CS 699
Assignment 4

Note: R is recommended since that is what is taught in class, but you may use any software you choose. As explained in submission guidelines on the next page, please provide your "final answers" in a document and provide a separate file containing your code or other detail that explains how you obtained your answers.

Caution: This assignment requests several training grids. Be careful what ranges you consider when forming the training grids. If your training grids don't cover wide enough ranges for each parameter, you may get suboptimal model performance. If your training grids cover too wide of a range, you may spend unnecessary compute time.

Problem 1 (60 points). Use the dataset *accidents1000.csv*. This data set has 999 observations and 11 features. The class feature is named MAX_SEV.

(1). Generate training and holdout partitions on the data set. Use 1/3 of the data in the holdout.

(2). Fit a discriminant analysis model. Use the discriminant analysis model to make predictions on the holdout data. Generate a confusion matrix and measure the model's performance using one or more appropriate metrics of your choice.

(3). Fit a neural network model, applying a training grid to tune the parameters. Use the neural network model to make predictions on the holdout data. Generate a confusion matrix and measure the model's performance using one or more appropriate metrics of your choice.

(4). Fit a random forest model, applying a training grid to tune the parameters. Use the random forest model to make predictions on the holdout data. Generate a confusion matrix and measure the model's performance using one or more appropriate metrics of your choice.

(5). Fit a support vector machine, applying a training grid to tune the parameters. Use the support vector machine to make predictions on the holdout data. Generate a confusion matrix and measure the model's performance using one or more appropriate metrics of your choice.

(6). Compare the four models. Which, if any, seems to perform better? Discuss.

Problem 2 (40 points). Use the dataset *restaurantdata.csv*. This data set has 8368 observations and 16 features. The numeric value to be predicted is named Revenue.

(1). Generate training and holdout partitions on the data set. Use 1/3 of the data in the holdout.

(2). Fit an XGBoost tree model, applying a training grid to tune the parameters. Use the XGBoost tree model to make predictions on the holdout data. Measure the tree's performance using one or more appropriate metrics of your choice.

(assignment continues on next page)

(3). Fit a random forest model, applying a training grid to tune the parameters. Use the random forest model to make predictions on the holdout data. Measure the forest's performance using one or more appropriate metrics of your choice.

(4.) Compare the XGBoost tree model to the random forest model. Which, if any, seems to perform better? Discuss.

Submission Guidelines:

- **Submit the solutions in a single Word or PDF document and upload it to Blackboard. You should separately upload the code or another document that explains how you arrived at the answers. Make sure each filename includes your name.**
    - **If you need more than 3-4 separate files, you might ZIP/RAR these files. Please still upload the Word/PDF separately. The separate Word/PDF greatly simplifies the facilitator's ability to comment on your assignment.**
- **Your submission should be organized well and easy to follow. Clearly list which question you are answering and provide the answer requested (one or more numbers, a screenshot, a plot, an explanation, whatever the question asks you to do). If a question contains multiple parts, you should clearly provide your answer for each part.**
- **If your facilitator tells you to submit the files differently than the above guidelines, you are expected to respect your facilitator's wishes starting on the next assignment.**
- **Facilitators can deduct up to 20% if you fail to follow these requirements (more if the questions are not actually answered).**
- **Facilitators can deduct 5% for each day the assignment is late.**
- **Unless your facilitator or the professor agrees, your assignment will not be graded if it is more than 3 days late (e.g., no credit will be given after Friday at 6 AM Boston time). The professor will usually ask the facilitator to make the decision but in rare cases (<1% of the time) has overridden a facilitator. Do not expect the professor to override in most cases.**