Computational Analysis of NBA Players with Machine and Deep Learning

Oliver Geun Hyung Park

Received December 02, 2024
Accepted April 30, 2025
Electronic access May 15, 2025

This paper analyzes the relationship between NBA (National Basketball Association) players' performance statistics and salaries. It uses machine learning (ML) and deep learning (DL) models to discover the impact of each player's statistics. The NBA stats are comprised of 384 players in the 2023-2024 NBA season, with 64 different performance statistics and injuries. The machine learning model resulted best with the Random Forest model. The model discovered the five key contributing statistics that align with the salaries: Point Per Game (PPG), Minutes Per Game (MPG), Field Goals Made Per Game (FGM/G), Country, and Age. The model discovered two subjective values: Country and Age, the stats that are not related to players' performances. Through the SHAP Plot, we found the percentage errors of key players to find if they are overpaid or underpaid due to other contributing factors.

Introduction

The National Basketball Association (NBA) is among the most successful and profitable sports franchises [1]. With its high-profit businesses, thirty different teams are also money-makers. The Golden State Warriors, a team in California, earns an annual revenue of 800 million dollars. Followingly, the New York Knicks produce 278 million dollars per year. As the yearly revenues of teams are skyrocketing, the players salaries reflect the notion. Jayson Tatum, a player in the Boston Celtics, contributed to the 2024 NBA Championship and signed a five-year contract extension this July, about 314 million dollars, which is higher than some franchises annual revenues. In professional sports, salary negotiations are one of the most critical factors for both team management and individual players, often involving major agencies and reflecting significant economic value. Therefore, predicting player salaries is not only a challenging task but also an increasingly important issue in the NBA [2] and various other professional sports leagues [3]. While traditional statistical methods have been widely used in the past, recent advancements in machine learning have led to more sophisticated approaches being actively applied to salary prediction [4].

In past NBA salary prediction studies, multiple regression models were used to predict salaries based on player statistical data. Additionally, statistical techniques such as ANOVA (Analysis of Variance) were employed to demonstrate that salary differences exist across different player positions [5]. Subsequently, machine learning-based algorithms, such as K-Nearest Neighbors (KNN), were introduced, leading to the adoption of various machine learning techniques for NBA salary prediction [6]. With the advancement of various algorithms [7], new models have been applied to similar datasets based on player statistics. However, it is important to recognize that player salaries are not solely determined by statistical performance but can be influenced by a wide range of external factors.

This research will explore the factors contributing to players signing major contract deals. Teams invest in players to have the best players in the league to ensure their winning percentages, and, as a result, players receive salaries to maximize the teams revenue. Players performance and skills are crucial to the team and the game. They are considered a production factor in the professional sports industry, as fans and teams heavily invest in them, and their salaries are affected by these factors. Players continue to enhance their skills as the teams aim to recruit the best players in the league. As their performance and skills increase, it correlates to positive results in their salaries. With the development of machine learning, this research will use Random Forest, k-nearest Neighbor, AdaBoost, Gradient Boost, Linear Regression, and Lasso Regression to present several visual charts and tables of how the variables affect players salaries. Data were collected from analyzing players biographies and statistics from the NBA, Fox Sports, and ESPN. Despite similarities with previous research [5-7], this research sets itself apart by adding injuries, ages, and international or United States players to dive deeper into the fairness and difference that stats make.

This research aims to find the relationship between players performance and NBA salaries and visualize fairness in the sports industry. To do so, we looked into statistical and biological variables affecting the wages players receive, finding the most impactful variable through Feature Importance. We also measured how valuable variables are to salaries using the

SHAP value. As international players compete with the highly-prospected United States for sports in the league, the influx of diverse players continues to increase. Understanding how dynamics influence fairness is essential, with factors such as geographic origin and individual achievements contributing to variations in player contracts and opportunities.

Method

Data Collection

In the process of collecting statistics for players. This research used Selenium, an open-source tool that automates web browsers. We used credible sports analytics sources like Fox Sports[1] and ESPN[2] to collect reliable and valuable data sets. For gathering information on salaries, we used Hoopshype [3] to get access to 2023 $\sim$ 2024 NBA Players salaries. To represent the best values that would contribute to the salaries of the NBA players. In total, we found that there are currently 387 active players and chose 64 variables that contribute to their salaries. These are essential variables to notice:

Table 1 List of variables collected contributing to the salary of players

Variable	Detail
PPG	$\text{Points Per Game} = \frac{\text{Total Points Scored}}{\text{Number of Games Played}}$
MPG	$\text{Minutes Per Game} = \frac{\text{Total Minutes Played}}{\text{Number of Games Played}}$
FGM/G	$\text{Field Goals Made Per Game} = \frac{\text{Total Field Goals Made}}{\text{Number of Games Played}}$
Age	Age of players from the dataset
Country	Whether the player is from the United States or International
PTS/48	$\text{Points Per 48 Minutes} = \frac{\text{Total Points}}{\text{Total Minutes Played}} \times 48$
TS%	$\text{True Shooting} = \frac{\text{Total Points}}{2 \times (\text{Field Goals Attempted} + 0.44 \times \text{Free Throws Attempted})}$
TPG	$\text{Turnovers Per Game} = \frac{\text{Total Turnovers}}{\text{Number of Games Played}}$
PPS	$\text{Points Per Shot} = \frac{\text{Total Points}}{\text{Total Field Goals Attempted (FGA)}}$

Injury Data Collection and Embedding

To collect the injury histories of NBA players, we used Selenium to gather each players384 players in totalinjury types and number of injuries. Analyzing injuries is crucial in predicting NBA salaries, as they can lead to a players physical ability decline. Therefore, they can affect their performance on the court, which may reduce their overall value to a team in our analysis. We first transform injury data into a multidimensional vector that captures the various injury types and their frequencies. The

1. https://www.foxsports.com/nba
2. https://www.espn.com/nba/
3. https://hoopshype.com/salaries/players/

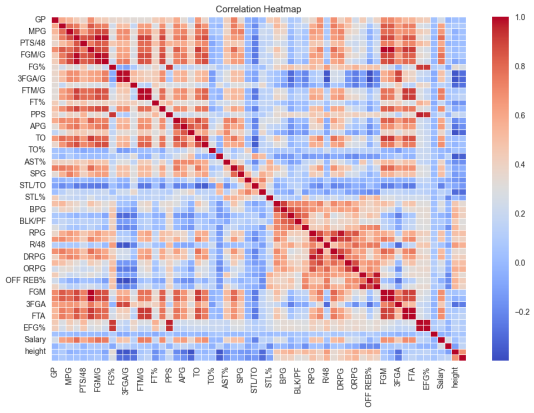

Fig. 1 Correlation plot Figure 1 presents the correlation plot of the collected data. The variable with the strongest correlation to Salary in this study is PPG (Points Per Game), which shows a high correlation coefficient of 0.77, indicating a strong positive relationship with salary.

injury data for players was collected based on when the injury occurred during the season and the type of injury sustained, as shown in Table 2. If a player had no injuries, the corresponding value was replaced with 0.

Table 2 Types and Frequency of Injuries of Bogdan Bogdanovic

Injury Type	Frequency
Illness	1
Ankle	8
Knee	15
Back	1
Hip	2
Quad	2
Rest	2
Conditioning	1
Hamstring	9
Right Ankle	1
Right Knee	1
Foot	1
Sore left hamstring	1
Sore right ankle	1

These injury types will turn into an injury vector: [15,8,7,2,2,1,1,2,1,1,1,1,1]. Afterward, we used the Deep Learning Embedded Layer to reduce the dimensionality of the vectors' components and maximize the variance in both positive and negative directions, resulting in an example of Bogdan Bogdanovic: -0.0238670. This embedded output would contribute to salary prediction by providing a summarized, lower-dimensional feature that covers the players injury history.

Exploratory Data Analysis (EDA) and Correlations between different variables

In NBA games, the most contributed variable in determining the Most Valuable Player is their statistics, which include Point Per Game (PPG), Minute Per Game (MPG), Assist Per Game (APG), Rebound Per Game (RPG), Games Played (GP), and other contributions to the game. In addition, the most-paid player in the NBA in the 2023 2024 season was Stephen Curry from the Golden States Warriors, who showed outstanding performance by GP of 79 out of 82 games. His MPG was 32.7 in 48-minute games, PPG was 26.4, APG of 5.1, and RPG of 4.5. In total, Stephen Curry received $55,761,216.

Data Preprocessing

Min-Max Normalization

Variables like salary consist of large numerical values, which would be hard to interpret. Normalization makes comparability across different player salaries easier. This helps prevent higher salary values from unfairly disproportionally influencing the analysis, which might lead to a biased approach in models relying on numerical inputs. Especially in the dataset collected for this study, certain star players receive exceptionally high salaries, which could be considered outliers. However, instead of removing these outliers, we chose to retain them to analyze their impact while mitigating their influence by applying Min-Max Normalization. Min-Max Normalization scales all values within a range of 0 to 1, making it easier to compare different variables and reducing the effect of extreme values (outliers). This method is particularly effective for variables with a wide range of values, such as salaries. By using Min-Max Normalization, we prevent the model from being overly influenced by high-salary players, ensuring a more balanced learning process.

0.1 Categorization of Player Country using Rule-Based Encoding

To collect the country data to determine any biases against being players from the United States or internationally, we used the Selenium tool to extract the country information from Fox Sports. In the raw dataset, countries and regions were represented by the state or government acronyms. If all country codes were used, countries with a small number of players would have minimal impact on the model during the variable transformation process. Nevertheless, to incorporate demographic information, we categorized the data into two broad groups to ensure meaningful analysis. To simplify these data, we implemented a rule-based approach to categorize entries as either U.S. or international:

- U.S. Entries: Players identified by a common state code, such as California, are assigned a binary label of 0.

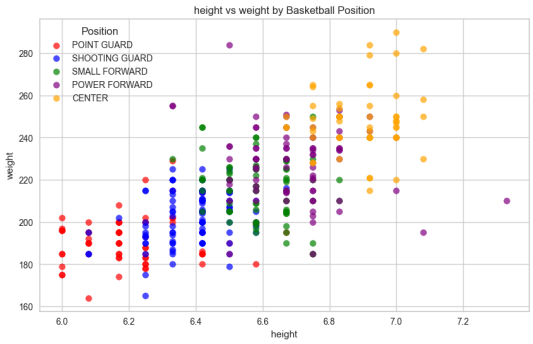

Fig. 2 Scatter Plot Height vs Weight by Basketball Position

Fig. 3 Box Plot for Salary Distribution by Position

- International Entries: For example, Spain was recognized as ESP and assigned a binary label of 1.

One-Hot Encoding Representation of Player Positions

In team sports, certain positions, such as goalkeepers in soccer or pitchers and catchers in baseball, are designated for specific tasks. In basketball, players are classified into five distinct positions, but in the dynamic nature of the game, strict positional boundaries can be difficult to maintain. Through pseudo-positioning or switching positions, players can strategically shift their roles during the game, making positional classifications more fluid. Figure 2 illustrates the distribution of height and weight based on player positions. The graph clearly shows that centers (yellow) tend to be taller and heavier than point guards (red). This suggests that in basketball, a player's physical attributes significantly influence their position.

Figure 3 presents a box plot of salaries by position, revealing that point guards generally receive higher salaries compared to other positions. To train a machine learning model, all prepared data must be converted into numerical values. There are two main methods for converting categorical text data into numerical values. One approach is to assign numerical values to each class, which is useful for binary classification or when classes have a natural order, allowing numbers to be assigned accordingly.

However, there are five positions in basketball: Point Guard, Shooting Guard, Small Forward, Power Forward, and Center. As the positions are represented by string variables in the data table, we would need more than this to create and analyze the data. Therefore, we used one-hot encoding to transform the string variable into a binary vector that machine learning algorithms can interpret. This ensures that no ordinal relationship is assumed between the categories, which is essential when dealing with no-numerical data. So, the variable position was defined like the binary vectors below:

Table 3 One-hot encoded table of Golden State Warriors starting roster

Player Name	Position	Transformed Vector
Stephen Curry	Point Guard	[1,0,0,0,0]
Klay Thompson	Shooting Guard	[0,1,0,0,0]
Andrew Wiggins	Small Forward	[0,0,1,0,0]
Draymond Green	Power Forward	[0,0,0,1,0]
Trayce Jackson Davis	Center	[0,0,0,0,1]

Modeling with Machine and Deep Learning

Machine Learning (ML) model is a subset of Artificial Intelligence (AI) focused on building algorithms that allow computers to learn from and make decisions based on data to find patterns and make predictions. Deep learning is a subset of machine learning that focuses on algorithms inspired by the structure and function of neural networks. These algorithms use multiple layers to extract higher-level features from raw data involving images, text, and speech.

Model with Machine Learning

We employed five machine learning models: Linear Regression, AdaBoosting Generator, Random Forest Generator, Lasso Regression, and k-Nearest Neighbors. Each model was evaluated using the PyCaret Regression Module to determine the performance metric that best represents the correlation between the variables and player salaries.

Linear Regression

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

$\hat{y}$: Predicted value (e.g., players salary)
β_0 : Intercept (the salary when all features are zero)
$\beta_1, \beta_2, \ldots, \beta_p$: Coefficients (weights) for each feature
$x_1, x_2, \ldots, x_p$: Values of the independent variables (features like points per game, assists, rebounds)
p : Number of features

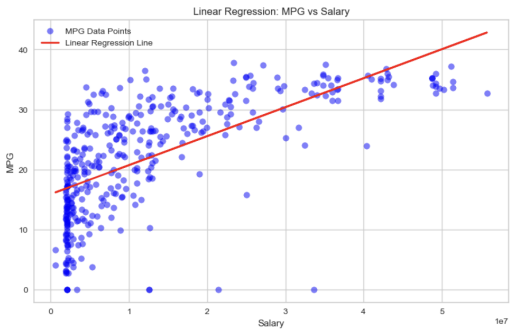

Fig. 4 Linear Regression Model between Salary and Minutes Per Game (MPG)

Random Forest Regressor

$$\hat{y} = \sum_{t=1}^{T} f(t)X$$

Given:
T : the number of decision trees in the forest $f(t)X$: the prediction from the t-th decision tree $\hat{y}$: the final predicted value obtained by averaging all tree outputs

We used a Random Forest model as it handles complex relationships between variables and performs well when interacting with the features. To improve the models accuracy, we use the Random Forests hyperparameters using PyCarets tuning model, optimizing to help the model perform better by finding the ideal number of trees and depths that minimize prediction errors. As a result, we can see the features that contribute highly to players salaries.

K-Nearest Neighbors (KNN) To predict a players salary, KNN compares the players stats to those of similar players. For example, as Jaylen Brown and Jayson Tatum have identical points per game, assists, and rebounds, KNN will predict that their salaries should be similar.

After analyzing the data based on players statistics, the KNN model predicted Jayson Tatums salary of $42,356,189.20, while he only received $34,848,340.0 0, giving the percent error of 21.54%. Meanwhile, the model predicted Jaylen Brown to be $41,671,999.40, receiving $49,700,000.00, giving a 16.15% percent error.

$$\text{Percent Error} = \frac{\text{—Predicted Value - Actual Value—}}{ActualValue} \times 100$$

Deep Learning

To integrate deep learning into our study, we used TensorFlow, an open-source machine learning framework, to build

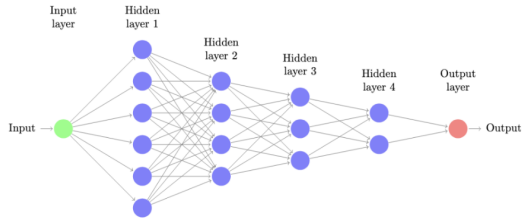

Fig. 5 TensorFlow Neural Network Model Architecture

and evaluate. Using TensorFlow enabled us to create a neural network that learned non-linear relationships between the players performance and salaries. The TensorFlow neural network was assessed using R^2, RMSE and MAPE. Despite the neural network capturing complex relationships between variables, it required longer training times than traditional machine learning models.

Table 4 TensorFlow Neural Network Architecture

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	4,160
dense_1 (Dense)	(None, 32)	2,080
dense_2 (Dense)	(None, 16)	528
dense_3 (Dense)	(None, 8)	136
dense_4 (Dense)	(None, 1)	9

We used the input layer as the entry point for our data, receiving various player performance metrics, such as GP (Games Played), PPG (Point Per Game), MPG (Minutes Per Game), and other relevant statistics. The hidden layer is initially composed of 64 neurons that process the input features and begin to extract patterns from the data. With 32 neurons, this layer captures more abstract representations of the data, building on the features identified in the previous layer. The third layer contains 16 neurons to refine the models understanding by complex combinations of features. The fourth layer enhances the models ability to distinguish subtle patterns, contributing to its predictive power with 8 neurons. The final layer is a single neuron that produces the predicted salary.

Performance Metric

R^2 (Coefficient of Determination)

In our research, we used R^2 to help understand how well our predictive models explain the variance in players salaries based on their performance metrics. A higher R^2 value indicates that the model effectively describes the relationship between player performance and salary.

$$R^2 = \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

SS_{res} represents the total squared difference between the observed data and the predicted data. It also SS_{tot} represents the total squared difference between the observed data and the mean of the observed data, which quantifies the total variability.

RMSE (Root Mean Squared Error)

When measuring the accuracy of our salary prediction models, RMSE measures the average magnitude of the prediction errors, estimating how far off the machine learning models predicted values are from the actual values. A lower RMSE value will indicate that the models predictions are more accurate, representing a lower difference between the predicted and actual values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

y_i is the actual value for the i-th players salary and $\hat{y}_i$ is the predicted value for the i-th players salary. n is the number of observations of players.

MAPE (Mean Absolute Percentage Error)

When using models like Random Forest, k-Nearest Neighbor, and other machine learning models, we applied MAPE to determine the accuracy and the precision of the model between the performance statistics. If we find that a model has a lower MAPE, it indicates that it is more accurate at predicting salaries as a percentage of actual values.

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

y_i is the actual value for the i-th players salary and $\hat{y}_i$ is the predicted value for i-th players salary. n is the number of observations of the players, and the absolute value will measure the error without regard to positive or negative.

Results

Result Table

We have evaluated each machine learning model through the PyCaret module. The Random Forest Regressor was the most accurate and reliable model for predicting NBA player salaries. Its combination of high accuracy, training time, and low error metric helps us to determine the best model. While other models like AdaBoost and Gradient Boosting performed well, Neural Network did not deliver the expected results, highlighting the importance of carefully selecting the suitable model for this study.

From Table 4, the Random Forest Regressor demonstrated the best performance in terms of R and RMSE. However, in

Table 5 Performance Metrics of Each Machine Learning Model with PyCaret Regression. Bold text represents the best result.

Model	R2	RMSE	MAPE	TT (Sec)
Random Forest Regressor	**0.6704**	**7239099.146**	0.8009	0.087
AdaBoost Regressor	0.6615	7372683.47	1.1702	0.023
Gradient Boosting Regressor	0.6586	7392961.423	**0.7902**	0.055
Light Gradient Boosting Machine	0.6465	7558976.873	0.8607	0.16
Neural Network	0.6265	8124748.24	2.408	4
Elastic Net	0.5971	8198550.043	1.0616	0.014
Ridge Regression	0.5805	8363087.914	1.0654	0.012
Lasso Regression	0.5652	8521800.742	1.1254	0.014
Lasso Least Angle Regression	0.5526	8628420.82	1.149	**0.01**
Linear Regression	0.5526	8628476.766	1.1491	0.333
Orthogonal Matching Pursuit	0.5381	8782784.175	1.1535	0.009
Decision Tree Regressor	0.3342	10230868.61	0.9179	0.013
K Neighbors Regressor	0.1829	11593246.1	1.2955	0.014

terms of MAPE, the Gradient Boosting Regressor achieved the best performance. The TT (Training Time) metric measures the time required to train each model, and the Lasso Least Angle Regression recorded the fastest training speed. In general, there is a trade-off between model accuracy and training time in AI models. Achieving higher performance often requires longer training times. In this experiment, the Neural Network required a relatively long training time. However, since the dataset size was insufficient to fully train a deep learning model, its performance was relatively lower than expected.

Therefore, to compare feature importance, we selected the Tree-based Random Forest Regressor. Although MAPE was the second best for this model, the difference compared to the best-performing Gradient Boosting Regressor was minimal (approximately 0.01). In terms of training time, as expected, training multiple trees requires more computational resources than simple regressors. However, given the relatively small dataset, the model still produced results quickly, even on a standard personal computer. More importantly, the primary objective of this study is to analyze the impact of various features on salary based on a highly interpretable and well-performing model. Considering these factors, we ultimately selected the Random Forest Regressor for further analysis.

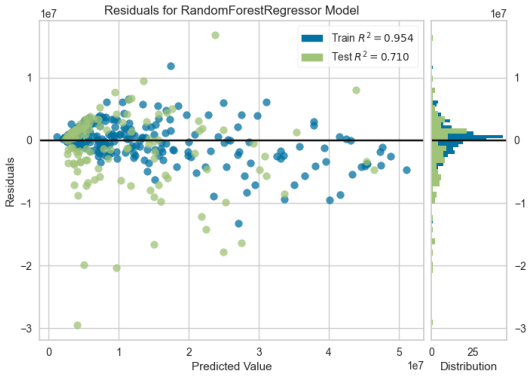

Fig. 6 Residual Plot on the performance of Random Forest Regressor model

Hyperparameter Tuning and Cross Validation

The Residual Plot shows the performance of the Random Forest Regression model, which includes errors plotted against the predicted salaries. The R^2 values for the training are 0.954, and the testing is 0.70. While the model fits the training data well, we cannot generalize its effectiveness to new data.

Table 6 Grid Search Parameter Selection for Random Forest Regressor

Parameter	Variables
n_estimators	[100, 200, 300, 500]
max_depth	[10, 15, 20, None]
max_features	['sqrt', 'log2']
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 2, 4]

The residual plot shows a significant performance gap between the training and test data, indicating that the model has overfitted to the training set and lacks generalization. To address this overfitting issue, we fine-tuned the Random Forest model by adjusting its hyperparameters to find the optimal configuration using Grid Search. Grid Search systematically tests each combination of predefined hyperparameter values to determine the best-performing parameters for the model. The hyperparameters used in the Grid Search process are listed in Table 5, and the final selected parameters are: 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100.

Since the dataset in this study is relatively small, splitting it into training and testing sets using a fixed ratio may lead

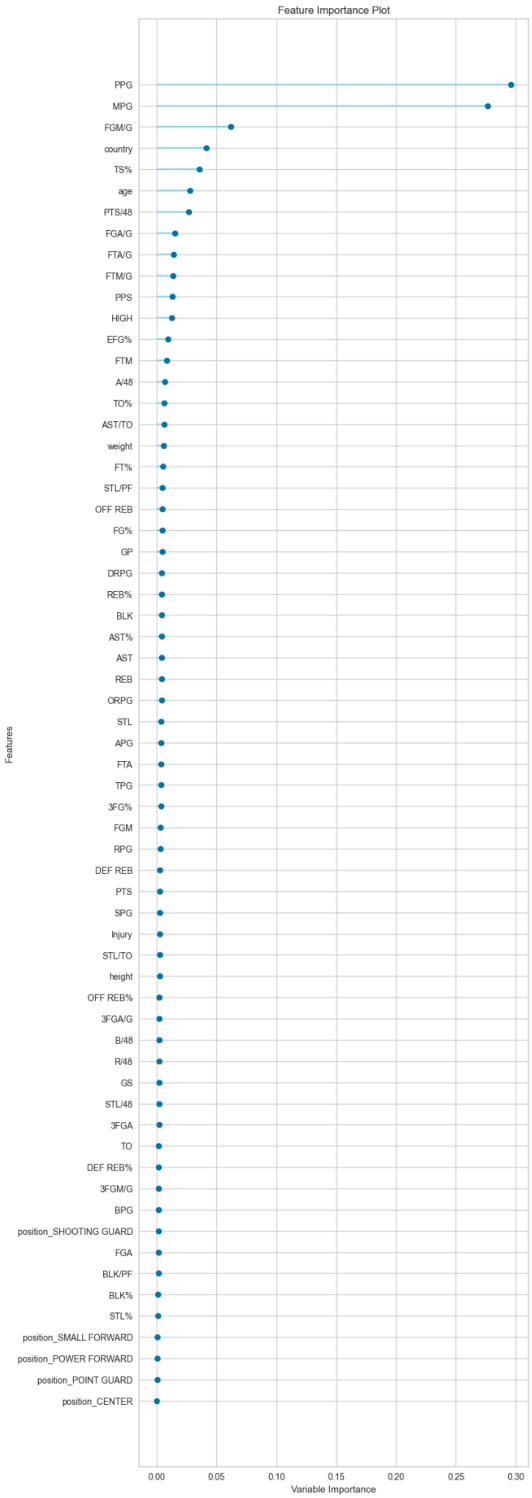

Fig. 7 Random Forest Feature Importance Plot on players statistics

Table 7 5-fold cross-validation results for the Random Forest Regressor

Fold	R2	RMSE	MAPE
1	0.8201	5924263	0.7553
2	0.698	7328756.4	0.7991
3	0.7782	7245229.8	1.0099
4	0.803	5486239.5	0.6935
5	0.6877	7172475.3	0.6456
Mean	0.7574	6631392.8	0.7807
Standard Deviation	0.0609	861291.29	0.1409

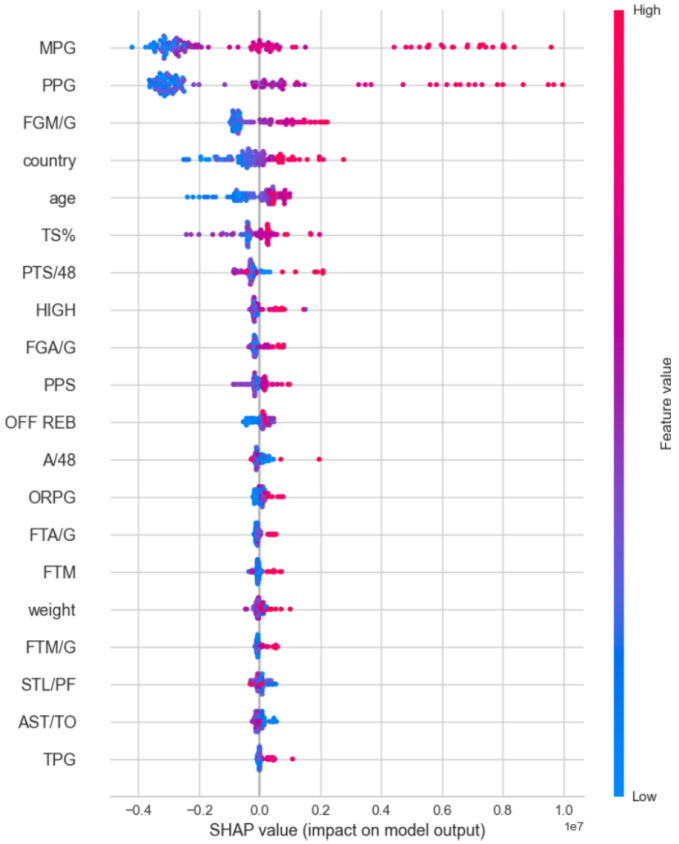

Fig. 8 SHAP Plot on players statistics

to overfitting, particularly if a randomly selected subset contains limited data points. To address this issue, we applied 5-fold cross-validation on the final Random Forest Regressor model. Table 6 presents the results of this process. The mean R score shows an improvement compared to the model before hyperparameter tuning, indicating better overall performance. Additionally, the lower standard deviation suggests increased stability and consistency across different data splits.

Feature Importance

We used a Random Forest model as it handles complex relationships between variables and performs well when interacting with the features. To improve the models accuracy, we use the Random Forests hyperparameters using PyCarets tuning model, optimizing to help the model perform better by finding the ideal number of trees and depths that minimize prediction errors. As a result, we can see the features that contribute highly to players salaries.

SHAP Plot

In addition to the Feature Importance plot, we used the tuned model to SHAP Plot from Figure 8 to determine different models impacts. The SHAP (SHapley Additive exPlanations) Plot provides insights into how each feature influences the Random Forest model's salary predictions. It quantifies the magnitude and direction of a features impact on the predicted salary. The x-axis represents the impact of each feature, as the positive values indicate the salary prediction is higher, and the negative will pull the prediction lower. The y-axis represents the players' performance stats. The four statistics that impact the players' wages are MPG (Minutes Per Game), PPG (Points Per Game), FGM/G (Field Goals Made Per Game), and Country. In addition, Age and Country influence the wages, indicating that players experience and country of origin play an essential role in determining them. As shown in the Feature Importance Plot, MPG, PPG, and FGM/G are critical in determining the salary. Unlike the Feature Importance Plot, the SHAP Plot regarded Age (Years of Experience) as higher, which can be interpreted as younger players who entered the league would get lower wages than veteran players. The country also plays a role in determining the salary, which doesnt impact on-court performance.

Player Prediction

We used SHAP (SHapley Additive expLanations) to interpret how individual features influence the Random Forest models salary prediction. The model predicts the salary for a specific player by removing irrelevant columns. It would compare the predicted salary with the actual salary to calculate the percent error and assess prediction accuracy. The overall percent error in the test dataset is 10.2%. For further analysis, we selected

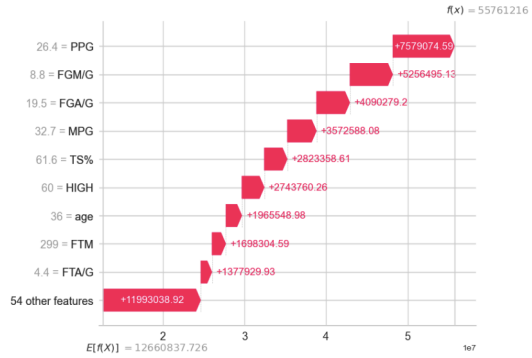

Fig. 9 SHAP Machine Learning Model Results for Stephen Curry

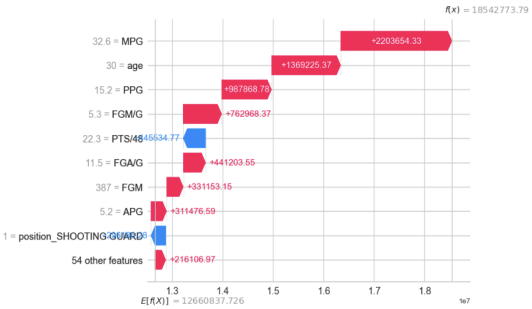

Fig. 10 SHAP Machine Learning Model Results for Derrick White

Stephen Curry, the player with the highest salary, and Boston Celtics Shooting Guard Derrick White, the player with the lowest salary. Stephen Curry (Point Guard for the Golden State Warriors) was evaluated as the highest-paid player in the 2023 $\sim$ 2024 NBA Season and the actual salary was $55,761,216. In our predictive model, his estimated salary was $51,054,933. These two data resulted in a percent error of 8.44%, indicating that our model was relatively close; it slightly underpredicted his actual by this margin. His performance was highly impacted by the Point Per Game (PPG), which was one of the highest-valued characteristics.

Boston Celtics Shooting Guard Derrick White received a predicted salary of $21,904,074.08, but he only received $20,071,429.0 with a percent error of 8.19%. Unlike Stephen Curry, Derrick Whites statistics showed the dominance of the second most important value: Minute Per Game. His position and PTS/48 (Points Per 48 Minutes) negatively correlated with his salary.

The difference between Stephen Curry's and Derrick Whites statistics can be found in the part of APG (Assist Per Game) and the position. In our SHAP graph, we could not find Stephen Currys APG positively correlating and impacting his salary. Still, in Derrick Whites SHAP graph, we witnessed APG positively correlating to his salary. This distinction is the reason for the

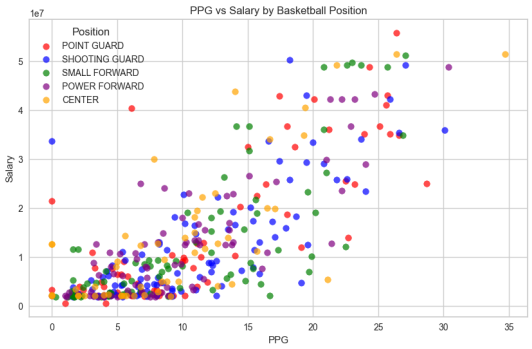

Fig. 11 Scatter Plot PPG vs Salary by Basketball Position

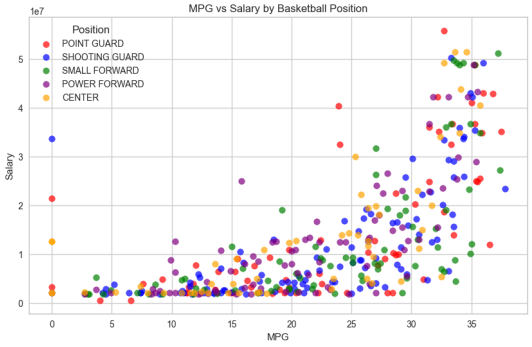

Fig. 12 Scatter Plot MPG vs Salary by Basketball Position

players position. Stephen Curry is a Point Guard, and Derrick White is a Shooting Guard. Point Guard is a player who tends to have the ball most of the time, and the Warriors playmaking is based on Stephen Curry, resulting in Curry with PPG (Point Per Game) contributing to his salary the most. Meanwhile, the Boston Celtics playmaking style doesnt rely on one players choice, which results in Derrick White having positive APG statistics.

In Figure 10, Derrick White's results indicate that the Shooting Guard position has a negative impact on salary. However, when analyzing the relationship between PPG (Points Per Game), MPG (Minutes Per Game), and salary across different positions, as shown in Figure 11 and Figure 12, there is no strong correlation between specific positions and these key salary-determining variables. This contrasts with Figure 2, where certain positions had significant differences in height and weight compared to others. Unlike other sports, where positions have fixed roles, basketball allows for greater positional flexibility during games. Therefore, rather than treating position as an independent categorical variable using one-hot encoding, a more advanced approach could be considered. Embedding layers could be used to capture similarities between positions, or Graph-Based Models could be applied to analyze relationships between different positions. These methods could serve as more effective ways to account for positional effects on salary

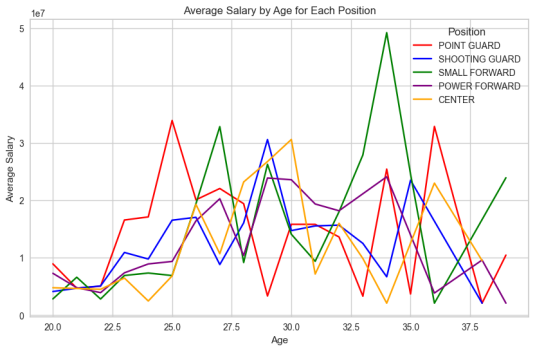

Fig. 13 Average Salary by Age for Each Position

in future research.

Both Stephen Curry and Derrick White, who are in their 30s, demonstrate that age has had a positive impact on their salaries. Figure 13 illustrates the average salary by age group for each position. While it is difficult to identify a clear pattern, players aged 2024 consistently have significantly lower salaries across all positions, confirming that rookie players who have just entered the NBA typically earn lower salaries. Among all positions, point guards experienced the fastest salary growth, reaching their career-high average salary around age 25. In contrast, small forwards show a huge salary increase after the age of 32, which appears to be strongly influenced by Stephen Currys high salary. This suggests that in the model, age became a key factor due to the contrast between younger and older players, with a notable salary increase around age 25. However, after age 25, there are no consistent salary trends across positions, indicating that salary variations become more position-dependent rather than strictly age-related.

Ethical Impact Analysis

In this study, demographic variables such as Country were included as potential factors influencing player salaries. However, even if the results indicate that the Country variable is partially correlated with salary, this reflects a complex interaction with multiple variables, making it difficult to establish causation. To address the ethical concerns regarding the inclusion of the Country variable, we conducted an additional experiment. As shown in Table 7, the results demonstrate that removing the Country variable from the dataset did not lead to a significant difference in model performance compared to when it was included. This suggests that Country alone does not have a direct impact on player salaries.

Table 8 Impact of Removing the Country Variable on Salary Predictions

Model	R2	RMSE	MAPE
Random Forest with country	**0.6704**	7239099.146	**0.8009**
Random Forest w/o country	0.6685	**7237635.669**	0.7923

Discussion

Lack of Data sets

As this research has only collected players statistics from the 2023 2024 season, players performance and salaries would be outdated as the NAB enters a new season. Many contract extensions or signing contracts happen during the off-season, while the data we collected havent been updated into the 2024 2025 season. For example, Stephen Curry, a point guard from the Golden State Warriors, is regarded as the highest-paid player in the 2023 24 season. Still, recently, Celtics power forward Jayson Tatum signed the largest contract ever in the NBA. However, relying on updated stats can also lead to a lack of accuracy since the leagues sponsorships and other factors contributed to the success of the franchises and changes in wages received by the players. Therefore, we seek improvements in collecting updated data and older statistics for more precise and accurate measurements.

In this study, we incorporated variables such as injuries to go beyond traditional performance-based statistics and demographic information, selecting factors that could influence a player's career longevity and ultimately impact their salary. However, player salaries are not solely determined by individual performance. A team's financial situation can also play a crucial role, and external factors such as marketing value or external influences may contribute to salary decisions, even for players with lower performance levels. Therefore, future research should consider incorporating a broader range of variables to provide a more comprehensive analysis of salary determinants.

Number of Fans for each NBA Team

Successful NBA teams, determined through abundant profits, often have large fan bases worldwide. For example, the Golden State Warriors account for a total value of 7.7 billion dollars, recorded as first in the ranking of NBA team values. The Warriors were able to pay Stephen Curry, the highest-paid player in our analysis, which we can interpret as the teams value can also result in a players salary.

Evaluating WNBA (Women National Basketball Association)

Despite the growing popularity of womens sports, players in the WNBA often earn substantially less salary than players in the NBA. Adhering to statistics and data in our models will not only help to find the disadvantages associated with gender but also reflect how the disparity in pay in media and sponsorship can contribute to players salaries. Additionally, we could find inherent bias associated with how womens sports are valued and how this can contribute to players' performance. This would play a significant role in fostering gender equity in sports. Similar to this study, data on female players can be collected to build an integrated dataset that includes both male and female players, allowing for an analysis of how much the gender variable impacts salaries and providing evidence of gender-based bias. Even without explicitly adding the gender variable, we can apply the same model from this study to female players. If no gender bias exists, the predicted salaries for female playersbased solely on their performance statisticsshould not significantly differ from the actual salaries. However, if there is a large discrepancy between the actual salaries of female players and the models predicted values, it would indicate that gender or other external factors are influencing salaries and negatively impacting the models accuracy.

Conclusion

This research used machine learning and deep learning models to identify the relationship between NBA players performance statistics and their salaries. Leveraging techniques such as Feature Importance and SHAP plots, the study discovered Point Per Game (PPG), Minute Per Game (MPG), Field Goals Made Per Game (FGM/G), Age, and Country as the top five variables influencing salary prediction. While PPG, MPG, and FGM/G are directly associated with on-court performance, Age and Country reflect subjective factors that raise questions about fairness and bias in salary allocation. After the analysis, injuries considered subjective variables, did not impact the salaries as two latter values. The Random Forest model was the most accurate and effective tool to predict players values. It handled non-linear relationships and feature interactions by crucial performance metrics R, RMSE, and MAPE. The models interpretability through feature importance of 64 statistics and SHAP plots provided insights into how various factors contribute to salary outcomes.

References

1 D. N. Mi, M. C. Radu and C. Punescu, Proceedings of the International Management Conference, 2021.

2 J. Zhang, *Highlights in Business, Economics and Management*, 2024, **24**, 1059–1064.

3 J. Park, J. Lee, H. Kim and S. Choi, *International Journal of Sports Marketing and Sponsorship*, 2024, **25**, 382–395.

4 J. P. Lautier, *A new framework to estimate return on investment for player salaries in the National Basketball Association*, 2023, https://doi.org/10.48550/arXiv.2309.05783, Preprint.

5 R. Xiong, Y. Zhang, F. Li and W. Chen, Proceedings of the 8th International Conference on E-business, Management and Economics, 2017.

6 W. Wu, Y. Liu, A. Chen and M. Yang, *Classification of NBA salaries through player statistics*, Sports analytics group at berkeley technical report, 2018.

7 E. zbalta, M. Yavuz and T. Kaya, *Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation: Proceedings of the INFUS 2021 Conference, held August 2426, 2021. Volume 2*, Springer International Publishing, 2021, pp. 189–196.