Assignment A03: Cross Validation
Due Date: 11:59pm on Saturday, July 25th, 2020

Problem Statement
You will use k-fold cross validation on the models you used in the past two labs in order to test how they are likely to perform with data they have never seen before. This assignment is a chance to practice cross-validation, but also to improve your models according to the feedback from myself regarding their strength and/or logic of your past models. If they weren't done very well, you have permission to recreate totally different models or better models on a different data set.

You have full permission of course to change your data set from the set of data you used in previous assignments. Please send me a one or two-line description of your proposed set and a screenshot of the excel file so I get a sense of what it is by Tuesday July 21st. In the final group project you will be using as many tools as possible to explore the same and/or related data sets, so I do recommend as much as is possible to stay with a similar topic as you have been working with to help you with the final project. However, it can also be beneficial to see and experiment with new data to see the various tools in different contexts so trying something new can be an advantage as well.

Expectations
• Choose the top 3* models (or so) you have created in past labs
• Evaluate their test errors with cross-validation techniques comparing with similar models of less and more complexity
• Use other approximations as well if available (cp, bic, aic) and comment on similarities or differences in the results these approximations give as compared to the CV process
• Comment on what you believe to be the strength of your model/models

Guidelines
Give a brief recap on your models and explain again why they are important. Make sure you take time to improve old ones or try something new if you did not score well on a previous lab. You will still be marked on the models and interpretations of those models make sense. You don't have to go as in depth this time around, but the models should still make sense, you should still use a sufficient # of data points (I suggest at least 100, but really at least 1000 is appropriate given all the data we can have access to), and so on.

The rest of the assignment is more technical than the others. We want to see that:

- you have executed cross-validation properly and understand the difference between test and training sets
- you understand how to perform k-fold cross validation properly and that you understand it is the preferred method of calculating realistic (although imperfect) test error rates
- that you may want to use other estimates of test error such as Cp, AIC, BIC, but that you realize these are not actual test errors, just training errors that have been modified to reflect what the true test errors likely are.
- You use and understand the one standard error rule to choose the "best" model.

Submission
1. You will create your solution in a tracked git repository.
2. The commit to be evaluated will include the message "Final Submission".
3. You will submit link to the solution git repository in D2L as submission to the assignment.

Grading Rubric

A03 – Cross Validation Grading Rubric				
	Very poor Value: 0	**Poor** Value: 0.5	**Average** Value: 1.5	**Good** Value: 2.5
Robustness of Models 25%	Only 1 or 2 categories on the far right completed at all or all three completed insufficiently well	1-3 categories to the right completed only reasonably well	At least 2 of the categories described to the right done quite well	1. Based on sufficient # of data points (at the very least 100, but ideally at least 500 or 1,000) 2. Interpreted well 3. Helped the reader understand the data better
Model performance vs. Model complexity 25%	The student hasn't illustrated that they understand the idea of complexity vs. performance	The student's performance vs. complexity graph contains significant logical or structural errors	The student has stopped short of creating too many different models and so, their performance vs. complexity graph has too few points and is of limited usefulness	The student has taken a model that shows promise and tried various combinations of possible predictors, and created a proper performance vs. complexity graph
Choosing the "best" model 25%	Little to no understanding of the 1SE rule	Understanding and use of the 1SE rule is weak	The student seems to have understood the 1SE rule, but the data looks doubtful	The student has understood and utilized the 1-standard-error rule properly
Methods 25%	Student has very little understanding of the purpose of CV or test error estimation	Student does the process mainly correctly but illustrates in their language they are unsure of what CV is or what Cp, AIC, BIC etc. are estimating	The student does not perform KFCV quite correctly. Somewhat confuses the idea of the test error estimate calculated with CV and those created by modification of training error rates	The student understands that KFCV is best and depends on this data the most. They realize Cp, AIC, BIC, etc. are estimates based on modifications of calculated training error and use these to confirm their KFCV