Pay vs. Performance: A Data Driven Analysis of the National Basketball Association's Salary Efficiency

Justin Hatch
B.S Computer Science, University of Oregon
jhatch3@uoregon.edu

ABSTRACT - For the 2025-26 season, players in the National Basketball Association (NBA) are projected to earn over $5 billion in total salary, averaging approximately $12 million per player [1]. Despite this scale of investment, questions remain about how efficiently teams allocate salary relative to on court performance. This study evaluates salary efficiency by modeling the relationship between player compensation and advanced performance metrics, including Offensive and Defensive Box Plus/Minus (OBPM/DBPM), Value Over Replacement Player (VORP), position, and minutes played. Using historical data from the 2020-21 through 2024-25 seasons, an XGBoost regression model is trained to estimate expected salary based on production. Residuals from this model are then used to identify players who appear overvalued or undervalued relative to performance based expectations. The results reveal substantial salary inefficiencies across the league, demonstrating how data driven modeling and residual analysis can support more informed roster construction and salary management decisions in a cap constrained environment.

I. INTRODUCTION

Roster construction in the National Basketball Association (NBA) is fundamentally constrained by the salary cap, requiring teams to balance star talent, roster depth, and financial flexibility. As player salaries continue to escalate, even modest inefficiencies in contract valuation can produce outsized competitive consequences. A single mispriced contract may restrict a team's ability to retain key contributors, pursue free agents, or adapt strategically over multiple seasons. Despite the growing availability of detailed player performance data, accurately translating on-court impact into financial value remains a persistent challenge for NBA front offices.

The growing role of analytics in professional basketball has reshaped how teams evaluate on court performance, yet player compensation remains influenced by factors such as market demand, positional scarcity, injury history, and narrative driven perceptions of value. These forces can introduce distortions between a player's true on court contribution and their contractual cost. As a result, the NBA labor market provides a compelling setting for examining whether data driven evaluation methods can uncover systematic inefficiencies in how teams allocate salary.

This study frames NBA player contracts as economic assets and evaluates them through a quantitative lens. Rather than asking whether a player is "good" or "bad," the analysis centers on whether a player's performance justifies their cost relative to peers. By comparing observed salaries to data driven estimates of expected compensation, the approach shifts the focus from raw performance to value efficiency, a perspective that is critical in cap limited environments.

Importantly, the objective of this study is not to propose a definitive salary model or to replace existing front office evaluation processes. Instead, it demonstrates how machine learning based valuation and residual analysis can function as diagnostic tools for identifying patterns of overpayment and underpayment across the league. In doing so, this paper highlights the practical role of interpretable data science methods in supporting informed decision making within professional sports organizations.

II. BACKGROUND

A. Residuals

In predictive modeling, a residual represents the difference between an observed value and the value predicted by a model. When applied to NBA salaries, residuals quantify the extent to which a player's compensation deviates from what would be expected based on measurable on court performance. Positive residuals indicate players earning more than model expectations, while negative residuals suggest potential undercompensation relative to production.

Residual analysis is widely used in economics and applied data science to identify inefficiencies and mispricing in constrained markets. Rather than evaluating players by performance or salary in isolation, residuals provide a comparative measure of value efficiency by isolating the portion of compensation not explained by observable productivity. This approach is particularly well suited to the NBA, where player contracts are influenced not only by performance but also by salary cap rules, free agency timing, positional scarcity, injury risk, and market driven perceptions of value.

By estimating expected salary as a function of advanced performance metrics such as Box Plus/Minus, Value Over Replacement Player, positional indicators, and minutes played, residuals capture the gap between production based value and actual compensation. Aggregating residuals across players and teams enables the identification of systematic patterns in salary allocation, highlighting potential overvaluation and undervaluation within a salary cap constrained environment. As such, residual analysis provides a transparent and interpretable framework for evaluating contract efficiency in professional basketball.

B. XGBoost

To estimate expected player salary, this study employs Extreme Gradient Boosting (XGBoost), a tree based ensemble learning method designed to capture complex, nonlinear relationships between predictors and outcomes. Unlike linear regression models, XGBoost constructs an ensemble of decision trees sequentially, with each tree trained to correct errors made by previous iterations. This iterative structure allows the model to capture interactions between performance metrics, positional effects, and playing time that may influence compensation in non productive ways.

XGBoost is well suited to salary modeling due to its ability to handle heterogeneous feature scales, robustness to multicollinearity, and strong performance on structured tabular data. These properties are particularly valuable in the NBA context, where advanced metrics such as Box Plus/Minus, and Value Over Replacement Player are often correlated and exhibit nonlinear relationships with salary. By optimizing a regularized objective function, XGBoost also mitigates overfitting, producing more stable and generalizable estimates of expected compensation.

The predicted salaries generated by the model serve as benchmarks for residual analysis. Comparing observed player salaries to XGBoost based expectations allows residuals to quantify deviations between market compensation and performance based valuation, forming the foundation for identifying salary inefficiencies across players and teams.

III. METHODS

A. Data Set

Player performance data were collected from Basketball-Reference advanced statistics tables [2], while salary data were obtained from ESPN's NBA salary listings [3]. The dataset spans five NBA seasons, from 2020-21 through 2024-25, providing a multi year sample capable of capturing stable relationships between on court production and compensation.

For each season, advanced player metrics were extracted from Basketball-Reference and stored as raw season level files. To maintain an

organized and reproducible data pipeline, performance data were initially saved as team level CSV files (one CSV per team per season) and subsequently aggregated across all seasons into a unified dataset.

Salary data were collected from ESPN's annual salary pages for the corresponding seasons. Salaries were cleaned and standardized by removing currency symbols and delimiters and converting values to numeric format. Player salary data were then merged with the advanced performance metrics using player name and season identifiers. The final merged dataset provides a comprehensive view of player compensation and advanced performance metrics across the league from 2020-21 through 2024-25, forming the basis for model training and residual-based evaluation of salary efficiency.

B. Features

The model incorporates a combination of performance, role, and contextual features to estimate an expected NBA player salary, reflecting how teams evaluate both on-court impact and usage. Minutes Played (MP) serve as a proxy for player availability, durability, and role size within a team's rotation; players trusted with higher minutes typically contribute more consistently and command greater compensation due to their sustained on court presence.

Offensive Box Plus/Minus (OBPM) and Defensive Box Plus/Minus (DBPM) are included to capture a player's estimated per possession impact on team performance at both ends of the floor [4]. These metrics estimate how many points per 100 possessions a player contributes relative to an average NBA player, adjusting for team context and pace. OBPM reflects scoring efficiency, playmaking, and offensive decision making, while DBPM captures defensive positioning, rebounding, and disruption effects. Together, they provide a more nuanced assessment of player value than traditional box score statistics alone.

Value Over Replacement Player (VORP) is incorporated as a summary metric that aggregates a player's total on court contribution relative to a hypothetical replacement level player [5]. By scaling Box Plus/Minus metrics by playing time, VORP accounts for both efficiency and volume, making it particularly useful for salary modeling where teams compensate not just peak performance, but sustained impact across a season. Including VORP allows the model to align more closely with front office valuation frameworks that emphasize marginal wins and opportunity cost.

C. Training

Model performance in XGBoost is highly sensitive to hyperparameters governing tree complexity, learning dynamics, and regularization. To avoid reliance on default settings and to systematically explore the hyperparameter space, this study employed a randomized hyperparameter search procedure with cross-validation. The search evaluated 700 unique hyperparameter configurations, each assessed using 4-fold cross-validation, resulting in a total of 2,800 model fits.

The hyperparameter search was parallelized across all available CPU cores (n_jobs = −1) to improve computational efficiency. Model performance was evaluated using the mean cross-validation score, and the configuration achieving the highest score was selected as the final model. This tuned model was subsequently used to generate salary predictions and conduct residual based salary efficiency analysis. A full table of evaluated hyperparameter ranges is provided in the cited work [8].

D. Performance

Model performance was evaluated using the coefficient of determination (R^2) and root mean squared error (RMSE). R^2 measures the proportion of variance in NBA player salaries explained by the model based on on-court performance metrics. While higher R^2 values indicate stronger alignment between performance and compensation, unexplained

variance is expected given that player contracts are also influenced by non performance factors such as market conditions, contract timing, and team specific strategy.

RMSE captures the average magnitude of prediction errors and is particularly sensitive to large deviations between predicted and observed salaries. Evaluating RMSE in log salary space allows errors to be interpreted proportionally across different salary levels, preventing high salary contracts from disproportionately dominating error metrics. Together, R^2 and RMSE provide a concise summary of predictive accuracy, while residual analysis offers complementary insight into how individual salaries deviate from performance based expectations.

E. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) was conducted using data from the 2024-25 NBA season to provide context for subsequent modeling and residual based evaluation. This analysis examined the distribution of player salaries across the league, including comparisons among top and bottom earners, positional salary distributions, and team level salary structures. Summary statistics were also computed to characterize the scale, spread, and skewness of NBA compensation.

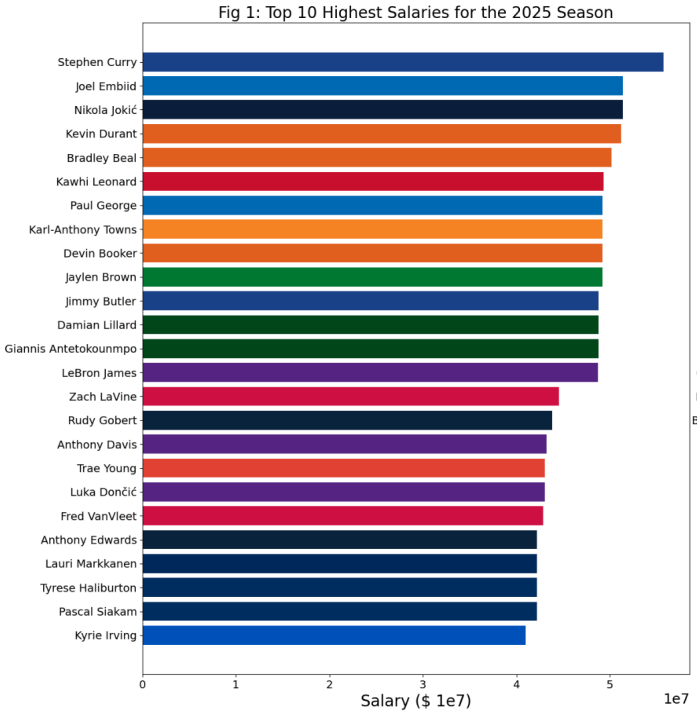


Fig 1: Top 10 Highest Salaries for the 2025 Season

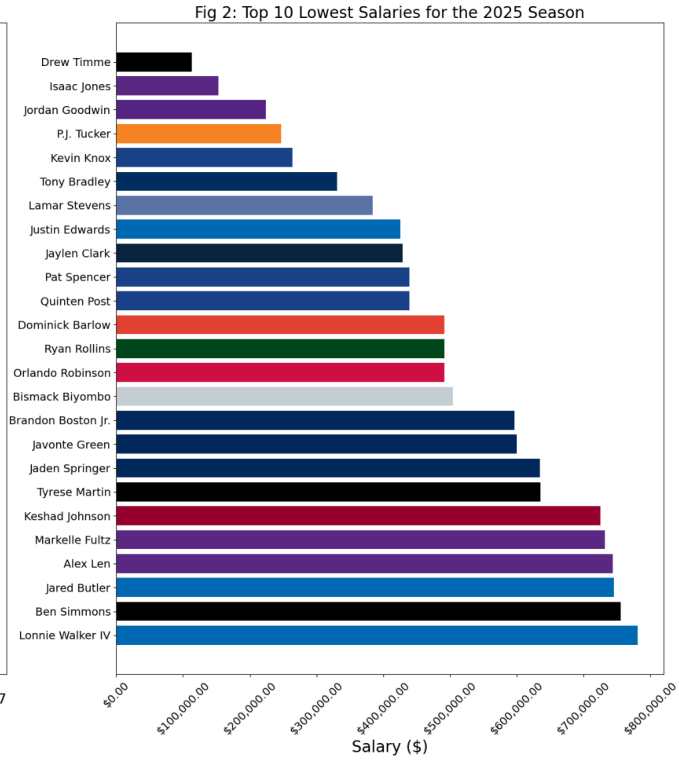
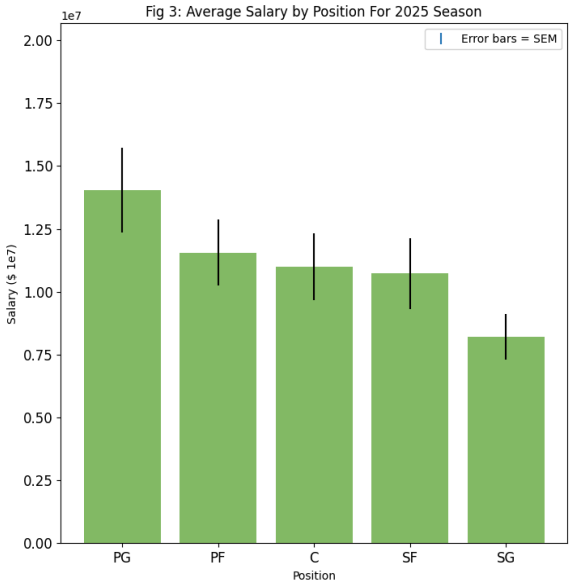


Fig 2: Top 10 Lowest Salaries for the 2025 Season



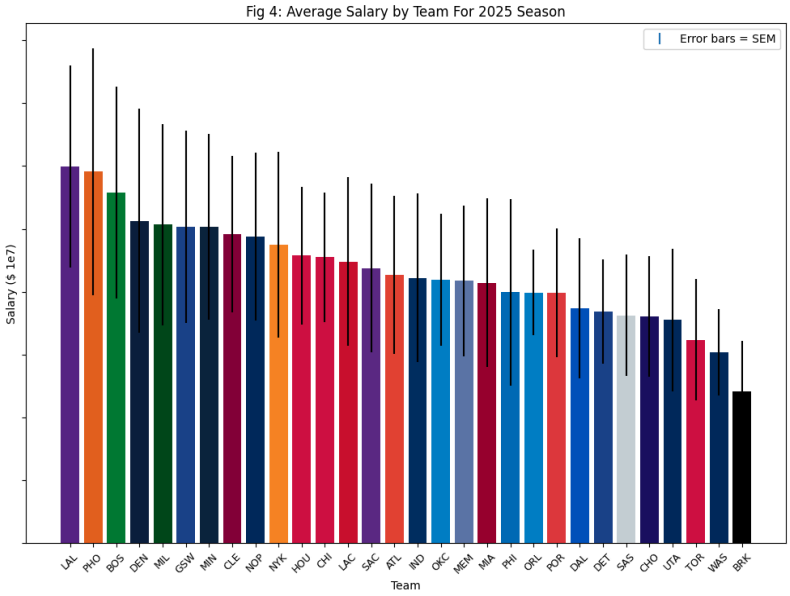
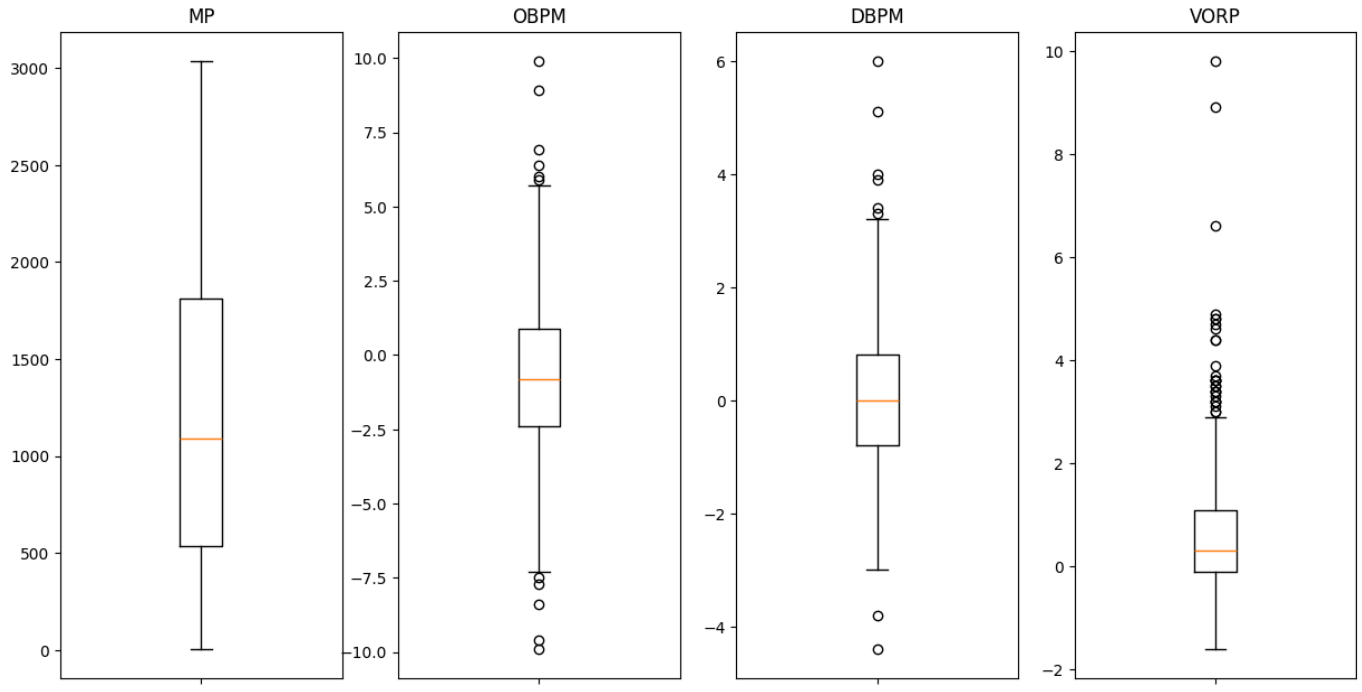


Figure 5: Distribution of Advanced NBA Data For The 2025 Season



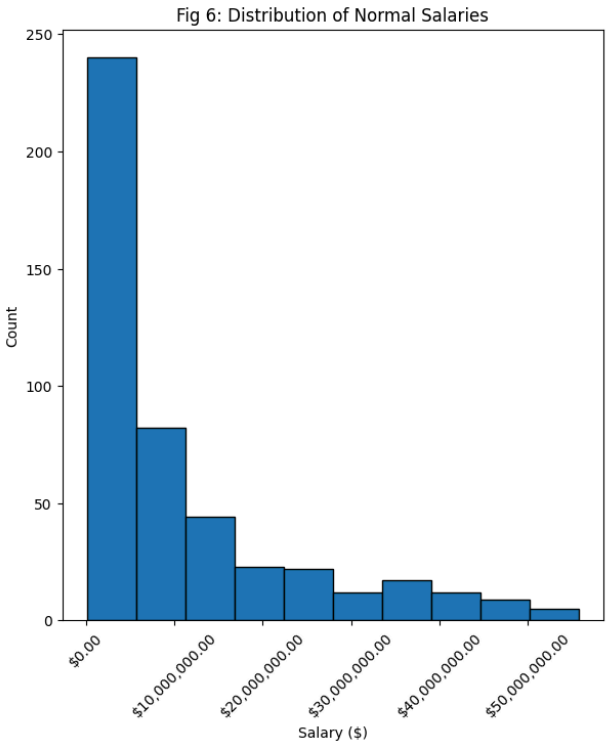


Fig 6: Distribution of Normal Salaries
Count
Salary ($)
$0.00
$10,000,000.00
$20,000,000.00
$30,000,000.00
$40,000,000.00
$50,000,000.00

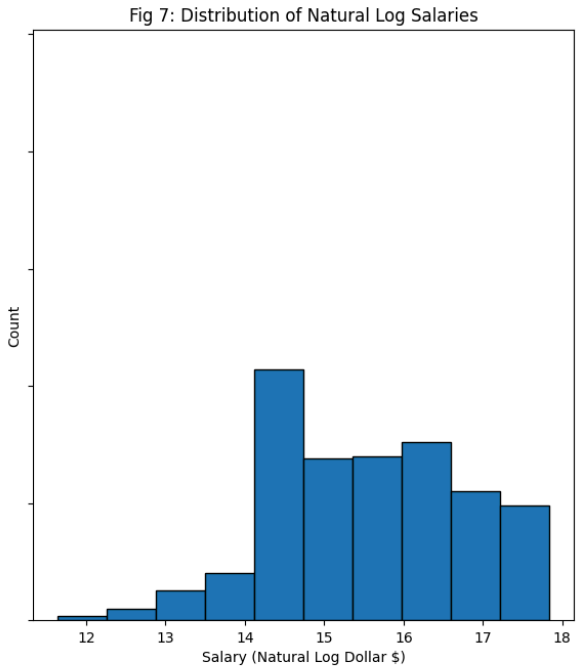


Fig 7: Distribution of Natural Log Salaries
Count
Salary (Natural Log Dollar $)

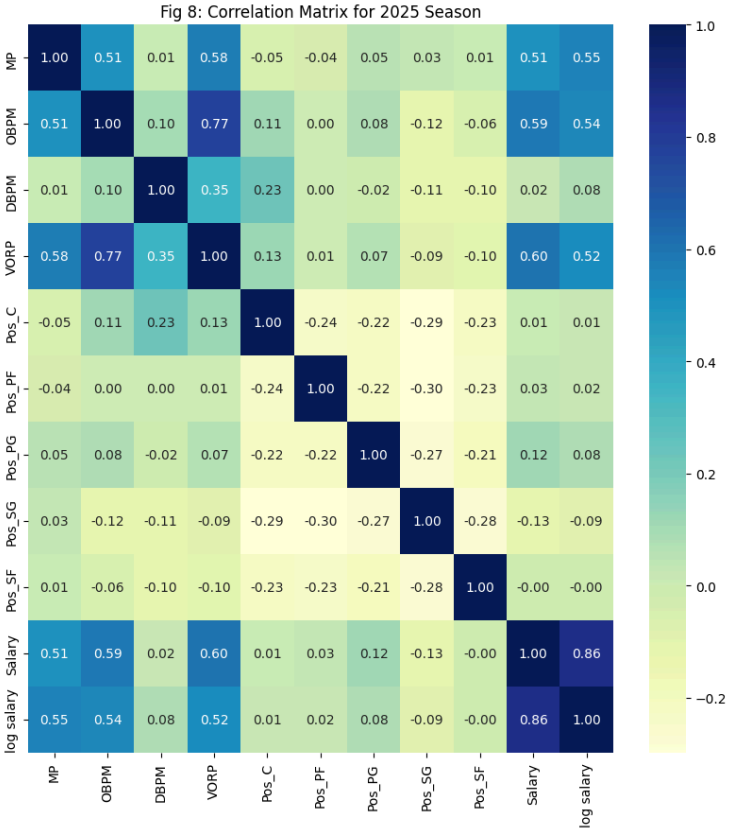


Fig 8: Correlation Matrix for 2025 Season
MP 1.00 0.51 0.01 0.58 -0.05 -0.04 0.05 0.03 0.01 0.51 0.55
OBPM 0.51 1.00 0.10 0.77 0.11 0.00 0.08 -0.12 -0.06 0.59 0.54
DBPM 0.01 0.10 1.00 0.35 0.23 0.00 -0.02 -0.11 -0.10 0.02 0.08
VORP 0.58 0.77 0.35 1.00 0.13 0.01 0.07 -0.09 -0.10 0.60 0.52
Pos_C -0.05 0.11 0.23 0.13 1.00 -0.24 -0.22 -0.29 -0.23 0.01 0.01
Pos_PF -0.04 0.00 0.00 0.01 -0.24 1.00 -0.22 -0.30 -0.23 0.03 0.02
Pos_PG 0.05 0.08 -0.02 0.07 -0.22 -0.22 1.00 -0.27 -0.21 0.12 0.08
Pos_SG 0.03 -0.12 -0.11 -0.09 -0.29 -0.30 -0.27 1.00 -0.28 -0.13 -0.09
Pos_SF 0.01 -0.06 -0.10 -0.10 -0.23 -0.23 -0.21 -0.28 1.00 -0.00 -0.00
Salary 0.51 0.59 0.02 0.60 0.01 0.03 0.12 -0.13 -0.00 1.00 0.86
log salary 0.55 0.54 0.08 0.52 0.01 0.02 0.08 -0.09 -0.00 0.86 1.00
MP
OBPM
DBPM
VORP
Pos_C
Pos_PF
Pos_PG
Pos_SG
Pos_SF
Salary
log salary
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

The EDA highlights substantial diversity in player salaries, both within and across positions and teams, reinforcing the need for a modeling approach that accounts for differences in role, playing time, and on-court impact. These descriptive findings motivate the use of a log transformed salary target and provide baseline intuition for interpreting model predictions and residual based measures of salary efficiency.

IV. EXPERIMENT

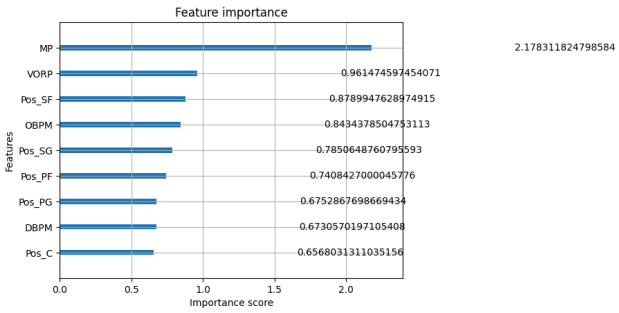
Model training was conducted using data from the 2020-21 through 2023-24 NBA seasons, while the 2024-25 season was held out for residual based evaluation and testing. This train-test split reflects a realistic forecasting scenario in which historical performance data are used to estimate expected salaries for a future season. Two baseline XGBoost regression models were trained using default hyperparameters to establish initial performance. The models differed only in their loss functions: **squared error** and **quantile error**. Squared error is the standard objective for regression and estimates the conditional mean of salary, but it is sensitive to extreme values. Because NBA salary data contain large outliers driven by max contracts, a quantile loss model was also evaluated, as it is more robust to outliers and less influenced by extreme salaries. Comparing these objectives provides a baseline assessment of how loss function choice affects salary predictions and residuals.

On the 2024-25 test set, the baseline models achieved RMSE values of 0.990 and 1.083 and R^2 scores of 0.325 and 0.192 for the quantile error and squared error objectives, respectively. These results indicate that advanced performance metrics explain a meaningful, though incomplete, portion of salary variation, consistent with the presence of non-performance factors in contract determination.

To improve predictive performance, hyperparameters were optimized using randomized search with cross-validation on the 2020-21 to 2023-24 season. The tuning process evaluated 2,800 model configurations across four folds. The best performing configuration achieved a mean cross-validation score of 0.3790. The selected hyperparameters balanced model flexibility and regularization, incorporating a moderate learning rate, shallow trees, and stronger regularization:

- ***Objective:*** *reg:quantileerror*
- ***n_estimators***: 1743
- ***max_depth***: 3
- ***learning_rate***: 0.0238
- ***subsample***: 0.95266
- ***colsample_bytree***: 0.9836
- ***min_child_weight***: 4

When evaluated on the 2024-25 test set, the tuned model achieved an RMSE of 0.954 and an R^2 of 0.372. This corresponds to a 3.6% reduction in prediction error relative to the quantile error baseline and an 11.9% reduction relative to the squared error baseline. The tuned model also exhibited notable gains in explained variance, reinforcing the value of hyperparameter optimization in capturing nonlinear relationships between on-court performance and player compensation.



Overall, the tuned XGBoost model provides a substantially stronger baseline for residual analysis. While a large share of salary variation remains unexplained, advanced performance metrics capture meaningful structure in NBA compensation. Feature importance analysis further supports this conclusion, with aggregate value measures such as VORP and minutes

played emerging as the most influential predictors, followed by offensive impact (OBPM), while positional indicators contribute comparatively less explanatory power. Additionally, residual diagnostics were conducted to assess the suitability of the model for residual based salary efficiency analysis. Visual inspection of residuals plotted against predicted salaries revealed no strong patterns, indicating that the model captures the dominant nonlinear structure in the relationship between performance metrics and compensation. The distribution of residuals was centered near zero and approximately symmetric, supporting their interpretation as relative measures of overvaluation and undervaluation rather than artifacts of model misspecification. While residuals exhibit heavier tails, this behavior is expected in salary data and does not undermine the use of residuals as a diagnostic tool in a predictive, non inferential setting.

Fig 9: Residuals vs Predicted

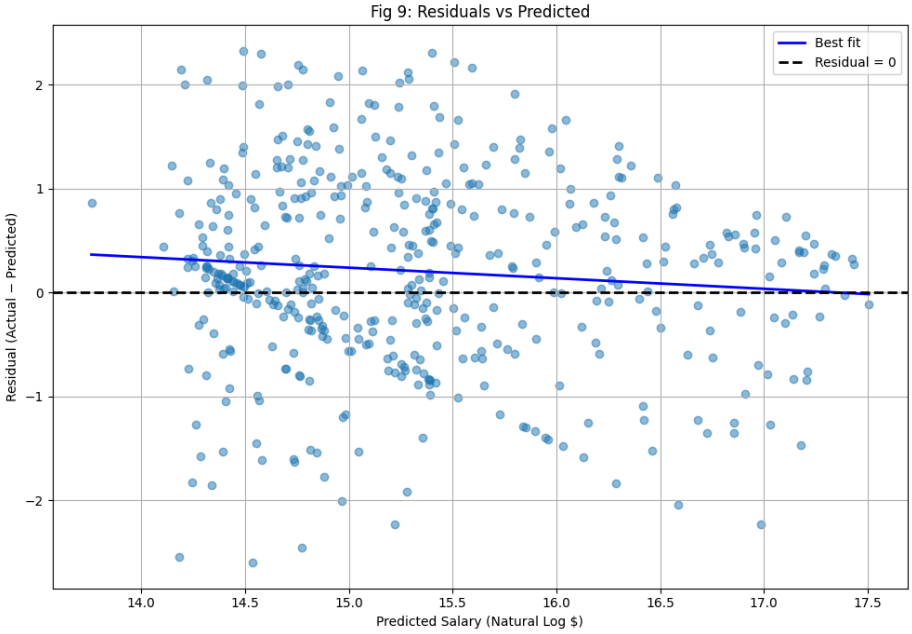


Fig 10: Actual vs Predicted

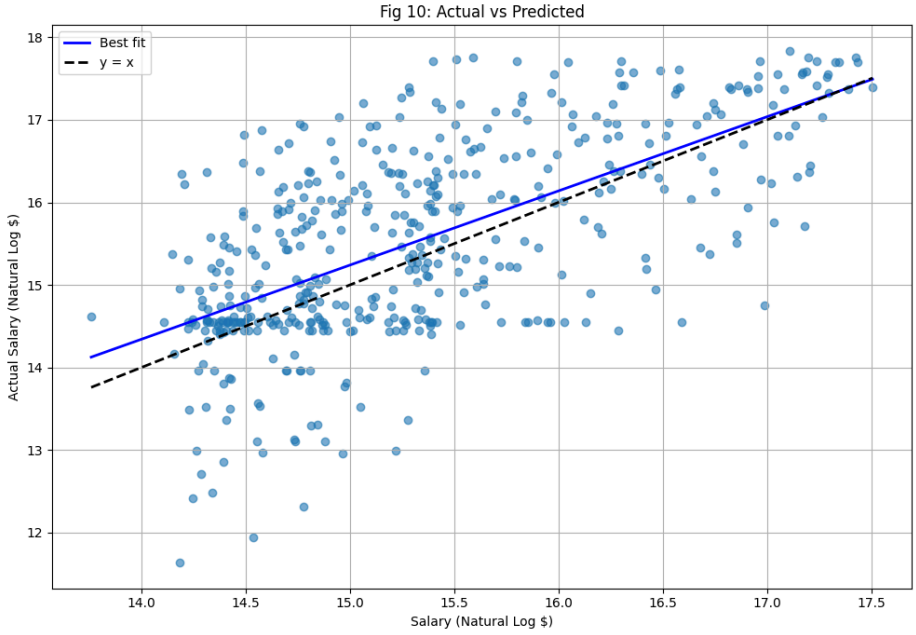


Fig 11: Distribution of Residuals

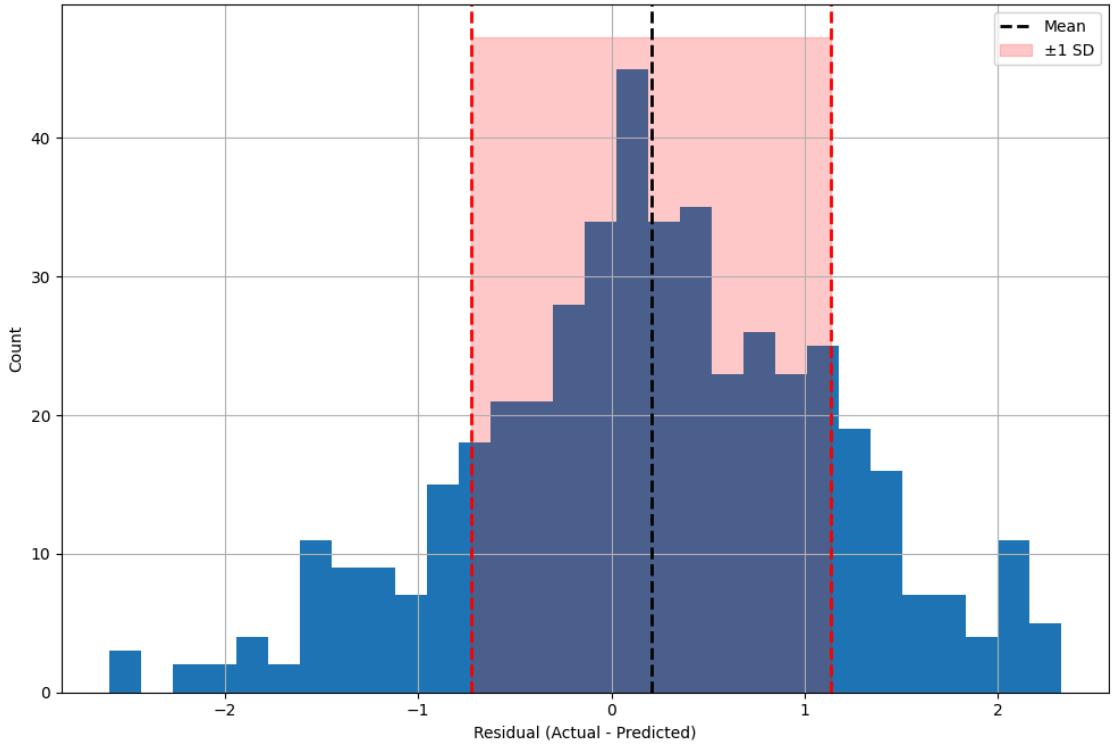
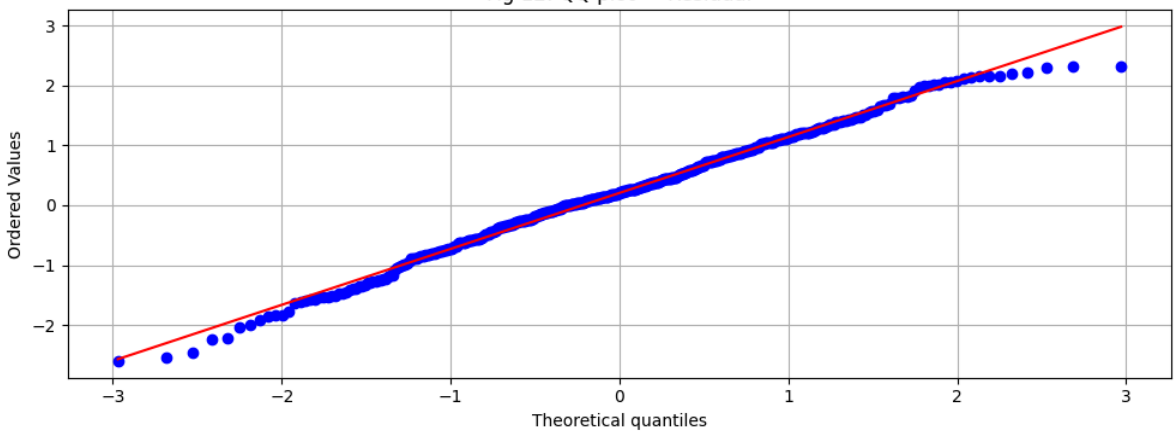


Fig 12: QQ plot — Residual



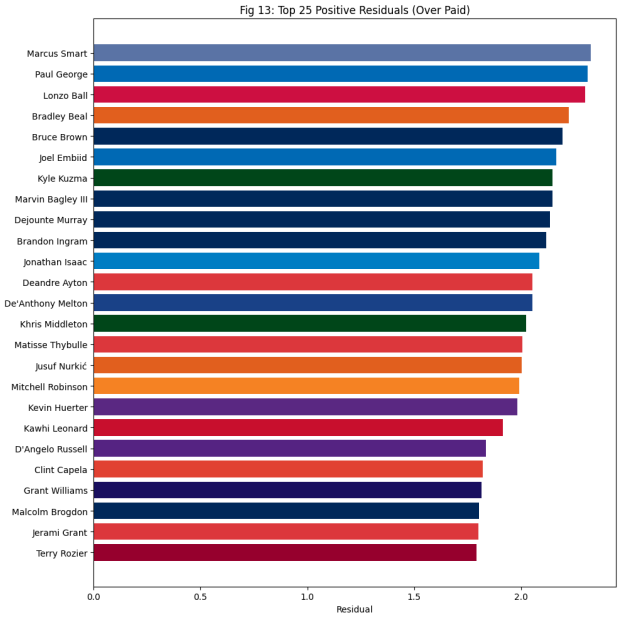


Fig 13: Top 25 Positive Residuals (Over Paid)
Marcus Smart
Paul George
Lonzo Ball
Bradley Beal
Bruce Brown
Joel Embiid
Kyle Kuzma
Marvin Bagley III
Dejounte Murray
Brandon Ingram
Jonathan Isaac
Deandre Ayton
De'Anthony Melton
Khris Middleton
Matisse Thybulle
Jusuf Nurkić
Mitchell Robinson
Kevin Huerter
Kawhi Leonard
D'Angelo Russell
Clint Capela
Grant Williams
Malcolm Brogdon
Jerami Grant
Terry Rozier
0.0
0.5
1.0
1.5
2.0
Residual

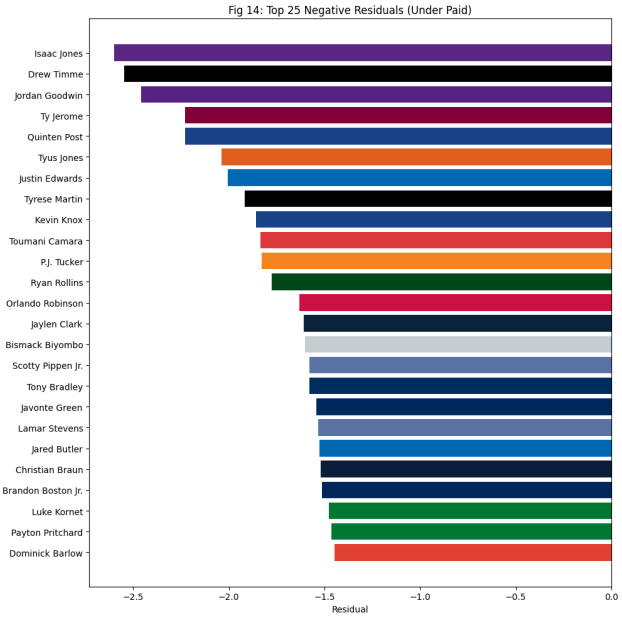


Fig 14: Top 25 Negative Residuals (Under Paid)
Isaac Jones
Drew Timme
Jordan Goodwin
Ty Jerome
Quinten Post
Tyus Jones
Justin Edwards
Tyrese Martin
Kevin Knox
Toumani Camara
P.J. Tucker
Ryan Rollins
Orlando Robinson
Jaylen Clark
Bismack Biyombo
Scotty Pippen Jr.
Tony Bradley
Javonte Green
Lamar Stevens
Jared Butler
Christian Braun
Brandon Boston Jr.
Luke Kornet
Payton Pritchard
Dominick Barlow
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
Residual

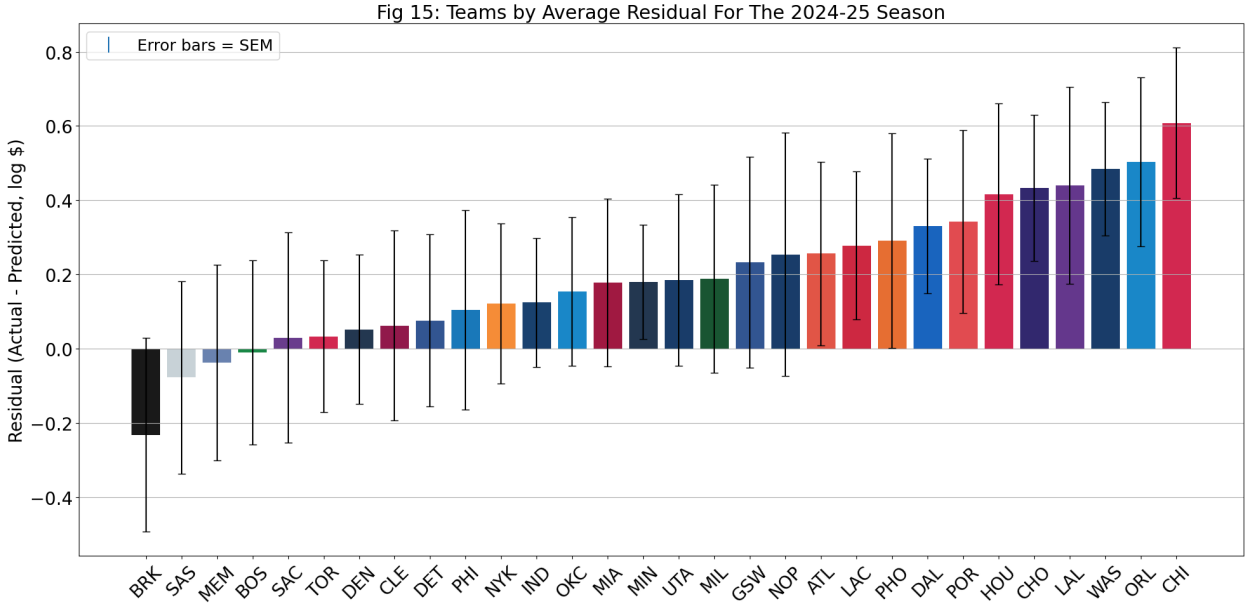


Fig 15: Teams by Average Residual For The 2024-25 Season
Error bars = SEM
Residual (Actual - Predicted, log $)
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
BRK
SAS
MEM
BOS
SAC
TOR
DEN
CLE
DET
PHI
NYK
IND
OKC
MIA
MIN
UTA
MIL
GSW
NOP
ATL
LAC
PHO
DAL
POR
HOU
CHO
LAL
WAS
ORL
CHI

V. RESULTS

As stated above, residuals are calculated as the difference between observed and predicted salaries, with positive values indicating players earning more than model expectations and negative values indicating potential undervaluation. These residuals form the basis of the previous 3 visuals [13, 14, 15], which highlight salary inefficiencies across players and teams.

Several high salary players appear as overpaid relative to performance based expectations in the 2024-25 season. **Lonzo Ball** emerges as one of the largest positive residuals, driven primarily by limited on court availability. Despite carrying a cap hit of approximately $21.4 million in 2024-25 [3], Ball returned after more than two years of injury-related absence and played a restricted role upon return, with additional time missed due to a wrist injury [4]. Because the model heavily weights minutes played and advanced impact metrics, Ball's limited availability substantially lowers expected salary estimates, resulting in a large positive residual. This outcome reflects injury and availability risk rather than a judgment of player

Deandre Ayton similarly appears as overpaid relative to model expectations. Ayton earned approximately $34 million during the 2024-25 season [3], placing him among the league's highest paid centers. While his per game production remained solid, averaging double digit rebounds and efficient scoring according to ESPN [5], his season was again affected by injury, limiting total games played. In addition, the model values aggregate impact measures such as VORP and minutes played more heavily than traditional box-score counting stats. As a result, Ayton's high, relatively inflexible contract combined with reduced availability produces a significant positive residual, highlighting the disconnect between market compensation and production based valuation.

In contrast, several low salary players appear strongly underpaid relative to their on court contributions. **Isaac Jones**, who played on a near minimum contract worth roughly $152,957 [3], registers the most negative residuals in the dataset. Even modest NBA level production can generate large negative residuals when paired with minimal compensation, as the model assigns nontrivial expected salary value to any meaningful on court contribution. Jones' case illustrates how minimum and two way contracts can produce extreme undervaluation signals, though such results should be interpreted cautiously given limited sample sizes.

Drew Timme also appears among the most underpaid players in the 2024-25 season. Timme signed a rest-of-season contract with Brooklyn valued at approximately $113,000 following strong performances in the G League [3], [6]. Given this extremely low cap hit, any player capable of delivering NBA rotation level minutes will necessarily produce a large negative residual. Although Timme appeared in only nine games during the 2024-25 season, he started two contests and averaged 28.2 minutes per game, indicating promising on-court development relative to his compensation. His inclusion therefore highlights the model's sensitivity to structural salary floors imposed by contract rules rather than purely informational inefficiencies. Importantly, this apparent inefficiency was short-lived: Timme's salary increased by approximately **1,630%** entering the 2025-26 season, rising to $1,955,377 [3]. This sharp correction suggests that the model successfully identified true salary efficiency in real time, which was subsequently reflected in market valuation during the following contract cycle.

Among more established rotation players, **Payton Pritchard** represents a particularly clean example of undervaluation. Pritchard earned approximately $6.7 million in 2024-25 [3] while delivering efficient production and positive advanced metrics on meaningful minutes for Boston [7]. Additionally he won 6th man of the year, an award given to the best performing player coming off the bench. Unlike minimum salary cases, Pritchard's residual reflects both sustained playing time and efficiency, making

his underpayment signal more robust and less sensitive to small sample concerns.

Aggregating player residuals at the team level reveals systematic differences in how organizations convert payroll into on court production during the 2024-25 season. Teams with the highest average residuals indicating relative overpayment include the **Chicago Bulls**, **Orlando Magic**, and **Washington Wizards**. Chicago's elevated average residual is driven in part by large contracts tied to limited availability, most notably Lonzo Ball's $21.4 million cap hit despite restricted playing time. Orlando's roster includes several high salary players whose compensation exceeds what the model predicts based on advanced performance metrics, reflecting the influence of role specialization, injury history, and market driven contract decisions. Washington similarly skews toward overpayment, with large veteran contracts, such as Jordan Poole's nearly $30 million salary [3], reflecting roster construction choices that prioritize flexibility, trade considerations, or rebuilding strategy over immediate production efficiency.

At the opposite end of the spectrum, teams with the most negative average residuals, indicating relative undervaluation, include the **Brooklyn Nets**, **San Antonio Spurs**, and **Memphis Grizzlies**. Brooklyn's negative average is driven by significant rotation minutes allocated to players on minimum or near minimum contracts, including Drew Timme, which lowers aggregate payroll relative to modeled performance expectations. San Antonio's roster structure is dominated by rookie scale contracts, most notably Victor Wembanyama's $12.7 million salary in 2024-25 [3], which generates substantial surplus value relative to impact. Memphis similarly benefits from meaningful contributions by lower cost players that offset the presence of high salary stars, demonstrating how team level residuals reflect the net balance of bargain contracts and expensive commitments rather than the value of any single player.

VI. CONCLUSION

This study evaluated salary efficiency in the National Basketball Association by modeling the relationship between player compensation and on court performance using advanced metrics and an XGBoost regression framework. Residual analysis applied to the 2024-25 season provides a performance based benchmark for identifying potential overvaluation and undervaluation under the constraints of the NBA salary cap.

The results reveal meaningful variation in salary efficiency at both the player and team levels. High salary contracts associated with limited availability, role changes, or declining production frequently appear overvalued, while younger players and rotation contributors on rookie scale or team friendly contracts often generate surplus value. At the organizational level, aggregated residuals highlight systematic differences in how teams convert payroll into on-court impact.

While NBA salaries reflect factors beyond measurable on court performance, the alignment between residual based signals and known labor market mechanisms supports the interpretability of this approach. Future research could extend this framework by incorporating additional data sources, such as player versatility measures, defensive tracking data, leadership proxies, or off court signals including social media presence, brand value, and marketability, to better capture components of player value that are currently priced by the market but not reflected in traditional performance metrics. Overall, residual based valuation offers a transparent and practical tool for diagnosing salary inefficiencies and informing data driven roster construction.

VII. Work Cited

[1] jhatch3. “Jhatch3/NBA-Salary-Minning.” *GitHub*, github.com/jhatch3/NBA-Salary-Minning. Accessed 03 Feb. 2026.

[2] “Basketball Statistics & History of Every Team & NBA and WNBA Players.” *Basketball*, www.basketball-reference.com/. Accessed

[3] “NBA Player Salaries .” *ESPN*, ESPN Internet Ventures, www.espn.com/nba/salaries. Accessed 03 Feb. 2026.

[4] Collier, Jamal, and Multiple Authors. “MRI Shows Bulls’ Lonzo Ball Has Grade 1-2 Wrist Sprain.” *ESPN*, ESPN Internet Ventures, www.espn.com/nba/story/_/id/42093562/bulls-lonzo-ball-least-10-days-sprained-wrist. Accessed 03 Feb. 2026.

[5] “Deandre Ayton 2024-25 Stats per Game - NBA.” *ESPN*, ESPN Internet Ventures, www.espn.com/nba/player/gamelog/_/id/4278129/type/nba/year/2025. Accessed 03 Feb. 2026.

[6] “Drew Timme.” *NBA G League Stats*, stats.gleague.nba.com/player/1631166/. Accessed 03 Feb. 2026.

[7] “Pritchard 2024-2025 Season.” *StatMuse*, www.statmuse.com/nba/ask?q=pritchard%2B2024-2025%2Bseason. Accessed 03 Feb. 2026.

[8] jhatch3. “Jhatch3/NBA-Salary-Minning/params.” *GitHub*, github.com/jhatch3/NBA-Salary-Minning/params. Accessed 03 Feb. 2026.

[Fig 1 - 15] jhatch3. “Jhatch3/NBA-Salary-Minning/Figs.” *GitHub*, github.com/jhatch3/NBA-Salary-Minning/Figs. Accessed 03 Feb. 2026.