Lachlan Sinclair

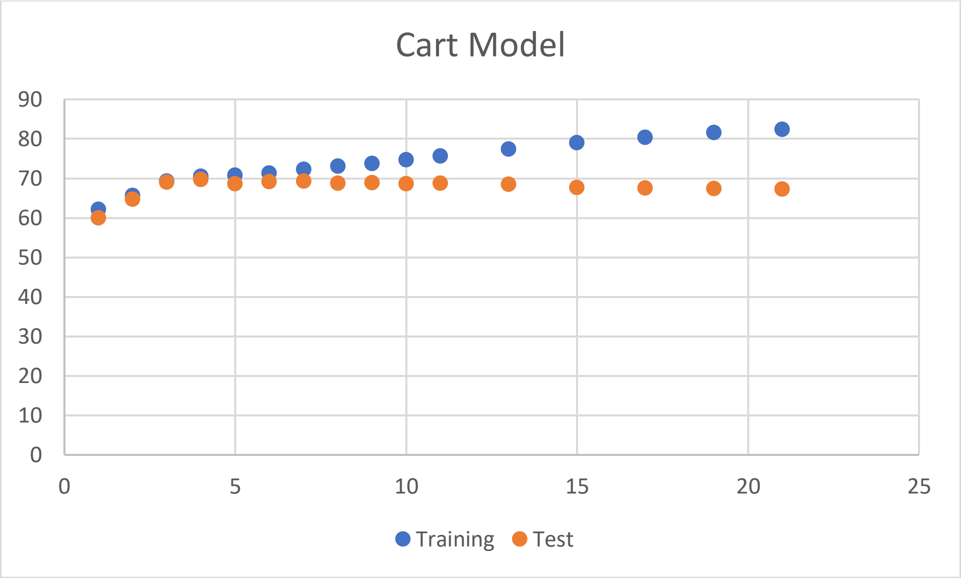
2/24/2020

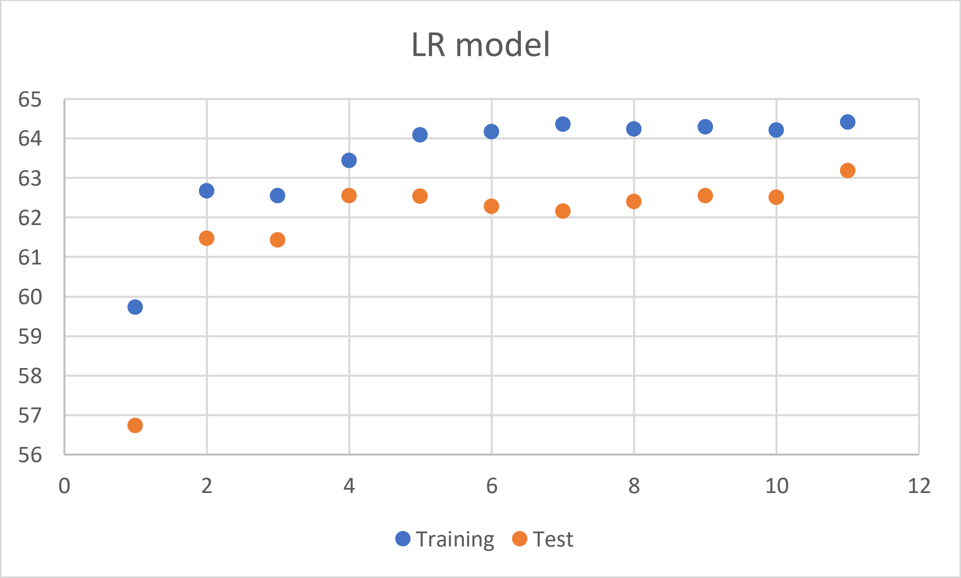
Week 7 Assignment 1

Reflections

a. The CART model's complexity is defined by the maximum tree depth in this assignment. More specifically the maximum tree depth is the property we use to control the complexity of the model. The tree depth determines how close the leaves can get to being pure given the configuration we are using. In the stopping rules we set the minimum records in parent branch and minimum records in child branch to 2 and 1 respectively which allows for leaves to be created that are entirely pure and contain a single record. If allowed to go deep enough the CART model will produce an entirely pure model with 100% accuracy on the training set. In the most general since the complexity of the model is still defined by the number of nodes in tree. For the logistic regression model, we used the attribute selection (inputs) to control the complexity as outlined in the textbook. The attribute inputs in the logistic node represent the number of variables the logistic regression model will use in its formula, the number of variables used is one way assess the complexity of the model. Therefor by controlling the number of inputs used we are directly controlling the complexity of the model.

b. I attached the graphs I made in excel to help me visualize the differences of the two models. Both models start off as expected when the complexity is very limited (underfitted/high bias), the model performs poorly on both the training set and the test set. One thing of note, in the underfitted section the performance of the CART model on the training set and test set produce significantly closer results to one another when compared to the LR models differences between its two subsets results. As the CART model progresses from underfitted towards overfitted there is the expected rise in both the training and test sets accuracies. Then after the complexity increases the overfitting (high variance) gradually begins to occur, the training sets performance continually increases while the test sets performance decreases. The LR model also exhibits the expected behavior up to the point in which overfitting begins, however after the "sweet spot" (naming convention from the textbook) the training performance slightly increases then sort of levels off. For the testing performance, after the sweet spot it decreases for a bit then begins to increase again. This differs from the CART models behavior and isn't what the generally expected behavior is, which is why I think is just an interesting trait of the specific dataset. It is possible though that the "sweet spot" will occur after the initial dip in test set performance.


Cart Model
90
80
70
60
50
40
30
20
10
0
0
5
10
15
20
25
● Training
● Test


LR model
65
64
63
62
61
60
59
58
57
56
0
2
4
6
8
10
12
● Training
● Test