11.3 Final Project Step 3

Katie Adams

08/05/2021

Click here to view R Markdown file

Final Project Step 3

You are now on to the final phase of your research paper. While this step does not require you build a model, you are welcome to do so if you feel you have the time. Instead, you need to make a recommendation for the approach you would take and what the remaining steps would be using the information you have learned in this course to take this project from simply being an analysis exercise to proposed implementation of a solution.

R Markdown Setup: Libraries and CSV Importing

```
library(tinytex) ## tinytex for PDF R Markdown
library(formatR) ## formatR to wrap the text in the PDF
library(tidyverse)
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80), tidy=TRUE)
```

```
# Load the combined 2016-2020 MLB Team Standard Batting dataframes to summarize
# the past 5 years in 1 dataframe
df_bat_5years <- read.csv("C:/Users/kadams/OneDrive - Suncor Energy Inc/DSC Masters Program/GitHub/DSC5
    header = TRUE, fileEncoding = "UTF-8-BOM")

# Changed column names to clarify what they are
df_bat_5years <- df_bat_5years %>%
    rename(Team = Tm, Wins = W, AvgNumPlayers = X.Bat, AvgBatAge = BatAge)

# Used mutate to add new variable Percentage Won which is total Wins/Games
df_bat_5years <- df_bat_5years %>%
    mutate(PercWon = Wins/G)

# Changed categorical columns to numeric, and made new dataframe to use with
# Correlation Matrix and Heatmap
df_bat_5years_2 <- df_bat_5years %>%
    sapply(unclass)
```

Introduction

Overall, write a coherent narrative that tells a story with the data as you complete this section.

1. Major League Baseball (MLB) has a problem: they spend too much money on players because of a flawed assumptions rather than hard data. MLB Teams buy (i.e. offering contracts) players who are perceived "better" (i.e. players like Johnny Damon) and assume that this will buy the team more wins. In actuality and according to the MLB Team Standing data, buying runs (i.e. exclusively looking at the players who have historically made a high number of runs in comparison to other plays) will buy the team wins. To understand what MLB Team Standing stats are correlated with the number of team wins, a correlation matrix was created, and it identified that the number of Team wins was postively correlated with the following MLB Team Standing varibles: R/G (Runs Scored Per Game), PA (Plate Apperances), R (Runs Scored/Allowed), 2B (Doubles), HR (Home Runs), RBI (Runs Batted In), BB (Bases on Balls/Walks), OBP (On Base Percentage), SLG (Slugging Percentage), OPS (On Base+Slugging Percentage), OPS+ (On Base+Slugging Percentage adjusted to player's ballpark), and TB (Total Bases). These positive correlations mean that as these variables go up or down, Team wins goes go and down. These highly correlated MLB Team Standing variables were then added to the linear regression model (see below) in relation to the number of Team wins, where it was found in the summary table that the variable BB (Bases on Balls/Walks) is statistically significant, but OPS has a larger positive effect on the number of Wins, where SLG and OBP have a larger negative effect on number of wins.

Top Correlated Variables to Number of Wins (from 10.3 Final Research Project - Step 2)

- `R.G - Wins  0.823618`
- `PA - Wins  0.783488`
- `R - Wins  0.817370`
- `X2B - Wins  0.604298`
- `RBI - Wins  0.808313`
- `BB - Wins  0.864829`
- `OBP - Wins  0.861863`
- `SLG - Wins  0.708363`
- `OPS - Wins  0.800968`
- `OPS. - Wins  0.684706`
- `TB - Wins  0.697107`
- `SF - Wins  0.619633`

Linear Regression Model of the 5 Year MLB Team Standing Dataset

```
# Fit Linear Regression model to the data set that predicts number of team wins

# Used the variables with the highest correlation to Wins in the correlation
# matrix
rm_bat_5years <- lm(Wins ~ R.G + PA + R + Doubles + RBI + BB + OBP + SLG + OPS +
    OPS. + TB + SF, data = df_bat_5years)
rm_bat_5years
```

```
##
## Call:
## lm(formula = Wins ~ R.G + PA + R + Doubles + RBI + BB + OBP +
##     SLG + OPS + OPS. + TB + SF, data = df_bat_5years)
##
## Coefficients:
## (Intercept)          R.G           PA            R      Doubles          RBI
##  -2.911e+02    7.969e-01   -4.600e-03    4.896e-01    4.497e-02   -5.106e-01
##          BB          OBP          SLG          OPS         OPS.           TB
##   8.537e-02   -9.556e+02   -1.751e+03    1.802e+03    2.731e-01   -6.623e-04
##          SF
##   4.297e-01
```

```
summary(rm_bat_5years)
```

```
##
## Call:
## lm(formula = Wins ~ R.G + PA + R + Doubles + RBI + BB + OBP +
##     SLG + OPS + OPS. + TB + SF, data = df_bat_5years)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.477  -8.652   2.762   9.665  18.894
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.911e+02  7.606e+02  -0.383   0.7067
## R.G          7.969e-01  2.456e+01   0.032   0.9745
## PA          -4.600e-03  3.566e-02  -0.129   0.8989
## R            4.896e-01  3.017e-01   1.623   0.1231
## Doubles      4.497e-02  6.810e-02   0.660   0.5179
## RBI         -5.106e-01  2.540e-01  -2.010   0.0605 .
## BB           8.537e-02  3.952e-02   2.160   0.0453 *
## OBP         -9.556e+02  1.572e+04  -0.061   0.9522
## SLG         -1.751e+03  1.575e+04  -0.111   0.9128
## OPS          1.802e+03  1.572e+04   0.115   0.9101
## OPS.         2.731e-01  1.761e-01   1.551   0.1394
## TB          -6.623e-04  7.861e-02  -0.008   0.9934
## SF           4.297e-01  2.426e-01   1.771   0.0944 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.2 on 17 degrees of freedom
## Multiple R-squared:  0.9101, Adjusted R-squared:  0.8467
## F-statistic: 14.34 on 12 and 17 DF,  p-value: 1.356e-06
```

```
# Split the data into training and test set
set.seed(123)
training.samples.bat_5years <- sort(sample(nrow(df_bat_5years), nrow(df_bat_5years) *
    0.8))
train.data.bat_5years <- df_bat_5years[training.samples.bat_5years, ]
test.data.bat_5years <- df_bat_5years[-training.samples.bat_5years, ]
```

```r
library(MASS)
# Fit LDA - Included highly correlated variables from correlation matrix
fit <- lda(Wins ~ R.G + PA + R + Doubles + RBI + BB + OBP + SLG + OPS + OPS. + TB +
    SF, data = df_bat_5years)

# Make predictions on the test data
predictions.bat_5years <- predict(fit, test.data.bat_5years)

accuracy <- mean(predictions.bat_5years$posterior[, 2])
accuracy
```

```
## [1] 0.03375268
```

```r
accuracy_false <- mean(predictions.bat_5years$posterior[, 1])
```

The problem statement you addressed: Summarize the problem statement you addressed.

1. Major League Baseball teams have a flawed assumption that buying (i.e. offering contracts) players who are perceived "better" (i.e. players like Johnny Damon) will buy the team more wins. When in actuality, buying runs (i.e. exclusively looking at the players who have historically made a high number of runs in comparison to other plays) will buy the team wins.

How you addressed this problem statement: Summarize how you addressed this problem statement (the data used and the methodology employed, including a recommendation for a model that could be implemented).

1. This research paper addressed this problem by analyzing the data using correlation and linear regression. From the 2016-2020 MLB Team Standings dataset, a correlation matrix was created to determine which variables were postively correlated with number of wins. From that, the top positively correlated variables were inputted into a linear regression model to calculate the significance codes for each variable, and determine which dataset predictors are significant in the linear regression model (statistically significant predictors). This addresses the problem of what MLB Team Standing variable predictors are statistically significant in relation to number of wins, and which variables has a greater positive effect on number of wins.

Analysis

Summarize the interesting insights that your analysis provided.

1. Positive Correlation: It was found that number of Team wins is highly and postively correlated to the following variables in the 2016-2020 MLB Team Standings dataset: R/G (Runs Scored Per Game), PA (Plate Apperances), R (Runs Scored/Allowed), 2B (Doubles), HR (Home Runs), RBI (Runs Battled In), BB (Bases on Balls/Walks), OBP (On Base Percentage), SLG (Slugging Percentage), OPS (On Base+Slugging Percentage), OPS+ (On Base+Slugging Percentage adjusted to player's ballpark), and TB (Total Bases).
2. Negative Correlation: It was also found the number of Team wins is not highly correlated to AvgNumPlayers (Average number of players used in games), 3B (Triples), CS (Caught Steeling), or SO (Strikeouts).

3. Statistically Significance: Within the linear regression model, BB (Bases on Balls/Walks) was found to be statistically significant. In looking at the estimates of the model, OPS has a larger positive effect on the number of Wins, where SLG and OBP have a larger negative effect on number of wins.

Implications

Summarize the implications to the consumer (target audience) of your analysis.

1. The implications of these findings to the MLB team scouts looking to recruit new baseball players is that they should look generally at the highly correlated variables of the model: R/G (Runs Scored Per Game), PA (Plate Apperances), R (Runs Scored/Allowed), 2B (Doubles), HR (Home Runs), RBI (Runs Batted In), BB (Bases on Balls/Walks), OBP (On Base Percentage), SLG (Slugging Percentage), OPS (On Base+Slugging Percentage), OPS+ (On Base+Slugging Percentage adjusted to player's ballpark), and TB (Total Bases). Also, they should highly focus of BB (Bases on Balls/Walks) as it is highly statistically significant within the linear regression model.
2. For the current 2021 MLB season, we can expect that the variable OPS (On Base+Slugging Percentage) will have a larger positive effect on the number of Team wins, where SLG and OBP have a larger negative effect on number of wins. In looking at the current MLB Team Standing data for 2021 (in which the MLB season is halfway done), we can expect that MLB teams the Boston Red Sox, Toronto Blue Jays, and Houston Astros have the highest probability of winning more games as they have the highest team OPS.

```
### Load the 2021 MLB Team Standard Batting dataframe
df_bat_2021 <- read.csv("C:/Users/kadams/OneDrive - Suncor Energy Inc/DSC Masters Program/GitHub/DSC520/
    header = TRUE, fileEncoding = "UTF-8-BOM")
order(df_bat_2021$OPS)
```

```
##  [1] 30 29 28 27 26 25 24 23 22 21 19 20 17 18 16 15 14 13 12 10 11  9  8  6  7
## [26]  5  3  4  2  1
```

```
head(df_bat_2021)
```

```
##                  TEAM LEAGUE  G   OPS   AB   R   H X2B X3B HR RBI  BB  SO SB CS
## 1      Boston Red Sox     AL 51 0.767 1723 264 447 110   4 68 243 140 449 22  5
## 2   Toronto Blue Jays     AL 50 0.760 1711 252 440  75   4 75 238 150 421 28  6
## 3      Houston Astros     AL 50 0.759 1732 258 462  94   5 57 244 164 345 12  6
## 4      Atlanta Braves     NL 49 0.759 1610 241 380  78   7 80 234 168 452 14  5
## 5 Los Angeles Dodgers     NL 51 0.749 1727 263 419  77  11 62 252 228 466 16  5
## 6     Cincinnati Reds     NL 49 0.748 1667 238 413  79   4 70 225 158 438 10 11
##     AVG   OBP   SLG
## 1 0.259 0.321 0.446
## 2 0.257 0.323 0.437
## 3 0.267 0.334 0.426
## 4 0.236 0.316 0.442
## 5 0.243 0.341 0.408
## 6 0.248 0.323 0.426
```

Limitations

Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

1. Because of the data outliers and assuming a normal distribution, there could be a wide distribution so scaling each variable to the same basic scale which could be done by converting each variable to a z-scale.

Concluding Remarks

1. In closing, there are many MLB Team stats (variables) that are tracked year after year in Major League Baseball (MLB). But, when it comes to strategically choosing the right players to improve the teams' chances of winning games, data analysis is the true indicator. Traditionally, baseball player "on base" stats have been largely ignored by MLB player recruiters, creating the problem of MLB teams buying perceived "better" players to get more wins. This research paper used MLB baseball stats and data exclusively to evaluate, analyze, and predict which team(s) will win the most games in the 2021 season. This directly addressed the problem statement by providing insights for improved MLB Team player selection to obtain future game wins. Specifically, the insights provided in this reseach paper are two fold: first, MLB player scouts should focus primarily on players' BB (Bases on Balls/Walks) stat when recruiting as it is highly statistically significant in relation to the number of wins within the linear model. Second, in the current 2021 MLB season, we can expect that three MLB teams have a higher probability of winning more games (including the final World Series game to win the Commissioner's Trophy in October) due to their high OPS stats which has larger positive effect on the number of Team wins: the Boston Red Sox, Toronto Blue Jays, and Houston Astros.