MICRO CREDIT DEFAULTER PREDICTION PROJECT

Submitted by:

SHIKHA KAMAL

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

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.

• Conceptual Background of the Domain Problem

Many microfinance institutions (MFI), experts and donor share 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.

• Review of Literature

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 be6(in Indonesian

Rupiah), while, for the loan amount of 10(in Indonesian Rupiah), the payback amount should be 12(in Indonesian Rupiah).

- # Motivation for the Problem Undertaken

The sample data is provided to us from our client database. It is hereby given us for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

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.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  The data that we execute containing the extreme amount of database , which take time to run and loosing of data can’t done in predicting any results. The mathematical solution containing the statistics results explanation so we run the database in describe function.

  Describe function predicts the statistics of data base and calculates their mean, mode, median, different stages of percentile, minimum values present, maximum values presents, etc. The include parameters addition to df defined variable execute the string type data value also. String data value have mode , their frequency and top and last value also display.

  The statistics show all counts equals to total number of rows, which we can say that there is no null values present in database. So , we can ignore the imputer function to fill nan values with these methods.

  The difference between standard deviation and mean of each columns show the large variations which we can say that there is spreading of data present which we can have skewness factor

  The difference between 75 % percentile and maximum value of columns having vast difference which we can say there is outliers present .

- Data Sources and their formats

  Info function execute the overall description of database with overall view of columns data value and type. The info function is showing all non- null value present in columns and its data type variable exists.

The database have variety of data type like int, float, object type variable. The integer have have only numeric value without any decimal point , and float having decimal points, and object having string type data type.

The string type data type later converted into numeric coding which can understand by python. Various method used to convert into numeric coding without affecting the target variable.

- Data Preprocessing Done

In data pre-processing , cleaning is being in database so that we can made data for training purpose and exclude some un- necessary values or data from data sets to have accurate results in prediction.

Here , we can analyse the data for by observing some values from statistics calculations or plotting the graph for certain more clear observation . here we put plotting of some scatter and count plot where we can check for the frequency of data values, where it show more role and where it show less interest in predicting the results.

we are predicting the label column as a results where we check whether the one paid the loan amount or not, or can say that he is a defaulter present or not.

After that , we check for the outliers, where we can check for certain data values which lie outside the normal distribution curve, and it need to be removed so that we can have proper data values which lie in normal distribution curve which great accuracy.

Now the step is taken to remove the outliers , we use the z score method from scipy library, it checks for the outliers but while removing it removes all the rows present in the outliers. So we

decide not to remove the outliers as they are object type outliers and we don't want to loose all data values for prediction.

Next step is to check for the skewness by distribution curve, by plotting the data in graph, we observe that there is certain columns which show the vast spreading of data in curve. To maintain its accuracy we need to remove all the skewness present in the database.

We use the power transform method for skewness removal , after that again we check for skewness. Then there is 4 or 5 columns show more skewness , then there we apply the log method .

After removing the skewness , we put data in the training for more certain prediction.

- Data Inputs- Logic- Output Relationships

Target column is taking the label one, which we need to predict and trained the rest of data. The label column is categorical type, we can cross check this value by unique method. Only 2 types of data value present 0 or 1.

0- Showing the success
1- Showing the failure

Here we treat the label column as categorical one and trained the data from categorical model only.

- Hardware and Software Requirements and Tools Used

Various libraries is used while executing different outputs and models and various library for cross checking for models. And most of the function is pre- installed in python for basic running of the code.

Libraries like pandas and numpy used for array and dataframe, formed of database.
Matplotlib is used for various kinds of graphs plotting and running and executing results. Matplot used in plotting graphs of boxplot, distribution plot, and figuring the data size .
Seaborn library used for plotting but it is advanced than matplot libarary function.
Warnings library used to just ignore the various warning that display in outputs while running some of the commands in python.

Various models training , different libaray used for different models, each models predicts the same results but their output values are different. Here we can compare whether the results is close to the actual data present in the label output.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  For training purpose, first step is to run all libraries of models so that later we can run all models together.

  Then we split the data into target variable (y) and rest of columns different (x), and then find out the max random state for better state for best accuracy.

  We run different models under one loop and find their accuracy score by having their prediction.

  - Logistic regression
  - Gaussian NB
  - Decision tree classifier

- Random forest classifier
- kneighbors classifier

here the best accuracy model and well trained model is random forest classifier which predicts the maximum score of 92.34 % , from rest of the model.

Then we cross validify all models prediction, by cross validation score method. Here also we execute the score and best and maximum score model is random forest classifier with maximum score of 92.11 % . So we decide to proceed with the random forest classifier only.

Then to cross the model again we run through the grid search cv method which fitted the best trained model under their best parameters . so here after running through grid search cv , and find out the best parameters of random forest model and then again we fit the model and predict their results. Here it is giving score of 90.41 % which we can afford to proceed further .

As the database is extense precious so, by deciding not to loose much data we come across the best fitted model under best parameter trained and execute.

- Testing of Identified Approaches (Algorithms)

After splitting of the database into x and y variable, where y is the target variable. The test size decide the database percentage that is included in testing data sets. Here we every where mention the 20 % test size database where the rest of the data base that is 80 % is for training.

For prediction of data , we compare the prediction value from the actual modified test size data. How much percentage it is giving the best results.

We can also do one thing, by adding the testing coding into the python. We can test the data by entering the data values and run their values to predict the results and compare whether it is giving the actual results or not.

- Run and Evaluate selected models

Different algorithms used for running models and training.

- Logistic regression – best accuracy score is 90.047 % and the confusion matrix and classification report is also calculated.
  Cross validation score is 89.58 %

- Gaussian NB- accuracy score is 75.74 % and confusion matrix and classification report also calculate.
  Cross validation score is 75.50 %

- Decision tree classifier- accuracy score is 88.33 %.
  Cross validation score is 88.36 %

- Random forest classifier- accuracy score is 92.34 %.
  Cross validation score is 92.11 %

- Kneighbors classifier- accuracy score is 90.12 %.
  Cross validation score is 90.08 %

  By observing these results we decide to proceed with random forest model for best accurate results.

- Key Metrics for success in solving problem under consideration

  Key metrics is used differently in different models types. Like for regression models we use the r2 score for comparing the results. For categorical models we use the accuracy score, confusion matrix and classification report for better comparision.

  Here the target variable is categorical type, where we proceed the accuracy score for prediction of results. Here the target variable that is label column is binary type which holds the value of 2 types of 0 or 1. So we can proceed with roc auc curve and roc auc score for having score check and curve plotting.

  In roc auc curve , we need to find:
  - Tpr(true positive rate)
  - Fpr (false positive rate)
  - Thresholds

- Visualizations

  There are various plots which can be drawn according to our preferences and choice so that we can have better visualization from each of the plots.

  Here we used count plot , scatter plot , box plot, distribution plot to have visual observation from different perceptions. Count plot is preferred as the database contains the extense amount of data and target variable is categorical type so we can have counts of data values and can observe which columns and data values is preffered more.

The label is plotted and it show the failure rate value is very high compare to the success rate value. Accordingly we can take the preffered data value by observing their count. And in count plot 2 horizontal lines is plotted which indicate the label category, and comparing which values lie more in which category.

The most of columns is spreading the data equally in both category but there were some columns in which spreading of data is less in one of the category. Likewise we have one column "cnt_ma_rech30 ", "amnt_loans90", "amnt_loans30", "" it showing less spreading of data in 0 category of label column. Which indicate the failure .

- Interpretation of the Results

Then we proceed with the box plot , which checks for the outliers. The data values which present outside the box plot, beyond their minimum value and maximum value called as outliers. The columns like "aon", "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_loans30", "amnt_loans30" , "maxamnt_loans30", "amnt_loans90", and many more columns which are having outliers.

We tried to remove the outliers by using z score method, so all rows is being removed and we loose all database values to proceed further for prediction. So we ignoring the outliers removal and some are object type so we can ignore for outliers. We proceed for further coding.

Now, plotting is being done on skewness checking, which indicate the spreadin of data in normal distribution curve. The normal distribution curve tends to flat as the spreading of data increases.

The columns like "aon", "daily_decr30", "daily_decr90", "rental30", "rental90", "medianmarechprebal30" ," fr_ma_rech90",

“sumamnt_ma_rech90”, and many more columns having skewness present. We need to removed the skewness by using the method of power transform.
After applying the power transform , there is no skewness removal in some of columns. Then we apply the log method for skewness removal in that columns.

CONCLUSION

- Key Findings and Conclusions of the Study

  After analysing the database and having visual observation, we proceed with the best fitted model and that is random forest model. After cross validation and grid search best parameters we find the random forest is best fitted and its accuracy score is 90.49 %.

  Then we have the roc auc curve and its score . we plot the roc auc curve and plot them.

  Finally saving the best prediction of the model so that it can check or testing by entering the data values . we got the best percentage by loosing very less percent of score.