Predicting NBA Player Salary - The Commission

dkharaba, dmadnani, ebabaogl, pkurani

Goal

How does an NBA team choose their player's salary? Should player statistics be a factor in the salary choice? The Commission was hired by a consortium of NBA franchises to develop a model that can predict NBA salaries based on the statistics of the players. The model would be used by these team owners to judge whether the players in their respective teams are overvalued or undervalued based on their statistics.

Data

We scraped two different datasets. The player statistics dataset, consisting of 29 features and 475 rows, was taken from Sports Reference (powered by 'sportrader', the official statistics partner of the NBA). The player salary dataset, containing 576 rows, was taken from Hoopshype (a sub-organisation of USA Today Sports). The datasets were joined on 'Player Names'. We lost 101 rows of players since they were not part of the season statistics dataset (they were on short term rookie contracts). We had to manually rename certain players due to their nicknames in the salary dataset. The combined dataset is ordered by salary in descending order.

Model + Evaluation Setup

Our ability to predict salaries accurately depended on understanding the relationship between player statistics and salaries. In our initial analysis, we sought to understand the linear relationship between our features through Pearson correlation and selected the top 8 features most related to salary which we further narrowed down to 5 using multicollinearity analysis (specifically, variation inflation factor). We implemented the multiple linear regression model and the random forest regressor to predict salaries. The data was randomly split 80-20 into training and test sets, ensuring that the models would train on a non-biased collection of salaries and statistics to understand the relationships and then use the test set to make predictions. The evaluation measures used to compare the performance of our models testing RMSE and adjusted R-squared.

Results and Analysis

Claim #1: Our two models, Multiple Regression and Random Forest, trained on features selected through backward-elimination of p-values and feature importance values respectively outperformed both the baseline models and the models trained with the full set of features.

Support for Claim #1: The table below shows the comparison of our testing RMSE metric and the adjusted R-squared value to those of the baseline model (always predicts the mean) and models trained on full set of features

Metrics	Final Multiple Regression	Final Random Forest	Baseline	Full Features Multiple Regression	Full Features Random Forest
Adj. R-squared	0.59	0.66	--	0.54	0.59
Test RMSE	5,456,244	4,893,930	8,658,439	7707201	5,550,000

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

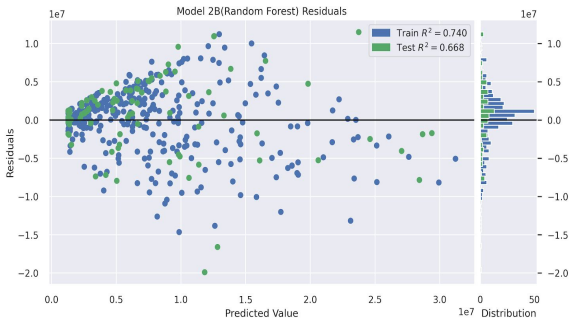
Support for Claim #2: We initially selected features for our models based on Pearson correlation coefficients and multicollinearity analysis ['PTS', 'FT', 'MP', 'TOV', '2PA']. The models using these features underperformed compared to the full set of features. The Pearson test did not fully capture nonlinear relationships or features that may have weak linear relationship with salary but have an impact on predictions (most prominently 'AGE') . Therefore, to improve the models we used Backward Elimination of features based on

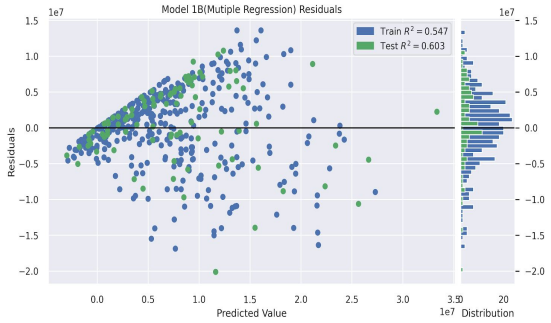
p-values select features ['Age', 'GS', '3P', 'FT', 'TRB', 'AST', 'STL', 'PF'] for Multiple Regression and the feature importance value function for Random Forest to select features (['Age', 'MP', 'PTS', 'FGA', 'GS', 'FTA', 'TOV', 'DRB']). Incorporating features that may have non-linear relationships greatly improved performance of the models.

Metrics	**Multiple Regression Pearson Features**	**Random Forest Pearson Features**	**Multiple Regression Final Backward Elimination**	**Random Forest Final Feature Importances**
Adjusted R-squared	0.49	0.50	0.59	0.66
Test RMSE	6,171,932	6,105, 915	5,456,244	4,893,939

Claim #3: Random Forest regressor is a better model for predicting the salaries of NBA players.

Support for Claim #3: As can be seen from tables above, Random Forest has better adjusted R-squared and testing RMSE value when compared to Multiple Regression. Furthermore, looking at the residual plots below, we can see the distribution of residuals is mostly random for Random Forest as compared to Multiple Regression whose plots show linear relations. This may be the case due to the fact that Random Forest is better at catching non-linear relationships and making more accurate predictions because it is an ensemble algorithm. However, it is worth noting that Random Forest has a bias in its predictions (it's residuals are positively skewed) compared to Linear Regression. Thus, Random Forest is a more accurate model but has more bias.





Claim #4: Efficiency statistics are poor predictors of salary even though they are commonly used as a measure to judge players.

Support for Claim #4: We selected three different combinations of features for our models using Pearson Correlation, Backward Elimination and feature importances. The **efficiency statistics** were in the bottom 10 out of 26 features in all cases and were not used in any of the three sets of features used to predict salary.

Measure	**FG%**	**3P%**	**2P%**	**eFG%**	**FT%**
Pearson Correlation Rank	Rank 22	Rank 25	Rank 26	Rank 24	Rank 23
p-value rank by order of elimination	Rank 3	Rank 9	Rank 4	Rank 6	Rank 7
Feature Imp. Rank	Rank 18	Rank 27	Rank 24	Rank 21	Rank 23

Feature Classification:

Group 1-Efficiency

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

Group 2- Player Statistics

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

Group 3- Per Game Statistics

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