Problem Statement

Loans account for a large portion of bank profits. Despite the fact that many people are looking for loans, finding a legitimate applicant who will return the loan is difficult. Banks must determine if the borrower will be able to repay the loan. It is crucial to know whether or not the borrower is going to be in safe hands. Therefore, machine learning models are created to predict whether or not the customer would repay the loan.

Objective

In the given problem statement, parameters like Gender, marriage, number of dependents, Education, Self Employment, Applicant Income, Co-applicant Income, Loan Amount, Property Area, Loan Amount Term and Credit History are used to determine whether the person should be lend loan or not. The data is visualized and a machine learning model is developed to predict loans.

Methodology

The given dataset has 614 rows and 13 columns. The missing values in the dataset are replaced with mean or median of feature. A model trained with the removal of all missing values creates a robust model. After the preprocessing of the data, the data is visualized and following insights are drawn:

Data Visualization and Analysis

1. Gender Distribution:

    • The approval rate for males is 69%, while for females it's 66.9%.

2. Graduation and Approval:

    • The approval rate for graduates is 70.8%, while for non-graduates it's 61.9%.

3. Residential Area Analysis:

    • The dataset includes people from different residential areas:

    • The highest approval rate is seen in semi-urban areas (76.82%), followed by urban areas (65.8%) and rural areas (61.45%).

4. Marital Status Impact:

    • Married individuals have a higher approval rate at 71.8% compared to unmarried individuals at 62.9%.

5. Credit History Influence

• Individuals with a credit history rating of 1 have a significantly higher approval rate at 79.04% compared to 7.8% for those with a rating of 0.

6. Employment Status Breakdown:

• Self-employed individuals have a slightly higher approval rate at 68.29%, compared to 68.79% for those who are not self-employed.

7. Approval Rate Across Categories:

• Graduates, males, and self-employed individuals have relatively similar approval rates, ranging from 68.29% to 70.8%.

These insights highlight the various factors influencing approval rates in the given dataset, such as gender, education, residential area, marital status, credit history, and employment status. Understanding these relationships can aid in making informed decisions for the coding project

Cleaning Dataset

The data is then split into train and test model. Label encoding is applied to all the selected test and train features. The label encoding approach is used to convert category variables into numerical format. The machine learning models can only function on numerical data that's why it is very helpful when working with methods that need numerical input. The train and test dataset are standardized. Variables that are measured at different scales do not contribute equally to the model fitting & model learned function and might end up creating a bias. Therefore, feature-wise standardization is often used before model fitting to address this potential issue.

Implementation of the model

The k-fold cross validation technique is applied for evaluating the predictive models. The dataset is divided into 10 subsets or folds. The 5 different algorithms- Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbour and Gaussian Naive Bayes are applied to find out the best performer. Logistic Regression had the highest accuracy of around 80%.

Conclusion

The program once given a large data set can accurately predict and correlate the various factors affecting the credit worthiness of a customer by looking and analysing different aspects about the customer's profile and records