Player Efficiency Prediction Model Using NBA Advanced Stats

Mohamed Sharif
Data Mining
Atlanta, GA
msharif9@student.gsu.edu

***Abstract*—This project was aimed to predict NBA Player Efficiency Rating (PER) using advanced metrics. The dataset included 11 years of NBA player data (2014-2024). The methods involved data preprocessing and modeling using linear regression. The model achieved stable predictive accuracy, providing practical insights into player performance.**

Keywords—NBA, Player Efficiency Rating (PER), Linear Regression, Data Mining, Machine Learning

I. Introduction

Background

Player Efficiency Rating (PER), developed by John Hollinger, is a summary statistic used in basketball analytics. It encapsulates a player's overall contribution to their team by combining various statistical measures like scoring, rebounding, and efficiency into a single value. PER adjusts for pace and other factors, making it an essential for evaluating player performance. Despite its utility, calculating PER involves a complex formula, making it less accessible for player analysis. Machine learning and data mining techniques offer new opportunities for analyzing sports data. These techniques can process large datasets to uncover patterns and predict outcomes with respectable accuracy. Predicting PER can have a real impact on player evaluation and team decision-making. By identifying key predictors of player efficiency, teams can better understand the factors driving success on the court. The reason I chose this topic is simple. For most of my life, I've loved basketball and the NBA. Being able to create this project helped me learn more about something that I care about. Using the data mining techniques I've learned over this semester along with the coding standards I already possess allowed me to create a program that can predict PER using NBA statistics.

Objectives

The primary objectives of this project were to:
- Identify the most impactful statistics contributing to PER.
- Build a predictive model for estimating PER.
- Develop a tool for real-time PER predictions based on player stats.
- Provide insights for basketball analysts.

Report Structure

This report covers the following:
1. Materials and Methods: Dataset creation, preprocessing, and modeling techniques.
2. Results: Model performance, feature importance, and predictive tool.
3. Discussion and Conclusion: Insights, limitations, and future work.
4. Table/Figures: Showing the visual results
5. Acknowledgments and References: Resources and tools used in the project.

Link for code: https://github.com/msharif9gsu/PER

II. Materials and Methods

Dataset

A statistical website called Basketball Reference was used to create this dataset, spanning 11 NBA seasons from 2014 to 2024. A normal NBA season will have 550 - 650 total players that participate. The entire dataset ended up having 5,792 player records, containing both basic and advanced stats. The 19 chosen stats were:

- **Basic Stats**: Minutes (MP), Field Goals Made (FG), Field Goals Attempted (FGA), Free Throws Made (FT), Free Throws Attempted (FTA), Rebounds (TRB), Assists (AST), Points (PTS), Passing Turnovers (BadPass Turnovers), Shooting Fouls Drawn (SFD), Points Generated by Assists (PGA), and total and1s (And1).

- **Advanced Stats**: True Shooting Percentage (TS%), Usage (USG%), Win Shares (WS), Box Plus-Minus (BPM), Value over Replacement Player (VORP), Offensive Rating (ORtg), and Defensive Rating (DRtg)

- **Target Variable**: Player Efficiency Rating (PER).

Each record represented a player's performance in a single season. The data was organized in a CSV format for analysis.

Preprocessing

- **Cleaning**: This dataset had multiple players with missing values in features like TS%, ORtg, and DRtg due to unqualified minute requirements. To clear all these players, everyone who played less than 200 minutes was removed.

Players with extreme outliers were also removed, as they would contribute to noise in the model. All players within these ranges were kept:

MP: Greater than 200, **PER**: Between 5 and 30, **TS%**: Between 0.3 and 0.7, **USG%**: Between 10% and 40%, **FG**: Between 0 and 15, **FGA**: Between 0 and 25, **FT**: Between 0 and 10, **BPM**: Between -10 and +10, and **VORP**: Between -2 and +10.

Afterward, the data type in each column was verified and the consistency throughout the dataset was checked. Lastly, for better readability, the rows were sorted alphabetically by the player's last name and the season year they played in.

- **Reduction**: Correlation analysis was used to identify the key predictors of PER. Defensive Rating (DRtg) was removed due to its low correlation (-0.1383216075), leaving 18 features:

Minutes: 0.4457379142, **FG**: 0.720283594, **FGA**: 0.6057710354, **FT**: 0.7068474488, **FTA**: 0.7243348529, **TRB**: 0.6536297654, **AST**: 0.4296414454, **PTS**: 0.7000733267, **PTOV**: 0.4305095142, **SFD**: 0.684044106, **PGA**: 0.4547931363, **AND1**: 0.6826519142, **TS%**: 0.5968459284, **USG**: 0.6180238693, **WS**: 0.757552787, **BPM**: 0.8544812516, **VORP**: 0.7813168753, and **ORTG**: 0.5967126026.

- **Transformation**: Applied min-max scaling to standardize all features to a range of 0-1. Doing this ensures compatibility with modeling.

Before preprocessing, the dataset had 5792 rows and 22 columns. After preprocessing, the dataset had 4490 rows and 21 columns, and all features were scaled between 0 and 1.

Modeling

The dataset was split into 80% training and 20% testing subsets. Linear Regression was trained on the normalized training data to predict PER. The model's coefficients highlighted important features, with metrics like Points (PTS) and Field Goals Made (FG) showing they have the highest influence. The model was evaluated against Random Forest, Gradient Boosting, and Support Vector Regressor, with Linear Regression showing the best performance. A 5-fold cross-validation approach was used to assess the consistency of the model. Metrics such as mean absolute error demonstrated the model's reliability across all folds, with low variance indicating strong generalization. All metrics used were:

Mean Absolute Error (MAE): Measures average prediction error.
Root Mean Squared Error (RMSE): Penalizes larger errors more heavily.
R^2: Indicates how much variability in PER is explained by the model.

III. Results

Predictive Tool

The predictive tool program that was created estimates a player's PER based on real-world stats. Users will input all 18 features that were listed for this project in order. Those features are then normalized using min-max scaling before being fed into the trained model. The output includes the real PER, predicted PER, and the difference between the two values. Table 1 summarizes the tool's performance for a set of 28 players ranging from 1997-2024. The tool's average difference for all players in that set was 1.81, demonstrating the utility it has for player evaluation. The reason these tests started in 1997 is because some of the basic features are considered "Play-By-Play." These include Passing Turnovers (PTOV), Shooting Fouls Drawn (SFD), Points Generated by Assists (PGA), and And1. The NBA began to track them that year.

Model Performance

Linear Regression accomplished an R^2 of 0.9542, explaining 95.42% of the variance in PER. The Mean Absolute Error (MAE) was 0.0283, and the Root Mean Squared Error (RMSE) was 0.0363, confirming high predictive accuracy.

Feature Importance

Feature importance analysis ranked the predictors of PER. Offensive metrics like Points (PTS), Field Goals Made (FG), Usage Percentage (USG%), Box Plus-Minus (BPM), and Free Throw Made (FT) had the highest influence in the model.

Visual Results

These charts were generated using Python libraries called Matplotlib and Seaborn.

Figure 2: **Actual vs. Predicted Plot**
Shows most predictions align closely with actual PER values.

Figure 3: **Residual Distribution**
Ensures errors follow a normal distribution.

Figure 4: **Residual Plot**
Confirms errors are randomly distributed, showing no systematic bias.

Figure 5: **Correlation Heat-map**
Discuss relationships between features and PER.

IV. Discussion and Conclusion

Understanding the Results

The project successfully predicted PER with reliable accuracy, validating the model's effectiveness. Offensive scoring statistics like Points (PTS) and Field Goals Made (FG) had the highest influence on the model. Passing statistics like Assists (AST), Passing Turnovers (PTOV), and Points Generated by Assists (PGA) had minimal impact, suggesting they contribute less to explaining variance in PER. The model closely predicted the PER of players such as Hakeem Olajuwon (real: 22.7, predicted: 22.9) and Michael Jordan (real: 25.2, predicted: 26.2). However, larger deviations for players like Peja Stojaković (real: 21.8, predicted: 17.8) shows that the model occasionally struggles with specific cases, potentially due to unique play-styles or limitations in the dataset.

Limitations and Future Work

The dataset spans 11 seasons but could benefit from adding more historical data for a broader generalization. Instead of starting in 2014, 1997 would be ideal because "Play-By-Play" stats were beginning to be recorded that year. Also, adding more advanced stat features along with player's shooting profile data could improve prediction accuracy. It may provide deeper insights into player efficiency by capturing aspects of performance not reflected in the current model. Finally, while PER is a good metric for summarizing player performance, exploring alternative metrics like Estimated Plus-Minus (EPM) could offer a more comprehensive evaluation. EPM considers on-court and off-court impacts, potentially having a better understanding of a player's contributions to team success.

Key Insights

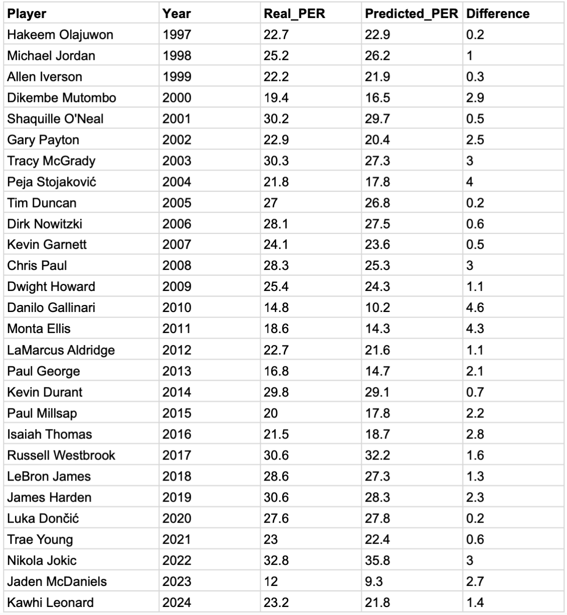
With basketball-specific insights in mind, the analysis of PER revealed that offensive scoring statistics play a more significant role in determining player performance than defensive or playmaking statistics. The coefficients with the highest influence aligned closely with elements that increase team success, such as offensive efficiency and shot creation. This highlights the importance of offensive contributions in overall player evaluations. As far as insights had with data mining, linear regression proved to be an ideal choice for this task due to its straightforward structure, allowing for clear identification of the relationships between features and PER. Also, the preprocessing steps, such as transformation and outlier removal, were vital in enhancing the model's performance. By normalizing features to a uniform scale and eliminating extreme values, the data was better suited for analysis. That contributed to the model's accuracy.

Conclusion

Overall, the project demonstrates how a combination of appropriate preprocessing and a focus on key offensive metrics can produce an effective tool for predicting and evaluating an NBA player's performance. It also displayed the power of machine learning in sports analytics, delivering a useful program for predicting PER. The model provides a foundation for future work, with the potential to improve player scouting, performance evaluation, and team strategy development.

Table/Figures

Table 1:



Player	Year	Real_PER	Predicted_PER	Difference
Hakeem Olajuwon	1997	22.7	22.9	0.2
Michael Jordan	1998	25.2	26.2	1
Allen Iverson	1999	22.2	21.9	0.3
Dikembe Mutombo	2000	19.4	16.5	2.9
Shaquille O'Neal	2001	30.2	29.7	0.5
Gary Payton	2002	22.9	20.4	2.5
Tracy McGrady	2003	30.3	27.3	3
Peja Stojaković	2004	21.8	17.8	4
Tim Duncan	2005	27	26.8	0.2
Dirk Nowitzki	2006	28.1	27.5	0.6
Kevin Garnett	2007	24.1	23.6	0.5
Chris Paul	2008	28.3	25.3	3
Dwight Howard	2009	25.4	24.3	1.1
Danilo Gallinari	2010	14.8	10.2	4.6
Monta Ellis	2011	18.6	14.3	4.3
LaMarcus Aldridge	2012	22.7	21.6	1.1
Paul George	2013	16.8	14.7	2.1
Kevin Durant	2014	29.8	29.1	0.7
Paul Millsap	2015	20	17.8	2.2
Isaiah Thomas	2016	21.5	18.7	2.8
Russell Westbrook	2017	30.6	32.2	1.6
LeBron James	2018	28.6	27.3	1.3
James Harden	2019	30.6	28.3	2.3
Luka Dončić	2020	27.6	27.8	0.2
Trae Young	2021	23	22.4	0.6
Nikola Jokic	2022	32.8	35.8	3
Jaden McDaniels	2023	12	9.3	2.7
Kawhi Leonard	2024	23.2	21.8	1.4

Fig. 2:

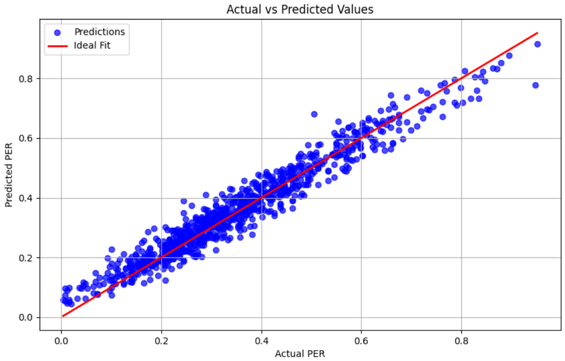


Fig. 3:

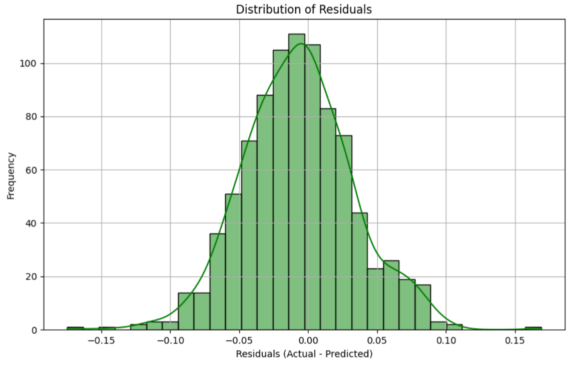


Fig. 4:

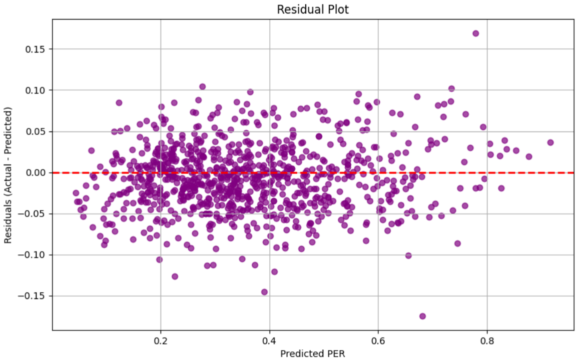
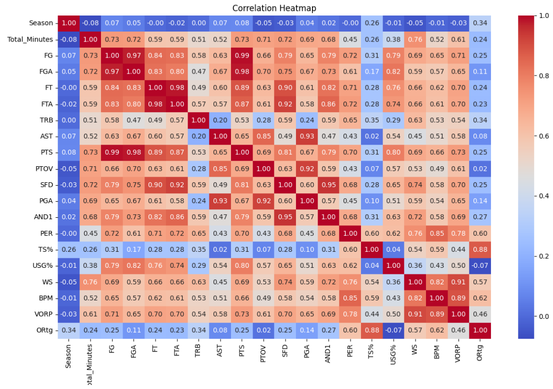


Fig. 5:



Acknowledgments

I appreciate Basketball Reference for making it easy to pull data from their site. It allowed me to create the dataset I wanted for this project. Creating my own dataset was preferred from the beginning, as it would give me all the context I needed going forward.

I also appreciate the developers of the Python libraries used in this project, which simplified data analysis, modeling, and visualization. These include:

1. **Pandas** for data manipulation and analysis.
2. **Numpy** for numerical operations.
3. **Matplotlib** and **Seaborn** for creating visualizations.
4. **Scikit-learn** for providing the tools for machine learning, including Linear Regression, evaluation metrics, and cross-validation.
5. **Joblib** for efficient model saving and loading.

Finally, I want to thank Professor Jingyu Liu for her teaching throughout the course, which gave me an understanding of some data mining concepts I used in my project.

References

1. Basketball Reference. "NBA Player Statistics." Retrieved from https://www.basketball-reference.com/leagues/NBA_2025_leaders.html
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
3. McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 56-61.
4. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95.
5. Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering, 13(2), 22-30.
6. Waskom, M., & the Seaborn development team. (2021). Seaborn: Statistical data visualization. Journal of Open Source Software, 5(52), 3021.