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

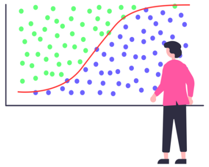
Super Loan is a local digital lending company that relies heavily on loans for revenue. The accuracy of loan default prediction can significantly impact the company's performance. To ensure the accuracy of our predictions, our team builds robust models that assess key drivers of default risk. This enables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

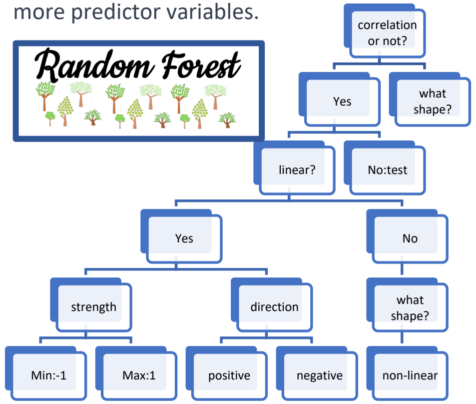
Utilized various visualization techniques, such as bar charts, stacked bar charts, heatmaps, scatter plots, donut chart, and violin plots, to conduct descriptive analysis and understand the relationship of different factors.



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

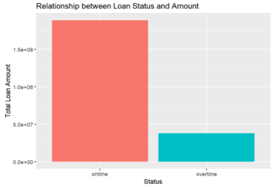
The objective is to predict whether a loan would default or not based on the demographic and performance data of the customer.

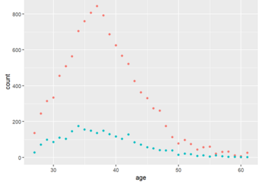
The dataset from Zindi contained three data table: **Demographic data , Performance data and Previous loans data,** containing fields such as customer ID, loan amount, repayment terms, and loan performance.

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

RESULTS

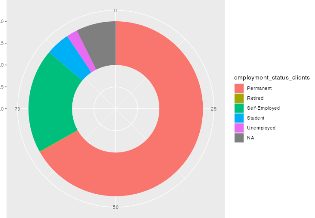
Descriptive Analysis : How Do The Defaulters and Non-defaulters Look Like?

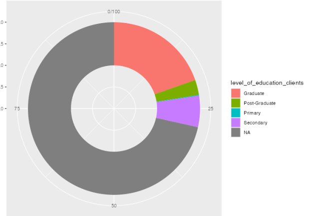




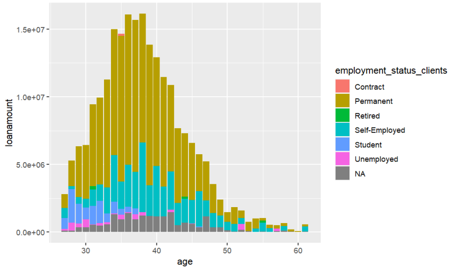
Let's explore the relationship between **loan amount** and **repay status**: the bar chart showed the total amount of on time repayment loans is much higher than overtime one.

To the relationship between age and delinquency, we further analysed the frequency of timely and late payments within different age groups.





In certain age groups and among individuals with certain educational backgrounds and employment statuses, there may be a higher frequency of late payments. It is possible that the higher number of late payments in certain age groups may be attributed to the higher number of loans taken out by individuals in those age groups.



Aged between 30-40 years old or having a **permanent working** status have a higher number of loans, leading to more cases of default. However, individuals with a **secondary education** and a **contract working** status have relatively smaller loans, yet they have a higher incidence of defaults.

RESULTS

Correlation Analysis : Do the Factors Matter?

Although we do plot some graphs that indicated the relationship between factors, but applying in linear regression model, it seems no significant linear between good/bad and other factors

Linear correlation

```
{r}
correlation_table(data=train_data, target="good_bad_flag_number")
```

Description: df [13 × 2]

Variable <chr>	good_bad_flag_n... <dbl>
good_bad_flag_nu...	1.00
Loanamount	0.12
Totaldue	0.11
Loannumber	0.09
Total_loan_number	0.09
Total_due	0.09
Total_loan	0.09
Employment	0.08
Age	0.06
Education	0.03

Prediction Analysis : Who Will Default?

Here is the machine learning prediction model:

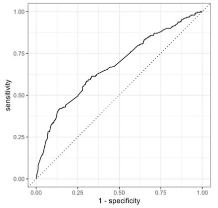
Logistic regression:

Data preparation: The first step is to prepare the data for the logistic regression model. This involves cleaning the data, handling missing values, and encoding categorical variables if necessary. The data is then split into training and testing sets.

Model training : The model is configured to predict the binary target variable "good" or "bad" based on a set of input features. During the training process, the model fits a linear equation to the training data that predicts the probability of the target variable given the input features. The equation is based on the logistic function, which maps the linear equation to a probability value between 0 and 1.

Model evaluation: The ROC AUC estimate of 0.6745 suggests that the model has moderate discriminatory power in distinguishing between good and bad instances. The ROC AUC is a metric that measures the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate at different classification thresholds.

Model refinement: If the performance of the logistic regression model is not satisfactory, it may be refined by tweaking the hyperparameters, feature selection, or data processing steps.



```
{r}
roc_auc(loan_results,
        truth = good_bad_flag,
        .pred_Bad)
```

A tibble: 1 × 3

.metric <chr>	.estimator <chr>	.estimate <dbl>
roc_auc	binary	0.6745446

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 exploring other models that could potentially be more effective.