NHL Win Prediction Model Visualization

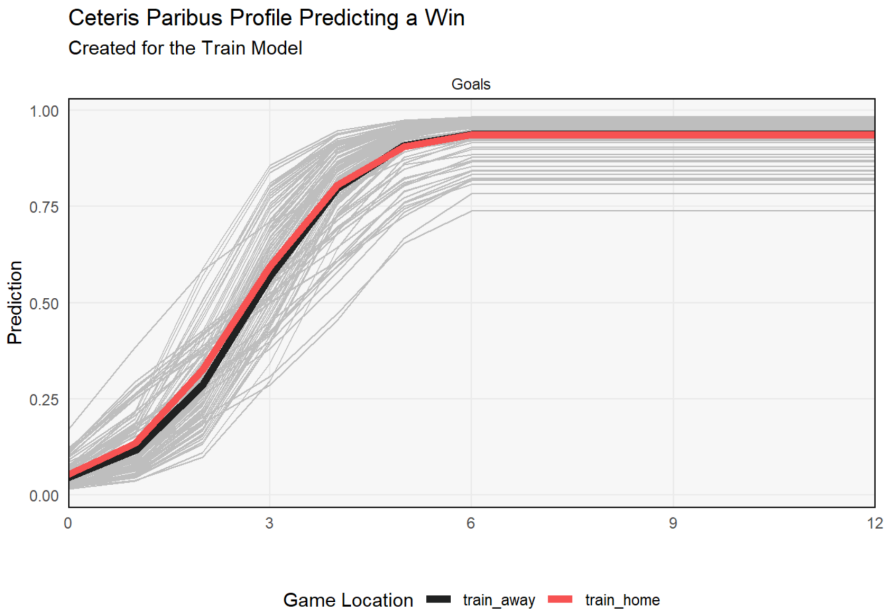
Maya Barron

R Script

Data Preparation:

I manipulated the data quite a lot before creating the interface. Since this data runs up to Covid times, I wanted to get rid of any games that weren't played so they didn't cause confusion on the plots. I kept all of my previous data manipulation from homework 2, although I didn't make use of all of it in this project. The team names appear in a cleaner format, which would make it easier if further analysis were to be done. I also removed a few columns from the dataset that I felt weren't helpful for what I was trying to visualize. Next, I transformed all of the character type columns into factor types so I could run train() using the "gbm" method. Finally, I dropped all of the duplicate rows that appeared in the dataset. At the end I was left with 47,432 rows, or data from 23,716 games. Full variable descriptions can be found inside of the R Script file.

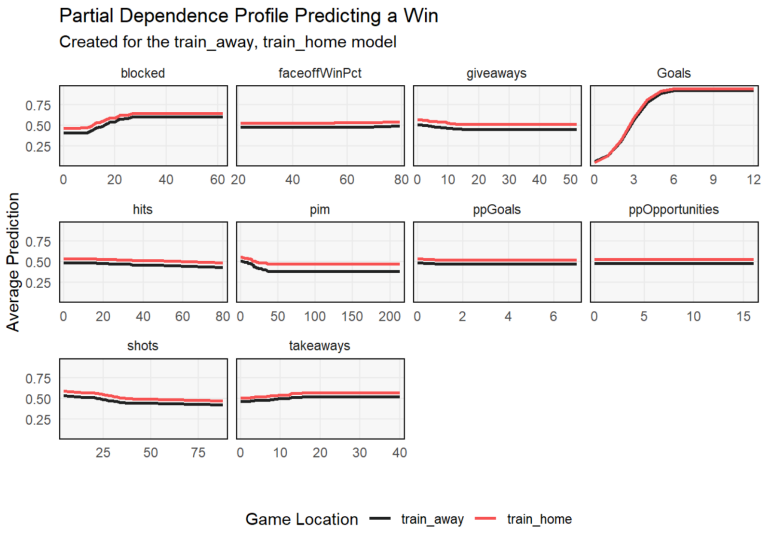
Visualization 1:

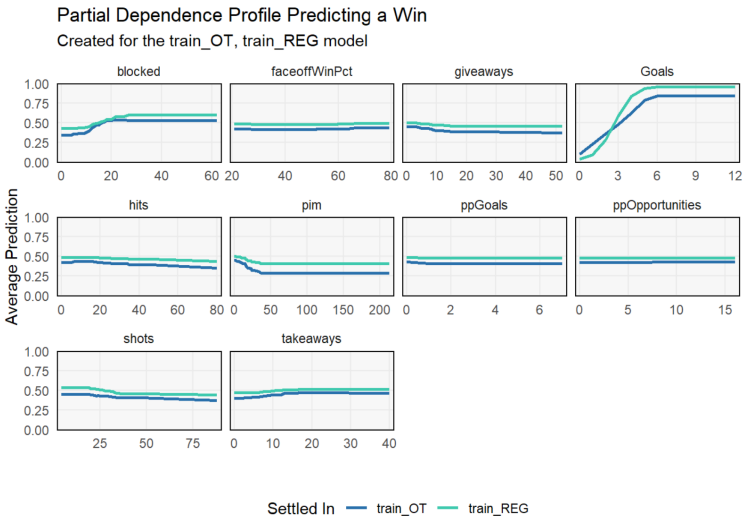


This visualization is meant to inform viewers about the likelihood of an NHL team winning a hockey game based solely on the team's goal count. The prediction was split by whether the game was played at home or on the road, in an effort to see if game location had any bearing on the model prediction. The colors I have chosen reflect the fact that home teams in the NHL almost always wear their colorful jerseys, while the away teams wear their white jerseys. The only trade-off type decision I had to make was selecting which variable to use to group and color the plot. My key finding is that there isn't much of a difference between how many goals the model predicts are necessary to win between games played at home and on the road. Overall, it looks like the model predicts that a team will win over 50% of the time if they score at least 3 goals in a game. The other key finding is that the model doesn't increase its prediction for a team winning past the 6 goal mark. I made this visualization by first creating a fit variable by training my entire dataset to predict the outcome of a game. I then used this fit variable to create my explanation variable

which I used to create the profile. I grouped the profile by the “HoA” column so I could split the prediction by whether the game was played at home or on the road. Finally, I plotted the profile with the arguments ‘geom = “profiles”’ and ‘variables = “Goals”’ to create a CP profile visualizing the impact of the number of goals on the game outcome, grouped by game location.

Visualizations 2 & 3:





These visualizations are meant to inform viewers about the likelihood of an NHL team winning a hockey game based on many game statistics taken into account individually. The graph on the left is split by game location (that’s why the bars are the same colors as the Ceteris Paribus Profile above), and the graph on the right is split by whether or not overtime was required. The only trade-off I had to make was choosing not to include some of the columns I considered to be irrelevant or that I thought would overcrowd the plot. The information in these columns included team name, head coach, and rink start side. The rink start side is irrelevant because the home team almost always starts on the left, so the information provided in this column should be directly linked to the information found in the HoA column. My key finding, as expected, is that the number of goals scored plays the biggest role in predicting whether a team will win or lose a game in the NHL. Blocked shots seem to have the second biggest role in predicting the outcome. It’s also interesting to see that as a team has more shots on goal, the model predicts that they have a lower chance of winning the game. This goes against intuitive logic. I created this plot by following the same steps as the Ceteris Paribus profile above, but instead of calling plot() with the argument ‘geom = “profiles”’, I used the argument ‘geom = “aggregates”’ to create the partial dependence profiles. For the plot on the right I also had to create a new profile before plotting where I grouped by the column “settled_in” instead of “HoA”.