

Micro Credit Loan

Submitted by:

Mayank Kumar

**ACKNOWLEDGMENT**

I would like to express my deepest appreciation to all those who provided me the possibility to take a step in this project. I gone through google and Kaggle for the references and my old projects helps me understanding the way to solve this use case.

**INTRODUCTION**

* Business Problem Framing

There are multiple telecommunication company that are providing micro loan credit to their customers which help the low financial population not to disconnect from the commination world. So, in this use case also there are multiple customers which taken loan and some don’t so I have to predict by taking important features that whether in future customers having this type of inputs and history will able to pay back the loan or not. So, this a real-world use case which will help telecom industry to provide the loan to which type of customers.

* Conceptual Background of the Domain Problem

From my prospective, primarily is to take the consideration of the region that from which place the customers are belonging. Because in the remote area most of the customers required the Micro credit and also check their history whether they are able to pay back the loan, In current scenario most of the people are able to pay back the loan in the remote area also because they are using the phones and also a good possibility to pay back the loan and this will increase the business for the telecom industry.

* Review of Literature

Micro Credit Loan policy is offered by MFI which is very useful for the low income population, In this current era everyone need to connect virtually so this policy provide the option to low income people to take a next step in this, In India also Jio and BSNL are providing this type of facility to their low income customers which are in need of loan if they don’t have sufficient amount and can payback the amount. So, I also go through some telecom sites and see that most of the remote people and low income population taking the benefit for this, and for sure in the future these telecom industry will be able to provide loan on their smart package i.e. 3 month package (include data and call) so this will help both customers and most important the industry, but before bringing this policy company should able to find a way that if they provide the loan on smart packages then how they will get the payback because 3 months is big time, customers may be change the network.

* Motivation for the Problem Undertaken

I choose this project because the dataset is very large and having multiple features which required lot of statistical analysis, feature engineering and Exploratory Data Analysis and primarily this project may help the telecom industry to find a way that to which customers did we provide the loan and is this policy is valuable for us or not? because this policy if works well then this will provide large scale benefit.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

1. Dataset contain around “209592” rows and 36 columns which make this dataset huge and required lot of feature engineering before applying the proper model on it.
2. As the data is numerical and most of the fields are continuous in nature so, I checked whether the data is normally distributed or not so, I came up with the result is that almost all features are skewed in nature.
3. Also, there are lot of outliers in the dataset which need to taken care of and most important is that most of features contain negative values which are not required according to requirement of the project use case.
4. There is proper mention that the loan amount is 5 and 10 and the payback amount is 6 and 12, but in the dataset this feature includes 0 value also which is consider as error in the dataset.
5. Also, all the amount mainly the sum amount of last 30 and 90 days and their median are giving on Indonesian currency.
6. Dataset is also imbalance and mostly contain the label as 1 which defines that most of customers pay back the loan, due to imbalance dataset the model is bias towards the label as 1, so this thing is also need to take care, so I used Tress which are not sensitive to imbalance dataset and does not bias.

* Data Sources and their formats

1. Features inside the data set is continuous and most of the features and numeric in nature.
2. In the dataset there are no null values that is good sign that gathering this data all fields and column are taken care off.

A picture containing table

Description automatically generated

So above image described the dataset i.e. total features, total count, their types and memory usage of the dataset.

* Data Pre-processing Done

1. Dataset involved the feature that loan amount is 5 and 10 and the payback amount is 6 and 12, but in the dataset this feature includes 0 value also which is consider as error in the dataset and I cleaned that unwanted data.
2. Also, most of the amount currency is Indonesian instead of “medianmarechprebal90”, so I converted it into Indonesian rupiah by multiplying it by 196.
3. Also, most of the features contain the negative values which does not play a significant role by seeing the use case, so I replace the negative value with the 0.
4. There are multiple features which are highly positive corelated with each other so by plotting the correlation heatmap I tried to remove the high positive correlated feature and using one of them.
5. Outliers can be responsible to distract the best fit outcome so need to be handle, So I used IQR and replace the outliers with the third standard deviation range which covers 99.7% of the data.

* Data Inputs- Logic- Output Relationships

By using seaborn heatmap I tried to plot correlation heatmap which show the covariance of each feature to each other.

There are 11 features which are positively correlated with each other more then 80%, means if one feature is increasing the second one is also increasing,

So, these features need to remove which reduce the problem of overfitting.

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

As I mentioned, features with negative value I removed that, these features are “age of network” age should not possible in negative etc.

Also, I assumed that all the features which contain the currency should be in same scale and their currency should be same which will help in feature scaling.

* Hardware and Software Requirements and Tools Used

Graphical user interface, text, application

Description automatically generated

Above are the libraries which I used to pre-process, predict and visualize the project.

Pandas - It is used to play with data frame and helps in to get more insight of the data, like describing the data and the types of the all features.

Numpy – It is used for doing the mathematical and statistical problem which are finding the mean, standard deviation and variance, through this we can find the presence of outliers.

Seaborn, Matplotlib – This is a visualization library which helps to plot different type of bar charts and line charts which help to get more information from the dataset.

Sklearn – This library is most important library which include all the metrics, accuracy, and the model which help to predict the output all classification and regression algorithms.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I followed the path of lifecycle of Data Science which includes:

Diagram

Description automatically generated

1. Read the use case and search from the google to know more about the use case and micro credit policy which are provided by the telecom industry whether they are helpful for the business growth or not and in future where this policy stand.
2. As data set is already there, so this step is excluded.
3. Data Preparation includes the data cleaning and describing the data which I followed and had done and remove the error data and replace the negative values in the dataset.
4. EDA involves the visualization which is helpful to get more insight from the data and get to know about the trend of features individually and in group, Also I had done the feature engineering and feature selection which involves to deal with the outliers and which features are more important and highly covariant.
5. Modelling involves creating the model with suitable algorithm which provide the best result, I tried multiple algo and also apply hyperparameter tuning and cross validation as the data set is imbalance so that’s why cross validation I used, Also to maintain the balance of the dataset I used SMOTE Tomek which handle the imbalance dataset.
6. Model Evaluation, for this I used confusion metrices and mainly focus on False positive and tried to reduce the False positive which is type 2 error and plot ROCAUC curve which is also covering most of the area under curve.

* Testing of Identified Approaches (Algorithms)

Algorithm Used:

1. Logistic Regression
2. Random Forest
3. Gradient Boosting
4. SVC
5. Extra Tree Classifier
6. KNN

* Run and Evaluate selected models

Table

Description automatically generated

A picture containing text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

A picture containing text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

A picture containing text

Description automatically generated

* Key Metrics for success in solving problem under consideration

1. In classification problem there are various metrics that are accuracy score, confusion matrix, classification repot, Roc Auc curve which help to check the efficiency of the model
2. Which metrices is useful is also depend and vary on domain, so as per the use case we must predict that whether the customer is defaulter or not
3. So, in this case accuracy score is good but most important is confusion matrix in which we must decrease the False Positive that is type 2 error.

* Visualizations

Plots:

**Bar Plot**

**Count Plot**

**Line Plot**

**QQ Plot**

**Heat Map**

Chart, bar chart

Description automatically generated

Most of the label is bias with label = 1, we can clearly see the count of label 1 is high more than 90%.

Chart, bar chart

Description automatically generated

From the above graph we can clearly see that there is no defaulter in 8th month i.e. August month of 2016 year.

In June month there are more defaulter as compared to non-defaulter and their age on cellular network is also more.

Chart, bar chart

Description automatically generated

Frequency of data account recharged in last 90 days is more in month July and least in June.

Also, in July month data account recharge in those account which are defaulter.

Chart, bar chart

Description automatically generated

Most of the count loan which are taken in last 30 days is in the month of July.

June and August is having somehow same scale of count in last 30 days.

Chart, line chart

Description automatically generated

Most of the users are taken loan in the month of July and decrease in august month and max count of loan is 1 and 2

Graphical user interface, application, Word

Description automatically generated

In the above count plot, most of the user take loan 1,2 and 3 time, these cover 90% of the total data. So, the data is for one year that is 2016 and especially for 3 months that are June, July an August so the count of taking loan is less.

A picture containing graphical user interface

Description automatically generated

These are the total loan amount that user had mostly are 6,12,15,24

Chart, line chart

Description automatically generated

I also use QQ plot to check whether the features are normally distributed or not there are 30+ features so I only attach graph for 2 plots, and we can clearly see that the graph is skewed. Checking the graph shape is useful when the features are continuous and numeric.

A picture containing text, scoreboard

Description automatically generated

By plotting heatmap, we can see the correlation of all feature between each other and if the 2 features are highly correlated then we can use one of them, so in this data set 8-9 features are highly correlated more than 80%. Below are highly correlated features.

{'amnt\_loans30',

'amnt\_loans90',

'cnt\_loans90',

'cnt\_ma\_rech90',

'daily\_decr90',

'maxamnt\_loans90',

'medianamnt\_loans90',

'medianamnt\_ma\_rech90',

'payback90',

'rental90',

'sumamnt\_ma\_rech90'}

**CONCLUSION**

* Key Findings and Conclusions of the Study

1. There are various aspect of the dataset and observation from my point of view:
2. From the statistical analysis it is clearly see that there is difference between the 75th percentile and max so this is an indication of the outliers in the dataset.
3. There is no defaulter in 8th month i.e. August month of 2016 year
4. In June month there are more defaulter as compared to non-defaulter and their age on cellular network is also more.
5. Removing unwanted column like msisdn number which does not play any imp role because that is unique.
6. Removing telecom circle and date column, which is also not an imp features, because telecom circle is same for all.
7. In this data set there are negative values which are from the total amount, age which can’t be negative.

* Learning Outcomes of the Study in respect of Data Science

1. In this project lot of feature engineering, feature selection and feature scaling are required, So, I learnt a lot about all this process which take most almost 60% of total time.
2. Also, because of outliers and some error in the data set lot of other statically analysis and function are used to make the data suitable to feed for the various model.
3. While creation of model I used multiple model with hyperparameter tuning and used cross validation because the dataset is imbalance so, cross validation is useful to select as sampling
4. Also, I used library named as SMOTE Tomek which is useful for oversampling and handle the imbalance dataset.
5. Because of large data set time of execution is very large which may up to 2 hours also.
6. In my finding, for imbalance dataset Random Forest perform good as compared to other model, as Random Forest is an boosting technique and because of using the tree internally it may also handle missing values and outliers and scaling is also not required.

* Limitations of this work and Scope for Future Work

1. The dataset contains multiple features and many rows which is good and from that we can find more information.
2. But the data is only for year 2016, if the data contain more information about 3-4 years then it will more precise and the output is more valuable and accurate.
3. Also, some features make also take into consideration that are region, city, occupation etc which will help to understand the data more.
4. To improve, there need to gather more data for defaulter also, so that by using their inputs we can able to predict the defaulter in future also.