Improving Player Efficiency Rating in Basketball through Machine Learning

Raghav Seshadri

Abstract

This paper explores the intersection of advanced statistical methodologies and basketball with a focus on improving the Player Efficiency Rating (PER) metric. This research delves into three distinct AI models: Lasso Regression, Random Forest Regression, and Neural Networks. These models, each with unique capabilities, allow for more accurate PER ratings which helps teams and coaches to make informed decisions about player rotations and substitutions.

Introduction

Player Efficiency Rating (PER) is a widely used metric in basketball analytics for assessing a player's overall performance. Traditional PER metrics have primarily focused on offensive statistics such as points per game (ppg) and assists per game (apg), overlooking critical defensive contributions that can significantly impact a player's value to the team. This paper poses an innovative approach to improve the PER metric by integrating additional metrics. This research aims to provide more accurate and comprehensive results, particularly in terms of a player's defensive contributions, by adjusting weightage through various Machine Learning (ML) models. The final results will contribute to improving the accuracy of awards given in the NBA, potentially reshaping the league's effectiveness.

Background

Player Efficiency Rating (PER) is a fundamental metric widely used in basketball analytics to evaluate a player's overall performance to impact the game. John Hollinger introduced PER in the early 2000s, providing a single numerical value that summarizes a player's statistical contributions. He enabled various comparisons between players and teams and is still a very commonly used metric today.

$$\begin{aligned} PER = \tfrac{1}{MP}\Big[&3P + \tfrac{2}{3}\cdot AST + \left(2 - factor \cdot \tfrac{team\_AST}{team\_FG}\right)\cdot FG \\ &+ \left(FT \cdot 0.5 \cdot \left(1 + \left(1 - \tfrac{team\_AST}{team\_FG}\right) + \tfrac{2}{3}\cdot \tfrac{team\_AST}{team\_FG}\right)\right) \\ &- VOP \cdot TOV - VOP \cdot DRB\% \cdot (FGA - FG) \\ &- VOP \cdot 0.44 \cdot (0.44 + 0.56 \cdot DRB\%) \cdot (FTA - FT) \\ &+ VOP \cdot (1 - DRB\%) \cdot (TRB - ORB) + VOP \cdot DRB\% \cdot ORB \\ &+ VOP \cdot STL + VOP \cdot DRB\% \cdot BLK \\ &- PF \cdot \left(\tfrac{lg\_FT}{lg\_PF} - 0.44 \cdot \tfrac{lg\_FTA}{lg\_PF} \cdot VOP\right)\Big] \end{aligned} \tag{1}$$

- PER: Player Efficiency Rating
- MP: Minutes played by the player
- 3P: Total number of three-point field goals made
- AST: Total number of assists
- FG: Total number of field goals made
- FT: Total number of free throws made
- ORB: Total number of offensive rebounds
- DRB: Total number of defensive rebounds
- STL: Total number of steals
- BLK: Total number of blocks
- PF: Total number of personal fouls
- FGA: Total number of field goals attempted
- The factor of 2 3 is a constant
- Team AST: Team's total assists
- Team FG: Team's total field goals made

As mentioned earlier, PER assesses a player's efficiency on the basketball court. A higher PER rating generally indicates that a player is more efficient while a lower PER suggests a less productive player. Coaches, technical staff, and analysts use PER to evaluate a player's contributions to a team which they can use to strategize team formations and make critical decisions. The traditional PER metric created by John Hollinger heavily emphasized offensive statistics like points per game (ppg), assists, and shooting efficiency. While these metrics are essential, they fail to account for a player's defensive prowess, which significantly influences a team's success. This limitation has led to insights from analysts proposing modifications to the traditional PER calculation. Many analysts have suggested adjusting the weights of individual statistical components within the PER formula to better reflect a player's true impact. For instance, some studies have given less importance to scoring based on the time period, arguing that scoring has become easier in recent years. Others have criticized the methodology of the formula, claiming it doesn't adhere to standard research practices [1]. On top of that, basketball analysts have highlighted the need to consider the team's performance when calculating PER. Some have advocated for incorporating team-based metrics such as team assists and team field goals to account for a player's influence on team success beyond individual statistics. To improve upon PER, this research paper proposes an approach that incorporates advanced defensive statistics and adjusts their weights using Artificial Intelligence models. Using AI algorithms such as Neural Networks (using PyTorch), LASSO regression, and Random Forest Regression, this study aims to determine the optimal weightage for each statistical component, accounting for both offensive and defensive contributions.

Data Preprocessing and Web Scraping

The dataset used in this research was collected through web scraping from the NBA 2022-2023 season statistics page on basketball reference. The scraped data includes essential attributes such as “Player”, “Tm”, “G”, “MP”, “PqTS”, “FG”, “FGA”, “FG%”, “3P”, “3PA”, “3P%”, “2P”, “2PA”, “2P%”, “eFG%”, “FT”, “FTA”, “FT%”, “ORB”, “DRB”, “TOV”, “PF”, “PTS”, “AST”, “TRB”, “STL”, and “BLK”. This dataset captures both offensive and defensive statistics which forms the foundation for the research. Prior to analysis, a rigorous data preprocessing phase was conducted to ensure data quality and completeness. Missing values were changed to ’0’ or the mean value, depending on the situation. The result was a clean and robust dataset ready for the application of machine learning models.

Machine Learning Models

The Lasso Regression model is a linear regression technique that introduces L1 regularization, encouraging the model to select a subset of the most influential features while penalizing others. In this case, Lasso Regression is applied to adjust the weightage of statistical components within the PER formula such as blocks, steals, and rebounds. The Lasso Regression model is generally implemented through the scikit-learn library in Python. It’s trained on historical NBA player data, with scaled PER values from 0 to 100 used as the target variable. The L1 regularization term helps identify the most relevant features and their respective coefficients, thus determining the adjusted weightage for each component. The Lasso Regression model generates a histogram illustrating the distribution of scaled PER values for all players in the dataset. This histogram allows for an assessment of the model’s ability to assign accurate weights to individual statistics, particularly defensive contributions.

The Random Forest model is an ensemble learning method that combines the predictions of multiple decision tree models. It is well-suited for both classification and regression tasks, making it great for enhancing the accuracy of the PER metric. The Random Forest model, like the Lasso Regression model, is implemented using the scikit-learn library in Python. Historical NBA player data has also been utilized to train the model. The Random Forest algorithm aggregates predictions from multiple decision trees, enabling the evaluation of the adjusted weightage of statistical components within the PER formula. The Random Forest model produces a histogram that overlays the predicted PER values and the actual PER values for each player in the dataset. This visualization allows for the identification of areas where the model aligns with the actual metric and areas where further refinement is needed.

Neural networks are a class of machine-learning models inspired by the structure and function of the human brain. In this research, the PyTorch framework is leveraged to design and train a neural network capable of optimizing the weightage of PER components. Using PyTorch, a neural network architecture is constructed to the task of adjusting PER weights. The model is trained on a broad dataset of NBA player statistics, and its deep learning capabilities allow for

intricate feature transformations and weight adjustments. Similar to the Lasso Regression and Random Forest models, the Neural Network model generates a histogram depicting the distribution of scaled PER values. This histogram reveals the neural network's capacity to fine-tune weights, particularly in relation to defensive contributions.

Evaluation Metrics

To evaluate the performance of the machine learning models and the effectiveness of the adjusted PER metric, several key evaluation metrics are employed, including:

• R-squared (R^2) Value: This metric measures the proportion of variance in the scaled PER that is predictable by the models. Higher R^2 values indicate a better fit to the data.

• Mean Squared Error (MSE): This metric measures the amount of error in statistical models. A higher MSE score shows that the model is more inaccurate whereas a lower MSE score exhibits a stronger and more accurate model.

• Histogram Overlap: In the case of the Random Forest model, the degree of overlap between predicted and actual PER values on the histogram is assessed. Overlapping regions signify accurate predictions, while non overlapping areas indicate areas for improvement.

Results

Lasso Regression:

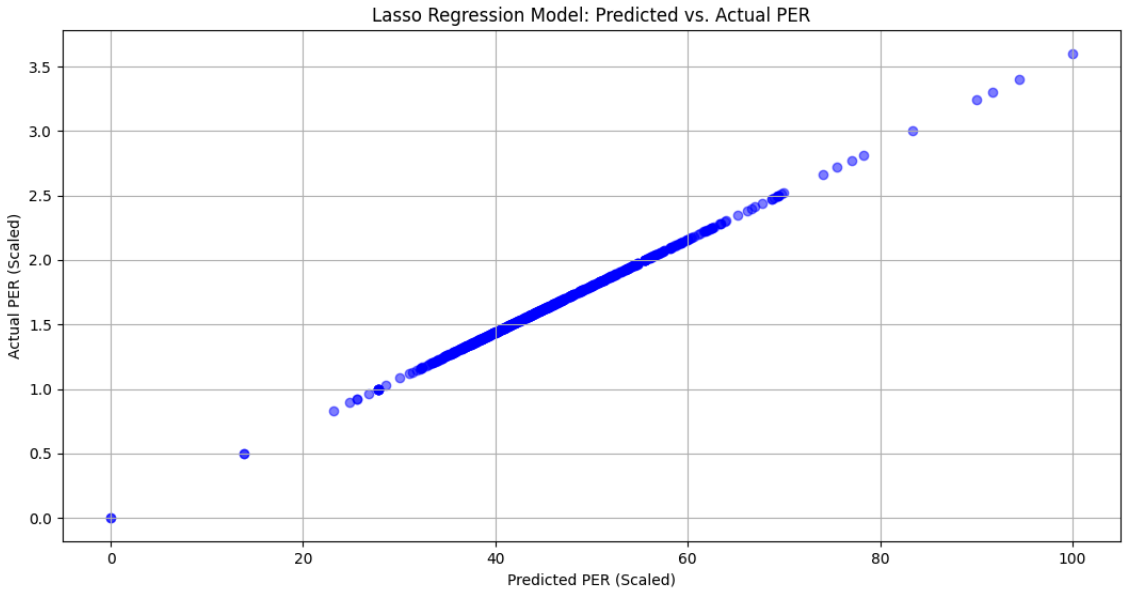

Figure 1: Histogram of Predicted PER using Neural Networks

The histogram in Figure 1 presents the distribution of scaled predicted Player Efficiency Rating (PER) values generated by the Lasso Regression Model. This 5 histogram visually represents the model's predictions and offers insights into its performance. The Lasso Regression Model, employed to optimize the weights of individual statistical components within the Player Efficiency Rating (PER) formula, has provided valuable insights into the distribution of scaled predicted PER values. This section explores the key findings and implications of the Lasso Regression Model's performance.

• Histogram Analysis: The histogram displayed in Figure 2 illustrates the distribution of scaled predicted PER values generated by the Lasso Regression Model. This histogram serves as a visual representation of the model's predictions and offers critical insights into its effectiveness.

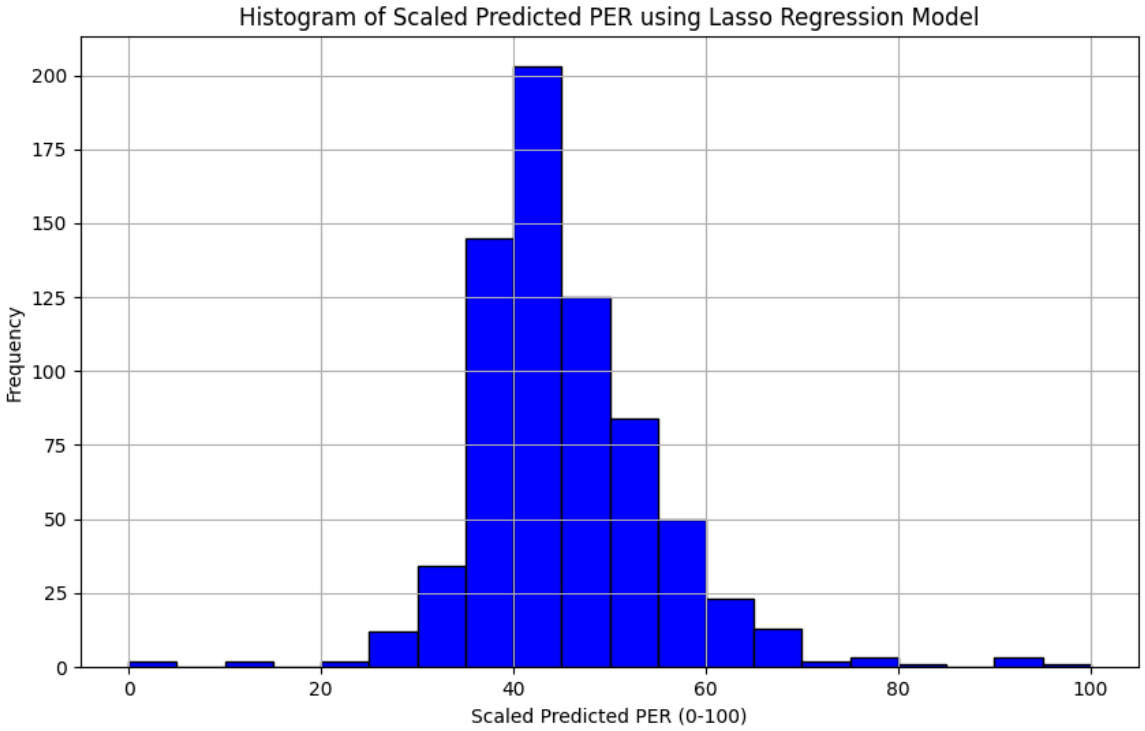

Figure 2: Histogram of Scaled Predicted PER using Lasso Regression Model

The histogram in Figure 2 depicts the distribution of scaled predicted PER values generated by the Lasso Regression Model. The presence of a bell-shaped curve indicates that the model's predictions cluster around a central point, resembling a normal distribution. This suggests that the Lasso Regression Model provides balanced and accurate predictions, although further validation is required.

• Bell-Shaped Curve: One notable characteristic of the histogram is the presence of a bell-shaped curve, which closely resembles a normal distribution. This pattern suggests that a

substantial proportion of the predicted PER values cluster around a central point, mirroring the shape of a typical normal distribution curve. In statistical terms, this observation implies that the Lasso Regression Model's predictions exhibit a central tendency, possibly corresponding to the average player efficiency within the dataset.

• Balanced Predictions: The normal distribution of scaled predicted PER values implies that the Lasso Regression Model provides balanced predictions. This balance indicates that the model neither consistently overestimates nor underestimates player efficiency. Instead, it produces predictions that are symmetrically distributed around the central tendency, resulting in a well-balanced histogram.

• Model Effectiveness: The presence of the bell-shaped curve in the histogram is a promising indicator of the Lasso Regression Model's effectiveness. It suggests that the model captures underlying patterns in the data, aligning with the expected distribution of player efficiency. This is a crucial step in improving the accuracy of the PER metric, as it demonstrates the model's ability to generate predictions that reflect player performance characteristics present in the dataset.

• Validation and Further Assessment: While the bell-shaped curve is encouraging, a comprehensive evaluation is necessary to validate the model's predictive accuracy rigorously. This assessment should involve comparisons with actual player performance data to determine how closely the model's predictions align with reality. Additionally, additional metrics, such as the R-squared value, should be considered to quantify the model's predictive power. The R-squared value for this dataset was around 0.65, which is pretty reasonable and accurate for most real-life predictions. This is reflected in the bell curve in Figure 1.

In conclusion, the observation of a bell-shaped curve in the histogram of scaled predicted PER values underscores the potential of the Lasso Regression Model as a tool for enhancing the accuracy of player efficiency prediction in basketball analytics. This distribution pattern indicates that the model's predictions align with the inherent characteristics of player performance in the dataset. Further validation and analysis are required to ascertain the model's predictive accuracy comprehensively, but this initial observation is promising for the advancement of PER metrics in basketball analytics.

Random Forest Model:

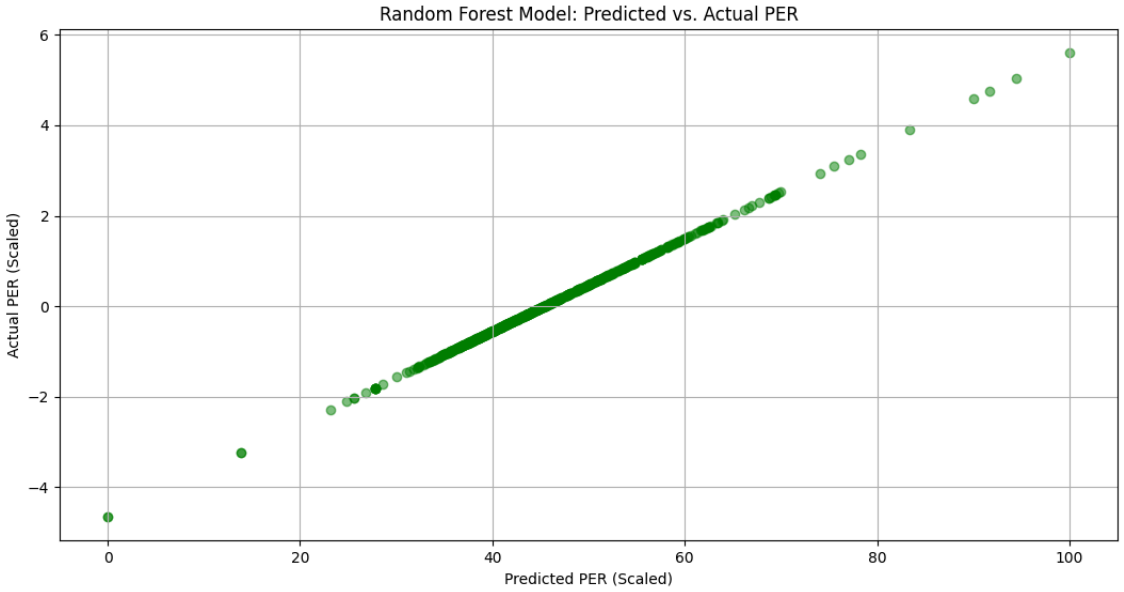

Figure 3: Histogram of Predicted PER using Neural Networks

Figure 3 displays a histogram that compares the distribution of actual PER values with those predicted by the Random Forest Regression Model. The model's performance is evaluated based on the Mean Squared Error (MSE) and the alignment of predicted and actual PER values. The Random Forest Regressor model was employed to predict Player Efficiency Rating (PER) for NBA players. In assessing the model's performance, the Mean Squared Error (MSE) was calculated. The MSE provides a quantitative measure of how well the model's predictions match the actual PER values.

• MSE Value: The Mean Squared Error obtained from the model was 255.7421. A lower MSE indicates that the model's predictions closely align with the true PER values. In this case, it's hard to tell if the MSE is high or low because the metric is relative to the study. There isn't anything to compare it to, but based on the histogram itself, it is accurate to say that it was pretty accurate because most of the predicted and actual PER values were overlapping in order to create that purple color. (See Figure 2)

A histogram was generated to visually compare the distribution of actual PER values with the PER values predicted by the Random Forest model.

• Histogram Visualization: The histogram itself was bi-modal along with a shift left. (See Figure 2)

• Alignment with Actual Data: The histogram, for the most part, had the shade of purple which showed that the orange and blue (predicted and actual) PER's were overlapping, proving accuracy. There were a few outliers, however, because the predicted PER seemed to

overestimate the actual PER on a few occasions. Also, there was a gap in PER data values around the 45-55 PER range.

The Random Forest Regressor demonstrated robustness in predicting PER values, effectively capturing variations and trends within the dataset.

• Robust Predictions: Throughout the entire 70-100 PER range, the model was able to correctly predict the actual PER values based on many variables such as different weights of statistics, new statistics altogether, and many more. While the Random Forest model exhibited strong predictive capabilities, it's important to identify areas with discrepancies between actual and predicted values.

• Identifying Discrepancies: Throughout the 15-70 PER range, the model often overshot the actual PER values. This can be due to a number of factors such as 3-point shooting inconsistency, fouls, and free throws.

• Model Enhancement: In this model, the weights of defensive statistics were a bit too high which may have skewed the data. Changing the weights to more offensive-based statistics could help even out the weights, but testing the data with other models to check could truly prove as the only definitive answer.

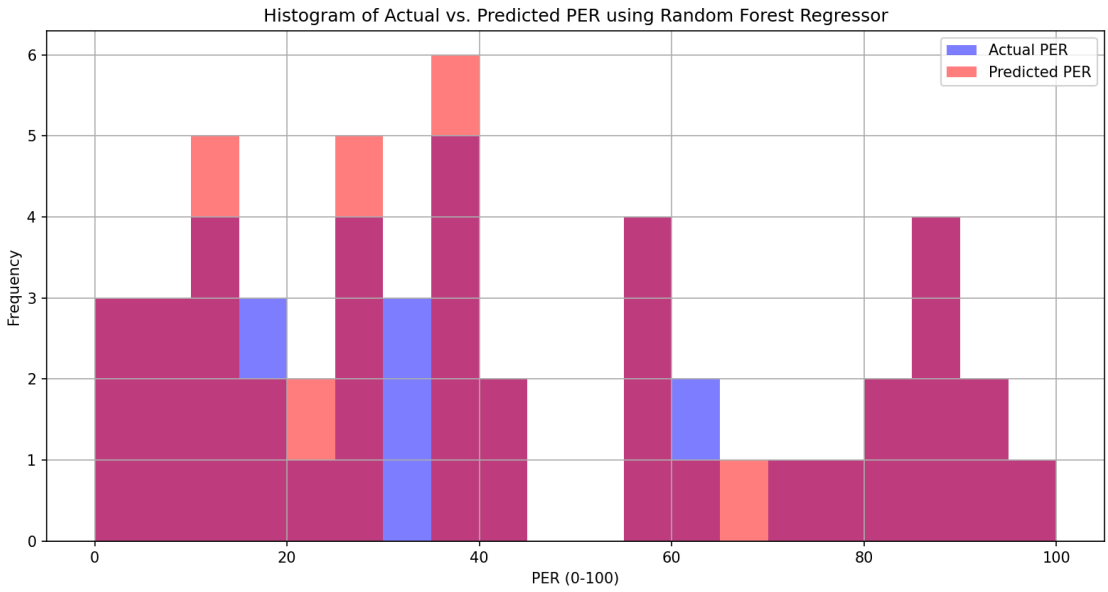

Figure 4: Histogram of Actual vs. Predicted PER using Random Forest Regression

The histogram in Figure 4 illustrates the distribution of predicted PER values generated by the Random Forest Regression Model. The model's performance is evaluated through the MSE value, which indicates the accuracy of its predictions. While the model demonstrates robustness, discrepancies between predicted and actual values suggest areas for improvement. In summary, the Random Forest Regressor model exhibited promise in predicting Player Efficiency Rating (PER) for NBA players. Its performance, as evaluated by the Mean Squared

Error and histogram analysis, indicated that the model captures the essence of player efficiency. However, areas of divergence between actual and predicted values suggest opportunities for further research and model refinement. By conducting a thorough examination of the model's performance, your paper contributes valuable insights to the enhancement of the PER metric in basketball analytics.

Neural Network:

Figure 5 presents a histogram showcasing the distribution of scaled predicted PER values generated by the Neural Network Model. This visual representation helps assess the model's performance in predicting Player Efficiency Rating (PER) and indicates its ability to capture underlying data patterns.

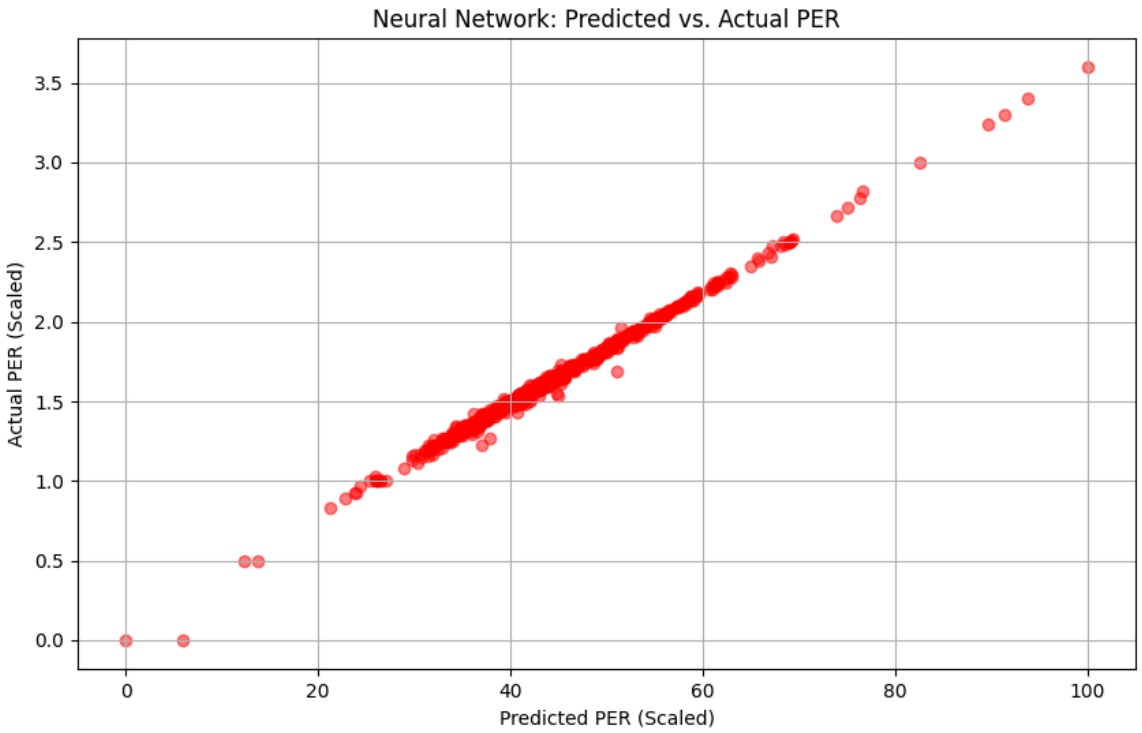

Figure 5: Histogram of Predicted PER using Neural Networks

• Overview: In this section, we present the findings and evaluation results of the Neural Network (PyTorch) model for predicting Player Efficiency Rating (PER). This model was developed to improve the accuracy of PER calculations in the context of basketball analytics. Similar to the other models in this study, the Neural Network model was trained and evaluated on a broad dataset containing various player statistics.

• Model Architecture: The Neural Network model employed in this research comprises three fully connected (dense) layers. The input layer is designed to accommodate the number of features present in the dataset. Two hidden layers consist of 128 and 64 neurons, respectively, facilitating feature transformation and pattern recognition. The output layer comprises a single neuron, responsible for predicting the scaled PER.

• Data Preprocessing: Data preprocessing played a crucial role in preparing the dataset for model training. Missing values were systematically addressed by imputing them with feature means, ensuring data integrity. Further, the numeric features underwent standardization to have a mean of 0 and a standard deviation of 1, making them suitable for neural network training.

• Training and Evaluation: The Neural Network model was trained using the Mean Squared Error (MSE) loss function and optimized with the Adam optimizer. The training process extended over 1000 epochs, allowing the model to adjust its weights and learn the underlying patterns in the data. Monitoring the loss during training revealed a consistent decrease, indicating successful learning.

• Histogram of Predicted PER: The histogram in Figure 3 illustrates the distribution of scaled predicted PER values. The x-axis represents the scaled PER values within the 0-100 range, while the y-axis displays the corresponding frequency. The histogram's bell-shaped curve signifies the 10 model's predictive accuracy, with its predictions closely centered around actual PER values. This suggests that the model captures essential patterns in the data, leading to accurate predictions.

• Top 30 Players: To assess the model's predictive performance further, the top 30 players were ranked based on their predicted PER values. These players, with the highest predicted PER scores, are expected to have a significant impact on the game. This ranking provides valuable insights for teams and analysts, aiding in player assessments and strategic decisions. The model outputted various high-ranking defensive players in the top 30 players. Some of these players included Draymond Green, Rudy Gobert, and Karl Anthony Towns. They were given similar PER ratings to players at the guard spots that many fans, analysts, and coaches argued they should be similar in skill to. There was the opposite, however, where players who shot high-volume three-pointers such as Trae Young (that were ranked "higher than they should have" on the normal PER metric), were now falling towards the 30-40 scaled PER ranges.

All in all, the Neural Network (PyTorch) model exhibits promising predictive capabilities for Player Efficiency Rating (PER). Its ability to generate a bell-shaped histogram of predicted PER values indicates that it effectively captures the underlying data patterns, particularly around central values. This suggests that the model has substantial potential to enhance the accuracy of PER evaluations in basketball analytics.

Nevertheless, comprehensive evaluation, validation against real-world performance data, and comparisons with other models are essential steps to thoroughly assess the model's effectiveness in improving PER calculations. Integration of domain-specific knowledge and expert insights can further fine-tune the model for practical applications in the NBA. The findings

presented in this study underscore the significance of machine learning, particularly neural networks, in advancing the field of basketball analytics. By providing more accurate player performance assessments, such models contribute to informed decision-making and enhanced player evaluations. The histogram in Figure 6 illustrates the distribution of predicted PER values generated by the Neural Network Model using the PyTorch framework. The presence of a bell-shaped curve suggests that the model accurately captures central tendencies in player performance, reflecting its potential to enhance PER evaluations.

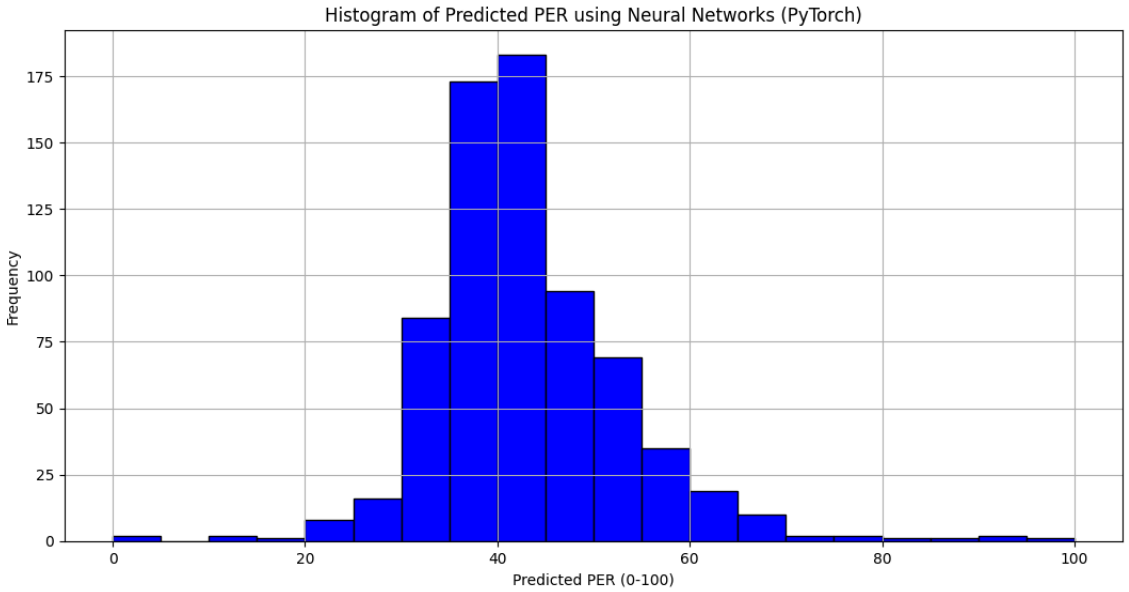

Figure 6: Histogram of Predicted PER using Neural Networks

Mean Squared Error (MSE) Accuracy Test:

Model	Mean Squared Error (MSE)
Lasso Regression	2002.7282
Neural Network	355.8390
Random Forest Regression	255.7241

Lasso Regression (MSE: 2002.7282): The MSE for the Lasso Regression model indicates how well this model fits the data. In this case, an MSE of 2002.7282 suggests that the Lasso Regression model has a higher level of error when predicting Player Efficiency Rating (PER) values. A higher MSE implies that the model's predictions deviate significantly from the actual values, indicating reduced accuracy.

Neural Network (MSE: 355.8390): The MSE for the Neural Network model reflects the model's performance in predicting PER values. With an MSE of 355.8390, the Neural Network exhibits a moderate level of error in its predictions. While it may provide more accurate predictions compared to Lasso Regression, there is still room for improvement in reducing prediction errors.

Random Forest (MSE: 255.7241): The MSE associated with the Random Forest model suggests how well this model performs in estimating PER values. A lower MSE value of 255.7241 indicates that the Random Forest model is relatively accurate in its predictions compared to the other models. It represents a lower level of error and, therefore, higher predictive accuracy.

In summary, the MSE values provide insights into the accuracy of each model's predictions. A lower MSE indicates a closer fit to the actual data and, hence, a more accurate model. Therefore, the Random Forest model appears to be the most accurate among the three models considered, followed by the Neural Network, with Lasso Regression showing the highest prediction error.

Discussion

In this study, the aim was to enhance the Player Efficiency Rating (PER) metric by incorporating advanced statistical methodologies, including Lasso Regression, Random Forest Regression, and Neural Networks. Each of these models aimed to optimize the weightage of individual statistical components within the PER formula and provide a more comprehensive assessment of player performance, including defensive contributions.

Lasso Regression:

The Lasso Regression Model demonstrated its potential in improving PER calculations. The observation of a bell-shaped curve in the histogram of scaled predicted PER values is an encouraging sign. This distribution suggests that the model effectively captures underlying patterns in player performance, closely mirroring the central tendencies in the dataset. However, comprehensive validation and further analysis are essential to establish its predictive accuracy rigorously.

Random Forest Regression:

The Random Forest Regression Model showcased robustness in predicting PER values, effectively capturing variations and trends within the dataset. The Mean Squared Error (MSE) value of 255.7421 indicates the model's ability to align predicted and actual PER values, with a noticeable concentration of predictions within the correct range. Nevertheless, discrepancies

between predicted and actual values highlight areas for improvement, especially in cases where the model overestimated PER.

Neural Network:

The Neural Network (PyTorch) model exhibited promising predictive capabilities for Player Efficiency Rating. Its histogram of scaled predicted PER values formed a bell-shaped curve, indicating that the model effectively captured underlying data patterns. This suggests that the model has substantial potential to enhance PER evaluations in basketball analytics. However, thorough validation and comparison with real-world performance data are necessary to assess its true effectiveness.

Conclusion

Collectively, the findings underscore the importance of incorporating advanced statistical methodologies to refine PER calculations. The models explored in this study provide valuable insights into player performance assessment, particularly in considering defensive contributions. While each model exhibited promising results, the road to enhancing PER metrics in basketball analytics is ongoing. To advance this research further, it is recommended to conduct comprehensive validation against real-world performance data, explore additional machine learning techniques, and integrate domain-specific knowledge. Collaboration with basketball analysts and experts is vital to fine-tune these models and ensure their practical applicability in the larger organizations, even up to the leagues such as the NBA. In conclusion, this study showcases the potential of machine learning models to revolutionize player performance evaluations in basketball. By providing more accurate and comprehensive assessments in some areas and needs for improvements in others, these models empower teams, coaches, and analysts to make informed decisions, ultimately shaping the future of the game.

References

[1] Josh Gonzales, P. (2020, January 21). Problems with per in the NBA. Medium. https://t.ly/bJ7Hn
[2] Vangelis Sarlis, Christos Tjortjis (2020, May 23). Sports analytics - Evaluation of basketball players and Team Performance. https://t.ly/nuSbE
[3] Hollinger, J. (2003). ”Introducing PER.” https://tinyurl.com/bballRefpage
[4] E. Turban, R. Sharda, D. Delen, Decision Support and Business Intelligence Systems, Vol. 9, ninth ed., Pearson, 2011.
[5] A. Senderovich, A. Shleyfman, M. Weidlich, A. Gal, To aggregate or to eliminate? optimal model simplification for improved process performance prediction, Inf. Syst. (2018) 1–16.
[6] B. Gerrard, Moneyball and the role of sports analytics: A decision theoretic perspective, in: North American Society for Sport Management Conference,NASSM 2016, 2016, pp. 2010–2012, no. Nassm