Institute of Systems Science
National University of Singapore

Continuous Assessment -1
Scope: SB and PA

August 2019

Instructions for Submission

Date: by 22 September 2019

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]

Loan Approvals

Background:

Lending Club (LC), which is a website that helps connect lenders with borrowers online. It also collects data on varied aspects of its lenders. Your task is to help lenders formulate a strategy to mitigate losses when the borrowers are unable to pay off their loans. The classification goal is to predict if the client will default (yes/no). Analyse the data using techniques learnt under SB and PA. In your report:

1. Comment on applicability of various raw variables from future business usage (production process standpoint)
2. Data preparation steps (including derived variables)
3. Sampling & validation considerations
4. Comparison of accuracy achieved by different techniques
5. Implementation issues

In addition, express your findings & suggested action plan in a 4-6 page slide deck for the CEO of LC.

Dataset:

Use subset of the dataset available online with 42,514 records and 20 fields provided various pieces of information on various loans to train/test our model.

Refer to the provided dataset (**loans.csv**) and data dictionary provided below for more information on the fields.

S.No.	Field	Brief Description
1	Id	unique id of customer
2	creditpolicy	1 or 0 depending on whether borrower meets certain initial credit criterion set by landingclub.com
3	loanamnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
4	term	The number of payments on the loan. Values are in months and can be either 36 or 60.
5	installment	The monthly payment owed by the borrower if the loan originates.
6	grade	LC assigned loan grade
7	emplength	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.

8	homeownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
9	annualinc	The self-reported annual income provided by the borrower during registration.
10	verificationstatus	Indicates if income was verified by LC, not verified, or if the income source was verified
11	targetloanstatus	Current status of the loan
12	purpose	A category provided by the borrower for the loan request.
13	dti	A debt-to-income ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
14	delinq2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
15	inqlast6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
16	openacc	The number of open credit lines in the borrower's credit file.
17	revolbal	Total credit revolving balance
18	totalacc	The total number of credit lines currently in the borrower's credit file
19	intrate	Interest Rate on the loan
20	revolutil	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.

Question-3 [8 Marks]

Planning for future Ridership

Background:

You are working for a transportation company in a megacity who are currently thinking of enhancing the city rail network. The current passenger pressure is increasing by the day. Therefore, they plan to add few more trains in line with the expected demand so that commuting becomes pleasant for riders. You have monthly ridership data from Jan 2005 - March 2018. The planning department also requested you to study the general rider trend pattern and the seasonal fluctuation pattern so that they can deploy more trains in those months when ridership is high and less trains when the ridership is low. The planning department requests to provide them with a 6 months forecast for ridership and help them formulate a strategy to optimize average number of trains they must deploy at a given period.

Submit your work in a slide deck containing no more than eight slides. In your deck:

1. Data preparation steps
2. Methods used
3. Comparison of accuracy achieved by different techniques
4. Findings

Note that you have no other data for this study and you do not know the data pattern yet.

Dataset: Refer to the file Amtrak.csv. Fields are self-explanatory so a dictionary is not provided.