Peter Emero, Rhea Halebeed, Eline Leske, Emalie Zhitomirsky

Professor Gold

Final Coding Paper

8 December 2023

Introduction

Sports are a big part of this world. An average of 1.59 million viewers tune in to watch NBA regular season games, so it is safe to say people care about watching the games and the outcomes of those games. In our project, we are trying to answer the question "Can we predict team success from team shooting splits (field goal %/ 3-point %/ free throw %)?" When trying to decide what we wanted to do for our project we narrowed our topics down to sports, economics, and topics regarding the medical field. However, after discussing with each member in the group we decided to focus our questions on sports because it was the most interesting to us and most relevant to what we each already know. We wanted to solve this question because we talked about how lots of people these days participate in sports betting so trying to figure out an answer to our question could help influence people's chances and choices in sports betting. We decided to look into teams in basketball, more specifically the NBA (National Basketball Association), to gather the data for our project. To break down the understanding of our question, team success means if a team wins, they are successful and if they lose, they are not successful. Team shooting splits include field goal percentage, 3-point percentage, and free throw percentage. In basketball, a field goal is a basket scored other than a free throw, and is worth two or three points depending on the distance of the attempt from the basket. So, field goal percentage is the ratio of field goals made to field goals attempted. 3-point percentage is the percentage of 3 point field goal attempts that a player makes. Finally, free throw percentage is found by dividing the player or team's

made free throws by the total number of free throws he or she attempted, and free throws have a value of one point.

Previous Works

There exist many questions regarding whether or not individual player's shooting splits from different areas of the court (such as field goal percentages, 3-point percentages and free throw percentages) can be analyzed as a whole to indicate or predict an entire team's success in a given season. The five previous works that we will further examine will guide us in our ability to conclude whether or not team success can be predicted from team shooting splits. With the help of these various research techniques, each source aids in finding each team's success along with the method's accuracy in doing so. The first source we can look at is "Game statistics that discriminate winning and losing at the NBA level of basketball competition" from the National Center for Biotechnology Information. In this article, a study took place to "examine differences in game-related statistical parameters between National Basketball Association (NBA) regular and postseason competitive periods and to determine which variables have the greatest contribution in discriminating between winning and losing game outcomes"(Cabarkapa). The study gathered that field goal percentage and defensive rebounding were the two factors that contributed to the winning and losing game outcomes for the teams. They also gathered that overall shooting efficiency helped account for the outcomes and variance. This study helps our project because they are looking into the same factors as us for what contributes to a team's success.

The next paper we looked at was, "Predicting Outcomes of NBA Basketball Games" by Eric Jones, who submitted this thesis to the Graduate Faculty of the North Dakota State University of Agriculture and Applied Science. This study took a stratified random sample of

144 NBA basketball games and developed models to predict in game statistics. This helps us because we are using in game statistics to help us graph and present our data to answer our overall question. In Osken and Onay's study "Predicting the winning team in basketball: A novel approach", they used "cluster memberships to train prediction models" which ultimately allowed them to show "that player stereotypes extracted from individual statistics are a valid approach to predict game winners" (Onay, Ceylan, and Cem Osken). Similar to how their study was able to prove that players' individual shots play a big role in predicting the entire teams' win, the study done by Anthony J. Onwuegbuzie titled "Factors Associated with Success Among NBA Teams' ' also arrived at the same conclusion. In this specific study researchers found that "winning percentages increased with field goal conversion percentage, number of assists per game, and number of points scored per game", essentially supporting the assertion we attempt to prove (Onwuegbuzie, Anthony J). Lastly, in the final study we have chosen to investigate, "Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics", Yuanhao Yang "analyzes the correlation between individual player's statistics and their team's performance, and develops a prediction model that can be used to forecast regular season results of NBA teams", arriving at a similar conclusion that team wins can be attributed to or "based on common player statistics" (Yang, Yuanhao).

Overall, with our examination of the five previous works, our group was able to learn more and contribute data to answer our question on if "Can we predict team success from team shooting splits (field goal %/ 3-point %/ free throw %)." While there might not be a solid answer, all of the previous studies we have presented help support our research, giving relevance to our overall question and topic that we seek to gather data on.

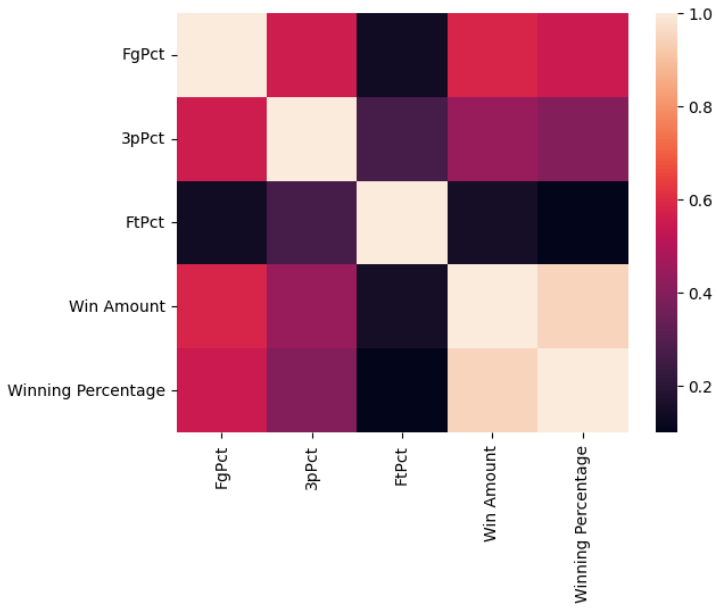
Methodology

Data Preparation:

When we started preparing the data, we took both datasets from data.world which involved team statistics and teams records. Our first step was to remove any years that were not shared between the two datasets, then to remove columns from the statistics dataset which did not involve shooting splits. This gave us the dataset seen below:

	No	Team	Fgm-a	FgPct	3gm-a	3pPct	Ftm-a	FtPct	Year	Record	Win Amount	Winning Percentage	Champion
0	6	Atlanta	38.4-84.5	0.455	10.0-28.8	0.348	15.7-20.0	0.784	2015-2016	48-34	48	0.585	0
1	4	Atlanta	37.9-82.3	0.460	9.8-26.6	0.369	16.2-20.9	0.778	2014-2015	60-22	60	0.732	0
2	18	Atlanta	36.8-81.3	0.452	9.5-26.4	0.361	17.3-22.0	0.785	2013-2014	38-44	38	0.463	0
3	16	Atlanta	37.3-80.8	0.461	8.4-22.9	0.369	14.4-20.2	0.712	2012-2013	44-38	44	0.537	0
4	11	Atlanta	36.4-80.9	0.450	7.3-20.0	0.365	15.3-20.6	0.743	2011-2012	40-26	40	0.606	0

Then, we looked at the correlation between the different variables we wanted to look at in our analysis using machine learning. To do so we created a heat map that shows the correlation between the different variables.



The correlation between “Win Amount” and “Winning Percentage” is high. The cause for this is because the percentage of winning is calculated using the amount of games won and the

total number of games. However, because the 1998-1999 season was shortened to 50 games instead of 82 like the other seasons, it was necessary to calculate and use win percentage to regulate our data. All other variables are not significantly related to each other.

Statistics:

We ran two statistical analyses of our data. The first one was a t-test to compare the mean shooting splits of teams with winning records with those of teams with losing records. To do this, we first created two separate datasets which contained statistics for teams with a win percentage above .500 (winning record) and below .500 (losing record). Then, we further split the data from the winning and losing record datasets into the shooting split variables. Next, we ran t-tests on those further split datasets between the same statistics. We hoped to determine which aspects of the shooting split were statistically significant in teams that had winning records, because a win percentage above .500 is generally considered a good season in the NBA

Our second statistical analysis was to run a chi-square tests for association to see if there is a significant relationship between team success categories like ⅖ games won or over ⅘ games won and teams with above average FT%. To do this, we first calculated the mean FT% so we could add a column determining which team seasons had above average FT%. After determining the above/below average seasons, we added an additional column which split the winning percentages of teams into 5 distinct levels: below 1/5 games won, between 1/5 to 2/5 games won, between 2/5 to 3/5 games won, between 3/5 to 4/5 games won, and over 4/5 games won. We then used the crosstab function to create a chi-square contingency table and then ran the test using the table.

Machine Learning:

Our first machine learning model was a K-Nearest Neighbors Regressor. We used the three parts of a shooting split: FG%, 3P%, and FT% to try to predict the winning percentage of a team's season. We felt a KNN model fit the problem because we would be finding the closest seasons to the datapoint, as every previous season statistic would be a potential neighbor. We decided to vary the n-neighbors parameter to try to find the best performing model. We first created the X and Y datasets of our model which was listed above. We then used the train_test_split function to create training, validation, and testing data. To find the best performing model, we created a for loop to run through different iterations of the model with a different n-neighbor value based on the accuracy of the validation set. Then, we scored the best performing model on the testing data to determine how accurate our K-Nearest Neighbors model really is.

Our second machine learning method was creating a Random Forest Regressor. We felt this was a good choice because we wanted to determine which of the shooting split features were the most important in predicting a team's success. We reuse the same features and prediction variables from the KNN model. Our X dataset was FG%,3P%, and FT% while we were attempting to predict the winning percentage. We decided to vary the n-estimators of the forest to tune for best performance. So, we created another for loop which scored the model with varying amounts of n-estimators. Similarly, we scored the highest performing model of the validation set on the test set to determine the accuracy of our model.

We also created a KNearestClassifier to determine whether or not given 3 features, the shooting split, if a team won a Championship. We once again used the same X dataset which was FG%,3P%, and FT%, however, this time the Y dataset was Champion whose column was a 1 if a

team won the championship. Most of this code overlaps with the KNN Regressor, because of the natural similarities between them both being KNN model's. That being said, predicting whether a team wins a championship is important in determining the success of a team, and we needed to create a classification model to do that.

Results

Statistics:

The result of our first t-test; comparing the mean shooting splits of teams with winning records with those of teams with losing records resulted in the following p-values:

```
#run t-tests for the mean shooting splits of the winning versus losing records
_, p =scipy.stats.ttest_ind(winning_FG,losing_FG)
print(f'FgPct: {p}')
_, p1 =scipy.stats.ttest_ind(winning_3P,losing_3P)
print(f'3pPct: {p1}')
_, p2 =scipy.stats.ttest_ind(winning_FT,losing_FT)
print(f'FtPct: {p2}')
```

```
FgPct: 3.063274446725167e-21
3pPct: 1.2453553624357497e-09
FtPct: 0.157568787072481
```

Therefore, there is a statistically significant difference in the mean FG% and 3P% of teams with losing and winning records. However, the t-test shows that there is no significant difference in the mean FT% of teams with winning and losing records. However, our chi-sqaure test was to determine whether there is an association between above-average FT% and winning percentage thresholds. Our results from the chi-square test are pictured below:

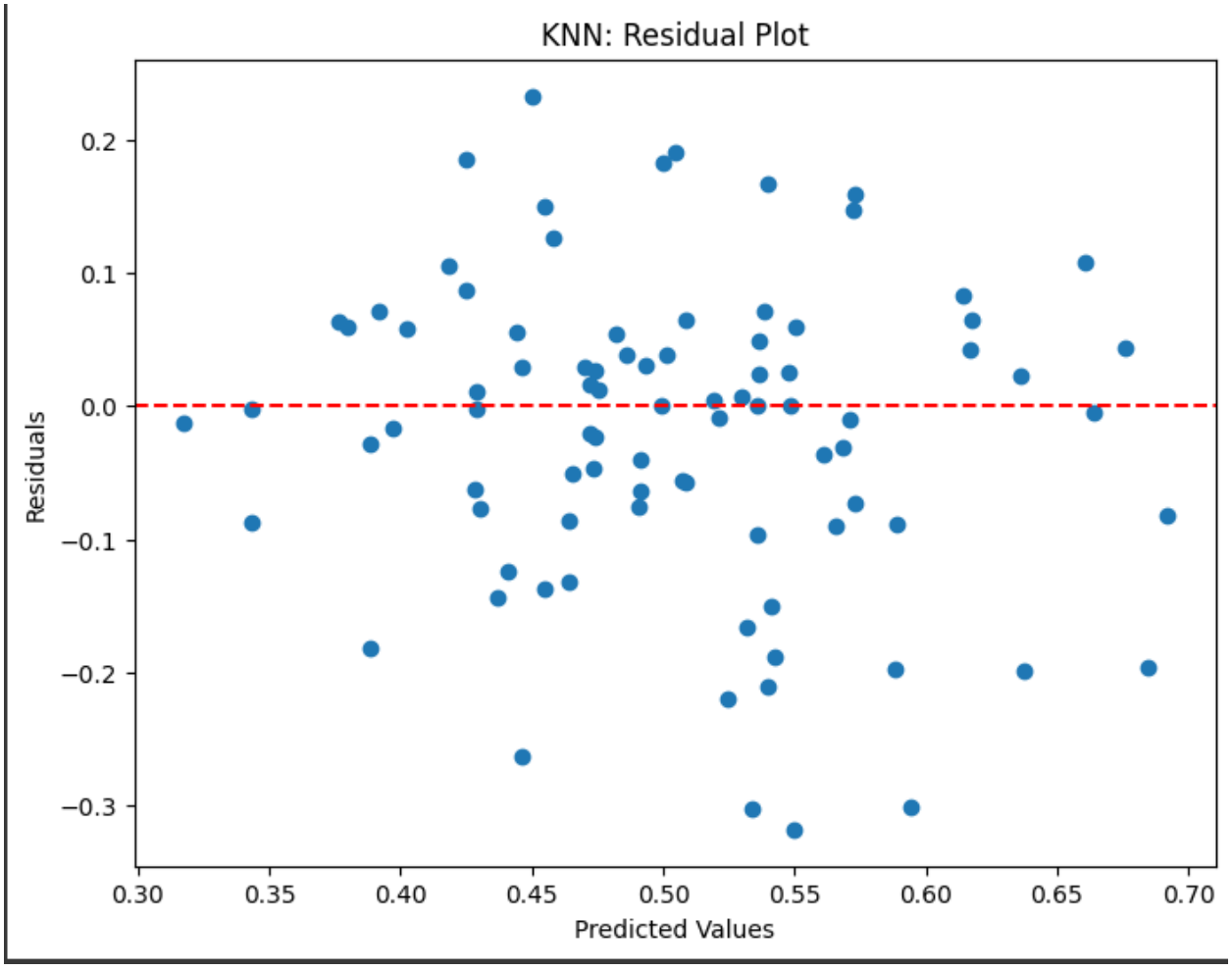
```
#run a chi-sqaure test to determine the significance of FT% to Win Level
_,p,_,_=scipy.stats.chi2_contingency(contingency)
p
```

```
0.04138923190508926
```

Therefore, as the p-value of .04 is less than .05 we can conclude that there is a significant association between having an above average FT% and the success of a team. Therefore, each of the features we are using for our machine learning models have impacts on the success of a team and are important to include in any potential future model.

Machine Learning:

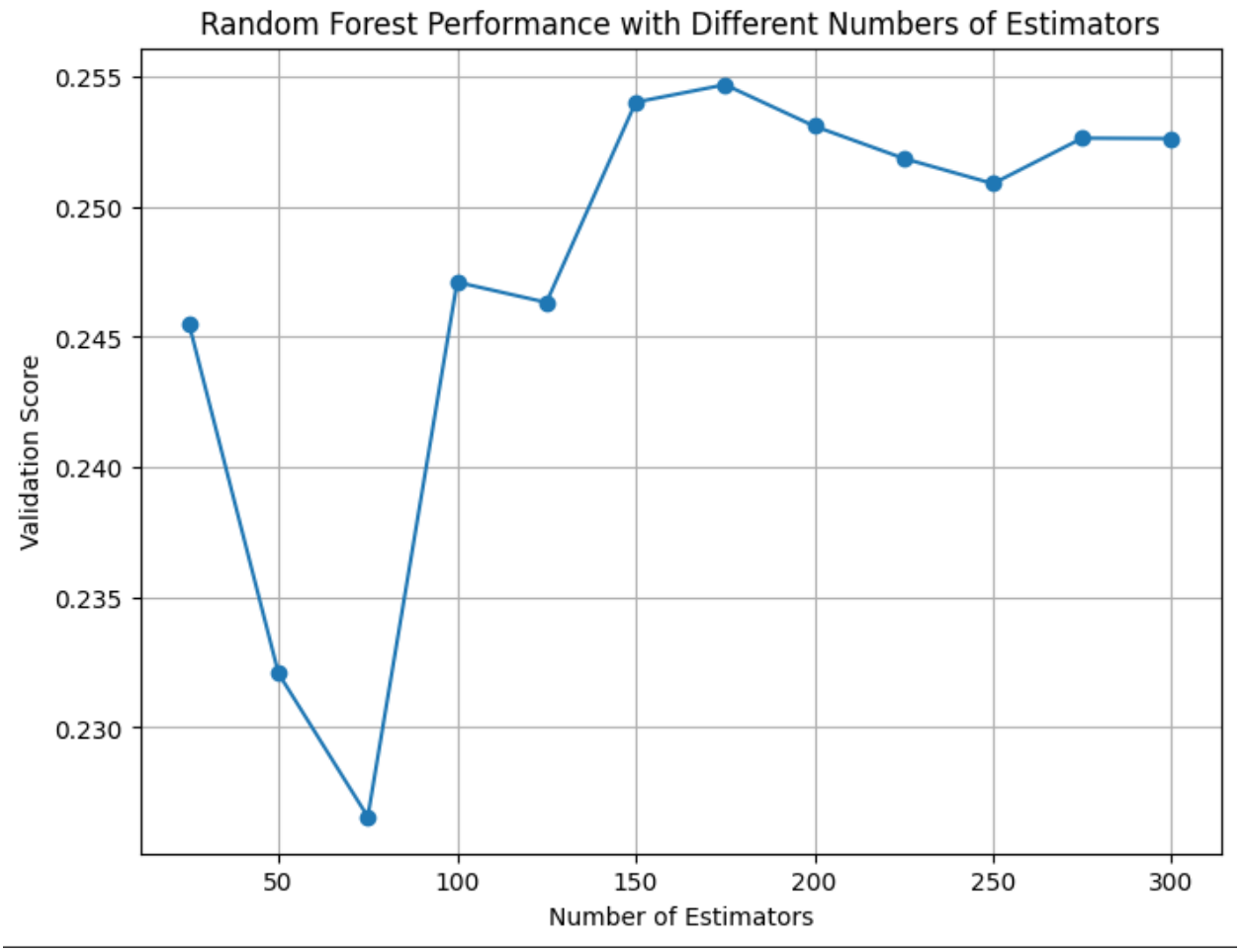
The K-Nearest neighbor model was unable to predict the winning percentage of a team very accurately. This residual plot of our model is pictured below:



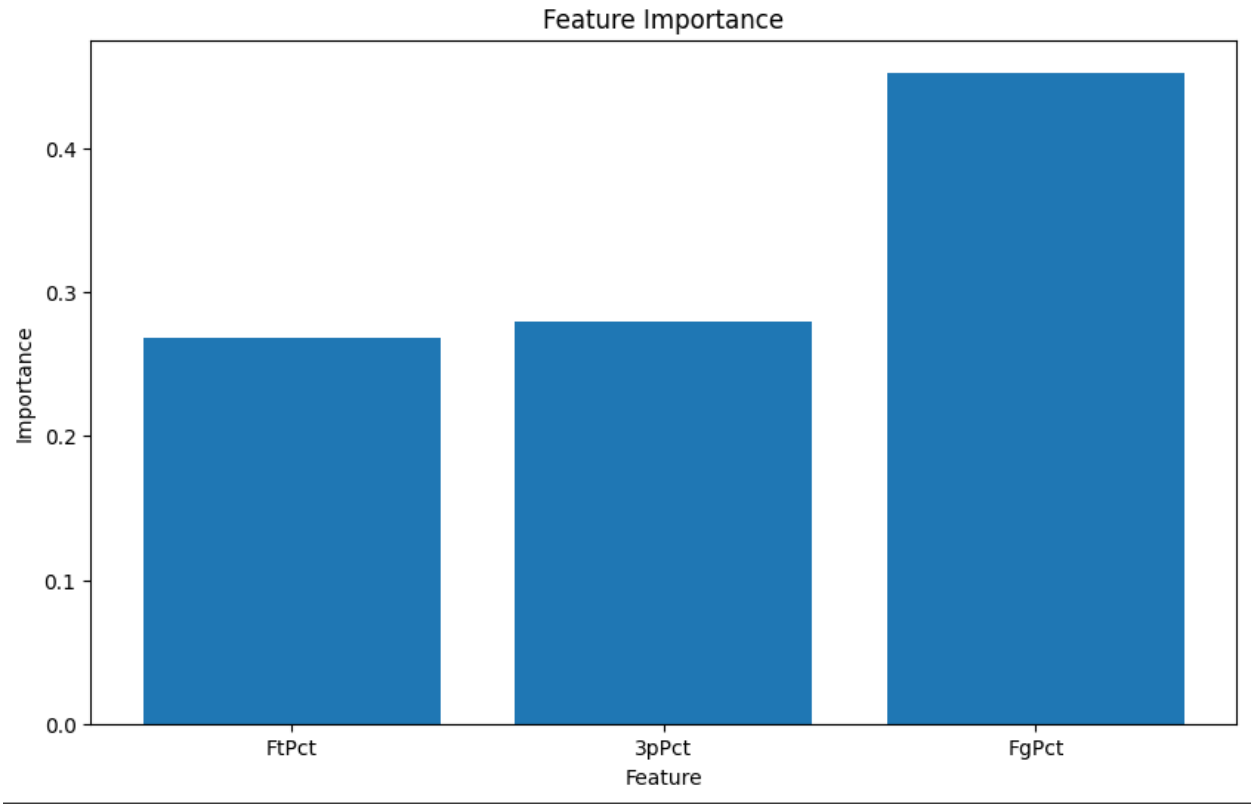
This clearly demonstrates that the machine learning model was unable to predict the outcome accurately. Ideally the blue point would scatter closely around the red line. However

this is not the case. They are scattered all over the graph and very few are close to the central red line. This portrayed why the best score for the model was about 0.265.

In addition the Random Forest Regression was also unable to accurately predict winning percentage. We looked at different n-estimators in the range of 25 to 325 in 25 step intervals. This is the outcome:

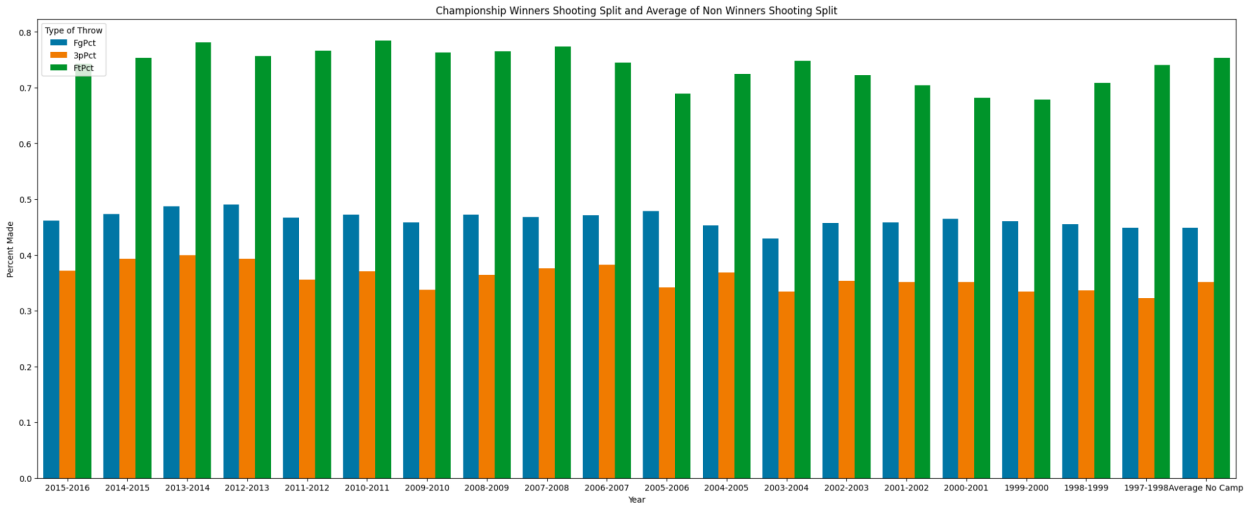


The model performed best using 175 estimators. The accuracy was about 0..277. Comparing it to the usual wanted accuracy of about 0.95 this is much lower. Meaning that the models are not able to accurately predict the outcome of the next NBA season using only the shooting splits.



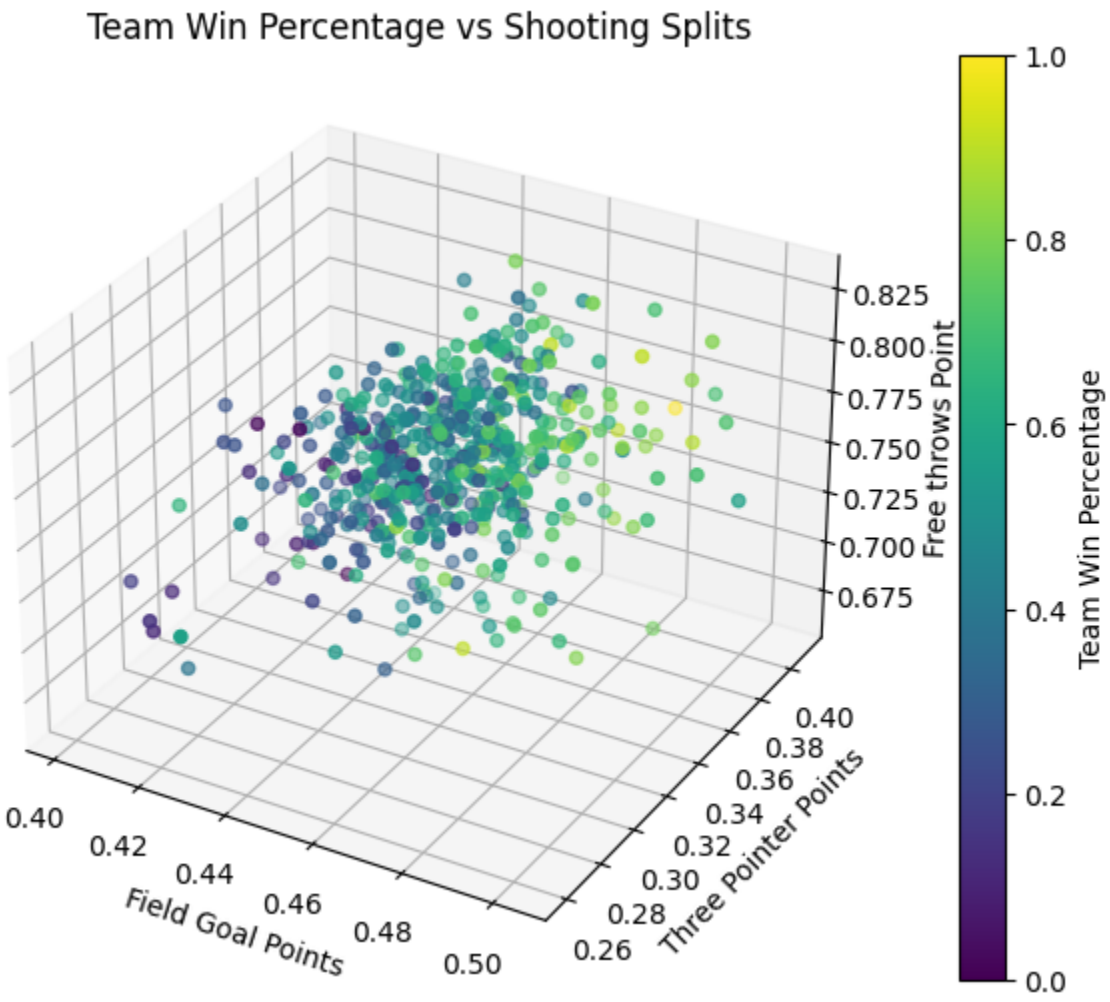
The Random Forest also allowed us to determine that FgPct was the most important feature it determining winning percentage.

Our KNN classifier model performed much better than the other two models. It performed the same for all n_neigbor values but we chose to use 2 which gave us an accuracy of approximately .973, a stark difference from our other two less successful models. Our performance was actually the same despite changing the n_neighbors parameter, which leads me to believe that the high success rate could be attributed to a lack of champions, considering we don't have an incredibly large dataset and there can only be one champion every year.

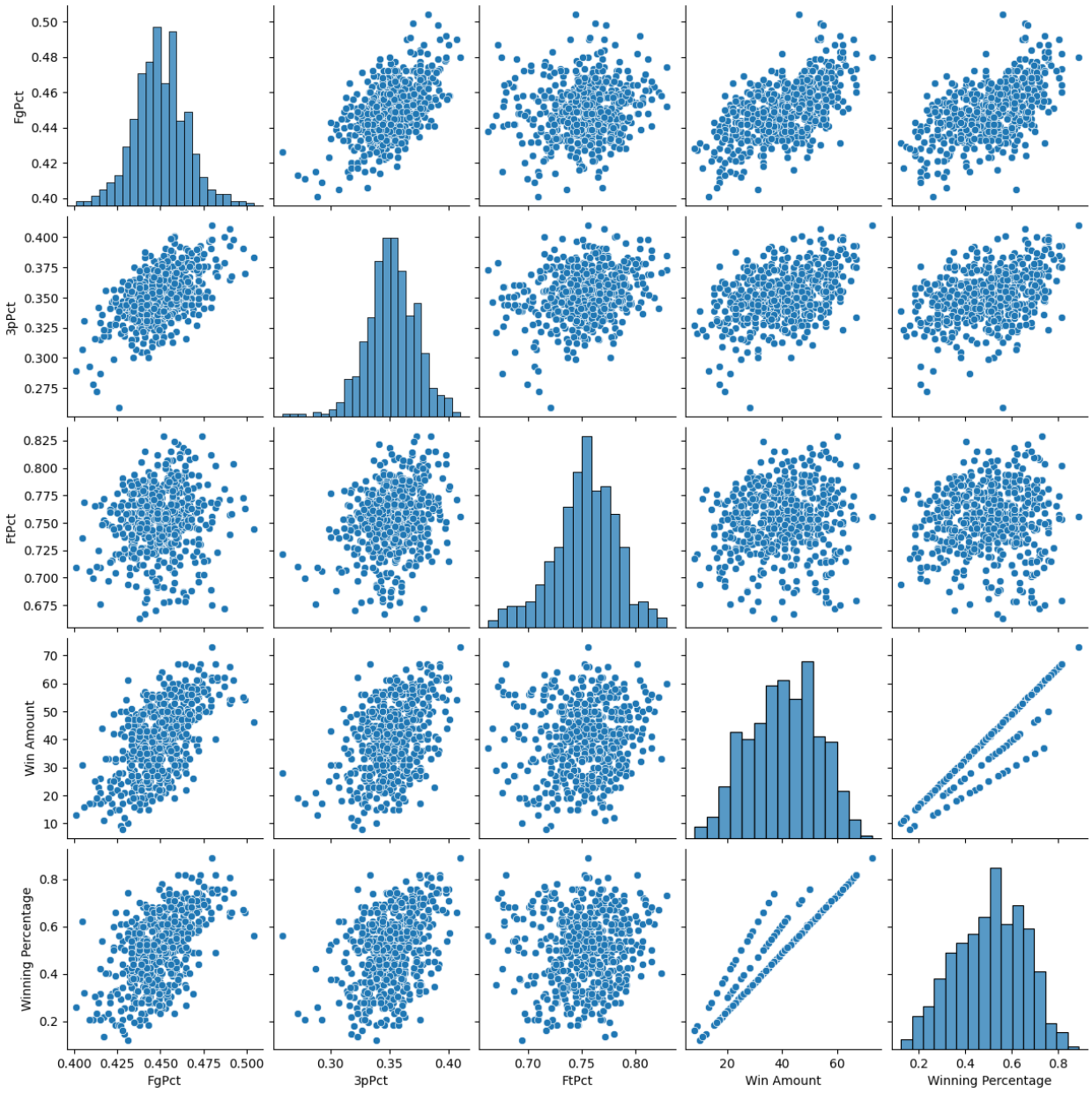


The above figure shows the championship winners' shooting split and the average of the teams that did not win the championship. The average of the losing team is higher than the splits of some of the teams that won the championship. This reinforced that it is difficult to accurately predict success based on exclusively shooting splits and the fact that the championship model might only be succeeding because of a lack of data.

Other Visualizations



This Figure is meant to visualize the relationship between team win percentage and the three different shooting split features simultaneously.



This Figure is meant to visualize the relationship between the different shooting split features and see if there are any distinguishable relationships.

Conclusions

Our central question was "can we predict team success from team shooting splits (field goal %/ 3-point %/ free throw %)?" I would say our results don't provide convincing evidence to say that we can predict team success from shooting splits. However, based on various pieces of

our report we believe that shooting splits are important to predict a team's success. For example, our statistical exploration concluded that there is a statistically significant difference in the mean FG% and 3P% of teams with losing and winning records. Additionally, FT% has an association with success level. So, all three shooting split variables are significant when it comes to the success level of teams. Additionally, our KNN Classifier for championship winners was very successful, but it is important to recognize the potential bias from a relatively small sample size. Our two regressor models were less successful but they represent a good starting point. Therefore, we are safe to conclude that if we added more features like rebounds per game, points scored, points allowed, etc. then we would expect to have more consistent and accurate results with our model. That being said, there is not substantial evidence to say that you can predict NBA team success ONLY with shooting splits.

References

Cabarkapa, Dimitrije et al. “Game statistics that discriminate winning and losing at the NBA level of basketball competition.” PloS one vol. 17,8 e0273427. 19 Aug. 2022, doi:10.1371/journal.pone.0273427 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9390892/

Jones, Eric S. “Predicting Outcomes of NBA Basketball Games.” April 2016, Fargo, North Dakota. https://library.ndsu.edu/ir/bitstream/handle/10365/28084/Predicting%20Outcomes%20of%20NBA%20Basketball%20Games.pdf?sequence=1&isAllowed=y

Onay, Ceylan, and Cem Osken. “Predicting the winning team in basketball: A novel approach.” Heliyon. 2022 Dec 9;8(12):e12189. doi: 10.1016/j.heliyon.2022.e12189. PMID: 36561688; PMCID: PMC9764182. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9764182/

Onwuegbuzie, Anthony J. “Factors Associated with Success Among NBA Teams”. February 13th, 2008. https://thesportjournal.org/article/factors-associated-with-success-among-nba-teams/

Yang, Yuanhao. “Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics”. May 2015, University of California at Berkeley. https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Stanley_Yang%20_Thesis.pdf