Problem Set 9

Karley Nadolski

April 2021

1 Splitting the data

- What is the dimension of your training data?

  The original training data has dimensions of 405 x 14. The prepped training data has dimensions 405 x 75.

- How many more X variables do you have than in the original housing data?

  After using the recipe function to transform the data (adding interaction terms, converting to factors, creating square terms, etc) there are 60 more X variables (74 as compared to 14).

2 Regression Results

- LASSO regression model

  This regression had an optimal λ (penalty) value of 0.00356 and an out-of-sample RMSE of 0.22.

- Ridge regression model

  This regression had an optimal λ (penalty) value of 0.0233 and an out-of-sample RMSE of 0.218.

3 Questions

- Would you be able to estimate a simple linear regression model on a data set that had more columns than rows?

  No. In order to estimate linear regression, there needs to be at least as many data points as there are variables. When p (predictors) is greater than n (observations), there is no longer a unique least squares coefficient estimate and the variance would be infinite.

- Using the RMSE values of each of the tuned models in the two previous questions, where does the model stand in terms of the bias-variance trade-off?

  The bias-variance trade-off has to do with how the model interacts with the data used to shape it. A model with high bias relative to variance often is an oversimplification of the data (the model pays almost no attention to the training data), leading to high errors on training and testing data. A model with high variance, however, pays too much attention to training data and doesn't generalize on new data that the model hasn't seen before (suggesting overfitting). Models with high variance perform well on the training data but poorly on the test data.

  For both models, the out-of-sample RMSE is reported. The LASSO model's RMSE is 0.22 and the Ridge model's RMSE is 0.218. These are both pretty small error values. This suggests that the models are not overly biased. Without being able to compare the RMSE values from the in-sample data to those of the out-of-sample data, it's difficult to say how high the variance is for either model. Because of the small RMSE value of the testing data, I would guess that this model is well positioned on the bias-variance tradeoff because of the extra steps taken for regularization.