Micro Credit Loan Use Case

Submitted by :- Parv Sharda

Introduction

- ## Business Problem Facing

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. 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 have been provided with this dataset by our 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 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).
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 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.

• Conceptual Background of the domain Problem

We understand that this is a classification problem as in our target variable with have only two categories where 1 denotes "Loan paid by the customer within 5 days period" whereas 0 denotes " Loan not paid by the customer within 5 days"

In the real world we can relate this problem with online loan / credit services provided by different applications on our smartphone and they are using ML to predict the success and failure rate.

• Review of Literature

During the EDA and visualization part we noticed that almost in our Target column 80 % of the values are 1 which denotes that almost 80% of the applicants are returning the loans within 5 days period.

Hence, we understood that success rate is high.

- ### Motivation for the Problem Undertaking

As we understand from our research that success rate is high, my motivation is to decrease the risk by building the best model which can provide us the most accurate answers whether we should offer the loan to the applicant or not so that our client's business doesn't suffer.

Analytical Problem Framing

- ### Mathematical/Analytical Modeling of the problem

On checking the statistical summary of the data set we found that there is a huge difference between the standard deviation and mean of each feature which denotes that the data is highly spreaded.

- ### Data Sources and their formats

Data is provided by the client in Excel format.

Feature 'Label' is our target column where 0 denotes: not paid and 1 denotes paid. We got to know that we have 2 object, 1 datatime64, 14 integer and rest 20 float type features.

- ### Data Pre-processing Done

In the data pre-processing part our first task was to handle the features which were not in float or integer data time as numeric values are mandatory for model building.
From data.info() function we got to know that we have 2 object, 1 datatime64, 14 integer and rest 20 float type features.

By checking unique values in Pcircle feature, we got to know that this feature only contains 1 unique value i.e. “UPW”, instead of converting it into a numeric column using Label Encoding I found dropping this column a good idea.

Then we had feature ‘pdate’ in DateTime datatype. Checked unique values using data.unique() function where I found that this data is for the year 2016. using .dt.day/dt.month function I successfully extracted the Date and Month from this and as the data is from the same year 2016 I didn’t extracted that. After extracting the required values I dropped ‘pdate’ column as well.

Our last non numeric column was ‘msisdn’ which was in Object datatype. After checking the unique values in this column we found that all the unique values were numeric however in the middle of the number there was I sign. It was not sure what I denote and hence instead of removing or replacing it with some other number was not a good idea. Hence, we dropped this column as well.

From above information, we can understand that we dropped all 3 non numeric columns.

We had one more feature under the name of “Unnamed: 0’, after checking unique values in this feature I found that this was just the indexing of each row. Hence dropped this feature as well.

There was huge skewness present in the data set.

I tried to remove the skewness using different methods such as np.log, np.sqrt, power transform, cuberoot transformation and min max scaler however no satisfactory output, hence I decided to continue with this.

By using boxplot method, we understand that there are extreme outliers in our dataset. By using z-score methods to trim/remove outliers we were losing 23% data, which was not good for our model.

Data Inputs - Logic – Output Relationships

We found relationship between our target variable and the features.

By looking at our output we understand that out target is most correlated with cnt_ma_rech30, cnt_ma_rech90 and so on.
Whereas on the other end features like fr_da_rech90, medianmarechprebal30 and so on are most negatively correlated with our target feature

Hardware and Software Requirements and Tools used

1. Anaconda Framework
2. Jupytor
3. Pandas
4. Numpy
5. Seaborn
6. Matplotlib
7. Warnings
8. Sklearn
9. Scipy
10. Pickle

Model Development and Evaluation

Our first step was to split our dataset into x and y variables.

After splitting the data-set we found best Random state for our model, it was 1 hence I continued with the model building process using the same random state

We used four different methods for our model building to check which one works the best.

1. Logistic Regression
2. Random Forest Classifier
3. Decision Tree Classifier
4. SVM

Performance of above models with Cross validation

1. Logistic Regression 86%
2. Decision Tree Classifier 88%
3. Random Forest Classifier 91 %
4. SVC 87%

From the above results we understand that the best model is Random Forest Classifier, I then performed hyper parameter tuning using GridSearchCV to check whether I can improve the model performance.

The result after applying GridSearchCV was same.