A linear Model of Home Loan Prophecy

Naeem Ul Hasan Chowdhury
Department of CSE
East West University
Dhaka, Bangladesh
nayemulhasan97@gmail.com

MD. Sharif Mulla Mahin
Department of CSE
East West University
Dhaka, Bangladesh
sbzm.mahin@gmail.com

Rayhan Ahmed
Department of CSE
East West University
Dhaka, Bangladesh
rayhanriyad97@gmail.com

I. Introduction

Loan is the lending of money by one or more individuals, organizations, or other entities to other individuals, organizations etc. Every Finance Bank deals with various types of loans. Banks want to give loan to every person who apply for loan. But there is a problem arising to give loan that all the people are eligible to take loan or not? To find the eligible people for loan, Banks want to automate the loan eligibility process. They have presence across all urban, semi urban and rural areas. The loan eligibility process based on customer detail provided while filling online application form. These details are Gender, Marital Status,Education, Number of Dependents, Income,Loan Amount, Credit History and others. By automating this process, those are eligible for loan, Bank can specifically target these customers. Customer first apply for loan after that bank validates the customer eligibility for loan.So, this is a standard supervised classification task. We are trying to generate a model using machine learning based concept to predict whether a loan would be approved or not. Our aim is checking the customer eligibility for getting loan.

II. Methodology

First of all, we choose a data set. Then we tried to analysis the data set. After analysing the data set, we found some null values in some entries. So we had to pre process the data set. To pre process the data set, we used mood operation in categorical values and mean operation in numerical value. Then we got a comfortable data set to train a model. For calculating purpose, we converted the all string data to numeric form. So our all data is in numeric form. It's help us to preparing our model. We splited the data set in test and train portion.The 70 percent data is in training portion and 30 percent is in testing portion. Then a model is generated that is trained by using LogisticRegression, DecisionTreeClassifier, RandomForestClassifier, ExtraTreesClassifier algorithm. At last we also added hyperparameter tuning for get better accuracy in our project model. A short diagram of our model is shown:

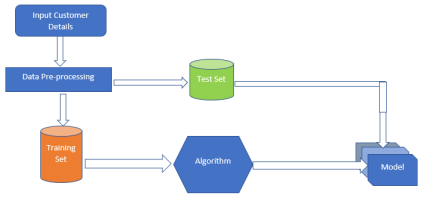

Fig. 1. Project Diagram

III. Implementation

A. Data Description

We collected a data set from kaggle website. In that data set, there were total13 columns, 614 entries, numeric value in eight columns and rest of the columns are categorical.We tried to understand the data set and found some null values present in the data set. The null values are corrected in data pre processing part.

B. Data Pre-processing

In the data set there are 614 entries and 13 columns. Some values are missing in the data set. Before we train the model through these data set, we had to fill the missing values in our data set. We fill the numerical missing values by using mean operation. And then We fill the categorical missing values by using mode operation.On the other hand, though there is 13 columns in our data set,we drop the Loan ID attribute in our model. From our perception survey Loan Id can't contribute much as an important attribute in the predictive model.Therefore, we drop this attribute.

C. Feature Selection And Classification Analysis

For feature selection analysis, we used Pandas library for read the data set. In our data set remain categorical values so we also use Scikit-Learn python library because Scikit-Learn python library require all input to be numeric because the machine can't read the character. We also use matplotlib library for creating2d graph and plots. We use seaborn library to visualise our data. To buildup the model, we have used ExtraTreesClassifier which is an ensemble learningmethod basically based on decision trees. In a previous study,random foreststssignificantly out-performed logistic regression. This motivated us to use RandomForest Decision Tree classifier. For training and testing purpose, we kept 70percent participants

in the train group and remaining 30 percent participants were kept in the test group.Here we also use cross validation to get better accuracy of the model.

D. Hyper-parameter Tuning

We had already built a random forest model to solve our machine learning problem. But we want to improve using hyper-parameter tuning. The best way to think about hyper-parameters is like the settings of an algorithm that can be adjusted to optimize performance, just as we might turn the knobs of an AM radio to get a clear signal. The best hyper-parameters are usually impossible to determine.This is not guaranteed to be optimal for a problem.

IV. Acknowledgment

We [1] like to thank the data set owner for his effort to construct this data set.We also thank to our faculty for his insightful comment on the submitted draft paper.

V. Limitations

Though we have tried to minimize the limitations, but because of time limit,there are still few limitations in our study which should be explored by future studies. In this project we may use log transformation for suitable distribution of the numerical data for better training the model.

VI. Conclusion

We conducted a study with the aim to develop a model which can classify loan prophecy of the customer. In our developed model after Compare all of the trees and predict output we see using hyper-parameter tuning in random forest model it gives customer's loan prophecy 80.61 percent accurately. The presented work is a step towards developing a better model in this area.