Winston Mok, Eric Verduzco, Jack Yu

Data 100

May 13, 2020

Final Project Narrative

Abstract

Basketball, like many other sports, has become subject to datification creating an environment where rich data is created after every game. The analysis of basketball data allows for researchers to find interesting trends to uncover underlying truths which in turn can become beneficial for managers when it comes to making decisions. In this research, we focused on manipulating post-game player box scores, to find correlations between a player's position in court and their game statistics. We then concluded our project by predicting an athlete's in-court position, based on their post-game statistics.

Introduction

In recent years, there has been an increase in the demand for data analysis in sports to optimize decision making during and after a game. Additionally, interest in load management has been at an increase in recent years, having many players advocating rest after a series of games. For that reason, our team decided to analyze player box score data to explore the data through different methods, in an attempt to find correlations between players' minutes played and their performance in that game. Furthermore, we used this idea to build a predictive model using Random Forest Classifier, to classify a player as a Center (C) , Guard (G), or Forward (F), based on their statistics of a given day. Our overall goal of this project was to discuss underlying correlations that could be found in players' box score data.

Description of data

The data we analyzed contained player box data: a combination of statistics (3 point percentage, number of rebounds, height, etc.) that each player achieved during a single game, these stats spanning from 2012 to 2018. In order to try to find a correlation between minutes played and player position, we grouped by the position. However, because some players were only marked as Forward or Guard rather than Small Forward or Power Forward and Point Guard or Shooting Guard, our team decided to just make 3 positions: guards, forwards, and centers. By grouping by these 3 positions, we were able to see the average statistic of each position for all the games listed.

In order to answer our first question of whether or not there is a correlation between player time and player position, we constructed a box and whisker plot to obtain the 25th to 75th percentile of player times at each position (Figure 1 at the bottom of this report). Even though the medians looked different for each position, the 25th to 75th percentile had a large amount of area overlapping, leading us to support the null hypothesis that there was little to no correlation between position and player minutes.

To make our data model accurate at guessing player position, we needed to find certain attributes that had a high correlation between player position and the statistic. We constructed a heatmap with each position and the statistic converted into standard units (Figure 2). With the use of standard units in the heat map, we could see which statistic had the largest difference from the mean, meaning the most variance and hopefully the most correlation between the different roles. Two attributes that stood out in the heatmap were height and weight; centers having around a one standard deviation above and guards with one standard deviation below for both categories. The box and whisker plots of these two attributes greatly supported our claims of a strong correlation, with the 25th to 75th percentiles of both attributes not overlapping at all (Figure 3 and Figure 4). Using this heatmap, we were able to choose which statistics seemed to have a large correlation with a player position in order to construct our model to be the most accurate it could be.

Description of Methods or Summary of Results

For predictive modeling, we wanted to predict a player's position. Originally, we decided to predict a player's position based on how many scores, assists, steals, and blocks they made, as well as the field goal, two-point, three-point, and free throw percentage. Once we cleaned our data to include only those columns, we split it into our train and test data. From there, we had to choose a model. At first, we weren't sure which model we wanted to use, so we decided to do several models and pick the one with the highest training accuracy. Out of all the models we ran, which included Linear Regression, Logistic Regression, Decision Tree, and Random Forest, the Random Forest model had the highest training accuracy (.841). Even though the Random Forest model had fairly high accuracy, we decided to do cross-validation to make sure that it does have high accuracy. It turns out that the model had an average validation score of .595 which is lower than expected. We decided to include more features so that our accuracy score could be higher.

We included the player's weight, height, turnovers, personal foul, field goal attempts, field goal made, two-point attempts, two-point made, three-point attempts, three-point made, free throw attempts, and free throw made to our existing features. Once we did this, our training model accuracy went up to .989 and our average cross-validation score went up to .873. From earlier visualizations (Figure 3 and Figure 4), we believe that the two features that helped make our model more accurate were the player's height and weight. After checking both training and the cross-validation accuracy scores, we decided to look at our test accuracy score and it had an accuracy of about .879. Thus, we can conclude that our model is fairly accurate. To further support our statement that our model is fairly accurate, we computed the root mean squared error for both train and test and both had extremely low error. Lastly, we plotted a graph (Figure 5) that indicated the root mean squared error depending on the number of features.

Seven Specific Questions

I. The two most interesting features we came across were the player's height and weight.

II. We didn't encounter a feature that we thought would be useful but turned out to be ineffective. In other words, all the features we selected were effective.

III. We had a difficult time choosing which model to use. We decided to go with the Random Forest Classifier model since it had the highest train accuracy and cross-validation accuracy score among all the models we tried.

IV. There weren't any limitations of the analysis we did. We had enough features to build our model, which had fairly high accuracy. We believe our predictive modeling is fairly straightforward; thus, none of our assumptions could be proven to be incorrect.

V. We did not encounter any ethical dilemmas with the player box score dataset in our project.

VI. Other data that would have contributed to our research would have been the location in court where a player passed the ball, took a shot, etc. This information would have been helpful to have created another feature such as a player's intelligence in the court.

VII. Although not present in this project, a growing concern is the data privacy of athletes and the extent to which data collection goes; it is concerning if an athlete's biometric data is available to the public. We have to remember that data is still people.

Data Visualizations

Figure 1:

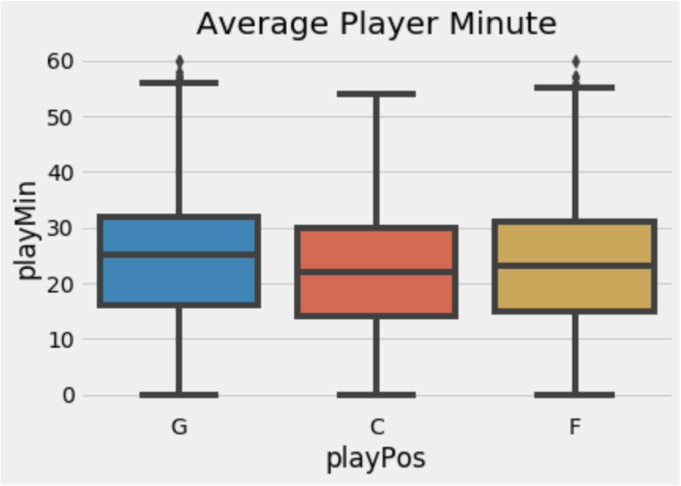


Figure 2:

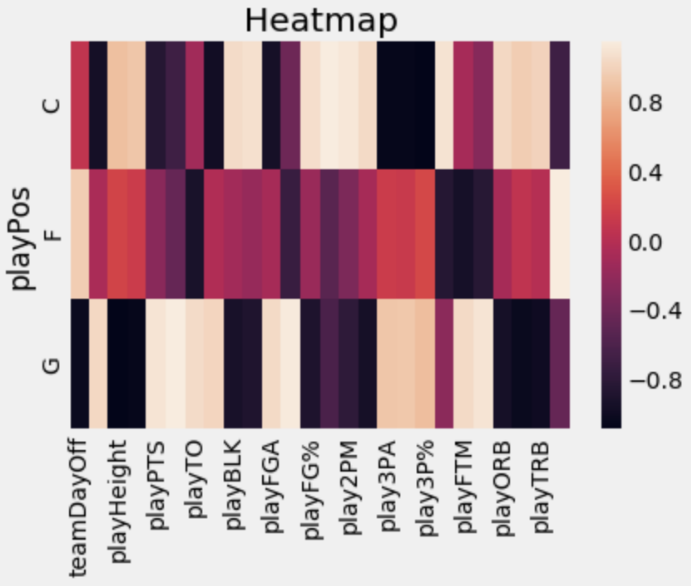


Figure 3:

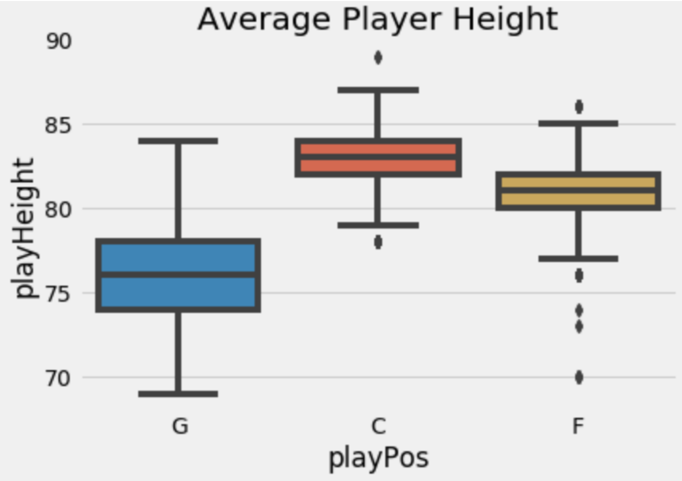


Figure 4:

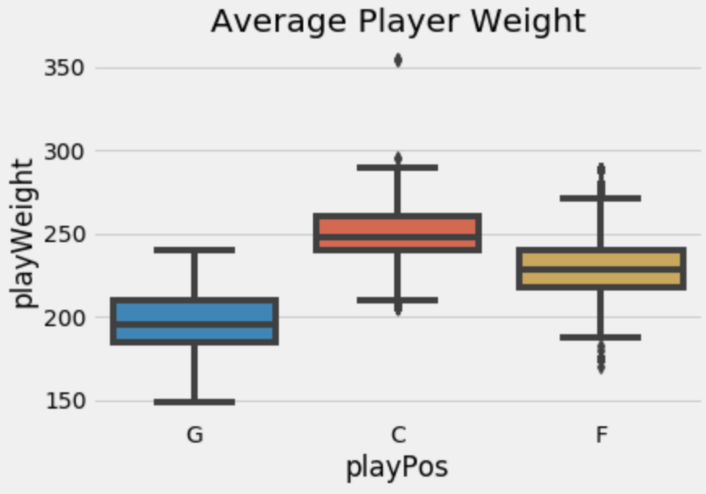
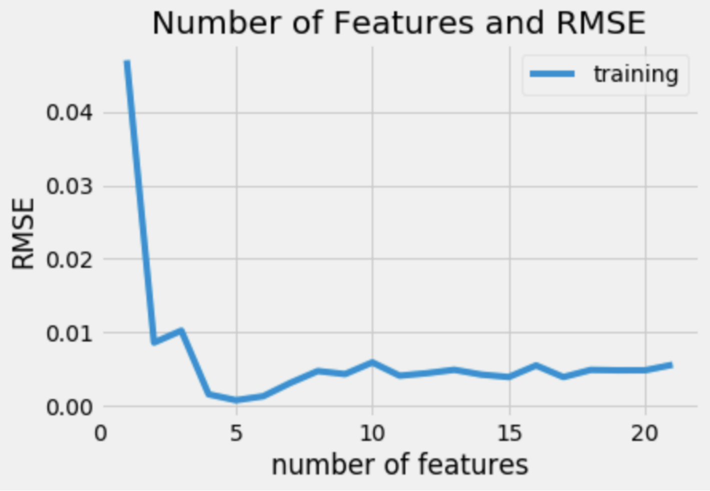

Average Player Weight
350
300
250
200
150
playWeight
G
C
F
playPos

Figure 5:


Number of Features and RMSE
training
0.04
0.03
0.02
0.01
0.00
RMSE
0
5
10
15
20
number of features