# Loan Risk Machine Learning Report Analysis

Maisie DeGraaf

## Purpose of Analysis

This analysis is to test and implement a machine learning model to predict loans that may either be a high-risk loan or a healthy loan based on financial information provided. This will help be able to predict loan defaults.

## Financial Information

The data used in the training contains information such as:

* Size of the loan
* The interest rate on the loan
* Income of the individual borrowing
* Debt to Income ratio
* Number of accounts
* Total Debt
* Derogatory marks on the individual
* Loan Status

## Variables

The loan status variable consists of two classes:

* Class 0: No default (Majority class)
* Class 1: Default (Minority class)

Value counts for the loan status variable:

* Class 0: 18765 instances
* Class 1: 619 instances

## Stages of Machine Learning

* Stage 1: Data Preprocessing: The data was split into features (X) and the target variable (y), then further divided into training and testing sets.
* Stage 2: Model Selection: Logistic Regression model was chosen for its simplicity and interpretability.
* Stage 3: Model Training: The Logistic Regression model was trained using the training data.
* Stage 4: Model Evaluation: The model was evaluated using the testing data, and performance metrics such as accuracy, precision, and recall were calculated.

## Methods

### Logistic Regression

Logistic regression is a statistical method used for binary classification tasks, where the outcome variable is categorical and has two classes (e.g., yes/no, 0/1). It models the probability that a given input belongs to a particular category.

## Results of Machine Learning Model 1 - Logistic Regression Model

### Class 0 (no loan default)

Precision: 100%

Recall: 99%

F1-score: 100%

### Class 1 (Loan Default high risk)

Precision: 85%

Recall: 91%

F1-score: 88%

### Overall

Accuracy F1-score (weighted average): 99%

## Summary

Logistic Regression model demonstrated a high accuracy and precision when used in predicting loan defaults. Overall, the F1-score accuracy is 99%, and is rated very high. In terms of no default (0) data, it indicates incredibly high precision scores and recall scores. This model is an excellent machine learning model particularly for no default, low risk loans. In terms of predicting high risk loans, the model still scores quite high with an 85% precision and a 91% recall. This is still decent in terms of performance and reaches a good threshold of 85% or more, but there is also still room for improvement in terms of predicting high risk loans in the (0) category.

Overall, considering the high accuracy and balanced precision and recall scores, the logistic regression model is recommended for predicting loan defaults in this scenario. The performance of the model may vary depending on the specific problem context, such as the importance of correctly predicting defaults versus non-defaults. It is a better fit model to use rather than a model such as a linear regression, since a linear regression predicts continuous values, whereas logistic regression will predict the probability of a binary outcome, between 0 and 1, a high risk or a low-risk loan.