Loan Approval Prediction System

CSE422 PROJECT REPORT

Artificial Intelligence

Group: 07

Ahasan Habib - 24141256

Raiyanul Islam Siam - 24141261

Table of Contents

INTRODUCTION

In contemporary financial landscapes, the efficient allocation of resources is crucial for economic growth and stability. One pivotal aspect of resource allocation is the granting of loans. However, the traditional methods employed for loan approval often suffer from inefficiencies, subjectivity, and sometimes bias. Machine learning (ML) algorithms present a promising avenue to enhance this process, offering the potential for more accurate, fair, and efficient loan approval decisions.

This report aims to delve into the development and implementation of a Machine Learning-based Loan Approval Prediction System. Through this system, we aim to leverage historical loan data to predict the likelihood of approval for future loan applicants. By harnessing the power of ML algorithms, we endeavor to create a robust, automated, and data-driven decision-making framework for loan approval.

DATASET DESCRIPTION

- Features: The dataset consists of 12 features, including no_of_dependents, education, self_employed, income_annum, loan_amount, loan_term, cibil_score, residential_assets_value, commercial_assets_value, luxury_assets_value, bank_asset_value, and loan_status.
- Classification Problem: The task is a classification problem since the target variable loan_status has discrete classes of "Approved" and "Rejected".
- Data Points: The dataset contains 4,269 data points.
- Feature Types: The features include a mix of quantitative and categorical variables.

- Correlation Analysis: A correlation analysis can be conducted to understand the relationships between the input features and the output feature loan_status. used seaborn heatmap, calculates the correlation matrix of the DataFrame between -1 to 1.

  -1 indicates a perfect negative correlation

  0 indicates no correlation

  1 indicates a perfect positive correlation.

DATASET PRE-PROCESSING

- Null Values: This dataset contains no null values, hence no requirement to handle null values.
- Encoding Values: Categorical values in features like education , self_employed and loan_status to 0 and 1
- Delete column: unnecessary column (loan id) which has no relation to the loan being approved.

FEATURE SCALING

The dataset provided contains a mix of quantitative and categorical features. Before training the machine learning models, it is crucial to perform feature scaling to ensure that all features contribute equally to the model training process.

DATASET SPLITTING

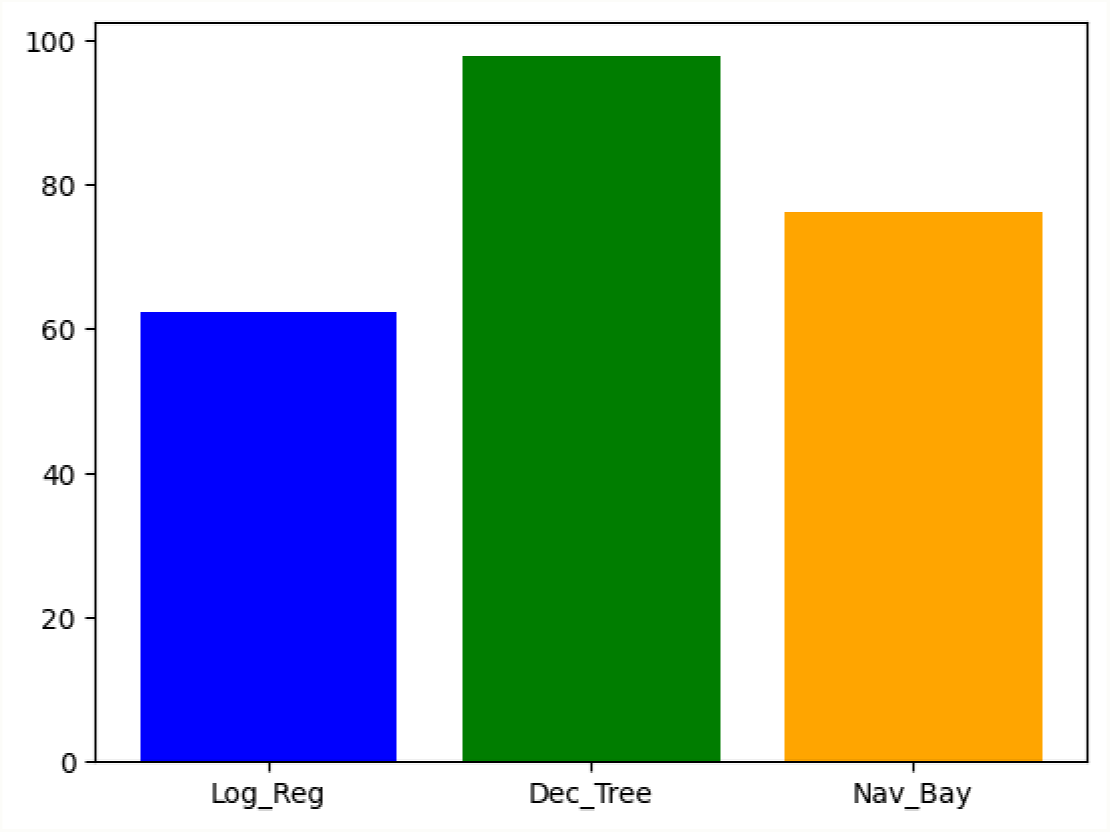
To evaluate the performance of the machine learning models, the dataset needs to be split into training and testing sets.The dataset contains a categorical target variable loan_status with two classes: "Approved" and "Rejected". Since the dataset may be imbalanced, a stratified split is preferred to ensure that the class distribution is maintained in both the training and testing sets.The dataset will be split into a training set (70%) and a testing set (30%) using the train_test_split function from the scikit-learn library. This will ensure that the models are trained on a representative portion of the data and evaluated on unseen data.

MODEL TRAIN AND TEST

1. **Logistic Regression** - Implemented with LogisticRegression from the sklearn.linear_model module. This model is used for binary classification and is based on the concept of odds and probabilities.
2. **Decision Tree Classifier** - Implemented using DecisionTreeClassifier from the sklearn.tree module. Decision trees model decisions and their possible consequences, including outcomes, resource costs, and utility.
3. **Naive Bayes Classifier** - Specifically, we have used the GaussianNB implementation from the sklearn.naive_bayes module. This model is based on applying Bayes' theorem with the assumption of independence among predictors.

MODEL COMPARE, SELECT, AND ANALYZE

Model Names	**Logistic Regression**	**Decision Tree**	**Naive Bayes**
Accuracy In %	*61.71%*	*96.72%*	*75.05%*



Given the above analysis, the Decision Tree model seems to be the best choice for this particular dataset because of its high accuracy and the ability to handle complex relationships in the data. However, the potential for overfitting should be considered, and techniques like pruning or setting a maximum depth could be used to mitigate this risk. While the Decision Tree outperforms the other models in terms of accuracy, a comprehensive approach involving cross-validation, model tuning, and perhaps ensembling different models could lead to even better performance and robustness. Understanding the domain and considering practical aspects of model deployment are also critical steps in the final model selection process.

CONCLUSION

This report has demonstrated the application of machine learning (ML) techniques in enhancing the efficiency and fairness of loan approval processes. By analyzing a dataset with diverse loan applicant features using Logistic Regression, Decision Tree, and Naive Bayes classifiers, the Decision Tree model proved superior, achieving an accuracy of 96.72%. This model effectively handles complex interactions between features but requires careful management to avoid overfitting. To ensure robustness, further steps such as cross-validation, hyperparameter tuning, and possibly exploring ensemble methods are recommended. Implementing an ML-based system can accelerate decision-making, reduce default risks, and minimize human biases, enhancing customer satisfaction. Continuously improving and adapting these models is crucial for maintaining a competitive edge in the financial services sector, thereby promoting economic stability and equitable resource allocation.

REFERENCES:

Our data set : Loan Approval Prediction DataSet