GLIMPSE -Journal of Computer Science •Vol. 2(1), JANUARY-JUNE 2023, pp. 29-31

LOAN PREDICTION USING MACHINE LEARNING

[1]Megha Gupta [2]Anant Tyagi

[1]Assistant Professor, Ajay Kumar Garg Engineering College,Ghaziabad, U.P., India

[2]Student 2nd CSE, Ajay Kumar Garg Engineering College,Ghaziabad, U.P., India

[1]guptamegha@akgec.ac.in, [2]anant2110008@akgec.ac.in

***Abstract*— Loan sanctioning and credit score forms a multi-billion-dollar industry. With a wide range of applicants ranging from students to large multinational companies for their expansion as an individual or organization. Processing these applications through traditional analysis is a complex task for the financial systems. The process of accepting or rejecting the application is quite time consuming as there are many variables to be considered. Making a machine learning model would reduce the human bias and delays in the application processing time. Banking systems have looked at the use of AI to determine lending risk and probability of repayment.**

***Keywords*— Machine Learning, Random Forests, XG Boost, Logistic Regression, Ensembling Techniques**

I. INTRODUCTION

With the increase in the banking sector, a mass of people is applying for bank loans but the bank due to its limited assets must grant to its assets to limited people only, so it would be a critical process for the bank to identify the safer options for granting loans.

So, banking institutions wish to reduce the risk factor involved in the process. Our model is based on reducing the risk factor using machine learning.

This model will help us identify the loan defaulters using the data mining algorithm.

The project may be deployed in five phases:

i) Strategy
ii) Data Preparation and Preprocessing
iii) Data Splitting
iv) Modeling
v) Model Deployment

II. EXSISTING SYSTEM

The existing system for loan prediction has been complex and time consuming where bank employees manually check all the variables related to issuance of loan and try to calculate the probability of its repayment. The ANN's model is made as a credit prediction system. The credit default is evaluated using Feed-forward propagation neural network. This existing system has a lot of human bias and unexplainable factors affecting the current system to provide authentic results [1].

Advantages

Time period will be reduced for loan sanctioning.

For avoiding the human error, the whole process will be automated.

Loan will be sanctioned to eligible applicants without any delay [3].

III. WHAT IS MACHINE LEARNING

Machine Learning is the branch of AI that deals with developing computer programs that can take an input set of data and identify patterns within the training data to develop a self-prediction model. ML's most important feature is self-learning. ML algorithms usually require a function which plays a significant role in success or failure. It can be done either by minimizing the loss function or maximizing the gains.

Machine learning algorithms can be classified into:

i) Supervised Learning
ii) Unsupervised Learning
iii) Recommendation system

IV. BRIEF INTRODUCTION TO THE MODEL

This method uses at least two weakly trained classifiers combined to form an ensemble model for better prediction. It involves the use of both bagging and boosting techniques. Random forest is used as a bagging technique and the result predicted by each model is sent for bootstrap aggregation [2] and uses XG Boost for boosting. Finally, we can say that assembling techniques improve the results.

V. ARCHITECTURE TECHNIQUES

Decision Tree:

It is a classification algorithm based on supervised Machine Learning. The decision tree is like a tree and contains branches which may represent outcomes, reactions or even the possible decisions. The leaf node of the tree is available with the answer.

Random Forest:

It is a bagging ensemble technique in which many weakly

trained decision trees are used to predict the outcome and at last bootstrap aggregating is done.

XG BOOST:
It is a boosting ensemble technique in which weakly trained models are arranged sequentially and predict better outcomes.

VI . DESCRIPTION OF THE PROJECT

Getting ready with system and data: In this project we have used python as the programming language and selectively used below listed libraries

Specification
Python: The language on which the model is to be written
Pandas: A python library used to make data frame.
Seaborn: Used to plot graphical representations to the data.
Sklearn: Used to import datasets.

2) ii) Understanding the data
The dataset for this loan repayment consists of twelve independent variables and a dependent variable (Loan Status) in the training data and all columns except for the Loan Status variable in testing data. The Y in Loan Status represents the loan was repaid and N suggests loan repayment is not done[4].

iii) Exploratory Data Analysis (EDA)
i. Univariate Analysis: The term univariate refers to analysis of a single variable. The purpose of this analysis is to understand the distribution with respect to a single variable.
ii. Bivariate Analysis: Bivariate Analysis is the analysis ofa data column with respect to another. It helps understanding the relationships between two variables. It is the simplest analysis available in statistics.

iv) Missing value imputation

Missing value imputations can be completed using
For numerical variables: If data skewed towards right or left: Median or Mode. If data is not skewed: Mean
For categoricalvariables: Mode is used.

v) Evaluation Metrics for Classification Problems
Helps in evaluation of the accuracy of statistical and Machine Learning models. Here it is used to check the accuracy of Logistic Regression, Decision Tree, Random Forest and XG Boost.

vi) Model Building (Part-1)
Our first model is to be scripted to predict target variables We begin with Logistic Regression to predict the binary outcome: Logistic Regression is the best suitable machine learning algorithm that can be used for providing a binary outcome. Linear Regression cannot be used to solve the outlier problems and logistic regression consists of sigmoid function.
Logistic regression is said as an approximation of Logit function that is the log of odds in favor.
The Logistic Function solves the problems of outliers due to its s shaped nature.

vii) Logistic Regression using stratified k-folds cross-validation
Validate to check the accuracy of the model.This technique involves reserving segments of the data set on which the model has not be entrained. H. The data are new to the model.The model is then test edon the data set before being finalized. Some of the common verification method sare listed below:

- The validation set approach
- k-fold cross-validation
- Leave one out cross-validation
- Stratified k-fold cross-validation
- Here we have used k fold validation to check if the model works fine or is to be trained better.

viii) FeatureEngineering
We use the knowledge of the domain to add new features that may help finding the dependent variable or the target variable.
ix) Model Building (Part-2)

After all the tasks are performed and new features get added. Now we train our model with algorithms. Initially we begin with random forest and then we move towards XG Boost, a more complex boosting technique.

Logistic Regression
Decision Tree
Random Forest
XGBoost

ALGORITHMS
i. Import all the required python modules
ii. Import the database for both TESTING and TRAINING.
iii. Check any NULL VALUES are exists
iv. If NULL VALUES exists, fill the table with corresponding coding
v. Exploratory Data Analysis for all ATTRIBUTES from the table
vi. Plot all graphs using MATPLOTLIB module
vii. Build the DECISION TREE MODEL for the coding.
viii. Send that output to CSV FILE.

VI . CONCLUSIONS & FUTURE SCOPE

To conclude I would say that this model is considerably productive for the banking sector. Even being trained on small data has shown a positive response with great efficiency. If trained for real time data it can prove to be a real breakthrough in the banking sector. Automation of the loan repayment pre-

diction will be beneficial for both the consumer and the banking authorities. As the government schemes have made loans easily accessible which leads to an increase in the number of applications for the bank to look for. Then the product would prove to be useful. Soon, this module of prophecy may be integrated with the module of automated processing systems. The system is trained on the old training data set. In the future, the software could similarly make new test data join the training data after a specified time.

REFERENCES

[1] Kumar Arun, Garg Ishan, Kaur Sanmeet, May-Jun. 2016. Loan Approval Prediction based on Machine Learning Approach, IOSR Journal of Computer Engineering (IOSR-JCE).

[2] Wei Li, Shuai Ding, Yi Chen, and Shanlin Yang, Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China, Key Laboratory of Process Optimization and Intelligent Decision Making, Ministry of Education, Hefei University of Technology, Hefei 2009, China.

[3] Short-term prediction of Mortgage default using ensemble machine learning models, Jesse C.Sealand on july 20, 2018.

[4] Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques, Dr. K. Kavitha, International Journal of Advanced Research in Computer Science and Software Engineering

ABOUT THE AUTHOR

Megha Gupta is an Assistant Professor in Department of Computer Science & Engineering at AKGEC. She is doing research in Machine Learning with an objective of providing various problem solution to society.

Anant Tyagi is a sophomore in BTech. (CSE). He is a Machine Learning enthusiast willing to get involved in software writings and cognitive model development. Working with an objective to cure human disability by using AI as a tool.