Springboard Capstone Project 2: Analysis of NBA Dataset

Michael Thabane

Kaggle Dataset Repository

August 7, 2020

Introduction

Business Problem

- In the last 30 years the NBA has been one of the fastest growing sports in the world
- With the advancement of sports analytics and the prevalence of gambling there is a high demand for player statistic predictions
- Problem: Creating a prediction model for player stats, in particular points; using previous game data
- Solution: Model the data using machine learning techniques to predict the number of points a player will score in an upcoming game

Approach

Data Acquisition and Wrangling

- Data was taken from the Kaggle Dataset Repository originally scraped from the NBA statistics website using the NBA API
- Dataset consists of four different spreadsheets:
  - First spreadsheet (games.csv) contains team boxscore data which consists of total team statistics.
  - The second spreadsheet (game_detail.csv) contains the player boxscore data which consists of individual player statistics for each game.
  - The third spreadsheet (players.csv) contains player information for which team they play for.
  - The fourth spreadsheet (ranking.csv) contains the nba standings for every day in the each season
- Have to manipulate data to create any variables that may have influence on the outcome of a players points.
- Created the running average to include the players average over the previous three and five games

Approach

Descriptive Analysis

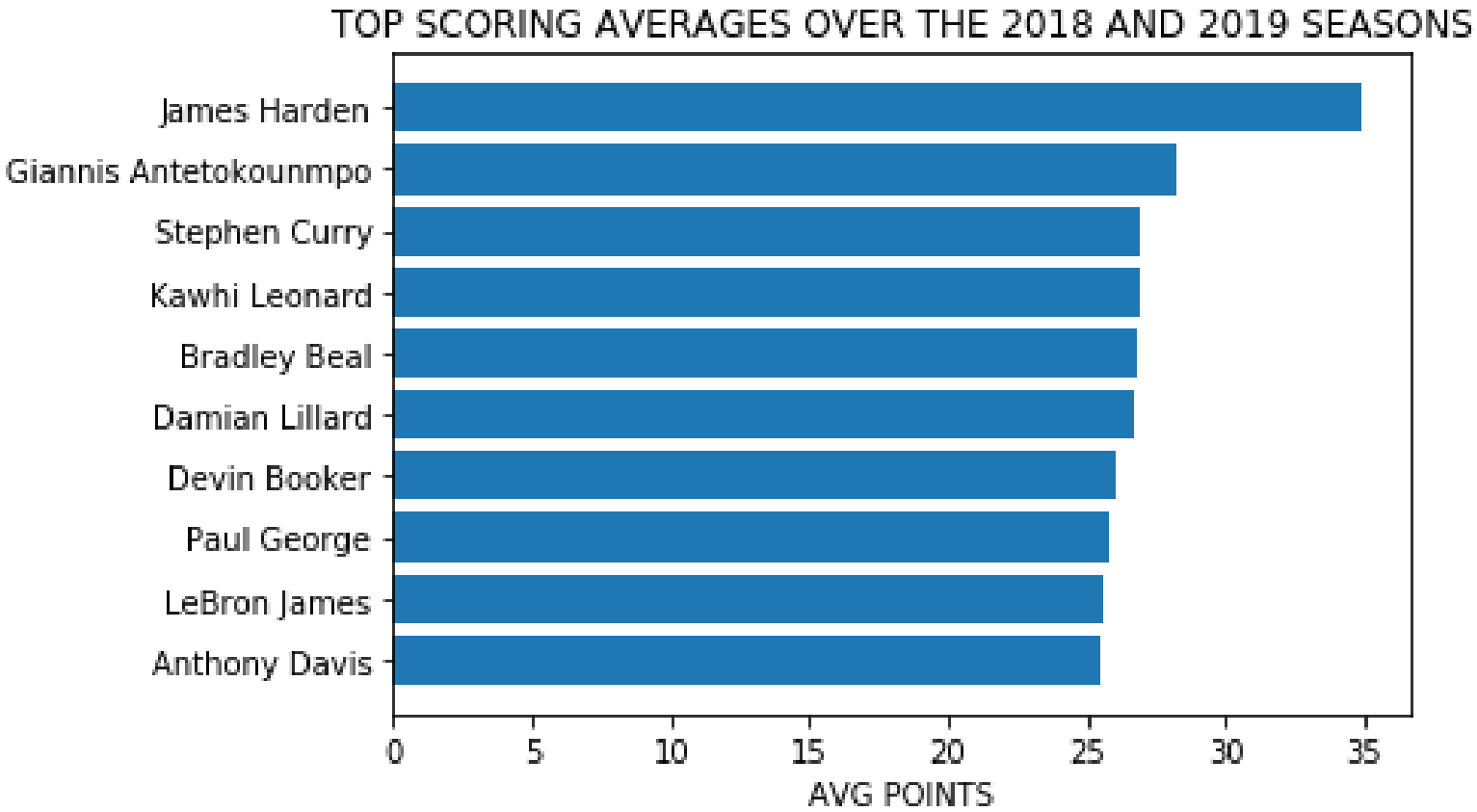


Figure: Top 10 Scorers Over 2018 and 2019 Seasons

Approach

Descriptive Analysis

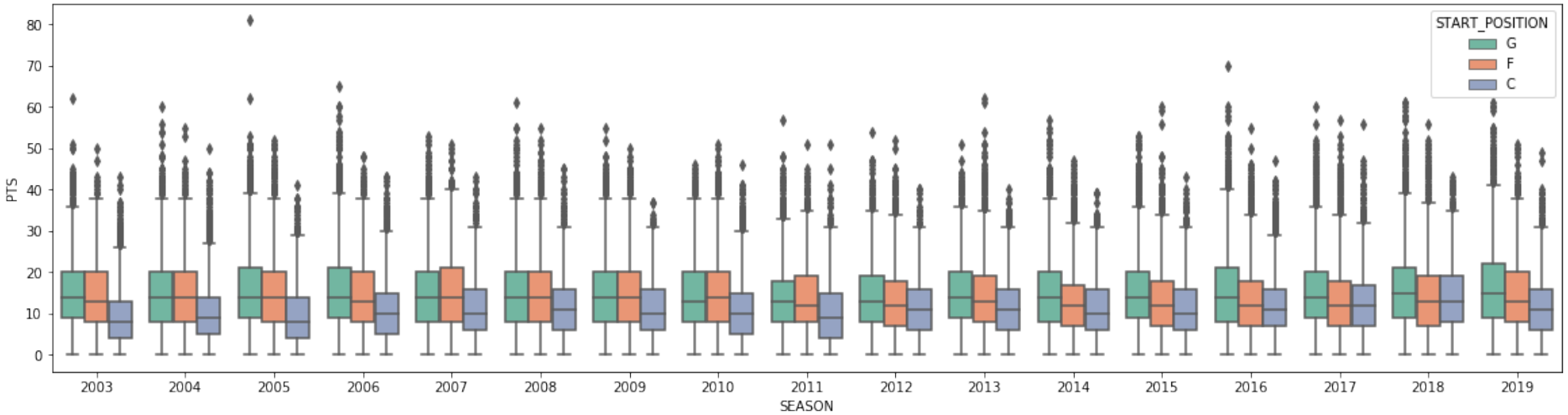


Figure: Boxplot of Starters Points

Approach

Descriptive Analysis

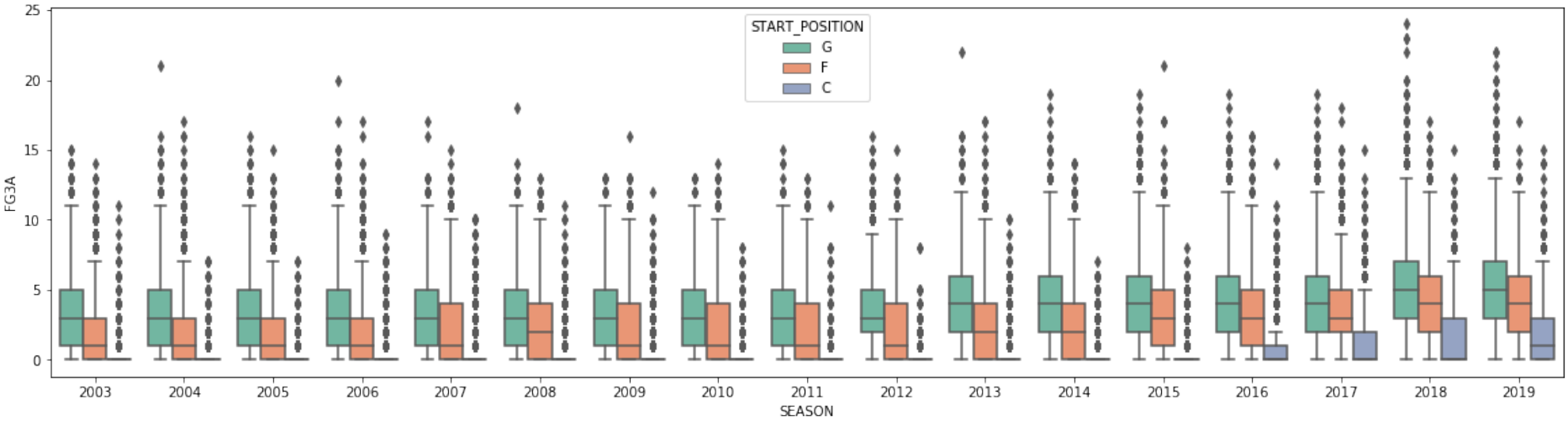


Figure: Boxplot of Starters 3PT Field Goals Attempted

Approach

Descriptive Analysis

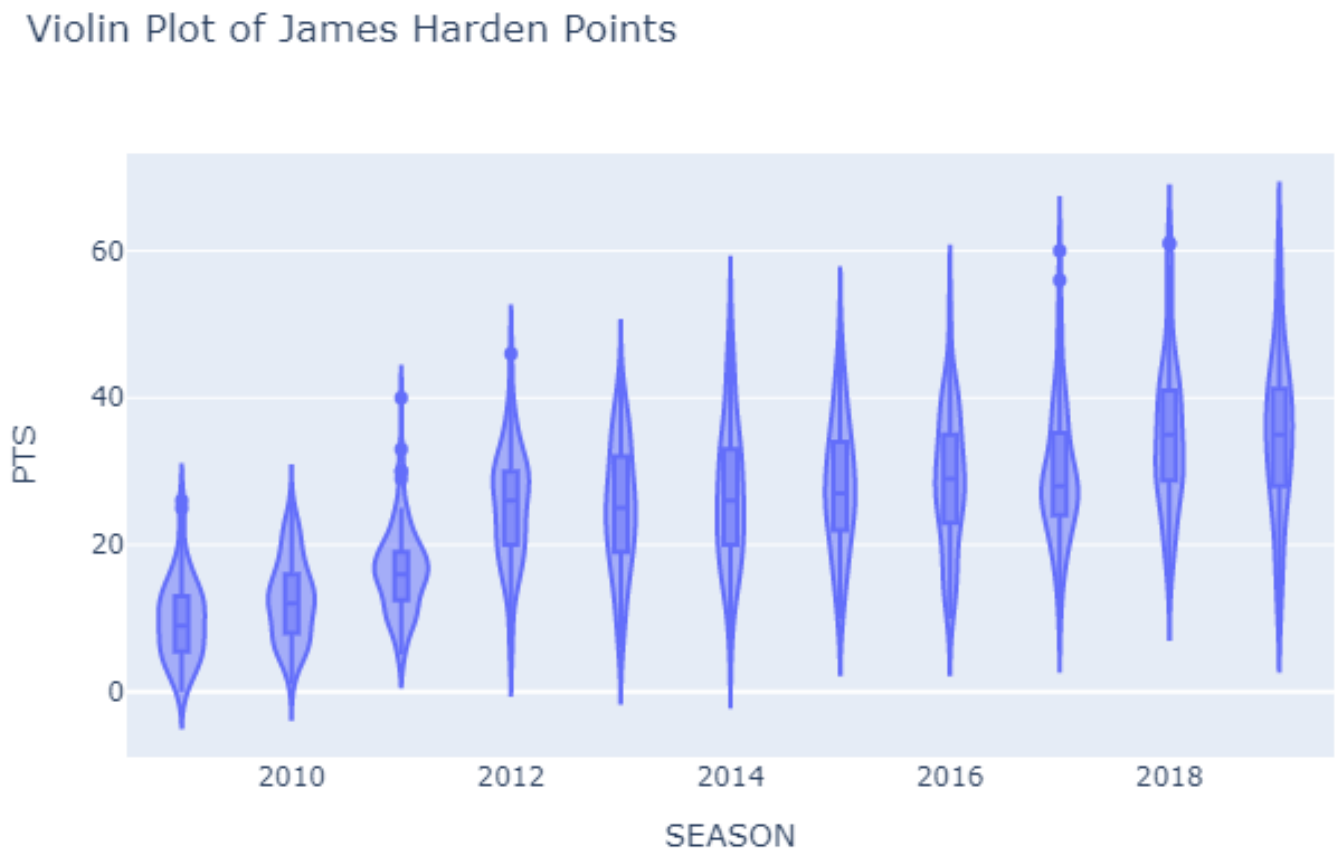


Figure: Violin Plot of James Harden

Approach

Descriptive Analysis

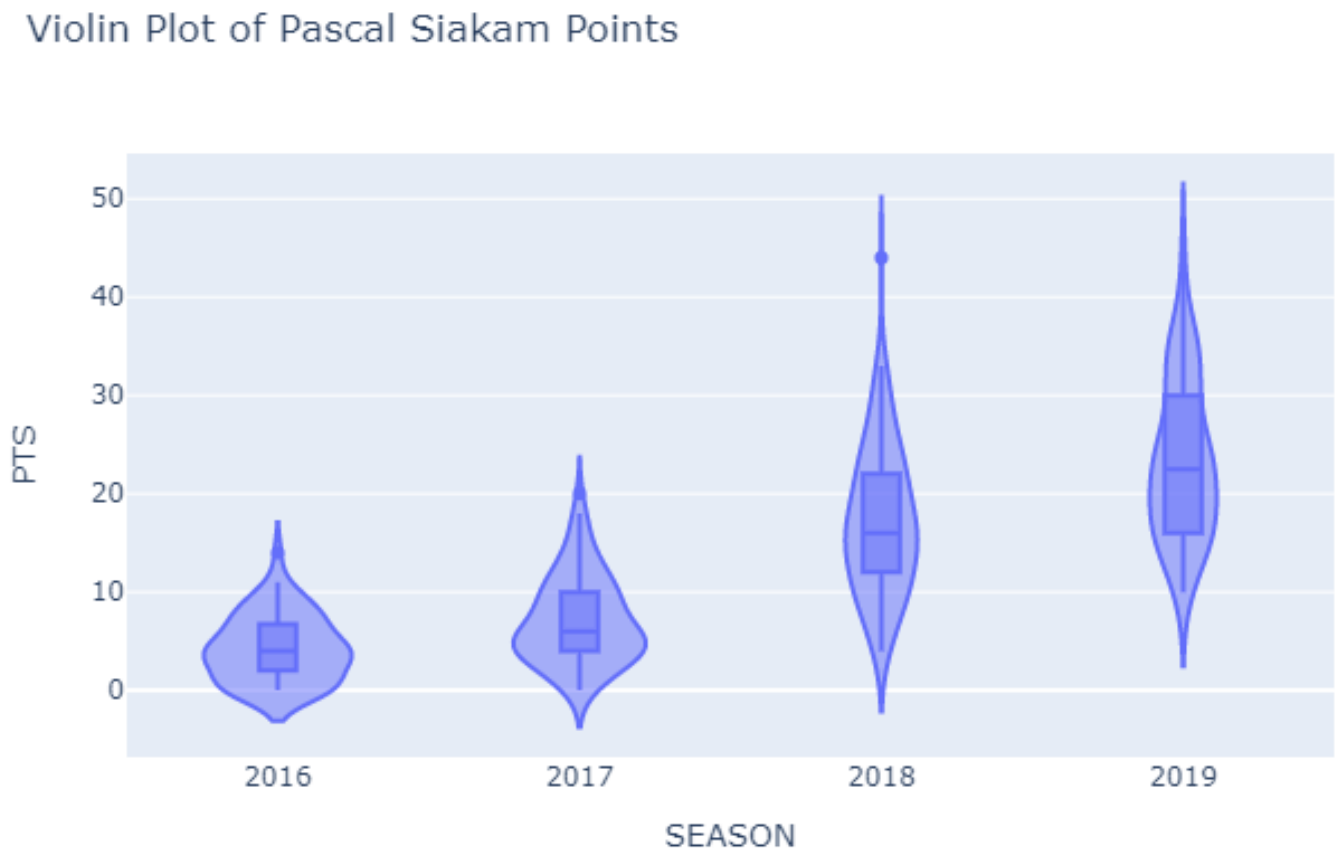


Figure: Violin Plot of Pascal Siakam Points

Approach

Descriptive Analysis

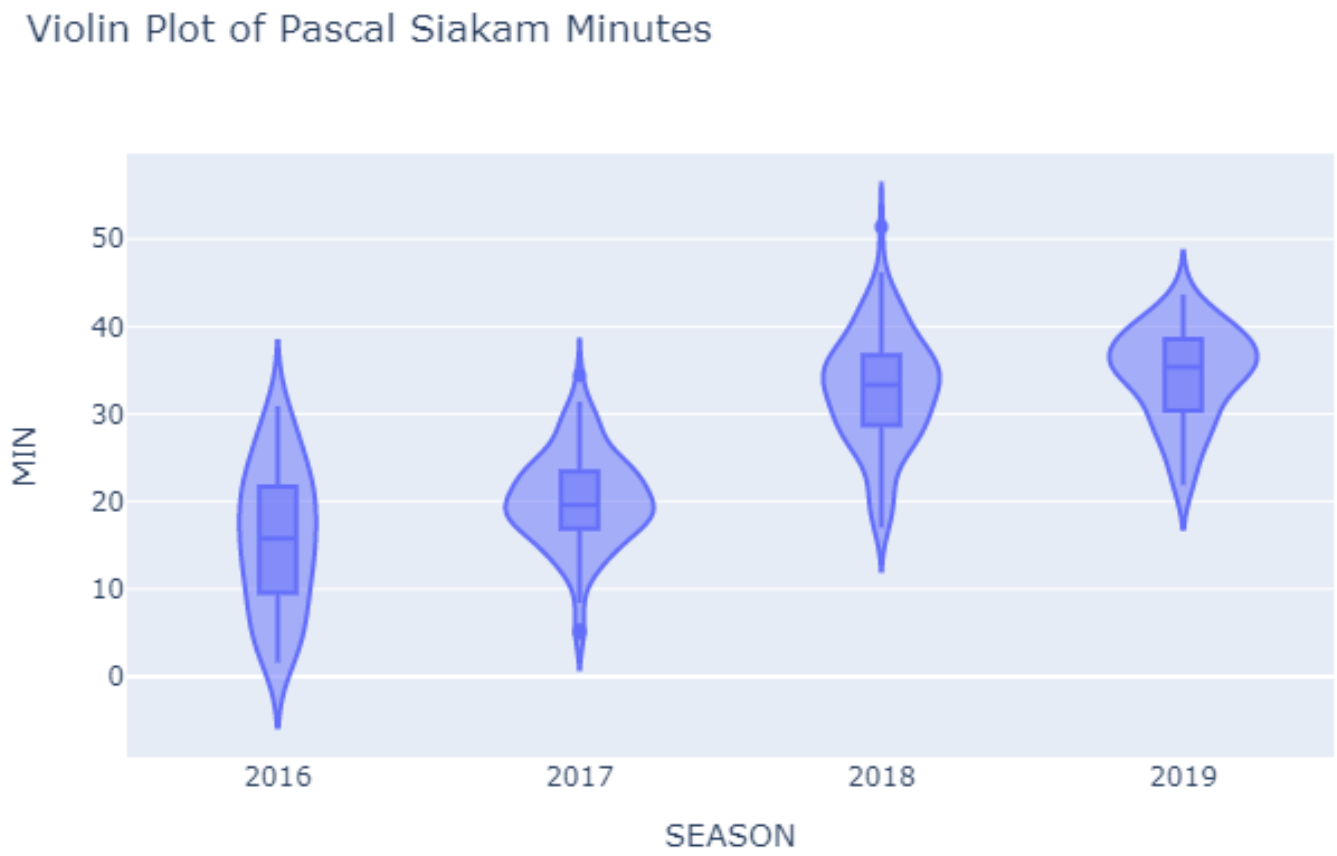


Figure: Violin Plot of Pascal Siakam Mins

Approach

Hypothesis Testing

- Test three hypotheses described below:
  1. The average points scored by playoff guards is equal to the average points scored by non-playoff guards
  2. The average points scored by playoff forwards is equal to the average points scored by non-playoff forwards
  3. The average points scored by playoff centers is equal to the average points scored by non-playoff centers
- There was significant evidence to reject all three hypotheses

Approach

Baseline Modeling

Model 1 - Lebron	Test - 2018 season
RMSE	7.753315379958686
r^2	-0.021090252249140296
MAPE	25.842116632470884

Approach

Baseline Modeling

Model 2 - Lebron	Test - Half 2018 season
RMSE	5.962926546520658
r^2	0.01502527934745001 6
MAPE	19.398868355111855

Approach

Baseline Modeling

Model 3 - Giannis	Test - Half 2019 season
RMSE	5.962926546520658
r^2	-0.04454537567408745
MAPE	23.254694325171055

Approach

Extended Modeling

Below are the tuned parameter values:

- n estimators = 700
- max depth = 5
- min samples split = 7
- min samples leaf = 11

Random Forest Model 1 - Lebron	Test - Half 2018 season
RMSE	6.069103756861786
r^2	-0.02036438087963477
MAPE	19.875482182850945

Approach

Extended Modeling

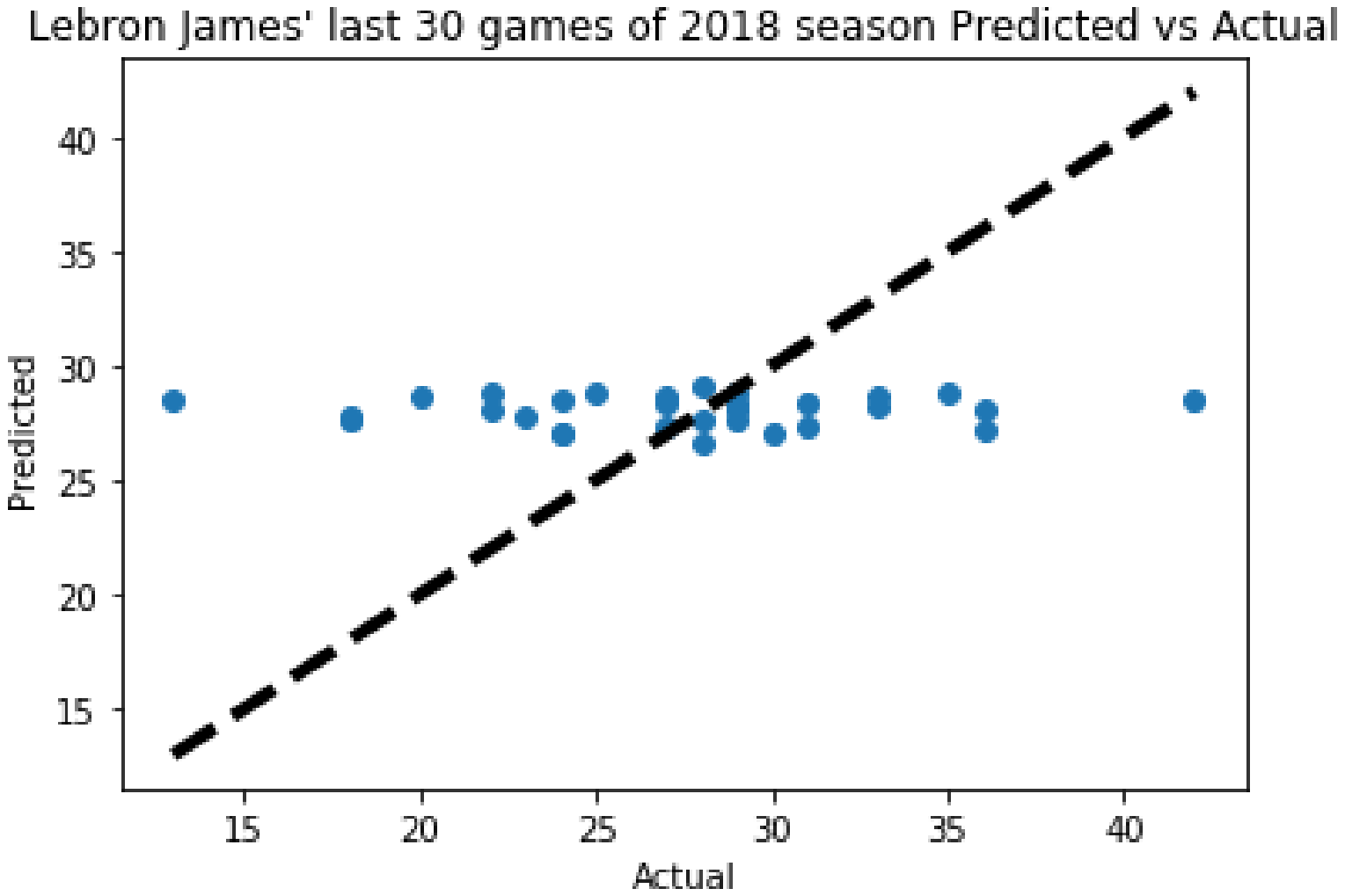


Figure: Lebron James' last 30 games of 2018 season Predicted vs Actual

Approach

Extended Modeling

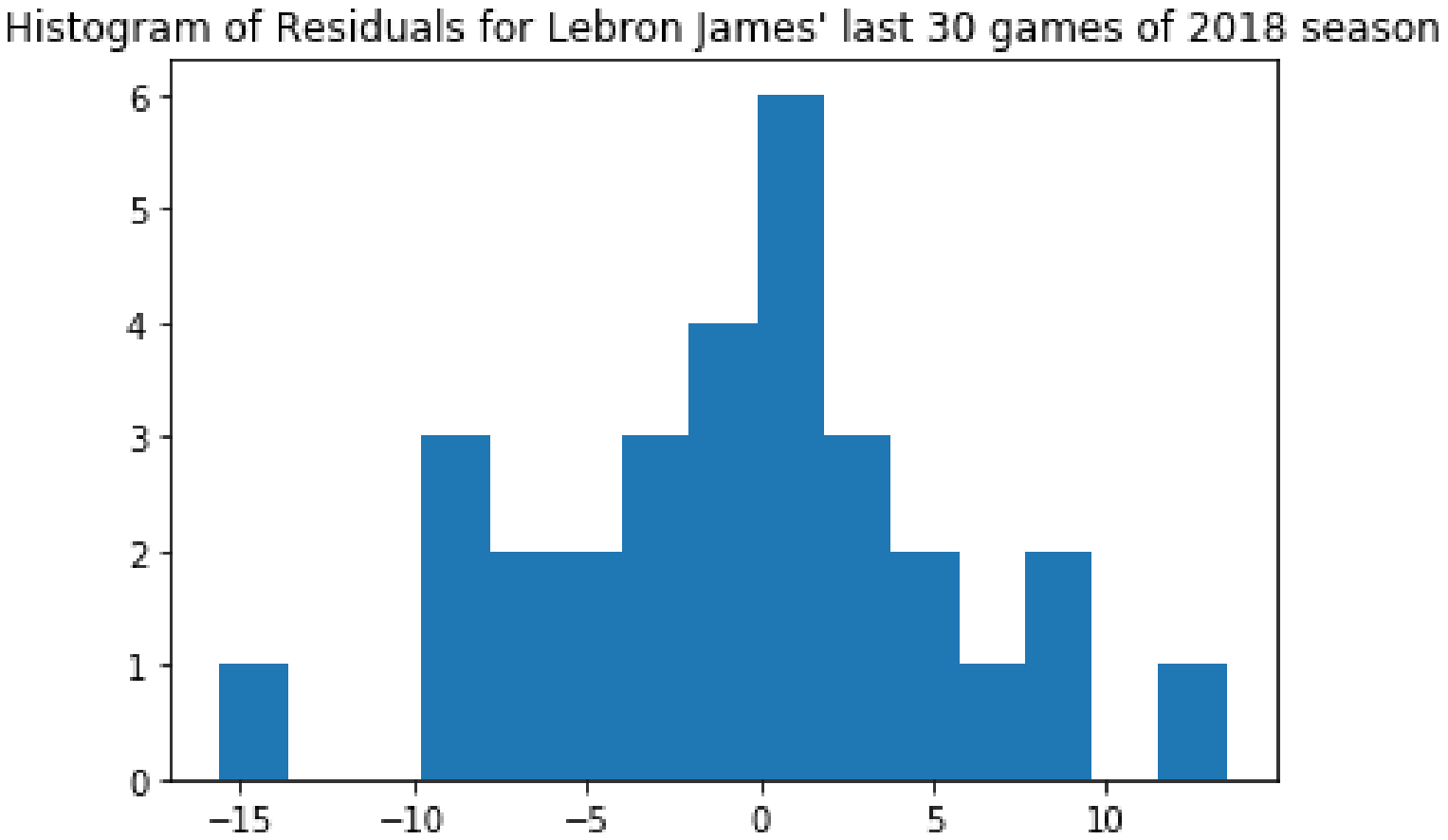


Figure: Histogram of Residuals for Lebron James' last 30 games of 2018 season

Approach

Extended Modeling

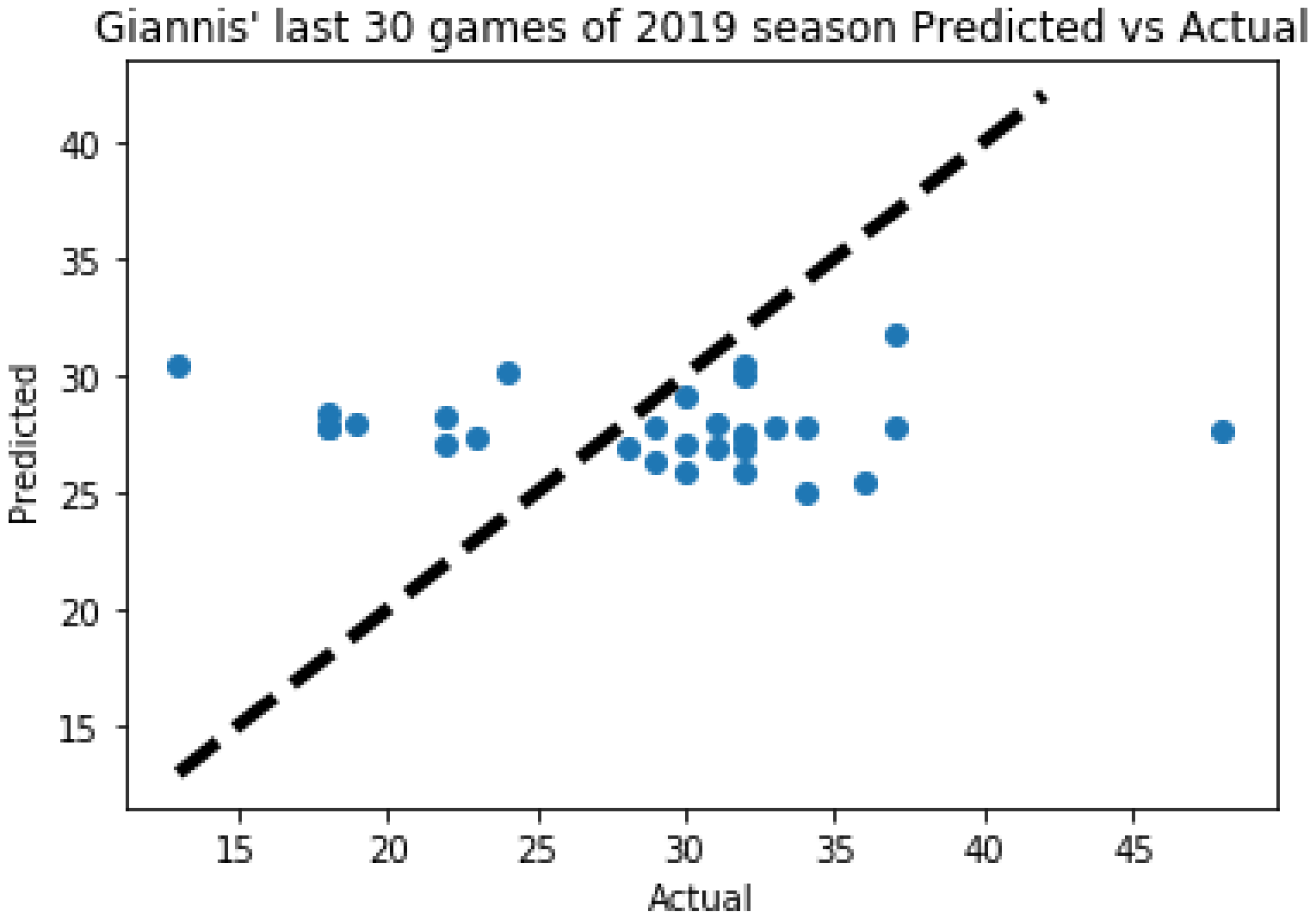
Below are the tuned parameter values for the model run using Giannis' data:

- n estimators = 100
- max depth = 5
- min samples split = 17
- min samples leaf = 3

Random Forest Model 2 - Giannis	Test - Half 2019 season
RMSE	7.667378481162914
r^2	-0.13970841582069826
MAPE	25.208935493860736

Approach

Baseline Modeling



Approach

Extended Modeling

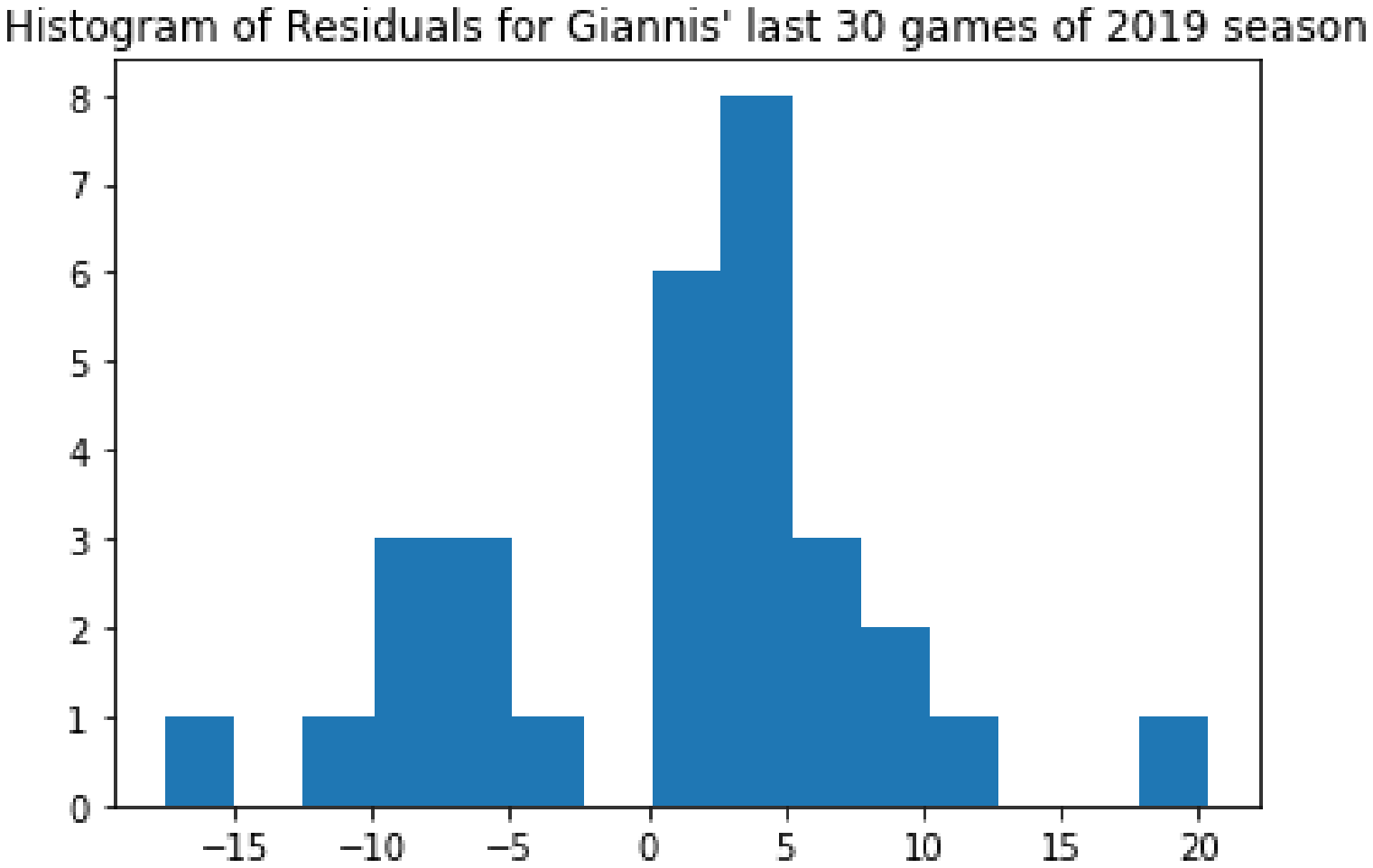


Figure: Histogram of Residuals for Giannis' last 30 games of 2019 season

Summary of Findings

Lebron Test - Half 2018 season

Statistic	Linear Regression	Random Forest
RMSE	5.962926546520658	6.069103756861786
r^2	0.015025279347450016	-0.020364380879634 77
MAPE	19.398868355111855	19.875482182850945

Summary of Findings

Lebron Test - Half 2018 season

Statistic	Linear Regression	Random Forest
RMSE	5.962926546520658	7.667378481162914
r^2	-0.04454537567408745	-0.13970841582069826
MAPE	23.254694325171055	25.208935493860736

Conclusions

- The Linear Regression model performs better than the Random Forest Regression model for both Lebron and Giannis' data
- Both models either slightly perform higher or lower than using the average number of points as a predictor for the number of points a player will score
- The time complexity of running the models for each player is of high order because you need to parameter tune the model for each players test set.

Future Work

- Since most of the models are close to performing as well as using the average number of points as a predictor I think it would be worth will to see if we could find trends using time-series model on player points.
- Implementing the model which performs the best on other statistics such as player assists and player rebounds

Recommendation for the Client

- My recommendation to the client would be to use only use point predictions that are high or lower than the posted point props by 6 points
- I would also recommend running running the code on large CPU in order to put it into production as the computing

Thanks