Homework #3 **Due: turned in by Monday 02/19/2018 before class**

__Chengyao (Leo) Ma__

(put your name above)

Total grade: _____ out of ___100___ points

Please answer the following questions and submit your assignment as a single PDF file by uploading it to the HW3 drop-box on the course website.

Hands-on predictive modeling (100 points)

Download the dataset on spam vs. non-spam emails from http://archive.ics.uci.edu/ml/datasets/Spambase. Specifically, (i) file "*spambase.data*" contains the actual data, and (ii) files "*spambase.names*" and "*spambase.DOCUMENTATION*" contain the description of the data. This dataset has 4601 records, each record representing a different email message. Each record is described with 58 attributes (indicated in the aforementioned *.names* file): attributes 1-57 represent various content-based characteristics extracted from each email message (related to the frequency of certain words or certain punctuation symbols in a message as well as to the usage of capital letters in a message), and the last attribute represents the class label for each message (spam or non-spam).

Task: The general task for this assignment is to build *two separate models* for detecting spam messages (based on the email characteristics that are given) using RapidMiner or any other tool you prefer (e.g., Python, Spark, etc.):

1. [Start with this task] The best possible model that you can build in terms of the *overall predictive accuracy*;
2. The best cost-sensitive classification model that you can build in terms of the *average misclassification cost*.

Some specific instructions for your assignment/write-up:

- Reading the data into RapidMiner (or your software of choice): Data is in a comma-separated-values (CSV) format. You may also want to add the attribute names (which are in *spambase.names* file) to the data file as the first line. It might be easiest to read data into Excel, save it as Excel file, and then import it to RapidMiner.
- Exploration: Make sure to explore multiple classification techniques. You should definitely consider decision trees, *k*-nearest-neighbors, and Naïve Bayes, but you are free to experiment with any other classification techniques you know (for example, you can try applying some meta-modeling techniques, etc.).
    - Also, explore different configurations of each technique (for example, you should try varying some key parameters, such as values of *k* for *k*-NN, etc.) to find which configurations work best for this application. You can use parameter optimization approaches to help you with that.
- Make sure to explore the impact of various data pre-processing techniques, especially normalization and attribute selection.
- When building cost-sensitive prediction models, use 10:1 cost ratio for different misclassification errors. (I think it should be pretty clear which of the two errors –

classifying a non-spam message as spam vs. classifying a spam message as non-spam – is the costlier one in this scenario.)

- In general, use best practices when evaluating the models: evaluate on validation/test data, discuss the confusion matrix and some relevant performance metrics (not just the required accuracy and average misclassification cost, but also precision, recall, f-measure – especially for your best/final models…), show some visual indications of model performance (such as ROC curves).
- Finally, as a deliverable, produce a write-up (i.e., a single PDF file) describing your aforementioned explorations. Report the performances of different models that you tried (i.e., using different data mining techniques, different attribute selection techniques, etc.) What was the performance of the best model in the cost-unaware task (i.e., in terms of accuracy)? What was the performance of the best model in the cost-aware task (i.e., in terms of expected cost)? Discuss the best models in two different tasks (as well as their performance) in detail, provide some comparisons. Draw some conclusions from the assignment.

Evaluation: 100 points total.

- Performance: 35 points (based on the performance achieved by your best reported models).
- Exploration/write-up: 65 points (based on the comprehensiveness of your exploration, i.e., when searching for the best performing model, did you evaluate and report just one or two techniques, or did you try a number of different variations, based on what you know from the class?).