Final Project

Benji Gold and Sam Alkalay

Packages and Data

```
library(tidyverse)
library(tidymodels)
library(glmnet)
library(caret)
library(MASS)
stats <- read.csv("data/stats.csv")


stats <- replace(stats, stats =="", NA)
stats <- stats %>%
  drop_na()
view(stats)
```

Lassos for Variable Selection

```
# LASSO Variable Selection Basic Stats
y <- stats$All.Star
x <- model.matrix(All.Star ~ player_age + b_ab + b_total_pa + b_total_hits +
                    b_double + b_triple + b_home_run + b_strikeout + b_walk +
                    b_k_percent + b_bb_percent + batting_avg + slg_percent +
                    on_base_percent, data = stats)
m_lasso_cv <- cv.glmnet(x, y, alpha = 1)
best_lambda <- m_lasso_cv$lambda.min
best_lambda
```

```
[1] 1.199787e-05
```

```
m_best <- glmnet(x, y, alpha = 1, lambda = best_lambda)
m_best$beta
```

```
15 x 1 sparse Matrix of class "dgCMatrix"
                           s0
(Intercept)      .
player_age      -0.0029714439
b_ab            -0.0066167989
b_total_pa       0.0039139641
b_total_hits     0.0075736864
b_double         0.0015694702
b_triple         0.0049747913
b_home_run       0.0128318399
b_strikeout     -0.0006560073
b_walk           0.0001915321
b_k_percent      0.0011541045
b_bb_percent     0.0026474619
batting_avg      0.9697000324
slg_percent     -0.0469204083
on_base_percent -1.3335030636
```

```
# LASSO Variable Selection Advanced Stats
y <- stats$All.Star
x <- model.matrix(All.Star ~ player_age + launch_angle_avg + sweet_spot_percent +
                    barrel + solidcontact_percent + flareburner_percent +
                    hard_hit_percent + avg_hyper_speed + z_swing_percent +
                    oz_swing_percent + meatball_swing_percent, data = stats)
m_lasso_cv <- cv.glmnet(x, y, alpha = 1)
best_lambda <- m_lasso_cv$lambda.min
best_lambda
```

```
[1] 0.006387699
```

```
m_best <- glmnet(x, y, alpha = 1, lambda = best_lambda)
m_best$beta
```

```
12 x 1 sparse Matrix of class "dgCMatrix"
                                  s0
(Intercept)             .
```

```
player_age             -0.0016947089
launch_angle_avg       -0.0001302131
sweet_spot_percent      .
barrel                  0.0079471445
solidcontact_percent   -0.0029771054
flareburner_percent    -0.0017692816
hard_hit_percent       -0.0017118515
avg_hyper_speed         .
z_swing_percent         .
oz_swing_percent        .
meatball_swing_percent -0.0021483152
```

Regressions

```
#Basic model
m1 <- glm(All.Star ~ player_age + b_ab +  b_total_hits +
                    b_double + b_triple + b_home_run + b_strikeout +
                    b_bb_percent + batting_avg + slg_percent +
                    on_base_percent,
  data = stats,
  family = "binomial"
)
tidy(m1)
```

```
# A tibble: 12 x 5
   term              estimate std.error statistic  p.value
   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)      -0.330      1.68      -0.196  0.844
 2 player_age       -0.0250     0.0503    -0.497  0.619
 3 b_ab             -0.0234     0.00686   -3.41   0.000647
 4 b_total_hits      0.0739     0.0246     3.00   0.00267
 5 b_double          0.0182     0.0378     0.481  0.630
 6 b_triple          0.00551    0.104      0.0530 0.958
 7 b_home_run        0.117      0.0449     2.60   0.00933
 8 b_strikeout       0.000198   0.00857    0.0231 0.982
 9 b_bb_percent      0.302      0.167      1.81   0.0700
10 batting_avg      27.3       16.4        1.67   0.0949
11 slg_percent       2.32       5.14       0.453  0.651
12 on_base_percent -39.9       19.1       -2.10   0.0361
```

```
m1_aug <- augment(m1) %>%
  mutate(prob = exp(.fitted)/(1 + exp(.fitted)),
         pred_leg = ifelse(prob > 0.5, "All-Star", "Not All-Star"))
table(m1_aug$pred_leg, m1_aug$All.Star)
```

```
              0   1
All-Star      9  18
Not All-Star 423  36
```

```
#Advanced model
m2 <- glm(All.Star ~ player_age + launch_angle_avg +
                     barrel + solidcontact_percent + flareburner_percent +
                     hard_hit_percent + meatball_swing_percent,
  data = stats,
  family = "binomial"
)
tidy(m2)
```

```
# A tibble: 8 x 5
  term                   estimate std.error statistic  p.value
  <chr>                     <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)             1.38       1.76       0.785 4.32e- 1
2 player_age             -0.0361     0.0468    -0.772 4.40e- 1
3 launch_angle_avg       -0.00881    0.0283    -0.312 7.55e- 1
4 barrel                  0.0852     0.0136     6.27  3.56e-10
5 solidcontact_percent   -0.0805     0.0943    -0.854 3.93e- 1
6 flareburner_percent    -0.0129     0.0379    -0.341 7.33e- 1
7 hard_hit_percent       -0.0256     0.0263    -0.974 3.30e- 1
8 meatball_swing_percent -0.0358     0.0162    -2.21  2.72e- 2
```

```
m2_aug <- augment(m2) %>%
  mutate(prob = exp(.fitted)/(1 + exp(.fitted)),
         pred_leg = ifelse(prob > 0.5, "All-Star", "Not All-Star"))
table(m2_aug$pred_leg, m2_aug$All.Star)
```

```
              0   1
All-Star      6  12
Not All-Star 426  42
```