Tianyang Huang

24 September 2022

**Using Machine Learning to Identify Optimal Physical Attributes for All-Around Basketball Players**

**Abstract**

**Introduction**

Therefore, there is an increasing need to identify the correlations between players’ height, weight, position, and experience with respect to their on court performance, so that we could find out how to evaluate and position a player in order to achieve maximal performance.

**Method**

Data Preparing: Reading the CSV file as dataframe. Fill all blank cells by the median of that column. Using onehotencoder to transform categorical data to numerical data. Position value is transformed from letter to corresponding letter from 1 to 5. Using Scikit-Learn to create a training set and a testing set.

Data Cleaning: Players that got traded have as many rows as their teams, therefore, we delete duplicates. We fill zero-value cells under “weight” and “height” with a median of nonzero values in respective columns. We remove abnormal stats, for instance, knowing Jokić.

Data Visualization: We first create a histogram for all attributes, then create a stratified graph for PER. We then implemented a correlation matrix based on PER as well as a scatter matrix based on five attributes.

Model Training: Try linear regression, logistic regression, random forest, cluster model, etc.

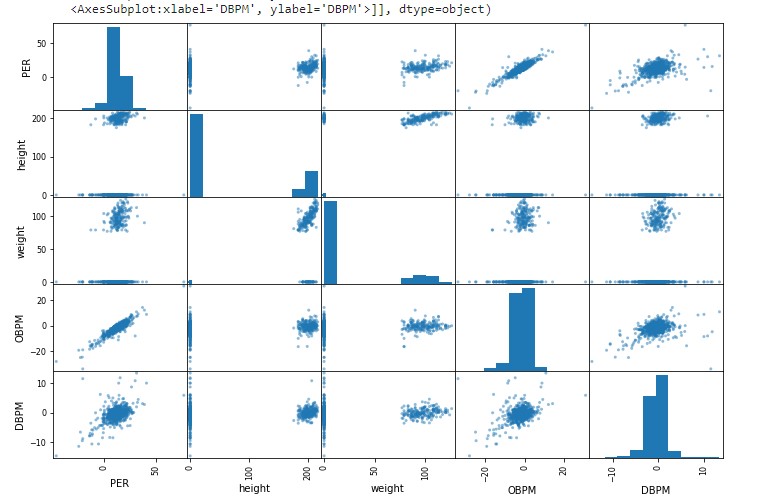
Testing and Validation: Use RMSE and K-fold validation to test out accuracy.

**Results**

Correlation Matrix Image:



Scatter Matrix Image:



From the correlation matrix we can see that OBPM has the strongest correlation with PER, and BPM follows, while DBPM falls a considerable number of places behind. Height and weight each only have 0.170 and 0.160 correlation measures. Similar to basketball critics, offensive effort is more influential to overall performance than defensive effort. Yet height and weight fail to display any visible correlation. When we switch to the correlation matrix, we can see that OBPM and DBPM basically form a direct linear relationship with PER, while height and weight individually with regard to other three variables, show no clear signs. After zooming in the heightVSPER graph(unable to show due to the limit), we could observe for different height the medians are close, but in the medium zone (6’5’’ to 6’8’’), the standard deviation is much smaller than the taller or shorter groups. And all high PER (>25) are achieved by players of 6’9’’ and above, except for Trae Young.

We trained a linear regression model, which has a decent success rate of 75%. But it couldn’t go any higher after rounds of tuning.

**Conclusion**

Height and weight are still dominating factors for basketball players. And taller players have higher upper limits. Medium-tall players display stable performance on both ends. And smaller players have to deliver high offensive performance to balance out their weakness in defense.Recruiting medium-size players is always safe, but big or small players make difference.

**Source & Credit**

Github: <https://github.com/kevinhcn/SSAC/tree/main>

Dataset: <https://www.kaggle.com/datasets/drgilermo/nba-players-stats>

<https://www.basketball-reference.com/leagues/NBA_2022_totals.html>

<https://www.basketball-reference.com/leagues/NBA_2022_advanced.html>