IS5312 Group Project Specification

A. Project Topic

You are free to choose any business prediction issues (e.g. Targeted Marketing,Customized Services, Housing Price Estimation, Credit Scoring, etc.) that might interest you.

B. Requirement

The objective of project is to understand how to build business analysis models using Python. You are required to collect the necessary data, and use pandas, seaborn and sklearn toolkit in Python to do data processing, visualization, modeling and model assessment. You might use one or several business intelligence techniques (i.e. decision tree, regression, neural network or others) to accomplish your analytical work.

The following outlines is for your report writing:

1. Introduction (state the problem, your purpose, expected solutions……)
2. Descriptions of Models
3. Model building
4. Business Analysis
5. Model

Assessment6. ….

7. Discussions/Conclusion
8. References (e. g. Theories, Data Sources, etc.)

C. Submission deadline

- Presentation slides (10~12 minutes talk) by Tuesday Week 13 (November 23)
- Project report by Sunday of Week 14 (December 5).

D. Dataset

- **Some useful data websites for possible data sources**
  1. Kaggle: *https://www.kaggle.com/datasets*
  2. Tianchi: *https://tianchi.aliyun.com/home/*

- **Available Dataset**

  In case you have trouble finding dataset that interests you, we provide a back-up dataset "**hmeq.csv**" for you to do group project. You can find the data description below.

  (**Note** that we encourage groups to do project with your own data, but there will not be any penalty point for groups using provided dataset.)

  - **Description for hmeq dataset:**

  A financial services company offers home equity lines of credit to its clients. The company has extended several thousand lines of credit in the past, and many of these accepted applicants (approximately 20%) have defaulted on their loans. By using geographic, demographic, and financial variables, the company wants to build a model to predict whether an applicant will default.

  After analyzing the data, the company selected a subset of 12 predictor (or input) variablesto model whether each applicant defaulted. The response (or target) variable BADindicates whether an applicant defaulted on the home equity line of credit. These variables,along with

their model role, measurement level, and description are shown in the following table：

Name	Model Role	Measurement Level	Description
BAD	Target	Binary	A value of 1 indicates that the client defaulted on the loan or is seriously delinquent. A value of 0 indicates that the client paid off the loan.
CLAGE	Input	Interval	Age of the oldest credit line, measured in months
CLNO	Input	Interval	Number of credit lines
DEBTINC	Input	Interval	Debt-to-income ratio
DELINQ	Input	Interval	Number of delinquent credit lines
DEROG	Input	Interval	Number of major derogatory reports
JOB	Input	Nominal	Occupational categories (six categories).
LOAN	Input	Interval	Amount of the loan request
MORTDUE	Input	Interval	Amount due on the existing mortgage
NINQ	Input	Interval	Number of recent credit inquiries
REASON	Input	Binary	DebtCon=debt consolidation loan. HomeImp=home improvement loan.
VALUE	Input	Interval	Value of the current property
YOJ	Input	Interval	Years at the applicant's current job

Appendix:

- **Some useful data collection tools for your possible reference**
  1. Octopus: *https://www.bazhuayu.com/*
  2. Python Crawlers:
     (1) DIY reference: *https://cuiqingcai.com*
     (2) Scrapy: *https://docs.scrapy.org/en/latest/*
     (3) Pyspider: *http://docs.pyspider.org/en/latest/*