ISOM5270 Group Project Guideline

General Goal:

In this project, you will have an opportunity to 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 should submit a project report, together with your data. Python 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, feature engineering/selection, learning algorithm selection, hyper-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 and discourage free-riding behavior, everyone will need to submit a confidential rating of the other teammates based on their contributions (from 0 [very insignificant] to 4 [very significant]).

Stages and Deadlines:

1. [Apr 23 11:59pm] Group formation: form your own group using Canvas -> "People" (4-5 students in each group)
2. [Apr 30 11:59pm] Submit your project idea (your business problem and data source) for approval.
3. [Jun 1 11:59pm] Project report: submit your final 10-page report following the structure below.

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 importance of the problem addressed in your project.

2. Data Understanding and Preparation

   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; etc. You may also perform some necessary preprocessing steps (e.g., feature normalization, feature discretization) before sending the data to model.

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 selection, hyper-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, MSE) 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

7. Appendix

   Include any supplementary tables or figures you want to add.

Possible Project Ideas and Data Sources:

1. Kaggle.com

   There are many publicly available competitions and datasets on Kaggle.com. Below are some examples:

   - Predict fraudulent credit card transactions: https://www.kaggle.com/dalpozz/creditcardfraud
   - Predict product backorders based on historical data: https://www.kaggle.com/tiredgeek/predict-bo-trial
   - Predict stock market movement based on daily news headlines: https://www.kaggle.com/aaron7sun/stocknews
   - Predict house price using regression: https://www.kaggle.com/harlfoxem/housesalesprediction/data
   - Predict employee attribution: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

2. UCI Machine Learning Repository

   - Predict client subscription after direct marketing campaigns: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
   - Predict default of credit card clients: http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
   - Predict whether an image is an advertisement or not: http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements
   - Predict daily total orders: http://archive.ics.uci.edu/ml/datasets/Daily+Demand+Forecasting+Orders
   - Predict approval of farm ads: http://archive.ics.uci.edu/ml/datasets/Farm+Ads
   - Predict company bankruptcy: http://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

3. Other useful links:

- https://www.analyticsvidhya.com/blog/2016/10/17-ultimate-data-science-projects-to-boost-your-knowledge-and-skills/