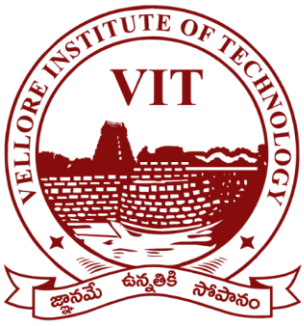

VIT-AP UNIVERSITY

VIT-AP University, G-30, Inavolu, Beside AP Secretariat Amaravati, Andhra Pradesh 522237

Machine Learning Project Report -
Loan Prediction using Machine Learning

Submitted By:

Aditya Bhupal – 20BCD7111

Submitted to:

Dr.Sachi Nandan Mohanty,
PostDoc(IIT Kanpur),
Ph.D.(IIT Kharagpur,WB)
Senior Member IEEE,
Fellow IE,
MACM

ABSTRACT

INTRODUCTION

Loan prediction involves the lender looking at various background information about the applicant and deciding whether the bank should grant the loan. Parameters like credit score, loan amount, lifestyle, career, and assets are the deciding factors in getting the loan approved. If, in the past, people with parameters similar to yours have paid their dues timely, it is more likely that your loan would be granted as well.

Machine learning algorithms can exploit this dependency on past experiences and comparisons with other applicants and formulate a data science problem to predict the loan status of a new applicant using similar rules.

Several collections of data from past loan applicants use different features to decide the loan status. A machine learning model can look at this data, which could be static or time-series, and give a probability estimate of whether this loan will be approved

The Prediction model not only helps the applicant but also helps the bank by minimizing the risk and reducing the number of defaulters

PROBLEM

This is a classification problem in which we need to classify whether the loan will be approved or not. Classification refers to a predictive modelling problem where a class label is predicted for a given example of input data

SOLUTION

We will be using Various Machine Learning Algorithm's namely Decision Tree Classifier, XGBoost, Ada Boost Classifier, Gradient Boost Classifier and Logistic Regression to predict if the loan should be granted to the applicant or not

RESULT

After testing on the test data provided below are the accuracy results that were obtained,

The accuracy of the Decision Tree (Post Pruning) = 83.464 %

The accuracy of the Decision Tree (Pre Pruning) = 82 %

The accuracy of Ada Boost = 76.44 %

The accuracy of Gradient Boost = 85 %

The accuracy of XGBoost = 88 %

The accuracy of Logistic Regression = 74.8 %

PROBLEM STATEMENT

There is a company named Dream Housing Finance that deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. However doing this manually takes a lot of time. Hence it wants to automate the loan eligibility process (real time) based on customer information

So the final thing is to identify the factors/ customer segments that are eligible for taking loan. How will the company benefit if we give the customer segments is the immediate question that arises. The solution is ….Banks would give loans to only those customers that are eligible so that they can be assured of getting the money back. Hence the more accurate we are in predicting the eligible customers the more beneficial it would be for the Dream Housing Finance Company.

SOLUTION TO THE PROBLEM

The above problem is a clear classification problem as we need to classify whether the Loan Status is yes or no. So this can be solved by any of the classification techniques like,

1. Decision Tree Classifier
2. Ada Boost Classifier
3. Gradient Boost Classifier
4. XGBoost Classifier
5. Logistic Regression

PROCEDURE OF THE PROPOSED SOLUTION

1. Loading the Essential Python Libraries
2. Load Training/Test Data Set
3. Checking the size of the dataset
    - 614 rows and 13 columns in the training dataset
    - 367 rows and 12 columns in the test dataset
4. First Look at the Dataset
    - Categorical columns : Gender(Male/Female), Married(Yes/No), Number of dependents (Possible values 0,1,2,3..), Education(Graduate/Not Graduate), Self Employed(Yes/No), Credit History(Yes/No), Property Area(Rural/Semi-Urban/Urban) and Loan Status(Yes/No)

- Numerical Columns : Loan ID, Applicant Income, Co-Applicant Income, Loan Amount and Loan amount Term

5. Data Pre-processing

    - Concatenating the train and test data for data pre-processing
    - Dropping the unwanted column
    - Identify missing values
    - Imputing the missing values
    - Next, we will be using Iterative imputer for filling missing values of LoanAmount and Loan_Amount_Term, So now as we have imputed all the missing values we go on to mapping the categorical variables with the integers.
    - We map the values so that we can input the train data into the model as the model does not accept any string values.

6. Exploratory Data Analysis (EDA)

    - Splitting the data to new_train and new_test so that we can perform ED
    - Mapping ‘N’ to 0 and ‘Y’ to 1
    - Univariate Analysis
    - The Display the Output

CONCLUSION

After proper analysis of the data, we can conclude that the XGBoost Classifier gave us the best results in terms of Accuracy. Extreme Gradient Boost (XGBoost) is an improvement over Gradient Boost and is very popular in binary classification algorithms. The decision trees are built in parallel in XGBoost than in series, giving it an edge over normal Decision Trees and Boosting algorithms.

- We did Exploratory data Analysis on the features of this dataset and saw how each feature is distributed.
- We did bivariate and multivariate analysis to see impact of one another on their features using charts.
- We cleaned the data and removed NA values
- We also generated hypothesis to prove an association among the Independent variables and the Target variable. And based on the results, we assumed whether or not there is an association.
- We calculated correlation between independent variables and found that applicant income and loan amount have significant relation.
- We created dummy variables for constructing the model
- We constructed models taking different variables into account and found through odds ratio that credit history is creating the most impact on loan giving decision
- Finally, we got a model with co-applicant income and credit history as independent variable with highest accuracy.
- We tested the data and got the accuracy of 88 %