Data Assignment 2 STAT 3341

Abstract:

Over the last decade, we have seen the advancement of data analytics in the sports world, but the MLB has incorporated this technology in many interesting ways. In this report, we will be analyzing three different models to predict base win probability using Statcast data: a logistic regression model, a penalized regression model, and a model using a decision tree. The logistic regression model allows us to estimate the probability of a binary outcome, in this case it is either the home team wins or loses. The penalized regression model allows us to regularize our models by adding a penalizing term into our model which will ultimately allow us to address multicollinearity issues. Furthermore, decision trees allow is to illustrate complex interactions and relationships within our data through a clear table. Each of these three models brings their own set of new information, and we are able to gain an even greater understanding of the small factors that increase and decrease winning percentage in baseball.

Introduction:

Baseball, often refered to as “America’s pastime,” is one of the most tradition-rich sports in the entire world. From Babe Ruth playing before World War II to Shohei Ohtani today, we have seen so many stars come through Major League Baseball, but aside from the immense amount of talent that has come through the league, it is also the first sport that utilized advanced data in the context of the game. With the growth in talent, we have also seen a huge growth in the use of technology and data analytics in the sport. What started as simply tracking outs, strikes, balls, RBI, and battling percentage, MLB teams have upgraded to using advanced camera and radar systems to precisely gauge different statistics within an MLB game, and this has allowed he league to track ball movement, release speed, spin rate and direction, and so many more metrics using Statcast. Statcast is a new way to track data analytics through the collection of statistics from MLB games.

Data Overview:

A high-level overview of the data gave us a lot of very useful information. The data is a fairly large dataset, as the full set has over 65,000 rows, and over 90 columns. By looking at the data, we are able to determine that 898 unique baseball games were accounted for in the data. We are able to see that the three most common events that occurred at every single at bat included over 25,000 cases of a field out, over 15,000 cases of a strikeout, and over 9,000 cases of a strikeout. When we fit the models, we are able to see that pitch acceleration is a key factor to the probability of the home team winning the game.

Analysis Task 1:

After performing the first analysis task, we are able to see that there were a total of 898 MLB games that were accounted for in this Statcast data ranging from 07/23/2020 to 09/27/2020. Next, through thr numeric summery of at-bats per game, we are able to see that the mean number of at bats per game is 38.11 with a median of 38.00. Lastly, after analyzing the different events that can occur at each at-bat, we found the three most common results. The most common result of an at-bat is a field out with over 25,000 occurrences. The second most common result is a strikeout with over 15,000 occurrences, and the third most common result is a single with over 9,000 occurences.

Analysis Task 2:

In our second analysis task, we built a base win probability model using logistic regression. It works by creating a linear relationship between the dependent variable and one or more independent variables. It is used to estimate the probability of a certain class or event occurring based on the values of the independent variables. When we do so, we found that the predictors which were statistically significant include inning, ay, post\_away\_score, post\_home\_score, and inning\_topbot\_Top. ‘Ay’ and ‘Post\_home\_score’ both increase the odds of the home team winning, and inning, sz\_bot, post\_away\_score, and inning\_topbot\_Top were the factors that decreased the odds of the home team winning.

Analysis Task 3:

In our third analysis task, we are builsing another base win probability model using penalized logistic regression instead. Penalized logistic regression is a variation of the logistic regression model in which a penalty term is added to the cost function. This penalty term reduces the complexity of the model, which helps to reduce the risk of overfitting. Furthermore, penalized logistic regression models are more robust to outliers than standard logistic regression models.When we run the code, we find that the optimal tuning parameter is Model 4 with a penalty of 0.00000830 and a mixture of 0.791.

Analysis Task 4:

In Analysis Task 4, we built a third base win probability model using a decision tree. To begin, I chose to create a grid Latin hypercube of size 10 as my tuning approach in this task. After this step, we were able to tune the model using the optimal tuning parameters that we found.

Analysis Task 5:

In Analysis Task 5, I am building several plots using the models that we have built in the previous three tasks.

Results:

The simple logistic regression model held up very well against the other two models. The simple logistic regression model has an accuracy of 75%, a kappa score of 50%, a sens score of 86%, a spec score of 65%, and a ROC AUC of over 86%. The penalized linear regression model that we fit boasted scores very similar to those of our simple linear regression model. All of these numbers were very comparable to the penalized regression model which had an accuracy of 75%, a kappa score of 50%, a sens score of 86%, a spec score of 65%, and a ROC AUC of 86%. On the other hand, the decision tree based model that we constructed has an accuracy of 74%, a kappa score of 47%, a sens of 85%, a spec of 63%, and an ROC AUC of 84%. This shows us that the model that we built using the decision tree based model does not do as well of a job of predicting base home win probability as the other simple and penalized linear regression models.

Discussion/Conlcusion:

In this assignment, we began by looking at Statcast data from MLB games from July of 2020 to September of 2020. We ventured into model building using a variety of different model building techniques in order to find the one that works best. We used a simple logistic regression model which is crucial to binary classification of data. Furthermore, we used a penalized linear regression model by adding a penalty term to the original model in order to reduce the amount of larger coefficients that we get in our data. We then used several different data visualization techniques to create plots of the Statcast data that we found. The use of these models shows us how we can use the data to predict the odds that the home team would win based on a certain variable. The research that we have done shows just how powerful data-driven insights are in the world of sports, and it shows just how we can use statistics and data visualization to find causes and effects of certain variables on win probabilities.