DSC 630 Milestone 3

Holly Erickson

Project Overview

Part One:

Using MLB data, use Kmeans to create pitch clusters. Find commonalities in the performance of the pitches in each cluster to offer insights to players so that they might improve their performance.

Part Two:

Predict the exit speed of a ball coming off of a bat using a random forest regressor. It had the best performance out of tested models.

Part One Steps

1. Split pitch data into pitcher handedness (right or left)
2. Use KMeans algorithm to divide pitches into clusters
   a. Determine optimal number for K
   b. Standardize data
   c. Divide pitches into clusters using pitch speed and movement data (pfx_x, pfx_z)
3. Perform data analysis on clusters to provide insights to pitcher based on their most common pitch clusters
   a. Determine zone location for each pitch based off of zone grid
   b. Find most common zone for key outcomes for each cluster: whiffs, soft contacts and ground balls

Part Two Steps

1. Feature engineering
    a. Find each stadium latitude, longitude
    b. Use API to get elevation for each stadium
    c. Use API to get historic weather for each pitch (temp, barometric pressure)
    d. Use parts a - c plus 9 features from original dataset to calculate the True Spin
2. Feature selection
    a. Combination of Pearson’s Corr, Kbest selector with f regression, random forest feature importance
3. Model selection - Random Forest Regressor
4. Hyperparameter tuning on cross validation set (next steps)
5. Train model and make predictions (next steps)

Challenges

My first major challenge was determining the optimal number of k. I looked into several techniques, including the elbow method, gap statistic and silhouette score.

The problem is that the data is very tightly grouped, so there are not easily identifiable clusters in the data and silhouette scores were negative.

In addition, the elbow method and gap statistic gave low numbers i.e. 3, but this did not split the pitches into enough clusters that I felt would represent the pitches well.

I decided to plot the clusters using a range for k to determine which k-value gave relatively tight coverage of the pitch movement without ‘micro-managing’ the data.

Challenges

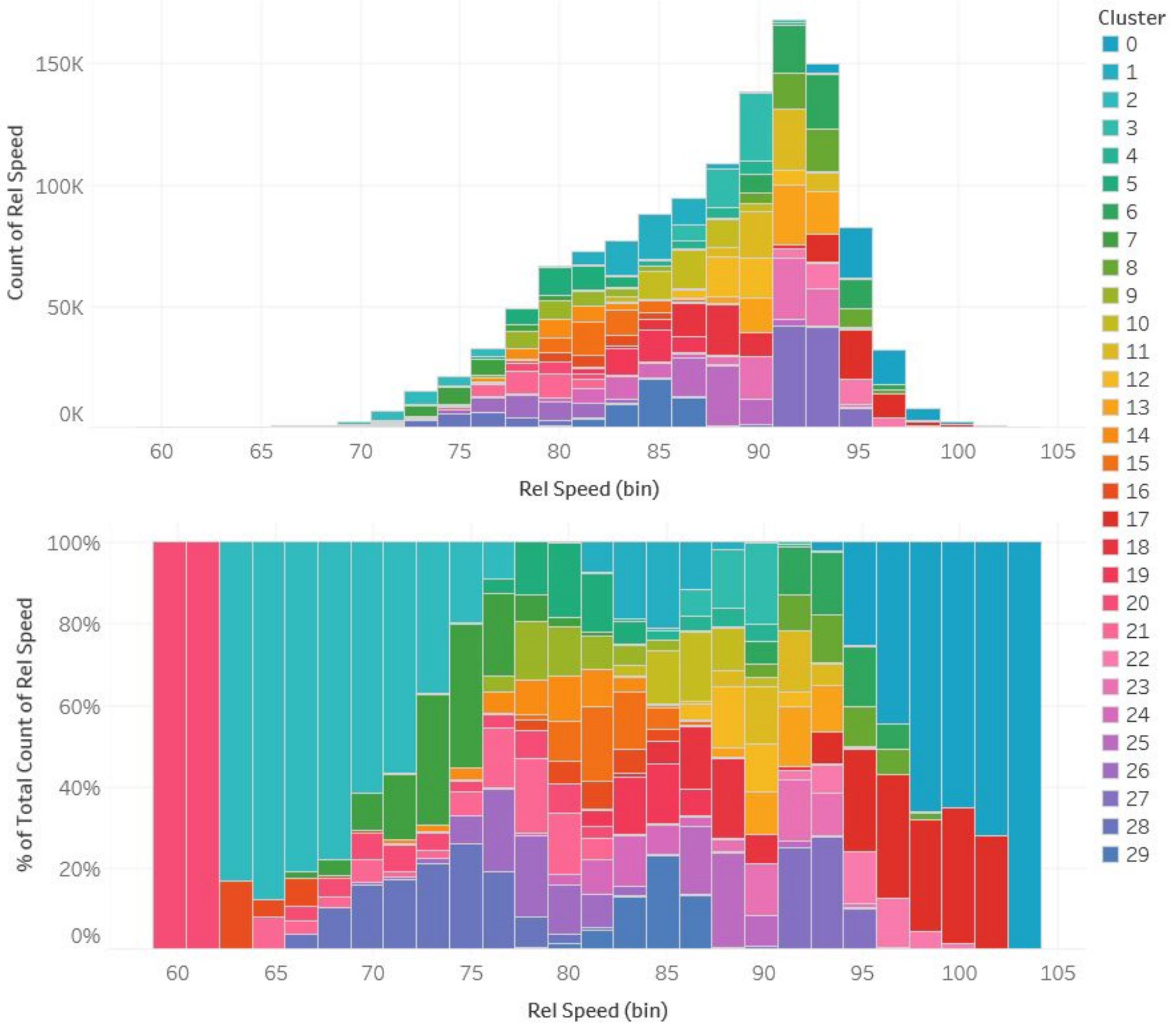
The second challenge is that there is a lot of overlap between pitch types. At first I was grouping pitches by handedness and again by pitch type before determining clusters. (So I would have a handful of clusters for each pitch type/handedness.)

I found that clusters were overlapping across pitch types, so I decided to run the clusters using all pitch types combined. For a batter, it doesn’t make a difference what the pitch is called, it just matters how it moves.

I decided on 30 clusters for left handed and 30 clusters for right handed pitches.

Left Pitchers Cluster Distribution

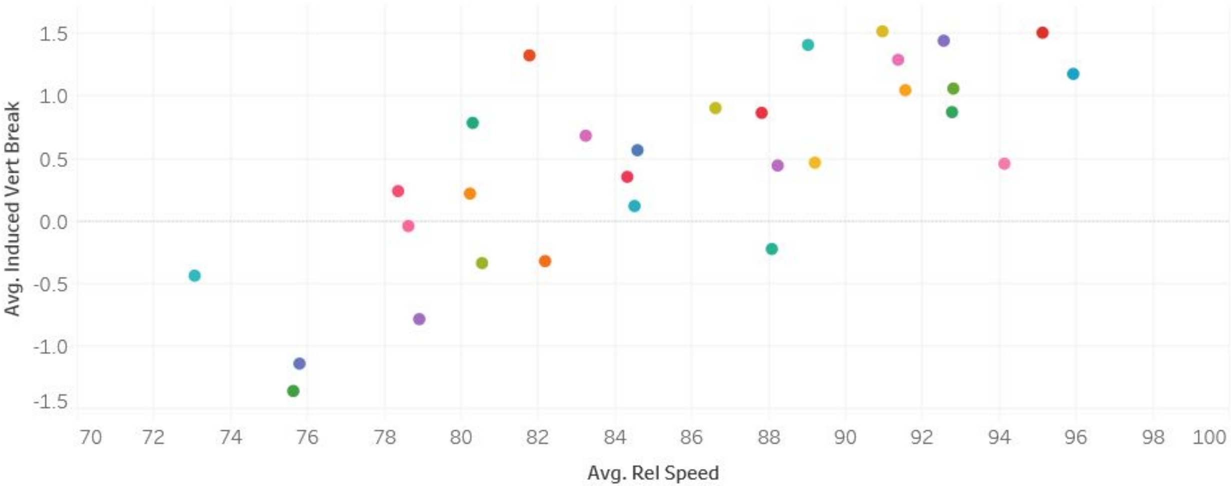
Release Speed



Left Pitchers Cluster Centers:

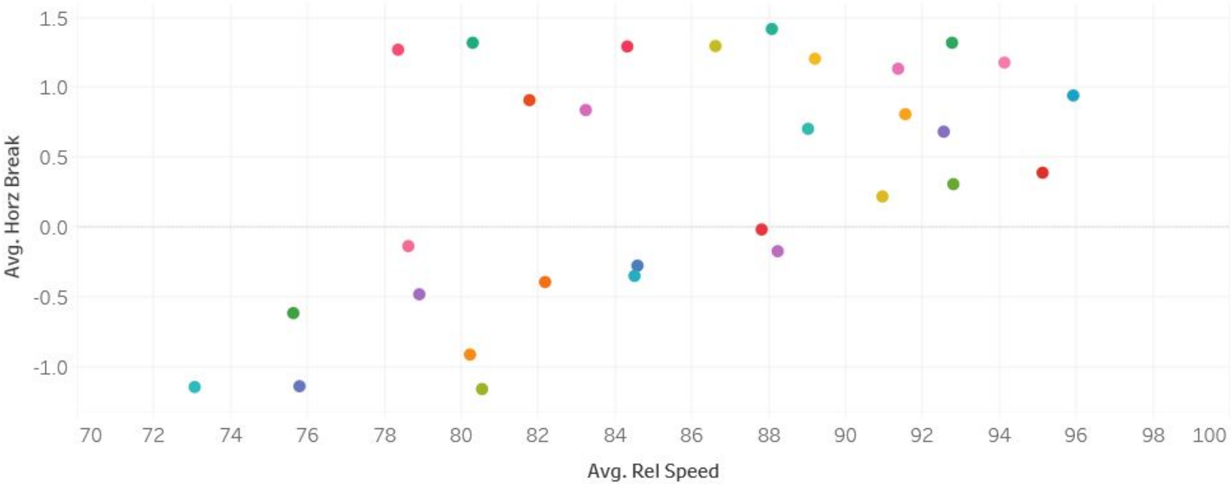
Induced Vertical Break

& Release Speed



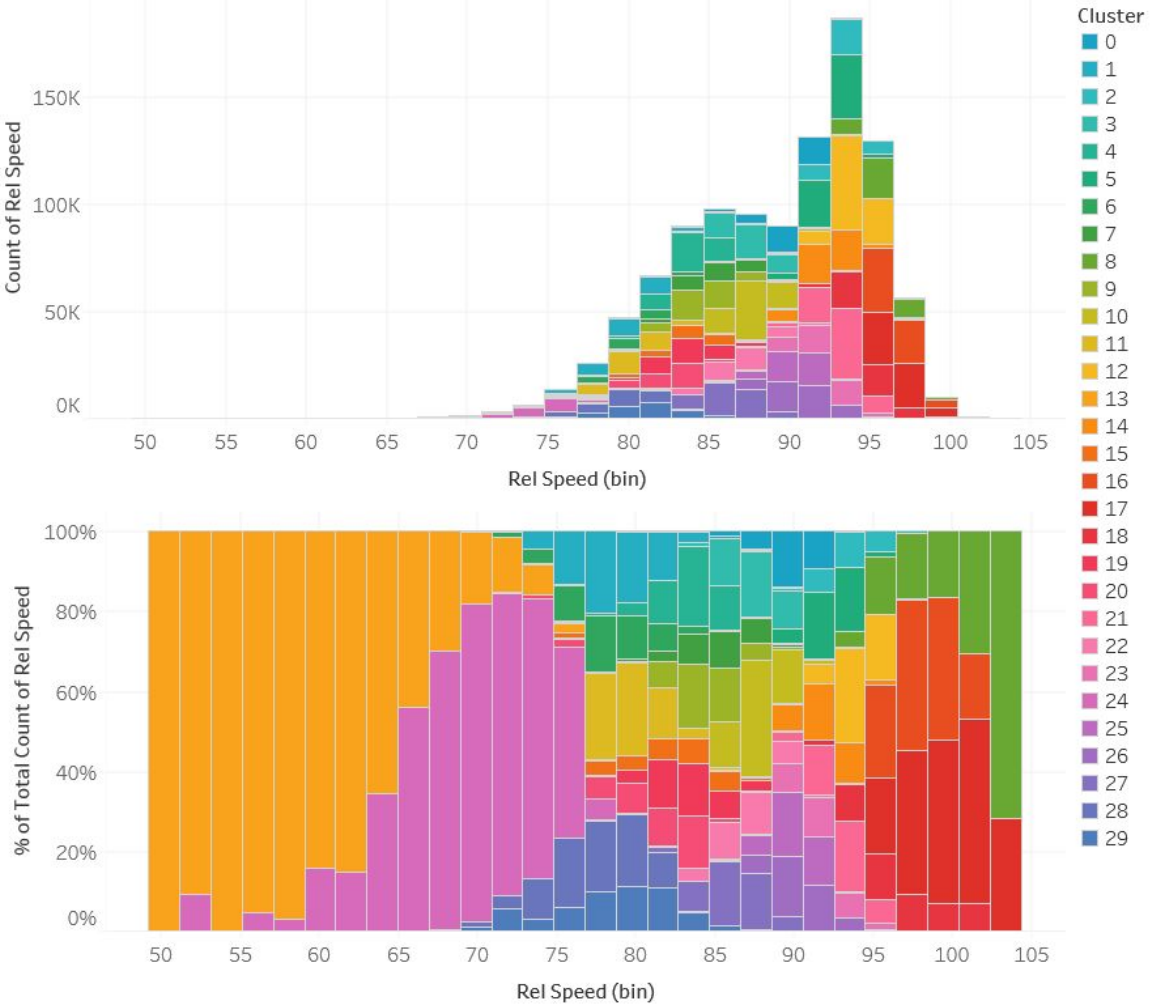
Horizontal Break

& Release Speed



Right Pitchers Cluster Distribution

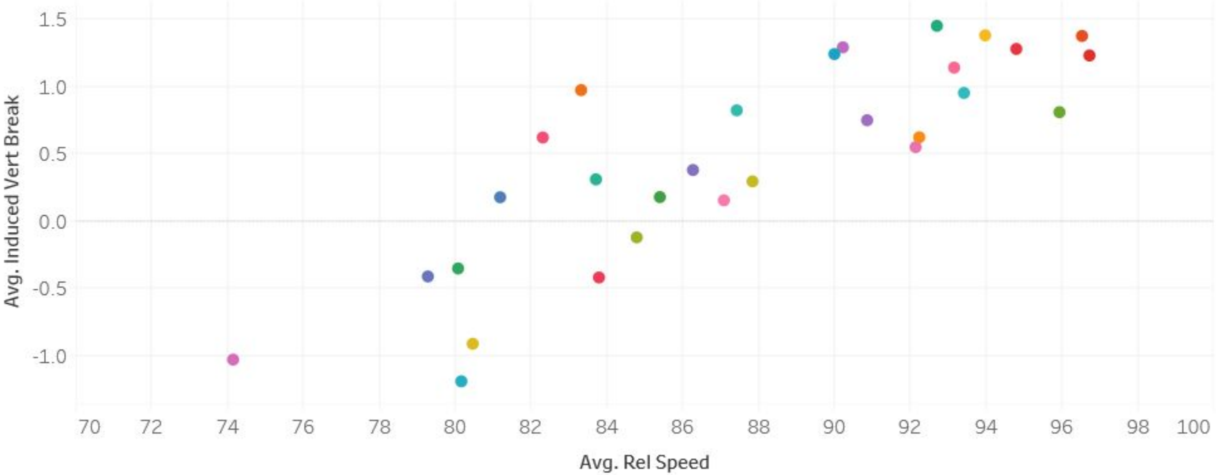
Release Speed



Right Pitchers Cluster Centers:

Induced Vertical Break

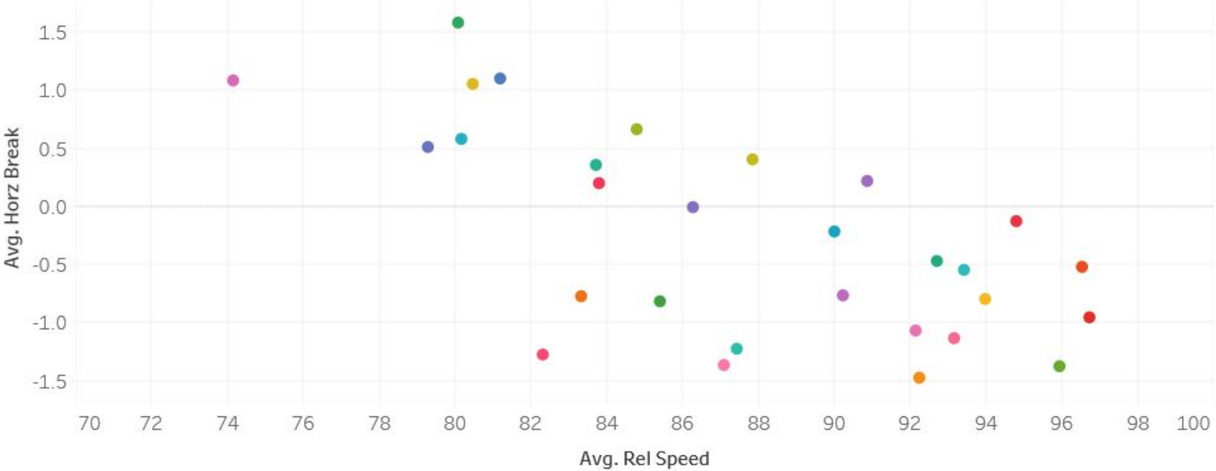
& Release Speed



Horizontal Break

& Release Speed

Notice the difference in horizontal break between right and left pitchers.


Avg. Horz Break
Avg. Rel Speed