ISOM3360 Spring 2020 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 can choose any business problem that you are interested in, and formalize it into a data mining task. Then, you need to get some data related to the task from your own sources or public sources. After that, you can apply some data mining algorithms to your data and evaluate the performance of your algorithms. Finally, you will submit a project report, together with your data and code. You are expected to use Python in the project.

Data sources:

There are several websites where you can find an interesting problem and obtain data.

Kaggle: https://www.kaggle.com/datasets an online data science community

Hong Kong public data: https://data.gov.hk/en/

Business Problem:

It does not have to be related to a business process. Any problem that has social and economic impact are encouraged. Keep in mind that ML is a predictive machine. Therefore, I expect you to build some predictive models. However, stock price prediction, horse racing prediction, soccer game result predictions are not recommended project ideas, because it is usually difficult to get useful data from public channel.

Data Requirement:

There is no general requirement for which data problem, or which data format you should work on. However, in order for the project to be substantial and manageable, we do have a data size requirement. **The number of examples should be greater than 5,000 and less than 500,000. The number of features should be greater than 20, before modeling.**

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:

0. [Mar 10th] Project announcement.
1. [Mar 17th 11:59pm] Group formation: form your own group using Canvas -> "People" (3-4 students in each group, and students form groups within their own section.)
2. [Mar 26th 11:59pm] Project Idea Report (non-graded): submit a 1-page proposal including the business problem and goal, possible data features, where to get the data, and detailed time table of your project.
3. [Apr 16th 11:59pm] Project Progress Report (non-graded): submit a 1-page project status report including what you have achieved so far, and what you plan to do next.
4. [May 14th 11:59pm] Project report: submit your project Python code, data and a final report.

Project Status Meeting:

I will have two project status meetings in the class time during the semester. Each team (can be either the whole team members or one team representative) will meet with me and describe project idea, and I will provide suggestions if necessary.

First meeting (Required to attend): Mar 31th

Second meeting (Required to attend): Apr 21st and Apr 23rd

Final submission

In the final submission, you need to submit your Python code, data and a final report. The report should follow the following structure. The length of report should be at least 5 pages but no more than 10 pages (Word, single column, single line space). You can include important figures in your report.

Report Structure:

1. Introduction

   Describe the problem you are going to tackle. Focus on problems that would be difficult to solve with traditional programming or simple heuristics. 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. 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.

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 (If any)