M4: Data Assignment

Jason Merten

Data Exploration/Preparation:

The data used in this analysis comes from two sets of data: a training dataset consisting of 2276 rows of data and test dataset consisting of 259 rows of data. These two datasets contain the same 17 columns of data, of which 6 columns contain missing values: **BATTING_SO**, **BASERUN_SB**, **BASERUN_CS**, **BATTING_HBP**, **PITCHING_SO**, and **FIELDING_DP**. Two of these columns (**BASERUN_CS** and **BATTING_HBP**) contain a large amount of missing data and were filled using a normal distribution of existing sample data. I decided to transform both datasets at the same time to reduce confusion in the code while simultaneously verifying the data after the transforms.

Building Models:

To assist with training the models, I split the training data into train_split and test_split data sets corresponding to 80% and 20% of the original data set. This will help the algorithms train the models based on known data and should provide a more accurate result than the champion model from Module 2.

I decided to create three separate models utilizing three different techniques. The first model was created using a best subset selection with cross-validation. The best MSE performance for this model was achieved by including interaction terms with the initial model.

The second model was created using the Lasso method. Again, including interaction terms yielded a strong result with the MSE.

The final model was manually created using Partial Least Squares. Again, I included interaction terms from the champion model from Module 2 to help improve the accuracy of the model.

Selecting a Model:

To select the most optimal model I primarily used the Mean Squared Error and r-squared values. Due to the way the models were build, I ended up creating two functions to compute the errors and output the results. Model 3 had the best performance out of all the models evaluated (including the previous champion model) based on lowest MSE and highest r-squared values.