Predicting which NBA Team will win the Finals

Introduction:
The NBA Finals is the ultimate showdown in professional basketball, marking the pinnacle of the National Basketball Association (NBA) season. It's the thrilling climax that determines the league's champion, and it has a fascinating history that stretches back to the inaugural season in 1946-1947. Back then, the NBA emerged from the merger of the Basketball Association of America (BAA) and the National Basketball League (NBL). Since then, the NBA Finals have become a stage for unforgettable moments, showcasing the brilliance of legendary players and sparking fierce rivalries that have become part of basketball lore. From the enduring dominance of the Boston Celtics and Los Angeles Lakers to the emergence of iconic superstars like Michael Jordan, Magic Johnson, Larry Bird, and LeBron James, the NBA Finals has been a platform for showcasing extraordinary talent. The series follows a best-of-seven format, with the first team to win four games crowned as the NBA champions. It captivates basketball fans worldwide, bringing together the most elite teams and players, and etching unforgettable memories in the annals of basketball history. For more information about the NBA Finals please visit this link.

Motivation:
By leveraging machine learning algorithms, I aim to delve deeper into the intricacies of the game, uncovering hidden patterns and factors that contribute to team success. The satisfaction of accurately predicting the outcome of the NBA Finals became a personal challenge, showcasing my analytical skills and expertise in merging data-driven insights with my knowledge in the sport. Moreover, the pursuit of predicting the NBA Finals winner can provide me a unique avenue for engagement and discussion within the basketball community, fostering meaningful conversations and sharing insights with fellow enthusiasts.

Constructing the Model:
In order to build a predictive model for the NBA games, several steps were undertaken. Firstly, the data was loaded using the pandas library, reading the game information from the "games.csv" file. The dataframe was then sorted by date to ensure chronological order. Empty entries and data before 2004, which contained NaN values, were dropped to ensure data integrity.
Since the "games.csv" file does not have the team name included but rather the team id, the "teams.csv" file was loaded, containing the team IDs and corresponding names. The "HOME_TEAM_ID" and "VISITOR_TEAM_ID" columns in the dataframe were

replaced with their respective team names using merge operations with the loaded team names dataframe.

The data was filtered to include only games within a specific time period, from August 2020 to July 2021. However, it is easy to change the time period in the code to look at multiple seasons. This focused the analysis on a particular season that has already been completed so we can test the model and see how accurate it is. Selected columns relevant to team statistics were identified, such as field goal percentage, free throw percentage, three-point percentage, assists, rebounds, points allowed and points scored. These columns were stored in the 'columns' variable.

To prepare the data for modeling, the winning team for each game was determined based on the comparison of points scored by the home and visitor teams. A new column, "Winning_Team," was created using a lambda function that assigned the home team name if they had a higher score, and the visitor team name otherwise.

The feature matrix (X) was created by selecting the columns of interest, and the target variable (y) was set as the "Winning_Team" column. The data was then split into training and testing sets using the train_test_split function, with a test size of 0.2 and a random state of 42.

To ensure consistent scaling, the numeric features in the training and testing sets were standardized using the StandardScaler. This step was essential for proper comparison and interpretation of the features.

By executing the provided code, a preprocessed dataset was obtained, with scaled features ready for modeling the prediction of winning teams in NBA games.

Training the Model:
In order to train the predictive model, a Support Vector Machine (SVM) classifier was utilized. The SVM model was instantiated with the parameter probability=True to enable probability estimation. The training data, consisting of the scaled feature matrix (X_train_scaled) and the corresponding target variable (y_train), were used to fit the SVM model using the fit() method. This process involved finding an optimal hyperplane that maximizes the margin between different classes in the data.

To evaluate the performance of the trained model, the accuracy of the predictions was calculated using the score() method on the testing data (X_test_scaled and y_test). The resulting accuracy value provided an estimate of how well the model generalized to unseen data, representing the percentage of correctly predicted outcomes.

Furthermore, the trained SVM model was utilized to predict the probabilities of each team winning in the testing set. The predict_proba() method generated the probabilities of each class (i.e., each team) based on the input data (X_test_scaled). Additionally, the

svm.classes_ attribute stored the corresponding classes (team names) for the predicted probabilities.

By executing this code, the SVM model was trained, its accuracy was evaluated, and the probabilities of each team winning in the testing set were predicted, providing valuable insights into the likelihood of success for each team.

Results:
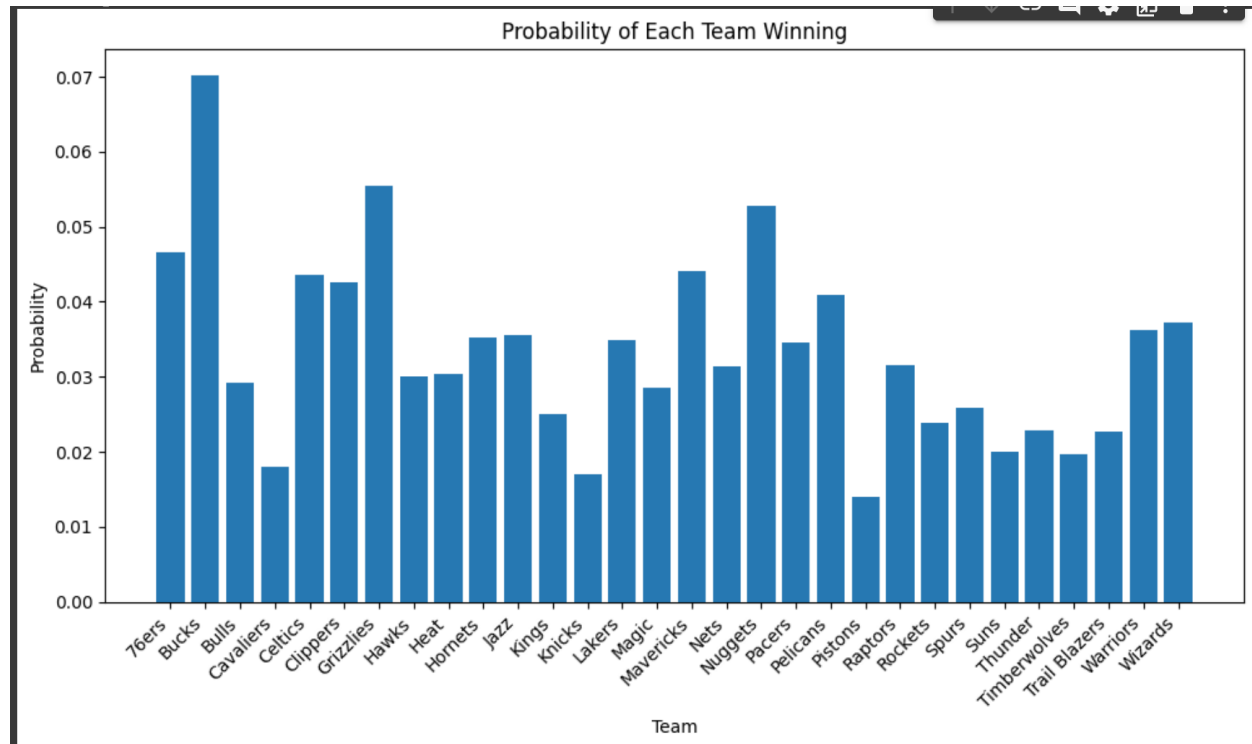
In the 2020 - 2021 NBA season, the model predicted that the Milwaukee Bucks had the best chance of winning the NBA Finals. This was correct, the Milwaukee Bucks did win the NBA Finals that season thus showing that the model could be accurate. The team with the best chance of winning in the Western conference was the Denver Nuggets. The Denver Nuggets in that season lost in the second round to the Phoenix Suns. The team that represented the Western Conference in the Finals that season was the Phoenix Suns, another team that had a low probability of making it to the finals.

Now let's test the model for the 2021 - 2022 NBA season.
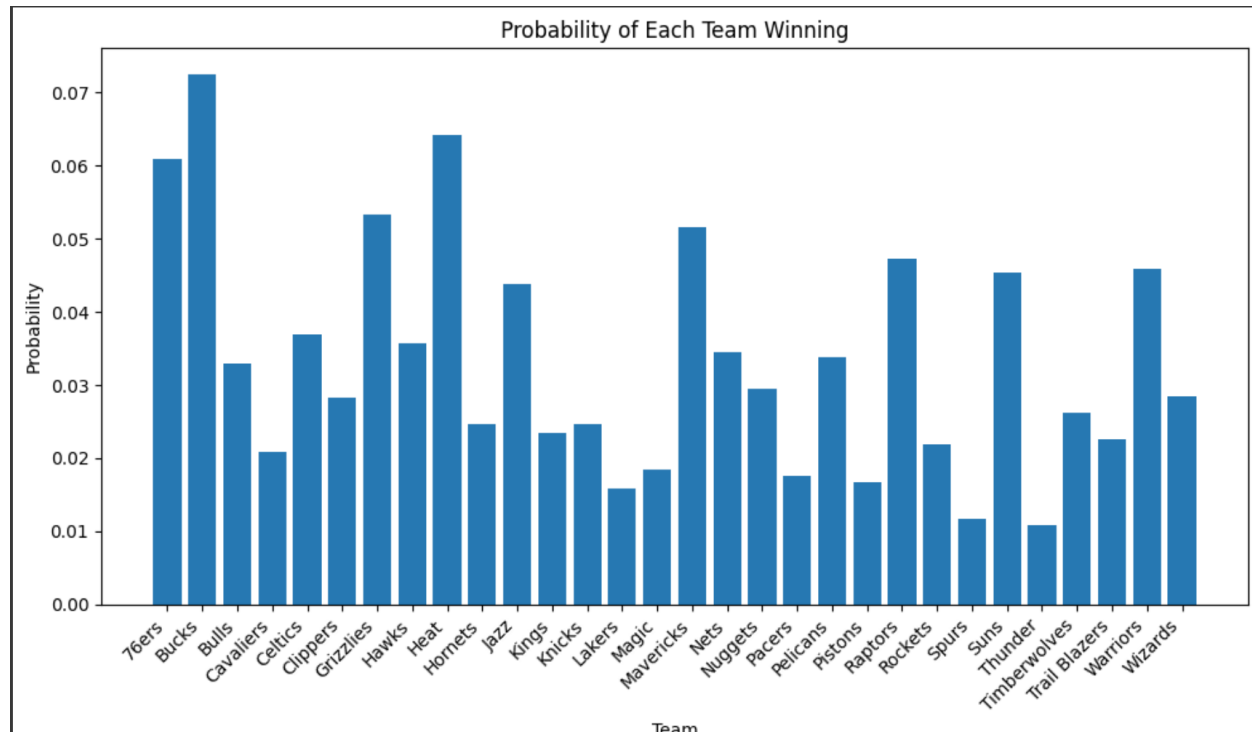
The Model has the Milwaukee Bucks to repeat as champions. In reality, the team suffered a tragic injury to one of its stars, Khris Middleton and lost in a closely contested second round series to the Boston Celtics. The team that won the NBA Finals that season was the Golden State Warriors defeating the Boston celtics.

Let's look at a couple other years.

Probability of Each Team Winning
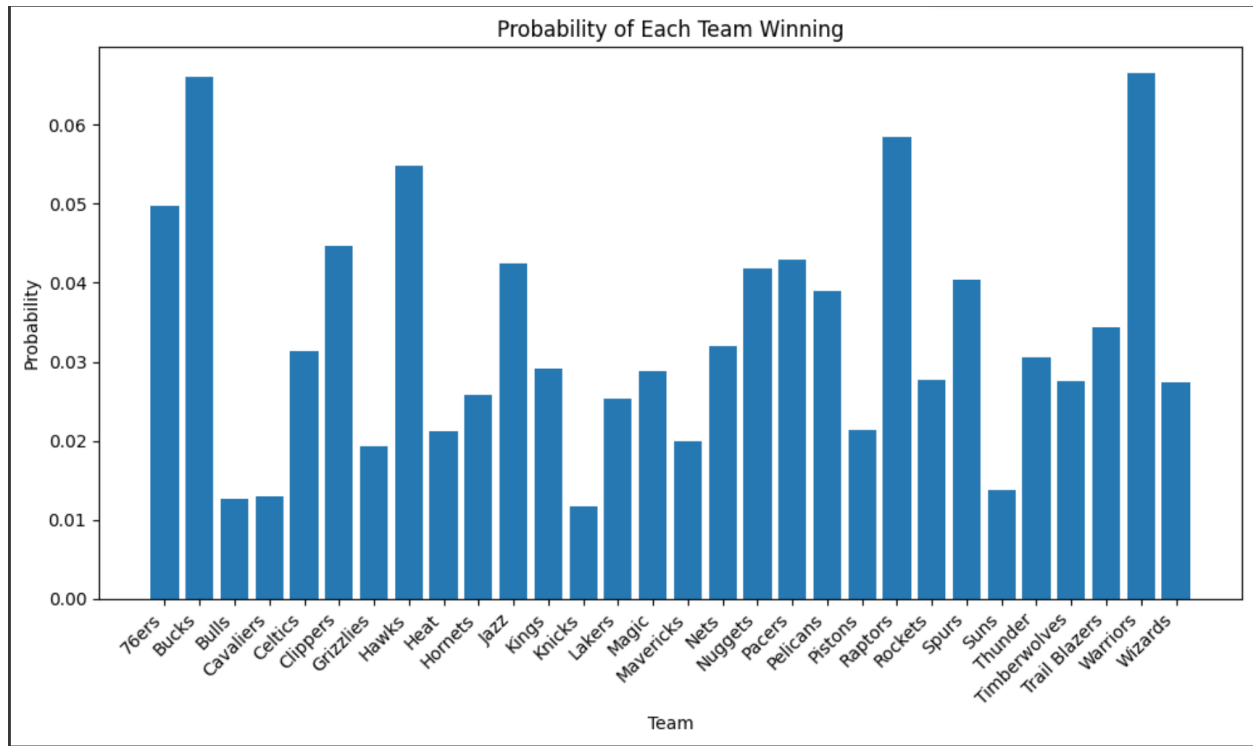
In the 2018-2019 season, the model predicted the Golden State Warriors to win but they were defeated by the 3rd most likely team to win, the Toronto Raptors. The Warriors dealt with major injuries during the Finals, including an Achilles tear by superstar Kevin Durant and a torn ACL by superstar Klay Thompson. The second most likely team to win, the Milwuakee Bucks were defeated by the Toronto Raptors in the Eastern Conference Finals.

Probability of Each Team Winning
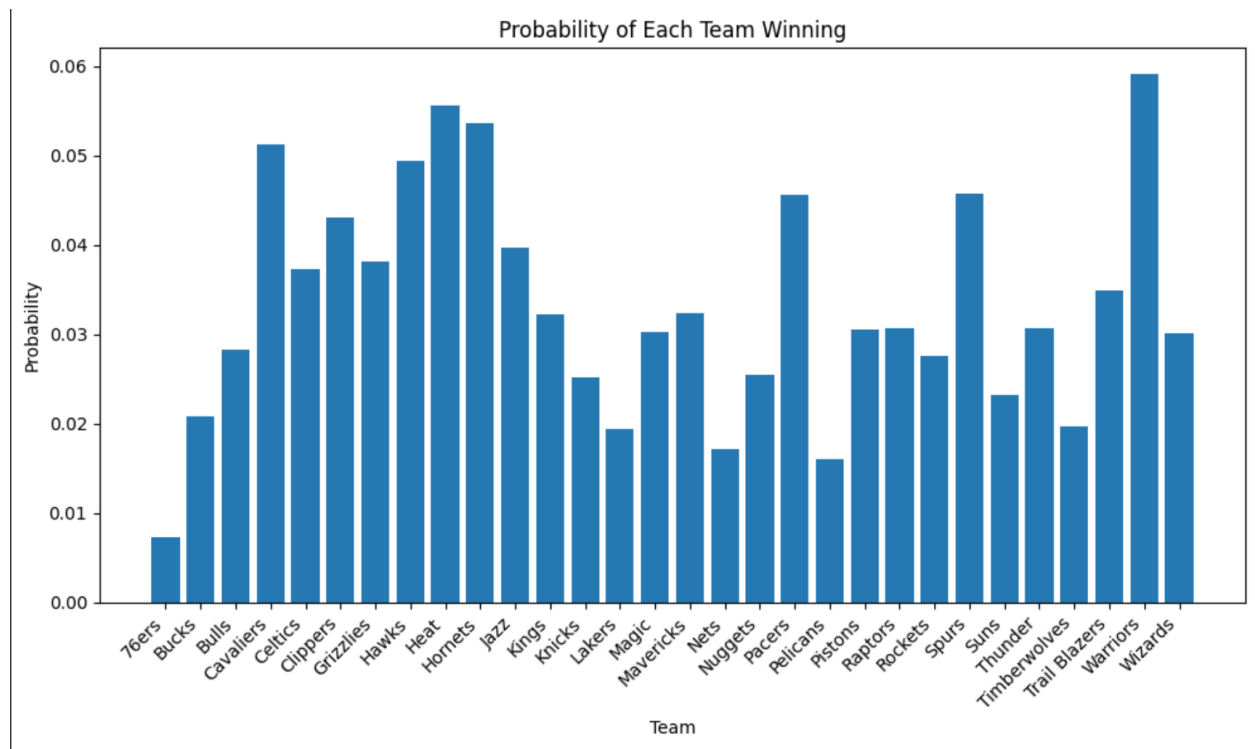
In the legendary 2015-2016 season, the model had the Golden State Warriors winning the finals. In reality they lost to the Cleveland Cavaliers, which the model had 4th most likely to win, in a close fought series. The Miami Heat, the 3rd most likely team to win the Finals were eliminated by the Toronto Raptors in the 2nd round while the Charlotte Hornets were eliminated by the Miami Heat in the 1st round.

Probability of Each Team Winning
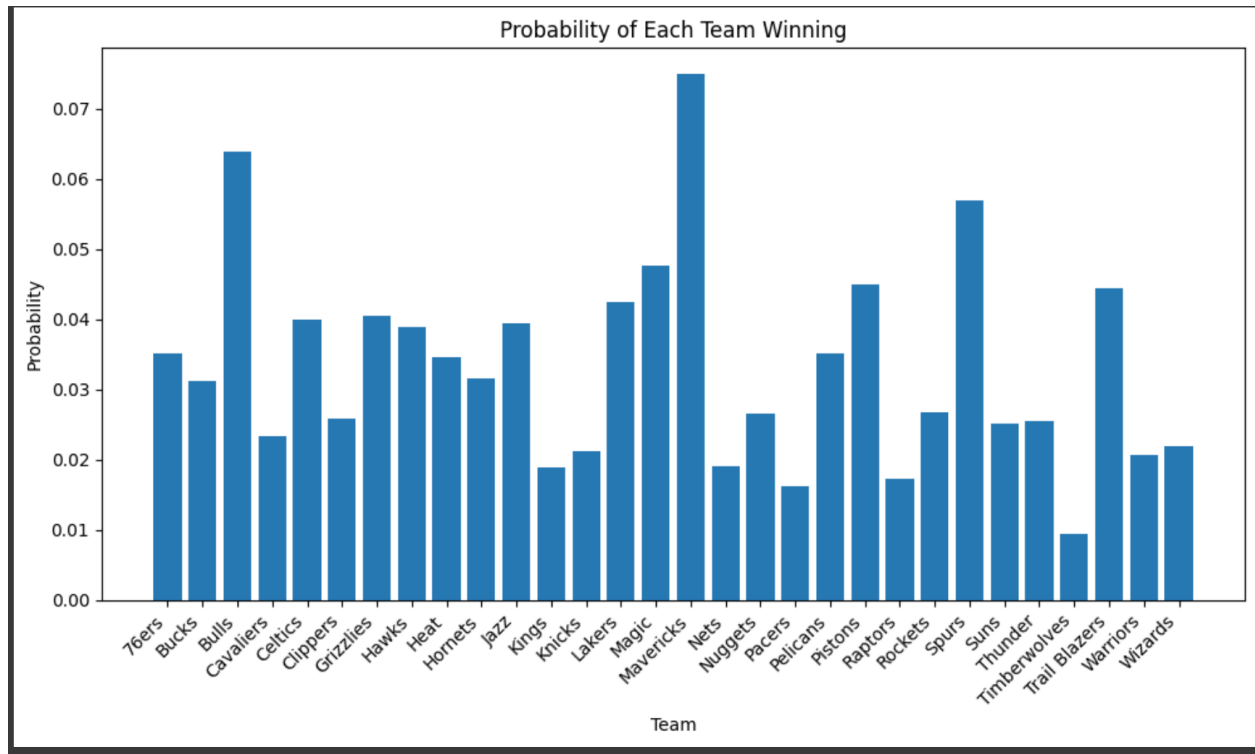
In the 2010-2011 NBA Season, the model correctly predicted the Dallas Mavericks to win the NBA Finals. This run by the Dallas Mavericks is considered one of the most unlikely runs of all time but was correctly predicted by the model. The Chicago Bulls, the second most likely team to win the Finals were eliminated by the Miami Heat in the Eastern Conference Finals, while the 3rd most likely team to win, the San Antonio Spurs, the best seed in their conference, were eliminated in the first round by the Memphis Grizzlies.

Now let's see what team the model thinks will win the Finals this season.

Probability of Each Team Winning
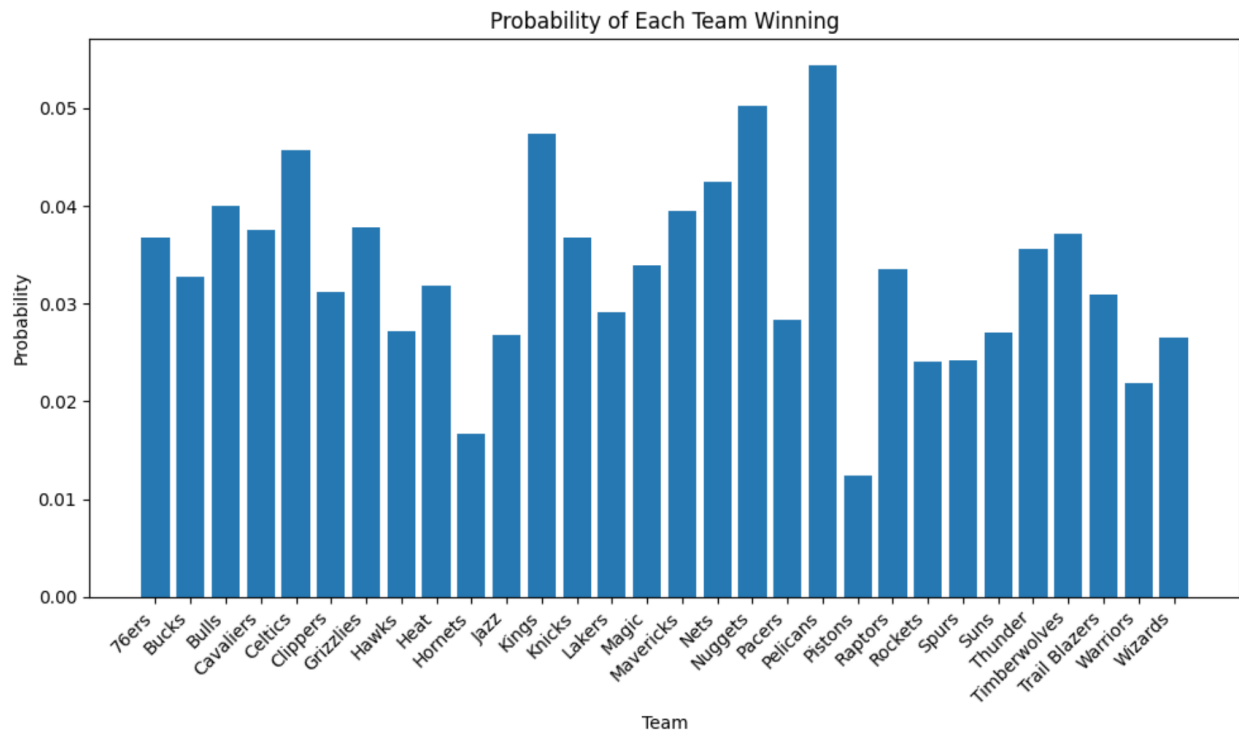
The model predicts that the New Orleans Pelicans will win the NBA Finals. However, the Pelicans lost in the play-in game while dealing with injuries all year to their superstars Zion Williamson and Brandon Ingram. After the Pelicans, the model is predicting the Nuggets, Celtics and Kings have the best chances of winning. The Nuggets and Celtics are currently still in the playoffs, playing in the Western and Eastern Conference Finals respectively, while the Kings were upset by the Golden State Warriors in the first round. The other teams that are still in contention are the Los Angeles Lakers, and Miami Heat which both have relatively low odds of winning. However, it is important to note that the Los Angeles Lakers and Miami Heat ended the regular season as the 7 and 8 seeded teams respectively in their conference and have upped their play drastically in the playoffs.

Conclusion:
In conclusion, the model was able to correctly predict the winner of multiple NBA finals. However, it is important to note that various factors play into the winner of the NBA finals including injuries and matchups which cannot be accounted for in this model. Some future work that can be done include predicting which players are likely to get injured in the playoffs and how that can impact the team's chances in the finals and predicting which player would win the Finals MVP for their team.