ML Group Project Proposal:

Predicting NBA Player Salaries: A Machine Learning Approach to Player Contract Valuation

Course: Introduction to Machine Learning in R (7,926,1.00)

Authors:

Hu Yliene: 25-602-848

Brigati Francesco: 19-750-694

Kündig Jeremy: 22-413-900

Oehler Flurin: 20-062-667

Picture: Gemini (2025)

Introduction

Analytics has become central to modern sports, as the movie Moneyball shows how data improves decision-making. The NBA is now one of the most data-rich leagues, offering decades of information on player performance and payroll. Combined with its salary cap and luxury tax, it provides a clear framework for studying how on-court output relates to financial valuation. From a machine learning perspective, we view the NBA as a supervised regression problem. Historical performance serves as a feature, while future salary is the target. Following the predictive mindset described by Pargent et al. (2023), our goal is to estimate out-of-sample salary predictions and test their generalizability. This project has two aims: first, to build models that predict future NBA salaries based on past performance, and second, to use these predictions to study team spending patterns and contract efficiency.

Motivation

We selected this project because of the league's rich data environment. The NBA publishes detailed salary and performance data, making it well-suited for supervised ML. Salary prediction is technically challenging and strategically relevant since contracts directly influence roster decisions under the cap system. This project allows us to apply course concepts in a real-world context.

Relevance

The NBA offers a valuable setting for studying salary prediction and team spending. Unlike most European sports leagues, it operates under a salary cap, a luxury tax, and standardized contract rules that strongly affect roster construction.

The NBA is a transparent labor market. Player performance is observable, and salary outcomes are public, which allows us to examine how performance is priced and whether inefficiencies emerge. Researchers such as Xu, Z. (2025), Papadaki (2020), and Berri and Krautmann (2024) highlight the existing interest in NBA analytics and machine learning.

From an ML perspective, the NBA is ideal for supervised prediction. Decades of structured data allow for resampling, benchmarking, and out-of-sample evaluation. Predicting future salaries is a straightforward, measurable task with real strategic implications.

Dataset

We combine four datasets covering player performance, compensation, and team spending from 1990 to 2022. We utilize CPI-adjusted values to normalize financial data across eras:

1. **Player Salaries (1990-2023):** Annual salary data used as the basis for our target variable: salary in season t+1.
2. **Season-Level Player Statistics (1950-2022):** Aggregated metrics like PER, Win Shares, TS%, and usage rate that capture long-term value.
3. **Salary Cap:** Annual league-wide salary cap limits (nominal and inflation-adjusted), used to normalize player salaries relative to the financial environment of each era.
4. **Team Records:** Regular-season win-loss records and winning percentages per franchise, enabling us to correlate financial spending with on-court success.

Linking the datasets

We merge the datasets using Player name, season, and team as common keys. We implemented a standardization protocol to map historical team names to their modern equivalents, ensuring consistency. This allows us to construct a unified player-season dataset that contains:

- Past performance indicators (season-level player stats).
- Contextual features (aggregated team payroll, team success, % of Salary Cap).
- Financial outcomes (salary in the following season).

Data Processing and Feature Engineering

Script 1 establishes a consistent dataset by merging sources and resolving inconsistencies. The main steps include:

1. **Filter period:** 1990-2022 (as all datasets cover this time span)
2. **Handle Nas:** Replace NAs with 0
3. **Standardization of Team History:** We map all historical franchise names to their modern equivalents to ensure longitudinal consistency across relocations and rebranding.
4. **Resolution of Traded Players:** To handle multi-team seasons without data duplication, we aggregate a player's total games played (G) but assign their statistical profile to the final franchise. This ensures performance is weighted against the payroll of the paying team.
5. **Merge Salary and player data:** Merge salary data based on player name
6. **Compute Effective Payroll & Share**: Account for traded players who do not take the full salary. Working with the share of the payroll a player uses we create a metric that is more inflation-robust
7. **Normalize Financials (Salary Cap)**: Merge historical salary cap data
8. **Integrate Team Success**: Include team records (Wins, losses, win percentage, and binary playoff indicator)

ML Analytical Approach

1. Key Modeling Assumptions

No future data leakage

All features reflect information available at the end of season *t*. Future values are removed, except for the contract-renewal indicator, which teams themselves know during negotiations.

Contract renewal identified by salary jumps

Because the dataset lacks official contract metadata, renewals are approximated using large salary changes (+20% / –15%). These seasons are excluded from training to avoid learning negotiation-driven distortions.

Player trajectory captured with lagged features

For each performance metric, the model includes lags (t–1, t–2) and performance deltas. This allows the model to account for both current production and long-term trends (improvement or decline).

Salary-related variables excluded

To prevent leakage, past salaries, payroll shares, and cap-percent metrics are removed. Only the league-wide salary cap is kept as a financial context variable.

Team context preserved

Team performance indicators (wins, losses, win%, conference rank, playoffs) are included, as they influence player visibility and perceived market value.

Multi-team seasons consolidated

When a player changes teams mid-season, game totals and effective salary allocations are aggregated so each player-season remains a single, consistent observation.

Overall, these assumptions ensure that the final dataset closely reflects the information NBA teams realistically use when determining player compensation.

2. Modeling Approach

Four models were trained and compared:

- **OLS:** basic linear regression without regularization.
- **Elastic Net:** glmnet with $\alpha = 0.5$ and tuned λ, balancing LASSO and Ridge to handle multicollinearity.
- **Random Forest (untuned):** 300 trees with default settings.
- **Tuned Random Forest:** hyperparameters (mtry, min.node.size) optimized through nested cross-validation.

This model set enables a clear comparison between linear methods and nonlinear ensemble techniques that capture complex interactions.

Cross-Validation (Train Set Only)

A 5-fold CV on non-renewal seasons shows the performance hierarchy:

- The baseline mean-salary predictor performs poorly.

- OLS and Elastic Net achieve similar accuracy (≈63–66% variance explained).
- Random Forest performs substantially better, confirming that salary formation involves nonlinear patterns (e.g., scoring × efficiency, usage × team success).

Nested CV (RF Tuned)

A 3×2 nested CV provides unbiased performance estimates for the tuned RF:

- $R^2 \approx 0.714$
- RMSE ≈ 3.64M USD
  These metrics demonstrate strong predictive capacity considering the natural volatility of NBA salaries.

Holdout Test (True Out-of-Sample)

On a 75/25 split (excluding renewal years), results are:

	model	rsq	rmse	mse
	<chr>	<dbl>	<dbl>	<dbl>
1	LM (OLS)	0.656	3972529.	1.58e13
2	Elastic Net (glmnet)	0.657	3969323.	1.58e13
3	Random Forest (RF, untuned)	0.715	3614357.	1.31e13
4	Random Forest (RF tuned)	0.724	3558177.	1.27e13

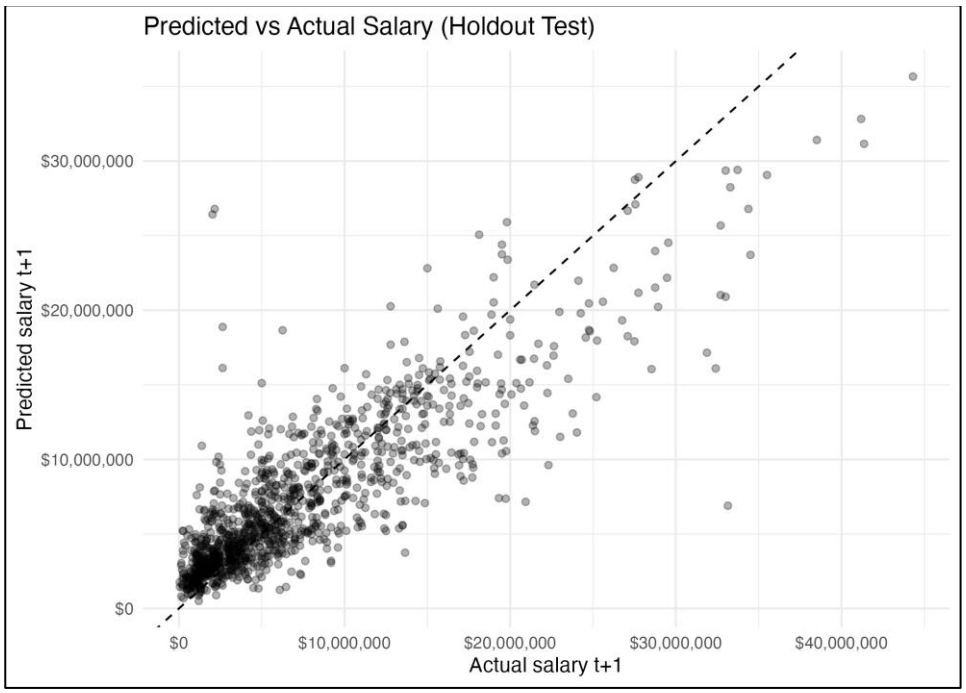
Graph 1: R-Output holdout test

The tuned Random Forest clearly outperforms all other approaches, explaining 72.4% of salary variation and achieving an average prediction error of roughly 3.5M USD, which is good for NBA contract modelling.

Interpretation of Diagnostic Graphs

Predicted vs Actual

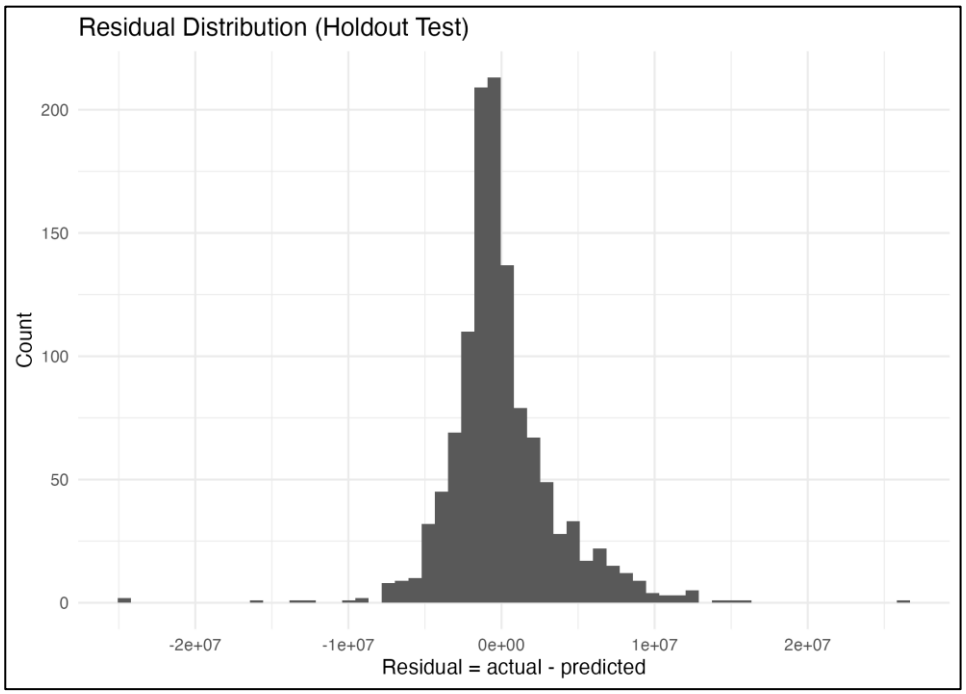
The scatterplot shows that predictions closely follow the 45° line for most players (salaries < ~$15M), indicating strong accuracy. Dispersion grows for high earners, which is normal because max-salary rules compress elite salaries and make extreme values harder to model. Despite this, the model captures overall trends well.



Graph 2: Plot of predicted vs actual salary from the holdout test

Residual Distribution

Residuals form a centred, symmetric bell curve, meaning the model has no systematic upward or downward bias. A few large outliers correspond to exceptional contract situations, such as supermax extensions or unusual negotiations, that cannot be inferred solely from performance data.



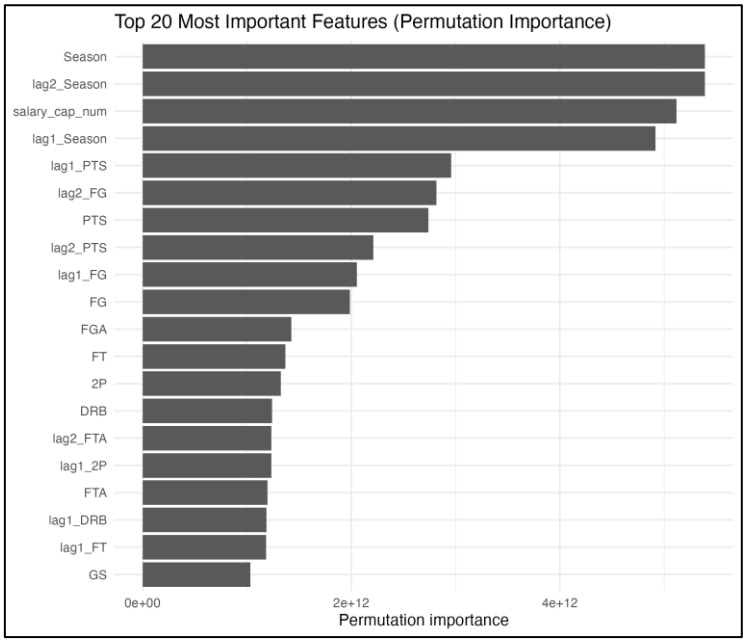
Graph 3: Plot residual distribution from holdout test

Permutation Feature Importance

The most influential variables are Season, its lagged versions, the Salary Cap, scoring metrics (PTS, FG, FGA), and rebounds and free throws. This reflects two realities:

1. Salaries depend heavily on macro factors, such as CBA changes and cap spikes.

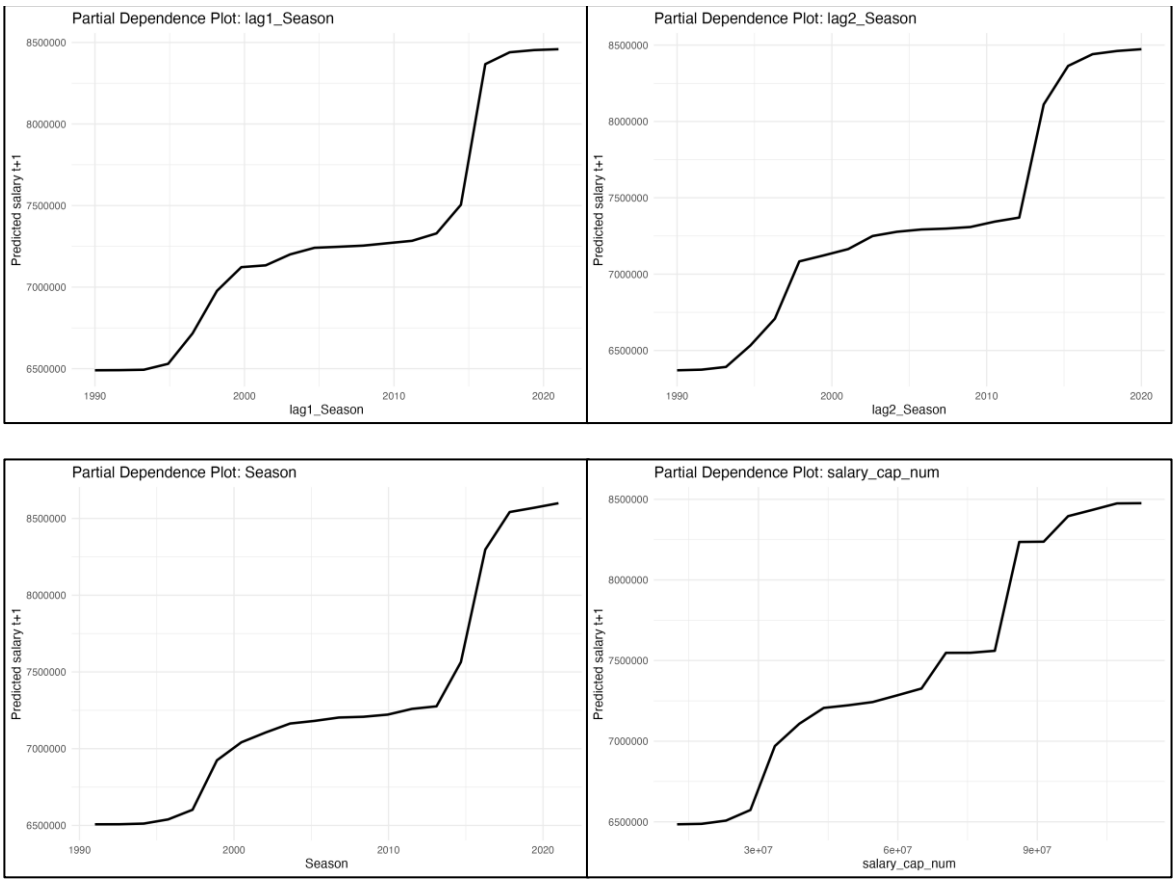
2. Scoring and core box-score production remain the main drivers of NBA contract value.



Graph 4: Plot top 20 most important features (permutation importance)

Partial Dependence Plots (PDPs)

The PDPs for Season, lag1_Season, lag2_Season, and salary_cap_num all show a clear upward trajectory: gradual growth from 1990–2015 and a sharp spike around 2016 (TV-rights boom). This confirms that the model correctly learned the long-run evolution of NBA salaries.



Graph 5: Partial dependence plots lag1 _Season, lag2_Season, Season and salary_cap_sum

Implications

The results suggest that NBA salary dynamics are largely rational and reflect a combination of individual performance, long-term player trends, and league-wide economic conditions. Scoring volume is the strongest driver of pay, while multi-year performance patterns (lags and deltas) also play a meaningful role in shaping contract expectations. Macro-level shocks, such as CBA changes or salary-cap jumps, are clearly embedded in compensation outcomes. Team success has a positive but secondary influence compared to individual production.

Limitations

Contract Information Limitations
The dataset lacks explicit contract metadata (rookie-scale tiers, max eligibility rules, Bird rights, and option years). The contract-renewal flag uses a heuristic that may misclassify some negotiation years.

Missing Performance Dimensions
Advanced metrics such as BPM, VORP, EPM, or RAPTOR are not included. Similarly, injury history, player role changes, and coaching strategies are absent.

Heteroskedastic Salary Structure
Top earners remain difficult to predict because of nonlinear collective-bargaining rules and negotiation dynamics that simple ML models cannot fully capture.

Conclusion

This project builds a realistic and leakage-free dataset and applies multiple models to understand NBA salary dynamics. The tuned Random Forest performs best, explaining over 72% of salary variation with an RMSE of about $3.5M, an excellent result given the volatility of NBA contracts.

The diagnostic graphs show a well-behaved, unbiased model: predictions are accurate for most players, errors grow logically in max-salary cases, and feature importance indicates that salaries are driven by scoring output, multi-year performance trends, and macroeconomic shifts such as salary-cap jumps. Overall, the model captures the essential structure of how NBA teams value players.

Outlook

Future improvements could include:

- Adding advanced metrics (TS%, BPM, RAPM), age, injuries, and positional data.
- Integrating real contract metadata (rookie scale, Bird rights, max-salary eligibility).
- Using hierarchical models to capture team-specific effects.
- Exploring temporal models (boosting on time sequences, RNNs) to better track career evolution.

These enhancements would bring the model even closer to real NBA decision-making and further improve predictive accuracy.

Our project begins by assessing the feasibility of predicting future NBA salaries. Moving forward, we could use these predictions to answer two additional questions:

- **The Cost of Success:** Do winning teams tend to pay a premium relative to actual player production?
- **Structure vs. Performance:** Do teams with concentrated, top-heavy salary distributions perform better than those with evenly distributed payrolls?

We started assessing these questions. After the first iterations, we were unable to support these assumptions, so further iterations on the model would be needed.

6. Literature

Berri, D. J., & Krautmann, A. C. (2024). The economics of NBA contracts: How performance, bargaining power and market forces shape player salaries. *Journal of Sports Economics*. Advance online publication. https://doi.org/10.1177/15270025251328264

Papadaki, I. (2020). *Estimating NBA players' salary share according to their performance*. arXiv. https://arxiv.org/abs/

Pargent, F., Schoedel, R., & Stachl, C. (2023). *Best practices in supervised machine learning: A tutorial for psychologists*. University of Zurich. https://doi.org/10.31234/osf.io/3j4bk

Xu, Z. (2025). *Enhanced prediction of NBA players' salaries using hybrid ensemble models and multi-objective optimization*. Expert Systems with Applications.