Capstone Project 1: Baseball and Linear Regression
By Scott W. Lew

Capstone Project 1 : Project Proposal

Predicting Baseball Team Wins In A Season

What is the problem you want to solve? The goal of this project is to predict the number of games a baseball team will win based on some key metrics of team performance such as batting and pitching statistics.This project assumes that these areas have a cause and effect on winning. For instance, a metric that measures pitching performance that has a positive correlation with wins is assumed to have causation. A baseball teams performance in terms of games won will be predicted using a regression analysis model which uses both pitching and batting statistics. Statistical analysis will determine which metrics have the most predictive power for team wins.

Who is your client? It would be useful for managers and coaches of baseball teams to have an understanding of which areas of team performance have a correlation with winning. As mentioned before, metrics that can predict wins are assumed to have causation for wins. These metrics would indicate where a team needs to improve performance to increase the number of games won by the team.

What Data: There is a vast collection of baseball batting and pitching statistics that are available from more than 100 years of professional baseball seasons. In this project, the batting statistics used consist of 28 features while the pitching statistics used consist of 35 features. For this project, the data from 1876-2018 was collected from the website www.baseball-reference.com using the Beautiful Soup program. The data was scraped and then converted into comma separated variable files (csv) for later use. For subsequent analysis, the data is in tabular form of a Pandas dataframe structure.

Problem Solving Methodology: Batting and pitching statistics will be studied for correlation with team performance in terms of games won, wins. Different regression analysis models will be created using different combinations of batting and pitching statistics and evaluated for their predictive power and accuracy.

Deliverables:

Deliverables will include documentation in the form of Google Docs and pdf files, code in the form of Jupyter Notebooks, and a slide presentation.