The Ideal Body Size and Skills Required for Each Positions in Basketball

Felicia Pramudji (SID: 3035033912)

Isabella Kurnia (SID: 3035115045)

University of California, Berkeley

Data 100: Principles and Techniques of Data Science

Prof. Ani Adhikari & Prof. Joey Gonzalez

May 13, 2020

I. Abstract

The National Basketball Association (NBA) is a competitive basketball game where only the most talented and fittest people are recruited and given the chance to play in an NBA-participating team. The players are often associated with a physically fit body with longer-than-average limbs and an extremely tall stature, where the average height of the players. However, each position has their own ideal body weights and heights. Therefore, in this report, we would like to prove that the height and weight of each player influences the position they will be assigned to the group. Consequently, this project may provide a prediction for future players of their potential positions using their height and weight.

II. Introduction

It is a well-known fact that every basketball players' dream is to play in the NBA championships. However, getting scouted by an NBA coach is an incredibly hard thing to achieve, since most coaches look for only the best of the best. If someone happens to be more than 7 feet tall and has some skills in the sport, they have a 28% chance of making it to the NBA since basketball is a sport dominated by very tall players. However, not all positions in the team are determined by the height of the player, some positions require more skills and high basketball IQ than a bigger body size. Steph Curry is not the tallest NBA player, but his proficiency and abilities are noted in a basketball court. Some positions are favored by height, while others are favored more by skills and intricate techniques. That's why there are players of different heights distributed in different positions. Therefore, in this project, we would like to prove the hypothesis that certain positions require certain body sizes. Additionally, we would like to figure

out the ideal body height and weight required for each position of the team. This will help aspiring basketball athletes to assess themselves and figure out the areas they can improve on.

III. Description of Data

The most important features of this project is the height and weight of each player and the positions they hold in the team, hence missing values may cause errors in our conclusion. Therefore, we decided to exclude data points where either values are missing. We also filled all the NaNs with 0 for all integer columns. We decided to clean our data in this way since we will only be using the top priority features. Since the weight and height of the players and the position are the core features of the experiment, these were the first features we cleaned. The other features that were all integer columns which we filled in 0 for all empty cells.

As we can see from our scatterplot, players with a shorter and lighter body mostly received a position of PG with a range of height 70 to 75 inches, while having weight from 150 to 210 pounds. As the players height and players weight increases, so does the change in their position. It goes from the shortest and lightest to the tallest and heaviest -- position G, SG, SF, F, PF, C. position. The scatterplot proves as a fact that there is a relationship between the height and weight of a player to the position that he/she would be in. A center should probably have a standing reach of at least 80 inch. The tallest and heaviest player would most likely be in the center position. This can also be proven in real life, for instance, the tallest player in the dataset -- Gursimran -- is in the range of height 82 inches from above and he is also in the center position, as agreed to the assumption from the scatterplot. Those measurements only include

out the ideal body height and weight required for each position of the team. This will help aspiring basketball athletes to assess themselves and figure out the areas they can improve on.

III. Description of Data

The most important features of this project is the height and weight of each player and the positions they hold in the team, hence missing values may cause errors in our conclusion. Therefore, we decided to exclude data points where either values are missing. We also filled all the NaNs with 0 for all integer columns. We decided to clean our data in this way since we will only be using the top priority features. Since the weight and height of the players and the position are the core features of the experiment, these were the first features we cleaned. The other features that were all integer columns which we filled in 0 for all empty cells.

As we can see from our scatterplot, players with a shorter and lighter body mostly received a position of PG with a range of height 70 to 75 inches, while having weight from 150 to 210 pounds. As the players height and players weight increases, so does the change in their position. It goes from the shortest and lightest to the tallest and heaviest -- position G, SG, SF, F, PF, C. position. The scatterplot proves as a fact that there is a relationship between the height and weight of a player to the position that he/she would be in. A center should probably have a standing reach of at least 80 inch. The tallest and heaviest player would most likely be in the center position. This can also be proven in real life, for instance, the tallest player in the dataset -- Gursimran -- is in the range of height 82 inches from above and he is also in the center position, as agreed to the assumption from the scatterplot. Those measurements only include

approximately 75% accuracy of the available players that can be seen from the model , and therefore are not 100% accurate. This is because there is a limitation of a scatter plot to depict the data, because some of the points are overlapped and therefore cannot be seen. Hence, this scatterplot only holds a generalized idea on what the height and weight of a player is in relation with his position.

In order to figure out the features we will be using, we used EDA as described below and pairplot to discover any possible correlations between features. The histogram diagonal of the matrix allowed us to see the distribution of a single variable while the scatter plots in the rest of the matrix show the relationship between the two variables. How we used the pairplot will be discussed further below.

IV. Description of Methods

We first loaded in all the necessary libraries for this project into our notebook, called pd.read_csv on all of our dataset to be used in our analysis and performed the data cleaning mentioned above. We then performed EDA, produced data visualizations, cross-validated our data and built a model in the end and checked the accuracy of each model. The paragraphs below will explain these steps in more detail.

Our initial goal was to prove that the height and weight of the players and their win rates would influence the position they would hold in the team. However, we found out that that was untrue. Therefore we decided to create a pairplot that displayed features that correlate with each other. We found that the 3-point field goals attempted per game (NCAA__3ptapg) and 3-point field

seen if there is a correlation between each combination of the two features. The figure below shows a section of the pairplot described previously:

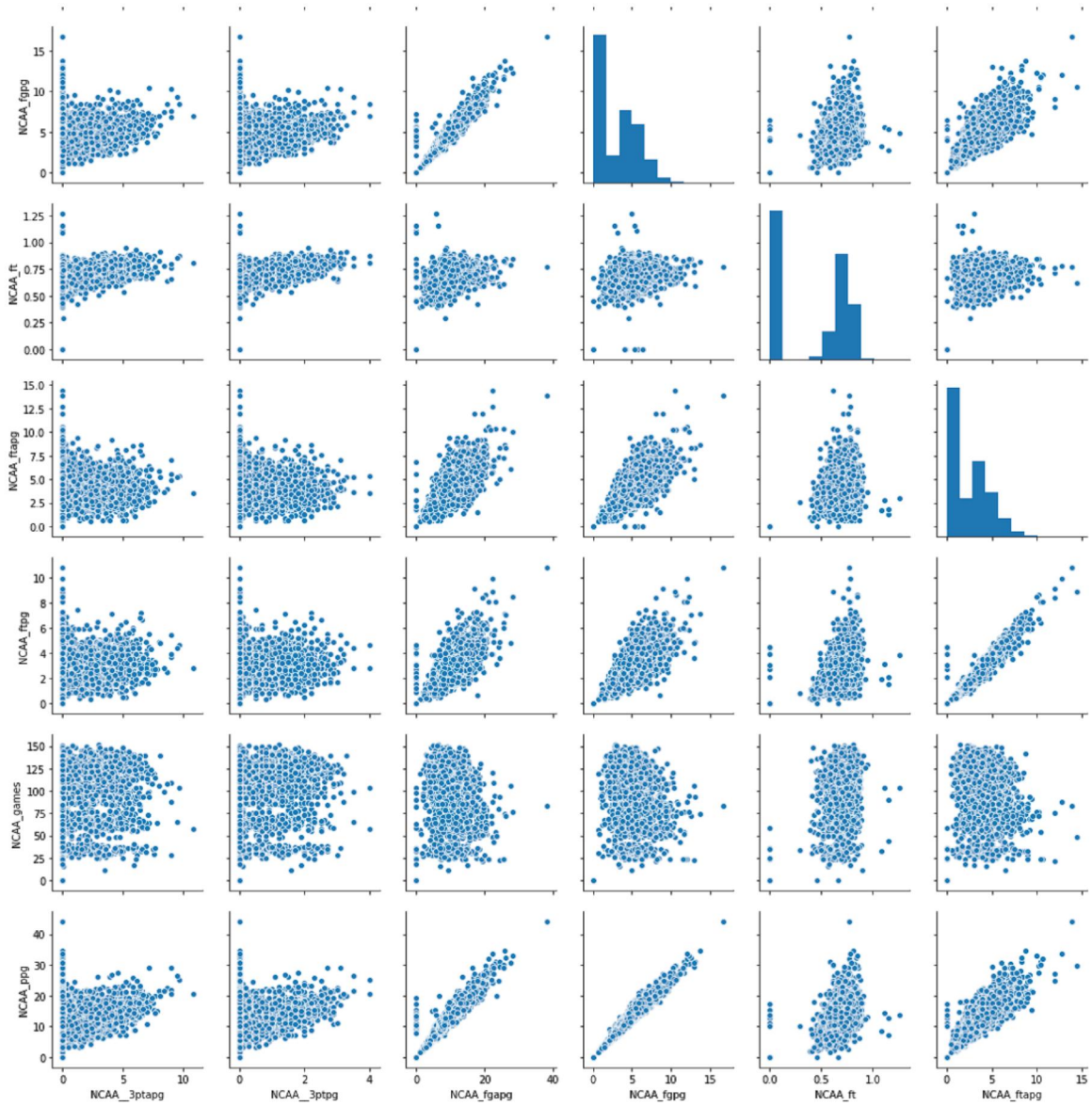


Figure 1. Snippet of Pairplot used to Determine which Features to Use

After cleaning our data and determining which features to use in our analysis, we modeled our data using Logistic Regression and Decision Tree, and compared their accuracies to determine

the best one. Before modelling, we split our data to training data and testing data using the train_test_split() method, where we took 20% of the data as test data. We scaled our data using StandardScaler(), which allows our data to have a mean of 0 and a standard deviation of 1, and is essential to reach the correct conclusion. We used LabelEncoder() on our position column, a categorical variable, assigning each position to a number in order to convert it into a model-understandable numerical data. We then used the split training and testing data for our models. The table below outlines all the models we used and the accuracy score we received.

	Training Accuracy	Testing Accuracy
Logistic Regression	64.0%	63.8%
Decision Tree	79.6%	79.3%
Random Forest	79.5%	79.8%

Table 1. Different Models Used and Accuracies of each Model

Logistic Regression on nba_train

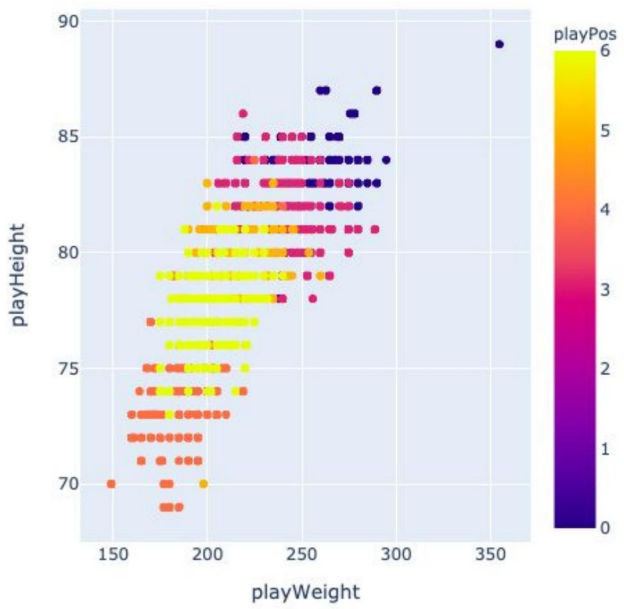
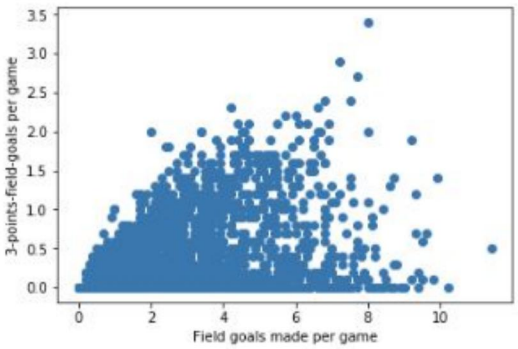
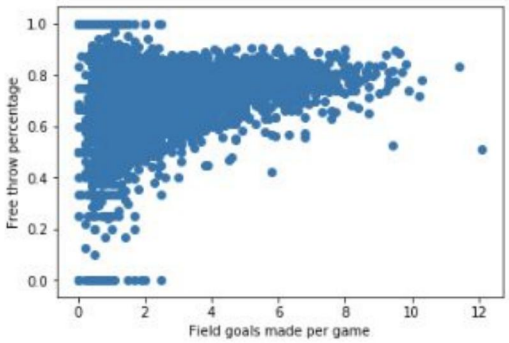


Figure 2. Logistic Regression on Relationship between Height and Weight with Position

features have a linear relationship with this feature. Some of the most interesting features that we came across is that “Field goals made per game” and “3 points field goals per game” will have a high correlation, but from the scatter plot it is clear that there is no relationship between these two, concluded from the scatter plot. Another one is the number three throws that we assumed will impact the field goals made per game, but in contradiction there is little to no relation with this two.





<u>(ii) Describe one feature you thought would be useful, but turned out to be ineffective.</u>

We thought the starting year they became active and the end year would be useful to calculate the number of years the athletes were active. We thought the years they were active would signify the experience they have in the business and experience would reflect their position, since we thought basketball was purely based on skills and nothing to do with body size.

<u>(iii) What challenges did you find with your data? Where did you get stuck?</u>

A challenge we found was when we had to clean our ‘height’ feature in the college dataset. The column was stored in ‘feet-inches’ format, which one of our members was able to convert it to only inches. However, for some reason, the same method did not work for the others and we

approximately 75% accuracy of the available players that can be seen from the model , and therefore are not 100% accurate. This is because there is a limitation of a scatter plot to depict the data, because some of the points are overlapped and therefore cannot be seen. Hence, this scatterplot only holds a generalized idea on what the height and weight of a player is in relation with his position.

In order to figure out the features we will be using, we used EDA as described below and pairplot to discover any possible correlations between features. The histogram diagonal of the matrix allowed us to see the distribution of a single variable while the scatter plots in the rest of the matrix show the relationship between the two variables. How we used the pairplot will be discussed further below.

IV. Description of Methods

We first loaded in all the necessary libraries for this project into our notebook, called pd.read_csv on all of our dataset to be used in our analysis and performed the data cleaning mentioned above. We then performed EDA, produced data visualizations, cross-validated our data and built a model in the end and checked the accuracy of each model. The paragraphs below will explain these steps in more detail.

Our initial goal was to prove that the height and weight of the players and their win rates would influence the position they would hold in the team. However, we found out that that was untrue. Therefore we decided to create a pairplot that displayed features that correlate with each other. We found that the 3-point field goals attempted per game (NCAA__3ptapg) and 3-point field

goals scored per game (NCAA__3ptpg) are positively correlated. This means that the teams in the dataset are very good in the sport since most, if not all, of the shots attempted actually got in. We also discovered that the field goals attempted per game (NCAA_fgapg) is positively correlated to the number of free throws attempted per game (NCAA_ftapg). Although this correlation has slightly more variation in the correlation, the features can still be considered as positively correlated. Thus we wanted to prove that the weight and height of the players and the aforementioned features influences the position they hold in the team.

We used exploratory data analysis (EDA) to find potentially useful patterns in our data from our College dataset for our analysis. In addition to just the body types of the players, we decided to include other factors such as certain skills and specific points scored. From our pairplot visualization, we can clearly see any potential correlations that may exist in our dataset and the combinations of various variables in each plot. If we use the selected columns and rows from the College table, we could not make much of the relationship between the variables and are limited to only a line-by-line understanding of these numbers individually. Pairplot is used in our first step to allow us to take in the data from the College table and discover any relationships that might exist between features, and also see the spread of each data point. Through this dataset, we selected a few features - 3-point field goals attempted per game, 3 point field goal per game, field goal per game, field goal attempted per game, free throw per game, free throw attempted per game, number of game played, points per game and visualizes if there is any relationship between each of these features. The matrix of relationships between each of the features can be

seen if there is a correlation between each combination of the two features. The figure below shows a section of the pairplot described previously:

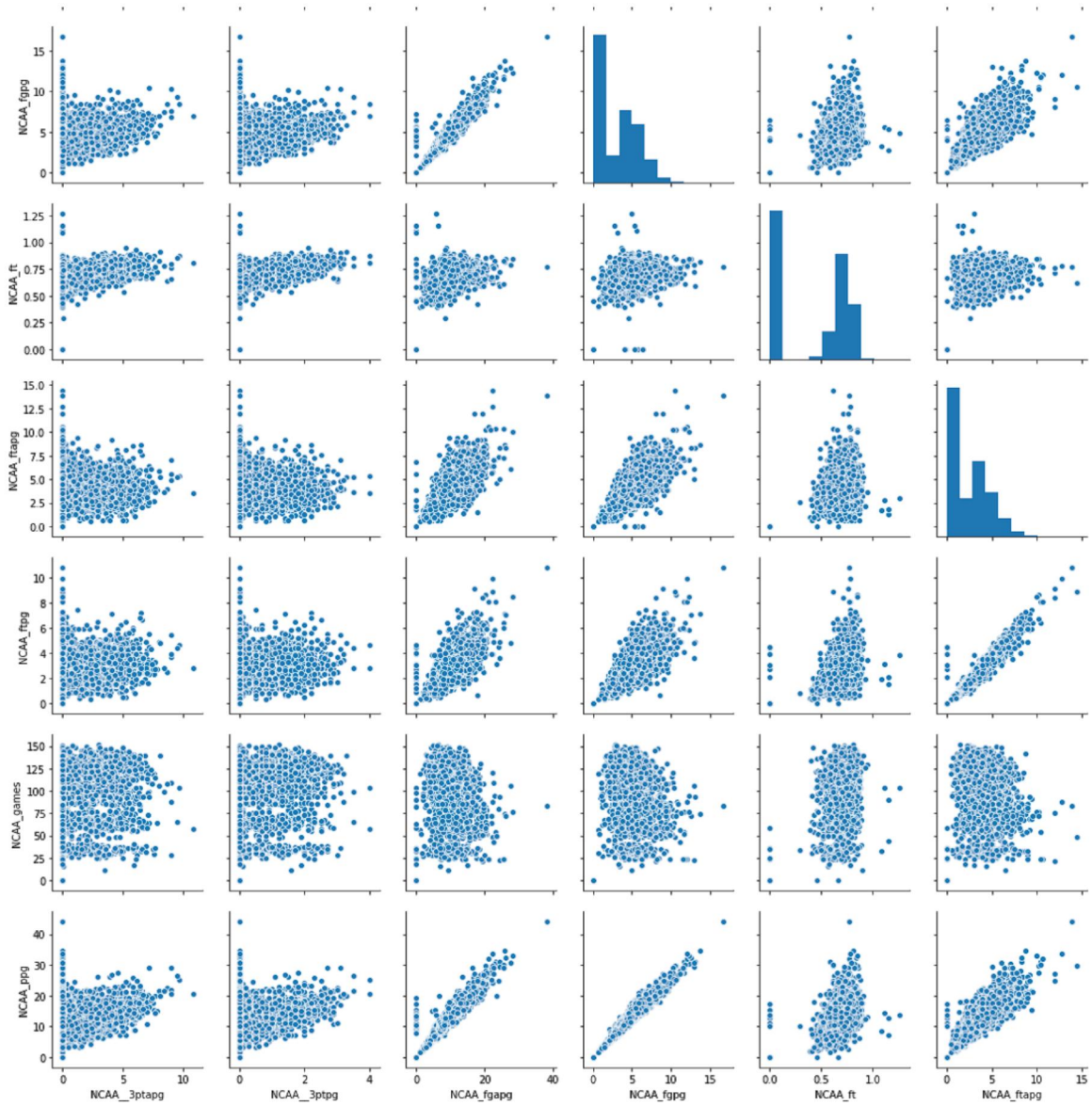


Figure 1. Snippet of Pairplot used to Determine which Features to Use

After cleaning our data and determining which features to use in our analysis, we modeled our data using Logistic Regression and Decision Tree, and compared their accuracies to determine

the best one. Before modelling, we split our data to training data and testing data using the train_test_split() method, where we took 20% of the data as test data. We scaled our data using StandardScaler(), which allows our data to have a mean of 0 and a standard deviation of 1, and is essential to reach the correct conclusion. We used LabelEncoder() on our position column, a categorical variable, assigning each position to a number in order to convert it into a model-understandable numerical data. We then used the split training and testing data for our models. The table below outlines all the models we used and the accuracy score we received.

	Training Accuracy	Testing Accuracy
Logistic Regression	64.0%	63.8%
Decision Tree	79.6%	79.3%
Random Forest	79.5%	79.8%

Table 1. Different Models Used and Accuracies of each Model

Logistic Regression on nba_train

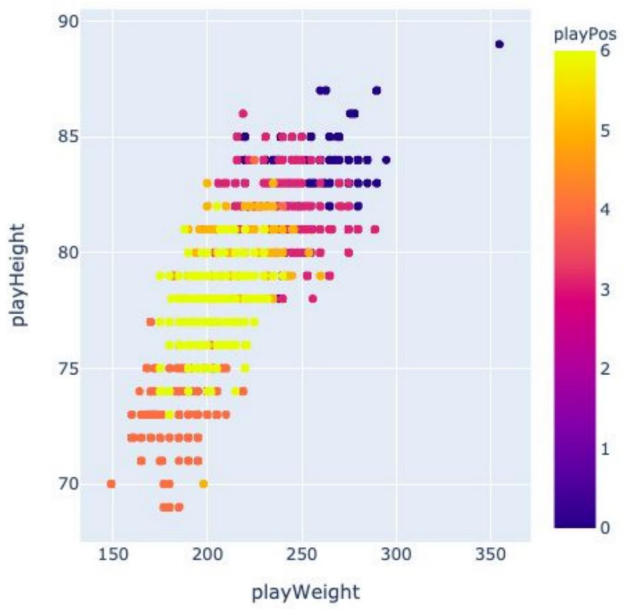


Figure 2. Logistic Regression on Relationship between Height and Weight with Position

From the table above, we concluded that the most accurate model would be to use Decision Tree. The correlation value between the field position with height and weight is -0.56 and -0.60. Since this is a multi-class classification, we are unable to use Logistic Regression because Logistic Regression is a binary classifier. Therefore, we used the Decision Tree model. An additional advantage of using a Decision Tree model is that it does not require feature scaling.

V. Summary of Results

Out of the three models, the decision tree performs the best on the training set's score of 79.6%, while the random forest classifier performs the best on the test's score of 79.8%. For the training and test accuracy of the three model, they vary from 0.15 to 0.35. For the decision tree, it has training error greater than test accuracy by 0.35, while for the random forest, it has training error greater than test accuracy 0.26. On the other hand, the logistic one versus rest, the training accuracy is greater than test accuracy by only 0.15.

From the decision tree and the random forest, we can see that the classes are not split by lines, and also that the decision boundaries are more spread out among the data, showing that the trait is more fitted to the training data. Therefore, we get such a high training scores, and also a relatively high test score. However, since these two classifiers are more fitted to the training data, test scores are much lower than the training scores.

From the table above, we concluded that the most accurate model would be to use Decision Tree. The correlation value between the field position with height and weight is -0.56 and -0.60. Since this is a multi-class classification, we are unable to use Logistic Regression because Logistic Regression is a binary classifier. Therefore, we used the Decision Tree model. An additional advantage of using a Decision Tree model is that it does not require feature scaling.

V. Summary of Results

Out of the three models, the decision tree performs the best on the training set's score of 79.6%, while the random forest classifier performs the best on the test's score of 79.8%. For the training and test accuracy of the three model, they vary from 0.15 to 0.35. For the decision tree, it has training error greater than test accuracy by 0.35, while for the random forest, it has training error greater than test accuracy 0.26. On the other hand, the logistic one versus rest, the training accuracy is greater than test accuracy by only 0.15.

From the decision tree and the random forest, we can see that the classes are not split by lines, and also that the decision boundaries are more spread out among the data, showing that the trait is more fitted to the training data. Therefore, we get such a high training scores, and also a relatively high test score. However, since these two classifiers are more fitted to the training data, test scores are much lower than the training scores.

VI. Discussion

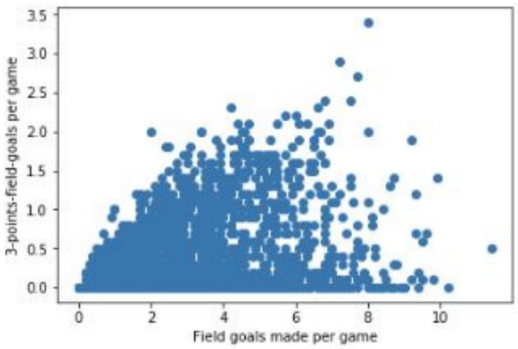
The tree model shows us that the position of a player on a team is influenced by the height and weight of the player, and certain skills they might have. Contrary to what some people might think, body size is not the only thing that matters for basketball players. These points have always been heavily discussed by basketball analysts and enthusiasts. Through this project, we hoped to address the issue and stereotype that all basketball players must have a certain body size in order to succeed. As an example, Steph Curry secured his place as one of the best point guards in the basketball court, despite having a shorter than average height for a basketball player.

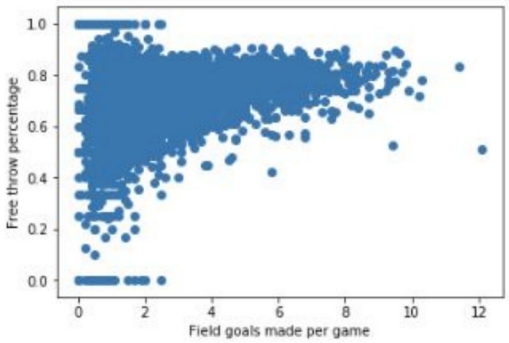
(i)What were two or three of the most interesting features you came across for your particular

Question?

Actually, if we look into the details of different games we can observe that whether a team will win or lose is dependent on many factors. We all know that people say that winning or losing is dependent on how the players perform but statistics tell us a different story. From our analysis, the performance of a team is based upon many other factors. We tried to justify the winning rate by looking from a player's position perspective. So, we tried to find features that affect the player's position, which are the height and weight of the player. This is the most-interesting conclusion which was drawn from the data after going through so many data visualizations from the dataset, and therefore we raised a particular discussion on the topic," What are the ideal height and weight at each position. Besides answering this question, since we know that having a higher 3 points field goals per game increases the winning rate, we would like to find which

features have a linear relationship with this feature. Some of the most interesting features that we came across is that “Field goals made per game” and “3 points field goals per game” will have a high correlation, but from the scatter plot it is clear that there is no relationship between these two, concluded from the scatter plot. Another one is the number three throws that we assumed will impact the field goals made per game, but in contradiction there is little to no relation with this two.





<u>(ii) Describe one feature you thought would be useful, but turned out to be ineffective.</u>

We thought the starting year they became active and the end year would be useful to calculate the number of years the athletes were active. We thought the years they were active would signify the experience they have in the business and experience would reflect their position, since we thought basketball was purely based on skills and nothing to do with body size.

<u>(iii) What challenges did you find with your data? Where did you get stuck?</u>

A challenge we found was when we had to clean our ‘height’ feature in the college dataset. The column was stored in ‘feet-inches’ format, which one of our members was able to convert it to only inches. However, for some reason, the same method did not work for the others and we

tried finding the reason behind it and found out that the 'height' column was formatted as 'date-month' format. We were unable to solve this, so we decided to use the 'playHeight' feature from the playerBoxScore dataset. We also got stuck when we tried assigning numbers to the positions. We initially thought of making a dictionary to store these numbers, but after researching, we found out about LabelEncoder() that does this instantly.

(iv) What are some limitations of the analysis that you did? What assumptions did you make

that could prove to be incorrect?

The dataset was best for time series analysis because some of the datasets did not contain a unique key. There were multiple values for the same entities. We have used scatter plots to show that the total number of successful shots does not depend upon the 3-point shoots, which contradicts what our initial assumptions are. Similarly, we have tried to show the relations between different shoot methods and the actual points scored by the team, which could be inferred from the pairplot method itself. There were many different factors that could change the outcome of each game, like the type of ground in which they are playing.

(v) What ethical dilemmas did you face with this data?

A team's ability to win should not change depending the location in which they are playing. However, when we were still playing around with the data, we found out that the teams are more likely to win when they are playing in their home base. After reading information about the possible reasoning behind this, we found out that break rooms for visiting teams are often made

intentionally smaller than normal, extremely hot during summer and extremely cold during winter. Additionally, the mentality of the players are also influenced since they are often jeered at and treated with disdain in the opposing team's home base. This raises the question as to the methods a team would go through to ensure their victory.

(vi) What additional data, if available, would strengthen your analysis, or allow you test some

other hypothesises?

Our hypothesis could be further developed by taking into account the skills a player has using a fixed metric. Although our analysis took into account some of the skills these players have, we did not completely explore it and a fixed metric would make our predictions more accurate.

(vii) What ethical concerns might you encounter in studying this problem? How might you address those concerns?

Since we are focusing on the physical features of the players, people might think that these are the standards of the industry. Consequently, people may simply give up on their dreams if they do not meet these standards or they might force their bodies to have these standards as well. They might have to undergo extremely strict and unreasonable diets and workouts in order to improve their chances of winning even by the slightest. Although we were not able to prove this concern, we are sure that there are some people out there that will be willing to do this for a greater chance of winning or getting noticed by an NBA coach. Despite the physical characteristics, there is more to an athlete than just their body. They need to also be able to make

split-second decisions that can turn a game around. Hence, a person that has an ideal body will not necessarily be a good player.