

Micro Credit  Project

Submitted by:

Kaushik Kalappa

ACKNOWLEDGMENT

My sincere thanks to my mentor and supervisor Swati Rustogi, a senior Data Scientist at Flip Robo ;Ram Prasad and Rohit from Data Trained for guidance throughout the Project phase.

Source of reference would include :

-https://www.geeksforgeeks.org/

-https://scikit-learn.org/stable/

-https://learning.datatrained.com/

INTRODUCTION

* Business Problem Framing

Loan sanctioning is a big decision for any financial organization since it can lead to a bad investment if proper investment research is not done. The organization can  invest in any corporation, or company or an individual and it comes with certain assurance when enough background research is done about the prospects. But the conventional parameters such as salary, credit history etc would not be enough to really make a good enough decision. More and more parameters needs to be check, but it cannot be done manually with conventional manpower and hence business needs to upgrade to work these paramters through machine learning.

* Conceptual Background of the Domain Problem

To understand this problem we need to understand how a loan process works in a Microfinance Institution. It 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.

* 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. 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.

They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days.

* Motivation for the Problem Undertaken

The main motivation is to help the client make a better and more informed decision in lending their financial resource so that there will be as low defaulters as possible. With the help of this model companies can find the confidence to invest in potential areas so that poor families are also  mitigated of their problems financial through this scheme

Analytical Problem Framing

* Mathematical/ Analytical Modeling of the Problem

Mathematically the mean and median ere calculated for each column to check for the distribution

Had to check for null values existing in the dataset

Had to use various univariate and bivariate plots to check for distribution and relationship between columns such as plot, scatterplot, displots.

* Data Sources and their formats

The data was shared by the client from a Telecom company in the form of csv file. They had about 209593 columns and 37 rows

* Data Preprocessing Done

Dropped the columns Unnamed: 0, msisdn and pcircle since they were not required

Dropped columns: last\_rech\_date\_da, cnt\_da\_rech30, fr\_da\_rech30, cnt\_da\_rech90, fr\_da\_rech90, medianamnt\_loans30,medianamnt\_loans90

since the values greater than 0 are very less and more than 75 percent are 0 and less

* Data Inputs- Logic- Output Relationships

The target variable is label which is a flag indicating whether the user paid back the credit amount within 5 days of issuing the loan. It is in terms of 0s and 1s and is a categorical column

The correlation suggests that: cnt\_ma\_rech30, sumamnt\_ma\_rech30, cnt\_ma\_rech9, sumamnt\_ma\_rech90have a positive corelation with the target column

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

None

* Hardware and Software Requirements and Tools Used

The models are run in a 4gb ram with pentium processor in Jupyter Notebook

The libraries used are:

Pandas for dataframe

Numpy, and Math  for arrays and mathematical functions

Seaborn and matplotlib for visual plottings

sklearn for importing models and classification metrics

Model/s Development and Evaluation

* Identification of possible problem-solving approaches (methods)

First dropped the unwanted columns and created new columns date

Then dropped categorical columns for skewness reduction

The checked for skewness and removed outliers and replaced them witth upper limit

Then applied log transformation to get the skewness in accceptable range

Then collaborated the categorical columns and then sent them for model testing

The applied various classification models to find the best fit model for the dataset

* Testing of Identified Approaches (Algorithms)

Decision Tree Classifier

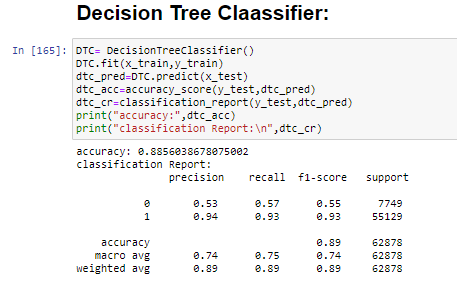
Linear Support Vector Classifier

RandomForestClassifier

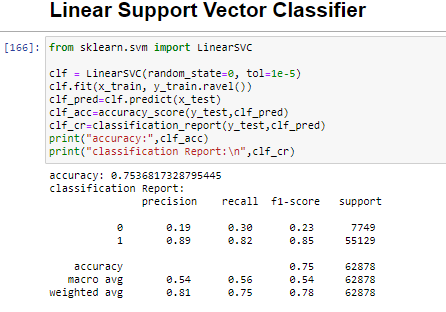
Naive\_Bayes

* Run and Evaluate selected models

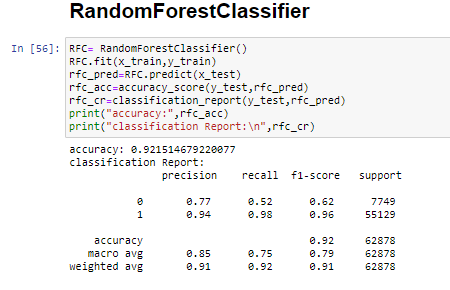
Decision Tree Classifier:



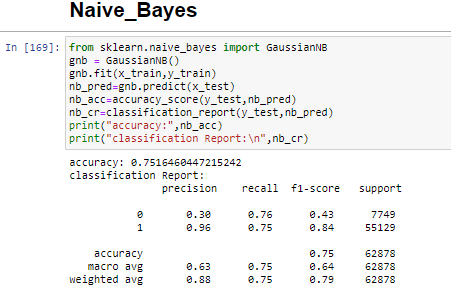
Linear Support Vector Classifier:



RandomForestClassifier:



Naive\_Bayes:



The accuracy of Random Forest Classifier was the highest which was the first level of key factor for best fit model

* Key Metrics for success in solving problem under consideration

The difference between Cross Validation and accuracy score of different models

Decision Tree Classifier:

(0.8856038678075002-0.883054304149387)\*100 = 0.25495636581132075

Linear Support Vector Classifier:

(0.7536817328795445-0.8648475095995234)\*100 = -11.116577671997884

RandomForestClassifier:

(0.921514679220077-0.9204839924931576)\*100   = 0.10306867269194164

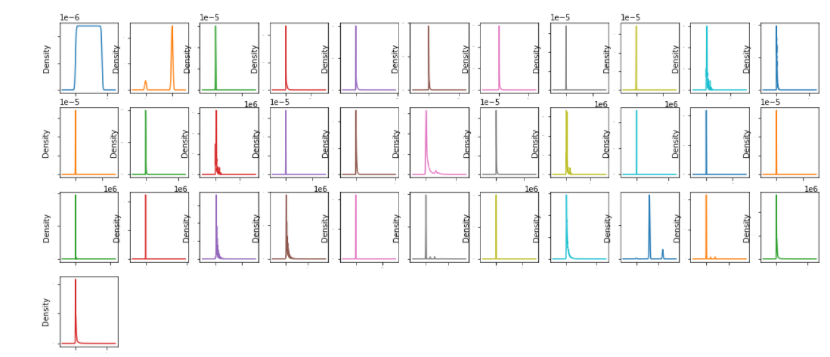
Naive\_Bayes:

(0.7516460447215242-0.7531167658602724)\*100 = -0.14707211387482433

Considering the f1-score, precison, recall and the difference between cross validation and accuracy of each model it can safely be concluded that

Random Forest Classifier is the best model

* Visualizations



These are the distribution plots of all the columns and none of them have a a normal distribution

Interpretation of the Results

The respective irrelevant columns were dropped and then after rectifying the skewness the dataframe was sent for modelling.

Upon trying out different models Random Forest Classifier came out as the best fit final model with an accuracy and  Cross validation score of 92.1514679220077 and 92.04839924931576  and a final accuracy  and Cross Validation score of 92.27869843188397 and 92.15765822963384 after Hyper Parameter tuning

CONCLUSION

* Key Findings and Conclusions of the Study

The key findings were spotting the numeric columns which had very little value for the model and dropping them so as to get the best model

Even though Random forest classifier had the most of the accuracy, it was the overall evaluation of classification metrics that gave an edge to the Random forest  model to become the best fit.

* Learning Outcomes of the Study in respect of Data Science

The thing that was learnt from this model was that outliers were spotted in data cleaning which had a big impact to skewness  when they were removed.

Also dropping the unwanted columns which had 75 percent of the data around 0 with minimum would not have contributed to the accuracy

* Limitations of this work and Scope for Future Work

The important thing to notice about this model is that it can work only with this telecom industry metrics as compared to other financial institutions

Even though this model has given a fair accuracy from a mathematical standpoint, there can still be unavoidable scenarios like market crash due to pandemics, human negligence , wilful defaulters, which the model would not be able to help.