Homework 1

Dhairav Chhatbar, Mael Illien, Salma Elshahawy

08/31/2020

Contents

Prepared for:

Dr. Nasrin Khansari
City University of New York, School of Professional Studies - Data 621

Prepared by:

Dhairav Chhatbar
Mael Illien
Salma Elshahawy

1 Introduction

The ability to analyize and predict performance of a professional baseball team using many dimensions is critical to competitive success for our organization. Therefore, we have analyzed the records of numerous professional baseball team from the years 1871 to 2006. Our hope is that the following report and the resulting predictive models will better inform the organization and assist in making data driven decisions moving forward.

"The goal of a baseball team is to win more games than any other team. Since one team has very little control over the number of games other teams win, the goal is essentially to win as many games as possible. Therefore, it is of interest to measure the player's contribution to the team's wins." Grabiner, B. D. [1] While we do not have the variables at the player's individual contribution level, we do have the entire teams contributions as an aggregate and will analyze that information.

2 Statement of the Problem

The purpose of this report is to determine the batting, baserun, pitching, and fielding effects on a baseball team's ability to win.

[1] (Grabiner, B. D. (n.d.). The Sabermetric Manifesto. Retrieved September 10, 2016 from http://seanlahman.com/baseball-archive/sabermetrics/sabermetric-manifesto/)

3 Data Exploration

Note that each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season. The following Table 1 - Descriptive Statistics provides the detailed descriptive statistics regarding our variable of interest - Number of Wins and our possible explanatory variables.

We noted that several variables were missing a nontrivial amount of observations and these variables are Strikeouts by batters, Stolen Bases, Caught stealing, Batters hit by pitch (get a free base), Strikeouts by pitcher, and Double plays. So we will need to address the missing values for further analysis.

Histograms of all of the variables have been plotted below so that the distribution of the data can be visualized. In the distribution for the number of walks allowed, only two bars exist due to the excessive number of outliers.

3.1 Imputing Missing Values

In order to address the missing values in our variables we used a nonparametric imputation method (Random Forest) to impute missing values. Several variables have a significant amount of skew, which include the number of base hits by batters and the number of walks allowed. Correspondingly, these two variables had a skew of 1.57 and 6.74 respectively. Therefore, we chose a nonparametric method due to several variables having significant skew and having a non-normal distribution.

3.2 Correlation Matrix

After competing the imputation, we can implement a correlation matrix to better understand the correlation between variables in the data set. The below matrix is the results and as expected, Number of Wins appears to be most correlated to Base Hits by batters (1B,2B,3B,HR).

4 Data Preparation

4.1 Outliers

4.2 Box Cox Transformation

5 Models comparasion

6 Selected Model

7 Prediction on Evaluation Data

8 Appendix A

8.1 R source code