NBA Salary Regression Analysis

Jonathan Yun

Professor Claudio Delrieux

Capstone 4449

25 April 2021

NBA Salary Regression Analysis

When it comes to professional sports, the salaries of players have always been a topic of discussion among fans and analysts. The NBA holds no exception to this scrutiny, as players receive blockbuster salaries every year. In addition to the league’s revenue increasing year after year, it is no surprise that player salaries have also skyrocketed as well. However, it is important to understand the factors that contribute to a player’s salary in the NBA. In this essay, we will use regression analysis to explore the relationship between various player statistics and demographics in respect to their salary. By examining the data, we can gain a deeper understanding of the factors that impact NBA player salaries and how they have evolved over time.

The dataset we will be analyzing consists of two separated datasets: one containing NBA player statistics and the other containing their corresponding salaries. The NBA player dataset contains information such as player id, birth date, birthplace, career assists, career field goal percentage, career field goal three percentage, career free throw percentage, career goals, career points per match, career tries, career win shares, career effective goal percentage, college, draft pick, draft round, draft team, draft year, height, high school, name, position, shooting hand, and weight. The NBA salaries dataset contained information on the annual salaries of each player in the league. This data includes player id, annual salary, season, season end year, season start year and team. By combining these two datasets we can examine the relationship between various player statistics and their corresponding salaries. This will allow us to gain insights into the factors that contribute to a player’s salary in the NBA and how these factors have evolved over time. The analysis of these datasets can help us better understand the economics of the NBA.

After importing the data into jupyter notebook I saw that for a lot of the features in the players dataset were null values. The columns that had a significant number of null values were the draft columns, career effective field goal percentage, career field goal percentage from three-point line, birthplace, and high school. Since I wanted to predict the salaries based off their career, I decided to drop columns that did not pertain directly to their career. I dropped all columns that were related to the draft, the high school they attended, the college they attended, birthplace and birthdate. I then merged the players dataset to the salary dataset by their unique player ID value. After merging the two datasets, I noticed that for the position’s columns, a player could hold multiple positions for any given year. Therefore, I got rid of the positions columns and instead one hot encoded the positions column. When preparing the data, I decided to split the data into three eras: classic, transition and modern forms of the NBA. “The classic era of the NBA focused on the inside the arc shots, the transition era of the NBA evidenced the migration from the rim to the three-point line and then the modern era of the NBA which is characterized by better shot selection and replacing inefficient two-point shots by more rewarding three-point shots.” (Rocha) By splitting the data in this way, I was hoping to see a trend in salary that increased as the game became more popular.

After preparation of the data, I applied multiple regression models to the three sets of data to see which accurately predicted the salaries of NBA players. I applied Linear, Logistic, Ridge, Lasso, and Random Tree Regression to all three of the datasets and found that the Rand Tree Regression model was the most accurate of all the models. The R-squared value was in the 0.60 – 0.75 range while most of the other models failed to get above a 0.5 R-squared value. After noting that the Random Forrest Regressor was the best performing model, I decided to apply both a grid search and a random search cross validation to the Random Forrest Regressor. I then applied the best parameters that both hyperparameter tuning algorithms found to the model and both only slightly increased the R squared value. I then created another model using only the most important five features for each. The most important features determined by both cross-validation techniques for the classic and transition era was season start year, season end year, career points, career win share and career points per match. Oddly for the modern era the most important features were career win share, career points, career points per match, career goals, and season end year.

Works Cited

Rocha da Silva, João Vítor and Rodrigues, Paulo Canas. ‘The Three Eras of the NBA Regular Seasons: Historical Trend and Success Factors’. 1 Jan. 2021 : 263 – 275.