Andrew Liu
7/27/20

The Effects of Weather on MLB Hitting

One category of variables that we considered using for our model was weather and wind data. This data was scraped directly from mlb.com and the four variables we scraped were the weather in degrees, the weather type, the wind speed in miles per hour, and the wind direction. The web scraping was done in BeautifulSoup4 and Selenium. We traversed through each day between the 2014-2019 seasons and scraped the box score, venue, and weather and wind data for each season game.

Data on the MLB website:

Weather: 76 degrees, Sunny.
Wind: 10 mph, Out To CF.

The rationale behind using these variables is that it might be easier or harder to get a hit during different weather and wind conditions. For example, a player may be more physically drained and thus have a more difficult time getting a hit if the weather is sunny and over 90 degrees. On the contrary, a player may have an easier time getting a hit if the wind is strong and blowing toward the outfield as it favors the batter.

In order to find out whether weather and wind data correlated with a player getting a hit, we matched the weather data with game logs and ran some statistical tests. For the continuous variables such as weather (degrees) and wind speed (mph), we first tried running logistic regression. However, there was almost no linear correlation for either variable or combined to be found with very small r^2 values close to 0.

Weather (Degrees):

Pseudo R-squ.: 0.001441

Wind Speed (MPH):

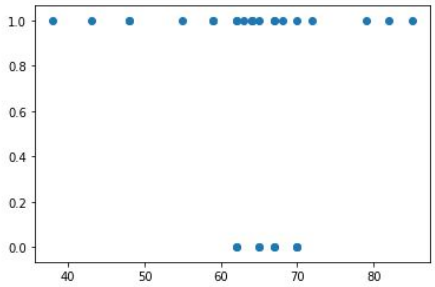
Pseudo R-squ.: 0.0005418

Combined:

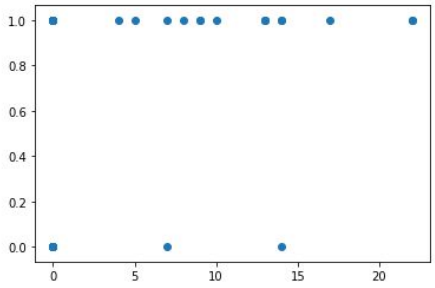
Pseudo R-squ.: 0.001849

Similarly, the scatterplots that we created did not provide much insight.

Weather (Degrees) scatterplot:



Wind (MPH) scatterplot:



We then looked at player-specific data where we looked at individual players to see if there was a larger correlation there since all players have different behaviors. We received similar results with very low pseudo-r^2 values when running logistic regression. Weather (Degrees) had a pseudo-r^2 value of only .02 while wind speed (MPH) had a slightly higher, but still small, pseudo-r^2 value of .14. Combined, they had a pseudo-r^2 value of .16.

Weather (Degrees):

Pseudo R-squ.: 0.02284

Wind Speed (MPH):

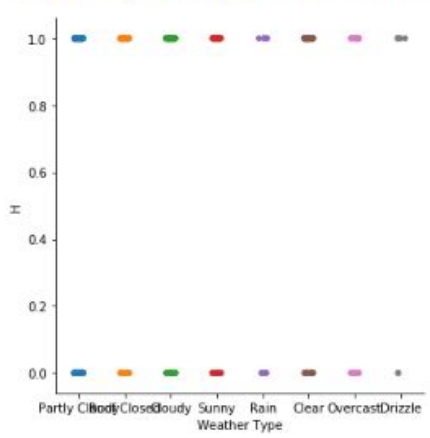
Pseudo R-squ.: 0.1424

Combined:

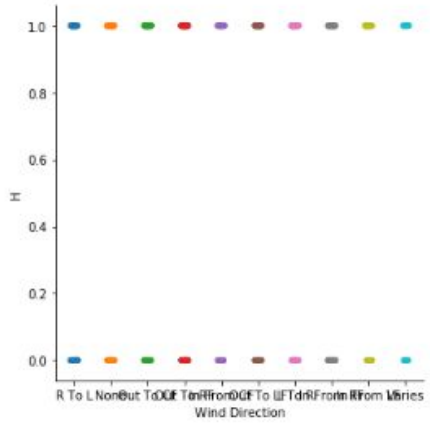
Pseudo R-squ.: 0.1583

When we created categorical plots for the categorical values of weather type and wind direction, we again received little insight regarding the general dataset.

Weather type categorical plot:

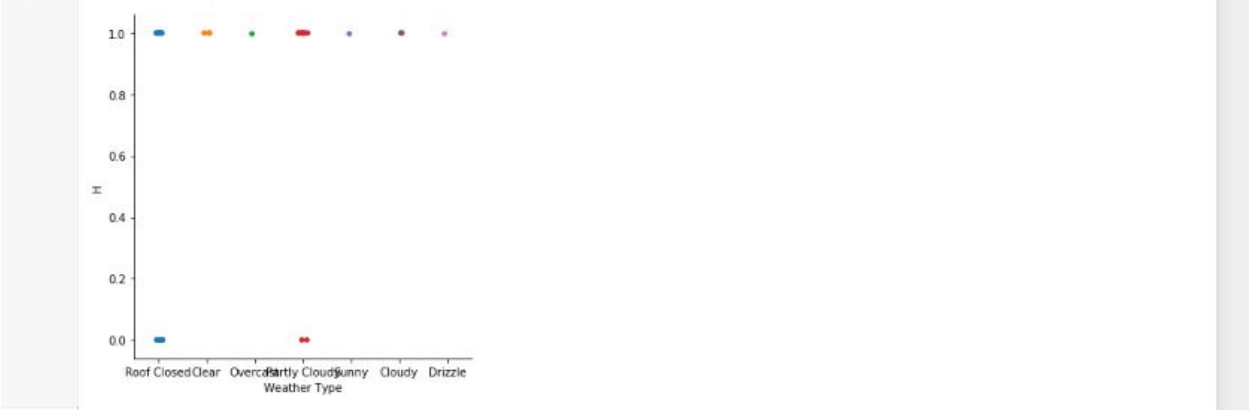


Wind direction categorical plot:



However, when we looked at the data for individual players, we saw more of a correlation between weather types and wind speed and whether the player got a hit, especially weather types. This indicated to us that weather and wind were more prominent factors on an individual basis. At a larger scale, there was significantly less significance and correlation as opposed to certain individual players, suggesting that some players are heavily influenced by weather and wind while others aren't.

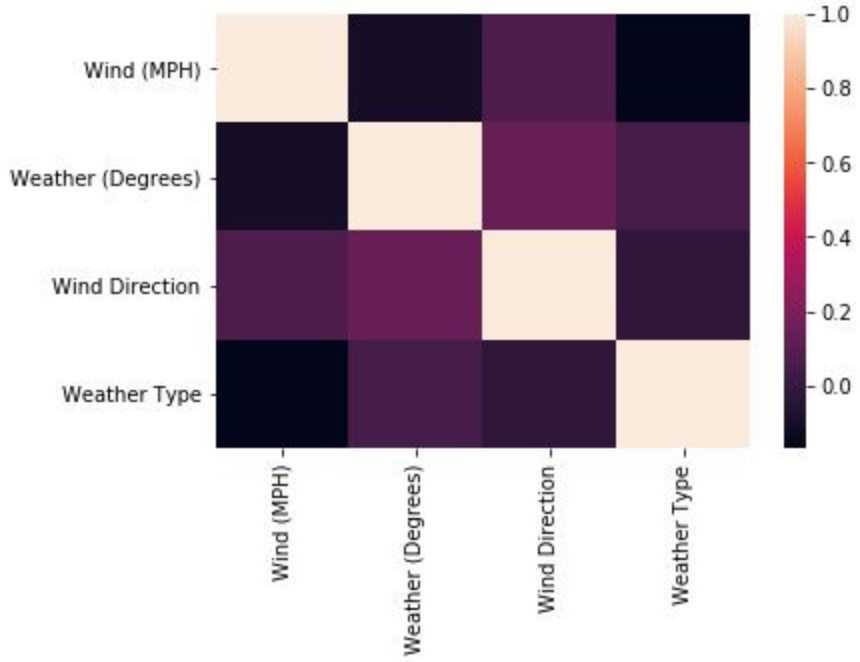
Individual player Weather Type categorical plot:



Individual player Wind Direction categorical plot:

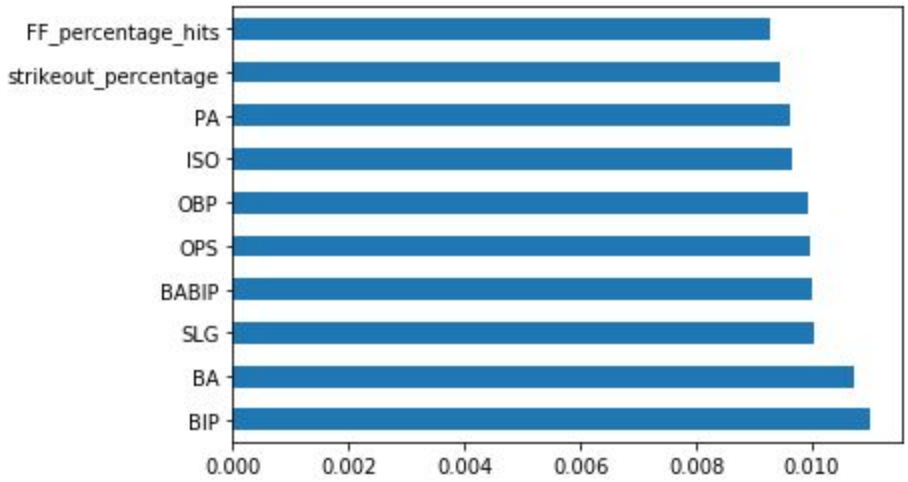
When looking at a heatmap of the correlation between features, we found that the wind variables and weather variables were very negatively correlated, which is most likely due to higher winds cooling down the temperature.

Heatmap of the correlation between wind and weather variables:

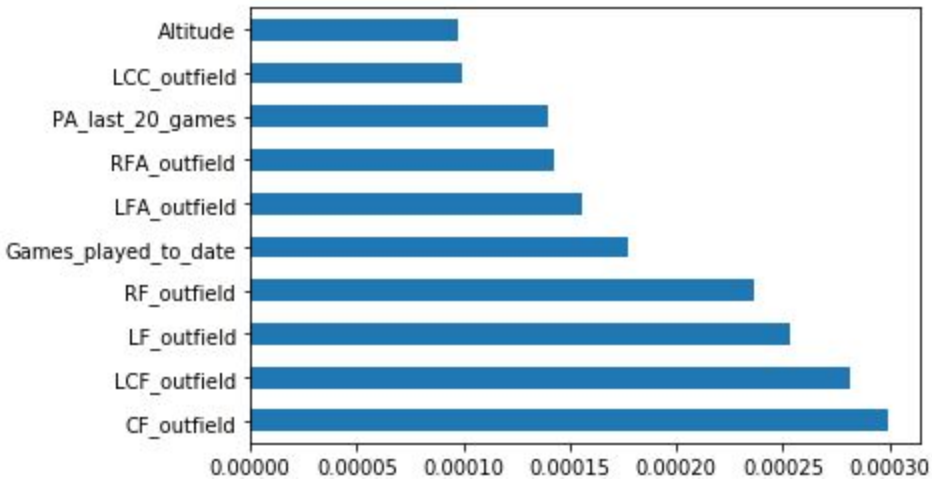


Lastly, we used the whole dataset to train a Random Forest model. Categorical variables were encoded. The weather and wind data weren't influential factors in our best performing models both as a whole and individually, adding very little value to the prediction. However, Weather (Degrees) did appear as one of the more influential features in Linear SVC, which was the least accurate model. Random Forest and Linear Regression performed the best but had high false positive rates while Linear SVC performed the worst but had low false positive rates. Overall, Random Forest was our best performing model and wind and weather data weren't particularly influential in that model.

Most influential features for Random Forest (our best performing model):



Most influential features for Logistic Regression:



Most influential features for Linear SVC: