**Module and Overview of the Analysis**

This analysis focuses on **predicting loan risk** using **machine learning models**, specifically **Logistic Regression**. The purpose of this analysis is to develop a model that classifies loans as either **healthy (0)** or **high-risk (1)** based on financial data.

**Data Overview**

* The dataset includes various financial factors, such as:
  + **Loan Size** (loan\_size)
  + **Interest Rate** (interest\_rate)
  + **Borrower’s Income** (borrower\_income)
  + **Debt-to-Income Ratio** (debt\_to\_income)
  + **Number of Accounts** (num\_of\_accounts)
  + **Derogatory Marks** (derogatory\_marks)
  + **Total Debt** (total\_debt)
  + **Loan Status** (loan\_status) - **Target Variable** (0 = Healthy Loan, 1 = High-Risk Loan)

The **goal** of the machine learning model is to predict whether a given loan falls into the **"healthy" or "high-risk"** category.

**Machine Learning Process**

1. **Data Preprocessing**:
   * Separated the **features (X)** and **labels (y)**.
   * Split the dataset into **training (80%)** and **testing (20%)** sets.
2. **Model Selection & Training**:
   * Applied **Logistic Regression** to classify loan statuses.
   * Trained the model using **training data** (X\_train, y\_train).
3. **Model Evaluation**:
   * Measured **accuracy**, **precision**, and **recall** using a **confusion matrix** and **classification report**.

**Results**

**Machine Learning Model 1: Logistic Regression**

* **Description**: A **binary classification model** predicting loan status.
* **Performance Metrics**:
  + **Accuracy**: **99%**
  + **Precision**:
    - **Healthy Loan (0)**: **1.00**
    - **High-Risk Loan (1)**: **0.86**
  + **Recall**:
    - **Healthy Loan (0)**: **0.99**
    - **High-Risk Loan (1)**: **0.94**
  + **F1-Score**:
    - **Healthy Loan (0)**: **1.00**
    - **High-Risk Loan (1)**: **0.90**
* **Confusion Matrix Insights**:
  + **Correct Predictions**:
    - **Healthy Loans (0)**: **14,924 correctly classified**
    - **High-Risk Loans (1)**: **476 correctly classified**
  + **Misclassifications**:
    - **False Positives** (Healthy loans misclassified as high-risk): **77**
    - **False Negatives** (High-risk loans misclassified as healthy): **31**

**Summary & Recommendation**

**Model Performance Summary**

* The **Logistic Regression model performed exceptionally well**, achieving an **accuracy of 99%**.
* It correctly classified almost all **healthy loans**, with **very few false positives (77 cases)**.
* It **identified 94% of high-risk loans correctly**, though it still **misclassified 31 high-risk loans as healthy**.

**Final Recommendation**

* The model **performs best at predicting healthy loans (0)**, ensuring that **low-risk borrowers are correctly identified**.
* However, for high-risk loans (1), **some misclassification occurs**. If catching all high-risk loans is crucial, the model’s **decision threshold** should be adjusted or another technique like **SMOTE (Synthetic Minority Over-sampling Technique)** could be used to improve recall.
* **If minimizing financial risk is the priority**, an alternative model like **Random Forest or a Neural Network** may be explored for better recall.

**Recommended Model**: **Logistic Regression with threshold tuning**  
 **Potential Improvement**: Consider **balancing the dataset** or trying **more advanced models** for better high-risk loan detection.