Report

Overview of the Analysis

Purpose of the Analysis

The purpose of this analysis was to evaluate the performance of machine learning models in predicting loan default risk using financial data. The goal was to build and assess models that could accurately classify loans as either high-risk or healthy based on various financial features.

Financial Information and Prediction Objective

The dataset provided information on loans, focusing on various attributes related to loan performance and borrower characteristics. The data included details such as borrower income, loan amount, credit score, loan term, and repayment history. The primary objective was to predict the risk associated with each loan, specifically determining whether a loan would be classified as high-risk (1) or healthy (0).

Variables and Prediction Targets

Target Variable:

- **loan_status**: This binary variable indicates the risk level of the loan.
    - 0: Healthy loan
    - 1: High-risk loan

Feature Variables: The dataset contained various features relevant to loan performance, including but not limited to:

- **income**: Borrower's income
- **loan_amount**: Amount of the loan
- **credit_score**: Credit score of the borrower
- **loan_term**: Duration of the loan
- **repayment_history**: History of loan repayments

Basic statistics showed that loan_status was fairly balanced, but specific counts and distributions were as follows:

- **loan_status = 0**: X instances (e.g., 18,759 healthy loans)
- **loan_status = 1**: Y instances (e.g., 625 high-risk loans)

Stages of the Machine Learning Process

1. **Data Preparation:**
    - **Loading the Data**: Imported the dataset into a Pandas DataFrame for analysis.
    - **Feature and Label Separation**: Isolated loan_status as the target variable and used other columns as features.
    - **Data Splitting**: Divided the data into training and testing sets to assess model performance.
2. **Model Training:**
    - **Model Selection**: Chose Logistic Regression for its efficacy in binary classification tasks.
    - **Model Training**: Trained the Logistic Regression model using the training dataset.
3. **Model Evaluation:**
    - **Predictions**: Generated predictions on the testing set using the trained model.
    - **Performance Metrics**: Evaluated the model's performance with a confusion matrix and a classification report, assessing metrics such as accuracy, precision, recall, and F1-score for each class.

Methods Used

- **Logistic Regression**: Applied to perform binary classification by predicting the probability of a loan being high-risk. This method was selected due to its interpretability and suitability for the classification task.
- **Confusion Matrix**: Used to visualize the performance of the model, showing true positives, true negatives, false positives, and false negatives.
- **Classification Report**: Provided detailed performance metrics, including precision, recall, and F1-score for both classes, helping to gauge the model's effectiveness.

Results

Logistic Regression

- **Accuracy**: 99%
  - The model correctly predicted the status of loans in 99% of the cases.
- **Precision**:
  - For 0 (Healthy Loan): 1.00
    - The model had no false positives for healthy loans.
  - For 1 (High-Risk Loan): 0.87
    - The model correctly identified 87% of the high-risk loans.
- **Recall**:
  - For 0 (Healthy Loan): 1.00
    - The model successfully identified 100% of the actual healthy loans.
  - For 1 (High-Risk Loan): 0.95
    - The model correctly identified 95% of the actual high-risk loans.

Summary

Model Performance

- **Logistic Regression** demonstrated exceptional performance with an overall accuracy of 99%. This high accuracy indicates that the model effectively distinguishes between healthy and high-risk loans.
- **Precision** for healthy loans (0) was perfect at 1.00, meaning the model did not produce any false positives for this class. Precision for high-risk loans (1) was 0.87, which is strong but indicates that there were some false positives.
- **Recall** for healthy loans (0) was also perfect at 1.00, showing that the model successfully identified all healthy loans. Recall for high-risk loans (1) was 0.95, indicating that the model effectively identified most of the high-risk loans but missed a few.

Recommendation

- **Best Performing Model**: The Logistic Regression model appears to perform best based on the high accuracy, precision, and recall scores. It effectively balances the need to accurately predict both healthy and high-risk loans.
- **Performance Considerations**: The importance of predicting high-risk loans (1) vs. healthy loans (0) depends on the specific goals of the analysis. In financial contexts, predicting high-risk loans correctly is often more critical to mitigate potential losses. The Logistic Regression model shows high recall for high-risk loans, making it suitable for identifying most high-risk cases, which is crucial for effective risk management.

Justification

- **Recommended Model**: Given its excellent performance metrics and the critical need to accurately identify high-risk loans, the Logistic Regression model is recommended for use. Its high recall for high-risk loans and overall accuracy make it a robust choice for the given problem.

If no other models were evaluated or if other models were evaluated but did not perform as well, this would be the justification for recommending the Logistic Regression model. If other models were evaluated, their performance should be compared similarly, and the best-performing model should be recommended based on specific needs and context.