QBUS2820 Assignment 1

This version was compiled on October 5, 2020

Task 1

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

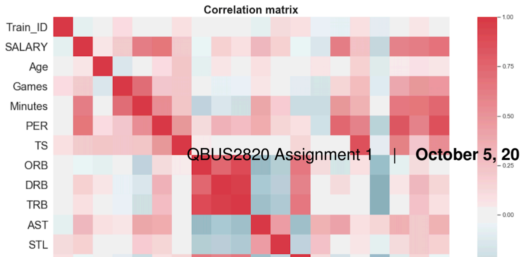
Data processing and exploratory data analysis

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



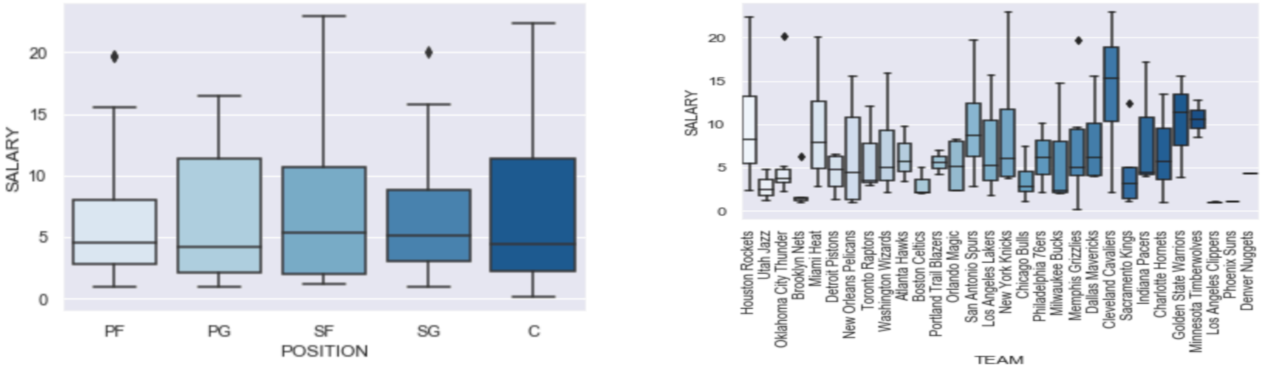
with salary at a correlation coefficient of 0.68. It also provides evidences of linearity between offensive win share, defensive win share and win share. Although other variables show mild linear relationship with salary, there can be other linkage between salary and these variables. Thus variables showing no linearity with salary can still be potentially informative and should be left for further feature selection when developing predictive models.

Moreover, colinearity is observed between several variables. One the one hand, the number of win shares which is linearly related to salary is also found to be correlated with number of offensive win shares and number of defensive win shares. One the other hand, total rebound is linearly related to both offensive rebound and defensive rebound.

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.

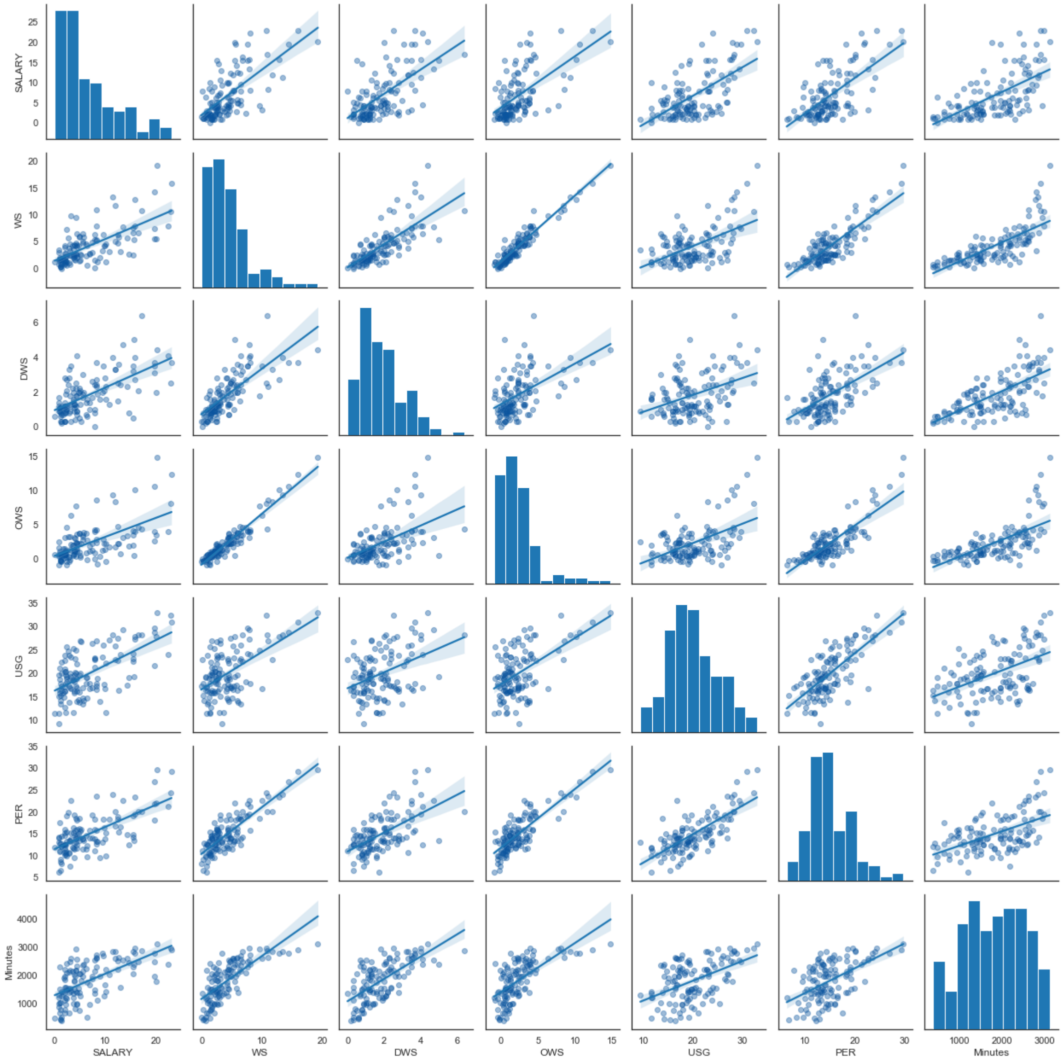
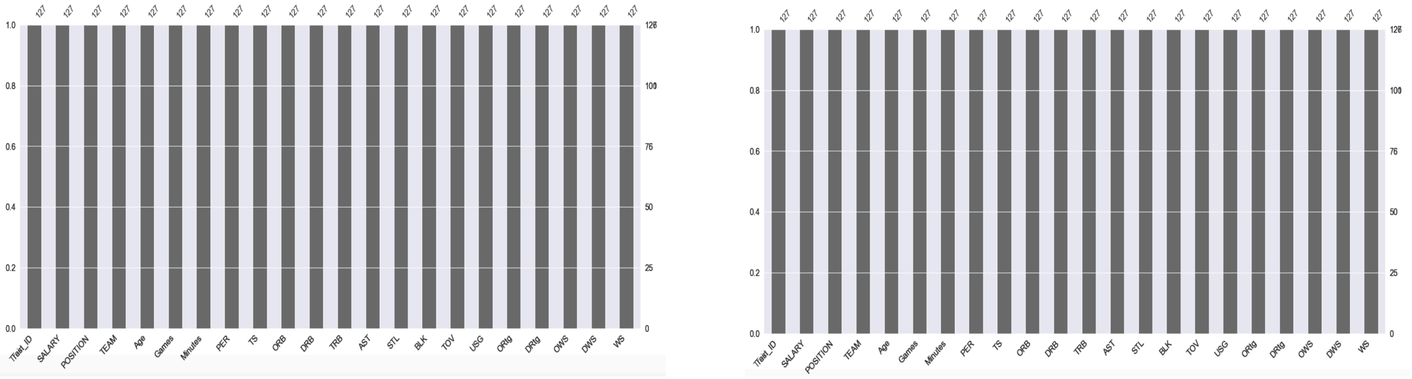


Fig. 2. Distribution of numeric variables.

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.

Methodology of K-nearest neighbour regression models

The K-nearest neighbour regression models are trained by one of the 19 numeric features, with different values of K ranging from 1 to 50. Totally 950 models are developed by changing the feature and the value of k. 5-fold cross validation is applied to assess negative mean square errors of models, followed by transferring negative mean square errors to root mean square errors. The model with the smallest root mean square error is selected as the optimal model.

The model trained by the number of win shares and a k of 19, which has a validation error of 4.2615 ($ Millions), is chosen to be the optimal K-nearest neighbour regression model. With this model, 19 neighbours are considered to examine the value of salary with specific number of win shares. Figure 5 visualizes this k-nearest neighbour regression model with observed data points of salary against number of win shares. Generally, the salary is predicted to increase as number of win shares increases, which is logical with the economic concern of the NBA. The more games a player win, the more valuable he is in the basketball team and thus the play deserves a higher salary. The model predicts salary to remain stable when 10 win shares

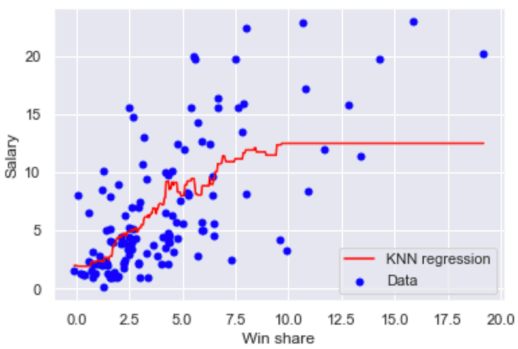


Fig. 5. Visualizing the K-nearest neighbour model with K = 19 and number of win shares as the feature.

is reached, which is not in line with observed data points.
This may result from the small number of data points with number of win shares greater than 10, resulting in insufficient neigbours for k-nearest neighbour model training.

Methodology of linear regression models

The polynomial regression models are trained by one of the 19 numeric features with a polynomial degree varying between 1 and 10. 190 polynomial regression models are developed by selecting different feature and the polynomial degree of the model.

Multiple linear regression models are also developed with different features. Recursive feature elimination (RFE) which is a feature selection method is applied. It removes weakest features based on corresponding coefficients of linear regression model until the required number of features is reached. The dependencies and colinearity between features are also captured and eliminated by this method. As discussed in the exploratory data analysis, there is colinerity between number of defensive win shares, offensive win shares and win shares, and between total rebound, offensive rebound and defensive rebounds. In order to avoid violating the assumption of multiple linear regression, the number of offensive win share and defensive win shares, offensive rebound and defensive rebound are removed from the feature pool for feature selection, leading to 15 numeric variables remaining in the feature pool. The number of features required for recursive feature elimination varies from 1 to 15, as such simple linear regression models with one feature are also estimated in this approach.

Negative mean square errors of these models are obtained by implementing 5-fold cross validation. Root mean square errors are then calculated to determine the optimal models of polynomial linear regression and multiple linear regression.

The optimal polynomial regression model contains number of win shares as the feature and a polynomial degree of 2. The predictive function of the model is: $Salary = 1.4213 + 1.4339\ WS - 0.0219\ WS^2$, where WS is the number of win shares. As shown in Figure 6, The polynomial regression better fits the observed data compared to the linear regression model as it captures more variance of salary with regard to different number of win shares.

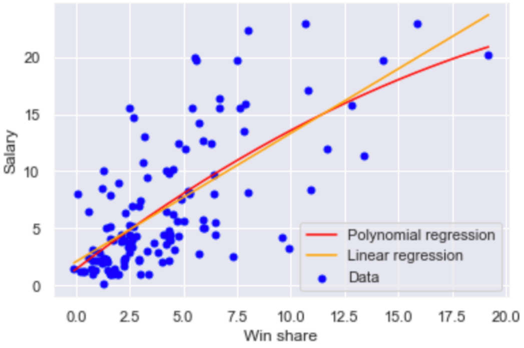


Fig. 6. Visualizing the polynomial regression model and linear model with number of win share as feature. The polynomial regression model has a polynomial degree of 2.

The optimal multiple regression model involves 14 numeric features, having the root mean square error of 3.792 ($ Millions). The predictive function is $Salary = 10.53 - 0.1\ Age - 0.16\ Games + 0.0037\ Minutes - 0.25\ PER - 0.098\ TS + 0.2\ TRB - 0.062\ AST + 0.58\ STL - 0.21\ BLK + 0.12\ TOV + 0.51\ USG + 0.11\ ORtg - 0.19 DRtg + 0.49\ WS$, where salary rises with incrase in number of minutes played, total rebounds, number of steals, turnover percentage, usage percentage, offensive rating, win shares, and decrease in age, number of games played, personal efficiency rating, true shooting percentage, assists, blocks, defensive rating.

With a smaller root mean square error of training model, the optimal multiple linear regression model including 14 features are finally selected as the optimal linear regression model.

Methodology of lasso regression models

Least absolute shrinkage and selection operator (Lasso), as an extension of liner regression analysis, conducts both feature selection and regularization. This helps to enhance the predictive performance and interpretability of the model developed.

The objective of a lasso regression model is to minimize $\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \alpha\sum_{j=1}^{p}|\beta_j|$, where α is a tuning parameter which represents how strong the L1 regularization penalise coefficients of the lasso regression model. Changing the value of α influences the number of features eliminated. When $\alpha = 0$, coefficients of features are not panalised such that no feature is removed. As the value of α increases, L1 penalty gets stronger and therefore more features are eliminated, vice versa. Furthermore, the value of α also affects the bias-variance trade-off. An increase of α leads to increase in bais while a decrease of α results in increase in variance.

Lasso regression model automates feature selection whereas multiple linear regression model require additional feature selection approach to deal with multicollinearity. Moreover, L1 regularization which penalizes the coefficients of linear regression model is performed with lasso regression. In the source data of this project, outliers and multicollinearity exist in several pairs of features. In this case, lasso regression with automated feature selection and regularization could be better-suited for the data compared to regular linear regression.

Lasso regression is a supervised learning technique while K-nearest neighbour regression is an unsupervised learning technique, which means a linear funcation is pre-deined for lasso regression when K-nearest neighbour regression observes pattern in the data without fix function. As such, lasso regression is more interpretable but less flexible than k-nearest neighbour regression. In this project, one of the goals is to discover dominant factors of salary for NBA players. To acheive this goal, a lasso regression model which can clearly define weights of features is more powerful compared to a KNN regression model.

19 numeric features are fed into lasso regression models with 7 values of alpha: 0.0001, 0.001, 0.01, 0.1, 0.5, 5 and 10. 5-fold cross validation is then applied to assess predictive performance of models. The optimal lasso model is selected based on negative mean squared error, the smaller the negative mean squared error, the better the lasso model performs.

The optimal value of α is 0.5 for this data. With this value of α, 10 features are selected with the predictive function: $Salary = 25.18 - 0.0044\,Age - 0.18\,Games + 0.0056\,Minutes + 0.1\,PER + 0.036\,DRB - 0.054\,AST + 0.098\,TOV + 0.27\,USG - 0.23\,RDtg + 0.17WS$. This indicates that salary increases with increases in number of minutes played, personal efficiency rate, defensive rebounds, turn over percentage, usage percentage and number of win shares while decreases in age, number of games played, assists and defensive rating result in an increase in salary. The validation root mean squared of the model is -12.8575 ($ Millions), which is the best performance among 7 lasso regression models developed.

Test set performance

Analysis and conclusions

KNN Discussion A potential solution to this issue is to choose a smaller values of k so that the salary is estimated based on fewer neibours. However, this could be a trade-off to worsen the predictive performance of the model.

Appendix

References

- Pedregosa, F., et al. (2011). “Scikit-Learn: Machine Learning in Python.” Journal of Machine Learning Research, vol. 12, pp. 2825–2830.

- Stephanie (2015). Lasso Regression: Simple Definition. [online] Statistics How To. Available at: https://www.statisticshowto.com/lasso-regression/#:~:text=Lasso%20regression%20is%20a%20type.
- 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/.
- Wikipedia Contributors (2019). Lasso (statistics). [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Lasso_(statistics).
- 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/.
- Basketball-Reference.com. NBA Win Shares. [online] Available at: https://www.basketball-reference.com/about/ws.html.

Task 2

Exploratory data analysis

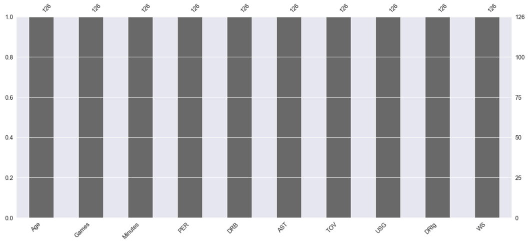


Fig. 7. Barchart of missingness for 'Boston housing' dataset.