CIS8695

**Homework II (Due: 2/4/2019, Midnight)**

**What to Turn In:**

Please write your answers and paste required results/reports in **a MS Word file**. **Note: Please do NOT submit any other file format, because it will cause grading inconveniency. Failing to submit in correct file format will cause the loss of homework grades! You can use screenshots if it is convenient for you.**

# Problem 1 (CART): Competitive auctions on eBay.com:

The file **eBayAuctions.csv** contains information on 1972 auctions that transacted on eBay.com during May-June in 2004. The goal is to use these data in order to build a model that will classify competitive auctions from non-competitive ones. A *competitive auction* is defined as an auction with at least 2 bids placed on the auctioned item. The data include variables that describe the auctioned item (auction category), the seller (his/her eBay rating) and the auction terms that the seller selected (auction duration, opening price, currency, day-of-week of auction close). In addition, we have the price that the auction closed at. The goal is to predict whether the auction will be competitive or not.

1. Note that in the dataset, the original variables of **Category** (11 categories), **Currency** (USD, nonUS), and **EndDay** (Weekend, Week) are categorical variables. Therefore, the dataset also contains their corresponding dummy variables.
2. Import the dataset and split the data into training and validation datasets using a 60%-40% ratio.
3. Fit a classification tree. Use ***Competitive*** as the output variable and the rest of variables as predictors. In the model, make sure that you exclude one dummy variable from each group of dummy variables (e.g. exclude Category\_SportingGoods, Currency\_nonUS and EndDay\_Weekend). To avoid overfitting, set the **maxdepth=6**.
   1. Report the tree (copy and paste the tree diagram).
   2. Report the prediction Confusion Matrix of Validation Data.
   3. What predictors are used by the tree?
   4. List the decision rules. For example, if variable1<0 AND variable2<2, class=0.
4. Are the rules practical for predicting the outcome of a new auction? Explain why (Hint: are you able to use the rules to classify a new auction before the auction ends? Do you know the values of all predictors in the rules before the auction ends? Some of them may not be known before the end of auction. What are them?). What variables should **NOT** be included in the predictor set? Explain why.
5. Fit another classification tree using the same setting in question 3. This time only use the predictors that can be used for predicting the outcome of a new auction.
6. Report the tree (copy and paste the tree diagram).
7. Report the prediction Confusion Matrix of Validation Data.
8. What predictors are used by the tree?
9. List the decision rules.

6. Examine and compare the summary reports in questions 3 and 5. Compare the overall performance (e.g., accuracy or error rates) between these two decision trees. Which model has better predictive performance? Explain why.

7. Build a random forest model for this prediction problem. Report:

1. Variable importance.
2. The prediction Confusion Matrix of Validation Data.

**Problem 2**. (Naïve Bayes Classifier).

Look at Problem 8.2 in Chapter 8 of the Textbook, page 202-203. (A picture of the problem is attached below). Complete (a), and (c) of Problem 8.2.

