SPA DRP - Predicting Strikeout Percentage Among MLB Relief Pitchers - Luke VanHouten

For my SPA DRP, I focused on baseball pitching statistics. I began the project by reading parts from the book “Mathletics” by Winston et al. to get an even better understanding of sports statistics related to baseball, football, and basketball. After this, I began looking at research papers from the MIT Sloan Analytics Conference from the last few years, and I found a paper written by Eric Martin that I found quite interesting, titled “Predicting Major League Baseball Strikeout Rates from Differences in Velocity and Movement Among Player Pitch Types”. In the paper he went over how he wanted to predict strikeout percentage based on pitch data such as aggregated pitch speed and movement among different pitch types for starting pitchers. In baseball there are different types of pitches, such as a fastball or slider; however, an issue arises where some players may throw different pitches that still look the same to the batter. Martin’s solution was to cluster pitches based on pitch data rather than name of pitch by using Gaussian mixture models. These identify probabilistic regions to separate pitches into different categories, as opposed to k-means clustering, where these clusters are set beforehand. This allows for there to be a varying amount of clusters (still between 1 and 9) which lines up with pitchers having a set number to the types of pitches they physically throw, which varies from pitcher to pitcher. This variability allows for the pitch types to be captured with a distribution denoted by Gaussian ellipses, which represent this distribution. There will be some overlap here, so he uses Mahalanobis distances to identify how close a pitch is to a given distribution.

For my project I wanted to use techniques found in the paper to predict strikeout percentage at the pitcher level, as opposed to per pitch type while still using the pitch data such as velocity and movement. I sought to have clustering be just an extension of the prediction model as clustering is very computationally intensive and I was limited by time. I wanted to focus on predicting strikeout percentage for relief pitchers, which was a topic of a future question brought up in the original paper. For my data I created an extremely large database from MLB Statcast, which tracks every pitch thrown over the years. I had data from 2015 to 2019 as well as 2021, which was 4.4 million pitches. I created a stat to track the platoon advantage (the advantage the player has when the handedness of the batter is the same as that of the pitcher) of the first batter the relief pitcher faces in an appearance, and calculated established stats such as strikeout and strike percentages directly from the database. I also joined from a FanGraphs dataset leverage index data that measures the change in win probability for a given play, which can be useful to see how high the stakes are for when a relief pitcher enters the game. With these handedness and leverage stats I wanted to see if the contextual nature of relief pitcher appearances was a predictor for strikeout percentage, as managers may use this information to choose when the relief pitcher enters the game. I then implemented the clustering algorithm, but only to identify the number of pitch types each pitcher had, which was also done in the original paper. However, due to time and complexity I was unable to match Martin’s level of removing outliers, so I am unsure of how well the clustering turned out. I split my joined data into features and a label (SO%), and then split it into training and testing data, with the testing data being the 2021 year. I didn’t include the 2020 season mostly due to it being a shortened season due to COVID (so less injuries could impact play) and that it easily allowed me to split my data into 2015-2019 and 2021. I then loaded all of this into an XGBoost machine learning model to predict SO% while also doing hyperparameter tuning to reduce error. My final mean absolute error ended up being 4.6%, which isn’t that bad in the context of the 15-35% range for SO%. My most important features ended up being strike percentage, which makes sense as if you throw more strikes you are likely to throw more strikeouts, as well as the leverage index per game event, which is related to SO% as the pitcher may perform better or worse in high stakes situations. Velocity was also quite important, with a higher velocity being necessary to get past the batter so they can’t get a hit. My platoon advantage stat was not very important, nor was the amount of pitch types. In the future I would like to work more with the pitch clustering, particularly when it comes to dealing with outliers as well as performing all of my analysis within each cluster. I could also add pitch sequencing, which is the way these pitch types are ordered in order to prevent the batter from being able to predict which pitch is going to be thrown next. Overall I had a really fun time with this project and I believe that it taught me a great deal about baseball analytics as well as how to run an effective mode and gain key insights from it.