ML Final

Michael Strohmeier

Motivation

NBA salaries are a function of player performance, market trends, and negotiation. Understanding these drivers can help teams, agents, and players make better decisions.

Literature Review

The Kaggle project titled "NBA Salary Predictor - Data Science Final 2024" by Dillon Timmer focuses on predicting NBA player salaries using machine learning techniques.

- Feature Importance: Certain player statistics, such as points per game, assists, and player efficiency ratings, were found to be significant predictors of salary.
- Model Performance: Among the models tested, **ensemble methods** like Random Forest and Gradient Boosting provided better predictive accuracy compared to simple linear regression.

Literature Review

The project demonstrated that machine learning models could effectively predict NBA player salaries based on performance statistics. The insights gained could be valuable for teams and analysts in assessing player value and making informed contract decisions.

Algorithm	RMSE
XGBoost	5.833194e+06
Random Forest	5.987596e+06
Voting Regressor	6.797299e+06
Decision Tree	8.044886e+06
Stacking Regressor	1.103614e+07
Stacking Regressor 2	1.130020e+07

Data Overview

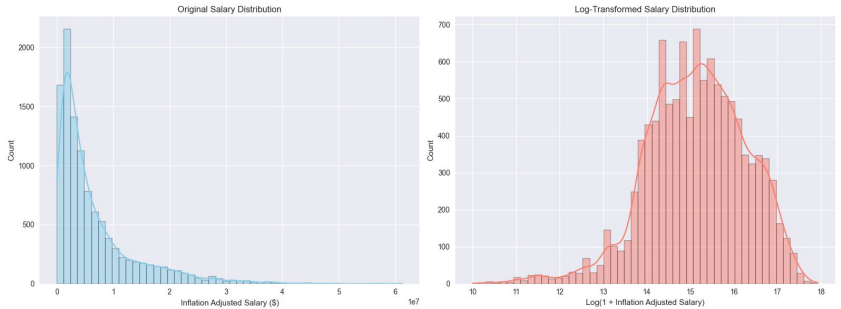
10,932 player-seasons (rows)

Key Features:

- Player Overview
  - Player
  - Season
  - Team
  - Position
  - Age
- Salary Info:
  - Salary (raw)
  - Inflation Adjusted Salary (target)
- Shooting:
  - FG, FGA, FG%
  - 3P, 3PA, 3P%
  - 2P, 2PA, 2P%
  - FT, FTA, FT%
- Game stats:
  - Games, Games Started,
  - Minutes Played,
  - Points
  - Rebounds, Assists
  - Steals, Blocks
  - Turnovers, Personal Fouls

Data Overview

- 34 player-seasons (rows) contained missing values and were dropped from the dataset
- Salary Summary
  - Min Salary : $21,899
  - Max Salary: $61,258,556
  - Med Salary : $3,893,046
  - Mean Salary: $6,501,424
  - Std Salary: $7,077,486

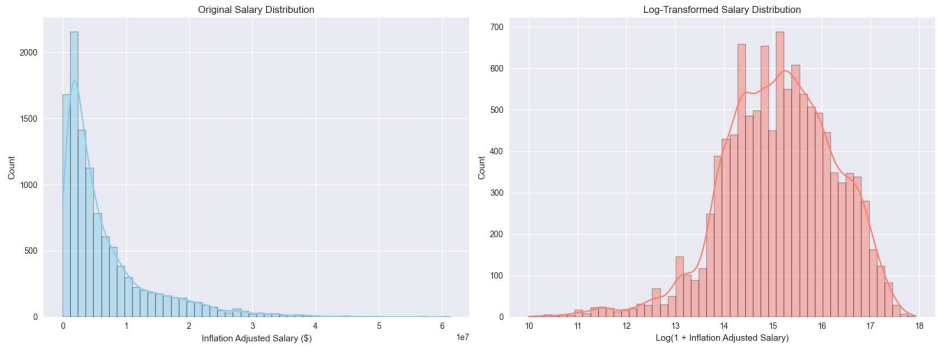


Salary Distribution

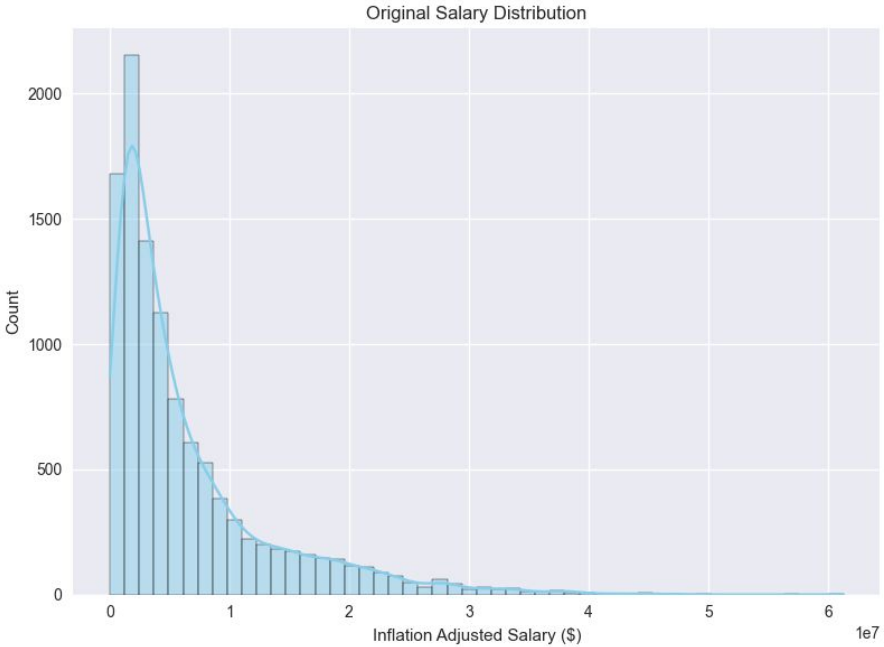
The left plot shows the real-world distribution of NBA salaries, highlighting the right-skewed nature and the prevalence of lower salaries.

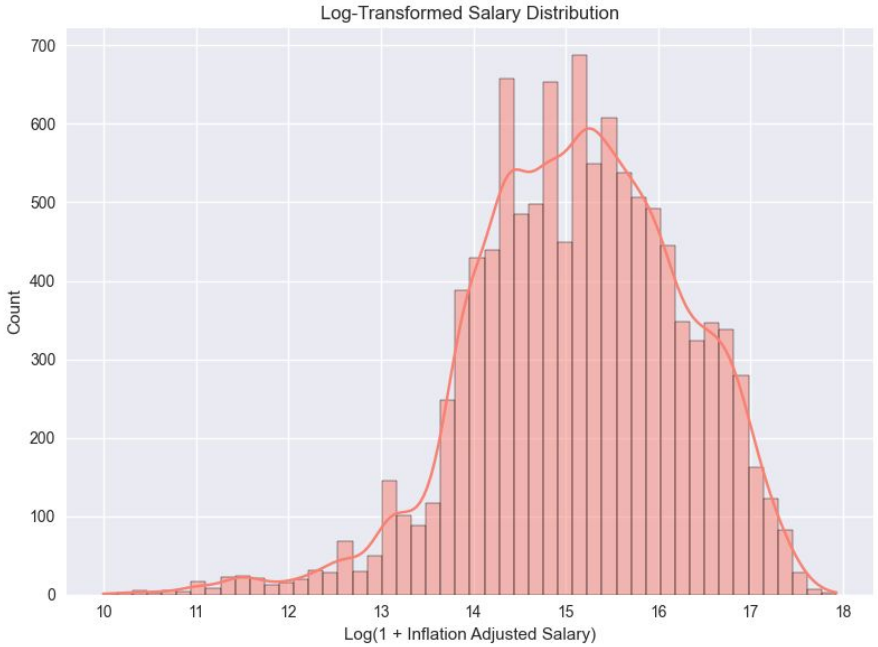
The right plot (log-transformed) reveals the underlying distribution more clearly for modeling purposes.

Log transformation is used in modeling to address skewness and improve regression performance.



Salary Distribution





Feature Engineering

Raw features: All box score stats, shooting percentages, and demographic info.

Created features:

- Per-game stats:
  - Points per Game
  - Assists per Game
  - Rebounds per Game, etc.
- Per-36-minutes stats:
  - Points per 36
  - Rebounds per 36
  - etc. (normalizes for playing time)
- Efficiency ratios:
  - Assist to Turnover Ratio

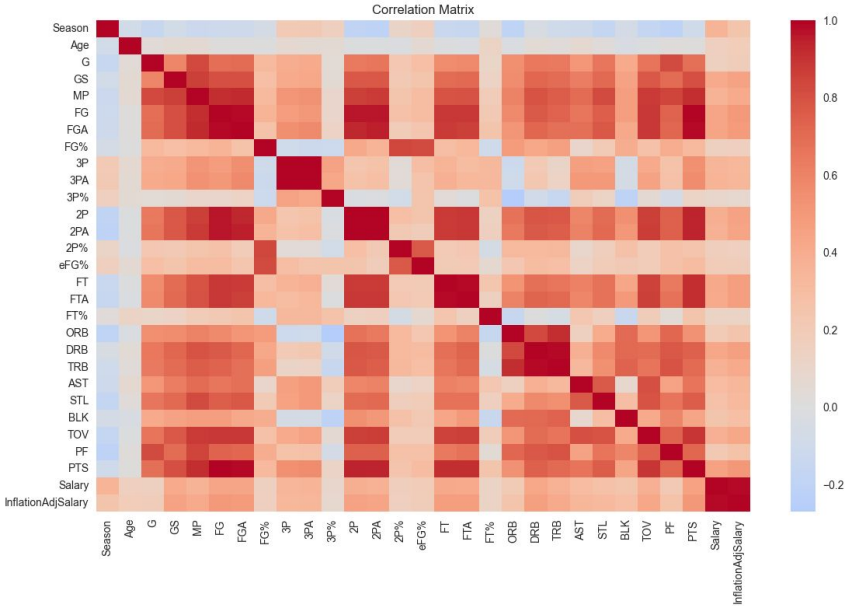
Multicollinearity handling

- Used Variance Inflation Factor (VIF) analysis to remove highly correlated features
- Final feature set: 7 features selected for modeling

Feature	Definition
Games Started	Number of games started
3-Pointers Made	Total 3-point field goals made
3-Point Percentage	Percentage of 3-point field goals made
Free Throws Made	Total free throws made
Blocks per Game	Average blocks per game
Rebounds per 36	Rebounds per 36 minutes
Assist to Turnover Ratio	Assists divided by turnovers

Correlation Analysis

Scoring and playing time are most strongly associated with higher salaries, but efficiency and advanced stats also matter



Data Preprocessing

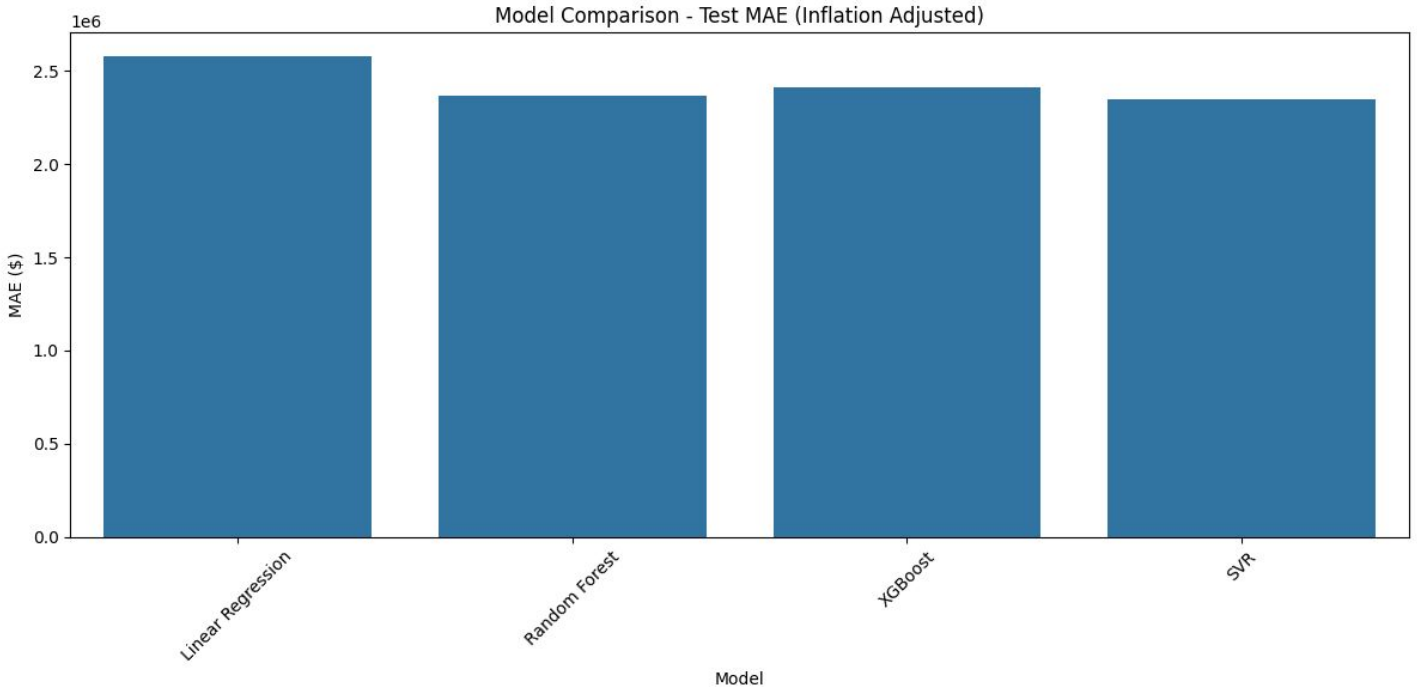
- Outlier handling: 0 salary outliers removed (after inflation adjustment)
- Missing values: 34 rows dropped (final: 10,016 rows)
- Standardization: All features scaled using StandardScaler
- Train/test split: 8,012 training, 2,004 test samples (80/20 split)
    - **Cross-validation is done on the training set**

Model Selection & Training

- Models evaluated:
  - Linear Regression
  - Random Forest Regressor
  - XGBoost Regressor
  - Support Vector Regression (SVR)
- Cross-validation: 4-fold
- Target transformation: log applied to salary to address skewness in response

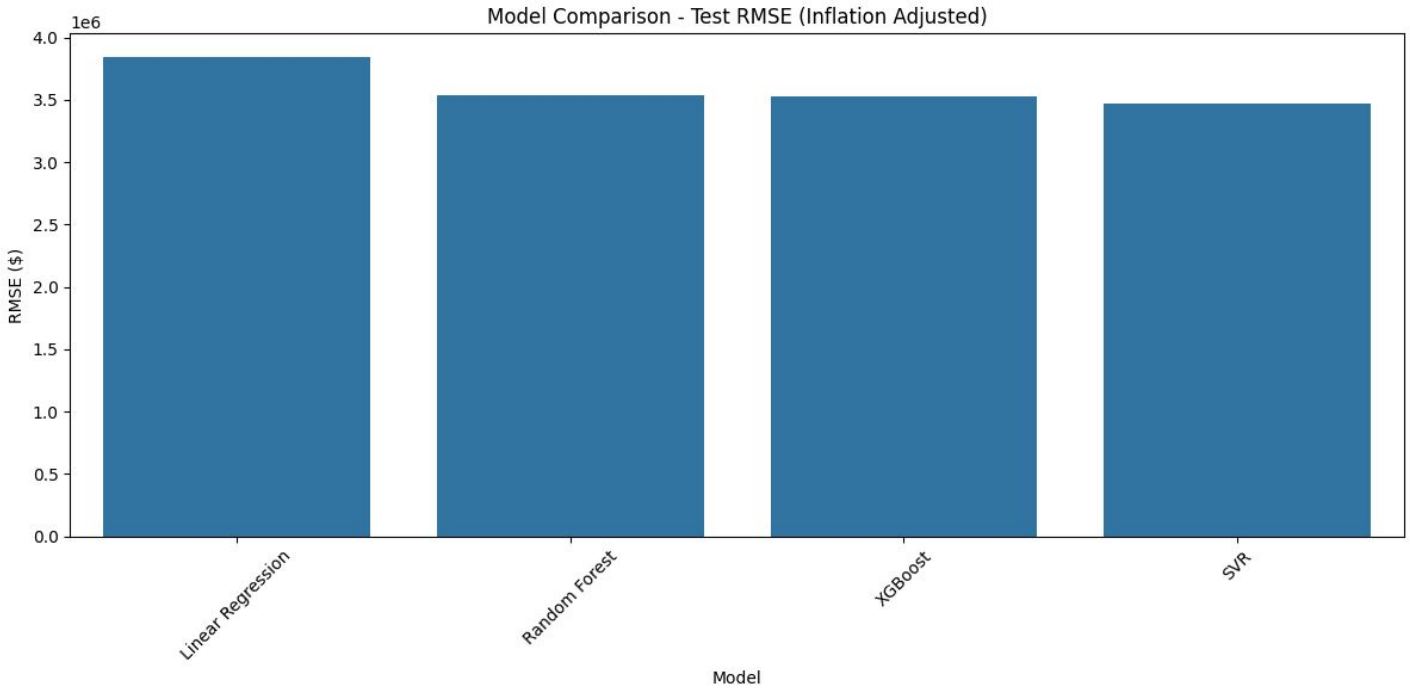
Model Selection & Training

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$



Model Selection & Training

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

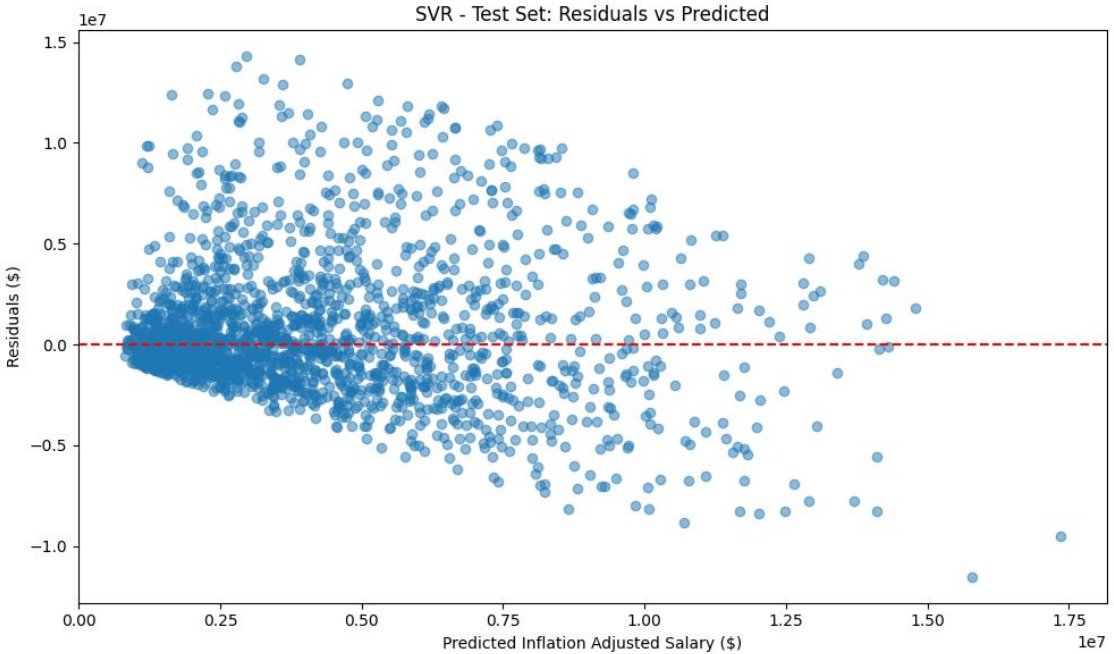


Best Model – SVR

- Why SVR?
  - Effectively captures non-linear relationships between features and salary
  - Handles outliers well and performs consistently with standardized data
  - Achieved better performance than tree-based models in this analysis
- SVR demonstrated better generalization, particularly for mid-range salaries, outperforming other models in capturing complex patterns.

Residual Analysis

Model is accurate for most players, but less so for outliers (superstars)



Conclusion

- Overall Performance: SVR accurately predicts most player salaries, effectively capturing non-linear patterns.
- Limitations: Less accurate for outliers, especially high-salary superstars.
- Recommendation: Avoid using SVR for superstars; consider specialized models for extreme salaries.