**Summary:**

I looked at the past 10 National Basketball Association (NBA) seasons (2012 – 2022). For each of these seasons, I looked at the players who were in the top 50 in total points scored. From here I examined the players who were in the top 50 in total points scored for every season within the 10-season time frame.

The result was three players: LeBron James, James Harden, and DeMar Derozan. For each of these players I accumulated statistics for games played, minutes played, field goal attempts, and points scored.

From these statistics, I charted the following:

* For each of the three players I charted, via line graphs, their total points scored vs each season, minutes played vs each season, games played vs each season, and field goal attempts vs each season.

*See nba\_project > Statistics > Individual Players > FGA VS Season*

*See nba\_project > Statistics > Individual Players > Games VS Season*

*See nba\_project > Statistics > Individual Players > Minutes VS Season*

*See nba\_project > Statistics > Individual Players > Points VS Season*

* For all players I calculated and plotted via scatter plots their average total field goal attempts per season, average total games played per season, average total minutes played per season, and average total points per season over the 2012-2022 timeframe.

*See nba\_project > Statistics > All Players > Player Averages*

* For all players I calculated and plotted via scatter plots their total field goal attempts, total games played, total minutes played and total points over the 2012-2022 timeframe.

*See nba\_project > Statistics > All Players > Player Totals*

* For each season and for all players I charted via scatter plots points vs field goal attempts, points vs games played, and points vs minutes played.

*See nba\_project > Statistics > All Players > Points VS FGA*

*See nba\_project > Statistics > All Players > Points VS Games*

*See nba\_project > Statistics > All Players > Points VS Minutes*

* For each of the three players, I charted a correlation heat map, comparing correlation of points, field goal attempts, minutes played, and games played for each player.

*See nba\_project > Heat Maps > Individual Players*

* For all players, I charted a correlation heat map, comparing correlation of points, field goal attempts, minutes played, and games played for all players.

*See nba\_project > Heat Maps > All Players*

From the data accumulated in data\_analyis.py for my final project I performed the following:

* Regression, both linear and quadratic
* Random Forest
* Decision Tree

Regression:

I used regression to predict total points based off the average field goal attempts, the average minutes played, and the average games played for each player. I also predicted total points based off a 5% increase and decrease of these averages. Additionally, I predicted total points based off last season’s total field goal attempts, minutes played, and games played, as well as for a 5% increase and decrease in these categories. For all the above mentioned, I performed both linear and quadratic regression, with all results, rmse, and r-squared printed to the system for each player and model.

For each model I created a chart for each player, and for each prediction grouping I created a chart for each player.

* The regression models charts show the regression model used with a scatter plot of actual points scored and a line plot for the regression line.

*See nba\_project > Regression > Regression Models > Points VS FGA*

*See nba\_project > Regression > Regression Models > Points VS Games*

*See nba\_project Regression > Regression Models > Points VS Minutes*

* The regression predictions charts show the predicted values for points scored based on the model used with a horizontal bar chart displaying all predictions for the specified model.

*See nba\_project > Regression > Regression Predictions > Points VS FGA*

*See nba\_project > Regression > Regression Predictions > Points VS Games*

*See nba\_project > Regression > Regression Predictions > Points VS Minutes*

Random Forest:

I used the random forest classifier to predict a range of points for next season based off each player’s field goal attempts, games played, and minutes played for each season. For each season, I assigned labels for below or above a certain point threshold. I worked in increments of 100 points starting at 1200 and ending at 2200. I used depths of 1-2 and subtrees of 1-5 and took the conglomerate of all these predictions to predict the player’s result at each point level. Additionally, at each point level I print the prediction (above or below) the point level, the accuracy of all predictions at that point level, the total amount of predictions, number of true positives, false positives, true negatives, false positives, and the true positive and true negative rates. If the model is undecided on if a player would score above or below a specified points level, the system will print this and the program acts conservatively in not identifying this as a positive result, but rather negative. Finally, based off each levels prediction, the system prints a 100-point range for predicted points scored, with the total accuracy of all predictions for the model.

For each player, at each point level, I plotted a scatter plot showing the predicted label for each subtree and depth combination for the random forest classifier.

*See nba\_project > Random Forest > DeMar Derozan*

*See nba\_project > Random Forest > James Harden*

*See nba\_project > Random Forest > LeBron James*

Decision Tree:

In addition to the Random Forest classifier, I used the Decision Tree classifier. The model will predict if a player will score above or below a certain points level based off of the labels assigned for that points level for the previous 10 seasons, similarly to that of the Random Forest classifier. The model will predict at each points level whether the player will score above or below that points level and provide the accuracy of the model at that level. Lastly, the model will provide a 100-point range for predicted points scored, with the total accuracy of all predictions for the model.

**Results:**

Based off the heat map for correlation of all players, field goal attempts should be the best indicator of total points scored in comparison to games or minutes, with a .95 correlation between points and field goal attempts. Interestingly, based off the heat map for LeBron James, he has a higher correlation between points and minutes played than games or field goal attempts, while the other players have a categorical correlation highest for field goal attempts in relation to points scored. (See Heat Maps folder)

When looking at the regression models, for each player, the linear (1 degree) model of field goal attempts vs points had the highest r-squared and lowest rmse for each player and would be the regression model I would be most confident in using to predict total points. While difficult, because it is hard to predict how many minutes, field goal attempts, and games played a player will play next season. The models yielded the following results (points rounded down to whole number).

**DeMar Derozan 1 Degree Points vs FGA (Average Points Per Season = 1674.3 *Over 10 season time frame, Last Season Points = 2118*)**

RMSE = 82.38651106763436, R-Squared = 0.9479667782610485

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Avg FGA | 5% Decrease Avg | 5% Increase Avg | Last Season FGA | 5% Decrease Last Season | 5% Increase Last Season |
| Points | 1673 | 1591 | 1755 | 2003 | 1904 | 2101 |

**James Harden 1 Degree Points vs FGA (Average Points Per Season = 2068.2 *Over 10 season time frame, Last Season Points = 1432*)**

RMSE = 40.42607011655252, R-Squared = 0.994972870960817

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Avg FGA | 5% Decrease Avg | 5% Increase Avg | Last Season FGA | 5% Decrease Last Season | 5% Increase Last Season |
| Points | 2064 | 1962 | 2165 | 1497 | 1423 | 1570 |

**LeBron James 1 Degree Points vs FGA (Average Points Per Season = 1801.7 *Over 10 season time frame, Last Season Points = 1695*)**

RMSE = 50.16978248156408, R-Squared = 0.9796064223715405

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Avg FGA | 5% Decrease Avg | 5% Increase Avg | Last Season FGA | 5% Decrease Last Season | 5% Increase Last Season |
| Points | 1798 | 1698 | 1898 | 1711 | 1615 | 1806 |

When looking at the random forest and decision tree classifiers, based off of the last runs and average field goal attempts, games played and minutes played.

**DeMar Derozan**

Decision Tree predicts between 1400 and 1500 points with a sum accuracy of 76.3636%

Random Forest predicts between 1500 and 1600 points with a sum accuracy of 71.4545%

**James Harden**

Decision Tree predicts between 1800 and 1900 points with a sum accuracy of 78.1818%

Random Forest predicts between 1900 and 2000 points with a sum accuracy of 73.4545%

**LeBron James**

Decision Tree predicts between 1600 and 1700 points with a sum accuracy of 83.6364%

Random Forest predicts between 1600 and 1700 points with a sum accuracy of 80.0%

**Conclusion:**

*Overall, the data set is somewhat small, and this leads to limitations, especially when looking at the Random Forest and Decision Tree classifiers. The player scoring from season to season can be somewhat volatile and depending on the run, the Random Forest and Decision Tree classifiers can predict results with high variances. When looking at player point outputs, given that this is over a ten-season time frame, it is hard to analyze a player’s potential output next season, with some players having much higher outputs in earlier seasons than in the more recent seasons, thus less likely to repeat these outputs in the later stage of their careers. For example, James Harden had very high point totals in earlier seasons and his point totals have diminished over the past few seasons. On the other end, DeMar Derozan had a very large point total last season in comparison to his average. Due to this, I would be more comfortable predicting next season’s points totals more heavily weighted towards the linear regression model tied to field goal attempts. For James Harden and LeBron James, I would focus more closely on last season’s statistics when making my prediction, while with DeMar Derozan, I would be more inclined to look at his averages when making my prediction.*