Big Data Analytics RSM8413 - Fall 2019

Group Assignment 1 Due date: 2019-10-20

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:
Universal Bank is a relatively young bank growing rapidly in terms of overall customer acquisition. The majority of these customers are liability customers (depositors) with varying sizes of relationship with the bank. The customer base of asset customers (borrowers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business. In particular, it wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors).
A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise smarter campaigns with better target marketing. The goal is to use k-NN to predict whether a new customer will accept a loan offer.
This will serve as the basis for the design of a new campaign.

The file UniversalBank.xls contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

Partition the data into training (75%) and test (25%) sets.

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. Report ONLY the two or three most important results you uncover, with overlay distributions or histograms.

Consider the following customer:
ID = 5001, Age=40, Experience = 10, Income = 84, ZIP Code = 90277, Family = 2, CCAvg = 2, Mortgage = 0, Securities Account = 0, CD Account = 0, Online = 1, and Credit Card = 1. Perform a k-NN classification using k = 1. Specify the success class as 1 (loan acceptance). How would this customer be classified?

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.