Abstract 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 player salaries.

I plan on exploring the relationships in the data using exploratory data analysis techniques, to explore the data, refine the data set, and to establish further hypothesis for testing.

After exploring the data, regression analysis will be conducted, both using time series where the most recent year is used to test the fit of the model, and standard 10-fold analysis. Some factors will also be analyzed using

social sequence analysis, to look at the causal time-linear relationships between player statistics and salary bumps.

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 pratical 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, 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.

    - 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'm going to use the enhanced mirror of the Lahman database, which has been recommended by Sean Lahman of the Lahman Database, 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.

The .csv files in at the above github link that I'm planning on using for this analysis include: AllStarFull, Appearances, AwardsPlayers, Batting, Fielding, FieldingOF, Master, Pitching, Salaries, and TeamsFranchises.

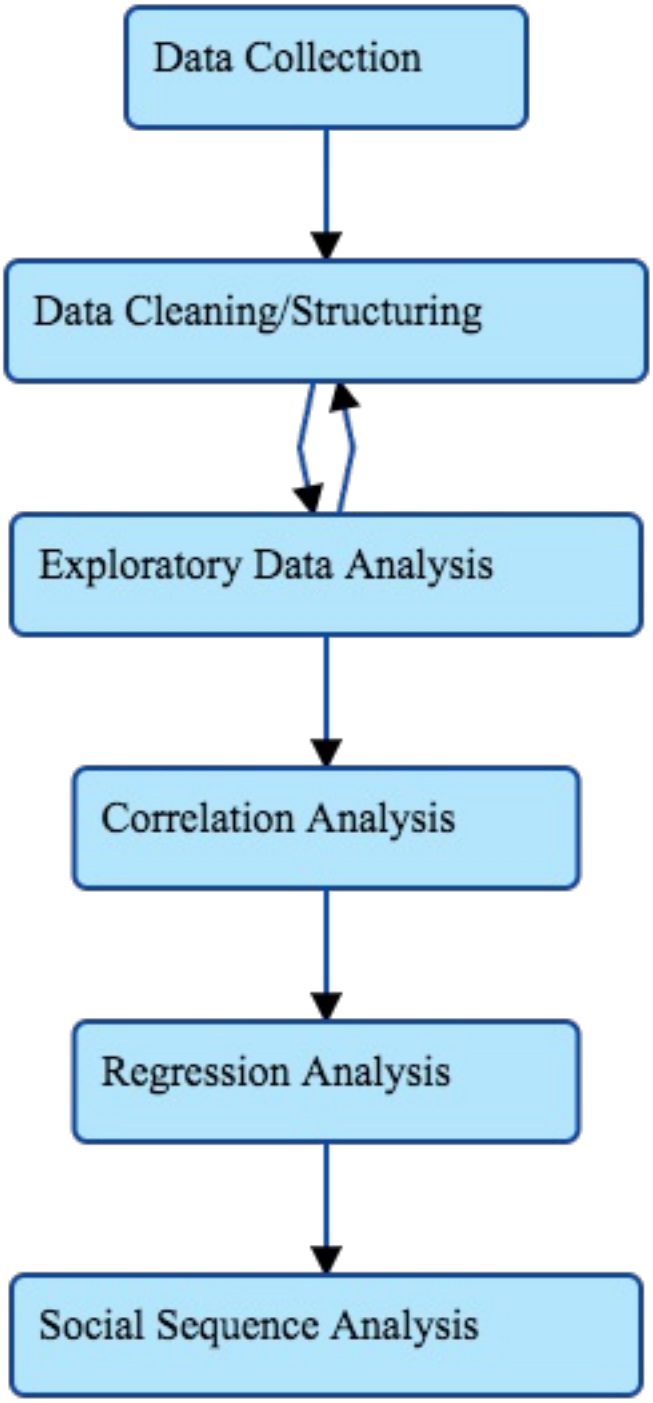
The AllStarFull, Appearances, Awards, Batting, Fielding, FieldingOF, Pitching tables provide attributes about the players.

I will be using games played in the post-season, games played though the regular season, positions played, batting average and fielding averages for my analysis. I will also be using attributes from these tables to calculate other sabermetric values such as WAR (Wins Above Replacement) ratings.

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

The Master and TeamsFranchises tables are both reference tables, that provide full text names for the player and team keys, as well as additional player information such as date of birth.

Approach



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

Step 2: 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
- Iterate over the fielding table to find the position(s) played by the player in a given season
- Add player pre-calculated statistics such as batting average to the master analysis table
- Calculate additional statistics such as WAR ratings and slugging ratings

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?
  - How does the number of positions a player plays affect salary?

Step 4: Correlation Analysis

Details:

- Using pre-determined factors, such as batting average, and any factors that prove relevant during the exploratory data analysis, check the correlation between the variables to establish that all the factors are independent before proceeding with regression analysis.

Step 5: Regression Analysis

Details:

- Conduct regression analysis using the independent factors identified in the exploratory and correlation analysis.
- Hold the most recent year out of the regression to test the fit of the model
- Test the fit of the model using 10-fold analysis.

Step 6: Social Sequence Analysis

Details:

- Examine the time ordered relationships between things like major league debut, post-season appearances, and increases in salary.