QBUS2820 Assignment 1

This version was compiled on October 4, 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.

summarizing findings

Data processing and exploratory data analysis

Two datasets `NBA_train` and `NBA_test` 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.

The `NBA_train` dataset is used for training and validating predictive models in this project while the `NBA_test` dataset is used for testing selected models. Therefore the exploratory data analysis is conducted based on the `NBA_train` dataset.

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. Although other variables show mild linear relationship with salary, there can be other linkage between salary and these variables. Thus variables showing no linearity with salary can still be potentially informative and should be left for further feature selection when developing predictive models.

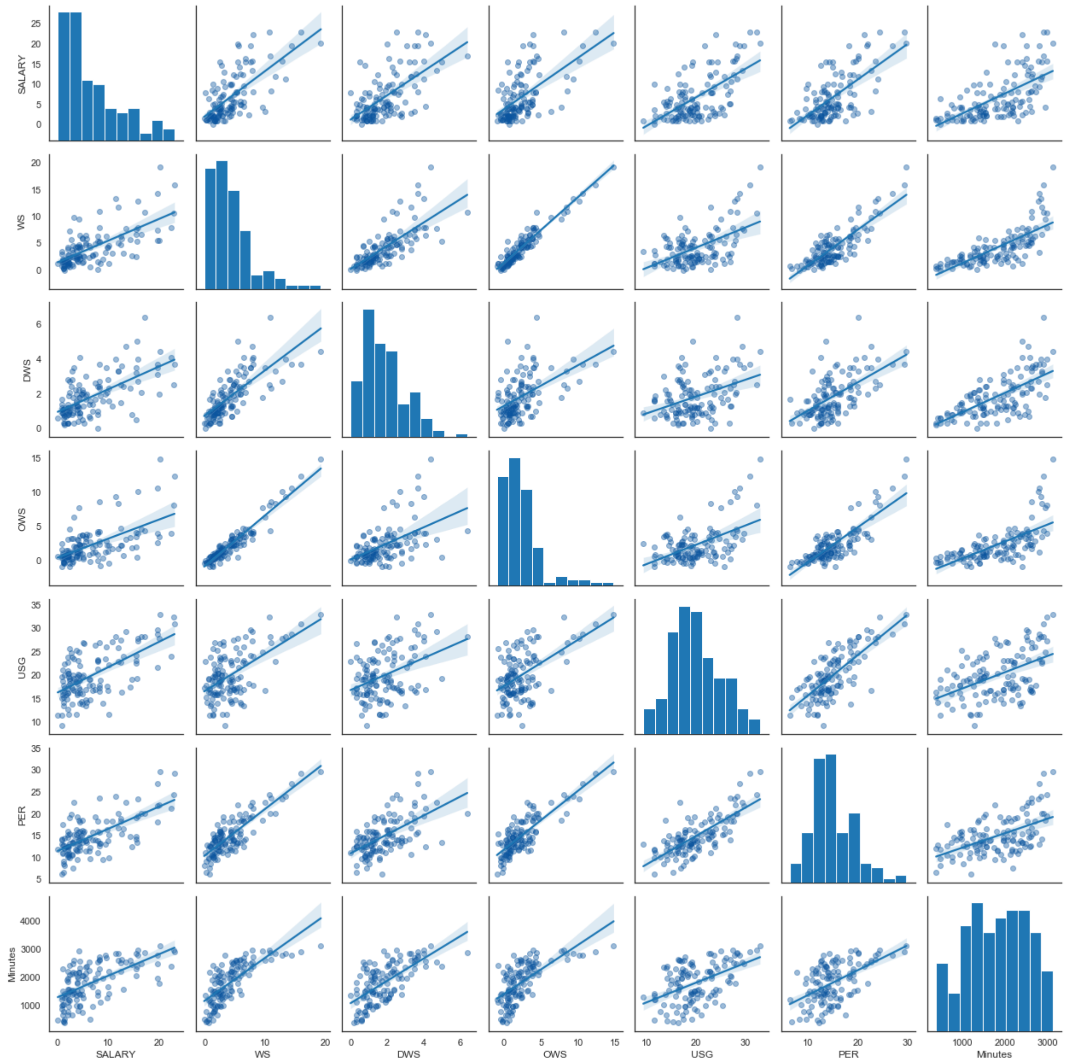


Fig. 2. Distribution of numeric variables.

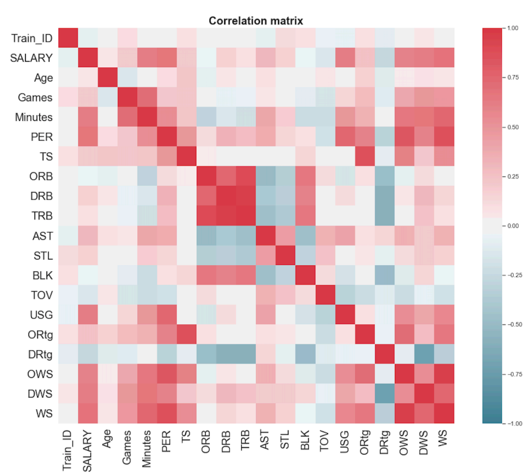
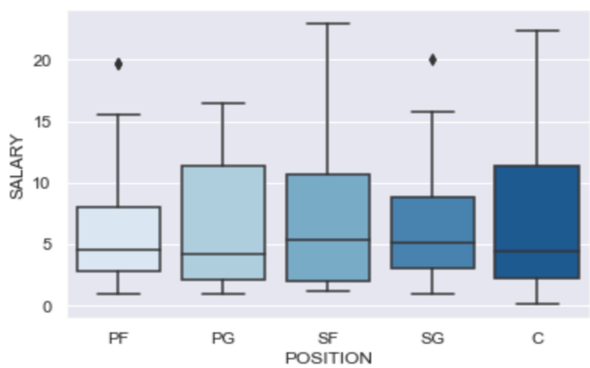


Fig. 1. Correlations between numeric variables based on correlation coefficients.

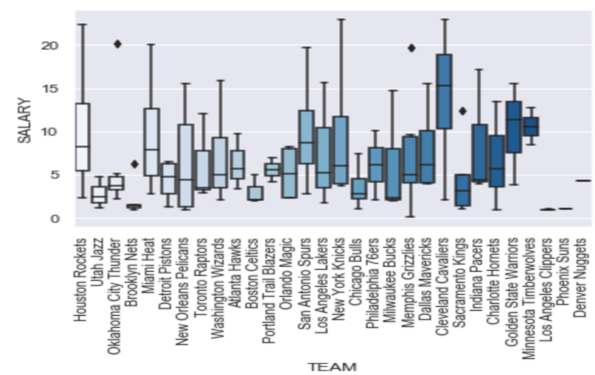
The relationships between salary and the six relative variables as well as the distribution of numeric variables are further visualized by a scatter plot matrix. In Figure 2, the linearity between numeric variables and salary shown is in line with the correlation matrix (Figure 1). Moreover, salary, win share, defensive win share and offensive win share are significantly right-skewed while usage percentage and personal efficiency rate are slightly right-skewed. In additions, the distributions demonstrates a small variance of number of minutes played.

To analyse the categorical variables, Figure 3 is generated to visualize the distribution of salary. Figure 3a describes how salary varies for players in different positions. Although the median salaries are similar at $4-6 millions for the five positions, variances of salary are slightly different. Salaries for players in the center and small forward position vary significantly without any outliers whereas variances of salary for both power forward and shooting guard are smaller with an outlier. Nevertheless, the distributions of salary for players in different positions are similar.

Outlining in Figure 3b, salary varies across different teams, which is reasonable in a business entity. There are many basketball teams within the NBA, resulting in small sample sizes of salary in each group. Some groups, such as Los Angeles Clippers, Phoenix Surs and Denver Nuggets, have information of only one player being recorded in this dataset. Furthermore, the team variable has no intrinsic order, and the teams in any unseen data can contain new teams. This makes the team variable less informative.



(a) Salaries for players in different positions.

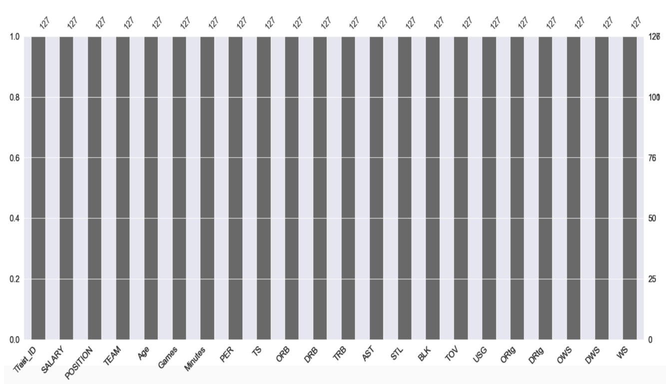


(b) Salaries for players in different teams.

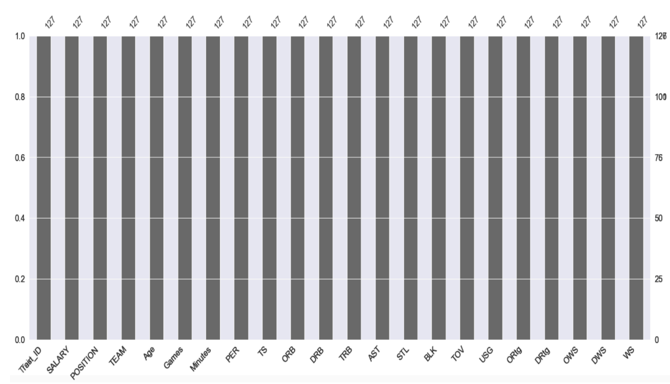
Fig. 3. Box plots of salaries for players in different teams and positions.

Feature engineering

To discover any missing values involved in the datasets, bar charts 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 removal or data imputation is performed.



(a) Missingness bar chart of the 'NBA train' dataset.



(b) Missingness bar chart of the 'NBA test' dataset.

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

According to the results of exploratory data analysis, the record ID, position a player in and team played are uninformative for predicting salary. Therefore these three variables are discarded whereas the salary is extracted from the `NBA_train` and `NBA_test` datasets to be the response. There are 19 numeric features including 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, win shares and team the player in, engaging in predictive model development.

After selecting and preprocessing the potential features and response, following the 80/20 rule, the NBA_train dataset is splitted into the training set and validation set. The training set is used to develop models while the validation set helps to select models developed.

Methodology of K-nearest neighbour regression models

Methodology of linear 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 lasso regression models

Test set performance

Analysis and conclusions

Appendix

References

- Bilogur, (2018). Missingno: a missing data visualization suite. Journal of Open Source Software, 3(22), 547, https://doi.org/10.21105/joss.00547.
- NBA Stats. (2018). NBA Stats. [online] Available at: https://stats.nba.com/.
- Rençberolu, E. (2019). Fundamental Techniques of Feature Engineering for Machine Learning. [online] Medium. Available at: https://towardsdatascience.com/.
- Richards, J. (2020). Why We Use an 80/20 Split for Training and Test Data Plus an Alternative Method. [online] Medium. Available at: https://towardsdatascience.com/.