REPORT ON LOAN PREDICTION USING

MACHINE LEARNING ALGORITHM

Pradakshina S B AI Department

Kongu Engineering College Erode, Tamil Nadu [pradakshinasbs.22aim@kongu.edu](mailto:pradakshinasbs.22aim@kongu.edu)

Nithish N

AI Department

Kongu Engineering College Erode, Tamil Nadu [nithishn.22aim@kongu.edu](mailto:nithishn.22aim@kongu.edu)

Yogananth J AI Department

Kongu Engineering College Erode, Tamil Nadu [yogananthj.22aim@kongu.edu](mailto:yogananthj.22aim@kongu.edu)

***Abstract***

***This study provides a general model for identifying and classifying loan prediction. The model uses a variety of machine learning algorithms, including Support Vector Machines (SVM), Random Forest, Ridge Classifier, AdaBoost, K Nearest Neighbors (KNN), Gradient Boosting, Decision Trees to identify the loan This model involves various technologies, transforming old data into designs, and improving the learning and ability of the algorithm. Additionally, the overall prediction is improved by using common methods such as loan predicting to use the combined results of separate operators .Results have proven useful indicating that fire models are used in detecting and classifying loan..***

1. Introduction

Using Loan approval prediction is a critical task in the banking sector. Machine learning models can provide insights into factors influencing loan approval decisions. This study utilizes a dataset with various attributes related to loan applicants and employs different machine learning algorithms to predict loan approval. Nowadays machine learning emerges as a pivotal player in forecasting loan prediction before its onset. Through the utilization of advanced algorithms, machine learning models can dissect data patterns, providing invaluable insights into potential risks associated with loan The capacity of machine learning to efficiently process vast amounts of information establishes it as a valuable tool for recognizing early indicators of loan and facilitating timely interventions. Within this context, a literature review scrutinizes existing research on loan prediction and the implementation of machine learning techniques. This review encompasses a thorough examination of prior studies, methodologies employed, and the findings derived, thereby laying the groundwork for the ongoing research. It contributes significantly to cultivating a comprehensive understanding of the subject matter.

1. LITERATURE SURVEY

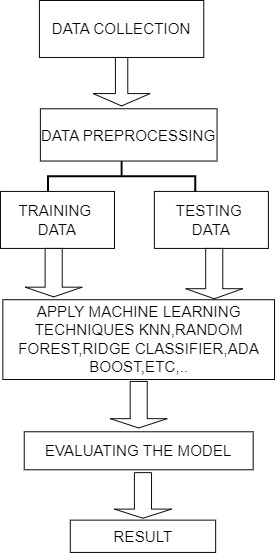
A variety of approaches can be used in loan prediction research to create and verify predictive models. Here are several research techniques for loan prediction.

K nearest neighbour, Decision Tree, Ridge Classifier, Bagging Classifier, Gradient Boost Classifier, ADA boosting, Random Forest.

The quantitative technique used in this study places a strong emphasis on measuring the available data. The steps involved in conducting research are carried out using flow chart given below. The dataset of loan prediction consists of 1000 data divided into 13 attributes. The dataset of loan is a publicly available dataset that was originally uploaded to Kaggle.com.

Accuracy, values will be compared between the test result in order to assess the outcome. By selecting the best value, the values will be utilized to evaluate the model.

1. SYSTEM ARCHITECTURE AND METHODOLOGY



Exploring speech discrimination using machine learning algorithms such as Support Vector Machines (SVM), Random Forests, Ridge Classifier, AdaBoost, K Nearest Neighbors (KNN), Gradient Boosting, Decision Trees and is a good way to go in natural language processing. Here are some steps you can take to create a distinctive search algorithm -

* 1. Load the dataset: To initiate the analysis, the dataset was loaded using Pandas' **read.csv** function. The dataset's first few rows were inspected using the **head()** method to gain a preliminary understanding of its structure and content. This step serves as the foundation for subsequent data processing and analysis.
  2. Preprocessing of data: Preprocessing requires cleaning the original text data and converting it into a format that can be used by machine learning algorithms.
  3. Feature extraction: This step will extract relevant features from the previous text. You can use embedding to create features that can be used by machine learning algorithms.
  4. Training model: Split the dataset into training and validation set. Use training methods to train your machine learning model.
  5. Validate the model: Use the validation process to evaluate the effectiveness of the model. Evaluation criteria usually include precision, recall, F1 score, and accuracy.

6.Reporting the model: Once you train and evaluate the model, you can use it to classify loan is approved or not.

1. DATA CLEANSING

We want our dataset to be “clean” before pre-processing the loan data and fitting the data into a model. We must ensure that there are no missing information, duplicates.

As a result of our preliminary analysis, no significance was found in any area. To check for duplicates, we compare specific values in the profile with all values in each column.

After looping through each row and checking the data accuracy ,we did not see any evidence of contradiction.

DATA PREPROCESSING

The preprocessing of the text data is an essential step as it makes the raw text ready for mining, i.e., it becomes easier to extract information from the text and apply machine learning algorithms to it. If we skip this step, then there is a higher chance that you are working with noisy and inconsis-tent data. The main objective of this step is to clean noise of what is less relevant in order to find the loan.

DATA AUGMENTATION

Data augmentation is a data oversampling technique used to increase the data size by adding new samples with a similar distribution to the original data or slightly modifying the original data. To be more efficient in distributing activities, the document needs to be updated to capture classroom notes.

DATA INTEGRATION**:**

The process of merging and integrating data from various sources to create a single, comprehensive representation is known as data integration. However, this step is not used because the diabetes dataset is already in the form of useful data for the analysis.

DATA REDUCTION**:**

The technique of using less data to provide the same or comparable analytical conclusions is known as data reduction. This is frequently done to address issues with enormous datasets, such as the need to simplify analysis, storage constraints, and processing efficiency. By reducing the size of the dataset without materially sacrificing its integrity or utility, data reduction techniques seek to preserve the pertinent information. Since the diabetic dataset already contains valuable information for the study, this step is not utilized.

DATA TRANSFORMATION**:**

The process of changing or modifying data from one format to another in order to satisfy certain needs is known as data transformation. Since the diabetic dataset already contains valuable information for the study, this step is not utilized

**MODEL TESTING AND EVALUATION:**

MODEL TESTING:

We replicated Jupyter Notebook in the cloud using Google Collaboratory for testing all the models. This is the test model that Google Collaboratory produced.

The study was conducted by comparing machine learning algorithms such as Support Vector Machines (SVM), Random Forests, Ridge Classifier, AdaBoost, K Nearest Neighbors (KNN), Gradient Boosting, Decision Trees, to predict loan based on many characteristics in the dataset.We are converting the categorical attributes into numerical attributes using label encoder for easier calculations.

MODEL EVALUATION**:**

We have successfully used the algorithms to predict loan based on several attributes.

Based on a number of Loan prediction characteristics, we have precisely predicted loan using the algorithms.

From the table below, it can be seen that from the six models, the best accuracy was 0.86 in the fourth experiment which is AdaBoost+Ridge Classifier. Next to that the AdaBoost and Ridge Classifier algorithm has produced a good accuracy of 0.82, while Random Forest has produced a good accuracy of 0.80.Then GradientBoosting Classifier has produced an accuracy of 0.78 and KNN has produced an accuracy of 0.78 . For the cross-validation section, we have added it as one to validate the results of the accuracy values we got.

.

1. ALGORITHM

* RIDGE CLASSIFIER:Ridge Classifier is a linear classification algorithm that extends linear regression for classification tasks. It adds a regularization term (L2 penalty) to the linear regression cost function, preventing overfitting by discouraging overly complex models. The regularization term helps the algorithm generalize well to new, unseen data. Ridge Classifier is suitable for scenarios where the feature space is vast, and there's a risk of multicollinearity among features. The regularization parameter, also known as alpha, controls the strength of regularization. Higher alpha values increase regularization, impacting the flexibility of the model. Ridge Classifier is particularly useful when dealing with high-dimensional datasets prone to overfitting in standard linear regression.

* KNN:For classification problems, K-Nearest Neighbors (KNN) is a simple machine learning technique.The algorithm learns labeled instances with corresponding attributes during the training phase.KNN determines the distance between each new, unlabeled data point and each point in the training set. The new point is then classified using the majority class of the k-nearest neighbors, which have been found. In essence, KNN predicts by examining the feature space classes of the adjacent data points, presuming that like points have similar properties. The algorithm's performance is heavily influenced by the distance metric selection and the value of k.
* DECISION TREE**:**Decision Tree is a versatile machine learning algorithm used for classification and regression. It recursively splits data based on key features, forming a tree structure with nodes, branches, and leaves. Each node represents a decision point, branches indicate rules, and leaves signify outcomes. The algorithm selects features to maximize data homogeneity, fostering interpretability. Decision Trees are transparent, making them valuable for understanding decision processes. They find applications in various domains and serve as building blocks for more advanced algorithms like Random Forests.
* RANDOM FOREST**:** Random Forest is an ensemble learning method used for both regression and classification tasks. It operates by constructing multiple decision trees during training and outputs the mode (for classification) or mean prediction (for regression) of the individual trees.The decision tree is the fundamental component of a Random Forest. A decision tree is a structure that resembles a flowchart, with each internal node representing a choice made in response to the input features, each branch representing the choice's result, and each leaf node representing the anticipated end output.A group of decision trees make up a Random Forest. Using a random subset of the training data and a random selection of features for decision-making, many decision trees are constructed during the training phase.
* ADABOOST **:**  AdaBoost, short for Adaptive Boosting, is an ensemble learning technique used for classification and regression tasks. It's a sequential learning algorithm that combines weak learners (models slightly better than random guessing) to create a strong learner.

1. Training Weak Learners in Sequential Order:A sequence of weak learners, such as decision trees or stumps (small, basic decision trees with a single split), are successively trained as part of AdaBoost's training process.

2. Pay Attention to Mislabeled Data:Every weak learner after that concentrates more on the training cases that the earlier models mislabeled. By giving these incorrectly categorized cases larger weights, it essentially gives hard-to-classify observations priority.

3. Weighed Learner Combination:The precision with which weak learners forecast the target variable determines how they are weighed. Higher weights are assigned to models that perform better in the end combination.

4. Final Forecast by Weighted Voting:To create the final prediction, AdaBoost aggregates the predictions of all weak learners using a weighted majority vote or weighted sum.

1. CONCLSION

In summary, the study concludes that AdaBoost+Ridge Classifier stands out as the most effective algorithm for loan prediction, providing accurate and robust results. To further improve predictive performance, continuous feature engineering, hyperparameter tuning, and exploring alternative ensemble methods are recommended. Balancing model complexity and interpretability is crucial for transparent decision-making. Continuous monitoring, adaptation to market conditions, and dataset expansion ensure ongoing model relevance. Implementing these suggestions enhances machine learning models for more accurate and reliable loan approval systems.

REFERENCES

* Kumar Arun, Garg Ishan, Kaur Sanmeet- Loan Approval Prediction based on Machine Learning Approach- IOSR Journal of Computer Engineering, p-ISSN: 2278-8727 PP 18-21.
* Kshitiz Gautam, Arun Pratap Singh, Keshav Tyagi - Loan Prediction using Decision Tree and Random Forest- (IRJET) Volume: 07 Issue: 08 | Aug 2020.
* Sujoy Barua, Divya Gavandi- Swindle: Predicting the Probability of Loan Defaults using CatBoost Algorithm- ICCMC 2020.
* Bhoomi Patel, Harshal Patil, Jovita Hembram- Loan Default Forecasting using Data Mining- (INCET) Belgaum, India. Jun 2021.