THE COMMISSION REPORT

dkharaba, dmadnani, ebabaogl, pkurani

Prediction Task

The Commission was hired by a consortium of NBA franchises to develop a model that can predict NBA salaries based on the players' statistics. The team owners would use this model to determine whether players in their respective teams are overvalued or undervalued based on their statistics.

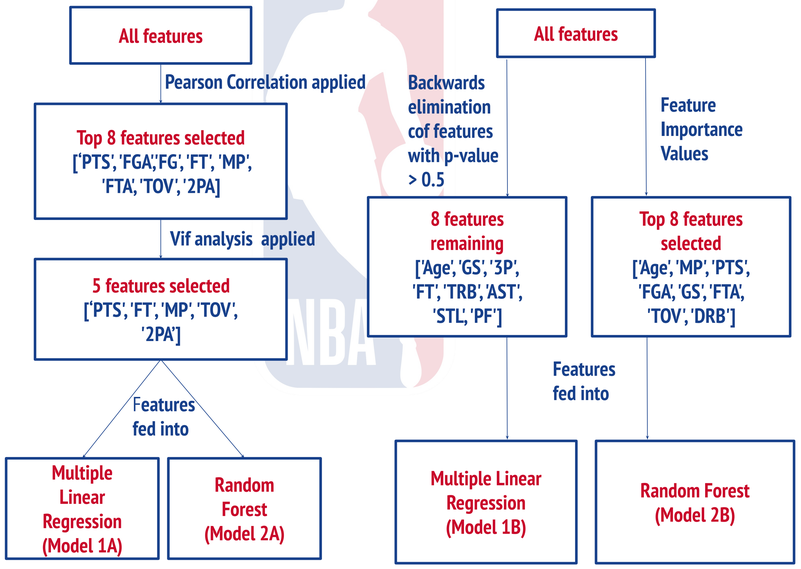
Data Collection

- The 2018-2019 Salary dataset has been scraped from Hoopshype (a sub-organization of USA Today). The dataset contained 576 rows of Players.
- The 2018-2019 per-game statistics dataset was scraped from Sport Reference. It contained 475 rows (Players) and 29 columns
- The two sets were combined making use of a left join in Pandas. They were joined on player names. In the process we lost 101 rows from the Salary dataset not present in the player statistics dataset. These rows consisted of rookie players that were on short term contracts.

Methodology

The data was randomly split 80-20 into training and test sets to ensure our models would train on a non-biased collection of salaries and statistics and then make predictions on the test set.

First Round of Analysis Flowchart



Second Round of Analysis Flowchart

Feature Classification

Efficiency Statistics:

FG% - Field Goal Percentage, **3P%** - FG% on 3-Pt FGAs, **2P%** - FG% on 2-Pt FGAs, **eFG%** - Effective Field Goal Percentage, **FT%** - Free Throw Percentage

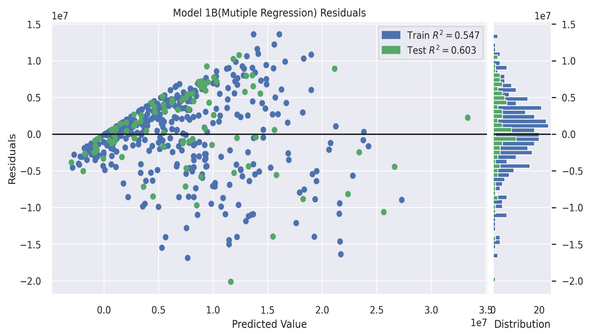
Player Statistics

Rk - Rank, **Pos** - Position, **Age** - Player's age on February 1 of the season, **Tm** - Team, **G** - Games, **GS** - Games Started

Per Game Statistics

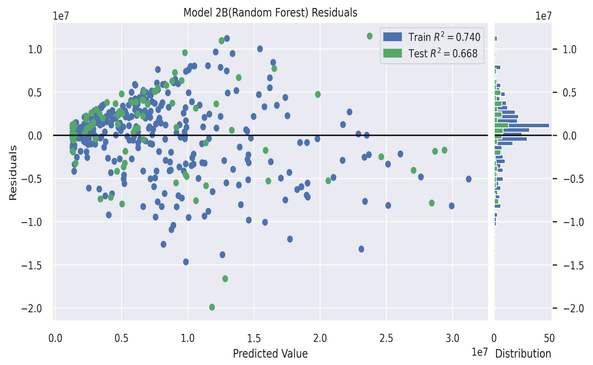
MP - Minutes Played, **FG** - Field Goals, **FGA** - Field Goal Attempts, **3P** - 3-Point Field Goals, **3PA** - 3-Point Field Goal Attempts, **2P** - 2-Point Field Goals, **2PA** - 2-Point Field Goal Attempts, **FT** - Free Throws, **FTA** - Free Throw Attempts, **ORB** - Offensive Rebounds, **DRB** - Defensive Rebounds, **TRB** - Total Rebounds, **AST** - Assists Per Game, **STL** - Steals, **BLK** - Blocks, **TOV** - Turnovers, **PF** - Personal Fouls, **PTS** - Points

Results



Plot 1 - Residuals for Model 1B

This model was trained and tested on the features selected by Backward elimination of features with p-values >0.05. This model outperforms Model 1A, Model 3 and the Baseline model.



Plot 2 - Residuals for Model 2B

This model was trained and tested on the features selected by looking at the features that made up 90 percent of the feature importance values. This is our model with the highest accuracy and lowest errors. It's hyperparameters were tuned through 4 fold cross-validation.

Table 1: Adjusted R-Squared and Test RMSE comparison of all models

Eval. Metrics	Model 1B	Model 2B	Baseline	Model 1A	Model 2A	Model 3	Model 4
Adj. R^2	0.59	0.66	--	0.49	0.50	0.54	0.59
Test RMSE	5456244	4893930	8658439	6171932	6105915	7707201	5550000

Plot 3: Top 10 features least correlated with Salary

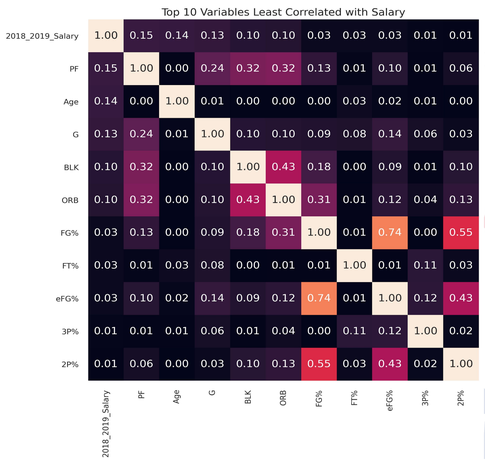
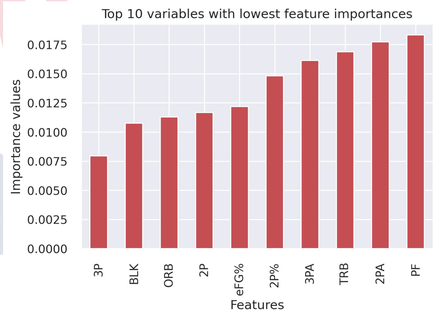


Table 2: Efficiency statistics ranked by order of elimination (p-value > 0.05)

p-value	FG %	3P %	2P %	eFG %	FT%
Elim. rank	3	9	4	6	7

Plot 4: Bar Chart of Top 10 features with lowest values of feature importances in Random Forest



Evaluation and Takeaways

To evaluate the accuracy and predictive power of our final models (**1B and 2B**), we compared them to each other, a baseline model (predicting the mean) and the two models that were run on the full set of features: **Model 3 (Multiple Regression)** and **Model 4 (Random Forest)**. Our **evaluation metrics** were **adjusted R-squared** and **testing RMSE** values. We also looked at residual plots to gain a better understanding of the bias and errors in our models and compare our final two models.

Takeaway #1: Models 1B and 2B outperformed both the baseline models and the models trained with the full set of features.

- **Table 1** shows that in terms of our evaluation metrics, Models 1B and 2B greatly outperform the baseline model and also Models 3 and 4.
- Feature selection has made these models more accurate and reduced errors in their predictions.

Takeaway #2: Using correlation to understand the relationship between the features and salary and select features leads to models underperforming.

- **Table 1** shows Models 1A and 2A underperformed compared to the models run on full set of features in terms of our evaluation metrics.
- The Pearson correlation test does not capture features that may have non-linear or very weak linear relationships with Salary
- Incorporating these features that may have non-linear relationships greatly improved performance of Models 1B and 2B

Takeaway #3: Random Forest regressor is a more accurate but more biased model for predicting the salaries of NBA players.

- **Plot 2** show us that distribution of residuals is mostly random for Model 2B as compared to Model 1B (**Plot 1**) which shows a linear relationship between residuals.
- This suggests that a non-linear model(2B) is a more applicable model for predicting NBA salaries..
- **Model 2B** has lower errors and higher accuracies but it is biased since its residuals are positively skewed

Takeaway #4: Efficiency statistics are poor predictors of salary.

- The **efficiency statistics** consistently ranked in the **bottom 10** in our feature selection methodology.
- This affirmed that idea that Player and Per-Game statistics are better indicators of Player salary (three different combination of those were in our sets of features).

Future Improvements and Limitations

- Incorporate statistics from more seasons so that our models can learn how change in performance across seasons can affect salary.
- Incorporate more features into our analysis and utilise ensemble algorithms such as AdaBoost and XGB to better understand the non-linear relationships in our data and improve prediction accuracy.
- Analyse the outliers in our data and do a statistical analysis of our actual residual and predicted salary values.
- A major limitation is that statistics don't always tell the whole story as intangible factors such as marketability and future potential are often key factors in salary decisions and cannot be captured in numbers.