BANK LOAN RISK PREDICTION USING MACHINE LEARNING ALGORITHMS.

ABSTRACT

Loan accounts are the majority of bank profits. Even though a large number of people are enquiring about loans, It's difficult to find a real borrower who will return the debt. When doing the procedure manually, there is a risk of making mistakes in selecting the right candidate. As a result, we're working on a machine learning-based loan prediction system that will automatically choose qualified applicants. Acceptance of a loan is a vital step for financial firms. The system determines whether the loan is authorised or denied. Loan recovery has a significant impact on a bank's financial results. It's tough to estimate a customer's ability to pay back a loan. In recent years, a large number of scholars have been researching on algorithms that anticipate loan acceptance. When dealing with massive volumes of data, machine learning techniques are particularly useful for predicting outcomes. Machine learning methods such as Logistic Regression (LR), Decision Tree (DT), SVC, KNeighborsClassifier(KNN), are used in this work to predict client loan acceptance. In terms of accuracy, the Logistic Regression machine learning algorithm surpasses KNN, SVC, and Decision Tree learning approaches, according to the results of the experiments.

Keywords: Machine Learning, Logistic Regression, Decision Tree, SVC, KNN.

1. INTRODUCTION

Banks' primary business is lending. Interest on the loan is the principal source of income. The loan businesses award a loan following a thorough verification and validation process. They remain pessimistic about the ability of the borrower to pay back the debt on schedule

To forecast the result, a Prediction Model employs data mining, statistics, and probability. In any model, predictors are variables that are projected to influence future outcomes. When data is collected from diverse sources, a statistical model is developed. A simple linear equation or a sophisticated neural network mapped with complicated software can be used. As additional data becomes available, the model becomes more sophisticated, and the error lowers, allowing it to forecast with the least amount of risk and in the shortest amount of time possible. The Prediction Model benefits banks by reducing the risk involved with the loan approval processes, and it benefits applicants by shortening the process time. The two major banking concerns are as follows: 1) What is the borrower's risk level? 2) Should we lend to the borrower, considering the risk? The interest rate, along with other indicators, is used to determine the riskiness of the borrower. The interest rate rises in proportion to the borrower's risk. We'll see if the application is approved for the loan based on the interest rate. Creditors are given loans by lenders in return for interest-bearing repayment guarantees. Only if the borrower repays the loan does the lender

receive reimbursement. The lender loses money if the borrower does not repay the loan. Customers are given loans by banks in exchange for repayment. Sometimes people would default on their payments due to several causes. In the case of a default, the bank has insurance on hand to mitigate the danger of a banking collapse. The insured amount might cover the full loan or only a portion of it. To determine if a customer is qualified for a loan, banking operations rely on traditional procedures. When there were a significant number of loan applications, manual techniques were most successful, but they were insufficient. Right now, making a decision would take a long time. As a result, the machine learning model for loan prediction may be used to assess and build strategies for a customer's loan situation. This model extracts and presents the essential borrower factors that define the consumer's loan status. Finally, the intended outcome is achieved

III.IMPLIMENTATION

Step 1: Pre-processing and loading the dataset. Importing relevant libraries and packages, as well as loading the dataset as a pandas data frame, is the first and most important step. We also do pre-processing by identifying null values and filling in missing values for numerical and categorical variables using the mean and mode, respectively.

STEP 2: Data and log transformation exploration

Any successful data analysis effort must include visualisation and an effective survey of existing variables. For each Applicants' Loan ID in the dataset, there are 12 characteristics.

The Target Variable is Yes(Y)/No(N), which must be forecasted using the test data. The depiction of numerical characteristics is univariable and uses a statistical bar graph. And we apply log transformation for numerical attributes to replace each variable with a(log used to remove or reduce skewness of our original data.

Step 3: Correlation matrix

A correlation matrix is a table that shows how two variables are related. The measure performs best when the variables are connected linearly. To see how well the data fits together, a scatterplot might be utilised. Variables are expressed as rows and columns in a correlation matrix.

STEP 4: Label encoding and train-test split. The process of encoding labels into numeric representation so that machines can read them is known as label encoding.

Train and test subsets should be separated into matrices or arrays.

Next (ShuffleSplit) is a little programme that encapsulates input validation and the following steps.

In a one-liner, combine split(X, y) and the application to enter data into a single call for data splitting.

STEP 5: Training a model, check the accuracy, cross accuracy

We must fit each model independently during the training process. Fit(X_train,y_train). Find the accuracy and cross-accuracy

IV. METHODS

Logistic Regression

Logistic regression is a supervised learning categorization method that allows us to estimate the output values for new data using correlations learned from previous sets of data.

K-Nearest Neighbor(KNN)

This technique is amongst the most fundamental Machine Learning algorithms based on Supervised Methods. The new model and existing cases are comparable, according to K-NN, and the new instance is placed in the category that is closest to the existing categories. This mechanism keeps track of all data and sorts new data into categories based on how similar it is to what's already there. This means that fresh data may be quickly sorted into the correct category using this method. It may be used for both regression and classification; however, the latter is the more common use. No data assumptions are used in the non-parametric approach.

SVC

Clustering vectors is encouraged (SVC). sklearn.svm is a C-based support vector classification system based on libsvm. Scikit-learn makes use of the SVC module. This class is in charge of providing one-to-one multiclass support.

Decision Trees

All qualities or choices should be dismissed according to the basic algorithmic rule of call trees. The greatest amount of information gained from the possibilities is utilised to choose qualities. Within the kind of IF-THEN rules, the data depicted in the call tree will be delineated. Quinlan's C4.5 classification techniques are an extension of this paradigm.

V. RESULT

Here are all of the techniques we created, and these methods use cross-validation to assess accuracy.

The values acquired for the various metrics using the various approaches are shown in the table below. Here, we pick accuracy; therefore, decision tree classification is the least accurate of all approaches.

As a result, we conclude that Logistic Regression performs well in predicting our data.

VI. CONCLUSION

When dealing with massive volumes of data, machine learning techniques are particularly useful for predicting outcomes. Logistic Regression, Decision Tree Classifier, SVC, and KNeighbors Classifier (KNN) were shown to produce the best results. Logistic Regression has the highest accuracy of these five techniques. After a thorough examination of the part's advantages and limits, it can be confidently concluded that the product will be a highly efficient component. This application is up to date and meets all Banker requirements. In many systems, this section is simply occluded. The experimental findings show that the Logistic Regression machine learning algorithm is better than the KNN, SVC, and Decision Tree machine learning techniques in terms of accuracy.