ISOM3360 Group Project Guideline

General Goal:

In this project, you will apply the data mining techniques you learned in the class to solve real-world problems. You need explore and search projects on Kaggle (https://www.kaggle.com/) or other websites and choose any business problem that you are interested in. You also need to make sure the related data sources are available. After that, you can apply some data mining algorithms to your data and evaluate the performance of your algorithms. Finally, you should submit a project report, together with your data. Python (Anaconda Jupyter notebook) is expected as your main tool to use.

Evaluation Criteria:

Your project report will be graded based on the **effort** instead of model performance. Therefore, please record your step-by-step progress clearly. You can start with a very simple model and improve the performance by trying different ways of doing the modeling. The possible efforts include data cleaning, missing data handling, outlier detection, feature engineering/selection, learning algorithm selection, parameter tuning etc. Your final model should be the best performer among the trials. To evaluate the performance, a proper evaluation scheme should be adopted. Clarity and organization of your written report are important when evaluating your project. Please explain why you believe the problem addressed in your project is important, describe the techniques you used to tackle the problem and the rationale behind your approaches clearly.

To encourage teamwork, each member in the same group will get the same score. But each group has the right to claim someone as a free rider. The score of the free rider will be lowered if sufficient evidence is provided.

Stages and Deadlines:

1. [Sept. 24th 11:59pm] Project proposal (non-graded): submit a 1-page proposal including the business problem and goal, the nature of the data mining task (e.g., classification, regression, clustering), possible data attributes, where to get the data, and detailed time table of your project.
2. [Nov. 1st 11:59pm] Project status report (non-graded): submit a 1-page project status report including what you have achieved so far, and what you plan to do next.
3. [Nov. 25th 11:59pm] Project report: submit your final 10-page report following the structure.

Report Structure:

1. Introduction

Describe the problem you are going to tackle. You may want to put your specific problem in a larger context and motivate the important of the problem addressed in your project.

2. Data Understanding

Indicate where you get your data (e.g., give a link to the web page from where you download your data) and describe your data. You may consider the following aspects: number of records; number of attributes and a brief description of their meanings, attribute type, range, mean, skewness; missing values; outliers; class imbalance.

3. Model Building

You should choose multiple data mining techniques to build models. You may dedicate a specific subsection to each data mining technique used. Each subsection should start by briefly summarizing the major ideas underlying the technique using your own words. For each model built, indicate the parameter values and describe the conclusions you can draw from it.

Some additional effort you can try to improve model performance: e.g., feature normalization, feature discretization, feature selection, parameter tuning. Provide the logical explanation of why you make such effort. You may present in more details any novel idea(s) you think interesting, which will bring your report to a higher level.

4. Performance Evaluation

Indicate the performance measures (e.g. accuracy, TPR, ROC, MAE) you have chosen to evaluate the performance of the models built. You should also indicate how the chosen performance measures were estimated (e.g. cross-validation, separate test set). You may want to summarize the performance of the built models, using the chosen performance measures, in a table. In this way, it is easy to compare the performance of different models.

5. Conclusion

Summarize the problem to be addressed and how the conclusions drawn from the built models help you to tackle the problem. List if any potential problems as future work.

6. References