ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

Ranking Metric Embedding-Based Multicontextual Behavior Profiling for Online Banking Fraud Detection (ReMEMBeR)

Mrs.S.Savithri
Department of Mca
Dhanalakshmi Srinivasan college of engineering and technology

Ms.Ashwini.K
Department of Mca
Dhanalakshmi Srinivasan college of engineering and technology

Abstract: As the loan amount is a percentage of the assessed value of the collateral, a high-value asset could mean more credit sanctioned for your use. The asset could be immovable (land or house) or movable (vehicle, inventory, equipment, investments, insurance policies, gold jewellery, art, and other such valuables). The goal is to develop a machine learning model for Bank Loan Approval Prediction, to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm.While Personal Loans (including credit card outstanding balance) are unsecured loans, approval for loan to purchase a car or a home, run a business, or study will not come through unless there is adequate collateral. Generally, banks are willing to fund up to 80% of the cost of purpose of the loan and expect the borrower to arrange for the balance. However, if you can put in more than 10-20%, the bank will not stop you. Rather, it will recognize that you are willing to reduce the bank's exposure to the default risk and approve your application sooner. The down payment you are able to make will have a huge impact on your home, education, car, or business loan eligibility.To implement and investigate how different supervised binary classification methods impact default prediction.

I. INTRODUCTION

Anomaly detection relies on individuals' behavior profiling and works by detecting any deviation from the norm. When used for online banking fraud detection, however, it mainly suffers from three disadvantages. First, for an individual, the historical behavior data are often too limited to profile his/her behavior pattern. Second, due to the heterogeneous nature of transaction data, there lacks a uniform treatment of different kinds of attribute values, which becomes a potential barrier for model development and further usage. Third, the transaction data are highly skewed, and it becomes a challenge to utilize the label information effectively. Anomaly detection often suffers from poor generalization ability and a high false alarm rate.. By doing so, the idea of collaborative filtering is implicitly used to utilize information from similar users, and the learned preference matrices and attribute embedding provide a concise way for further usage.

II. DISADVANTAGES OF EXISTING SYSTEM:

- They had proposed a mathematical model and machine learning algorithms is not used
- Class Imbalance problem was not addressed and the proper measure were not taken.

III.PROPOSED SYSTEM:

Exploratory Data Analysis of loan approval

Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

Data Wrangling

In this section of the report will load in the data, check for cleanliness, and then trim and clean given dataset for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

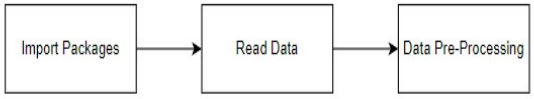
Data collection

The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set**Data Pre-processing**

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer.

MODULES:

1. Data Pre-processing
2. Data Analysis of Visualization
3. Comparing Algorithm with prediction in the form of best accuracy result
4. Deployment Using Flask



IV.MODULES DESCRIPTION

Module Diagram
Given Input Expected Output

Input : data
Output : removing noisy data

Exploration data analysis of visualization

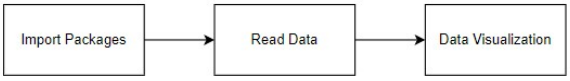
Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Module Diagram
Given Input Expected Output

Input: data
Output: visualized data

Comparing Algorithm with prediction in the form of best accuracy result

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine



learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data.

In the example below 4 different algorithms are compared:

- Logistic Regression
- Random Forest
- Decision Tree Classifier
- Naïve Bayes

Algorithm Explanation

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithms learn from labeled data.

Logistic Regression

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

In other words, the logistic regression model predicts P(Y=1) as a function of X. Logistic regression Assumptions:

- Binary logistic regression requires the dependent variable to be binary.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is, the model should have little.

Module Diagram

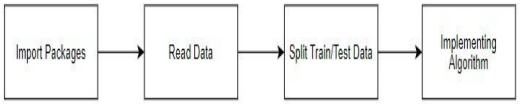
Given Input Expected Output
Input : data
Output : getting accuracy

Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees
The following are the basic steps involved in performing the random forest algorithm:

- Build a decision tree based on these N records.
- Choose the number of trees you want in your algorithm and repeat steps 1 & 2

Module Diagram



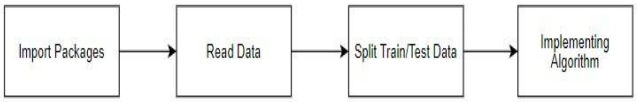
Given Input Expected Output
Input : data
Output : getting accuracy

Decision Tree Classifier

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Assumptions of Decision tree:

- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are

Module Diagram

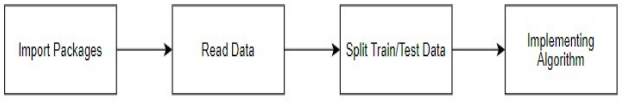


Given Input Expected Output
Input : data
Output : getting accuracy

Naive Bayes algorithm:

The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modeling problem probabilistically.

Module Diagram



Input Expected Output

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

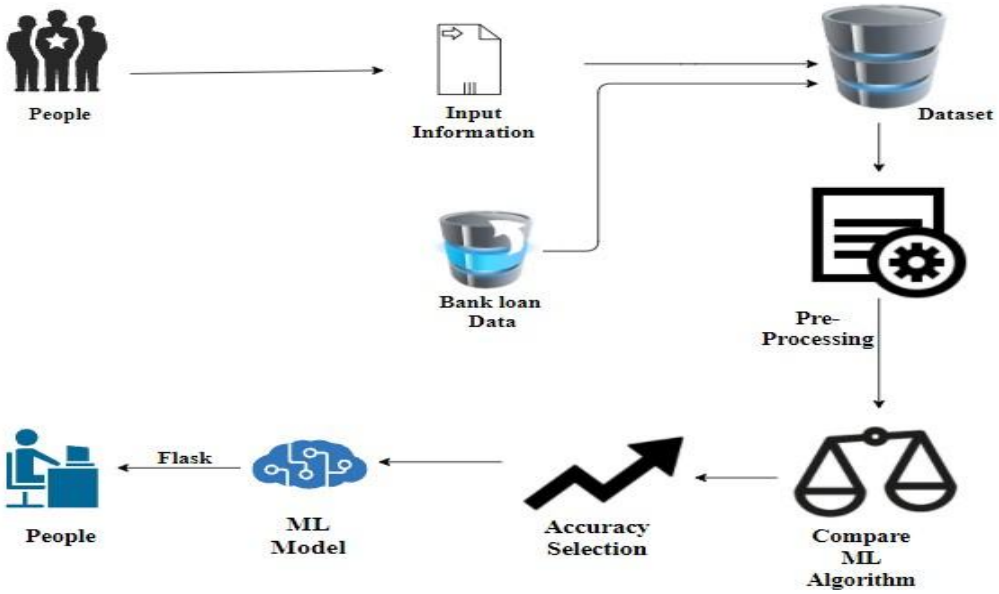
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

Input: data
Output: getting accuracy

System design is described as a process of planning a new business system or more to replace or to

V. SYSTEM DESIGN

SYSTEM ARCHITECTURE



VI. SYSTEM TESTING

9.1 Unit Testing

Unit testing focuses verification effort on the smallest unit of software design – the software component or module. Using the component level design description as a guide, important control paths are tested to uncover errors within the boundary of the module. The relative complexity of tests and uncovered scope established for unit testing. The unit testing is white-box oriented, and step can be conducted in parallel for multiple components. The modular interface is tested to ensure that information properly flows into and out of the program unit under test.

9.2 Validation testing

Attempts to find the errors in the following categories

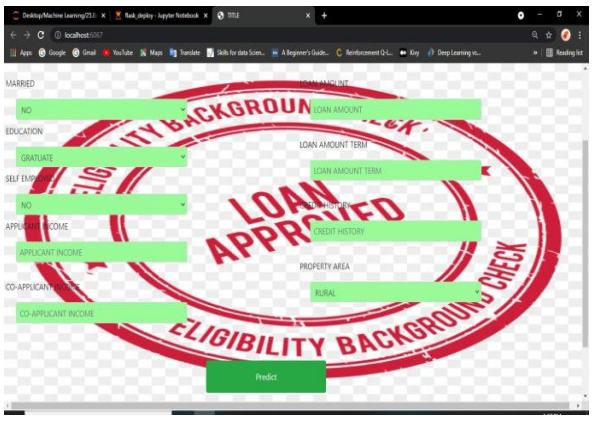
- Incorrect or missing functions
- Performance errors and initialization
- Termination errors
- Interface errors
- Error in data structure

PERFORMANCE AND LIMITATION

VII. CONCLUSION

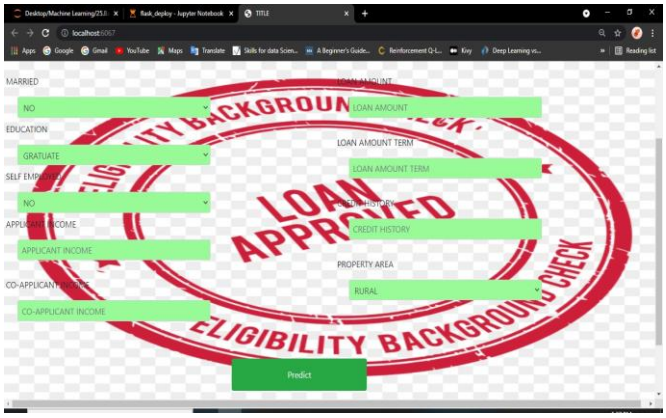
- The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation.
- The best accuracy on public test set is higher accuracy score will be find out. This application can help to find the Prediction of Bank Loan Approval.

REPORTS



ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022



REFERENCE

- **C**.Sankar Ph.D," Impact of Personal Loan Offered by Banks and Non Banking Financial Companies in Coimbatore City",2017.
- M. Cary Collins," Improving Information Quality in Loan Approval Processes for Fair Lending and Fair Pricing",2013.
- Kumar Arun, Garg Ishan, Kaur Sanmeet," Loan Approval Prediction based on Machine Learning Approach",2016.
- Sivasree M S, Rekha Sunny T," Loan Credibility Prediction System Based on Decision Tree Algorithm" 2015.
- Jiří Doležal, Jiří Šnajdr, Jaroslav Belás , Zuzana Vincúrová," Model of the loan process in the context of unrealized income and loss prevention",2005.