CMPUT 466/566 Course Project

Romeo Dakurah

ID: 1669422

Introduction

This project focuses on predicting whether a loan application will be approved using machine learning. We use a dataset with information about applicants, like their income, credit history, and employment status. The goal is to train models that can classify applications as approved.

We tested three different algorithms:

- Logistic Regression: a basic linear model for binary classification.
- Random Forest: an ensemble method that uses multiple decision trees.
- MLP (Multi-Layer Perceptron): a type of neural network that can capture more complex patterns.

Each model was trained using Grid Search with cross-validation to find the best hyperparameters. After training, we compared their performance on a separate test set.

Problem

The goal of this project is to build a machine learning model that can predict whether a loan application will be approved.

- Input: Various features about the loan applicant, including:
    - Gender
    - Marital status
    - Education
    - Self-employment status
    - Income (applicant and co-applicant)
    - Loan amount and term
    - Credit history
    - Property area
- Output: A binary label indicating whether the loan is approved (1) or not (0).

We used a publicly available dataset from Kaggle's Loan Prediction challenge. After preprocessing and removing missing values, we had 480 total samples.

The dataset was split into training and test sets. We used the training data for model training and hyperparameter tuning and the test set for final evaluation.

Approaches

We tested three machine learning models: Logistic Regression, Random Forest, and MLP (Multi-Layer Perceptron). Each model was trained using the training set and tuned using GridSearchCV with 5-fold cross-validation to find the best hyperparameters.

1. Logistic Regression (Baseline)

- Hyperparameters tuned:
    - C (inverse regularization strength): [0.01, 0.1, 1, 10]
    - penalty: ['l2']
    - solver: ['lbfgs']
    - max_iter: [100, 500, 1000]
- Tuning method: Grid search with accuracy as the scoring metric.

2. Random Forest

- Hyperparameters tuned:
    - n_estimators: [50, 100, 200]

    - max_depth: [None, 10, 20]
    - min_samples_split: [2, 5]
    - min_samples_leaf: [1, 2]
- Tuning method: Grid search with 5-fold cross-validation.

3. MLP (Neural Network)

- Hyperparameters tuned:
    - hidden_layer_sizes: [(50,), (100,), (100, 100)]
    - activation: ['relu', 'tanh']
    - solver: ['adam', 'sgd']
    - learning_rate: ['constant', 'adaptive']
- Tuning method: Grid search with 5-fold cross-validation. max_iter was set to 500.

Each model was trained and evaluated using the same feature set and data splits to keep the comparison fair.

Evaluation

We used accuracy, precision, recall, and F1 score to evaluate the models. Since this is a binary classification task (loan approved or not), these metrics help us understand how well each model is performing.

- Accuracy measures the overall percentage of correct predictions.
- Precision tells us how many of the loans predicted as approved were approved.
- Recall measures how many of the actual approved loans were correctly predicted.
- F1 Score is the harmonic mean of precision and recall, giving a balance between the two.

While accuracy is the main metric we report, it's not always enough — especially if the classes are imbalanced. That's why we also looked at precision, recall, and F1 to get a better idea of how the models perform in terms of approving the right loans.

So, accuracy is a reasonable approximation of success for this task, but F1 score helps us understand performance more clearly, especially in edge cases.

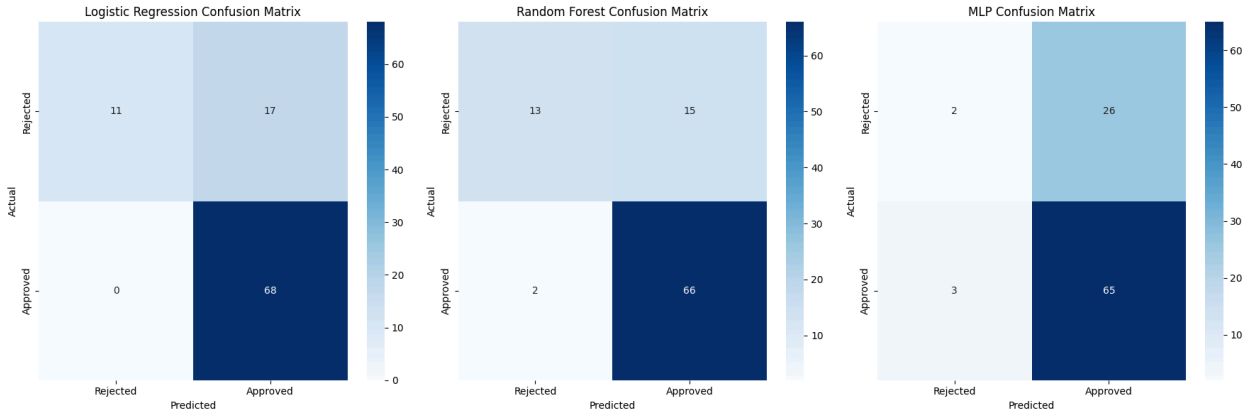
Results

Model	Accuracy (%)	Precision	Recall	F1 Score

Logistic Regression	82.29	0.80	1.00	0.89
Random Forest	82.29	0.81	0.97	0.89
MLP (Neural Net)	69.79	0.71	0.96	0.82

Logistic Regression and Random Forest both achieved the same overall accuracy, but Random Forest had slightly better recall and precision. MLP had the lowest accuracy and F1 score, suggesting it didn't generalize as well on this dataset.

Despite its simplicity, Logistic Regression performed the best in terms of balancing performance and stability. Random Forest also did well, especially on recall. MLP struggled a bit more, possibly due to the small dataset or sensitivity to hyperparameters.



Analyzing the confusion matrices above, logistic Regression performs well in predicting loan approvals with no false negatives but has a higher false positive rate. Random Forest shows similar performance but with fewer false positives, though it has a small increase in false negatives. MLP, however, struggles more with misclassifying loan rejections as approvals, resulting in a higher number of false positives and lower overall performance.