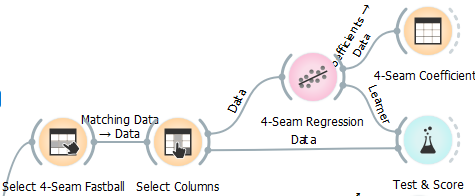
ASE – 370

Week 7 Final Project Write-Up

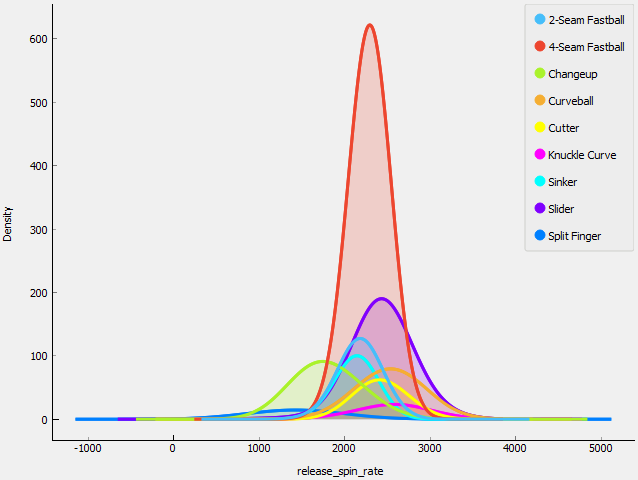
Jeremy Maschino

**Overview:** The goal of this project was to mine data on pitches from Baseball Savant, to come up with equation to use to predict the spin rate of a baseball from multivariate linear regression. I then would be implementing these equation into my Senior Capstone, to predict the spin rates of users. I used the Application Orange, as it allowed me to do data preprocessing, many different regressions at once, keep track of my workflow, and keep everything neat and organized. This goes against the grain of what other pitch trackers do as they actually measure the spin rate, but if proven effective, we could cut the cost down on producing such pitch trackers.

**Data and Model:** The data, as stated earlier is from MLB’s Baseball Savant. I actually had a couple different files to work on as the ball moves different ways, depending on what hand you throw it with. For the files, I had to do a few different preprocessing techniques to narrow the data down to what I needed. First, I did a feature reduction (this was done in MS Excel). I selected pitch\_name, release\_speed, release\_pos\_x, release\_pos\_z, plate\_x, plate\_z, and release\_spin\_rate. I then cleaned the data of any null values. I did this by removing the rows with null values. This should not hurt me as I have almost 40,000 records to work with for both my right handers data and left handers data. I then performed a two feature constructions, by creating horizontal\_break by the formula, horizontal\_break = release\_pos\_x – plate\_x, and created vertical\_break by the formula, vertical\_break = release\_pos\_z – plate\_z. Now for right handers, the horizontal break will be negative and for left handers the horizontal\_break will be positive. After this, I used orange to do a row selection, to select pitches that are a specific pitch type, then did an feature reduction to select only release\_speed, horiztonal\_break, and vertical\_break to be fed to my linear regression model, with my target release\_spin\_rate. I selected these features as I was constrained to them due to the outputs of my radar.

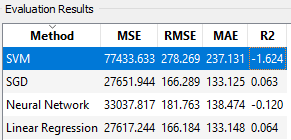


*Figure 1: Orange Workflow Example*



*Figure 2: Distributions of pitches by spin rate*

**Results:** Although, my results aren’t great. They are much better than if I were to use a different algorithm, when testing different algorithms, linear regression always performed better than the others that inputted numeric values.



*Table 1: Test Scores of Right-Handed 4-Seam Fastball*

But what I found more helpful, was when I would input similar pitches from different handed pitchers into their respective regression equations, I would get very similar results. For example, a pitch from a right-handed pitcher that has a velocity of 93 mph, a vertical break of 3 feet, and horizontal break of negative half a foot produces around the same spin rate as a left-handed pitcher that has the same velocity, same vertical break and a horizontal break of positive half a foot.

**Conclusion:** Due to the constraints I had to deal with, using regression may not be a viable solution for my current radar. But, if I were to go forward with marketing the radar setup and web app, I would contract someone out to built a custom radar that generates more inputs, similar to Trackman. But, for now it does work as a substitute for producing spin rates.