Machine Learning Analytics RSM8413 – Fall 2022

Group Assignment 2 Due date: 2022-11-13

Deliverables (to be submitted on Quercus):
1. Report including a detailed description of your findings. At most 10 pages + appendices.
2. All Python source code either in a Jupyter Notebook (*.ipynb) or a Python file (*.py). One file!

Include an Executive Summary (at the beginning) describing your most salient findings. Explain all steps and results clearly and cogently, so that a reasonably intelligent though statistically naïve manager could understand it. You need to include all graphics in your report. Your narrative should be clear and concise, accompanied by supporting evidence in the form of graphics and tables. All tables and graphics should be well formatted (e.g., tables should not run over from one page to another).

The Case:
Competitive Auctions on eBay.com. The file eBayAuctions.xlsx contains information on 1972 auctions that transacted on eBay.com during May–June 2004. The goal is to use these data to build a model that will classify auctions as competitive or non-competitive. A *competitive auction* is defined as an auction with at least two bids placed on the item auctioned. The data include variables that describe the item (auction category), the seller (their 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 at which the auction closed. The task is to predict whether or not the auction will be competitive.

Tasks:

Data Preprocessing:
Create dummy variables for the categorical predictors. These include *Category* (18 categories), *Currency* (USD, GBP, Euro), *EndDay* (Monday–Sunday), and *Duration* (1, 3, 5, 7, or 10 days). Perform any other preprocessing steps that you might consider to be necessary and explain briefly.

Split the data into training and test datasets using a 60%:40% ratio.

Fit a classification tree using all predictors (if feasible). To avoid overfitting, set the minimum number of records in a terminal node to 50.
Write down the results in terms of rules.

Describe the interesting/unexpected and uninteresting (= rather obvious) information that these rules provide.

Suppose you had to slightly reduce the number of predictors due to software limitations, or for clarity of presentation, which would be (a) good variable(s) to choose?

Is this model practical for predicting the outcome of a new auction?

Fit another classification tree (with a minimum number of records per terminal node = 50), this time only with predictors that can be used for predicting the outcome of a new auction. Describe the resulting tree in terms of rules.

Plot the resulting tree as a tree diagram.

Plot the resulting tree on a scatter plot: Use the two axes for the two best (quantitative) predictors. Each auction will appear as a point, with coordinates corresponding to its values on those two predictors. Use different colors or symbols to separate competitive and non-competitive auctions. Draw lines (you can sketch these by hand) at the values that create splits.

Does this splitting seem reasonable with respect to the meaning of the two predictors? Does it seem to do a good job of separating the two classes?

Examine the classification table for the tree. What can you say about the predictive performance of this model?

Based on this last tree, what can you conclude from these data about the chances of an auction obtaining at least two bids and its relationship to the auction settings set by the seller (duration, opening price, ending day, currency)?

What would you recommend for a seller as the strategy that will most likely lead to a competitive auction?