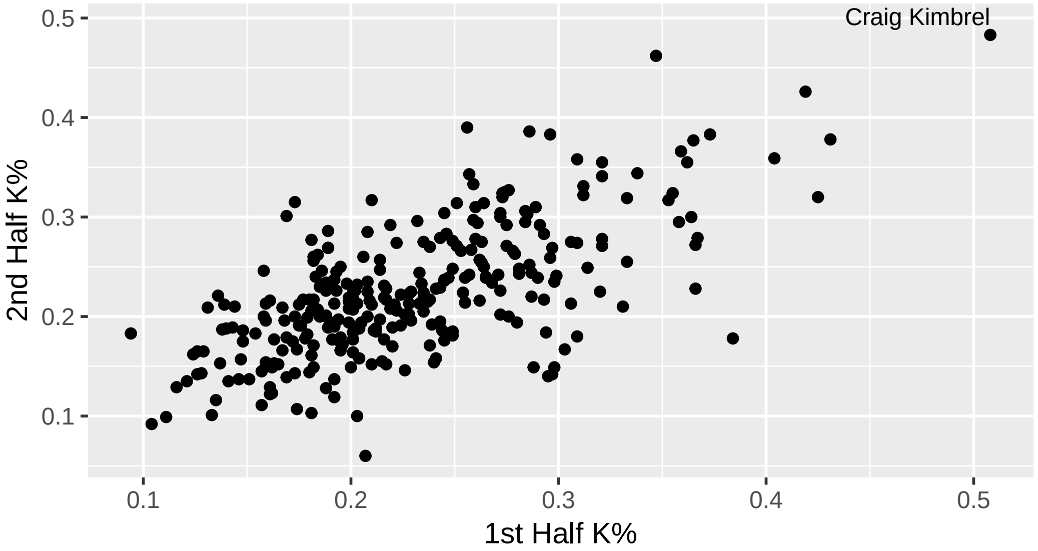
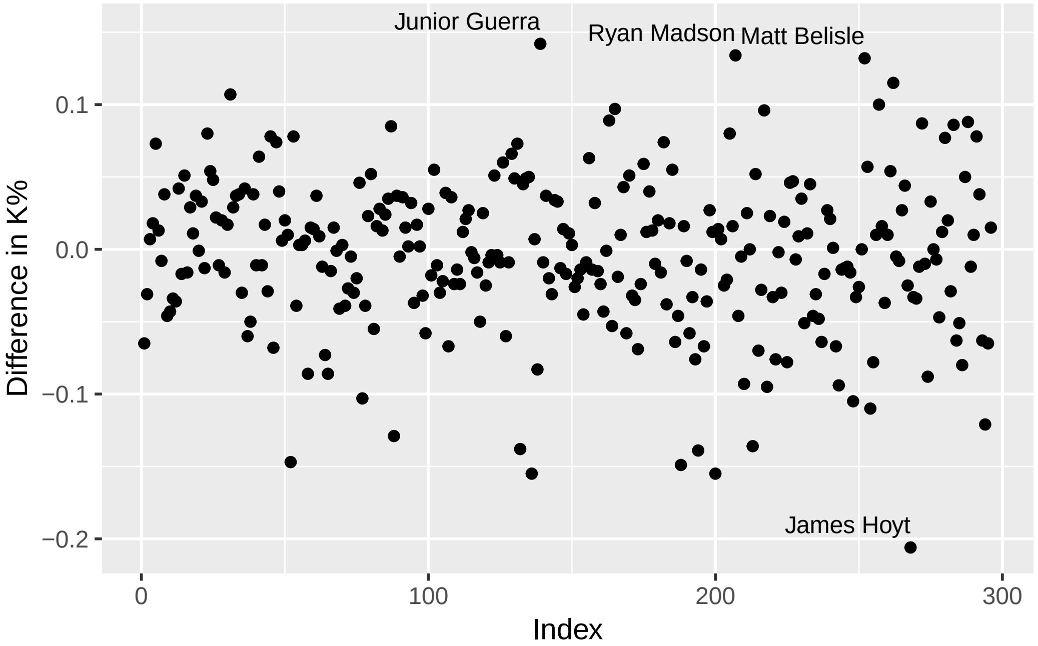
Predicting Strikeout Percentages

Nolan McCafferty

The goal of this analysis is to predict the strikeout percentage in the second half of the 2017 season for each player given his pitching statistics in the first half of 2017. The statistics for each pitcher include innings pitched, ERA, FIP, first half strikeout percentage, first half swing percentage, and many more. My initial step was to do some exploratory analysis of the variables. The plot below shows the relationship between first half strikeout percentage and second half strikeout percentage. This relationship is a positive one which makes sense because pitchers with a higher strikeout percentage in the first half should typically also have a high percentage in the second half of the year. Also, wow Craig Kimbrel had a **ridiculous** second half.



Looking at this plot I noticed significant differences between the first and second half K% for a couple pitchers. Wanting to explore this more, I then plotted the differences between the two:



From the plot above we can see that Junior Guerra, Ryan Madson, and Matt Belisle were the three pitchers that improved their strikeout rate the most from the first to the second half of the 2017 season. Additionally, Jame Hoyt was by far the worst case of a strikeout rate decreasing in the second half. Looking at the data for these four pitchers, the only major difference I can see between the Hoyt and the other three is his much lower contact rate in the first half.

Name	ERA	FIP	xFIP	AVG	K.	BB.	Swing.	Contact.	GB.	LD.	FB.	X2ndHalfK.
Junior Guerra	4.78	7.16	6.00	0.247	0.173	0.138	0.439	0.768	0.342	0.235	0.423	0.315
Ryan Madson	2.17	2.62	2.74	0.186	0.256	0.044	0.471	0.765	0.565	0.250	0.185	0.390
Matt Belisle	5.82	4.83	5.29	0.259	0.169	0.117	0.496	0.791	0.411	0.234	0.355	0.301
James Hoyt	4.91	3.04	2.24	0.254	0.384	0.044	0.483	0.621	0.416	0.208	0.377	0.178

Now to make the predictions we will compose several models. We will subset the data into training and testing datasets and use root-mean-squared-error to evaluate each model. First, linear regression:

```
## [1] 0.0542503
```

Then Ridge Regression and LASSO:

```
lambda.grid <- 10^seq(5,-5, length = 100)
fit.ridge.cv <- cv.glmnet(train.x, train.y, lambda = lambda.grid, alpha = 0)
ridge.fitted <- predict(fit.ridge.cv, newx = test.x, s = "lambda.min")
rmse(test.y, ridge.fitted)
```

```
## [1] 0.05296212
```

```
fit.lasso.cv <- cv.glmnet(train.x, train.y, lambda = lambda.grid, alpha = 1)
lasso.fitted <- predict(fit.lasso.cv, newx = test.x, s = "lambda.min")
rmse(test.y, lasso.fitted)
```

```
## [1] 0.05309137
```

Finally, we will use the powerful XGBoost algorithm:

```
RMSE <- c()
for (i in 1:200) {
  modelxg <- xgboost(data = train.x,
   label = as.matrix(train.y),
   objective = "reg:linear",
   eval_metric = "rmse",
   max_depth = 2,
   nrounds = i,
   verbose = FALSE
   )

xg.fit <- predict(modelxg, test.x)
RMSE[i] <- rmse(test.y, xg.fit)
}

best <- which.min(RMSE)

bst <- xgboost(data = train.x,
   label = as.matrix(train.y),
   objective = "reg:linear",
   eval_metric = "rmse",
   max_depth = 2,
   nrounds = best,
   verbose = FALSE
   )

xg.fit <- predict(bst, test.x)
rmse(test.y, xg.fit)
```

```
## [1] 0.0548927
```

```
knitr::kable(xgb.importance(model=bst))
```

Feature	Gain	Cover	Frequency
K.	0.6209242	0.3154706	0.2758621
Contact.	0.1677623	0.1987418	0.1724138
BB.	0.0810367	0.1414259	0.2068966
LD.	0.0426679	0.0955266	0.1034483
Swing.	0.0293365	0.0890028	0.0689655
IP	0.0238722	0.0689655	0.0689655
FIP	0.0214523	0.0470643	0.0689655
ERA	0.0129477	0.0438024	0.0344828

We can see that the XGBoost model actually performed worse than the previous models in terms of RMSE. From the feature importance table we can see that strikeout percentage was by far most important followed by contact percentage. This again makes sense because the strikeout percentages in the first and second half should typically be very similar. Of the models above, the Ridge Regression and LASSO models gave us the lowest RMSE. Using the RR model, the first few predictions can be seen below:

Name	Actual	Predicted
Nick Vincent	0.190	0.213
Adam Warren	0.224	0.233
Edwin Diaz	0.331	0.265
Chris Beck	0.103	0.193
Alex Wilson	0.189	0.178
Pat Neshek	0.320	0.250
Carl Edwards Jr.	0.355	0.317
Luis Garcia	0.208	0.217
Drew Storen	0.217	0.185
James Pazos	0.239	0.251