NBA writeup

Prepared by

Rania Almneie
Hanadi Alshahrani
Najd Alqahtani

The National Basketball Association is a professional basketball league in North America.

To measure the performance of their players and predict their future performance, statistic will help us to determine that.

Abstract

The National Basketball Association is a professional basketball league in North America. To measure the performance of their players and predict their future performance, statistic will help us to determine that.
So, we'll use NBA season to build a basic linear regression model to predict number of points for players.
We obtain the data from the NBA website using web scraping, then apply the EDA to this data, and it is now ready to predict our target using a linear regression model for more details, refer to the following components:

Design

We wanted as a team to analyze the performance of NBA players and predict their performance in the future based on the data of the current season.

Algorithms

The first step to achieve our goal, use beautifulSoup and selenium for web scraping to get data from the NBA website, then followed EDA which included data exploration, cleaning, handling the outliers and visualization data using seaborn, matplotlib and tableau.
After that build a linear regression model, we use sk-learn.

Data

Data was extracted from the NBA website for players composed of 536 observations and 7 features which are:

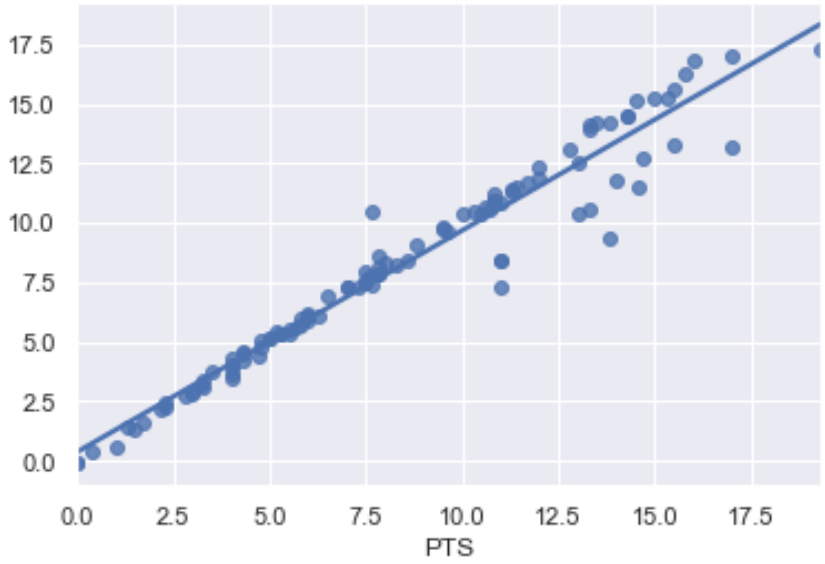
- PLAYER: Player names.
- MIN (Minutes Played): The number of minutes played by a player or team.
- PTS (Points): The number of points scored.
- FGM (Field Goals Made): The number of field goals that a player or team has made. This includes both 2 pointers and 3.
- 3PM (3 Point Field Goals Made): The number of 3 point field goals that a player or team has made.
- FTM (Free Throws Made): The number of free throws that a player or team has made.
- FP (Fantasy Points): The number of fantasy points a player accumulates.

The y-variable will be the number of points (PTS), and the x-variable will be determined by the best model.

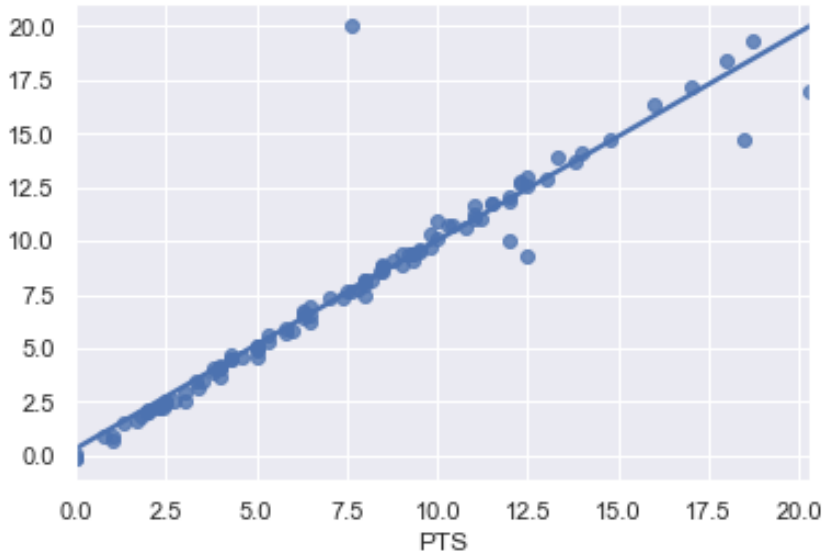
Communication

Based on the modeling results, we consider Model 6 is the best of them depend on R Squared, MAE, MSE and RMSE, which includes all features in the x-variable and PTS in the y-variable.

We are plotting in validation set and test set to make sure the model 6 predicted is good and we achieved that.



Validation dataset



Test dataset

Tools

- Pandas
- NumPy
- Sklearn
- Seaborn
- Tableau
- matplotlib
- Selenium
- BeautifulSoup