Baseball Data Analysis

Introduction

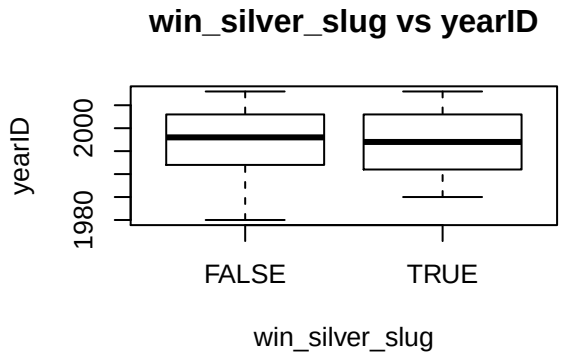
Model of Silver Slugger Award

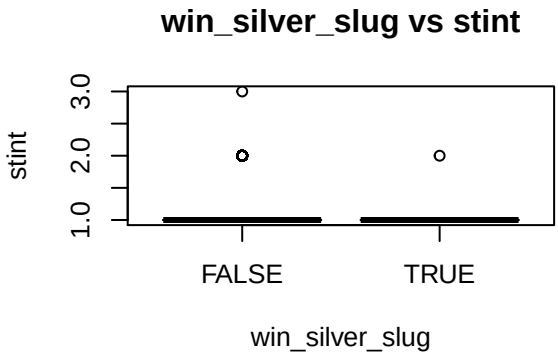
Full Model

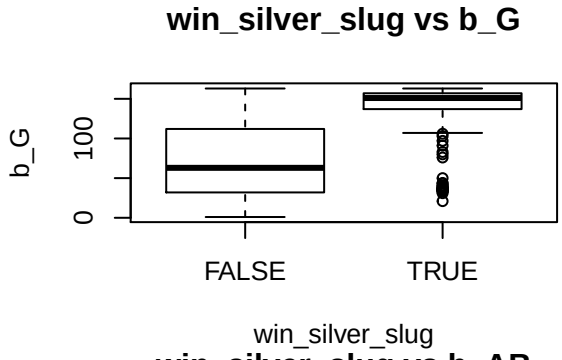
The original model was: **win_silver_slug = yearID + stint + b_G + b_G_batting + b_AB + b_R + b_H + b_2B + b_3B + b_HR + b_RBI + b_SB + b_CS + b_BB + b_SO + b_IBB + b_HBP + b_SH + b_SF + b_GIDP + b_G_old**

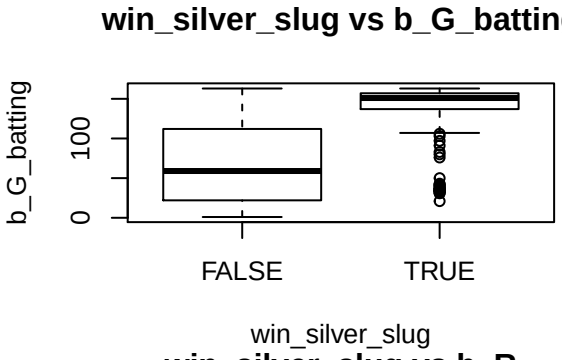
Analysis

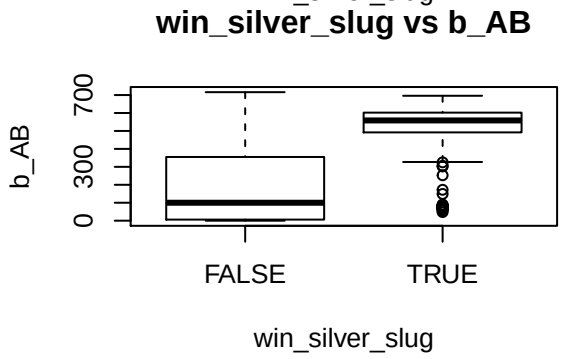
```
plot_all_box(which(colnames(batting_data) == "win_silver_slug"), batting_data)
```

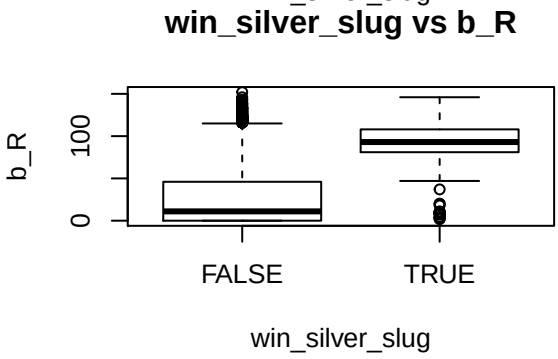

win_silver_slug vs yearID
yearID
1980
2000
FALSE
TRUE
win_silver_slug

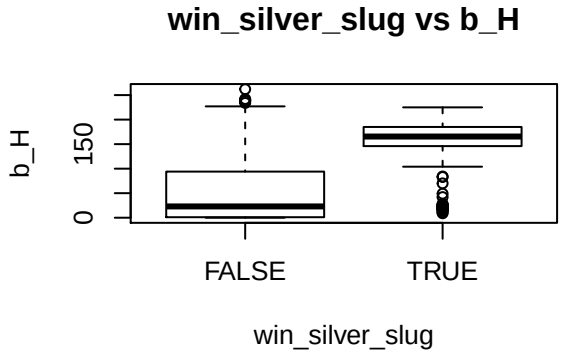

win_silver_slug vs stint
stint
1.0
2.0
3.0
FALSE
TRUE
win_silver_slug

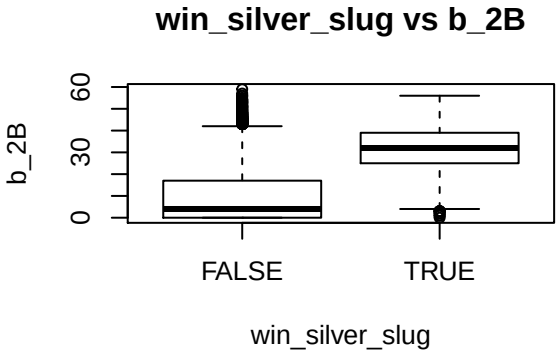

win_silver_slug vs b_G
b_G
0
100
FALSE
TRUE
win_silver_slug

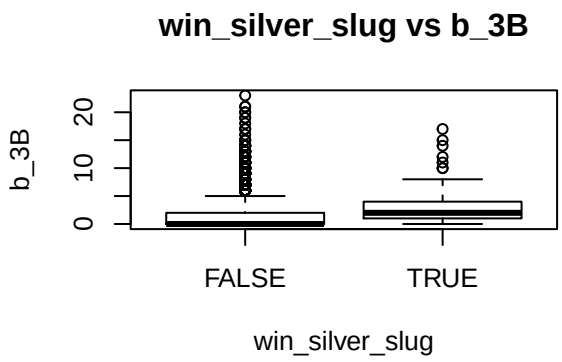

win_silver_slug vs b_G_batting
b_G_batting
0
100
FALSE
TRUE
win_silver_slug

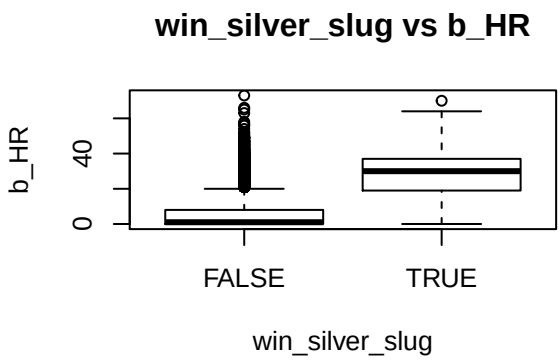

win_silver_slug vs b_AB
b_AB
0
300
700
FALSE
TRUE
win_silver_slug

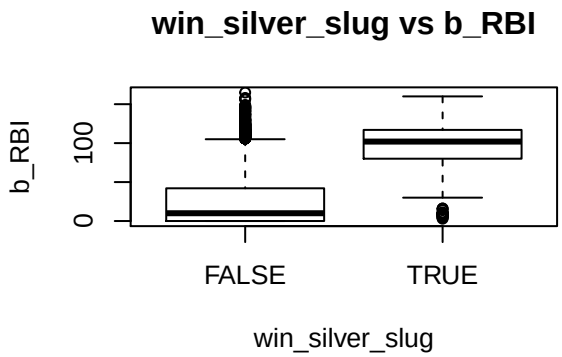

win_silver_slug vs b_R
b_R
0
100
FALSE
TRUE
win_silver_slug

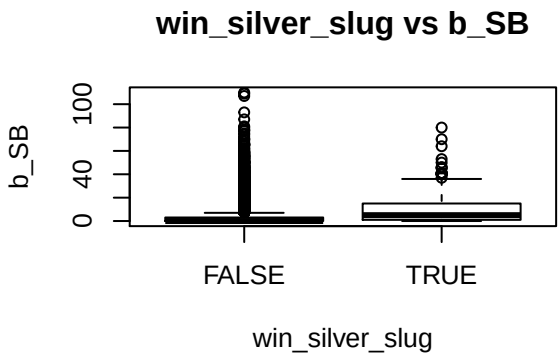

win_silver_slug vs b_H
b_H
0
150
FALSE
TRUE
win_silver_slug

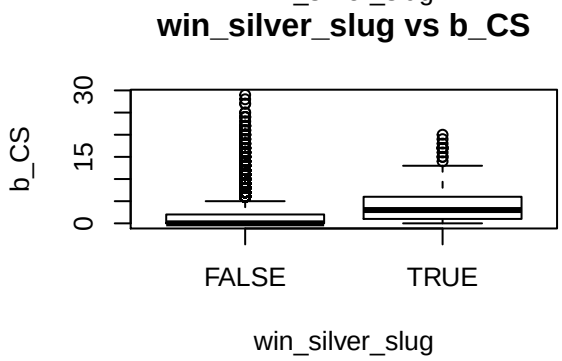

win_silver_slug vs b_2B
b_2B
0
30
60
FALSE
TRUE
win_silver_slug

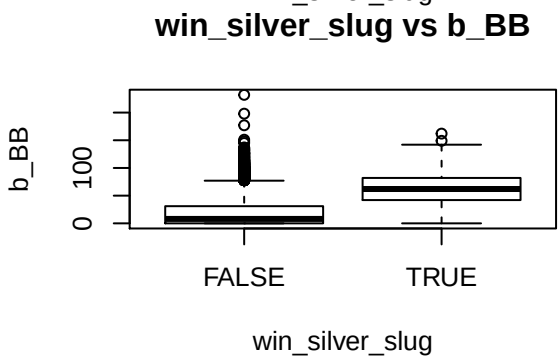

win_silver_slug vs b_3B
b_3B
0
10
20
FALSE
TRUE
win_silver_slug

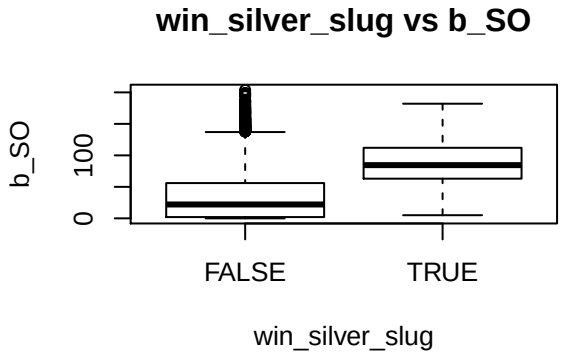

win_silver_slug vs b_HR
b_HR
0
40
FALSE
TRUE
win_silver_slug

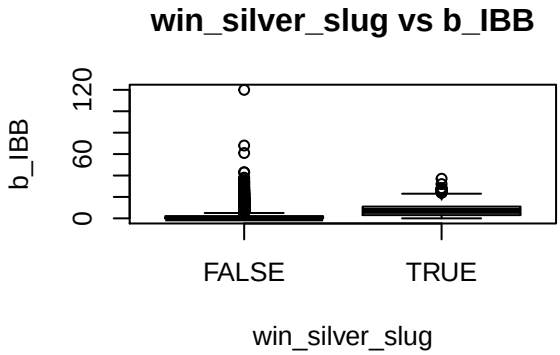

win_silver_slug vs b_RBI
b_RBI
0
100
FALSE
TRUE
win_silver_slug

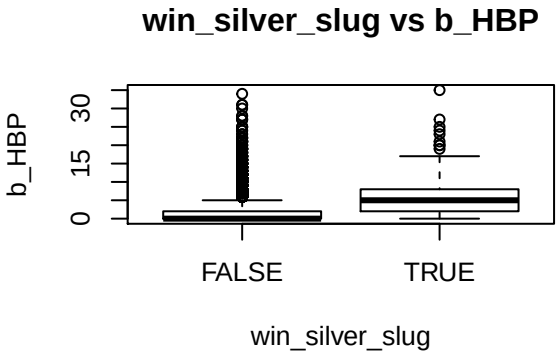

win_silver_slug vs b_SB
b_SB
0
40
100
FALSE
TRUE
win_silver_slug

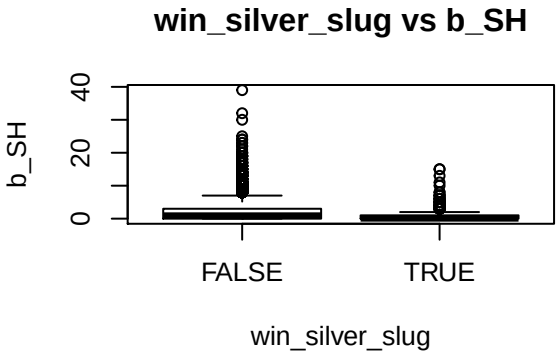

win_silver_slug vs b_CS
b_CS
0
15
30
FALSE
TRUE
win_silver_slug

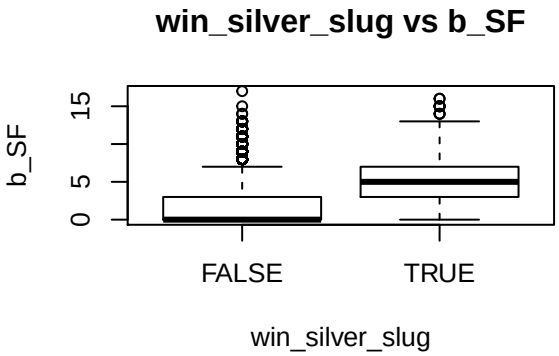

win_silver_slug vs b_BB
b_BB
0
100
FALSE
TRUE
win_silver_slug

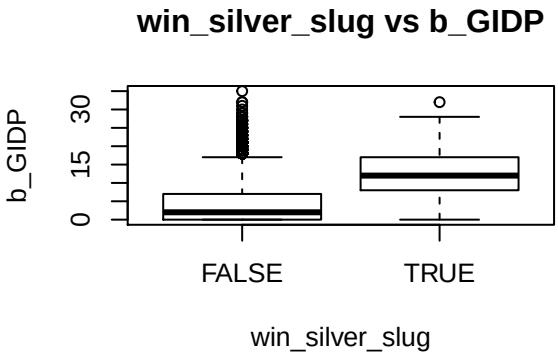

win_silver_slug vs b_SO
b_SO
0
100
FALSE
TRUE
win_silver_slug

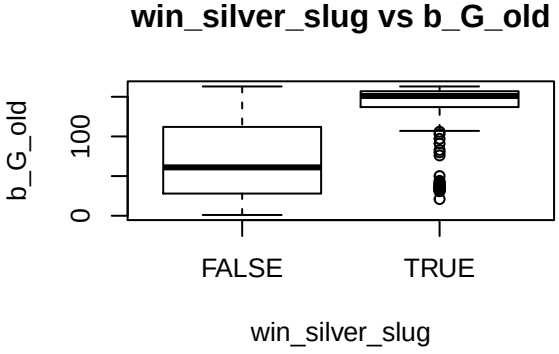

win_silver_slug vs b_IBB
b_IBB
0
60
120
FALSE
TRUE
win_silver_slug









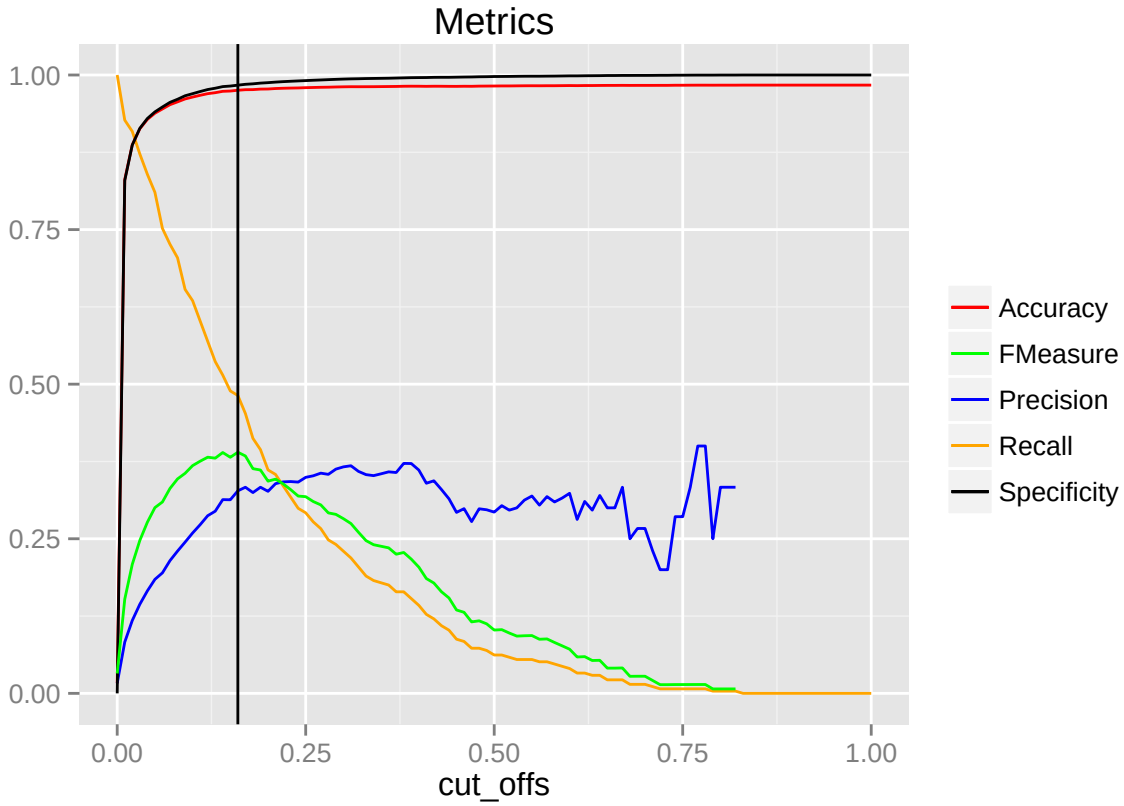


Model Selection

After backward selection the model that was chosen was: **log(odds(win_silver_slug)) = 110.502946507652 + -0.0580991182515473 * yearID + -0.0493472166980932 * b_G + 0.039786343060648 * b_H + 0.0477089384066037 * b_HR + 0.0208714277476128 * b_RBI + 0.00864433771119618 * b_BB + 0.0459306086282769 * b_HBP**

Finding a Optimal Cut Off

```
plot_of_cut_offs
```



After searching for a good cut off value, 0.16 was chosen.

```
best_cut_off
```

```
##   cut_offs true_positive true_negative false_positive false_negative
## 1     0.16           132         16185            271            142
##      recall  accuracy precision specificity f_measure      prior     lift
## 1 0.4817518 0.9753138 0.3275434   0.9835318 0.3899557 0.01637776 19.99928
```