Roni Kaakaty

DSC530-T301

Databall

My final project topic was attempting to re-create the premise of the famous “Moneyball” movie. I set out to capture “counting stats” of baseball free agents from the time-period of 2015-2019 in order to see if the criteria of the past is still what teams look for when deciding to sign free agents for big money. Baseball purists had the belief that a sign of a good hitter was a batting average above .300. I wanted to see if batting average was still the gold standard that separated modestly paid free agents from ones that secured the most money that free agency period. When compiling the data set, I only factored in the players last 3 years before free agency because I wanted to avoid outliers. Whether the player was just getting into the league or just had a great contract year, using three years as the sample size would give me a better idea of the hitter’s actual ability and will help me avoid a larger variance in data. I also left off pitchers because I wanted an apples-to-apples comparison. Comparing pitcher salaries to position player salaries would skew the data. I found the 5 highest annual salary contracts given out by teams to position players each year from 2015-2019 and took each player’s 3 year mean for all the stats I compiled. Another factor that I excluded was players resigning with their previous teams. I did not want the possibility of a hometown discount impacting the data. A true free market was the only way to accurately capture the value that player had to teams.

My EDA found that while batting average is still an important factor, there are other counting stats that matter to teams more now. The average batting average from the top paid players in the years I compiled came out to .274, which resulted in my null hypothesis being rejected. I found the variable that had the greatest correlation with increased compensation was WAR, also known as wins above replacement. My linear regression model should a strong linear correlation between WAR and a player’s annual salary. One item that I found surprising was the lack of concern for a player’s defensive abilities. I assumed players with a higher number of defensive runs saved would have resulted in more defensive minded players ending up on the highest paid lists. However, a majority of the highest paid free agents in the time-period I compiled had a negative defensive runs saved number.

If I were to do the analysis differently, I believe I should have chosen to do pitchers instead of position players. With position players, there was not a drastic difference between the counting stats from players in the time frame I accumulated. I believe identifying advanced pitcher metrics such as spin rate, would have been a better barometer of how the game has changed. Starting pitchers used to go above 6-7 innings in a game regularly, however with this analytics shift, there are now openers and relievers starting games. Another thing I would change would be to identify more advanced metrics for position players. Due to the fact that there wasn’t a great variance in the stats that I captured, I believe that if I went more in depth, I could find some advanced metric that would have created that separation I needed to identify a significant trend. I assumed that the counting stats that I identified would be enough to give me that separation trend I was looking for, but I was wrong. I also assumed that annual salaries would be increasing from 2015 to 2019 just based on inflation, but that was incorrect as well. The salaries were all over the place and not dependent on the year.

The hardest part about this project was just trying to figure out the best tests to run for my data set. I did not know which analytical distribution would be the most appropriate for the data I was working with. It was also hard to factor in other variables such as positions of the players, whether their league has a designated hitter, or other items that would influence the annual salary.

While this was a very challenging project, I really like the foundation it helped me build and I’ve gained more confidence in attempting to do my own analysis to other sports analytics items that I come across.

References:

Agrawal, Y. (2019, March 07). Hypothesis testing in Machine learning using Python. Retrieved from <https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce>

Felix. (2019, December 10). Exploratory Data Analysis in Python from DataCamp. Retrieved from https://datascience103579984.wordpress.com/2019/12/10/exploratory-data-analysis-in-python-from-datacamp/2/

Lecture 03 - Data Science. (n.d.). Retrieved from https://sites.google.com/site/datascience14/lectures/lecture-03

MLB Stats, Scores, History, & Records. (n.d.). Retrieved from <https://www.baseball-reference.com/>

Positions: Glossary. (n.d.). Retrieved from http://m.mlb.com/glossary/positions

<https://www.youtube.com/watch?v=_kb2X85DkKk&ab_channel=KindsonTheTechPro>