Project Report: Loan Prediction Using Machine Learning

Objective

The primary goal of this project is to create a machine learning model that predicts whether a loan application should be approved based on various features provided in the dataset. This classification task aims to differentiate between approved and denied loan applications.

1. Dataset Preparation

1.1 Overview

The dataset includes multiple features describing loan applicants, such as demographic details, loan information, and financial status. The target variable is binary, indicating whether a loan was approved or denied.

1.2 Data Cleaning and Preprocessing

- Handling Missing Values: Missing values were addressed appropriately. For numerical features, the mean or median was used for imputation, while categorical features were filled using the mode.

- Categorical Encoding: Categorical variables, such as gender, marital status, and education level, were transformed using techniques like one-hot encoding or label encoding to prepare them for model training.

- Feature Scaling: Numerical features were standardized to ensure uniform contribution to model performance.

- Splitting Data: The dataset was divided into a training set (80%) and a testing set (20%) for evaluating model performance on unseen data.

2. Model Selection and Training

2.1 Model Choice

Various machine learning models were considered, including:

- Random Forest Classifier

- Naive Bayes

- Decision Tree Classifier

- K-Nearest Neighbors (KNN)

These models were selected due to their effectiveness in classification tasks.

2.2 Hyperparameter Tuning

- Random Forest Classifier: Hyperparameters like the number of trees (n_estimators) and maximum tree depth (max_depth) were optimized through grid search and cross-validation.
- Naive Bayes: The Gaussian Naive Bayes variant was employed, suitable for continuous data.
- Decision Tree Classifier: Parameters including max_depth, min_samples_split, and min_samples_leaf were fine-tuned to prevent overfitting.
- K-Nearest Neighbors (KNN): The ideal number of neighbors (k) was identified through cross-validation.

2.3 Model Training

Each model was trained using the training dataset, learning to map input features to the target variable (loan approval status) through their respective algorithms.

3. Model Evaluation

3.1 Performance Metrics

The models were assessed using several metrics on the test data:

- Accuracy: The proportion of correct predictions made by the model.
- Precision: The ratio of true positive predictions to the total predicted positives.
- Recall: The ratio of true positive predictions to the total actual positives.
- F1 Score: The harmonic mean of precision and recall, providing a balanced measure between the two.

3.2 Evaluation Results

- Random Forest Classifier: Achieved the highest accuracy at 82%, with a precision of 0.80, recall of 0.78, and F1 score of 0.79.
- Naive Bayes: Showed moderate performance with an accuracy of 76%, but had lower precision (0.72) and recall (0.70).
- Decision Tree Classifier: Attained an accuracy of 78%, but exhibited overfitting, resulting in poorer generalization.
- K-Nearest Neighbors: Recorded an accuracy of 75%, with performance varying based on the chosen value of k.

The Random Forest Classifier was determined to be the most effective model for this task, demonstrating a strong balance between bias and variance, making it well-suited for loan prediction.

4. Conclusion

This project successfully created a machine learning model to predict loan approvals using various classification algorithms. The Random Forest Classifier emerged as the top performer, excelling in accuracy, precision, recall, and F1 score. The findings highlight the significance of data preprocessing, model selection, and hyperparameter tuning in developing effective predictive models. Insights from this project can be valuable for financial institutions aiming to assess loan applications efficiently.