2002 Oakland Analysis

By Sarthak Jain

(1987 words)

Abstract

The game of baseball has notoriously been a stat driven game. However, the game was revolutionized by a man named Billy Beane with his use of data analytics to build a high-quality team with the budget of a small market team. Beane's journey to changing sports as we now it was most famously depicted in the movie and book "Money Ball". In the past managers of baseball teams would look at strange factors to determine worth, for example, if a pitcher does not have an attractive girlfriend he would not be considered confident and lose value. Beane began to throw abstract factors like this out and began to look at statistical evidence of well play such as on base percentage and walks. This resulted in a huge success for the Oakland Athletics and the game was changed. Now the game has continued on this path and even some stats previously thought of as the most important, like homeruns, have become less valuable then stats like slugging. my goal was to continue the work of Billy Beane and examine how data analytics effect the Oakland Athletics.

Findings:

- Slugging and walks were the most significant stats that contribute to wins
- Run Differential was the biggest contributor to making the playoffs
- Money spent on the team does not always translate to wins

Conclusion:

When trying to do the "money ball" approach, finding players that succeed in getting walks, as well as high slugging percentage, will lead to the team having a better run differential

which is the best way to make the playoffs. Stats in significant areas are the driver of baseball, so players who are cheap but good in those areas could be more beneficial then expensive players.

Introduction

I used the datasets Original.Baseball, Team, and Salary, which I obtained from Kaggle and baseball-reference to see which stats truly are significant in baseball. The dataset contained play by play data dating all the way back to 1871 and had data from things such as wins and losses, to more advanced statistics such as slugging, on base percentage, and ERA (Earned run average). In addition to that the data set also included information on the managers for the teams and the salaries of the players. While all this data was not used, it allowed me to look at this problem from any angle I wanted to. The goal of this project was to do a modern day "Money ball" type analysis and try to find what stats should be most highly regarded when signing a player with the intent to win games. This would not only give the team incentive to find players with these skill sets, but can also give a better indication as to how much they should be paid.

To find this information out I selected data from the years 1961-2016 to give a very broad view of the subject and a way to see how the money ball era has impacted baseball. I ran a linear regression to test the significance of specific stats to see how much each one contributed to wins, as well as figure which team had historically won the most and what steps they took to get there. First, I wanted to see how Billy Beane alone impacted the team and compared the stats of his team to other Athletics teams. As expected Beane ran a much more efficient team and got a significantly higher number of wins. Next, I found that the New York Yankees had won the most and decided to compare them with the Oakland Athletics to see which stats were similar, or how they were different. For both organizations the significant stats that lead to the most wins for

each team were similar, the only difference was the Yankees performed better in those areas. Knowing the market size of Oakland, I wanted to see how much it "costs to win" by looking at the salary of the team at the time and seeing if the more money paid translated to a better record. Surprisingly more money paid did not always result in more wins.

As a manager Billy Beane's main goal each year was to make the playoffs and I wanted to look at this problem through his eyes. Winning simply was not enough if the season did not end with a trip to the playoffs. I made a linear regression model with playoffs as the Y variable and found that run differential was one of the most significant stats to a teams playoff birth I found that teams with at least 95 wins or a run differential of +133 seem to make it the most. Finally, I created a Neural Network that included every MLB team in addition to the significant stats that I found to see how the inputs relate to wins. Due to the high number of inputs I included multiple layers with multiple nodes to try to get the best output possible.

Variable Definitions

Year – Year which the data is from

G - Games

W – Wins

L – Losses

RD - Run differential

RA – Runs allowed

RS – Runs scored

SLG – Slugging

OBP – On base percentage

OSLG – Opponent slugging

OOBP – Opponent on base percentage

BB – Walks

BA – Batting average

Playoffs – Did the team make the playoffs or not

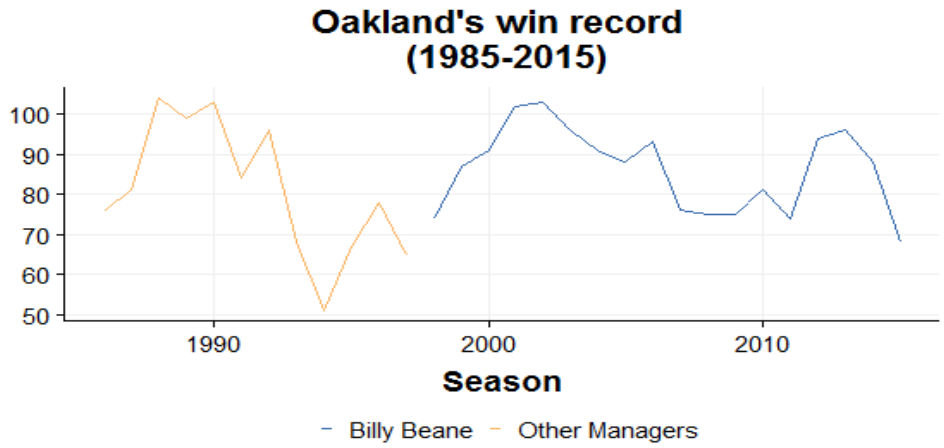
Methods and Analysis

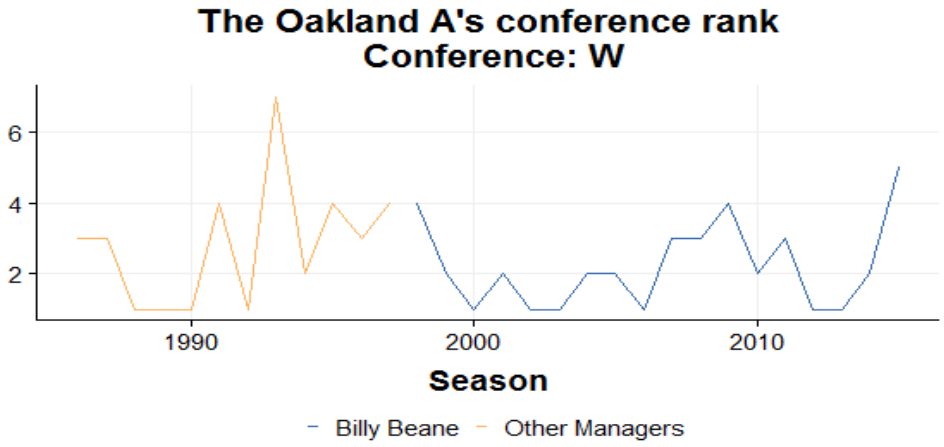
For my statistical analysis the first model I tried to perfect was my linear regression model. The model was runs allowed in relation to: hits, walks, and slugging

```
Call:
lm(formula = RS ~ h + bb + +SLG, data = NYY2002T)

Residuals:
    Min      1Q  Median      3Q     Max
-38.133 -12.439  -0.186  10.181  43.694

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -523.53139   43.41381 -12.059 1.22e-13 ***
h              0.25400    0.06151   4.129 0.000233 ***
bb             0.28446    0.05750   4.948 2.15e-05 ***
SLG         1839.86587  205.86938   8.937 2.50e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.25 on 33 degrees of freedom
Multiple R-squared:  0.9676,    Adjusted R-squared:  0.9647
F-statistic: 328.6 on 3 and 33 DF,  p-value: < 2.2e-16
```

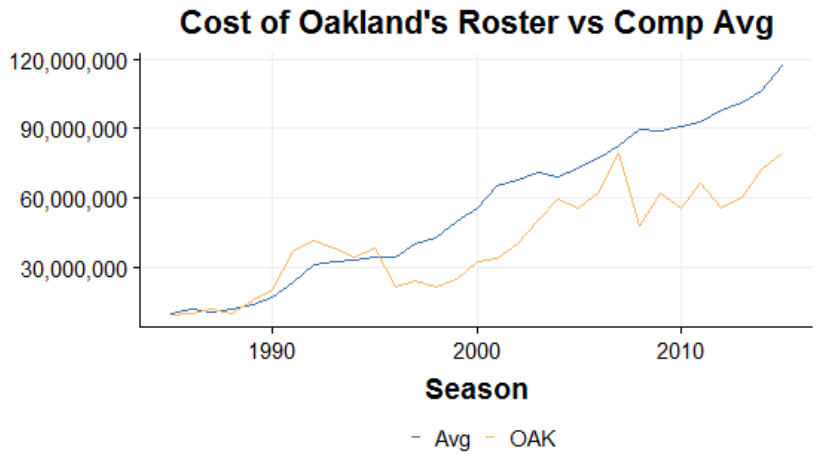
```
Call:
lm(formula = RS ~ h + bb + +SLG, data = OAK2002T)

Residuals:
    Min      1Q  Median      3Q     Max
-58.158 -11.637  -0.829  20.574  35.416

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -609.17438  111.66974  -5.455 1.02e-05 ***
h              0.40372    0.13838   2.918  0.00718 **
bb             0.47927    0.09803   4.889 4.51e-05 ***
SLG         1271.50332  388.07895   3.276  0.00298 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.43 on 26 degrees of freedom
Multiple R-squared:  0.9281,    Adjusted R-squared:  0.9198
F-statistic: 111.9 on 3 and 26 DF,  p-value: 5.516e-15
```

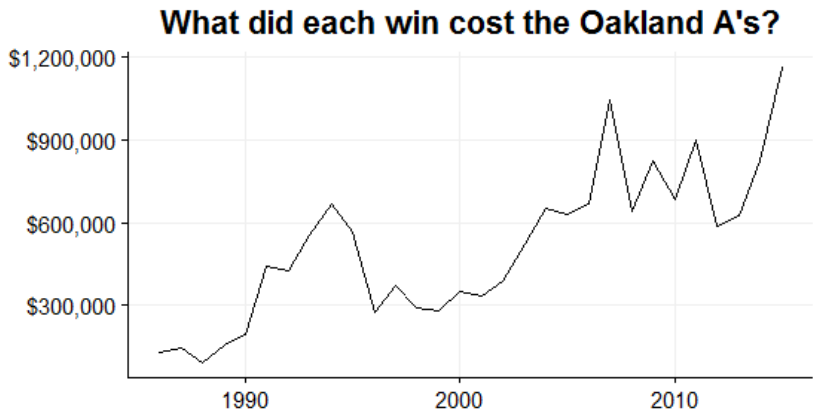
This compared what went into how significant certain variables are to runs scored for a team. The two teams used are the New York Yankees, the best team over the time period looked at, and the Oakland Athletics. For both teams the same data points were highly significant except for the Yankees they were more significant due to the fact that those translated to more runs scored during that time. In addition, the models had a R-squared of .9647 for the Yankees and .9198 for the Athletics. The goal behind this regression was to see if there was statistical difference between historically highest paid & most winning roster and one of the lowest paid & an average performing roster.



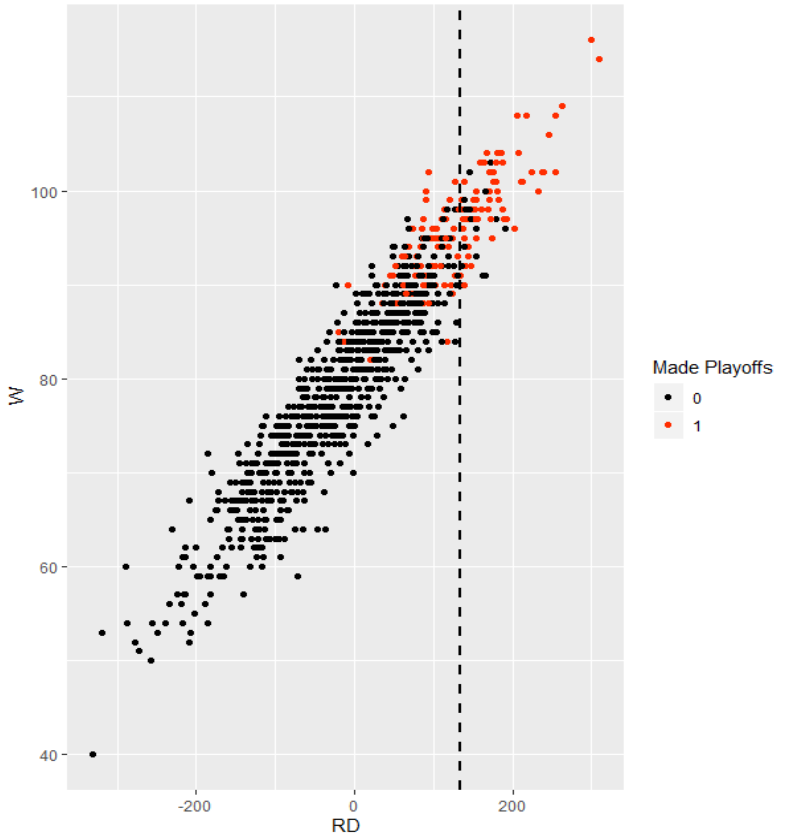


This shows the impact of Billie Beane's money ball approach. It is evident that the team was in a slump that Beane was able to turn around in a very impressive amount of time once implementing the money ball approach, proving the impact of the method.





Oakland is considered a small market team compared to other places such as New York or Houston so being able to be successful while keeping a low payroll is the key to success. These clearly show the impact of the money ball approach. The team was paying more than average with no success, but as soon as Beane began his money ball approach (year 2002) there is a drop in the cost per win, however the team was more successful during this era.



RD(Run differential)

W(Wins)

lm(formula = W ~ RD, data = dataBefore2002)

Coefficients:

(Intercept) RD

80.8814 0.1058

W = 80.8814 + 0.1058(RD)

RD = (W - 80.8814) / 0.1058

RD = (95 - 80.8814) / 0.1058

RD ≈ 133

The two graphs above show the impact of what was my most important stat to making the playoffs, run differential. The average number of wins needed is 95, since this is a baseline for how to make the playoffs it is important to understand which stats lead to most wins. I used 95 as x intercept because when the line was placed at 90 on the x-axis there would, there would too many black dots on the right side of the line. I determined that run differential was the best way to get wins. The run differential

is a tightly packed cluster of teams that made the playoffs that all had a significant positive run differential against their opponents with the exception of a few outliers. This shows that similar to wins, if you get a run differential of around +133 you are extremely likely to make the playoffs.

```
> runsScoredRegNoBA <- lm(RS ~ OBP + SLG, data = dataBefore2002)
> summary(runsScoredRegNoBA)

Call:
lm(formula = RS ~ OBP + SLG, data = dataBefore2002)

Residuals:
    Min      1Q  Median      3Q     Max
-70.838 -17.174  -1.108  16.770  90.036

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -804.63      18.92  -42.53   <2e-16 ***
OBP          2737.77      90.68   30.19   <2e-16 ***
SLG          1584.91      42.16   37.60   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.79 on 899 degrees of freedom
Multiple R-squared:  0.9296,    Adjusted R-squared:  0.9294
F-statistic:  5934 on 2 and 899 DF,  p-value: < 2.2e-16
```

To answer this, I needed to know which stats were the most statistically significant when trying to predict the number of runs scored in a season. my data set contains three stats that I focused on OBP, Slugging, and Batting average. At first I ran the regression with batting average variable (adj. $R^2 = .93$), but the variable wasn't significant so I took the batting average variable out and ran the regression again (adj. $R^2 = .9294$). Since, the adj. R^2 was only affected by 0.06% I decided to proceed forward with above model.

During the time Beane and DePodesta were performing these calculations, traditional major league scouts placed a heavy emphasis on batting average (BA). Before looking at the data, it seems like it would be the most directly correlated with scoring runs, but that is not the case.

Traditional scouts had no clue that they were looking for the wrong attributes all along. That blissful ignorance was one of the main reasons Beane and DePodesta were able to gain such an advantage.

```
> runsAllowedReg <- lm(RA ~ OOBP + OSLG, data = dataBefore2002)
> summary(runsAllowedReg)

Call:
lm(formula = RA ~ OOBP + OSLG, data = dataBefore2002)

Residuals:
    Min      1Q  Median      3Q     Max
-82.397 -15.178  -0.129  17.679  60.955

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -837.38      60.26 -13.897  < 2e-16 ***
OOBP         2913.60     291.97   9.979 4.46e-16 ***
OSLG         1514.29     175.43   8.632 2.55e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.67 on 87 degrees of freedom
  (812 observations deleted due to missingness)
Multiple R-squared:  0.9073,    Adjusted R-squared:  0.9052
F-statistic: 425.8 on 2 and 87 DF,  p-value: < 2.2e-16
```

Both of the model that I prepared have extremely strong adj. R^2. I had some down time so I performed some algebra and to find out what OBP and SLG do I need to achieve a run-differential of +133.

RA = -837.38 + 2913.60(OOBP) + 1514.29(OSLG)

The OOBP and OSLG for the A's in 2001 were:

OOBP = 0.308

OSLG = 0.380

we get: RA = -837.38 + 2913.60(0.308) + 1514.29(0.380)

= -837.38 + 897.39 + 575.43 ≈ 600

So basically A's need to score 733 runs (600 + 133)

A's 2002 season stats: OBP=.339; SLG=.432; OOBP=.315; OSLG=.384. When I plug the values in the runs allowed equation I get:

RA = -837.38 + 2913.60(0.315) + 1514.29(0.384) ≈ 662. The actual runs allowed were 654; Plugging in the values to the runs scored regression equation we get: RS = -804.63 + 2737.77(0.339) + 1584.91(0.432) ≈ 808 The actual runs scored in 2002 was 800 & the actual RD was 146. my equation predicts 96 wins W = 80.8814 + 0.1058(146) and the actual season wins for A's were 103.

Next I made a Neural network that included every MLB team to see which teams, combined with the RD, OBP, SLG, BA in relation to wins. This showed how those inputs when applied can lead to more wins. Due to the large number of inputs used, I used three hidden layers with six nodes in each layer.

```
> table(actual, prediction)
      prediction
actual   0   1
     0   9   1
     1   4 166
```

The confusion matrix gave mainly true positives and very few true negatives which shows the accuracy of the model due to the small number of false positives and negatives. The accuracy of the model is at 96.67% and the error rate is only at 3.33% which shows that the model performed very well.

Conclusion

In conclusion, the money ball approach works for a reason. By finding which stats are important to a teams success, the run differential, and finding the stats that help get the best run differential, Slugging and walks, will inevitably lead to the ultimate goal of making the playoffs. By focusing on these significant variables instead of outside abstract variables it is easier to find the exact kind of players that can lead your team to victory. In addition, this also helps find the budget players that may have the ability to impact the team more than the super expensive players.

The ability to focus on the goal of 95 wins or +133 run differential allows teams to hand pick players who fill certain roles to achieve this goal. The models ran pointed with very high confidence that these variables are the key to a good season. The money ball approach may truly be the most efficient way to win the game.

References

Michael Lewis, 2010, *The Big Short: Inside the Doomsday Machine*

https://www.baseball-reference.com

https://www.kaggle.com/wduckett/moneyball-mlb-stats-19622012