Predicting NBA Rookie's career

Yao Wateba Appeti, Riccardo Corrente, Pietro Sciabbarrassi, Francesco Sbordone, Santiago Vessi

December 2023

SAPIENZA
UNIVERSITÀ DI ROMA

Table of Contents

1 The Dataset

- The Dataset
- EDA
- Regression Model
- Clustering

The DataSet

1 The Dataset

The Dataset we worked on was obtained from the Basketball Reference website using **scrapers** to import 3 different datasets. These datasets were then merged to create the final dataset. The ultimate dataset includes the following features:

The final dataset has the following features of the rookies from 2010 to 2023:

- ID of the player.
- 32 statistics for the rookie's first season.
- Player Efficiency Rating (PER_mean) for the entire career, which serves as the target variable we aim to predict. It represents a numerical value indicating a player's skill level.

Table of Contents

2 EDA

- The Dataset
- EDA
- Regression Model
- Clustering

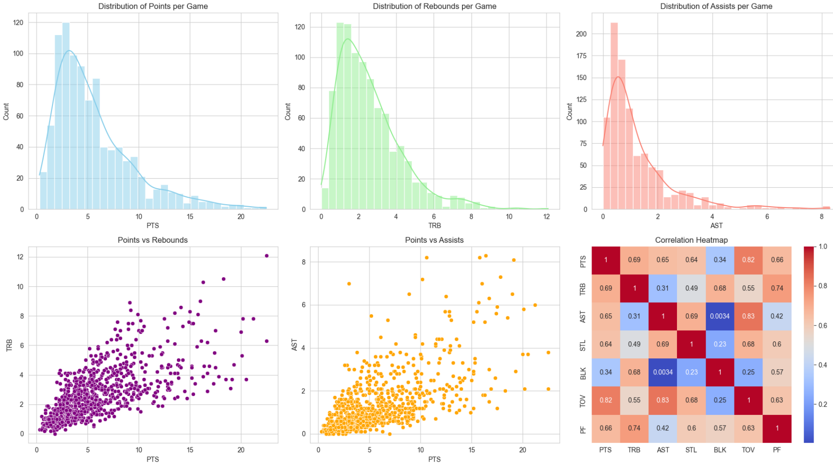
Exploratory Data Analysis (I)

We have applied various data analysis techniques to summarize the main characteristics of the dataset and identify correlations between variables.
Specifically, we have:

- Created charts to understand the distribution of some key variables (PTS, AST, TRB).
- Used a correlation matrix to investigate how our target variable may be influenced by others.
- Utilized other visualizations such as boxplots and scatterplots to explore the variations in PTS, AST, and TRB over the years.

Exploratory data analysis (II)

2 EDA



Distribution of Points per Game

Distribution of Rebounds per Game

Distribution of Assists per Game

Points vs Rebounds

Points vs Assists

Correlation Heatmap

Table of Contents

3 Regression Model

- The Dataset
- EDA
- Regression Model
- Clustering

Regression Model

We decided to use a linear regression model to predict the PER career value based on first year rookie's statistics.

- 80% of our players to train our OLS model
- 20% of our players to test our OLS model
- R^2 to evaluate performance
- Predict 2023's rookie career PER

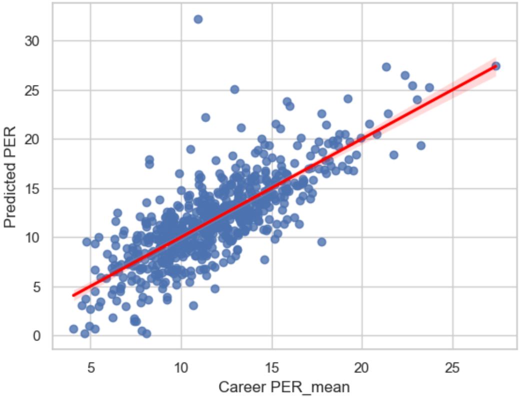


Figure: predicted PER vs actual PER

Evaluation of the Model

3 Regression Model

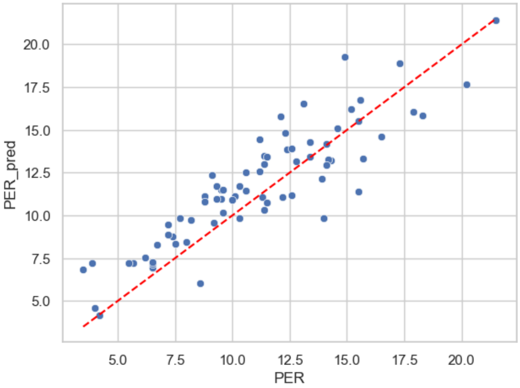


Figure: Predicted PER vs Rookie's PER

Let's make the prediction for the 2023 rookies.
The graph illustrates how the model predicts which players will maintain their current performance level and who is likely to see an increase or decrease in their Player Efficiency Rating (PER) over the course of their careers.

- R-squared: 0.658

Table of Contents

4 Clustering

- The Dataset
- EDA
- Regression Model
- **Clustering**

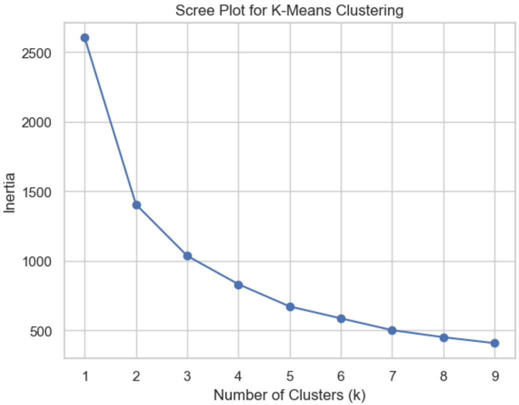
Clustering for a better regression model

The linear regression model does a good job given its simplicity, but it also leaves a lot of variance unexplained. This because one linear regression model is unlikely to succesfully predict all kind of different players.

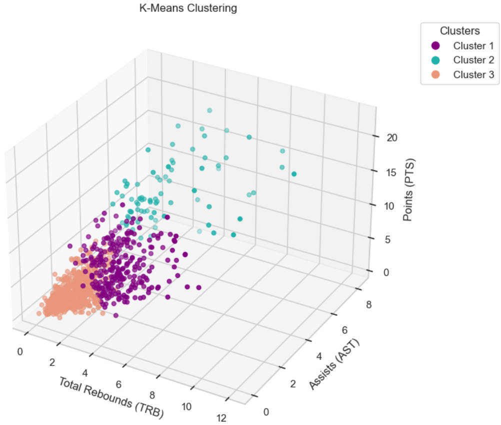
A way to solve this problem is grouping them in clusters and examine how groupings impact ability to predict career performance.

Selection of the number of clusters for k-means

- Grouping player by their main statistics (PTS, TRB, AST)
- Cluster algorithm used: k-means.
- To decide the number of clusters we can use the scree plot for the k-means clustering algorithm.
- The optimal number of clusters to choose is 3.



The 3 clusters

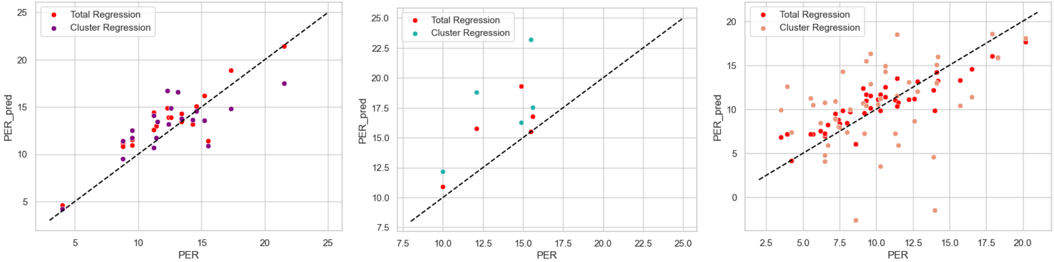


- We’ve identified the 3 clusters.
- The players are now grouped by their stats based on their skills.
  - Group 1: Top Players
  - Group 2: Average Players
  - Group 3: Weak Players
- Now we can expect a better regression for each cluster than the previous one.

Clustering

4 Clustering

After categorizing each 2023 rookie into a cluster based on their first-year statistics, we can now observe the updated predictions:



(a) Cluster 1 (Average)

(b) Cluster 2 (Top)

(c) Cluster 3 (Bottom)