CARLSON
SCHOOL OF MANAGEMENT
UNIVERSITY OF MINNESOTA

PREDICTIVE ANALYTICS
MSBA 6420

Homework #3

Hao-Chun, Niu

(put your name above)

Total grade: ________ out of ___100___ points

There are 2 numbered questions. Please answer them all and submit your assignment as a single PDF file by uploading it to the course website.

Quation 1. (50 points) Use numeric prediction techniques to build a predictive model for the HW3.xlsx dataset. This dataset is provided on the course website and contains data about whether or not different consumers made a purchase in response to a test mailing of a certain catalog and, in case of a purchase, how much money each consumer spent. The data file has a brief description of all the attributes in a separate worksheet. Note that this dataset has two possible outcome variables: Purchase (0/1 value: whether or not the purchase was made) and Spending (numeric value: amount spent).

Your tasks:

(a) (20 points) Build numeric prediction models that predict Spending based on the other available customer information (obviously, not including the Purchase attribute among the inputs!). Use linear regression, k-NN, regression tree, SVM regreesion and Neural Network and ensembling models. Briefly discuss your explorations and present the best result (best predictive model) for each of these techniques. Compare the techniques; which of them provides the best predictive performance? Please make sure you use best practices for predictive modeling. (I.e., do you need to set which hyper-parameter? Normalize?)

(b) (20 points) As a variation on this exercise, create a separate "restricted" dataset (i.e., a subset of the original dataset), which includes only purchase records (i.e., where Purchase = 1). Build numeric prediction models to predict Spending for this restricted dataset. All the same requirements as for task (a) apply.

(c) (10 points) For each predictive modeling technique, discuss the predictive performance differences between the models built for task (a) vs. task (b): which models exhibit better predictive performance? Why do you think that is?

Quation 2. (50 points) Download the dataset on spam vs. non-spam emails from the following URL: 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 already 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 different models for detecting spam messages (based on the email characteristics that are given): (i) the best possible model that you can build in terms of the overall predictive accuracy (i.e., not taking any cost information into account), and (ii) 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:

- **Start working on the assignment early.**
- **Make sure to explore multiple classification techniques (we have learned quite a few of them in the class by now).**
  - **Also, make sure to explore different hyper-param for each technique (for example, try several different values of k for k-NN) to find which configurations work best for this application.**
- **Make sure to explore the impact of various data pre-processing techniques (e.g., normalization).**
- **When building cost-sensitive prediction models, use 10:1 cost ratio for different misclassification errors. (It should be pretty clear which of the two errors – false positive or false negative – is the costlier one in this scenario.)**
- **In general, use best practices when evaluating the models: nested CV, discuss the confusion matrix and some relevant performance metrics (accuracy, precision, recall, f-measure, AUC, average misclassification cost…), show some visual indications of model performance (ROC curves, lift charts).**
- **As a deliverable, produce a write-up describing your aforementioned explorations. Report the performances of different models that you tried. Discuss the best models in two different tasks in detail (which parameters worked best, what was the performance), provide some comparisons. (upload your code as well) Draw some conclusions from the assignment.**

- **Evaluation: 50 points:**
  - **Performance: 30 points (based on the performance achieved by your best reported models).**
  - **Exploration/write-up: 20 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?).**