Loan Defaulter Prediction: Random Forest

Concept:

A loan defaulter prediction system using random forest machine learning algorithm is a system that uses a **supervised learning technique** that combines the output of multiple decision trees for classification and regression problems. The system can help lenders to assess the credit risk of potential borrowers and make informed decisions on whether to approve or reject their loan applications. The system can also help borrowers to improve their credit score and financial behaviour by providing feedback and recommendations.

Problem Statement:

When borrowers do not pay back their loans on time, it causes serious problems for both them and the lenders. Borrowers risk legal consequences, fines, and lower credit ratings, while lenders suffer financial losses and reputation damage. Hence, a trustworthy and precise system that can forecast the probability of loan default is essential to assist lenders and borrowers in making smarter decisions.

Objective:

This project aims to create a system that uses random forest machine learning algorithm to predict loan default probability based on loan applicants' data and history. The system will also give feedback and suggestions to borrowers to enhance their credit score and financial habits. The system will measure and compare the model's performance with other algorithms using various metrics. The system will help lenders and borrowers make better decisions by providing prediction results, risk scores, and personalized advice.

Data Source:

https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior

Data Analysis Software:

Python & Jupyter Notebook Libraries used:

- Numpy: To solve complex mathematical problems
- Pandas: Use for dataframe manipulation
- Seaborn: To create data visualization
- Matplotlib: To create data visualization
- ipywidgets: Interactive analysis
- sklearn: Implement complex machine learning algorithm
- Imblearn: Provides tools for dealing with classification and imbalanced campfire.

Data Visualization:

Bar Graphs and Line Charts will be used using for better visualization.

Microsoft PowerBI shall also put a good impact on understanding the parameters of the project.

Methodology:

The methodology behind the project of loan defaulter prediction system using random forest algorithm can vary depending on the type of system, the data sources, and the models. However, a general methodology can be summarized as follows:

- **Data collection:** The system collects various data from the loan applicants, such as their personal information, income, expenses, assets, liabilities, credit history, employment status, etc. The system also collects historical data of loan repayments and defaults from previous borrowers.
- **Data preprocessing:** The system preprocesses the data and transforms it into numerical or categorical features that can be used for machine learning models. The system also handles missing values, outliers, and imbalanced classes in the data.
- **Model building:** The system selects a number of decision trees that it wants to build and randomly selects a subset of data points and features for each tree. The system builds and trains the decision trees using the selected data points and features, using a split criterion such as gini index or entropy to create the nodes. The system uses random forest machine learning algorithm to combine the output of multiple decision trees for classification and regression problems.
- **Model evaluation:** The system evaluates the performance of the model using metrics such as accuracy, precision, recall, F1-score, ROC curve, etc. The system also compares the results of random forest with other machine learning algorithms such as logistic regression, decision tree, neural network, etc.
- **Model deployment:** The system deploys the model and uses it to predict the probability of default for each loan applicant and assigns them a risk score or a risk category (such as low, medium, high). The system provides the prediction results and the risk scores or categories to the lenders and the borrowers, as well as personalized feedback and recommendations to the borrowers based on their data and prediction results.

Probable Outcome:

The probable outcome of the loan defaulter prediction model using random forest algorithm under machine learning is a reliable and accurate system that can predict the likelihood of loan default and help lenders and borrowers make better decisions. It can handle complex and large datasets and mitigate overfitting and bias-related inaccuracy. It can provide feedback and recommendations to the borrowers to improve their credit score and financial behaviour. It can help lenders to assess the credit risk of potential borrowers and make informed decisions on whether to approve or reject their loan applications. It can help borrowers to understand their credit risk and take actions to improve their credit score and financial behaviour.

Thank You
Team ArtiSci'09

Sri Sri University