**Credit Risk Analysis Report**

**Overview of the Analysis:**

The purpose of this analysis is to assess the performance of a machine learning model in predicting credit risk. Specifically, we aim to determine whether a loan is likely to be low-risk (healthy) or high-risk (prone to default). This analysis helps financial institutions make more informed lending decisions, potentially minimizing financial losses due to high-risk loans.

The dataset used in this analysis consists of historical loan application information, with features such as loan amount, applicant income, credit history, debt-to-income ratio, and other relevant financial variables. The target variable (loan status) is binary, where:

* + 1 represents high-risk (default-prone) loans.
  + 0 represents healthy loans.

The distribution of loan status in the dataset is highly imbalanced:

y.value\_counts() - 0 75036

1 2500

This shows that 97% of the loans are classified as healthy (low-risk), while only 3% are classified as high-risk (prone to default). No additional measures were taken to address this class imbalance during the analysis. This imbalance significantly impacted the model's performance, as the model is biased towards predicting healthy loans, given their overwhelming majority in the dataset.

**Machine Learning Process:**

1. Data Preprocessing: The data was separated into labels (y) and features (X).

2. Modelling: The primary model used for this analysis was Logistic Regression, which is commonly employed for binary classification problems.

3. Performance Metrics of Logistic Regression: The model's performance was assessed using standard metrics, despite the imbalanced nature of the data:

Accuracy: 99% — The model correctly classified 99% of loans as either high-risk or healthy.

However, this high accuracy is misleading due to the class imbalance, as the majority of loans are healthy (low-risk).

**Precision and Recall:**

**For healthy loans (label 0):**

* + Precision: 1.00 (100% of loans predicted as healthy were actually healthy).
  + Recall: 0.99 (99% of actual healthy loans were correctly identified).

**For high-risk loans (label 1):**

* + Precision: 0.85 (85% of loans predicted as high-risk were truly high-risk).
  + Recall: 0.91 (91% of actual high-risk loans were correctly predicted).

**Results Summary:**

Logistic Regression Model:

* + Accuracy: 99%
  + Precision (Healthy Loans): 1.00
  + Recall (Healthy Loans): 0.99
  + Precision (High-Risk Loans): 0.85
  + Recall (High-Risk Loans): 0.91

**Strengths:**

High Accuracy: The model achieved high accuracy, correctly classifying most loans. However, this result is heavily influenced by the large number of healthy loans.

Excellent Performance on Healthy Loans: The model demonstrated perfect precision (1.00) and near-perfect recall (0.99) for healthy loans, meaning it is highly reliable in identifying low-risk loans.

Good Recall on High-Risk Loans: The model identified 91% of high-risk loans correctly, which indicates it is somewhat effective in recognizing default-prone loans.

**Limitations:**

Impact of Class Imbalance: The significant class imbalance (97% healthy loans, 3% high-risk loans) has affected the model’s performance. The model is biased towards predicting healthy loans due to their overwhelming majority. While the overall accuracy is high, this is largely due to the correct classification of healthy loans.  
For high-risk loans, the precision (0.85) indicates that 15% of loans predicted as high-risk were actually healthy. These false positives could lead to creditworthy customers being incorrectly flagged as high-risk, potentially impacting lending decisions and customer relations.

Moderate Precision for High-Risk Loans: The precision for high-risk loans (0.85) reveals that the model is less reliable in correctly identifying high-risk loans. This could result in missed opportunities to minimize default risk and unjustified denial of loans to healthy applicants.

**Summary and Recommendations:**

The Logistic Regression model performed well in predicting healthy loans, with high accuracy, precision, and recall for this majority class. It is suitable for deployment in environments where most loans are expected to be healthy, and accuracy in identifying those loans is crucial.

However, the model’s performance for high-risk loans is not optimal. The class imbalance has significantly affected the model, leading to a tendency to predict loans as healthy. This affects both the precision and reliability of high-risk loan predictions.

**Potential Improvements:**

Addressing the Class Imbalance: Consider implementing techniques such as oversampling the minority class(high-risk loans) or under-sampling the majority class (healthy loans) to create a more balanced dataset.

Exploring Advanced Models: More complex models such as Random Forest or Gradient Boosting could help capture non-linear relationships in the data and improve the prediction of high-risk loans.

Threshold Tuning: Adjusting the classification threshold could strike a better balance between precision and recall for high-risk loans.

**Final Recommendation:**

The model can be deployed with caution, particularly for classifying healthy loans. However, improvements should be made if accurate predictions of high-risk loans are critical to the lending process. In high-stakes environments, where false positives or false negatives have serious financial consequences, further enhancements are necessary to mitigate risks.