Department of Statistics, The Chinese University of Hong Kong
STAT 4011 Statistics Projects (Term 1, 2018–19)

Project Outline (Part 1)

Preamble

The course is to provide an opportunity for students to apply their knowledge to solve real-world problem. Students will be required to complete a group project, conduct a final project presentation and submit a written report, on which their assessment will be based. The final report should summarise findings and results obtained by individual group members while it should also provide a summary of justifications for their final conclusion.

Learning Objectives*

At the end of this course, students are expected to

i. develop the ability to understand and implement ideas proposed/discussed in recent Statistics research literature,
ii. illustrate some complex ideas/procedures with laymen terms and
iii. present their findings and conclusions in a concise and professional manner.

* Students are expected to spend 150 hours (5 in-class hours and 145 out-class hours) for this course.

Project Description

This project works with a dataset describing insurance transactions publicly available at Oracle Database Online Documentation (2015) at

http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/anomalies.htm.

The dataset describes insurance vehicle incident claims for an undisclosed insurance company. It contains 15,430 claims; each claim comprises 33 attributes describing the following components:

- Customer demographic details (Age, sex, martial status, and so on)
- Purchased policy (Policy type, vehicle category, number of supplements, agent type, and so on)
- Claim circumstances (day/month/week claimed, policy report filed, witness present, past days between incident-policy report, incident-claim, and so on)
- Other customer data (number of cars, previous claims, driver rating, and so on)
- Fraud found (yes and no)

The goal is to make use of four different machine/statistical learning methods for identifying fraudulent cases. Each member in the team should work on his/her own classification tool while the project presentation and final report should discuss and compare the pros and cons of using various methods. Combinations of ideas are highly recommended.