Credit Risk Analysis Report

# Overview of the Analysis

Lending companies operate on the premise of providing financial resources to borrowers with the expectation of repayment. However, inherent in this process is the risk of borrowers defaulting on their obligations, leading to financial losses for the lender. To mitigate this risk, lenders employ various methods, one of which is using machine learning algorithms to analyze historical lending data. In this analysis, we delve into the application of the Logistic Regression algorithm to assess credit risk based on a dataset from a peer-to-peer lending services company.

# Methodology

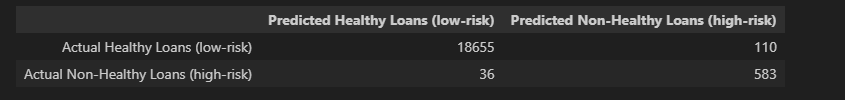
The Logistic Regression algorithm is chosen for its effectiveness in classification tasks, particularly in predicting the probability of a target variable, which in this case is the creditworthiness of borrowers. The dataset provided by the lending company is used to train and test the model, with the aim of distinguishing between healthy (low-risk) and non-healthy (high-risk) loans.

# Model Performance

Upon fitting the Logistic Regression model to the dataset, an impressive accuracy score of 99% is achieved. However, a deeper examination reveals a disparity in the recall values between healthy and non-healthy loans. While the model exhibits a recall value of 0.99 for healthy loans, indicating strong performance in identifying them correctly, the recall value for non-healthy loans is lower at 0.94. This discrepancy can be attributed to dataset imbalance, where the majority of the data belongs to the healthy loan category.

# Evaluation Metrics

The confusion matrix provides further insight into the model's performance. Out of 18,765 healthy loans, the model accurately predicts 18,655, with only 110 misclassifications. Conversely, for the 619 non-healthy loans, the model correctly identifies 583, but misclassifies 36. This asymmetry highlights the model's tendency to label loans as healthy more frequently, leading to a higher rate of misclassification for non-healthy loans.

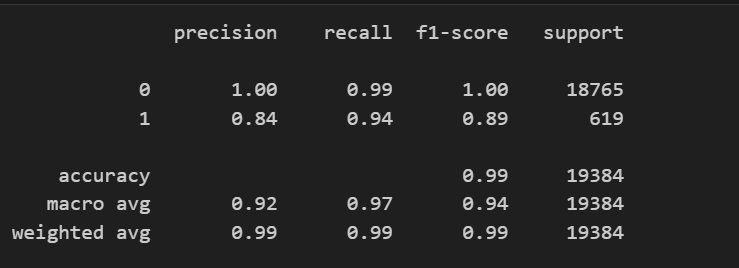


# Results

The Logistic Regression model fitted with the DataSet predicted healthy loans 100% of the time and predicted non-healthy loans 84% of the time.

* The model fitted with data has a higher possibility of making these mistakes:
  + a healthy loan (low-risk) is classified as a non-healthy loan (high-risk).
  + a non-healthy loan (high-risk) is classified as a healthy loan (low-risk).

According to the models recall scores, the model made 1% of mistakes when predicting healthy loans and made 6% of mistakes when predicted non-healthy loans.



The model generated an accuracy score of 99% .

# Summary

In this analysis, we aimed to develop a machine learning model to assess the creditworthiness of borrowers using historical lending data from a peer-to-peer lending service. We utilized the Logistic Regression algorithm, commonly employed for classification tasks, to predict whether loans are healthy (low-risk) or non-healthy (high-risk) based on loan status provided by the lending company.

The Logistic Regression model achieved an impressive accuracy score of 99%. However, upon deeper analysis, we found that the model's performance was affected by the imbalance in the dataset, with a majority of instances belonging to one class label. This resulted in lower recall values for non-healthy loans compared to healthy loans, indicating that the model is better at predicting healthy loans than non-healthy ones.

Specifically, out of 18,765 healthy loans, the model correctly predicted 18,655 and incorrectly predicted 110. Conversely, out of 619 non-healthy loans, the model correctly predicted 583 and incorrectly predicted 36. This imbalance led to a higher rate of misclassification for non-healthy loans.

The model exhibited a higher probability of classifying loans as healthy (low-risk) rather than non-healthy (high-risk), with healthy loans being predicted correctly 100% of the time and non-healthy loans predicted correctly 84% of the time.

While the model achieved a high accuracy score, its performance, particularly in identifying non-healthy loans, could be improved, possibly through techniques such as data resampling or using alternative algorithms better suited to handle imbalanced datasets.