P8106 - Final Project - NBA Players Salary Prediction

Mingkuan Xu, Mengfan Luo, Yiqun Jin

Introduction

For teams in the National Basketball Association (NBA), a key strategy to win more games is to properly allocate their salary cap - an agreement that places a limit on the amount of money that a team can spend on players' salaries. How to evaluate the performance of each NBA player and give a suitable level of salary is a complicated problem. In this project, we intend to predict the salary of NBA players in the 2021-2022 season based on their game statistics. We collected game statistics that are commonly used to evaluate players from the NBA official website, built both linear and non-linear models, including linear regression, elastic net regression, principle component regression (PCR), generalized addictive model (GAM), multivariate adaptive regression spline (MARS) model, random forest and neural network on selected feature variables, and compared these models to determine a final predictive model.

Data Preprocessing

We will conduct data analysis and model construction based on two datasets on NBA players' contracted salary [1] and performance statistics per game [2] in 2021-2022. The following steps are included in our data preparation:

- Two original datasets were inner joined by players and teams
- Kept only one record with most number of games played for each of players, given a player may transfer to other teams during the session and have multiple records.
- Removed 5 variables with missing values caused by division of other existing variables.
- Divided count variables (`field_goal`, `free_throw`, etc.) by variable `minute` to convert them to efficiency

The final cleaned dataset has 442 records and 24 variables, including 2 categorical variables, 21 numerical variables and 1 numeric response variable `salary`.

- `position` – Position of the player (5 categories)
- `age` – Player's age on February 1 of the season
- `team` – Team that the player belong to. (30 categories)
- `game` – Number of games played per minute
- `game_starting` – Number of games played as a starter per minute
- `minute` – Minutes played per game
- `field_goal` – Field goals per minute
- `fg_attempt` – Field goal attempts per minute
- `x3p` – 3-point field goals per minute

- `x3p_attempt` – 3-point field goal attempts per minute
- `x2p` – 2-point field goals per minute
- `x2p_attempt` – 2-point field goal attempts per minute
- `free_throw` – Free throws per minute
- `ft_attempt` – Free throw attempts per minute
- `offensive_rb` – Offensive rebounds per minute
- `defenssive_rb` – Defensive rebounds per minute
- `total_rb` – Total rebounds per minute
- `assistance` – Assists per minute
- `steal` – Steals per minute
- `block` – Blocks per minute
- `turnover` – Turnovers per minute
- `personal_foul` – Personal fouls per minute
- `point` – Points per minute
- `salary` – Salary of the player in million

Exploratory Analysis

Univariate Analysis

The following plots show distribution of each univariable. For categorical variables `team` and `position`, they are distributed quite evenly. There are 30 unique values in `team`, which may result in too many dummy variables in the model. Therefore, we may consider exclude `team` or cluster it into fewer classes in selected models.

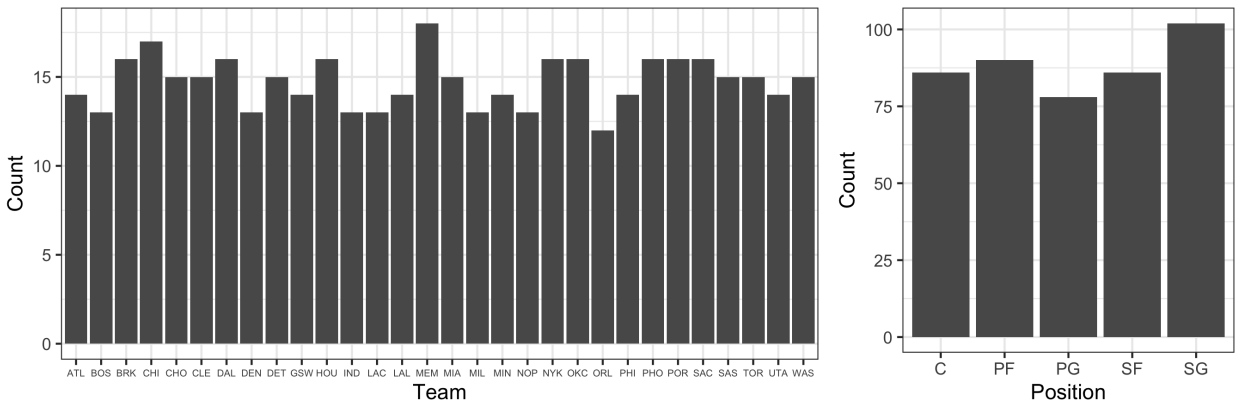


Figure 1: Histograms of categorical predictive variables

For numeric variables, some of them (`gs`, `ft`, `orb`,`blk`), including response `salary` are skewed, with some players have extremely high salary. Visualization for all variables are enclosed in Appendix A.

Correlation Analysis

From the correlation heat map, it is obvious that multicolinearity could be a problem, which we may consider using penalized models (ridge, lasso) or ensembled models (random forest, boosting, neural network) to fix. The feature maps demonstrated that some correlations are non-linear, which we may consider using GAM or MARS to address.

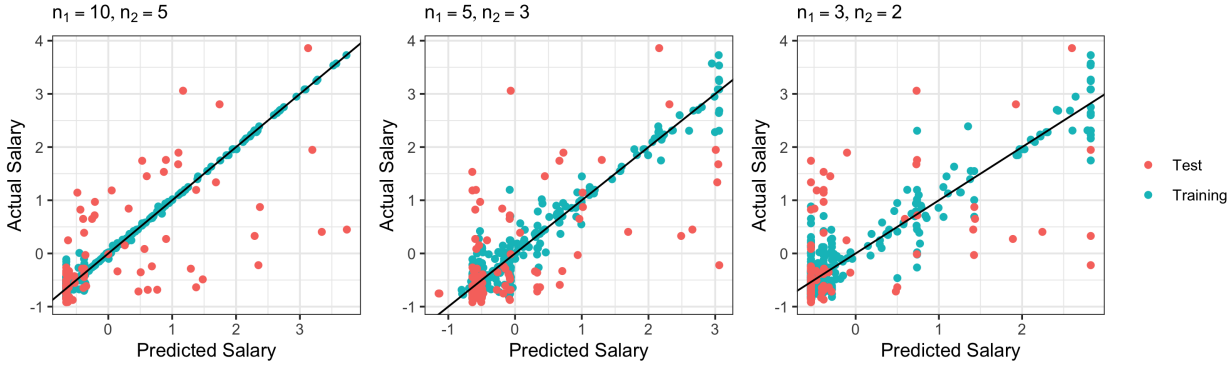
From categorical variable `position` and `team`, extremely high values and large variance in salary show in all positions and some teams.

Model Construction

Neural Network

Several 2 hidden-layer neural networks were built to fit the data. Despite trying different number of nodes and applying regularization techniques (L2 and dropout), the resulting models still have a noticeable overfitting problem. Given the size of the dataset is very small ($n = 442$), the performance of neural network is not as good as some traditional statistical models. It is more useful when the size of dataset is much larger with more variables.

As shown in the figure, as the number of nodes in the first and second hidden layers increases, neural networks can provide very accurate fittings of the training data, with much lower MSEs compared to other methods. However, the predictions are not satisfying when the models are applying to the test data.



Model Comparasion

Conclusion

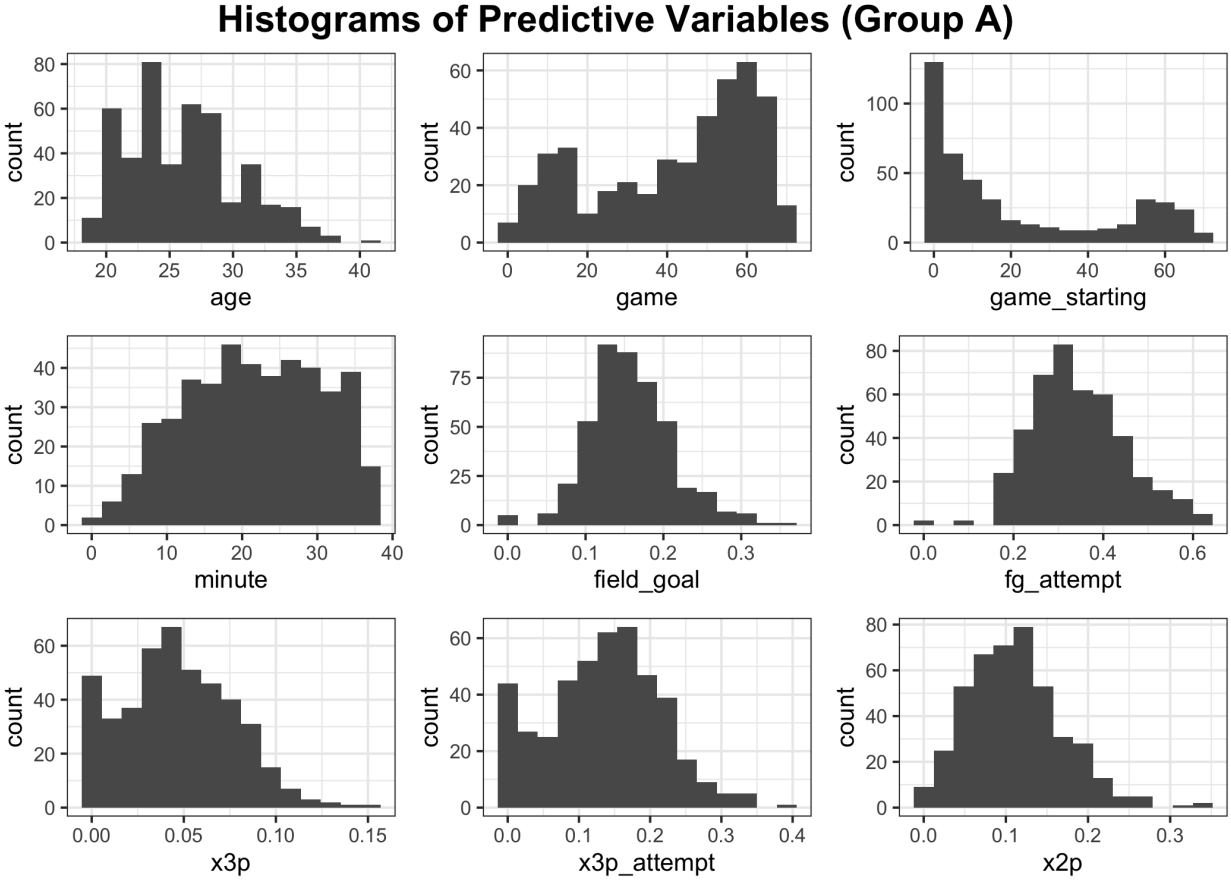
References

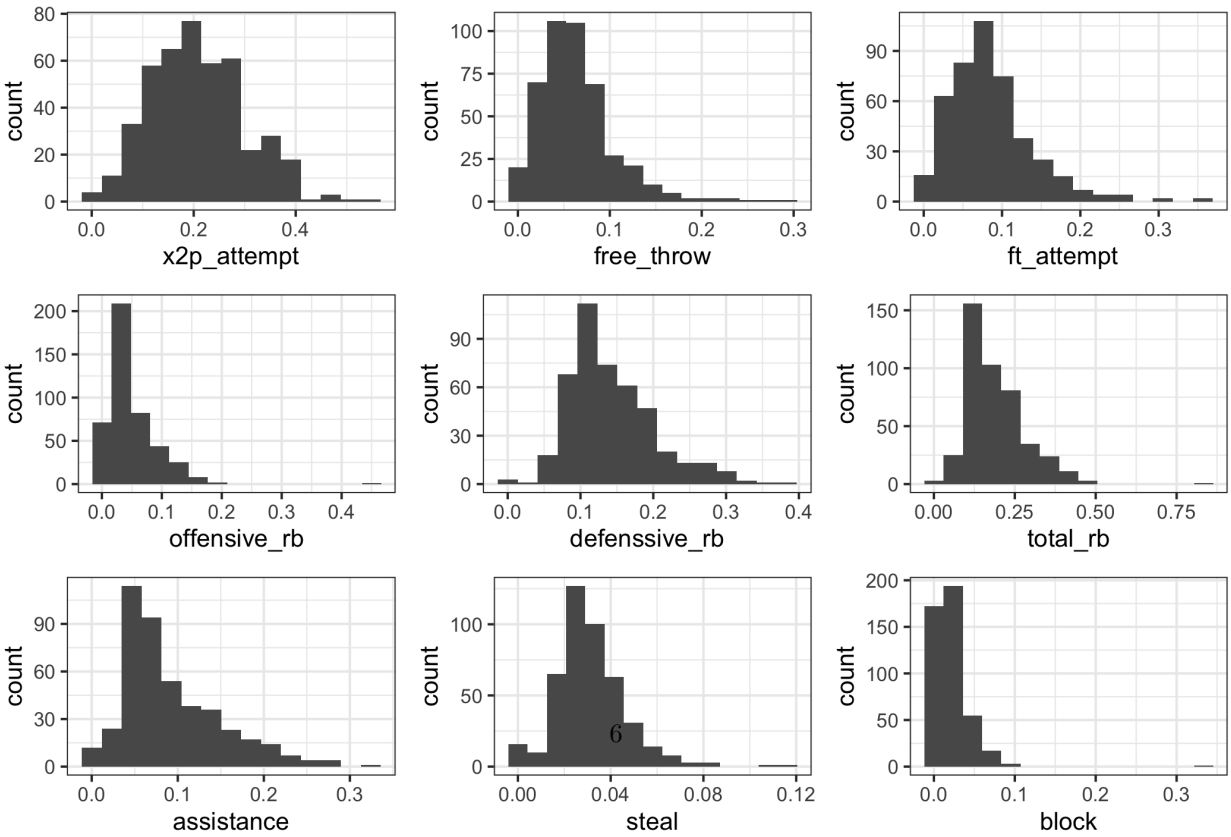
[1]https://www.basketball-reference.com/contracts/players.html

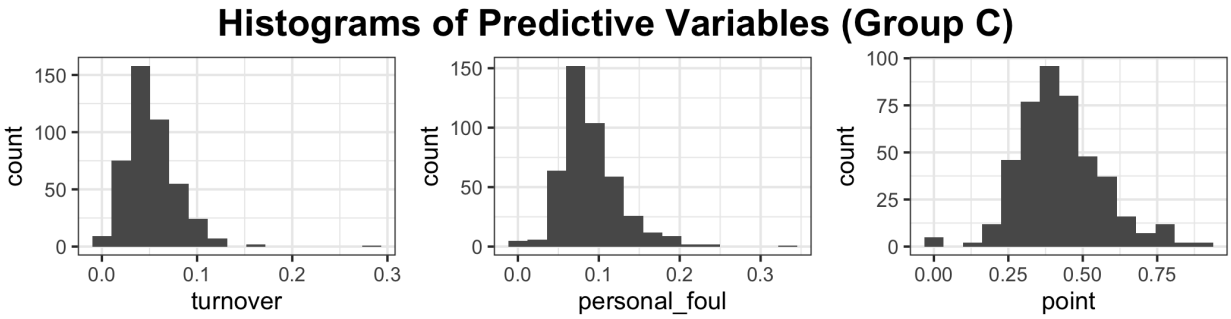
[2]https://www.basketball-reference.com/leagues/NBA_2022_per_game.html

Appendices

Appendix A - Numeric Variable Distribution







Appendix B - Numeric Variable Distribution

Appendix C - Neural Network Model