QBUS2820 Assignment 1

This version was compiled on September 29, 2020

Introduction

Although the NBA is known for being a sport league across globe, it is a vast economic entity as well. Undoubtedly, it has been a major impact in the past decades, and it does not seem to be slowing down anytime soon. Hence, economy is also a huge part of the business. Beside the League's branding, its commercial success is contributed by the players at large as they make the trends on social media and attract costumers to buy their products in a constance. However, the most important attribute of a player is none other than his performance on the court. Performance is what NBA players thrive for as it decides their salary level. How much salary a player is worth can be a hard estimation to the teams because the performance of athlete fluctuates. Furthermore, the salary cap of the League as a whole, too, fluctuate every year. Fortunately, the League records players' data in various categories which include field goal attempted, field goal percentage, offensive and defensive ratings, etc. Data is a powerful tool because it can reflect a player's contribution on the court with precision. Accompanied by the comparison of the salaries given to a certain level of player, data can serve as a strong reference that allows objective calculations.

This project aims to develop several predictive models of salary for NBA basketball players. Three models including , k-nearest neignbour model, a linear regression model and a lasso regression model are involved.

summarising findings

Data processing and exploratory data analysis

Two datasets `NBA_train.csv` and `NBA_test.csv` are analysed in this project. The data is collected by NBA, with the corresponding raw data and metadata being publicly accessible on the NBA websites. There are 2 categorical variables and 19 numeric variables regarding players' personal information and game performance, with an additional unique ID of each record in the datasets. The numeric variables includes salary, age, number of games played , number of minutes played, personal efficiency rate , true shooting percentage , offensive rebounds , defensive rebounds , turnover percentage , assists , steals, blocks , turnover percentage , usage percentage , offensive rating, defensive rating and win shares while the categorical variables are the position and the team a player in.

Figure 1 illustrate that win share, defensive win share, offensive win share, number of minutes played and personal efficiency rate show linear relationships with salary, with win share having the strongest linear relationship with salary at a correlation coefficient of 0.68. It also provides evidences of linearity between offensive win share, defensive win share and win share.

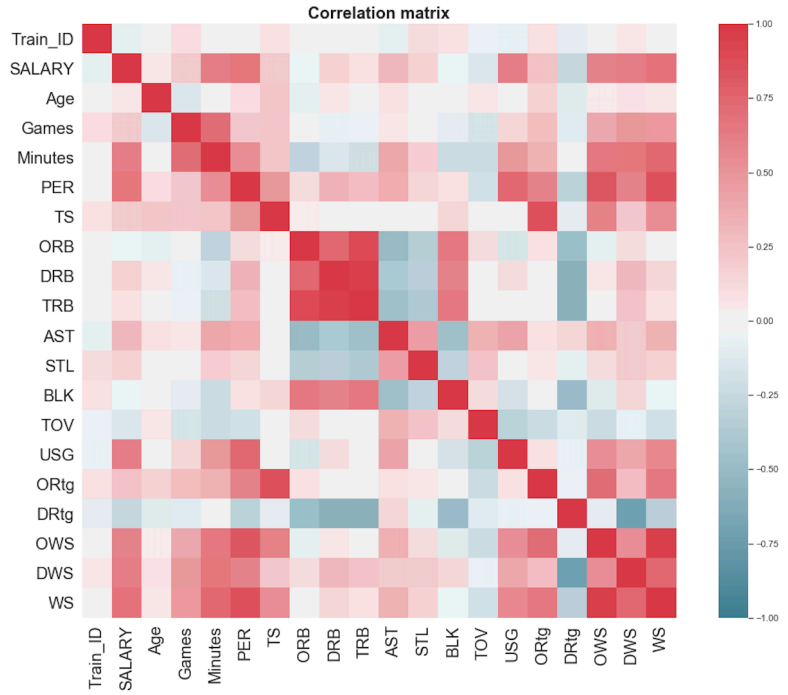


Fig. 1. Correlations between numeric variables based on correlation coefficients.

The relationships between salary and the six relative variables as well as the distributation of numeric variables are further visualized by a scatter plot matrix Figure 2. The linearity between numeric variables and salary shown in Figure 2 is in line with the correlation matrix. Salary, win share, defenssive win share and offensive win share are significantly right-skewed while usage percentage and personal efficiency rate are slightly right-skewed. Moreover, the distributiotion demonstrates a small variance of number of minutes played.

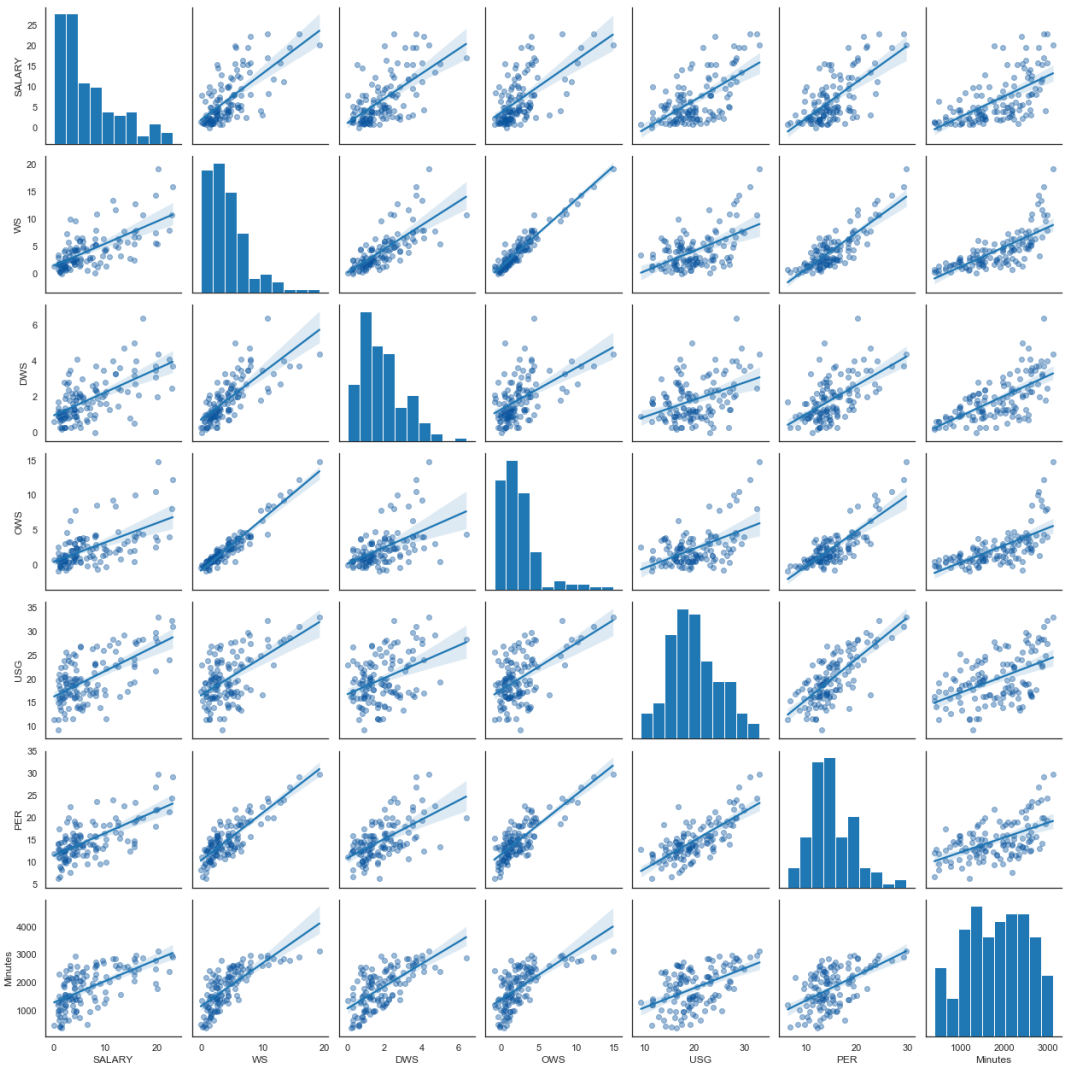
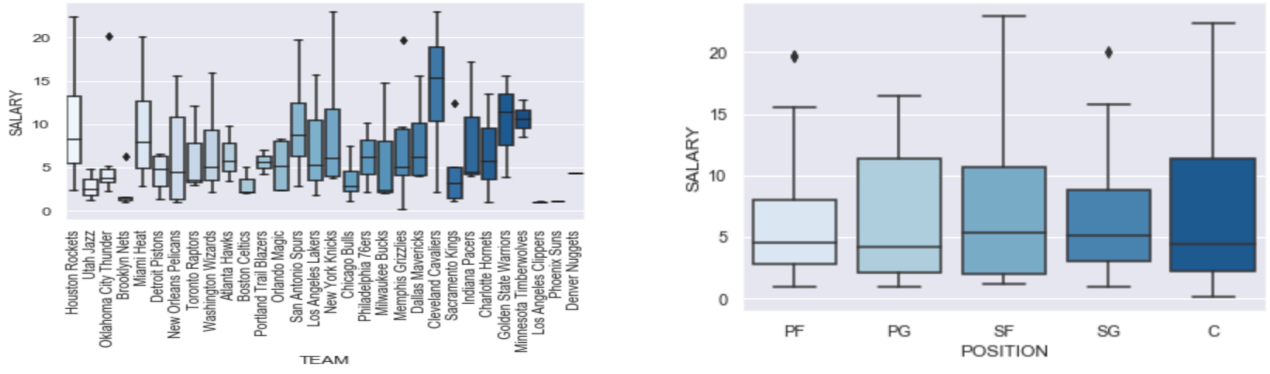


Fig. 2. Distribution of numeric variables.



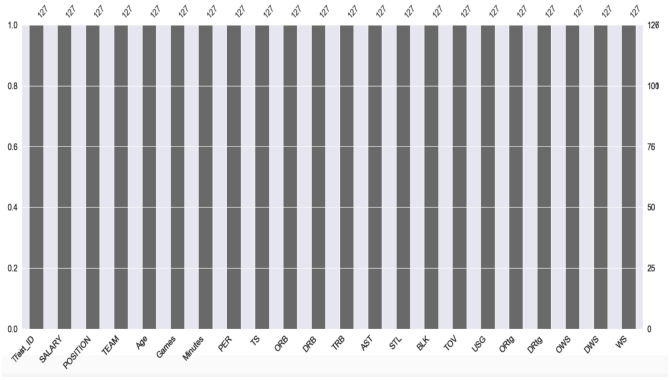
(a) Salaries for players in different teams.

(b) Salaries for players in different positions.

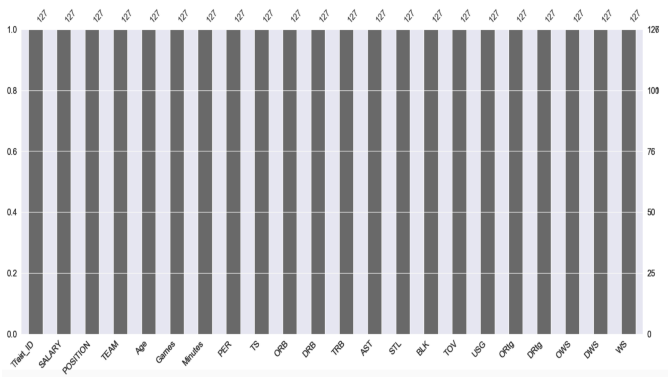
Fig. 3. Box plots of salaries for players in different teams and positions.

Feature engineering

To discover any missing values involved in the datasets, barcharts of missingness are generated to visualize missingness. As shown in Figure 4, both `NBA_train` and `NBA_test` are complete without any missing values. Therefore no data cleaning process is performed.



(a) Missingness barchart of the 'NBA train' dataset.



(b) Missingness barchart of the 'NBA test' dataset.

Fig. 4. Visualizing the missingness of two datasets used.

Methodology of the linear regression and kNN regression models

Variable	Number of neignbours	Validation error
DWS	20	4.1311
WS	8	4.1856
Minutes	9	4.2967
OWS	27	4.5046
PER	18	4.7217

(a) Top 5 K nearest neighbor models with the highest validation errors.

Polynomial degree	Validation error
[2, 3]	4.340583
[2, 3, 4]	4.350843
[2]	4.361685
[2, 3, 4, 5]	8.342030
[2, 3, 4, 5, 6]	8.407928

(b) Top 5 polynomial linear regression models with the highest validation errors.

Fig. 5. Validation errors of 10 models developed.

Methodology of the model that is not covered in this unit

Test set performance