Final Report for Capstone Project for Ryerson CKME136

Margaret Anderson Kilfoil

https://github.com/Deadlysmurf/Capstone

Introduction

Most sabermetric analysis has been focused on predicting expected player performance and the expected performance record of teams. As there is more focus on increasing the competitive balance in major league sports through salary caps and luxury taxes, teams need to spend more time analyzing their payrolls. This includes looking at the monetary value that a player contributes to a team objectively through analytics, rather than relying on perceived values.

My key research question is to define what factors are the most relevant in predicting current fielding player salaries.

Pitchers and catcher were eliminated from this analysis, because their metrics for success are substantially different from a fielding player. Post-season play was removed from the analysis, because it biases the analysis to the teams that have been more successful, hurt the abilty of the model to find the salaries of players who are with less successful teams.

Literature Review

In the past four months, I've been heavily reading literature and watching webinars posted on Data Science Central, which has been mainly focused on work place applications, although a great deal of the regression literature will carry over into this project.

I've also been watching other webinars, such as the BrightTalk Data Science summit series as part of workplace enrichment, which cover a wide range of practical applications of data science techniques.

Literature about Sabermetrics

- Adler, Joseph. Baseball Hacks. O'Reilly Media, 2006. Print.
- This book is meant to be a self-directed program for learning more about baseball, Perl, MySQL, and R. It provides lots of exercises about how to analyze baseball data. It's not very useful for this project, as the project goes beyond the exercises in the book, but it's a excellent introduction guide.
- Chang, Jason, and Joshua Zenilman. "A Study of Sabermetrics in Major League Baseball: The Impact of Moneyball on Free Agent Salaries." Honors in Management (2013). Washington University in St. Louis, 13 Apr. 2013. Web. http://olinblog.wustl.edu/wp-content/uploads/AStudyofSabermetricsinMajorLeagueBaseball.pdf.
    - An undergraduate management thesis about free agent salaries since the early 2000s, with a focus on economic theory. Interesting to see their approach, although there are some flaws in their methodology. (The salary data they are using appears to be incomplete, they sampled three years of data only, and they only took into account the previous year's performance of a player)

Literature about Python

Prior to this project, I have never worked in Python. I'm excited for the opportunity to add another language to my education, especially as it gets more popular in the data science community.

- McKinney, Wes. Python for Data Analysis. 1st ed. O'Reilly Media, 2012. Print.
    - McKinney is the primary author of Pandas for python and a globally recognized expert in python. This older publication is in python 2.7, while I am working in python 3.6, so it has been more valuable for concepts than an actual coding resource.

- Learn R, Python & Data Science Online. Computer software. Datacamp. Datacamp.com. Web.
    - Datacamp is a great online based learning software that combines videos and interactive lessons to teach python, R and data analysis. Just learning python, it has proved a fairly invaluable resource.

Literature about Exploratory Data Analysis

- Brillinger, David R. "Data Analysis, Exploratory." International Encyclopedia of Political Science. Web. <www.stat.berkeley.edu/~brill/Papers/EDASage.pdf>
    - This article summaries a variety of Exploratory Data Analysis techniques, mostly inspired by the work of Tukey, whose's textbook in 1977 popularized the phrase 'Exploratory Data Analysis'. This article focuses on the importance of looking for structure and patterns in the data before hypothesis testing the data.

- Tukey, John W. The Future of Data Analysis. Ann. Math. Statist. 33 (1962), no. 1, 1--67. doi:10.1214/aoms/1177704711. http://projecteuclid.org/euclid.aoms/1177704711.
    - This publication is old (written before the common use of computers for data analysis), but most of publication remains valid and important today. Tukey, who would go on to create the box plot in his 1977 book on data analysis, spends most of the article discussing what he sees as the future of data analysis as we move into the computer age. He talks about the importance of keeping an eye to the art, as well as the science of data analysis, that automation should be encouraged because it allows the sophisticated data analyst to spend more time with new areas of exploration, and the importance of flexibility in approaches to data analysis.

Literature about Regression

- "An Introduction to Machine Learning with Scikit-learn¶." Scikit-learn 0.18.1 Documentation. Scikit-learn 0.18.1. Web. http://scikit-learn.org/stable/tutorial/basic/tutorial.html.
    - Documentation and guides on one of the primary machine learning packages for python. Includes the primer on the code for the package, as well as concrete examples of the package being used on the iris and digits data sets.

Literature about Social Sequence Analysis

- Abbott, Andrew. 1995. “Sequence Analysis: New Methods for Old Ideas.” Annual Review of Sociology 21:93-113. Web.

    - An early paper about the application of sequence analysis, traditionally a methodology used only in the hard sciences, in the social sciences (examples: sociology and economics). There is some light discussion of techniques towards the end of the paper, although a lot of the methodology is outdated as computers have become more powerful.

- McVicar, D. and M. Anyadike-Danes (2002). "Predicting Successful and Unsuccessful Transitions from School to Work by Using Sequence Methods." Journal of the Royal Statistical Society 165:317-34. Web.

    - This review of predicting the success of transition from school to work in Northern Ireland has a strong methodology secion that discusses the merits of clustering analysis, social sequence analysis, using the output from a cluster analysis in a regression analysis, and the importance of data prep for the applicable methods.

Dataset

I used the Chadwick Bureau enhanced mirror of the Lahman database, recommended by Sean Lahman, as the Lahman database hasn't been updated for the season rollover and is missing salary information. The github mirror can be found here: https://github.com/chadwickbureau/baseballdatabank.

I've imported the .csv files in at the above github link, and have used the Appearances, Batting, Fielding, Master, and Salaries tables in this analysis.

The Salaries table will provide the dependent variable (salary) for analysis. It provides salaries by player, team and year.

The Master table provides attributes about the players, such as their date of birth, height, weight, and playing hand.

The Appearances, Batting, and Fielding tables provide playing statistics about the players.

Other statistics have been calculated from the available playing statistics, such as OPS (On-base plus slugging).

Approach

All steps of this analysis have been brought together at this Github link:
https://github.com/Deadlysmurf/Capstone/blob/master/Juypter%20Notebooks/Bring%20It%20All%20Together.ipynb

Please find the step diagram attached as Appendix A.

Step 0: Prepare Environment

Details:

- Import required Python packages
- Set the Alpha value for analysis
- Set Target Variable

Github Link: https://github.com/Deadlysmurf/Capstone/blob/master/Scripts/Packages.py

Step 1: Data Collection

Details:

- Download the Chadwick Bureau database so that a static copy of the data is maintained
- Import the data into python, specifically into Pandas dataframes

Github Link: https://github.com/Deadlysmurf/Capstone/blob/master/Scripts/Data_Import.py

Step 2: Preliminary Data Cleaning/Structuring

Details:

- Tie the salary information to the player information
- Calculate relevant player statistics, such as age of the player and number of seasons played
- Left join the master table to the player statistics tables tables
- Calculate additional statistics such as OPS
- Drop unnecessary fields

- Removed any infinite or NaN values

Github Link: https://github.com/Deadlysmurf/Capstone/blob/master/Scripts/Data_Cleaning.py

Step 3: Exploratory Data Analysis

Details:

- Explore relationships in the data, such as sub-questions to examine:
  - How have salaries changed year over year?
  - How do salaries differ between National and American leagues?
  - How do salaries differ between teams?
  - How does player nationality affect salaries?
- Conduct testing to determine if there are difference between the means of categorial variables.
  - Eliminate the variables that aren't statistically significant

Githib Link: https://github.com/Deadlysmurf/Capstone/blob/master/Scripts/Data_Testing/StdDev_Test.py

Githib Link: https://github.com/Deadlysmurf/Capstone/tree/master/Juypter%20Notebooks/Exploratory%20Data%20Analysis

Step 4: Secondary Data Cleaning

Details:

- Use K-means clustering on categorial variables with more than three categorial.
  - For example: This clusters teams into very high, high, mid, low and very low spending teams. This makes the linear regression model significantly more accurate than trying to include all the teams individually
- Dummy categorial variables
  - Creates dummy variables, allowing the categorial variables to be included in the analysis

Githib Link: https://github.com/Deadlysmurf/Capstone/blob/master/Scripts/Dummy_Strings.py

Step 5: Correlation Analysis

Details:

- Calculate the Spearman correlation of all variables and salary
  - Spearman is required because all of the independent variables are not continuous
- Remove any variables that are not statistically correlated to salary.
- Create heatmaps of the Rho and P-Values to see if there are cross-correlated independent variables

Githib Link: https://github.com/Deadlysmurf/Capstone/blob/master/Scripts/Correlation.py

Step 6: Regression Analysis

Details:

- Split the data set into Test and Train data sets, using 50% in each sample.
  - Also removes any NaN values

- Apply a linear regression model to the data.

Githib Link: https://github.com/Deadlysmurf/Capstone/blob/master/Scripts/SplitToTrainTest.py Githib Link: https://github.com/Deadlysmurf/Capstone/blob/master/Scripts/LinReg.py

Results

With the variables included in the analysis, the regression predicts approximately 70% of the variance within the salaries of baseball players.

Unexpectedly, the most predictive variable is the birth country of the individual player. Depending on the home country of the player, their salary can be expected to be between $4M more, or $6M less. More analysis should be conducted to determine if the is a result of stronger baseball training in certain countries that is driving this, if there are regional characteristics that are impacting the quality of players from a country, or if there are underlying biases that are impacting how much they are paying players.

The analysis shows that the is clear difference between the have teams, and the have not teams, in terms of their salaries. Depending on the grouping of team, the top and bottom team clusters have a $4M range between their coefficients. This is factor that the salary cap is meant to address, so that they teams than have more money to spend on players are able to buy talent that the other teams can't afford. This shows that there is still significant variance between the teams ability or willingness to pay.

The effect of a player having pinch run or pinch hit was unexpected. Pinch hitting or running in a season adjusted the expected player salary down by approximately $2M. This is likely due to the trend of using young, rookie players as pinch runners, because they don't want to risk injury to their sluggers.

The most valuable playing statistic is the on base percentage plus slugging (OPS). The better the OPS, the more money that a player can expect to make. This makes sense, as the goal of the game is to get on base and score runs, and this factor is the primary indicator of a player's ability to do this.

Right-handed throwing is more valuable according to the regression coefficients than left-handed throwing, although that could be bias by the number of players that are right-handed.

The model also shows that the batters who can bat from both side of the plate have a higher predicted salary than those players with the same metics who are only single sided batters. left-handed batters have a positive impact on salary, while right-handed batters have a negative impact on the player salary. Being a switch hitter is a valuable commodity as managers match the hitting direction of the batter to the throwing hand of the opposing pitcher. Left-handed batters are less common than right-handed batters, therefore appear to command a premium on the market.

The least impactful variable is the height of the player, which has a minimal impact on the model. As a fielder, taller is generally better, but the impact on the salary of the player is minimal.

Conclusions

There are other factors that should be explored, including a time analysis of the career of baseball players and an analysis of why certain countries produce higher paid players than others, this analysis provides a strong base analysis.

This analysis shows that there are factors that the player cannot impact, such as their country of birth, age, and number of years in the major league.

There are also factors that the player can impact that will increase their salary. For example, targeting contracts with specific teams and increasing their OPS.