Institute of Systems Science
National University of Singapore

Continuous Assessment -1
Scope: SB and PA

August 2020

Instructions for Submission

Date: by 20 September 2020

Time: by midnight

Place: LumiNUS folder: “CA-1 Submission”

1. This is a group assignment. Not more than five (5) members per group. Single student submissions are fine.
2. Attempt ALL the questions.
3. Submit a single zip containing solutions to both the questions.
4. **Write your name(s) on the front page of each of your submission document**. Use the format shown on next page.
5. Answers should be given in the form of a either a report or power point presentation or both as stated under individual question.
6. All data sets used for each problem are provided.
7. You may use any software to solve the problems.
8. State clearly any assumptions you make in answering any question where you feel the requirement is not sufficiently clear.

[Team Name]

[CA-1: SB & PA]

1. TEAM MEMBER
2. TEAM MEMBER
3. TEAM MEMBER
4. TEAM MEMBER

Question-1 [4 Marks]

Journal Summary

Please summarize a journal that utilizes any predictive analytics method that you have learnt in class into one PowerPoint page. You will need to include the following:

1. Business issues/problem
2. Objectives
3. Data & Process
4. Model
5. Outcome/implication
6. Limitation

Do not repeat journal summary listed in https://analyticsandintelligentsystems.wordpress.com

In addition to avoid any duplication of article a forum is created under BAP (Predictive Analytics) folder of Luminus. Please upload your article link to avoid duplication among yourselves.

Question-2 [8 Marks]

Reporting Frauds

Background:

Company XYZ Pvt. Ltd. has its salespersons making sales every month in their designated markets. The salespeople are free to set the selling price according to their own policy and market. At the end of each month, they report back to the company their transactions. These salespersons have sold a set of products for the company and have reported the sales back to the company at the end of each month. Data on reports submitted by salespersons to the company over a period is under scrutiny. The shared data, for this assignment, is a subset of larger dataset pertaining to a short period. Note that the salespersons were allowed to set the selling price as per their own observations on market.

There are 401,146 such records in the dataset (reports.csv). Initially the company has looked through the records in a random fashion and the inspection outcome is given in one of the columns. Particularly in the column 'Insp' you will find records which are already manually labelled by the company experts as either 'Ok' or 'Fraud'. It is your task to automate this huge task with predictive models. The already labelled records will serve as your source of training & testing data for building a model which later can be applied to score the rest of the records which are labelled 'unkn' under column 'Insp'.

1. Your task is to verify the veracity of these reports given past experience of the company that has detected fraud attempts in these reports?
2. Propose a methodology to score/rank these reports labelled 'unkn' under column 'Insp' with increasing severity of possible fraud so that limited resources of the company can be utilized judiciously.
3. Also (where applicable) comment on:
   a. Data preparation steps (including derived variables)
   b. Sampling & validation considerations
   c. Comparison of accuracy achieved by different techniques
   d. Implementation issues

Variable	Meaning
id	unique id of salesman
prod	id of the sold product
quant	number of product units sold (as reported)
val	monetary value of sales (as reported)
insp	3 values ('ok': if found not fraudulent otherwise 'fraud', unkn: if not inspected)

Question-3 [8 Marks]

Labour Force Participation

Background:

A dataset on labour force participation rates of various countries has been provided. From the given data, make a comparative study of India, China and Singapore. Also, develop a time-series forecasting model for the countries of choice predicting numbers for next 2 years. Explain clearly the choice of model algorithms and the accuracy of the results.