Big Data Analytics RSM8413 – 2021/2022

Group Assignment 1 Due date: 2021-10-30

Deliverables (to be submitted on Quercus):
1. Report including a detailed description of your findings. At most 10 pages + appendices.
2. All Python source code either in a Jupyter Notebook (*.ipynb) or a Python file (*.py). One file!

Include an Executive Summary (at the beginning) describing your most salient findings. Explain all steps and results clearly and cogently, so that a reasonably intelligent though statistically naïve manager could understand it. You need to include all graphics in your report. Your narrative should be clear and concise, accompanied by supporting evidence in the form of graphics and tables. All tables and graphics should be well formatted (e.g., tables should not run over from one page to another).

The Case:
COVID-19 Behavior Data (COVID-19BehaviorData_CAN_USA.csv)
Fresh from the Rotman MMA program, you are hired into your dream job: You join the Public Health Agency of Canada. You are tasked to study a data set collected by the Institute of Global Health Innovation Imperial College London. This dataset includes information about symptoms, testing, isolation, social distancing, and other COVID-19-related behaviors for tens of thousands of individuals across 29 countries. The data come from interviews, with around 21,000 individuals interviewed each week, from the end of March until the data were downloaded in mid-August.
Note: This is a very large dataset, therefore, it has already been restricted to Canada and USA only and a much smaller number of variables.

Tasks:

Perform data preparation on the data set. Make a table of the data types for every variable in the data set.

Perform EDA, especially with respect to the relationship between the predictor variables and the target, which you must choose! It must be a categorical variable. Clearly explain your choice and the underlying business objectives. Report only the most important results you uncover supported by suitable graphs.

What is a choice of k that balances between overfitting and ignoring the predictor information?

Tune the model.
Carefully consider all points that we discussed in class with respect to k-NN.

Show the classification matrix for the test data that results from using the best model parameters.