MATH 4322 - Introduction to Data Science and Machine Learning
Group Project Report - Fall 2022 December 2, 2022

Directions:

- Due December 2nd to be uploaded in BlackBoard before 11:59 pm.
- Needs to be at least 6 pages, no more than 10.
- On the cover page please put the name of all of your group members.
- Provide only the most critical pieces of code in this report, without including too much output and clogging up the report. Simply attach the full R source code to your submission in order for me to have access to the whole thing.
- For each section/subsection, make sure to specify the names of students who mostly worked on that part of the report. E.g. "2. Random Forest (Johnny Doey, Brenda Parker)".
- Your report should include introduction, methods, results of your methods, conclusion, and references (bibliography).
- See examples on Blackboard.
- This report is worth 75 points, points each for each sections mentioned above is in the descriptions below. You will also be graded on correctness of the models used, correct interpretation of the results, professional looking, and any typos. Each of these points are based on a 0 - 5 scale, total 20 possible points for this part.

Points	Reason
0	No work done
1	The group completed the task the result was inappropriate or inaccurate.
2	The group completed the task with some understanding.
3	The group completed the task partially correct.
4	The group completed the task fully correct.
5	The group completed the task fully correct with further and deeper understanding of the task.

- The following is what I am expecting to see in the report.

1. Introduction(10 points)
   - Short description of your data. Including your the inputs and the outputs of the data.
   - The question you are wanting to answer. Just one question with a response variable.
2. Methods (25 points)

- Use two models to answer your question. Give the reason why you are using these two models and the advantages and disadvantages of each of the models.
- For the two models provide the following:
  (a) Write down the model formula.
    * Linear or Logistic regression - you should know how a model equation looks like.
    * Tree-based models (decision trees, random forests) -just write it in the form of $response \sim x1 + x2 + \ldots$
    * Neural networks draw a picture of the layered structure (like we did in class, exam and HW), denoting all the nodes appropriately.
  (b) Give the thought process of your considerations while fitting the model. For example,
    * For linear/logistic regression - would you consider excluding some predictors from consideration, and why? Some predictors might not make sense from domain knowledge, or their p-value might be really large after fitting the model.
    * For single decision trees -make sure to conduct pruning and explain the reason for it, why is it done.
    * For neural networks, depending on your task -explain your choices for the of hidden layers, number of nodes per layer, activation functions
  (c) Randomly subdivide your full data set in 80% for training, and 20% for testing.
    * Proceed to train your model on the 80% training data, and then record its prediction error on left out 20% testing data.
    * Repeat this subdivision 10 times and provide the mean test prediction error of the model for all of the 10 iterations.
    * Makes use to use *set.seed*() in order for both models to work on the same 10 train/test data subdivisions.

3. Results (10 points)
   - This is the results based on your question you are answering.
   - If you wanted to predict, then answer your question using the training/testing errors.
   - If you wanted to make an inference, then fit your model on the full data set (no train/test).
     (a) Output the results, providing the most important model summaries and images that resulted from your model fitting.
       * For linear/logistic regression it would be the significance table that one gets from *summary*() output, with all coefficient estimates and significance values.
       * For single-tree models -picture of the tree.
       * For random forests - variable importance plot.
       * For neural networks -the progress of training/validation loss and metrics.

(b) Provide the interpretation of results and your conclusions as it pertains to the original overall question. E.g. if your question was similar to "which factors affect the response" or "which variables are significant" -provide the appropriate answer.

4. Conclusion (5 points)
   - Recap the results.
   - Which model was the best to use.
5. Bibliography (5 points)