Mid-Term Project

MSCS 6520 Business Analytics
Spring 2018
Assigned: March 5, 2018
Due: March 26, 2017 (by beginning of class)

For this project, you will:

1. Find a publicly available dataset suitable for machine learning. Kaggle, KD Nuggets, and the UCI Machine Learning Repository are all good places to find datasets.
2. Identify the dependent (output) variable that you are trying to predict and the independent (input) variables that you can use as predictors.
3. Perform exploratory data analysis to generate hypotheses about which variables are the best predictors. This will involve creating scatter plots, box plots, or jitter plots. Choose 10 predictors (features) that you think will perform well.
4. Perform one round of forward feature selection for a kNN Model. After this, discard the 5 worst-performing predictors (features). Continue the forward feature selection with the remaining 5 predictors until you have optimized your model.
5. Describe what you learned about the data from this process, including any cases where the result was different from your initial expectations.

This project should be the equivalent of two homework assignments in scope. There will be no presentations – just a short, written report documenting your work. I will not be grading you on the accuracy of your model. I will be grading you on how well you follow the process we developed in class (using the appropriate types of plots, splitting data into training and test sets, scaling the data, using forward feature selection, and evaluating your models. Mostly, I want to see that you can apply what we've learned in class on your own) as well as your ability to craft a narrative about the data (your hypothesis and what you learned).

Written Report

The written report should contain:

- Name, link, and brief description of the public dataset you used in your analysis (half of a page). Include the output of the `head()` command so that I can see the various columns and their types.
- Write up your questions, analyses, and interpretations; forming a coherent story. Describe the results of your feature engineering process in terms of which features were and were not predictive. Include plots of your real data and predicted data, demonstrating the improvement from specific features. (2 pages)
- Reflection (at least half of a page)
    - Were you successful at building a predictive model?
    - Did you face any challenges?
    - In this project, what skills did you employ? What skills do you think you can improve upon in the future? How might you go about improving those skills?

Total length: At least 3-4 pages, longer if necessary. The report should be single or 1.5 spaced. Footnotes or endnotes should be used for URLs and references instead of including them inline.

Submit the report as a PDF to D2L. You may work in pairs; in which case, you should only submit one PDF per group.