**IST 707 Applied Machine Learning**

**Course Project Instruction**

1. Project Format

The objective of the project is to use the main skills taught in this class to solve a real data mining problem. Students can choose to work individually or pair up with another student.

1. Choose a data mining problem and data set

For this project, you must choose your own dataset. It can be one that you created yourself or found from other resources, such as the Kaggle competitions and the UCI repository (<http://archive.ics.uci.edu/ml/>).

Some rules/tips about choosing data sets:

1. Do not choose the data sets that we have analyzed in class, such as the Titanic data, the zoo data, etc.
2. It should not be a small or made-up dataset.For this semester, “small” is defined as fewer than 100 examples in the dataset.
3. Choose a data set that does not require excessive data preprocessing.
4. Experiment design

Define a problem on the dataset and describe it in terms of its real-world organizational or business application. The complexity level of the problem should be higher than homework assignments.

The problem may use one or more of the types of machine learning algorithms that we have studied this semester: Classification, Clustering and Association Rules, in an investigation of the solution to the problem.

This investigation must include some aspects of experimental comparison: depending on the problem, you may choose to experiment with different types of algorithms, e.g. different types of classifiers, and some experiments with tuning parameters of the algorithms. Alternatively, if possible, you may use more than one of the algorithms (Clustering + Classification, e.g.). If there are a larger number of attributes, you can try some type of feature selection to reduce the number of attributes. You may use summary statistics and visualization techniques to help you explain your findings.

No need to use all of them. But explanation and statistics are needed to justify your choice of algorithms and methodologies, such as feature selection or parameter tuning.

1. Project idea proposal

Submit and present a 1-2-pager to describe the data mining problem, the data set, and your initial strategies for data analysis. This document describes:

1. What real-world problem do you want to solve?
2. What data mining task(s) would you model it as, association rule mining,

clustering, and/or classification?

1. What data set do you plan to use? Provide a sample data table to include

instances and attributes. Provide important data summary that is relevant to your problem modeling. For example, for classification task, what is the target variable, how many classes, what do the class labels mean, how many examples in each class, etc.

1. How would you evaluate your data mining result?

You are encouraged to read each other’s proposals to learn from each other.

It’s OK to choose a data set that another student also chooses to use, as long as your work is independent from each other. This means you should propose a different research question to address.

Feedback will focus on whether the problem modeling is appropriate, whether the project complexity is appropriate (if not, suggestion for adjustment), and whether the initial data analysis strategy is reasonable.

1. Project progress report

You are expected to have finished all experiments and a draft project report. This round of feedback is used to help you fix major problems in your research design and result analysis.

1. Final project report

To complete this project, write a final report that conforms to general academic paper format. Make sure to cite relevant work appropriately.

Your report should be within 8 pages, 1 inch margin on all sides, and at least 12-point Arial or Times New Roman.

1. How you will be evaluated

* Research question(s) appropriate within context
* Data description provides context and a basic understanding of the data
* Try at least 3 different models and explain why they were/were not useful
* Visualizations convey the results in an easy-to-understand manner
* Interpretation of results are actionable (as opposed to just interesting)
* Problem is sufficiently complex
* In-class presentation presents findings in an easy-to-understand way (e.g., no data science lingo)

Speeding Violations: Most customers did not have any speeding violations, however, those who did not submit a claim had customers with more speeding violation frequencies.

* Removed NAs columns, 1939
* Ended with 8149 rows
* Removed Age, Race, and ID columns
* Ended with 16 columns
* Added column for rank
* Ended with 17 columns
* Good Rank = past accidents = 0
* OK Rank = past accidents = 1
* Bad Rank = past accidents > 1
* Created a new dataset for clustering and scaled data
* Includes columns
* DUIS
* Speeding Violations
* Past Accidents
* Annual Mileage

Add in that we changed data for factoring