COMP 562 Final Project

Brianna Li, Srikar Chedalavada, Neha Jakkinpali, Anuttam Perumal, Anthony Wang, Griffin Wheeler

University of North Carolina at Chapel Hill
Department of Computer Science

Abstract

The NBA is a league that revolves around its players, and each organization's most important basketball decisions revolve around which players they select to play for their team as well as their respective salaries. In this report, we have sought to run Linear, Ridge, and LASSO regressions with k-fold cross validation on the player statistics that predict player salaries; this will prospective NBA organizations and General Managers with the data that determines which players are being under and over paid based on their statistics. We were able to complete this using the 2016 NBA player statistics- a dataset that can be used to extrapolate findings to NBA players in future years.

1 Introduction

1.1 Motivation

There are a variety of factors that may affect the outcome of an NBA game; ultimately, this comes down to the caliber of players in each franchise and their respective statistics. While some players may perform situationally and others may perform consistently from game to game, a player's salary can be used as a measure of his value to the franchise. The goal of this project was to determine which types of player statistics have the largest impact on a player's salary by using a variety of regression techniques, and then performing k-fold cross validation in order to provide the exact player statistics that have the largest bearing on a player's salary.

1.2 Data

The dataset used for this project was found on Kaggle, and its link can be found in the References section of this paper. It contains the player statistics of 342 unique athletes, and provides 36 different variables for each athlete corresponding to a type of statistic. The main statistics used were (not comprehensive):

- Two Point shots scored
- Free Throws scored
- Total Points
- Wins (Real-Plus Minus)
- Minutes Per-Game played

1.3 Plan

The first section will focus on the motivations of the data, what the data set is, and how it was compiled. The second section will further analyze and begin the pre-processing of the data. The third section will discuss findings from the linear regression, lasso regression, and ridge regression. The fourth section will use k-fold cross validation to improve and estimate future NBA salaries. In the fifth section, the conclusions drawn from the regression(s), validation, and final predictive model will be explored in addition to any future work that may be done on the project.

2 Pre-Processing the Data

2.1 Closer Look at Data

This dataset contains several statistics, but it is most probable that only a subset of these statistics is of interest when trying to predict a player's salary. The initial subset of statistics included was intentionally broad in order to be as inclusive of all factors as possible, and was then narrowed down to a final subset by identifying multicollinearity as seen in the correlation.

2.2 Underlying Assumptions and Limitations

When analyzing the data, it was assumed that salaries were influenced by a player's performance on the court; in other words, relatively high statistics indicated that a player's salary would be higher while the contrary would indicate a lower salary. Extraneous variables, such as those that may arise from location, were not taken into consideration (such as cost-of-living adjustments). It was also assumed that a player's salary was determined only based on their previous year's performance. One limitation that was not accounted for was the presence of one player impacting the outcome of another player. Additionally, any missing or undefined values were replaced with a zero. In general, the final predictive model only considers individual player statistics despite the additional features that may affect those statistics such as team chemistry and playing conditions.

2.3 Feature Engineering

From our data analysis, we were able to narrow down our selection by identifying multicollinearity, as seen in the correlation. P values greater than the absolute value of .7 are used to show multicollinearity. In this scenario, however, we acknowledge the real-world practicality of basketball statistics and decided that multicollinearity above 0.91 should be removed, as NBA player statistics overlap significantly. From these variables, we reasoned that reasonable and minimal- if any- multicollinearity existed. Alongside this, we made an effort to split our regression modeling processes into two parts: 5-feature engineering and all-feature engineering.

3 Regression

3.1 Linear Regression

We ran a linear regression on both the 5-feature and all-feature datasets. With player statistics we realized all the values were already previously standardized and did not require further modification. The variables we found to be statistically significant to the 95 percent confidence level are Age, Total Rebounds, Personal Fouls, and Points. The remaining variables were not statistically significant, although they still played a part in our regression outcome. Positive coefficients indicate that an increase in the statistic leads to an increase in player salary. While it is logical that personal fouls are a "negative" characteristic, particular attention must be drawn to the negative coefficients of 3P percent and 2P percent. The p values of both of these are higher than .05, but there is also a potential explanation: higher paid players take more shots and some of which are higher risk, therefore would not have extremely high 3P percent and 2P percent. The cross-validation accuracy test is 0.39.

3.2 Lasso Regression

We performed a lasso regression to the dataset as well, which eliminates less useful variables by setting their coefficients to 0. The cross-validation accuracy test is 0.41.

3.3 Ridge Regression

We also ran a ridge regression on the dataset to shrink the coefficients of variables that do not contribute much to the outcome. The cross-validation accuracy test is 0.42.

4 Modeling

4.1 5-Fold Cross Validation

In order to validate the information we had derived from the regressions, we used a 5-fold cross-validation for each individual regression done on each individual df. We found that after cross validating each of the regressions, the most accurate of the regressions came out to be from the linear regression model of the 5-feature df.

5 Summary

5.1 Best Models and Accuracy

The best models we were able to conclude from the regressions tests were the models created using the 5-feature df; specifically the linear regression. It's accuracy was 0.317.

5.2 Conclusion

The overall goal of the project was to determine if machine learning could be applied to determining which player statistics contribute the most towards predicting a player's salary in the NBA. The most accurate predictive model was concluded to be the linear regression, with an r-squared metric of 0.317. Overall, the determination of these statistics would allow any NBA franchise to manage their player's salary budgets more effectively.

5.3 Future Work and Lessons Learned

There were a few assumptions made before progressing through the feature engineering, regression, and validation steps in the project, and there is no concrete way of determining the degree to which these affected the final prediction model's outcome that lies within the scope of this project. One of the more significant limiting factors of this project was a limited amount of data available. Although statistics for every listed player in the 2017 season were used, the amount of data was limited to the 2017 season. The NBA is known for large discrepancies between the player statistics of more experienced athletes as opposed to those of newer athletes, and having a larger time range of data would allow for the implementation of additional time-based features, such as percent change of a certain (basketball) metric. A lack of data could also lead to inaccuracies in the fit of the model, particularly leading to problems with overfitting on the various types of regression. In addition, the metrics for the 5-fold cross validation were inaccurate for the same reasons listed above. The second limiting factor in this project was the multicollinearity between certain variables, which lies in the nature of the dataset. For example, "MP" and "MPG" are acronyms for minutes played and minutes played per game, which are similar in nature but may convolute the final model because of this.

5.4 Github Code

https://github.com/anuttamp/final562

5.5 References

Dataset: “NBA predicting player salaries” by Rodrigo
https://www.kaggle.com/rodrimc/nba-predicting-player-salaries/data