# Loan Default Prediction Challenge

Luo Zheng — Xu Jiajie — Chen Yiman — Advised by: Prof. Kam Tin Seong

## INTRODUCTION

As the economy develops, the public's attitude towards loans has changed. Loans have become an important part of our daily lives, applying and using appropriately can improve our standard of living. Loan prediction is a crucial tool for financial companies, as accurate predictions can help avoid financial losses. To improve our loan prediction accuracy, we use a range of visualization techniques and analytics to assess the impact of different factors. This allows us to make informed lending decisions that benefit both customers and company.

## CHALLENGE

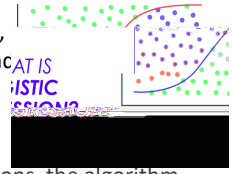
Super Loan is a local digital lending company that relies heavily on loans for revenue. The accuracy of loan default prediction can significantly impact the company's performance. To ensure the accuracy of our predictions, our team builds robust models that assess key drivers of default risk. This enables us to accurately predict repayment odds and make informed lending decisions.

## DATA & APPROACH

**1.Exploratory Data Analysis (EDA) via Data Visualization**

Exploratory Data Analysis (EDA) in the form of visualization techniques will be applied to provide the user portrait.

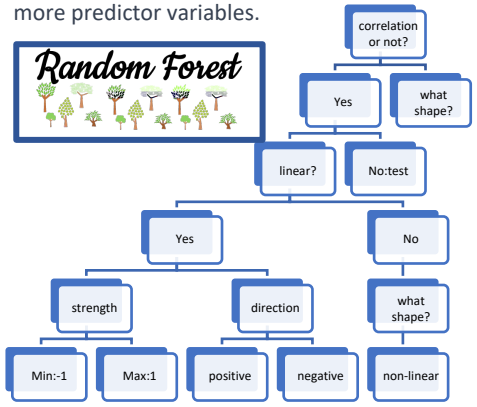
Utilized various visualization techniques, such as bar charts, stacked bar charts, heatmaps, scatter plots, donut chart, and violin plots, to conduct descriptive analysis and understand the relationship of different factors.



**2. Built prediction model's technique**

**Random forest:** a statistical model when making predictions, the algorithm aggregates the predictions of all the decision trees for the final prediction.

**Logistic regression:** a popular statistical model used for binary classification tasks, where the goal is to predict a binary outcome variable based on one or more predictor variables.




**3. R Shiny & Quarto** Utilized R programming to process and perform analyses , as well as developed a website. Key packages used including shiny, readxl, ggstatsplot, ggplot2, shinydashboard, plotly, lubridate, corrplot, tidyverse, leaflet, leaflet.extras, tmap, sf, and tmaptools. The team would use R and Shiny to build up an interactive webpage that can show the different statuses of clients and the prediction.

**Dataset & Outcome**

The objective is to predict whether a loan would default or not based on the demographic and performance data of the customer.

The dataset from Zindi contained three data table: **Demographic data , Performance data and Previous loans data,** containing fields such as customer ID, loan amount, repayment terms, and loan performance.

There would be 2 outcomes, "good" or "bad," indicating whether the loan was settled on time or not.

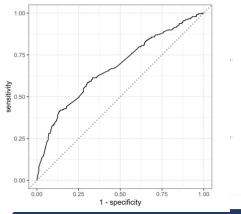
## RESULTS

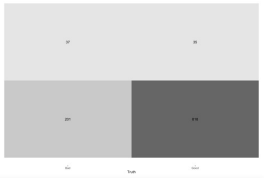
**Descriptive Analysis : How Do The Defaulters and Non-defaulters Look Like?**

## RESULTS

**Correlation Analysis : Do the Factors Matter?**

**Prediction Analysis : Who Will Default?**



## FUTURE WORK

In future, we plan to further optimize the exploratory data analysis process by taking into consideration the customer's residential address. This will enable us to conduct a more in-depth analysis of credit defaults in different regions and identify any potential patterns or trends that may exist. By doing so, we can gain a better understanding of how location impacts credit defaults and use this information to inform our future decision-making.

In addition, we will also focus on improving the accuracy of our prediction model by