Hoops or Hashtags: Do on-court skills or social media influence predict basketball salaries? A statistical analysis

Diya Boyapati

Abstract

NBA player salaries can be affected by many things, from how well players play on the court to their popularity off the court. For teams, knowing what affects these salaries can help them make better decisions. For players, it can be useful to understand what they should focus on to boost their earnings. In this project, we have explored which factors have the biggest impact on NBA player salaries. We'll look at different aspects like performance stats and social media presence to see what really matters. This analysis aims to give both teams and players a clearer idea of how salaries are determined.

Introduction

Our primary goal is to uncover which factors have the most significant influence on player compensation. Are salaries predominantly driven by a player's performance metrics, such as points per game, minutes played, and defensive contributions? Or do external factors, such as social media presence and public image, play a more substantial role? We aim to address these questions by analyzing a selection of relevant features and constructing a predictive model to estimate player salaries.

To avoid complications from multicollinearity—a situation where predictors are highly correlated—we have chosen a streamlined set of performance indicators: minutes per game (MPG), points per game (POINTS), and both offensive and defensive real plus-minus values (ORPM and DRPM). These metrics are integral in assessing a player's on-court effectiveness and overall contribution to the team.

In addition to performance metrics, I also consider non-performance-related factors such as player position and social indicators, including Twitter following and Wikipedia page views. The inclusion of these social dimensions aims to shed light on whether a player's marketability and public image contribute significantly to their salary, or if the emphasis remains solely on their on-court achievements.

For players themselves, this analysis can give a perspective: Should one focus on enhancing their performance and game statistics to secure a higher salary, or should they also invest in building their social media presence and public profile? By examining the relationship between these various factors and player salaries, we provide insights that could help players and teams make more informed decisions about where to allocate their efforts and resources.

I used Data from the Kaggle Dataset “Social Power NBA” combined on-court performance data for NBA players in the 2016-2017 season, alongside salary, Twitter engagement, and Wikipedia pageviews.

Methods

The initial step in our analysis involved performing an Exploratory Data Analysis (EDA) to examine the distributions, correlations, and patterns within the data. I began by reviewing summary statistics for key variables, such as player salaries, performance metrics (MPG, POINTS, ORPM, DRPM), and social indicators (Twitter engagement and Wikipedia page views). This process allowed me to identify potential outliers, trends, and relationships between variables, providing a foundation for building predictive models.

I visualized the relationships between player salaries and various performance and social indicators using scatter plots and correlation matrices. This helped in identifying which factors might have the most significant influence on salaries and guided the selection of variables for our predictive models.

To quantify the impact of each factor on NBA player salaries, I built predictive models using Linear Regression and Polynomial Regression. The goal was to assess how well performance metrics and social indicators explain the variability in player salaries.

Linear Regression(Figures 1.1-1.4): We applied this model to determine the linear relationship between each predictor (e.g., MPG, POINTS, ORPM, DRPM, Twitter engagement, and Wikipedia page views) and player salaries. The model provided coefficients that indicate the expected change in salary for a one-unit change in the predictor, while holding other variables constant.

Polynomial Regression(Figures 1.5-1.6): Recognizing that some relationships might not be strictly linear (e.g. salary), we extended the analysis using Polynomial Regression. This model allowed me to capture more complex relationships and better fit the data where necessary.

To confirm the significance of our findings, statistical tests were conducted, including:

One-Way Analysis of Variance: This test helped to determine whether there are statistically significant differences between the means of salaries across different player positions.

Coefficient of Correlation:The coefficient of determination or R2 represents the proportion of the variance in the dependent variable that can be explained by the independent variable.This was used to measure the strength and direction of the linear relationship between player salaries and each predictor.

Multivariate regression(Figure 3.1): The multivariate regression analysis evaluates the combined influence of various factors on NBA player salaries. By examining these variables together, the model provides a comprehensive understanding of how both on-court performance and off-court popularity contribute to player compensation, highlighting the relative importance of each factor in determining salaries.

Results

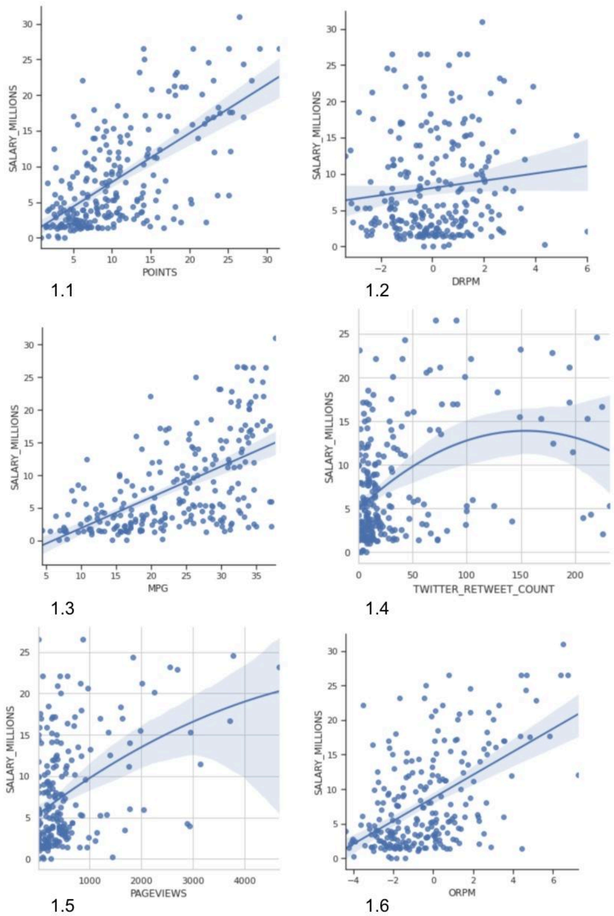

Figure 1

Linear regressions between each predictor and effect on player’s salaries

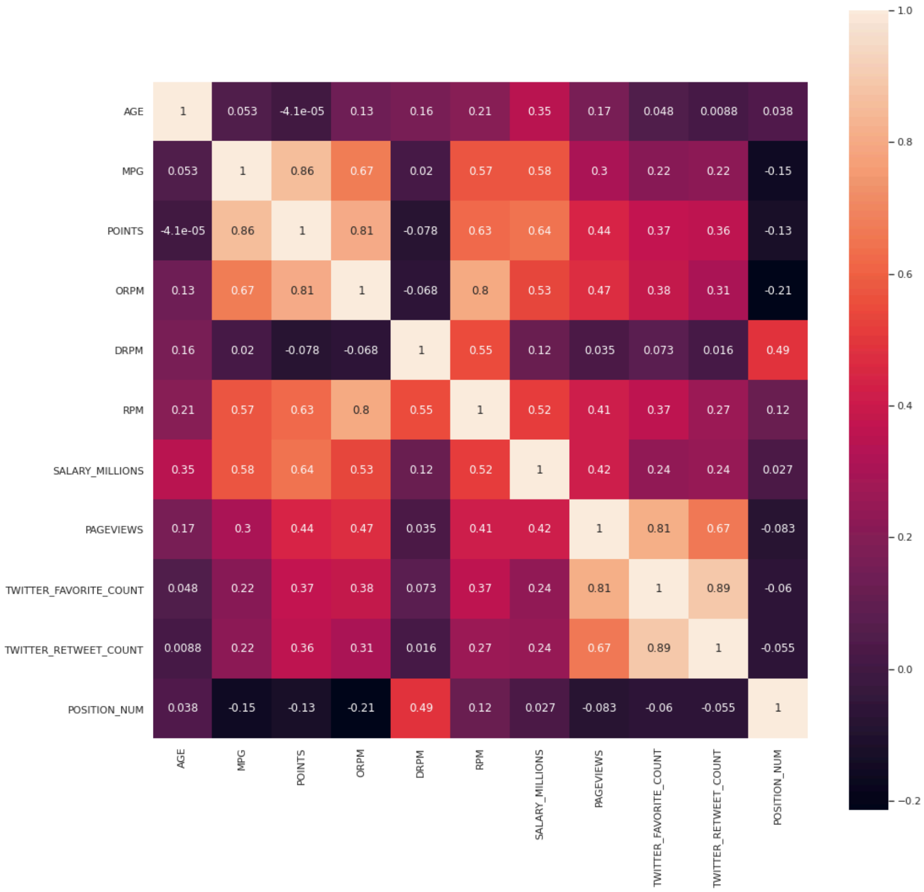

Figure 2.1

Correlation heat map showing the strength and direction of relationships between variables, with color gradients indicating the degree of correlation

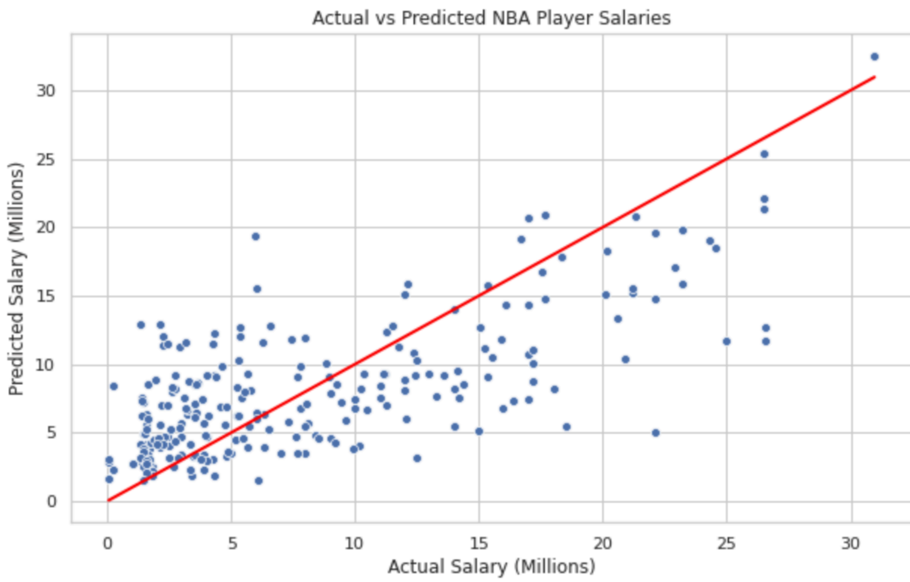

Figure 3.1

Comparison between the player’s actual salary to our model’s predicted salary in millions

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        SALARY_MILLIONS   R-squared:                       0.452
Model:                            OLS   Adj. R-squared:                  0.441
Method:                 Least Squares   F-statistic:                     38.51
Date:                Thu, 29 Aug 2024   Prob (F-statistic):           9.90e-29
Time:                        01:08:26   Log-Likelihood:                -730.19
No. Observations:                 239   AIC:                             1472.
Df Residuals:                     233   BIC:                             1493.
Df Model:                           5
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                    0.4031      1.353      0.298      0.766      -2.263       3.069
PAGEVIEWS                0.0018      0.000      5.144      0.000       0.001       0.003
TWITTER_FAVORITE_COUNT  -0.0020      0.001     -3.483      0.001      -0.003      -0.001
MPG                      0.3008      0.054      5.604      0.000       0.195       0.407
ORPM                     0.6154      0.222      2.769      0.006       0.178       1.053
DRPM                     0.5543      0.212      2.609      0.010       0.136       0.973
==============================================================================
Omnibus:                        7.472   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.024   Jarque-Bera (JB):                7.336
Skew:                           0.421   Prob(JB):                       0.0255
Kurtosis:                       3.164   Cond. No.                     8.34e+03
==============================================================================
```

Figure 4.1

Multivariate regression table with coefficient of correlation among predictors

Minutes Per Game(MPG)

The correlation between minutes per game (MPG) and salary was found to be moderate, with an R^2 value of 0.333, indicating that approximately 33.3% of the variance in player salaries can be explained by the time players spend on the court. The Root Mean Square Error (RMSE) for this model was 5.668, which suggests a moderate level of prediction accuracy. While MPG is a relevant factor, it is not the most powerful predictor of salary. The coefficient for Minutes Per Game of 0.3008 suggests that for each extra minute a player averages per game, their salary increases by about 0.3 million dollars This moderate influence implies that although playing time is important, other factors also play significant roles in determining player compensation. Figure 1.1 illustrates how the MPG, as the independent variable, affects the Salary in millions, the dependent variable, using a linear model.

Points Per Game(PPG)

Points per game (POINTS) demonstrated a stronger correlation with salary, with an R^2 value of 0.413, indicating that players who score more points tend to earn higher salaries, with about 41% of the variance in salaries explained by their scoring performance. The Root Mean Square Error (RMSE) for this model was 5.316, which is lower than the RMSE for MPG, reflecting better prediction accuracy. Points per game emerged as a key performance indicator with a significant impact on salary, underscoring the importance of scoring ability in determining player compensation. Figure 1.2 Illustrates the relation between PPG and Salary. The figure and both the R2 and the RMSE indicate that it is more strongly related with the salary than the minutes per game

*We excluded Points Per Game from the coefficient analysis and multivariate regression results because points scored are closely influenced by Offensive Real Plus-Minus (ORPM). Including both POINTS and ORPM in the model would introduce multicollinearity, which could distort the coefficients and lead to inaccurate interpretations of each variable's impact on salary. By removing POINTS, we ensure that the regression results provide a clearer and more accurate representation of how each independent factor, such as ORPM, contributes to player salary without overlap.

Offensive and Defensive Real Plus Minus(ORPM/DRPM)

Offensive Real Plus-Minus (ORPM) measures a player's average impact on his team's offensive performance, quantified by the points scored per 100 offensive possessions. Conversely, Defensive Real Plus-Minus (DRPM) assesses a player's average impact on his team's defensive performance, based on the points allowed per 100 offensive possessions. ORPM showed a moderate influence on salary, with an R^2 value of 0.285 and an RMSE of 5.896, indicating that a player's offensive contributions are somewhat valued in salary determination. This suggests that players who positively influence their team's offensive output are likely to earn higher salaries, reflecting the premium placed on scoring and offensive efficiency in the NBA. Additionally, The coefficient for Offensive Real Plus-Minus (ORPM) of 0.6145 indicates that, on average, each additional point in ORPM is associated with an increase of approximately 0.6145 million dollars in player salary. This suggests that offensive contributions, as measured by ORPM, are significantly valued in salary determination, highlighting the importance teams place on a player's ability to positively impact offensive performance.On the other hand, DRPM had a weaker impact on salary, with an R^2 value of just 0.013 and an RMSE of 6.894. This indicates that while defensive contributions are important, they are less influential in determining player salaries compared to offensive metrics.The coefficient for **Defensive Real Plus-Minus** (DRPM) is 0.5543, which, while still positively related to salary, is weaker than the coefficient for Offensive Real Plus-Minus (ORPM). This indicates that although defensive contributions are important, they are not as strongly valued in salary determination as offensive contributions. Teams may prioritize players' offensive impact when negotiating salaries, reflecting the higher emphasis on scoring and offensive effectiveness. Figures 1.4 and 1.5 represent the models for DRPM and ORPM, respectively, illustrating the different impacts these metrics have on salary determination.

*DRPM is highly correlated with other predictors (like offensive metrics or points), its individual contribution to explaining salary might be low, which reflects in a low R-squared when used alone. However, when all predictors are included together, multicollinearity can inflate the variance of the coefficients, resulting in some coefficients (like DRPM) appearing larger

Twitter Engagement

The data contained noticeable outliers in Twitter interaction, making it unlikely that a linear model could accurately capture the relationship between Twitter engagement and player salary. To better explore this relationship, a second-order polynomial regression model was constructed, showing a quadratic model. Twitter engagement demonstrated limited predictive power in salary determination, with an R^2 value of 0.235 and an RMSE of 6.070, and The coefficient for **Twitter Favorite Count** is -0.0020, which suggests a small negative relationship with salary. This indicates that an increase in the number of Twitter favorites is associated with a slight decrease in salary. This could imply that social media popularity, as measured by Twitter favorites, does not significantly influence player salaries when other performance metrics are considered. While a social media presence might be important for personal branding, it has a limited direct influence on salary. However, in these polynomial models, Twitter engagement showed potential overfitting, indicating that its relationship with salary is not straightforward and may be influenced by other factors. Figure 1.3 illustrates the dispersion of results between Twitter engagement and salary, showing that while the second-order polynomial model reduced bias, it also increased the variance, leading to a higher risk of overfitting and reduced generalization. Despite these adjustments, Twitter interactions remained a poor predictor of salary compared to POINTS or even ORPM.

Wikipedia Engagement

To better fit and display the results, I built a second-order polynomial regression model for Wikipedia page views, similar to the approach used for the Twitter model. While this non-linear model better explains the relationship between Wikipedia page views and player salary, it is likely to perform poorly when applied to new players and unseen data due to its potential to overfit. Wikipedia page views emerged as a better predictor of salary than Twitter engagement, with an R^2 value of 0.309 and an RMSE of 5.769. This suggests that public interest, as measured by page views, plays a more significant role in salary determination than social media engagement. The coefficient for Wikipedia Page Views is 0.0018, showing a small positive relationship with salary. While this influence is minor, it is stronger than that of Twitter favorites, indicating that public interest has a slightly greater impact on salaries than social media engagement. However, despite its stronger predictive power compared to Twitter, Wikipedia page views still ranked secondary to performance metrics like points per game. This indicates that while public image and visibility matter, on-court performance remains the most critical factor in determining player salaries. Figure 1.6 illustrates the polynomial regression model for Wikipedia page views.

Discussion

The study found that points per game (POINTS) is the most important predictor of NBA player pay, emphasizing the importance of scoring ability in deciding salary. Among the performance measures examined, POINTS had the greatest association with salary, emphasizing its importance in salary calculation. Wikipedia page views were identified as the second most significant component, demonstrating that public interest and visibility, as assessed by page views, play an important effect in player profits. In contrast, Twitter participation had a less significant impact on salary prediction, indicating that while social media presence is important, it is not as influential as other indicators. The study also emphasized the relative value of offensive and defensive contributions. Offensive Real Plus-Minus (ORPM) had a better link with salaries than defensive Real Plus-Minus (DRPM), indicating that teams may emphasize scoring and offensive impact when evaluating player pay. This research implies that offensive capabilities are more highly valued than defensive skills in today's current NBA salary structure. According to the data, athletes who want to maximize their earnings should prioritize improving performance measures, specifically scoring. While establishing a social media presence and public profile might lead to increased pay, these elements are secondary to on-court accomplishments. Players should prioritize their on-court performance, particularly in terms of scoring, in order to increase their market significance. Teams can use these data to guide their salary decisions, focusing more on players' offensive contributions and performance measures during contract discussions. However, they must also evaluate a player's marketability and public image, especially when investing in star players who may boost the team's brand and fan engagement. By combining performance data and social indicators, teams may make more informed decisions that benefit both on-court success and off-court marketability.

References

1. Gift, N. (2017). *Social Power NBA* [A data set containing combined on-court performance data for NBA players in the 2016-2017 season, alongside salary, Twitter engagement, and Wikipedia traffic data]. Kaggle. https://www.kaggle.com/datasets/noahgift/social-power-nba

2. Hayes, A., & Scott, G. (2024). *Multicollinearity: Meaning, Examples, and FAQs*. Investopedia. Retrieved August 20, 2024, from https://www.investopedia.com/terms/m/multicollinearity.asp

3. Pennsylvania State University. (2018). *7.7 - Polynomial Regression | STAT 462*. Applied Regression Analysis. Retrieved August 18, 2024, from https://online.stat.psu.edu/stat462/node/158/

4. Laerd Statistics. (n.d.). *One-way ANOVA - An introduction to when you should run this test and the test hypothesis*. Laerd Statistics. Retrieved August 20, 2024, from https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php

5. Johnson, R., & Wichern, D. (n.d.). *Getting started with Multivariate Multiple Regression | UVA Library*. UVA Library. Retrieved September 1, 2024, from https://library.virginia.edu/data/articles/getting-started-with-multivariate-multiple-regression

6. Lyons, D. R., Jackson, D. E. N., & Livingston, D. A. (2015). Determinants of NBA Player Salaries. *The Sport Journal*. https://thesportjournal.org/article/determinants-of-nba-player-salaries/

7. Berri, D. J., Brook, S. L., Schmidt, M. B. (2007). Does one simply need to score to score? *International Journal of Sport Finance, 2* (4) 190-205.

8. Curcic, D. (2024, March 25). *NBA salaries analysis (1991-2022)*. RunRepeat. Retrieved September 1, 2024, from https://runrepeat.com/salary-analysis-in-the-nba-1991-2019

9. Sigler, K., & Compton, W. (2018). NBA Players' Pay and Performance: What Counts? *The Sport Journal*. https://thesportjournal.org/article/nba-players-pay-and-performance-what-counts/

10. Mertler, C. R. and Vannetta, R. A., Advanced and Multivariate Statistical Methods: Practical Application and Interpretation, 3rd Edition, Glendale, CA: Pyrczak Publishing.

11. Racine County Eye Staff. (2023, May 10). *The Impact of Social Media on the NBA*. Racine County Eye. Retrieved September 7, 2024, from https://racinecountyeye.com/2023/05/10/impact-of-social-media-on-the-nba/