Research Archive of
Rising Scholars (preprint)

Where bright minds share their learnings

NBA Player Evaluation Using a Value-Performance Salary Index | VSPI

Aditya Rath

Abstract

This study aims to develop a new metric to analyze the value of NBA players by analyzing various statistical categories. The study incorporates data from the last three NBA seasons, including regular stats, advanced metrics, and salary information, to create a predictive assessment of player value. The methodology involves building datasets, generating visualizations, and applying machine learning techniques to uncover patterns and relationships within the dataset.

This study examines how these factors contribute to a player's overall impact on their team's success and financial worth to their organization. The results indicate that combining these data points allows for a more nuanced understanding of player value, going beyond traditional evaluation methods that often rely on singular statistics.

The study's findings reveal that the newly developed metric offers a robust tool for NBA teams to make more informed decisions regarding player signings and contracts. This metric can influence roster management and financial strategies in professional basketball by providing a monetary value for assessing the effectiveness of contracts and return on investment. Overall, this research highlights the importance of a data-driven approach in optimizing team performance and salary-cap allocation.

Introduction

In the competitive world of professional sports, accurately evaluating a player's value is crucial for making informed decisions about contracts and team composition. The National Basketball Association (NBA) is no exception, where teams invest substantial resources in acquiring and retaining talent. They aim to put together a roster with the right players to win an NBA title or eventually compete for one. With the NBA salary cap and luxury tax regulations in place, teams must carefully allocate their financial resources to maximize their competitiveness while avoiding financial penalties. This makes efficient player valuation a key aspect of long-term success.

Traditional methods of assessing player value often focus on individual statistics such as points scored, rebounds, or assists. While these metrics provide insight into a player’s performance, they may not fully capture a player's overall contribution to the team's success or their financial worth to the organization. A high-scoring player is not necessarily the most impactful player on the court, just as a defensive specialist or playmaker may provide immense value despite modest scoring numbers. To create a more accurate evaluation, teams must consider advanced metrics that reflect a player's all-around impact on winning.

A great example of two unique players and their impact can be seen in Draymond Green of the Golden State Warriors and, historically, Dennis Rodman of the Chicago Bulls. These players defy the conventional expectation that a high-value player must be a dominant scorer or a one-dimensional defensive specialist. Instead, they bring a diverse skill set that contributes to team success in unique ways. When taking a closer look at Green, one can see how low his scoring numbers are compared to other stars in the league—he averaged just 8.5 points per game (PPG) in 2022, which is significantly lower than the top scorers, who put up around 25 PPG. However, his value extends far beyond scoring. He also averaged 7.2 rebounds, 6.8 assists, and 1.8 combined steals and blocks per game while maintaining strong durability. Looking deeper, his Player Efficiency Rating (PER) is well above average, and his Win Shares (WS) contribute positively to his team's success. These advanced metrics highlight how Green's strengths—defense, playmaking, and leadership—elevate his overall impact beyond traditional box score stats.

This research aims to bridge this gap between pure scoring and other advanced metrics by developing a comprehensive metric that integrates various performance indicators and financial data. By leveraging advanced analytics and data-driven models, this study will quantify player impact in a way that aligns with modern team-building strategies. Such research is valuable to all NBA teams as owners and general managers strive to optimize spending and maximize performance. Ultimately, this project will contribute to the broader field of sports analytics, offering a data-driven framework for assessing player value with greater accuracy.

Background Research

The assessment of player value has been a subject of interest in both sports analytics and economics. Previous studies have explored various methods to quantify a player's impact, from traditional box score statistics to more advanced metrics like PER and WS. These methods often rely on specific statistical inputs and may not consider the full scope of a player's influence on the game, which is necessary to provide a more accurate value.

Some works, like the Dunks and Threes platform, pioneered Estimated Plus-Minus (EPM), a statistic measuring a player's offensive and defensive impact using play-by-play data (Dunks & Threes). EPM's regression-based approach to isolating individual player effects influenced this study's methods by taking several years of data. Similarly, the DARKO projection system by AP Analytics uses Bayesian modeling to predict player performance over time (Medvedovsky). Its dynamic approach informed this research's emphasis on continuously updating player valuations with new data, ensuring greater accuracy in predictive modeling.

Aaron Cole Smith's exploration of clustering NBA players using unsupervised machine learning offered valuable insights into identifying player archetypes, such as scorers, facilitators, or defenders (Smith). Inspired by this, this study employs clustering techniques to segment players into performance-based groups to identify patterns in their overall value. Jesse Blant's salary regression analysis further contributed to this work by illustrating the relationship between on-court performance and financial compensation, a critical aspect of linking player contributions to market valuation (Blant). Organizations like the NBA Official Website and Basketball Reference have also been instrumental in advancing player analytics through comprehensive data collection and visualization tools ("NBA Advanced Stats"; "NBA Player Contracts"). Their platforms provided crucial datasets and analytical techniques that improved the depth and scope of this research.

Recent advances in machine learning and data analytics have opened new avenues for player evaluation. This research builds on these existing methodologies, incorporating multiple years of statistical and financial data to create a new metric that aims to provide a more comprehensive evaluation of player value.

Research Methodology

Data Collection

The data for this study were collected from publicly available sources, such as NBA official statistics, salary records, and advanced analytics databases. Most of the data collected came from two main statistical trackers: NBA.com and BasketballReference.com. Three main datasets were created to facilitate the analysis:

1. A dataset for player salaries, which included detailed salary information for NBA players over the past three seasons, covering various contract types.
2. A dataset that included standard performance metrics such as points, assists, rebounds, field goal percentage, and minutes played, which are commonly used to assess player performance in games.
3. An advanced metrics dataset, which focused on more complex metrics like Player Efficiency Rating, Win Shares, Player Impact Estimate (PIE), Value Over Replacement Player (VORP), and Net Rating.

The datasets from these sources were organized into CSV files to ensure easy accessibility and allow for straightforward analysis. These datasets formed the foundation of the research process to determine the factors that contribute to player valuation and salary decisions.

Data Preprocessing

The initial stage of the methodology involved the cleaning and preprocessing of the data to ensure its accuracy and consistency. Blocks of text were plugged into a sorting program I created called DataHelper, where stats scattered through the text were output to use in the CSV files. This step was crucial for ensuring the validity of the analysis. Several preprocessing techniques were applied, including the handling of missing values. For missing values, rows with incomplete data were removed because certain players who were being paid didn't have meaningful playing time or contributed no statistics. Additionally, the performance data for each player was aligned with their corresponding salary information, ensuring that each player's salary was matched to their performance for the correct season. These steps ensured that the data was comparable across seasons, teams, and players.

Data Visualization

Following the preprocessing phase, various visualizations were created to explore the relationships between player performance metrics and their salaries. Various graphical methods were employed, including scatter plots and bar charts.

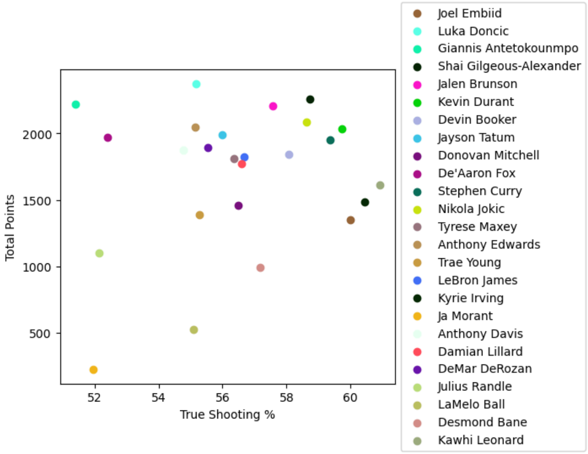


Figure 1: True Shooting % vs Total Points

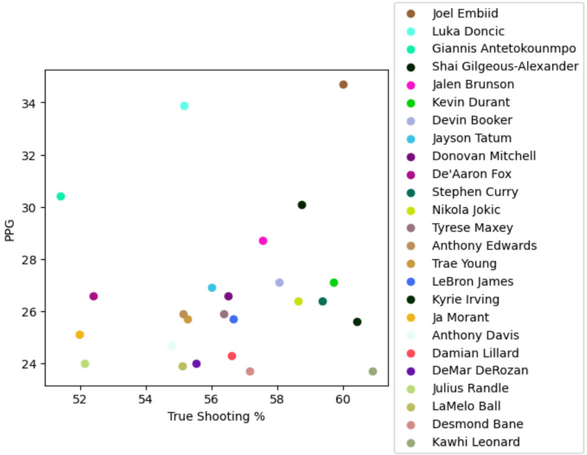


Figure 2: True Shooting % vs PPG

First, I projected the common metric of True Shooting of the 25 top scorers and took note of inconsistencies like Joel Embiid being highly efficient in Figure 2, but is somehow an overall low scorer like in Figure 1, most likely due to injuries.

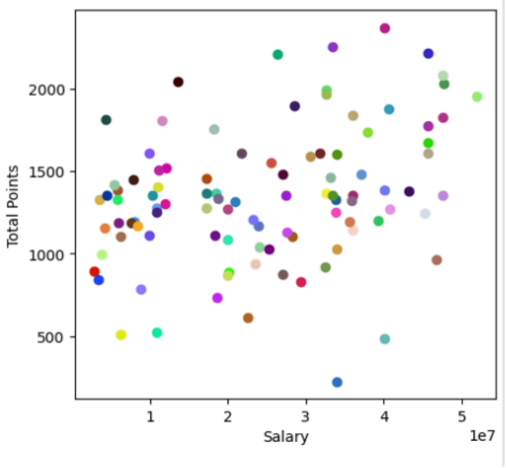


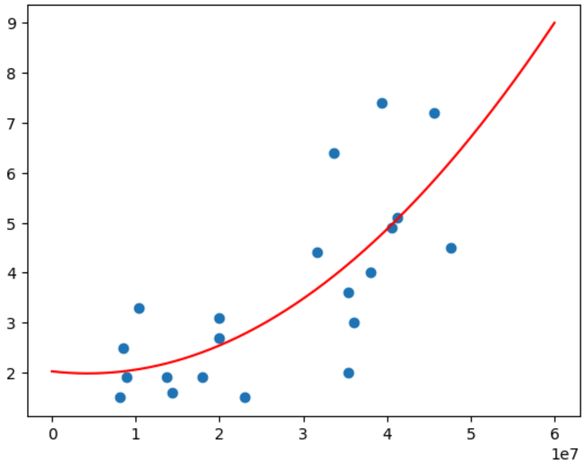
Figure 3: Salary vs Total Point (Larger Sample Size)

By exploring these visualizations such as the ones above, it became apparent that certain performance metrics were more strongly correlated with player salaries than others, which provided important insights into how teams value players based on their contributions on the court.

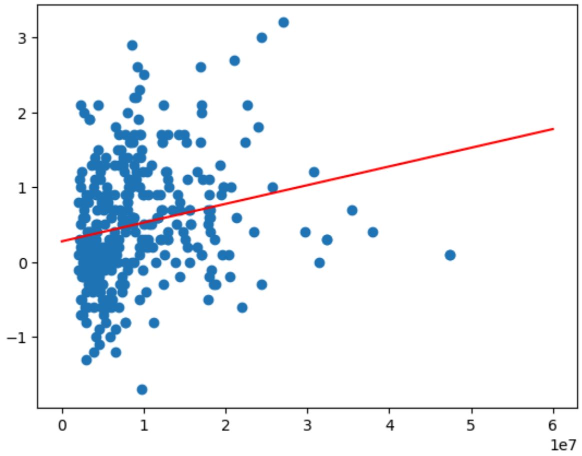
Model Development and Testing

Machine learning techniques were employed to model the relationship between player performance metrics and their salaries. The goal was to determine the most effective way to combine these factors into a single, reliable metric that could serve as a predictor of player value and salary. Several models were tested during this process. Linear regression models ($y = \beta_0 + \beta_1 x$) were initially applied to identify baseline relationships between performance metrics and salaries, and then polynomial models thereafter ($y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_{□} x^k$). This method helped to determine how individual metrics, such as points or rebounds, related to salary, providing a simple but useful understanding of how basic performance statistics might influence a player's value. The final comparison chosen was Value Over Replacement Player (VORP) on the y-axis against salary on the x-axis because VORP had the best fit with salary and was one of the few metrics that valued each player individually.

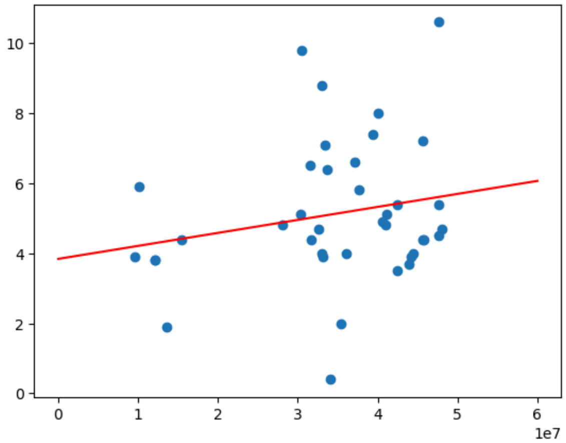
C1 - Superstar Big Men: 21



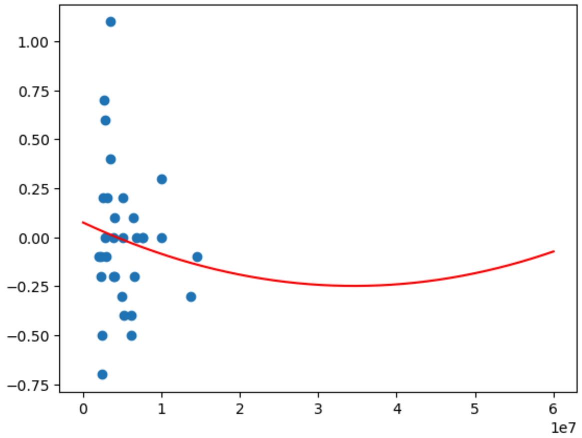
C2 - Role Player Guards: 348



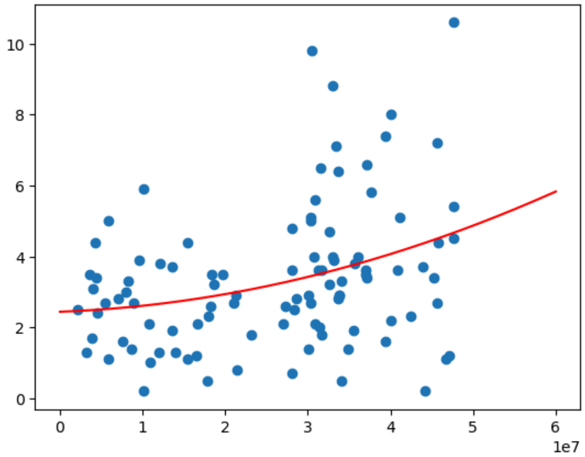
C3 - Superstars: 39



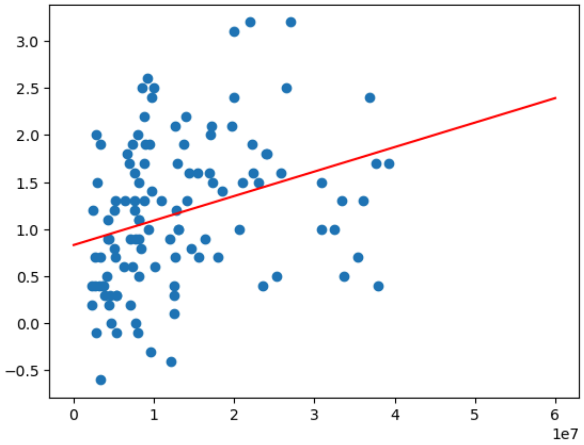
C4 - Benchwarmers: 32



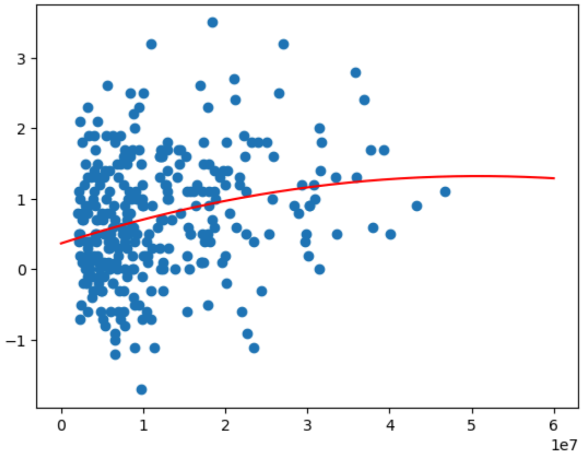
C5 - Star Guards: 101



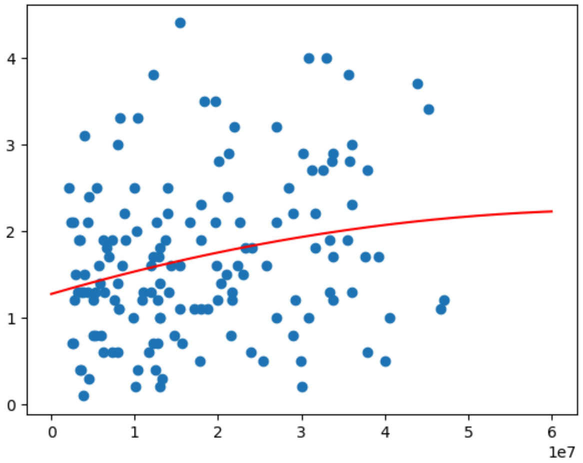
C6 - Role Player Big Men: 116



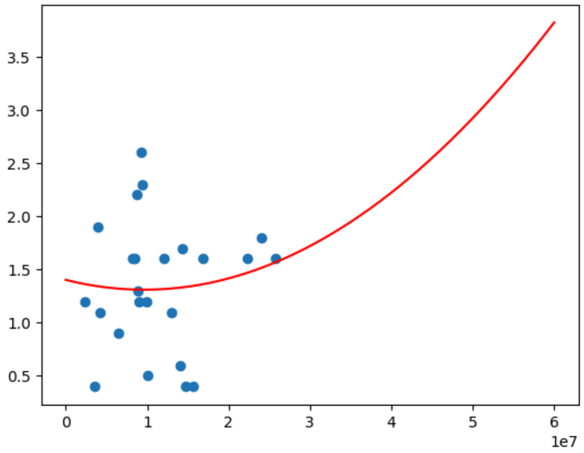
C7 - All-Around Role Players: 308



C8 - Stretch Bigs: 149



C9 - Defensive Specialists: 25



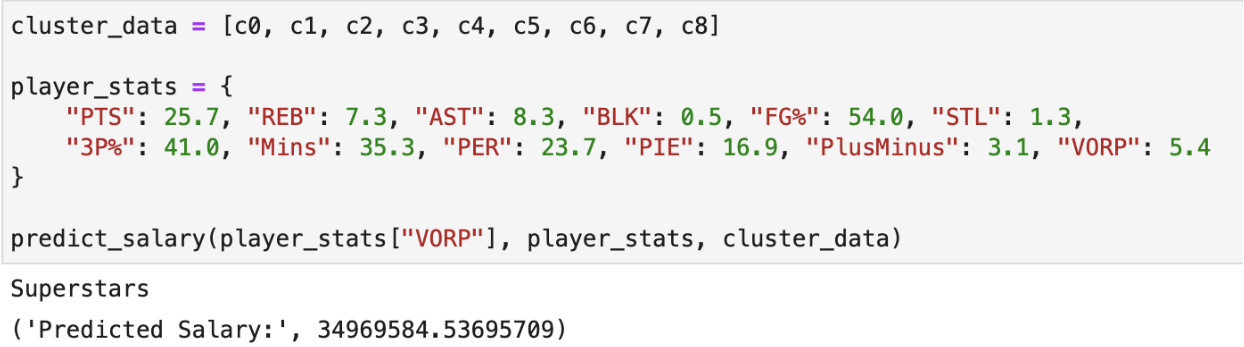
Final Metric Selection

After iterating on various models and approaches that compared salary to other advanced data metrics, it became clear that VORP was the most reliable and comprehensive metric for assessing player value. It was determined when all the total Mean Squared Errors (MSE) were added up for both PIE and VORP. VORP had a lower score than PIE, which meant a more accurate fit and that the fit curve had values closer to the real data. It had also become evident that PER and PIE would be a very important and crucial metric to use to determine clusters because it decreased the MSE for some of the clusters. VORP incorporates both offensive and defensive contributions, adjusts for playing time, and is strongly correlated with team success. This makes it a more holistic metric than others, such as PER and PIE, which focus on individual performance in isolation. By comparing VORP with other advanced metrics, it became evident that this metric was the most consistent in predicting player salaries across different teams and seasons. Below is the comparison of VORP to PIE, which had the second closest total MSE to VORP, to put into perspective just how well VORP performed with a difference of 23.8462.

	Ideal M Value	MSE (VORP)	Ideal M Value	MSE (PIE)
Cluster 0	2	1.4292	2	3.2323
Cluster 1	1	0.6130	1	4.2481
Cluster 2	1	3.8302	2	4.9193
Cluster 3	2	0.1271	3	3.2011
Cluster 4	2	3.5982	2	8.1986
Cluster 5	1	0.5570	1	3.0730
Cluster 6	2	0.6934	2	2.6760
Cluster 7	2	0.7990	2	2.1526
Cluster 8	2	0.3506	2	4.1429
Totals	N/A	11.9977	N/A	35.8439

Final Testing

One of the key aspects of this analysis was the use of predictive modeling based on actual performance statistics. The study aimed to predict an expected salary that aligns with a player's statistical impact by mapping player achievements to corresponding VORP values. Using the curve and statistics given for a certain player, they were matched to the correct cluster (curve). Once that had been done, the point where the VORP intersected the curve was found and the salary was output. The predictions were then compared to actual wages, evaluating how accurately the model captured real-world salary distributions.



```
cluster_data = [c0, c1, c2, c3, c4, c5, c6, c7, c8]

player_stats = {
    "PTS": 25.7, "REB": 7.3, "AST": 8.3, "BLK": 0.5, "FG%": 54.0, "STL": 1.3,
    "3P%": 41.0, "Mins": 35.3, "PER": 23.7, "PIE": 16.9, "PlusMinus": 3.1, "VORP": 5.4
}

predict_salary(player_stats["VORP"], player_stats, cluster_data)
```

```
Superstars
('Predicted Salary:', 34969584.53695709)
```

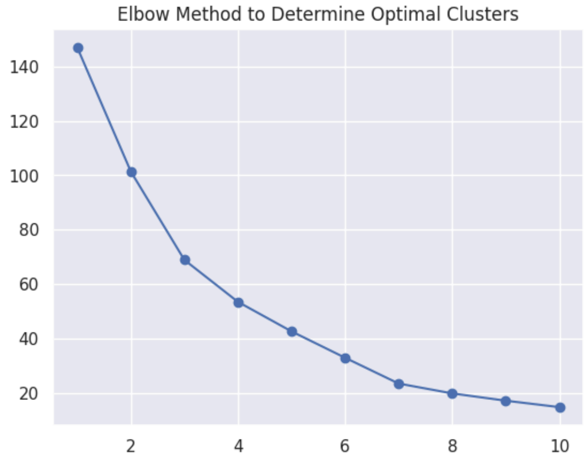
Figure 1: LeBron James 2023 Stats Plugged Into the VSPI With a Predicted Salary

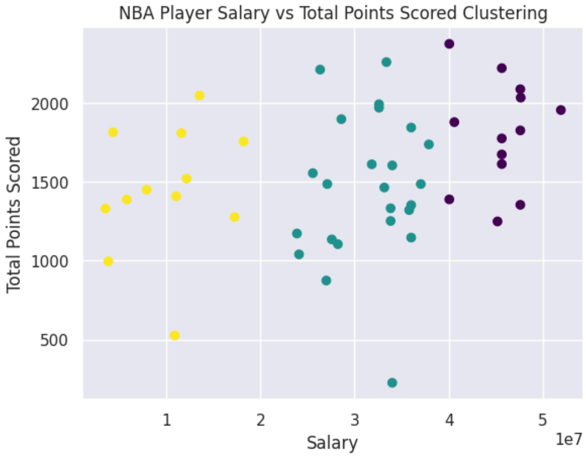
The findings showed that the model performed well for players within mid-tier and common archetype clusters. These players' predicted salaries closely matched their actual earnings, demonstrating that the methodology effectively captured the financial worth of standard contributors. However, the model exhibited larger errors when evaluating players on the extreme ends of the spectrum—players such as superstar LeBron James, whose numbers from last year are listed above. This suggests that while the model works well for most players, it struggles to accurately capture outliers (external factors such as marketability, contract structures, and team priorities play a larger role), like with LeBron who is one of the most valuable players in the NBA (generally considered the most marketable player in the sport), easily getting paid more than about 35 million dollars in annual earnings.

Analysis

The results of this study reveal significant insights into the relationship between player performance, value, and salary. The clustering process identified nine distinct player archetypes, each representing a unique style of play and contribution to their respective teams. These clusters were derived from a combination of traditional and advanced metrics, ensuring that both box-score statistics and deeper analytical measures were considered. Some examples of players in their clusters were Giannis Antetokounmpo in “Superstar Big Men”, Luka Doncic in “Superstar”, and Alex Caruso in “Defensive Specialist”.

During the research process, several challenges were encountered. One major issue was overfitting in machine learning models. Some of the models, particularly base CNN models, that used TensorFlow to split into train and test data, performed exceptionally well on training data but struggled to generalize when tested on new data. To address this, techniques such as splitting data into clusters to generalize effectively were implemented. Clusters were first brought to attention when they were experimented with early on with small test sizes. First, an Elbow method was developed to see where the biggest bend in the data was and became my cluster total. Then, it was applied to the data to visualize what they looked like, but clusters based on salary were not used because they didn't fit in as well with the regression models as the clusters that sectioned data based on statistical ranges.





The inconsistencies observed at the high and low ends highlight an important limitation of the study. Superstars often receive salaries surpassing their statistical contributions due to factors such as fan engagement, branding, and leadership impact. Conversely, lower-tier players may receive wages influenced by minimum contract agreements or roster construction needs rather than pure statistical output. These findings suggest that while statistical analysis is crucial in player valuation, additional factors beyond performance metrics must be considered to achieve a comprehensive evaluation.

Exception: VORP should increase as salary increases and should not fall to the Salary axis ever

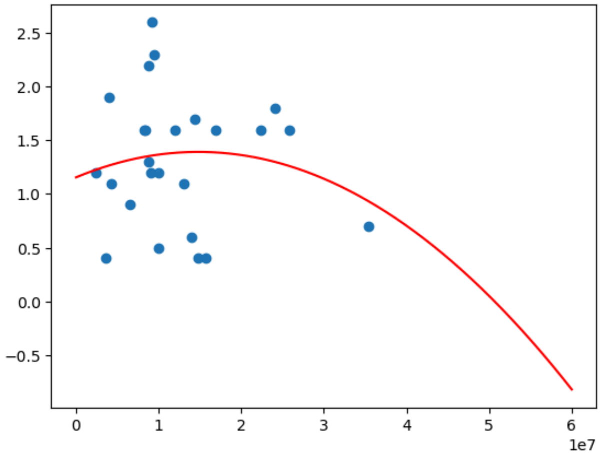


Figure 1: Cluster Before Exception Removed

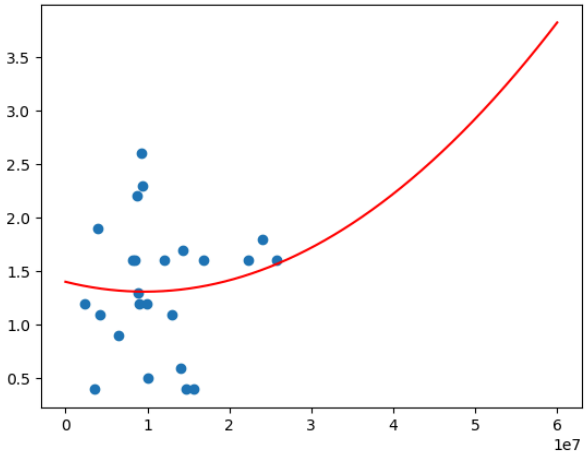


Figure 2: Cluster After Exception Removed

Overall, this analysis confirms that the developed metric provides a solid framework for estimating player value based on performance. Using clustering and regression models allowed for a more structured approach to salary prediction, offering NBA teams a data-driven method for evaluating contract decisions.

Conclusion

With more time and resources, this project could be refined by improving how the model handles outliers by using a more advanced model (An LLM, for example), especially in cases where factors outside of raw performance influence player salaries. Future refinements could involve integrating additional non-statistical variables such as endorsement value and leadership qualities to improve predictive accuracy for high-end and low-end players. I could develop new adjustments or incorporate additional variables to capture these factors better, ultimately improving the accuracy of my curves and predictions.

Even with its current limitations, the findings from this project are valuable. The results prove that it is possible to build a model capable of predicting NBA player value based on performance metrics. While extreme data points may not always align perfectly, this research lays the foundation for future improvements and more sophisticated modeling techniques. Expanding on this idea, similar models could be applied beyond the NBA to the other major sports leagues in the U.S., such as the NFL, MLB, and NHL. With further refinement, the approach could be shifted globally to analyze contracts and player value in soccer, cricket, and other major international sports industries.

This project is a stepping stone towards a more data-driven approach to evaluating athletes. With continued work, better models, and larger datasets, this kind of analysis could help teams make smarter financial decisions and even change how player contracts are structured in the future.

Works Cited

Blant, Jesse. "NBA Salary Regression Modeling." *Medium*, medium.com/@blant.jesse/nba-salary-regression-modeling-4846e53a1d3b.

"Estimated Plus-Minus (EPM)." *Dunks & Threes*, dunksandthrees.com/epm. Chart.

Medvedovsky, Kostya. "Daily Adjusted and Regressed Kalman Optimized Projections | DARKO." *Ap Analytics*, apanalytics.shinyapps.io/DARKO/.

"NBA Advanced Stats / NBA Traditional Stats." *NBA*. Table.

"NBA Player Contracts." *Basketball Reference*. Table.

Smith, Aaron. "Clustering NBA Players Based on Statistics — an Intro into Unsupervised Machine Learning." *Medium*, 27 May 2020, medium.com/@aaroncolesmith/clustering-nba-players-based-on-statistics-an-intro-into-unsupervised-machine-learning-597ba8ea795a.