STAT2604 Personal Project

Course: STAT2604 Introduction to R Programming and Elementary Data Analysis
Total marks: 100
Due date: 23:59, Dec. 17, 2021

This is a personal project and each student should work on their own to finish the project. Please pack your Rmd source code file together with the output pdf file as one compressed file, and submit only that compressed file onto Moodle. Name the compressed file in the format `(Name)_(UID)_P`. Your final submitted pdf file should be limited to at most 8 pages (including all the codes, figures and tables).
(*Note: As our course is more about statistical modeling in R, the marking of this project will mainly focus on your thoughts and reasoning in solving the problem, rather than the perdition performance of your model.*)

1 Problem and Objectives

A bank is now facing a trade-off between accepting customers so that it retains its share in the mortgage loan market and incurring losses due to providing loans to customers that default. The bank managers are interested in the following questions:

1. What is the proportion of good customers that can be granted loans while ensuring that $\alpha\%$ of the bad customers are wrongly identified? ($\alpha = 5\%, 1\%, 0.5\%$)
2. What is the top 3 most important explanatory variables that affect whether a customer is good or bad?

Your task is to undertake a thorough investigation of the dataset provided by the bank, which contain information about past bank customers.

2 Data Description

This dataset contains all the relevant information collected by the bank of their 2000 mortgage loan customers. In total there are 14 explanatory variables and the class label variable indicating whether a customer proved to be good or bad. A bad customer is one that has missed three or more payments during the first year of the mortgage.

Table: Data Description

ID	Customer ID
Annual Income	Annual Gross Income in $s
Credit History	Loan applications in past five years

Credit Cards	Credit cards currently held
Amount	Loan amount
Number of Dependants	Number of family members that rely on the customer
Employment	1 Other 2 Self Employment 3 Part time 4 Full time private sector 5 Full time public sector
Installment Percentage	Monthly installment as percentage of monthly gross earnings
Time at Current Employment	in years
Time at Address	in years
Age	in years
Delayed or Missed Payments	0 No missed/delayed payments over last 3 years 1 Delayed payments only over last 3 years 2 Missed payments over last 3 years
Residential Status	Rent Own Live with Family
Existing Credits	Additional lines of credits
Area indicator	Location of branch receiving application
Good Customer/ Bad Customer (Target variable)	Yes No

3 Task

Write an report that contains your R code, answers and detail explanations to the questions below. The report should be in **pdf** file format (pre-installation of TeX distribution required).

Exploratory Data Analysis:

1. Briefly summarize the data with descriptive statistics. Draw one or two most interesting findings based on your summary statistics. (10 marks)
2. Suggest the top 3 most important explanatory variables that affect whether a customer is good or bad. Support your claims with appropriate visualizations. (15 marks)
3. Are different variables related, and which variables convey information similar to that provided in other variable(s)? (10 marks)
4. Do you find evidence of outliers or other issues with data quality (e.g., incorrect observations)? If there is any, find a proper way to handle the problem. (25 marks)

Modeling:

1. Split the data into training and testing sets. Choose a model to fit the training set. Tune the model if there is any tunning parameters. (10 marks)
2. Choose an appropriate evaluation measure based on the project objective. Explain the reason and calculate that measure on the testing set. Give your answer to the first question in **Problem and Objectives**. (20 marks)
3. Base on your fitted model, find the top 3 most important explanatory variables. Do they conform to your previous suggestion? Provide your final choice of the top 3 variables and comment on the contribution. (10 marks)