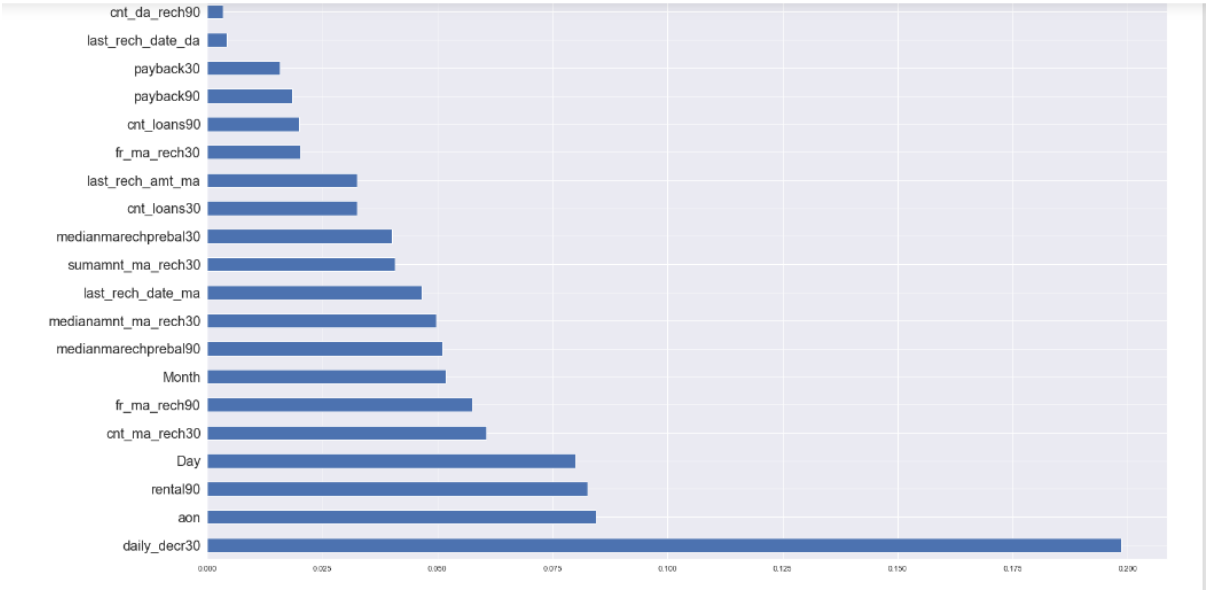


This countplot shows that the more than 80% of the user have maximum amount of loan 6(Indonesian Rupiah) while around 10% have a loan of 12.



Above plot shows the comparison of three months and gives the count of users identified as non-defaulter(1) / defaulter(0). We can infer that month 7 has seen high number of defaulters whereas June has varying number of amount of defaulters.

We can also infer that after 21/08, no user has issued a loan amount.



The above plot represents the features on y-axis and x-axis represents the importance indicator with respect to Extra Tree Classifier algorithm.

- Interpretation of the Results

  Since the problem is based on binary classification, the ultimate goal is to check on which independent feature affects the model predictive capability. For, that we have taken couple of measures. The first being the relation within independent features and to what extent it affects the prediction. For this we have constructed heatmap which excludes target variable.

  The second important thing we have implemented is to build a barchart of feature importance under the Extra Tree Classifier which gives machine learning an added benefit to build generalized model to avoid variance.

  Based on this insight we can select feature having highest importance and analyse the data of that feature and try to infer any pattern generating out of it. Results can vary based on the parameters provided and apply cross-validation to build error-proof model.

  Since, it is a classification problem and assuming that the dataset is balanced, momentarily we can rely on f1-score and ROC score.

CONCLUSION

- Key Findings and Conclusions of the Study

  Describe the key findings, inferences, observations from the whole problem.

  The key findings are most of the users have taken a loan amount of 5(Indonesian Rupiah) and the generally users have issued the loans in the 2^{nd} week where we find the max amount of defaulters.

  Given this fact, there are other data which shows that they differ from the normal group of users.

- Limitations of this work and Scope for Future Work

  We are restricted to see if the accuracy obtained stands to the mark on unseen data difficult to filter out the noise as it could affect the balance on bias and variance. We can study the analysis by plotting the relation plot to link the pattern but the main problem lies to train the model since the dataset is imbalance which increases the chance of overfitting and that is the case.