A Comparison of the Ability of Regression Models to Predicting Total Points for NBA Players

by Alyaqadhan Al Fahdi

Data & Research Questions

Data

- The main source of data for this study will be the NBA Players Stats for the 2023 season, available on Kaggle (NBA Players Stats (2023 Season), 2023)

- This dataset provides comprehensive statistics for each player, covering various aspects of their performance across 539 players and 30 different predictors.
  - 27 Quantitative variables
  - 3 Categorical variables

Research Questions

1. Can historical player data be used to predict the points scored by NBA players in the upcoming season?
2. Which model are most effective in forecasting player points scored?

Possible Methods

Prediction methods:

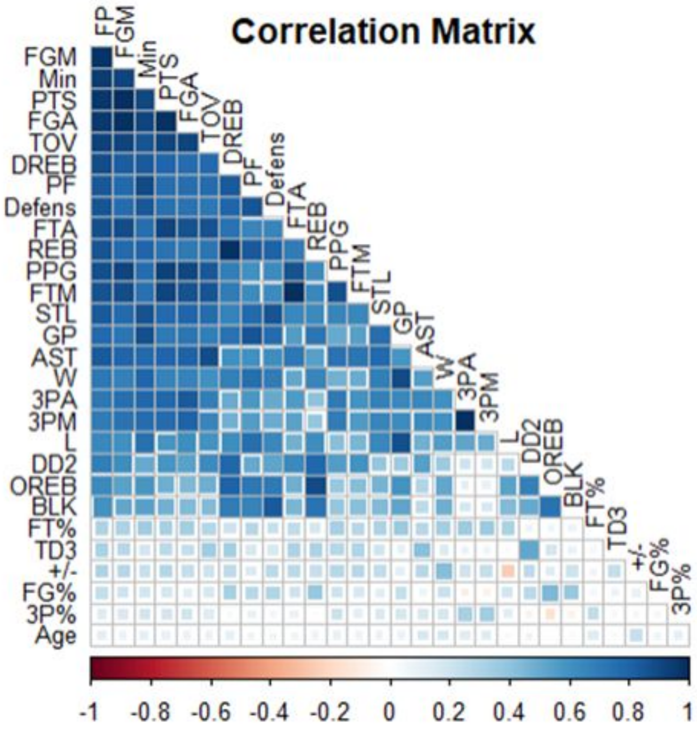
- Linear Regression
- SVM
- KNN
- Random Forest

Selecting Predictors:

- EDA
- PCA

Exploratory Data Analysis

Exploratory Data Analysis



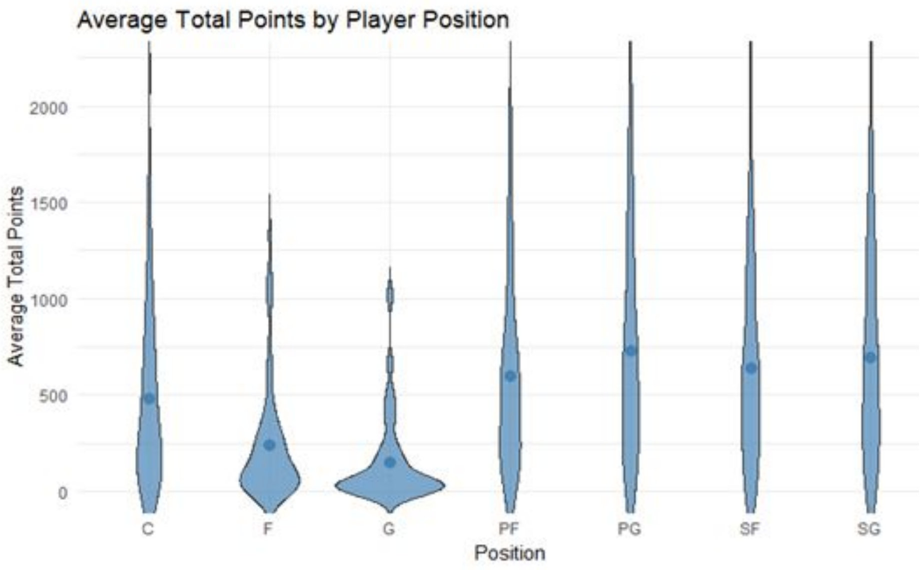
The correlation matrix shows that most of variables have highly correlated between them and the Total Points.

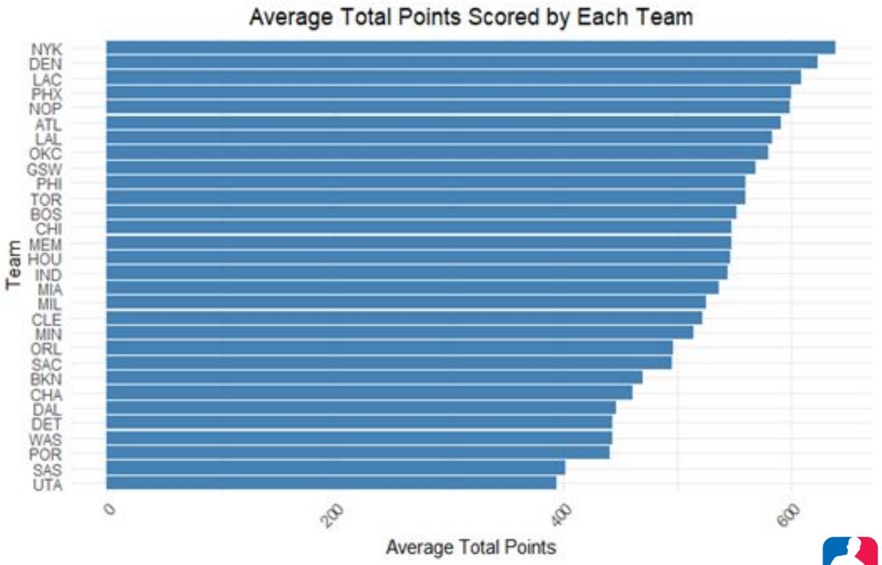
Highly correlated variables with PTS:

-GP: Games Played
-W: Wins
-Min: Minutes
-FGM: Field Goals Made
-FGA: Field Goals Attempted
-3PM: Three-Point Field Goals Made
-3PA: Three-Point Field Goals Attempted
-DEFENS: Blocks + Steals
-FTM: Free Throws Made
-FTA: Free Throws Attempted

-DREB: Defensive
-REB: Rebounds
-AST: Assists
-TOV: Turnovers
-PF: Personal Fouls
-FP: Fantasy Points

Exploratory Data Analysis



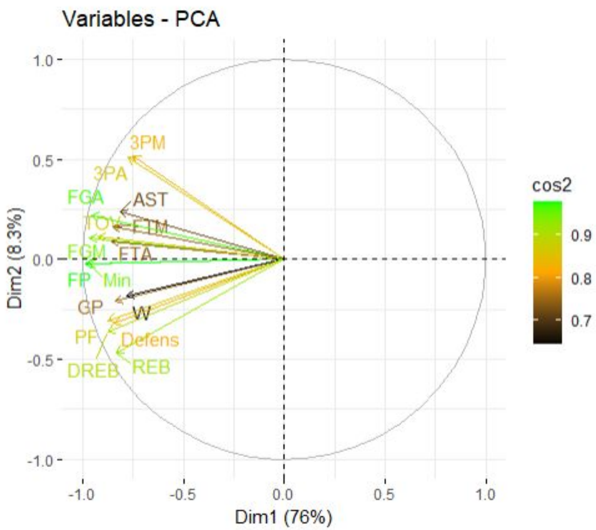


Principal Component Analysis (PCA)

PCA

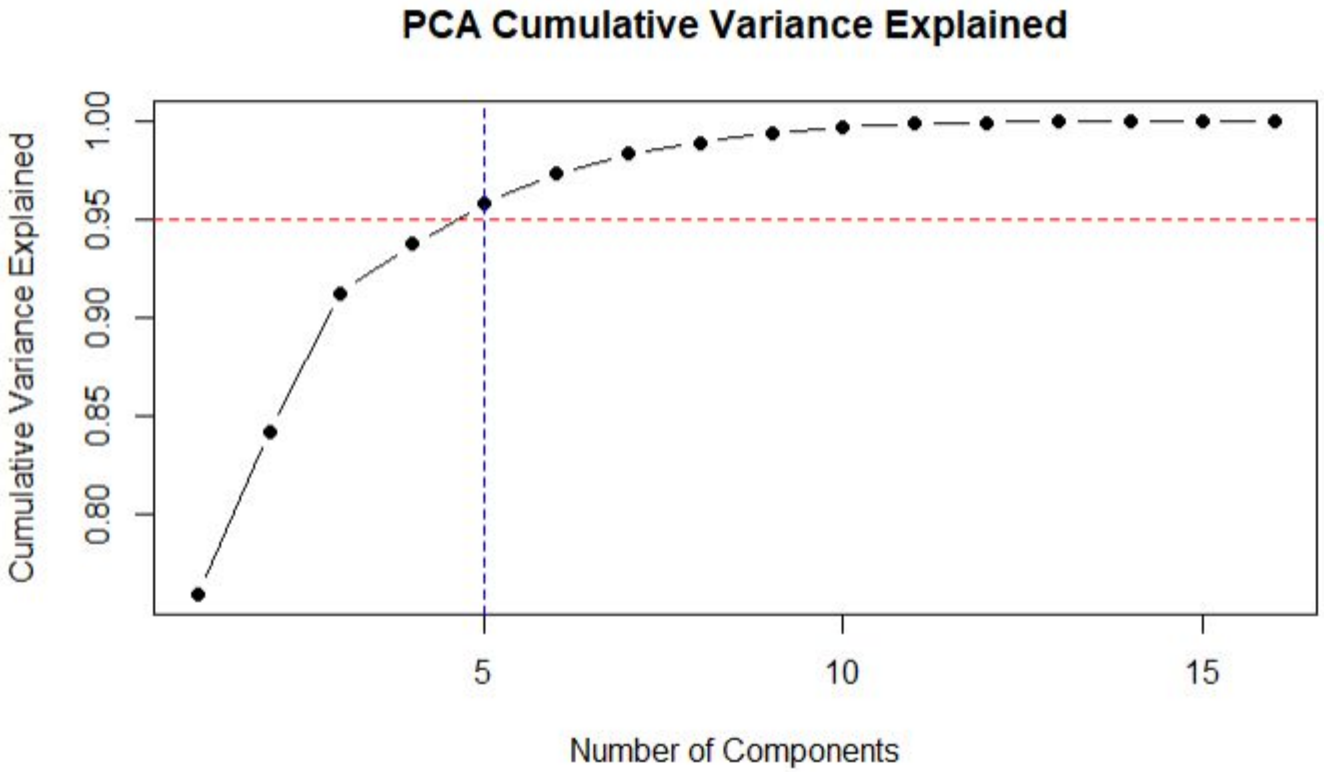
Overview

PCA: is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.



Why is it useful here?
Reduce multicollinearity between features, leading to improved performance and stability of regression

Principal Component Analysis (PCA)



Principal Component Analysis (PCA)

Loadings of the First Five Principal Components

	PC1	PC2	PC3	PC4	PC5
GP	-0.24	-0.18	0.38	0.15	-0.31
W	-0.22	-0.16	0.41	0.22	-0.53
Min	-0.28	-0.02	0.14	0.02	0.07
FGM	-0.28	0.09	-0.13	-0.12	0.00
FGA	-0.27	0.19	-0.06	-0.10	0.02
3PM	-0.21	0.45	0.34	-0.26	0.14
3PA	-0.22	0.44	0.31	-0.24	0.13
DEFENS	-0.24	-0.28	0.05	0.08	0.20
FTM	-0.24	0.14	-0.40	-0.12	-0.42
FTA	-0.25	0.08	-0.41	-0.12	-0.39
DREB	-0.25	-0.32	-0.09	-0.26	0.23
REB	-0.24	-0.41	-0.09	-0.26	0.21
AST	-0.23	0.21	-0.14	0.74	0.27
TOV	-0.27	0.10	-0.20	0.26	0.14
PF	-0.25	-0.27	0.15	0.00	0.09
FP	-0.28	-0.02	-0.11	0.01	0.09

-**PC1**: Strong negative loadings across many variables

-**PC2**: Dominated by 3-point shooting (3PM, 3PA) and negatively associated with defensive variables (DREB, REB).

-**PC3**: Positive loadings for games won (W) and a strong negative for free throws (FTM, FTA); may reflect the impact of successful plays excluding free throws.

-**PC4**: Highlighted by a very strong positive loading on assists (AST)

-**PC5**: Negative loadings for win (W) and a mix of positive and negative loadings elsewhere

Models Evaluation

Models Evaluation

Data Splitting:

- Training set: 80%
- Testing set: 20%

Hyperparameter Tuning:

- Cross-Validation (5 Fold)
- Grid Search

Evaluation Metrics:

- RMSE
- R-squared

Comparing the Models

Comparison of Model Performances

	Performance Metrics	
Model	RMSE	R-squared
Linear Regression	0.758	0.596
Random Forest	0.136	0.979
K-Nearest Neighbors	0.132	0.982
Support Vector Machine	**0.088**	**0.991**

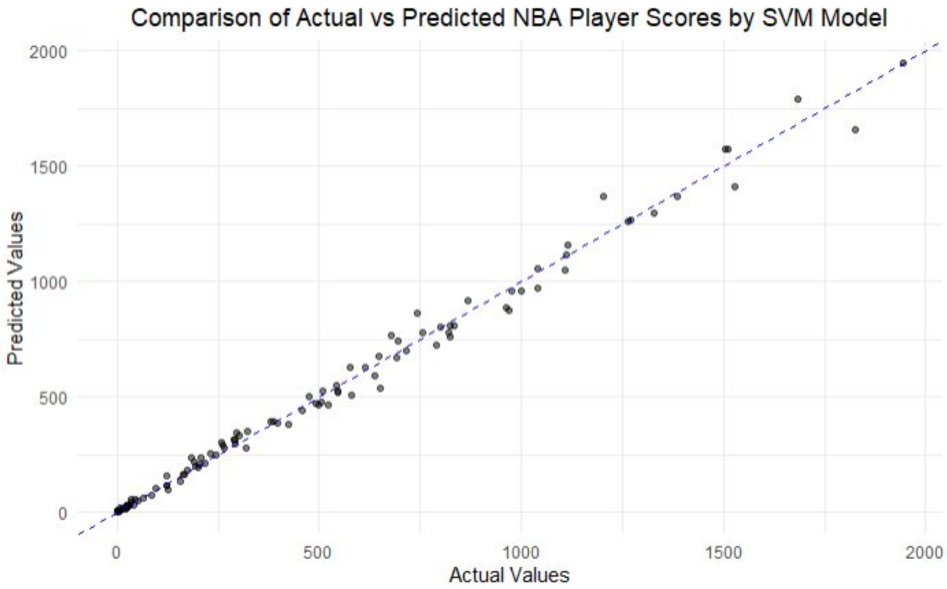
Support Vector Machine (SVM)

Support Vector Machines with Linear Kernel - Summary

C	RMSE	Rsquared	MAE
0.1	0.1210705	0.9869914	0.0863701
1.0	0.1181809	0.9873764	0.0787123
10.0	0.1184802	0.9874071	0.0793816
100.0	0.1184689	0.9874221	0.0793724
150.0	0.1184788	0.9874207	0.0793368

The ‘C’ parameter controls the trade-off between the insensitive loss and the sensitive loss. A larger value of ‘C’ means that the model will try to minimize the insensitive loss more, while a smaller value of C means that the model will be more lenient in allowing larger errors.

Support Vector Machine (SVM)



Side-by-Side Comparison of Actual vs Predicted NBA Player Scores by SVM Model

Actual Top	Predicted Top	Actual Bottom	Predicted Bottom
1946	1945	24	20
1826	1657	22	23
1683	1792	20	24
1529	1411	20	14
1510	1572	10	17
1505	1576	9	20
1385	1370	9	11
1329	1296	9	11
1271	1267	9	7
1263	1260	4	7
1204	1367	4	3
1114	1157	3	11
1113	1117	2	6
1109	1047	2	8
1041	1053	0	5

Conclusion

- **Historical Data as a Foundation:** Using historical player statistics is a possible method for predicting future performance (forecast points scored in the upcoming NBA season).

- **PCA for Dimensionality Reduction:** Principal Component Analysis effectively condensed our feature set, reducing multicollinearity while retaining 95% of the data variance.

- **SVM More Effective Than Other Models:** The Support Vector Machine with a linear kernel showed up as the most effective model with an RMSE of 0.088 and an R-squared of 0.991.

- **Next Steps:** Moving forward, we recommend the continued improvement of the SVM model with big data that contains at least 3 seasons and exploring the use of additional factors, such as player health for even more accurate predictions.

References

- GeeksforGeeks. (2023, January 30). Support Vector Regression (SVR) using Linear and Non-Linear Kernels in Scikit Learn. GeeksforGeeks. https://www.geeksforgeeks.org/support-vector-regression-svr-using-linear-and-non-linear-kernels-in-scikit-learn/

- *NBA Players stats(2023 season). (2023, August 4). Kaggle. https://www.kaggle.com/datasets/amirhosseinmirzaie/nba-players-stats2023-season*