20 Years Later?

Ethan Duncan

6/11/2021

Abstract

Logistic regression model to estimate an MLB team's likelihood of making the playoffs based on the depth of their payroll, the strength of their performance on the road, and strength of performance against teams with a winning record. Utilizes data from Baseball Reference, Baseball Prospectus, MLB.com (official stats), and Spotrac. Estimations made for Seattle Mariners and Boston Red Sox based on current season stats up until date of writing.

Background

As a long time Mariners fan who has been waiting now 19 years for the team to make the playoffs, I continue to wonder, will we be able to do it 20 years later? While I will never give up on the Mariners, it is hard to watch them have the longest playoff drought among the entire MLB. As I continue to anxiously watch the Mariners try to maintain a win % of .500 (win as many games as they lose), I decided to find out how much hope I should really hold for my hometown team to finally make the playoffs this season.

Moneyball is a popular film from 2011 that highlights the true story of Billy Beane's attempt to create a playoff team out of the 2002 Oakland A's using the small bank account they had. Unfortunately the MLB is the only U.S. professional sport that does not have a salary cap which puts small market teams (Seattle, Oakland, Baltimore, etc.) at a heavy disadvantage. Thus, I wanted to understand how much of an effect (if any) that a team's salary had on their ability to make the playoffs, given the success of certain underdogs such as Oakland and Tampa Bay who both have had similar payrolls to the Mariners but have made the playoffs when we have not.

Since making the playoffs is a discrete outcome, I utilized classification to best analyze MLB teams ability to make the playoffs. I included a vast amount of features in order to ensure I had the best approach to properly classifying a team's ability by thoroughly considering all potential factors.

Data Aggregation

Fortunately enough, the MLB has made a great effort over recent decades to make as much data accessible to the public as possible. While the MLB - Official Stats[3] has a great amonut of data points for each team and player, alternative sites (Baseball Reference, Baseball Savant, FanGraphs, etc.) quickly grew more popular since they were able to gather and provide additional data points and overall more information (whether descriptive or unhelpful). Salary data is one of the pieces of information that is not readily available on most baseball data sites so it was necessary to do a bit of data aggregation from multiple sources to get all of the information I wanted in one place.

I personally prefer to use Baseball Reference[1] for all of my primary baseball data needs so I used their data to build the structure of my final data frame. Unfortunately the salary data provided by Baseball Reference is not as accurate as I hoped so I had to find another site for this information. Spotrac[4] is arguably the most accurate source for MLB payroll/salary data, however they require a paid subscription to access data from more than 3 years ago so we had to find the most accurate yet free site. Thankfully Baseball Prospectus[2] provides payroll/salary data back through 1979 free of charge, and after a quick validation using 2019, 2020, and 2021 data from Spotrac[4], their salary data was far more accurate than Baseball Reference[1].

Using Baseball Reference[1] and Baseball Prospectus[2], I created a master data frame containing data from each of the 30 teams over the 5 most recent seasons (2015-2019). 2020 data was not included as the season length was greatly reduced and playoff rule structure was different. The following custom variables were calculated using data provided:

> `made-playoffs`: 1 if team made playoffs; 0 if not
>
> `road-win-pct`: team's winning % for away games
>
> `vsLHP-win-pct`: team's winning % when facing left-handed starting pitcher
>
> `vsWO-win-pct`: team's winning % when facing winning/strong teams
>
> `payroll-M`: team's total payroll (in millions of $USD)

The data aggregation process can be evaluated by looking at the Excel workbook, MATH 447 - Project Data Agg.xlsx[5]. The output of this process and data frame used for analysis/modeling is MLB Team Stats 2015-2019.csv[6].

Methods

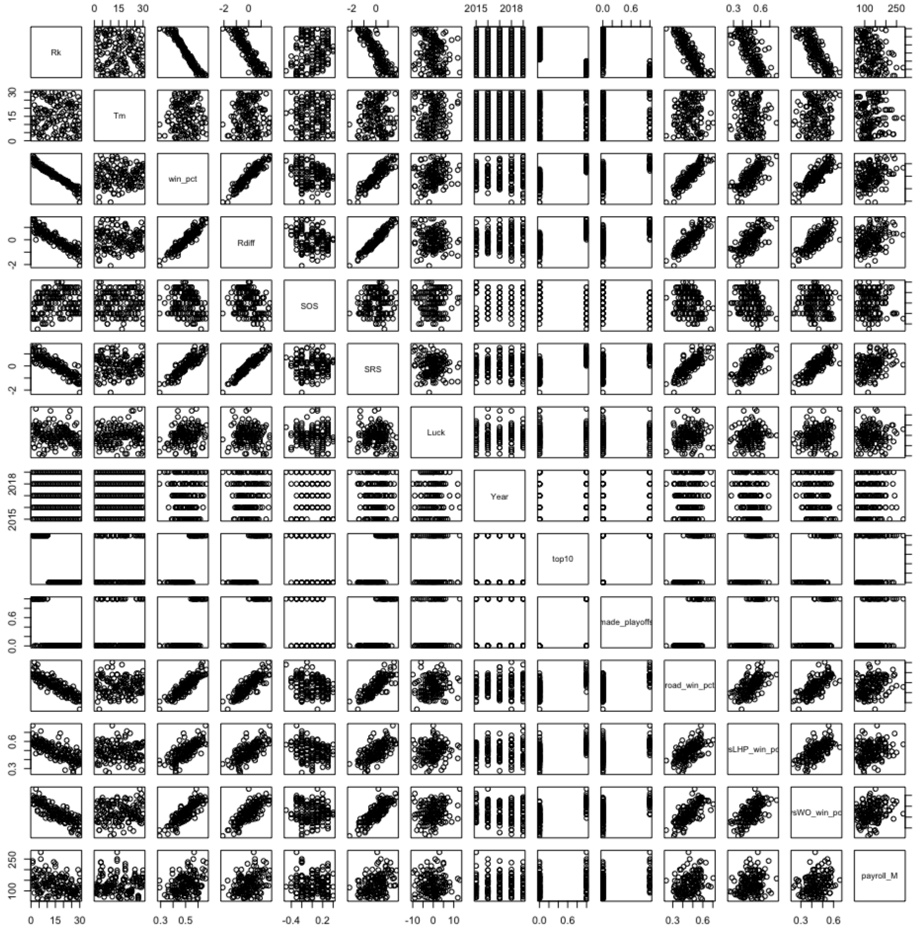
After completing basic EDA on the data set, there was a clear relationship between a team's winning % and their total payroll/player salaries. When looking at what features best describe a team's ability to make the playoffs, it was no surprise that certain features that measure their strength were most significant.

Moving from EDA to the actual modeling, I initially constructed simple and multiple linear regression models to determine what features could best estimate a team's winning %. Interestingly enough, the SLR models performed much better than the MLR models, with the team's payroll being the most influential feature.

Once I had a chance to understand what features affect a team's winning %, I re-focused my modeling efforts towards a team's ability to make the playoffs. After lots of trial and error, I found a high level of satisfaction with the performance of a logistic regression model using `road-win-pct`, `vsWO-win-pct`, and `payroll-M`.

Results

EDA



Given the number of potential predictors, I used a correlation plot matrix to identify the most significant relationships visually. There is a false sense of correlation among all of the variables that represent some type of a team's win %, however there are a lot of interesting relationships to note visually as it relates to `win-pct`. There is a very strong positive relationship between `win-pct` and `Rdiff` (total runs scored - total runs scored by opponent), but not as strong between `win-pct` and `payroll-M` although there is still some discernible trend. When it comes to `made-playoffs`, there is no surprise that `win-pct` has a strong relationship with it. Interestingly enough though, some of the more descriptive win % stats such as `vsWO-win-pct`, `road-win-pct`, `vsLHP-win-pct`, and `Rdiff`.

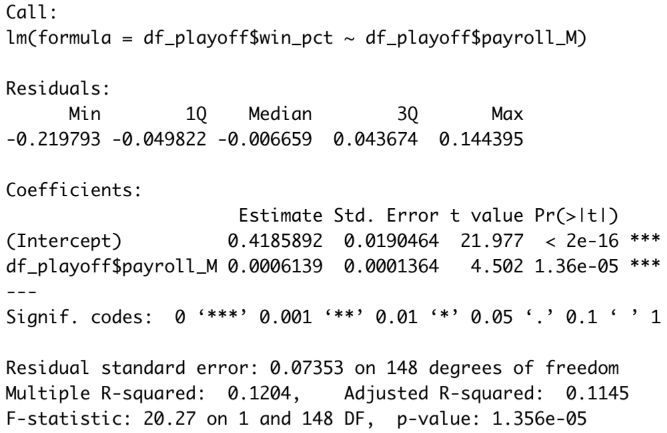
Results (cont.)

Estimating a Team's Winning %

Utilizing SLR and MLR, I found that the model that answered my research questions the best was a SLR model using only `payroll-M` as the sole predictor:

$$\texttt{win-pct} = \beta_0 + \beta_1\texttt{payroll-M} + \epsilon$$

$$\texttt{win-pct} = 0.418589 + 0.000614\texttt{payroll-M} + \epsilon$$



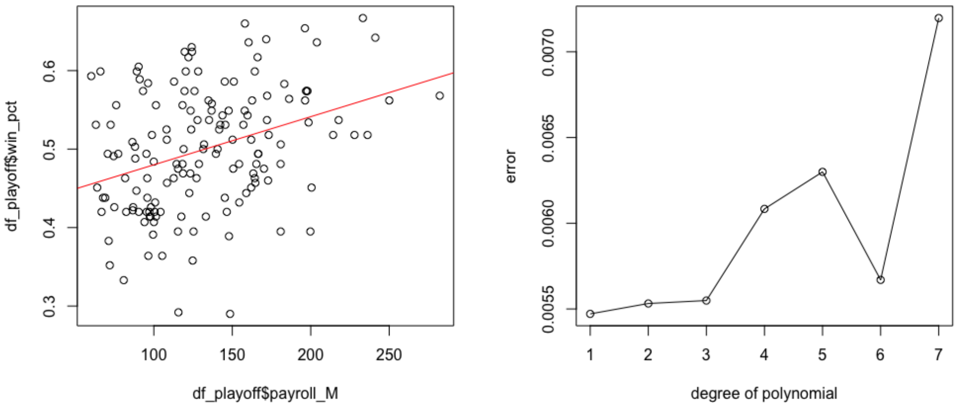
```
Call:
lm(formula = df_playoff$win_pct ~ df_playoff$payroll_M)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.219793 -0.049822 -0.006659  0.043674  0.144395 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          0.4185892  0.0190464  21.977  < 2e-16 ***
df_playoff$payroll_M 0.0006139  0.0001364   4.502 1.36e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07353 on 148 degrees of freedom
Multiple R-squared:  0.1204,	Adjusted R-squared:  0.1145 
F-statistic: 20.27 on 1 and 148 DF,  p-value: 1.356e-05
```

Using R, we fit this model to produce the summary above. We can see that although the model overall is significant along with the intercept and single predictor (`payroll-M`), our adj. R^2 is unfortunately quite low at 0.1145. At the end of the day, this does make sense due to the fact that so many other factors contribute to a team's winning %, but it is interesting to still be able to draw valuable insights from the information at hand.

Based on this model, we can infer that a team's `win-pct` will typically increase by 0.0006139 for every additional \$1M (USD), or 0.006139 for every additional \$10M, that is spent by the team on total player salaries (for one season). Since our `win-pct` is in decimal form, we can re-write our findings to say that a team's winning % will typically increase by 0.6139% for every additional \$10M that is spent on total player salaries. While the accuracy of the model's coefficients may not be perfect, it is extremely interesting to note that should a team want to sign an all-star level player (at approx. \$10M per season), they would only be able to increase their winning % on average by 0.6139, which roughly equates to one single game over a 162-game season. Therefore, there is indication that adding all-star/luxury players to a team's roster may not actually increase their winning % by as much as many may think. While it is fair to argue that a team's ability to make the playoffs could come down to 1-2 games, it seems like a negligible improvement for the amount of money being spent.

Results (cont.)

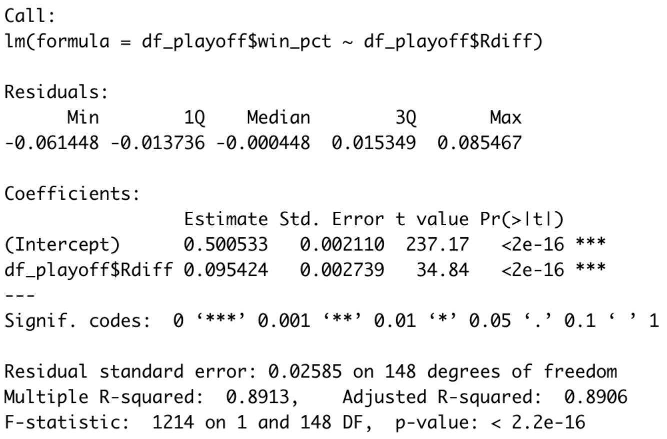
After plotting the data with the model overlaid, we can see that there is a decent amount of deviance from the prediction line. I ran LOOCV on the model for polynomial degrees of 1,2,3,4,5,6,7, but found that the first order of `payroll-M` produces the lowest error.



I did not find any significant MLR models, however I found another SLR model to be quite interesting. I utilized the well known statistic of a team's run differential (total runs scored - total runs scored by opponent) as the sole predictor:

$$\texttt{win-pct} = \beta_0 + \beta_1\texttt{Rdiff} + \epsilon$$

$$\texttt{win-pct} = 0.418589 + 0.000614\texttt{Rdiff} + \epsilon$$



```
Call:
lm(formula = df_playoff$win_pct ~ df_playoff$Rdiff)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.061448 -0.013736 -0.000448  0.015349  0.085467 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      0.500533   0.002110  237.17   <2e-16 ***
df_playoff$Rdiff 0.095424   0.002739   34.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02585 on 148 degrees of freedom
Multiple R-squared:  0.8913,	Adjusted R-squared:  0.8906 
F-statistic:  1214 on 1 and 148 DF,  p-value: < 2.2e-16
```

Results (cont.)

Estimating a Team's Playoff Appearance Potential

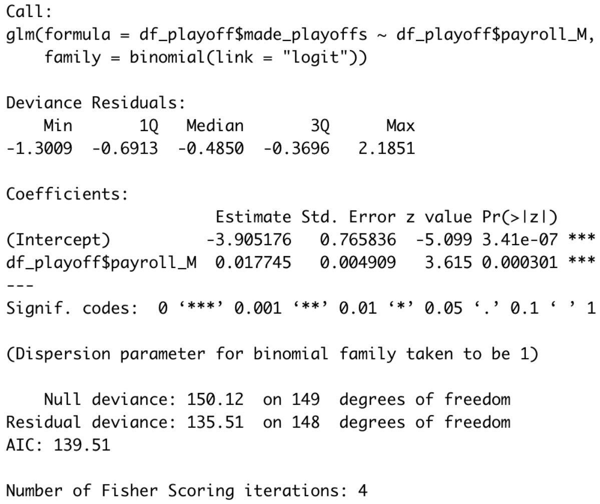
Since whether or not a team makes the playoffs is a binary outcome (either they make it or they don't), I utilized logistic regression models to estimate the probability of a team making it to the playoffs.

I initially created a logistic regression model using `payroll-M` as the sole predictor (let θ represent the odds of making the playoffs):

$$ln(\theta) = \beta_0 + \beta_1\texttt{payroll-M} + \epsilon$$

$$ln(\theta) = -3.905176 + 0.017745\texttt{payroll-M} + \epsilon$$

Interestingly enough, it actually performed surprisingly well on it's own. Using R, I obtained the following model summary along with the test error of **0.1770** generated using LOOCV:



```
Call:
glm(formula = df_playoff$made_playoffs ~ df_playoff$payroll_M,
    family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3009  -0.6913  -0.4850  -0.3696   2.1851

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.905176   0.765836  -5.099 3.41e-07 ***
df_playoff$payroll_M  0.017745   0.004909   3.615 0.000301 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 150.12  on 149  degrees of freedom
Residual deviance: 135.51  on 148  degrees of freedom
AIC: 139.51

Number of Fisher Scoring iterations: 4
```

While I was satisfied with the performance of this model, I wanted to test the compatibility of other meaningful predictors to see what improvements may be possible.

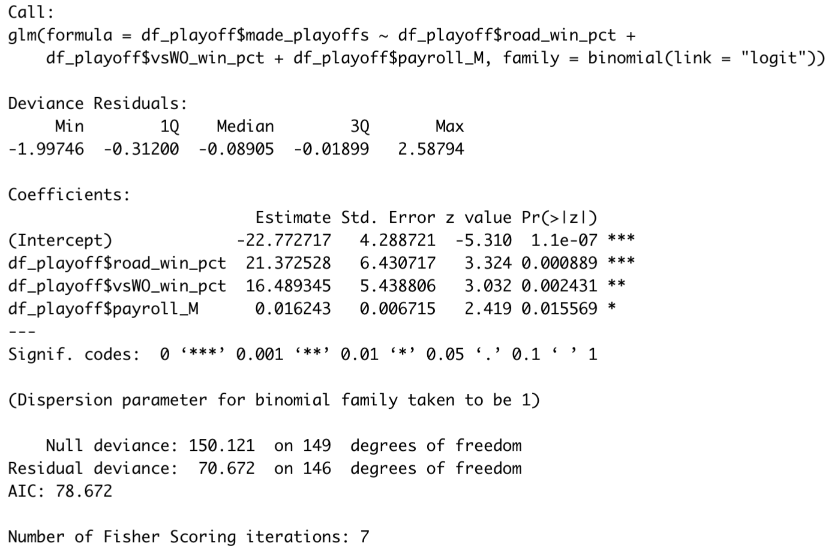
Results (cont.)

This exploration led me to the specific model including `payroll-M` along with `vsWO-win-pct` and `road-win-pct`:

$$ln(\theta) = \beta_0 + \beta_1\texttt{payroll-M} + \beta_2\texttt{vsWO-win-pct} + \beta_3\texttt{road-win-pct} + \epsilon$$

$$ln(\theta) = -22.77 + 0.02\texttt{payroll-M} + 16.5\texttt{vsWO-win-pct} + 21.37\texttt{road-win-pct} + \epsilon$$

Interestingly enough, it actually performed surprisingly well on it's own. Using R, I obtained the following model summary along with the test error of **0.2467** generated using LOOCV:



```
Call:
glm(formula = df_playoff$made_playoffs ~ df_playoff$road_win_pct +
    df_playoff$vsWO_win_pct + df_playoff$payroll_M, family = binomial(link = "logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.99746  -0.31200  -0.08905  -0.01899   2.58794

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -22.772717   4.288721  -5.310  1.1e-07 ***
df_playoff$road_win_pct  21.372528   6.430717   3.324 0.000889 ***
df_playoff$vsWO_win_pct  16.489345   5.438806   3.032 0.002431 **
df_playoff$payroll_M      0.016243   0.006715   2.419 0.015569 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 150.121  on 149  degrees of freedom
Residual deviance:  70.672  on 146  degrees of freedom
AIC: 78.672

Number of Fisher Scoring iterations: 7
```

This model performed the best among all surveyed in terms of the AIC (while also having a relatively low error) which makes sense since it includes arguably the most relevant features related to a team's ability to make the playoffs.

Estimation

Winning %

Utilizing the first SLR model I constructed, I estimated the 2021 Mariner's win % using this year's salary data. According to Spotrac[4], the Mariner's total team salary is $81,257,217. We use the following linear regression equation to make our estimation:

$$\texttt{win-pct} = \beta_0 + \beta_1 \texttt{payroll-M} + \epsilon$$

$$\texttt{win-pct} = 0.418589 + 0.000614 \texttt{payroll-M} + \epsilon$$

$$\texttt{win-pct} = \boxed{0.468}$$

Although we are only just over a third of the way through the 2021 season, the Mariner's current win % is 0.477 (based on a 31-34 record as of EOD 6/11/2021) which is actually quite close to our estimation.

Chance of Playoff Appearance

Utilizing the second logistic regression model I constructed, I estimated the 2021 Mariner's probability of making the playoffs using their current season stats. Again, the Mariner's total team salary is $81,257,217. According to MLB - Official Stats[3], the Mariner's away/road record is 14-20 and their record against winning opponents is 15-20 (as of EOD 6/11/2021). These records are used to calculate `road-win-pct` and `vsWO-win-pct` within R. We then use the following logistic regression equation to make our estimation:

$ln(\theta) = \beta_0 + \beta_1\texttt{payroll-M} + \beta_2\texttt{vsWO-win-pct} + \beta_3\texttt{road-win-pct} + \epsilon$

$ln(\theta) = -22.77 + 0.02\texttt{payroll-M} + 16.5\texttt{vsWO-win-pct} + 21.37\texttt{road-win-pct} + \epsilon$

$ln(\theta) =$ -5.5856

$P(\text{made-playoffs}) = \frac{e^\theta}{1+e^\theta}$

$P(\text{made-playoffs}) = \boxed{0.0037}$

Estimation (cont.)

At first I thought I incorrectly calculated the est. probability of the Mariners making the playoffs this year given it being so low. However after some careful consideration, there was unfortunately no bug or error committed, just an accurate representation of the little hope that Mariners fans usually have each season.

While validating my process to ensure it was correctly implemented, and in an attempt to raise my spirits, I estimated the probability of the Boston Red Sox making the playoffs.

According to Spotrac[4], the Red Sox's total team salary is $176,846,501. According to MLB - Official Stats[3], the Red Sox's away/road record is 20-10 and their record against winning opponents is 17-11 (as of EOD 6/11/2021). These records are used to calculate `road-win-pct` and `vsWO-win-pct` within R. We then use the following logistic regression equation to make our estimation:

$ln(\theta) = \beta_0 + \beta_1\texttt{payroll-M} + \beta_2\texttt{vsWO-win-pct} + \beta_3\texttt{road-win-pct} + \epsilon$

$ln(\theta) = -22.77 + 0.02\texttt{payroll-M} + 16.5\texttt{vsWO-win-pct} + 21.37\texttt{road-win-pct} + \epsilon$

$ln(\theta) = 4.3594$

$P(\text{made-playoffs}) = \frac{e^\theta}{1+e^\theta}$

$P(\text{made-playoffs}) = \boxed{0.9874}$

Discussion

Based on the significance of the models constructed, I am pleased with the results obtained. While I derived probably the worst possible answer to my original question of whether or not the Mariners have a chance of making the playoffs this year, I am glad to have an accurate result rather than not.

When it comes to estimating a team's probability of making the playoffs, there is an endless list of factors that contribute to the overall outcome. With the models I created, there are a number of assumptions that need to be considered such as injuries, strength of division, coaching/management etc., however I feel that the models still provide quite a descriptive estimation. Throughout a season, one could utilize the 2nd logistic regression model to track their team's chances of making the playoffs and also understand what attributes might increases their chances the most based on where they are currently at.

Although the models constructed were fairly rudimentary, they provide validation that there is predictive capability when it comes to estimating MLB team's chances of making the playoffs. In the future I plan to look for any other features that are significant/influential in predicting playoff appearance probabilities, in addition to incorporating more complex classification models to see if there is any improvement. I also specifically want to continue evaluating the effect/impact of a team's salary to ensure I understand the entirety of the influence that money has on the MLB.

While a lot can happen in a 162-game season, it is fun to make predictions but more importantly understand what improvements could be made that could actually put the Seattle Mariners in a position to finally make the playoffs.

References

1. Baseball Reference (https://www.baseball-reference.com/leagues/MLB/2019-standings.shtml)

2. Baseball Prospectus (https://legacy.baseballprospectus.com/compensation/index.php?cyear=2019)

3. MLB - Official Stats (https://www.mlb.com/standings)

4. Spotrac (https://www.spotrac.com/mlb/payroll/)

5. MATH 447 - Project Data Agg.xlsx (https://wwu2-my.sharepoint.com/:x:/g/personal/duncane6_wwu_edu/EVGgbrGkh8dKnpy5aBfFwEsBJwEPrl2tO3SubxwWbnB-Fg?e=UJMMgN)

6. MLB Team Stats 2015-2019.csv (https://wwu2-my.sharepoint.com/:x:/g/personal/duncane6_wwu_edu/ETbl7fe8zLREkuEe3f5eTUOBoMIczQMxYqwAHpeoTKhcyg?e=oG9Ex5)

Appendix - R Code

```
#install package for cv.glm()
library(boot)

# load the data
df = read.csv('MLB Team Stats 2015-2019.csv', header = TRUE)
# reduce data to variables of interest
df_playoff = df[-c(3:6, 8:9, 13, 15:20, 27)]

# quick EDA
pairs(df_playoff) ### {win_pct vs. run_diff}
# plot made_playoffs against likely predictors
par(mfrow=c(2,2))
plot(df_playoff$win_pct, df_playoff$made_playoffs)
plot(df_playoff$vsWO_win_pct, df_playoff$made_playoffs)
plot(df_playoff$road_win_pct, df_playoff$made_playoffs)
plot(df_playoff$vsLHP_win_pct, df_playoff$made_playoffs)



# SLR model for win_pct = payroll_M
slr_model_payroll = lm(df_playoff$win_pct ~ df_playoff$payroll_M)
slr1_summ = summary(slr_model_payroll)
# plot data against regression eq.
plot(df_playoff$payroll_M, df_playoff$win_pct)
abline(a=slr1_summ$coefficients[1], b=slr1_summ$coefficients[2], col='red')
# make prediction for Mariner's win_pct
pred_1 = slr1_summ$coefficients[1] + slr1_summ$coefficients[2]*(81.257217)
pred_1

# SLR model for win_pct = Rdiff
slr_model_runDiff = lm(df_playoff$win_pct ~ df_playoff$Rdiff)
summary(slr_model_runDiff)
```

```
# LOOCV on orders of payroll_M
x = df_playoff$payroll_M
y = df_playoff$win_pct
# df consisting of payroll_M and win_pct only
df = data.frame(x,y)
# vector to store error for each model
errors = c()
# for loop over the 7 models we want to fit
for (i in 1:7){
  set.seed(99)
  # fit model based on degree of polynomial
  fit.glm.i <- glm(y˜poly(x,i))
  # extract and append error to storage vector
  i_err = cv.glm(df, fit.glm.i)$delta[1]
  errors = c(errors, i_err)
}
# plot the errors
plot(c(1:7),errors, xlab = "degree of polynomial", ylab = "error")
lines(c(1:7),errors)
```

```
# logistic regression model to predict odds of making playoffs using only payroll_M
logit_modelX = glm(df_playoff$made_playoffs ~df_playoff$payroll_M,
                   family = binomial(link = "logit"))
summary(logit_modelX)
cv.glm(df_playoff, logit_modelX)$delta[1]

# logistic regression model to predict odds of making playoffs
logit_model = glm(df_playoff$made_playoffs ~ df_playoff$road_win_pct
                  + df_playoff$vsWO_win_pct
                  + df_playoff$payroll_M, family = binomial(link = "logit"))
logit_1 = summary(logit_model)
cv.glm(df_playoff, logit_model)$delta[1]

# calc. current season stats for Mariners needed for prediction
SEA2021_road_win_pct = 14/(14+20)
SEA2021_vsWO_win_pct = 15/(15+20)
# prediction for Seattle Mariners
lnOdds = logit_1$coefficients[1] + logit_1$coefficients[2]*(SEA2021_road_win_pct)
      + logit_1$coefficients[3]*(SEA2021_vsWO_win_pct)
      + logit_1$coefficients[4]*(81.257217)
pred_prob = exp(lnOdds)/(1+exp(lnOdds))

# calc. current season stats for Boston Red Sox needed for prediction
BOS2021_road_win_pct = 20/(20+10)
BOS2021_vsWO_win_pct = 17/(17+11)
# prediction for Boston Red Sox
lnOdds_BOS = logit_1$coefficients[1] + logit_1$coefficients[2]*(BOS2021_road_win_pct)
      + logit_1$coefficients[3]*(BOS2021_vsWO_win_pct)
      + logit_1$coefficients[4]*(176.846501)
pred_prob_BOS = exp(lnOdds_BOS)/(1+exp(lnOdds_BOS))
```