swstr_modeling

Owen McGrattan

10/21/2018

```r
# load in packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(baseballr)
library(readr)
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(stats)
library(glmnet)
```

```
## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-16
```

```r
library(caret)
```

```
## Loading required package: lattice
```

```r
# read in grouped df
df <- read_csv("swstr_pitchers.csv",
    col_types = cols(X1 = col_skip()))
```

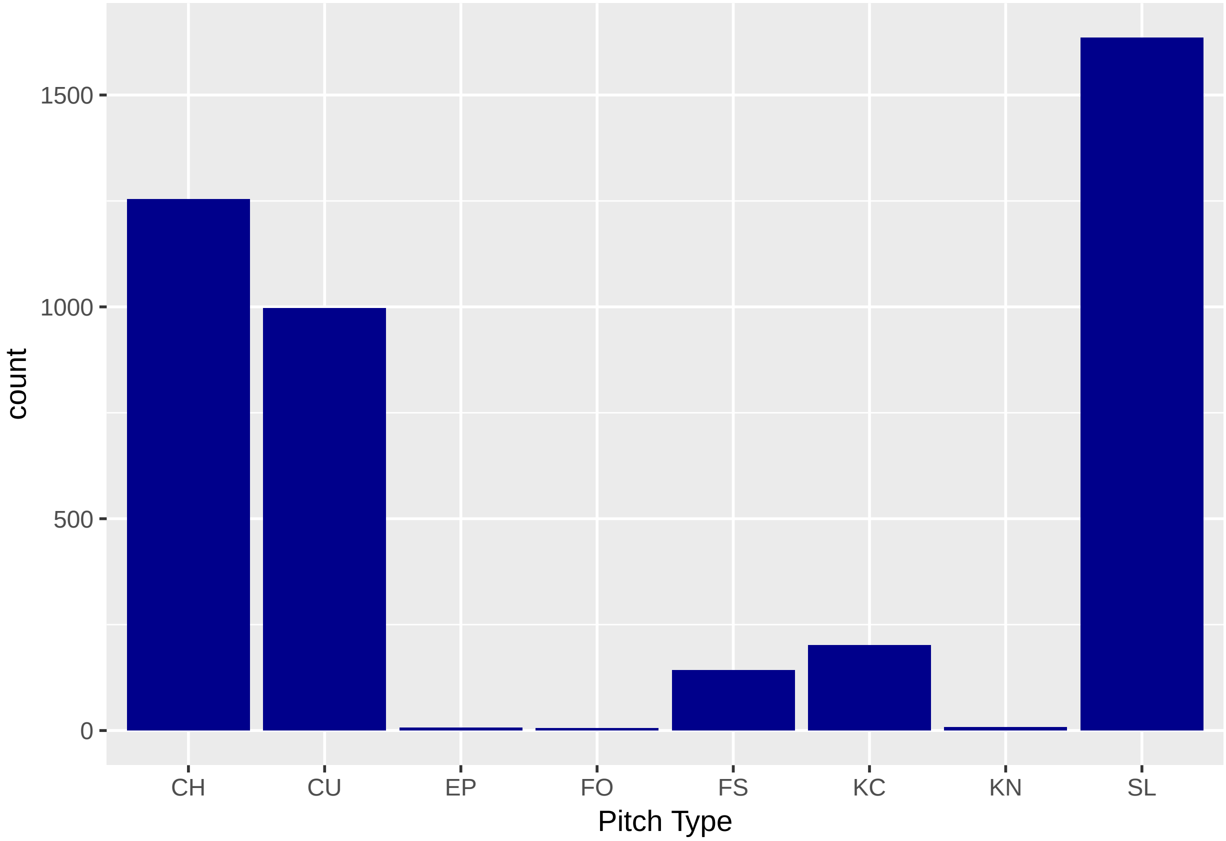
```
## Warning: Missing column names filled in: 'X1' [1]

## Warning in read_tokens_(data, tokenizer, col_specs, col_names, locale_, :
## length of NULL cannot be changed
## Warning in read_tokens_(data, tokenizer, col_specs, col_names, locale_, :
## length of NULL cannot be changed

## Warning in read_tokens_(data, tokenizer, col_specs, col_names, locale_, :
## length of NULL cannot be changed
```

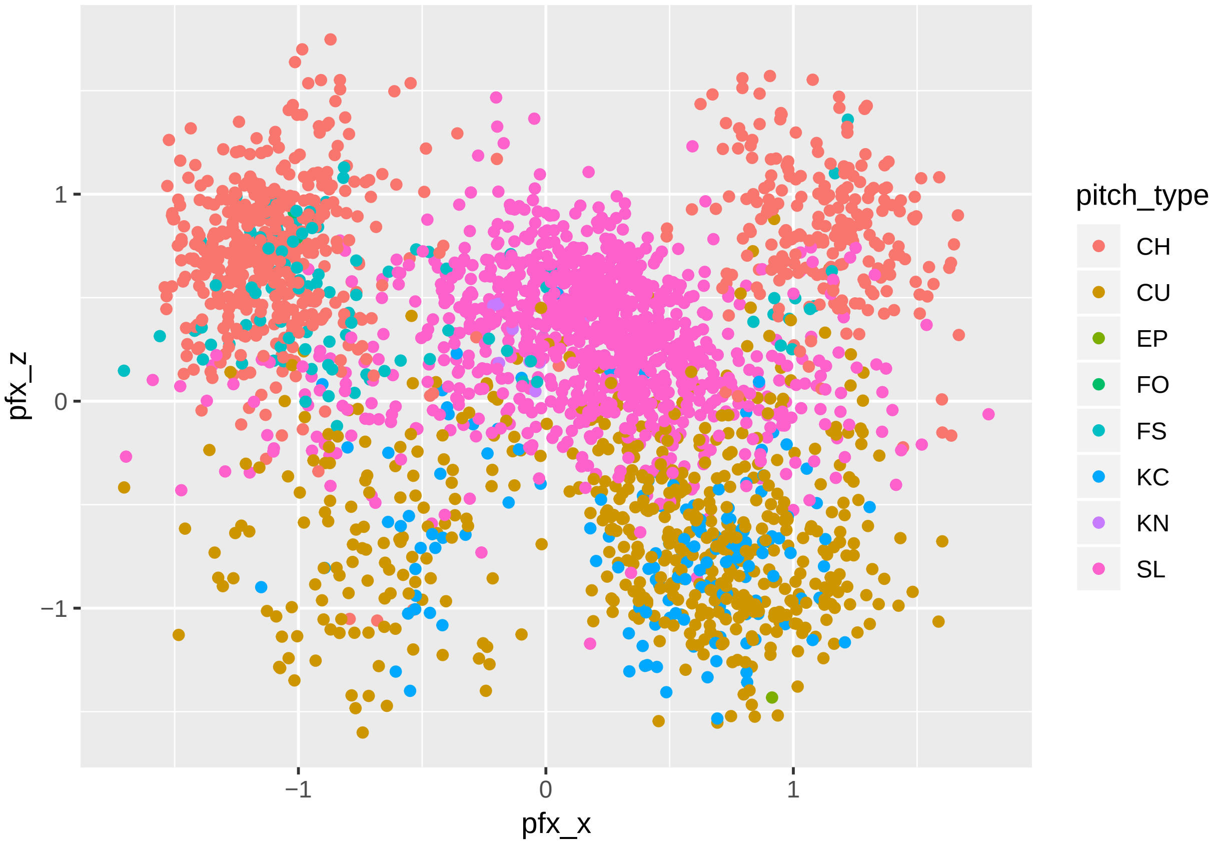
```
# rename our fastball velo column appropriately
colnames(df)[which(names(df) == "fb_diff")] <- "fb_velo"
```

```
ggplot(data.frame(df$pitch_type), aes(df$pitch_type)) + geom_bar(fill = "blue4") + labs(x = "Pitch Type
```



- If we're fitting a model it may just be best to leave the forkballs, knuckles, and ephuses out.
- How do these pitches look on a movement spectrum?

```
ggplot(filter(df, len > 150)) + geom_point(aes(x = pfx_x, y = pfx_z, color = pitch_type))
```

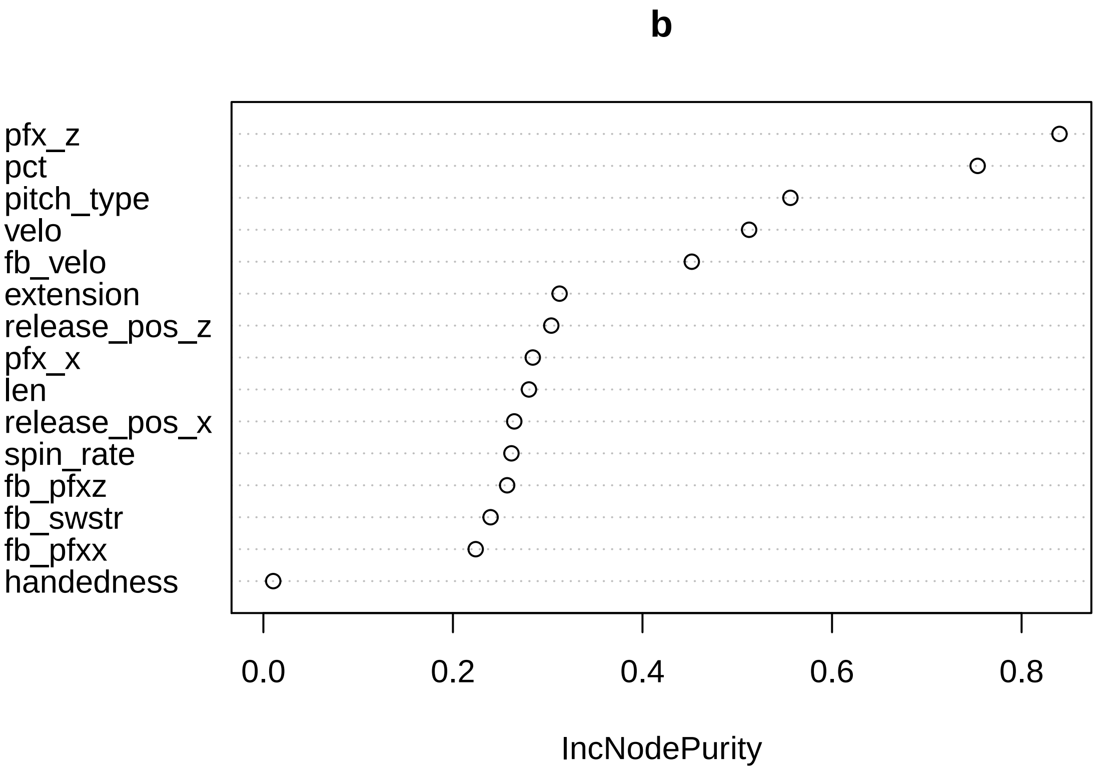


```r
# shrink our sample down to those who have thrown at least 150 of an offspeed pitch, none of the more o
samp <- df %>% filter(len >= 150, pitch_type != "KN", pitch_type != "FO", pitch_type != "EP")
```

```r
# random forest
samp$pitch_type <- as.factor(samp$pitch_type)
samp$handedness <- as.factor(samp$handedness)
b <- randomForest(formula = swstr ~ .,
                  data = samp[,-c(1, 3)], na.action = na.omit)
print(b)
```

```
##
## Call:
##  randomForest(formula = swstr ~ ., data = samp[, -c(1, 3)], na.action = na.omit)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##           Mean of squared residuals: 0.001188163
##                     % Var explained: 47.7
```

```r
varImpPlot(b)
```



- Our most important variable in this situation was vertical movement which isn't too surprising considering you want pitches that move up and down rather than those that travel more left-right along the barrel. Also not surprising was the fact that velocity, both of the offspeed pitch and average FB velo was incredibly important. I was curious to see that release height played such a factor, moreso than horizontal movement. It could be some sort of tuning due to extreme outliers in release point or more that horizontal movement doesn't tell us much at all for whiffs.
- Taking a general overview of our model, an R^2 of ~ 0.49/.48 tells us a strong deal given just pure "stuff" and not taking the context of each individual pitch. Of course there's a good deal of variation that we should not expect to be explained in the data I fed into the model. While I do give pct usage and total num of pitches thrown, I do not give any information about pitch location, usage by count, sequencing, batter handedness, leverage, so on and so forth.
- We've put together a decent model but it has a finite application. Given that swstr% is something that'll stabilize fairly quickly it doesn't necessarily help identify current MLB candidates that will see a jump in performance, we can simply look at MLB pitchers who generate a strong amount of whiffs on one or more misused offspeed pitches. This prediction is something that better serves identifying the minor league talent (in an ideal scenario where we have movement data on these pitches) who project to produce above average whiffs and therefore strikeouts. It'd greatly benefit our model to include swstr%'s for given players in our data set who have pitched in either the AA or AAA level. Unfortunately, while we can grab swstr% for pitchers, we do not have publically available pitch-by-pitch data with each pitch type tracked.