NBA salary prediction

(Crest Gold)

Jamin Wong

Adrian Ng

Jack Mok

Kinson Chow

Report: <https://rpubs.com/Jamin/nba_player_salary_202223>

All documents: <https://github.com/jamingreen/NBAPredictions>

Summary:

This report aims to find the correlation of the player statistics in the year before and the salary of that player in the year after.

We found player statistics for 2021-2022 season and their respective salary in 2022-2023 season. We have collect the player statistics and salary of 384 NBA players that played in 2021-2022 and have salary in the 2022-23 season. The player statistics is collected from [Basketball Reference](https://www.basketball-reference.com/) and [NBA official website](https://www.nba.com/). The detailed links are listed below in the Source of files section.

We started with merging the the data from different source and removing repeated variables.

In the exploratory data analysis, we start with univariate analysis of the response variable salary and then explore the correlation of individual variables with salary.

In the modelling section, we use the RMSE score on the train set to determine the final model. We have 10 fold cross validation on each model (the 10 fold all use RMSE to hyperparameter tuning). The root mean squared error (RMSE) is a measure of prediction error while R-squared is a measure of the proportion of variation explained by the model.

We have included the following modelling methods:

\* Linear Regression (Train RMSE: 0.4996, R-squared: 0.7285)

\* Lasso Regression (Train RMSE: 0.5246, R-squared: 0.7076)

\* Elastic Net Regression (Train RMSE: 0.5134, R-squared: 0.7178)

\* Stepwise Regression (Train RMSE: 0.5072, R-squared: 0.7535)

\* Decision Tree (Train RMSE: 0.5096, R-squared: 0.6438)

\* Random Forest (Train RMSE: 0.2264, R-squared: 0.7097)

\* Neural Network (Train RMSE: 0.6170, R-squared: 0.6624)

As the Random Forest Model has the lowest RMSE, we chose it as the final model for our prediction. It has 0.4835 of estimate out-of-bag RMSE.

In the interpretation part, we will mainly use linear regression and principle component analysis to analyse. We will not use the random forest model as it is hard to give interpretation on a random forest model.

In the analysis, imperial units are used as it is the common unit systems in the NBA. All the salary is record in USD.