Maxwell Snodgrass

An Analysis of Goal-Scoring Variables in Professional Soccer

1. Introduction

Expected goal models are a tool that has gained popularity in sports analytics. These models aim to calculate the probability that a shot will go in based on the characteristics of the shot. An accurate expected goal model will be able to reveal which teams over perform or under perform during a particular season. These models also apply to individual player performance. Therefore, each player will have their own quantity of expected goals for the entire season and their ability to score goals will decide if they are able to outperform their expected goals. Outperforming expected goals means that the player should aim to maximize their ability to convert shots to goals. However, this is a difficult task for players to achieve since there are many variables that play into how effective of a scorer can be. This research will attempt to find what the most significant variables are for maximizing goals/shot. When these variables become more apparent, teams can focus on maximizing them in order to turn more goal scoring opportunities into actual goals.

2. Data Collection and Cleaning Process

a. Data Collection

Two datasets were used in this project. The first dataset is the player statistics from the Bundesliga 2018-2019 season. The data was collected from fbref.com which is a soccer statistics website that partners with Opta, an organization that owns a high-quality expected goal model.

The second dataset that was used was the player ratings from FIFA 18. This game provides ratings on a plethora of different aspects of soccer based on the past performance of the player.

b. Data Cleaning and Transformation

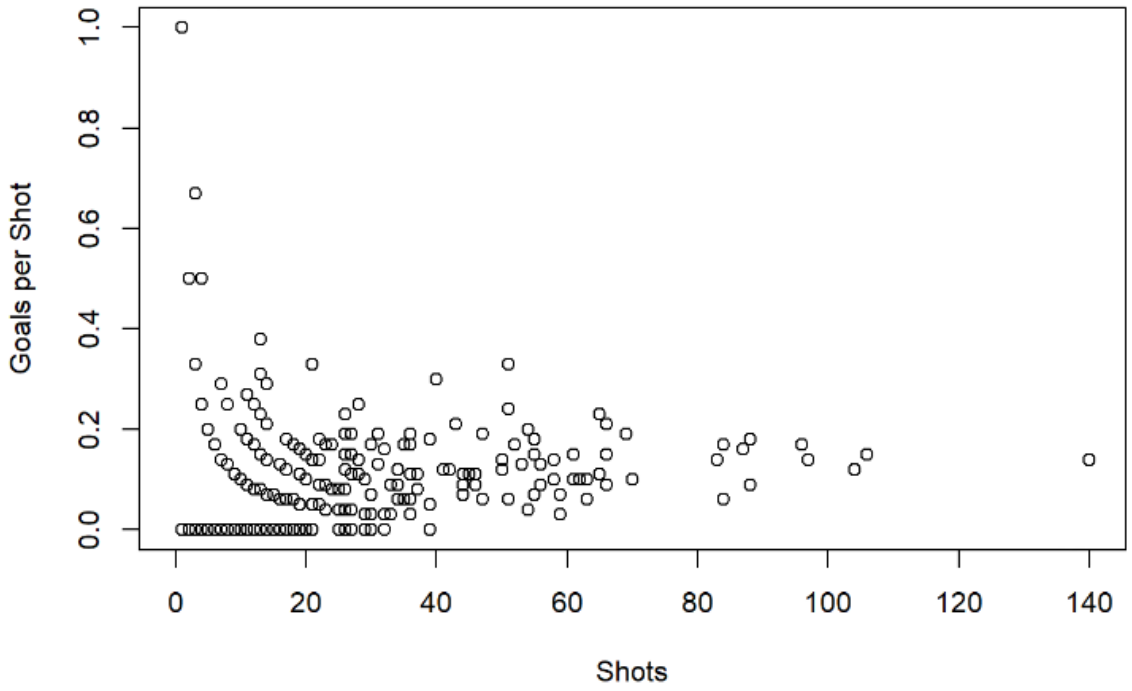
The first step in the data cleaning process was to join the Bundesliga player stats dataset together with the FIFA 18 player ratings dataset. It is worth noting that the FIFA 18 player ratings dataset bases its skill ratings off the seasons before the 2018-19 season. This is important because it will allow for a better prediction since the player's 2018-2019 performance will not be factored into skill rating, eliminating multicollinearity. To join the datasets together, I used SQL and its union function to join the two datasets together based on the name of the player. Then, I eliminated all players who did not play in the 2018-2019 Bundesliga season.

After that was completed, the position variable had to be transformed into a numerical value so that it can be used in the final regression. The position scale (PosScale) variable is measured on a scale of 1-5, with 1 being a defensive player and 5 being an offensive player. This was measured simply by transforming the position (Pos) variable from the player dataset to a 1-5 scale. Defenders were assigned 1, midfielders 3, and forwards 5. Players with a mixed position, like defense and midfield or midfield and forward, were assigned 2 and 4 respectively.

Finally, players with both the defender and forward position were assigned 3.

The last step of cleaning the data includes removing the observations that do not qualify for the regression. There are two components to this process. Firstly, I removed all players who did not have ratings in the FIFA 18 dataset. Most of these players had a low age which indicates they were new to professional soccer in this season. Since they do not have a history of playing in professional leagues, they cannot be included in the final regression since they have no available ratings. The second part of the process is creating a minimum cutoff for players in the dataset. The variable I chose to represent an accurate cutoff was total shots. This is because there are a large amount of players who had very little playing time, and may have had a small amount of shots. This is problematic for the data since a small amount of shots means that a single goal can dramatically increase the goals/shot variable. In figure 1, it can be seen that a convergence to 1 and 0 occurs below 20 shots. Therefore, the minimum cutoff for total number of shots a player must have to be qualified for the dataset is 20.

Figure 1: Total Shots vs. Goals per Shot



3. Describing the Dataset

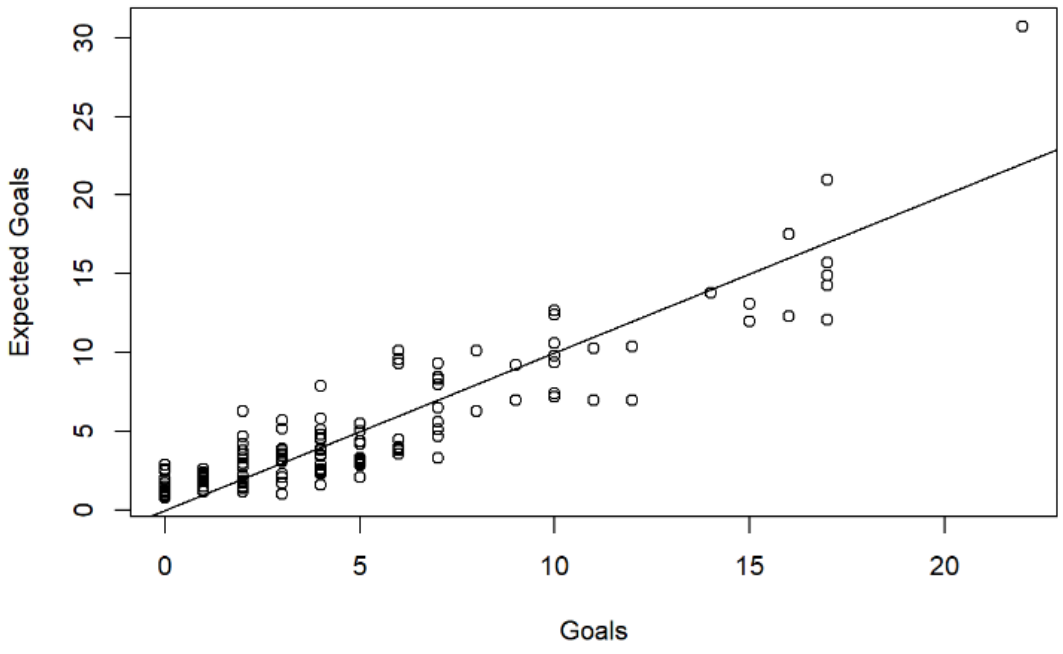
a. Player Expected Goal Models

As mentioned above, expected goal models are a measure of how many goals a player should score during the entirety of their season. These models are able to identify players who underperformed or overperformed when it comes to converting shots into goals. A player with more actual goals than expected goals means they were able to convert shots to goals more often than the average player. On the other hand, a player with less actual goals than expected goals means they were not able to score as much than what would be expected for an average player. The first part of this research will focus on comparing the players' expected goals to the players' actual goals to identify if the players as a whole over performed or under performed. The expected goal model used

for this is from Opta, which is an expected goal model that includes variables like location of shooter, body part of shooter, type of pass, type of attack, shot angle, clarity of the shooter's path to the goal, position of goalkeeper, pressure from defense, and more.

Figure 2 is a scatterplot that represents actual goals vs expected goals provided by the Opta model. The line going through the middle is not a regression line, rather a simple y=x which represents a world where the expected goal model accurately predicted the amount of goals every player scored. Therefore, the higher the data points are, the more the player overperformed, and the lower the points are, the more the player underperformed.

Figure 2: Total Goals vs Expected Goals



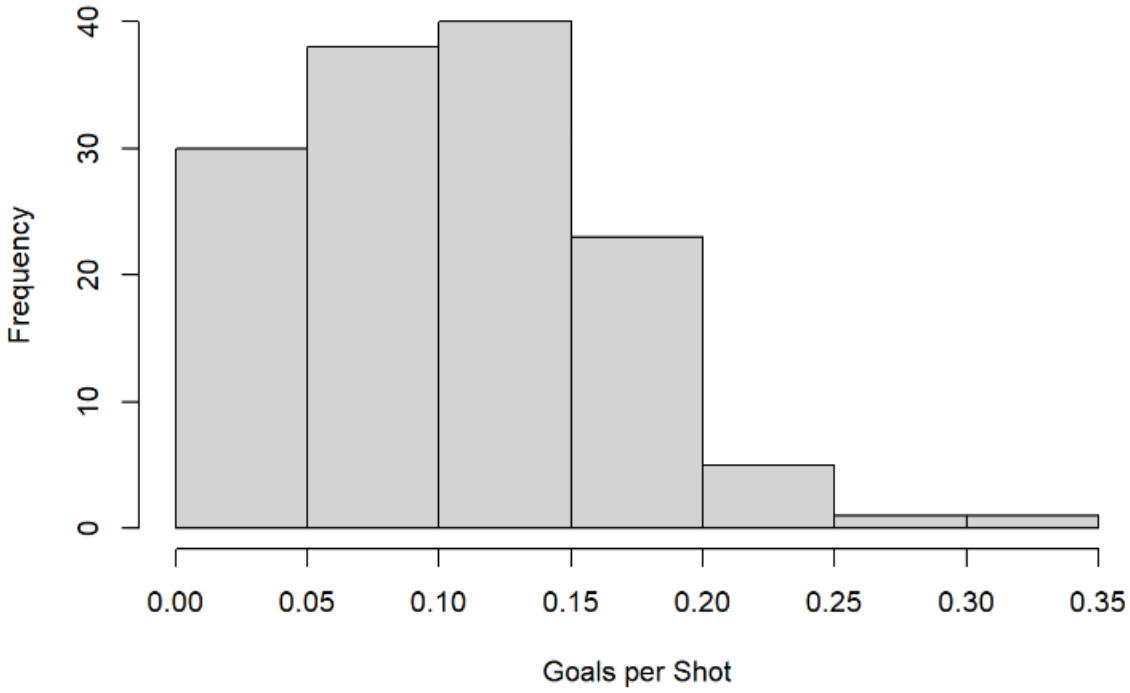
An interesting insight to be taken is that it is easier to underperform than overperform based off this model. This could be because of multiple

psychological and physical factors like nerves, physical fitness of player, focus, and other. These are all variables the model does not account for. To explain what is the most driving factor into converting shots into goals, a regression must be run that includes individual player data.

b. Analysis of Response Variable

The response variable for the final regression will be goals/shot (G.Sh). The reason for choosing this variable as the response is because it is the most accurate variable available for explaining how well a player can convert shots into goals. It is important to measure goals per shot and not total goals since this gets rid of the confounding variable of how many minutes a player has played during the season. A player who has played more minutes will naturally have more goals, therefore goals/shot measures accuracy and how well a player can convert shots to goals alone, and is not dependent on how many minutes an individual plays. The mean amount of goals/shot scored by the players in the new and cleaned dataset was .1075. The standard deviation was .0646. The histogram in figure 3 reveals that goals/shot is right skewed.

Figure 3: Histogram of Goals/Shot

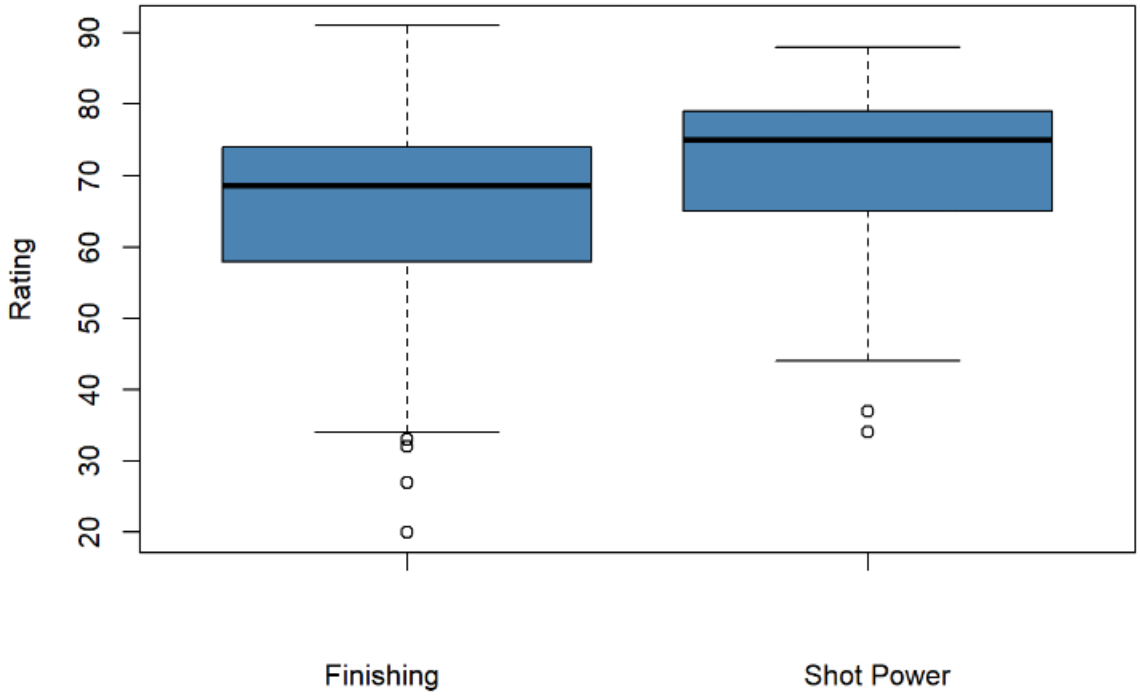


c. Analysis of Predictor Variables

The predictor variables that will be used in the final regression include the player's overall finishing rating (Finishing), shot power (Power), shot on target % (SoT.), average shot distance (Dist), position scale (PosScale), age (Age), and penalty kick attempts (PKatt). The regression will reveal which of these variables are significant in predicting goals/shot. The first two variables that should be compared to each other are the two variables that were taken from the FIFA18 dataset: finishing and shot power. A side-by-side boxplot (figure 4) shows that shot power is on average higher than finishing. Also, the variability of shot power is less.

This makes sense because at a professional level, it is expected that shot power among all players will be similar and at a high level. Finishing has more variance since there are more skills that factor into it than sheer strength.

Figure 4: Side by Side Box Plots of FIFA 18 Ratings



The next two variables to compare is shot distance and position scale. The scatterplot in figure 5 shows that players who are more offensive tend to shoot at a closer distance as compared to midfielders. Defenders seem to have a more random shot distance and do not have apparent clusters like midfielders and forwards. But it is also worth noting that there are more midfielders and forwards included in the dataset than

defenders, since many defenders most likely did not make the cutoff of having 20 shots. And this can be seen in the bar graph in figure 6.

Figure 5: Position Scale vs. Average Shot Distance

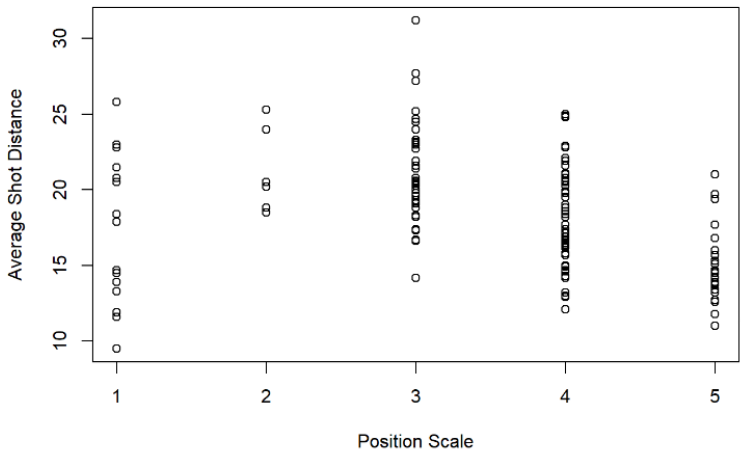
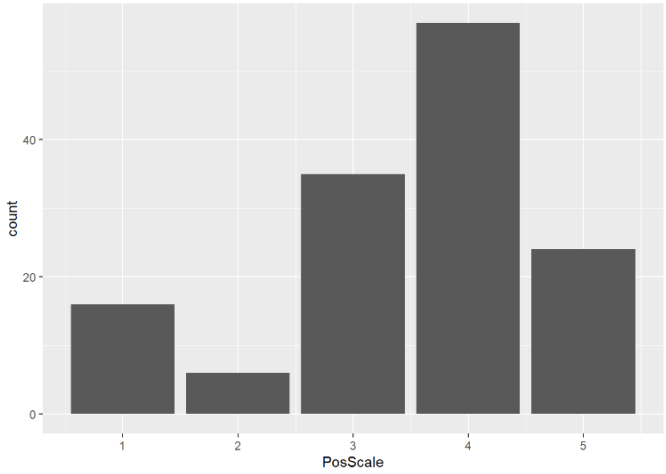


Figure 6: Bar Graph of Position Scale



The correlation matrix in figure 7 shows that there are no strong correlations between any of the variables, but there are some moderate correlations. The variables with a positive moderate correlation include finishing rating and position scale, finishing rating and shot power rating, and shot power rating and age. The correlation between finishing and shot power makes sense because shot power is one of the factors that determines overall finishing rating. Also, finishing rating and position also makes sense because players who are offensive tend to be better finishers since that is what they practice more.

Figure 7: Correlation Matrix of Predictor Variables

4. Analyzing Goal-Scoring Variables through Regression

The following regressions will be run to see what variables are significant in predicting the amount of goals/shot a player scores in the Bundesliga. The first regression will be comprised of all variables. Then, backwards elimination based on AIC and partial F-tests will be used to eliminate variables that do not contribute significantly to the response variable. After that, confidence and prediction intervals will be created for certain variables. And finally, residuals will be calculated to see the variability of the data.

Firstly, a full regression model was computed using all of the variables. At first glance, it seems that average shot distance and shot on target percentage are the two most important variables in predicting goals/shot. This was because they were the only two variables with significant p-values in the regression. Therefore, the model must be reduced to eliminate insignificant variables. The model selection process involved two techniques. The first technique was backwards elimination using AIC. This process revealed that finishing rating, shot power rating, age, and PKatt were insignificant and should not be included in the regression. The variables that should be included are average shot distance, shot on target percentage, and position scale.

The second technique that was used was a partial F-test to verify that position scale should be used in the final regression. This is because this variable is noticeably the weakest of the three. The full model in this test is:

$$G.Sh = \beta_0 + \beta_1 * Dist + \beta_2 * SoT. + \beta_3 * PosScale$$

The reduced model in this test is:

$$G.Sh = \beta_0 + \beta_1 * Dist + \beta_2 * SoT.$$

The null and alternative hypotheses are as follows:

$$H_0: \beta_3 = 0 \ \ vs. \ \ H_A: \beta_3 \neq 0$$

Using the ANOVA tables for both, the F statistic was calculated:

$$F = \frac{.291962 - .28288}{\frac{.2802}{134}} = 4.34$$

The critical value for this at the 95% level was 3.91 with degrees of freedom of 1 and 134. Since the F statistic is greater than the critical value, we can reject the null hypothesis and conclude that position scale contributes

significantly to the model that already includes average shot distance and shots on target %.

With the conclusion of these tests, the final model for predicting goals/shot is:

$$G.Sh = .06966 + .00294 * SoT. - .00507 * Dist + .00733 * PosScale$$

This model can be interpreted using the three variables. For every 1 percentage increase in a player's shots on target %, their goals/shot proportion increases by .00294. For every one-yard increase in average shot distance, the goals/shot proportion decreases by .00507. And for every 1 unit increase in position scale, the goals/shot proportion increases by .00733. Therefore, in order for a player to maximize their goals/shot, they should improve their accuracy, take shots from closer distances to the goal, and play an offensive position like a forward.

Next, I created 95% confidence and prediction intervals for the average player in the dataset. I did this by using the mean value of the three predictor variables, which were 3.486 for position scale, 18.25 yards for average shot distance, and 35.68% for shots on target percentage. The 95% confidence interval revealed we have 95% confidence that the average player in the dataset will have a goals/shot proportion between .09983 and .11523. The prediction interval revealed that we are 95% confident that the true goals/shot proportion of an average player is between .01676 and .19830.

The final step of the regression analysis is to create a residual plot that explains the variability of the data. The residual histogram and plot below in figures 8 and 9 indicate that there is normality in the data. A jackknife residual

plot was also created, and a Shapiro-Wilk test was run which also confirmed the normality of the data since there was a p-value of .2463 and a failure in rejecting the null hypothesis of the data being normally distributed. Homoscedasticity was not found because the residual plot did not have any patterns.

Figure 8: Histogram of Residuals

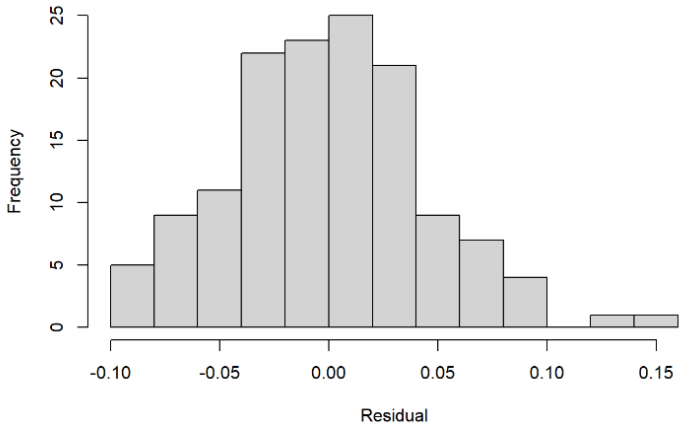
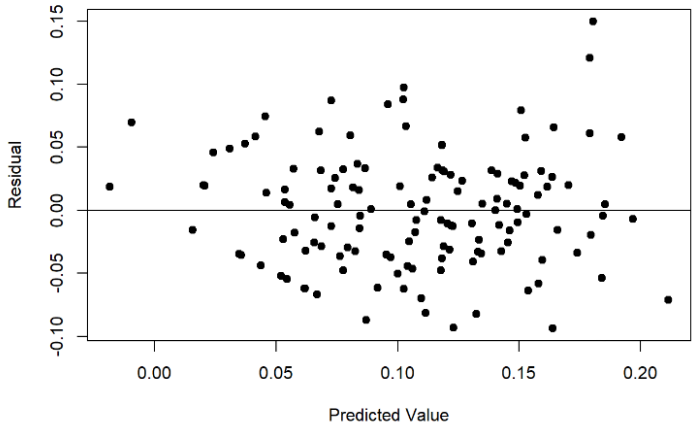


Figure 9: Residual Plot



5. Conclusion

The ultimate takeaway from the final regression is that shot distance, accuracy, and position are the most important variables in converting shots into goals. Forwards who have good accuracy with their shots and who are close to the net when they shoot tend to have a large goals/shot proportion. This regression primarily explains the importance of positioning when it comes to converting shots into goals. The players who are in the best position to take shots will score more goals, but only if they are accurate enough. Therefore, teams should focus on training forwards to make plays where they are in the best

and closest position relative to the net to score. Improving how runs are made towards the goal and how through balls are played by the midfield could help.

It is interesting to see that the two variables from the FIFA 18 dataset are insignificant. It could be that this rating system is not accurate to real life performance or that the most effective way to convert shots to goals is about positioning, which is why variables like average distance and position scale are significant. Of course, skill is important in converting shots to goals which is why shots on target % is also significant. Age and penalty kick attempts were also insignificant and had no predictive power. This may be more apparent since age has no bearing on how well a player can make shots. Also, the amount of penalty kicks would not be significant since it is an entirely different mindset than making a shot during a live game. Some of the roadblocks I stumbled upon was not being able to find data on certain variables that may have been important. For example, the dominant foot of a player, relationship with team, years with a certain club, and other variables can potentially affect the amount of goals per shot he scores. Also, a finishing rating still has the ability to be significant in the regression if it was taken from a different and more accurate source than FIFA 18. Potential areas of research that could branch from this would be what the most important variables are for a team as a whole for converting shots to goals. This would include variables involving possession, assists, passing ability, etc… I believe that the more effective a team is at delivering assists, the more goals will be scored per shot.

Sources

2018-2019 Bundesliga Shooting Stats. https://fbref.com/en/comps/20/2018-2019/shooting/2018-2019-Bundesliga-Stats

Shrivastava, Aman. *Fifa 18 Complete Player Dataset*. https://www.kaggle.com/datasets/thec03u5/fifa-18-demo-player-dataset/