Abstract for Capstone Project for Ryerson CKME136

Margaret Anderson Kilfoil

https://github.com/Deadlysmurf/Capstone

(a) Problem and Theme

Most sabermetric analysis has been focused on predicting player performance and the performance record of any given team. 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. more analysis is required verify that players are being paid commensurate with their performance.

(b) Research Questions

How many years back are salaries comparable?

What factors are the most relevant in predicting modern salary?

Are there factors that contribute to outliers? For example: Does the number of years with a team explain a player accepting a lower salary than the regression would suggest they would command?

(c) Data Sources

There are three main data sources for sabermetric data:

- http://www.baseballprospectus.com/
- http://www.retrosheet.org/
- http://www.seanlahman.com/baseball-archive/statistics/

I'm going to use the enhanced mirror of the Lahman database recommend on the Lahman Database website for the majority of my work, as the Lahman database hasn't been updated for the season rollover. The github mirror can be found here:

- https://github.com/chadwickbureau/baseballdatabank

(d) Expected Techniques and Tools

The primary tool expected to be used for this analysis is Python 3.5.2, with the packages pandas, matplotlib (2.0), numpy and weka.

The expected techniques are to use exploratory data analysis to analyze how far back salaries appear to be comparable. A simple k-means clustering may be performed to validate if there are any groups that vary significantly enough that they should be eliminated from the general analysis. Then 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.