CIS 8695 Big Data Analytics

Term Project Description

This group project is one of the most valuable components to you as a future data scientist. You will be using the techniques that you have learned in this class to develop a predictive model and address a related business problem. I expect you to work hard on this project and take it very seriously. Do not underestimate the time it will take to complete this project. It is never too early to start on this project!

General Guidelines: Four students in a group. At the end of the semester, each group will present the model and results in class. Each group submits a copy of the final report (no more than 10 pages).

Project: You find your own business problems by acquiring a dataset. You can acquire datasets from online sources, such as Kaggle.com. You can also use any dataset that you may face in your working experience. From the data, you develop business problems to address (e.g., what to predict? Why?), and then develop analytics solutions.

Group Project Deliverables

The deliverables are written reports.

Written Report. Each group should submit a written report on iCollege by the due date. The report should be in MS WORD format. It should be no more than 10 pages in length (single-space), in 12 point Times New Roman font, 1 inch margin. Please include screenshots as appropriate.

If you refer to outside sources (such as a webpage, a news article, a report etc.), please make sure that you provide appropriate citations to those sources within your written deliverables.

In your project submission, please include both the project report and the relevant data files.

Make your reports as professional as possible. You may follow the outline below to generate your reports.

Suggested Outline of Final Report

1. *Table of Contents* (including page numbers)
2. *Project description*
    - Describe the context, the related business problem, and the goals of the project.
3. *Data Exploration and Preprocessing*
    - Identify the output variable and input variables.
    - Any graphs or plots to explore the data
    - Any missing data? Any data errors? How to deal with them?
    - Any redundant data (correlation is high)? How to deal with this problem?

- Any data transformation (categorical variables to dummy variables, standardization, etc)?
- How do you partition the data (e.g., training and validation)?

4. *Model*

   Throughout the semester, we have covered the following 6 classes of prediction models:

   - **Regression models**, e.g., linear regression, logistic regression, ridge regression, lasso regression, etc.
   - **Tree-based models**, e.g., classification tree, random forest, boosted tree, etc.
   - **Naïve Bayes model.**
   - **SVM models**, with different tuning parameters and kernels (linear, polynomial, etc).
   - **KNN models**, with different K values and distance measures (rectangular, triangular, etc.).
   - **Neural network models**, with different architectures (e.g., hidden layers and nodes) and tuning parameters.

   Your analytics solutions should cover at least **neural network and another 3 classes of model**. For each model class, you do not have to try all different models in the class, but you should at least use one of them.

   We have also learned **ensemble models**, with different ensemble methods (e.g., voting, averaging, and weighted averaging). **You should also use at least one ensemble technique in your solution.**

   Your report should clearly specify:

   - Why you choose certain analytics models;
   - What are the specifications of your models, e.g., what predictors are used? Any important parameters used to build the models?
   - Any variable selection techniques used?
   - Compare between different models and recommend the best model.

   Note: For practice purpose, you are expected to employ multiple different prediction models, such as regression, decision trees, neural networks, etc. You are also expected, although not necessarily required (depending on the nature of your problem), to employ some unsupervised learning techniques, such as PCA, SVD, and clustering.

5. *Results and discussion*

   - Report the estimation/prediction outputs and evaluation matrices.
   - Interpret and discuss the results and summarize the managerial insights (e.g., the actions that one can take in response to the findings).
   - You can use appropriate visualization approaches to better present your results and discussion.

6. *Summary*

   - The lessons learned from doing the project.