Predicting MLB Games - Midterm Report

Brian Won (bw436) and Jiacheng Wu (jw2545)

November 7, 2019

Our Data

Our current dataset only consists of 1615 games from the 2018 season, with each row representing a game. For each game, we have hitting and pitching stats for the two teams. We decided to use stats on a rolling window basis; for example, for hitting, we only look at how teams have been hitting over their previous 10 games. Our reasoning is that recent performance and trends are more predictive than past season stats for predicting the outcome of a single game. As of now, the only pitching info we have are the rolling window stats of the two starting pitchers of a game. We elected to represent stats of starting pitchers individually because they have a much greater effect on a game than their bullpen counterparts.

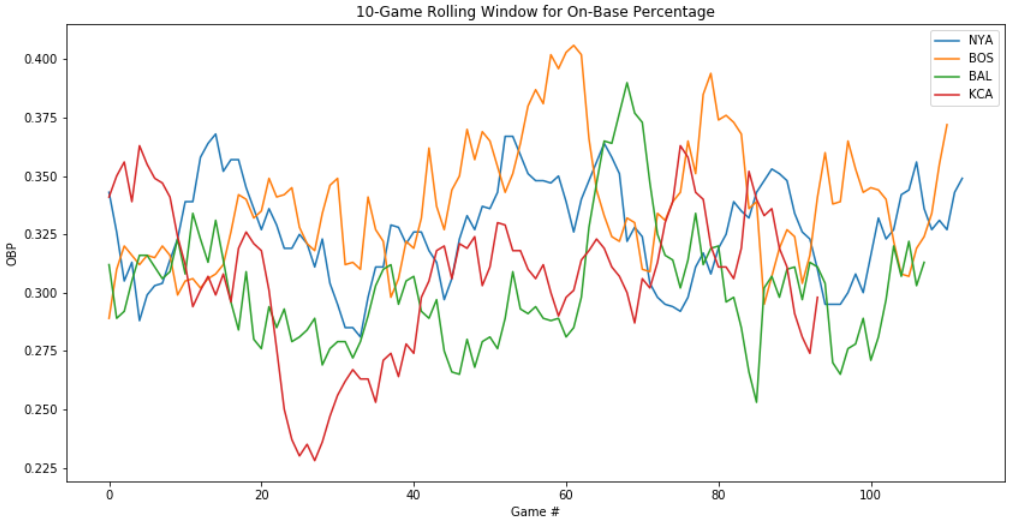
Initial Analysis

Class Balance

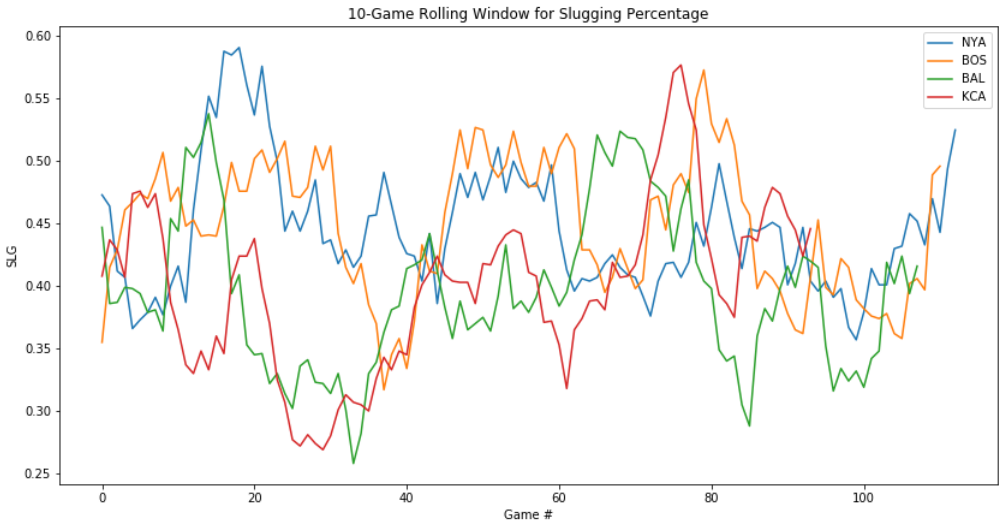
Of the 1615 games, the home team won about 53% of the time. It makes sense that home teams tend to win slightly more and it's great that we don't have to deal with much class imbalance for this classification problem.

Plots

Below is a plot for rolling on-base percentages (OBP) for a few teams:



Here we've plotted rolling OBP for the Yankees, Red Sox, Orioles, and Royals. OBP is simple: it's the number of times a hitter reaches base safely divided by total number of plate appearances. OBP is a "good" stat because it represents efforts towards a goal in baseball and only accounts for the skill of the featured player or team. The goal in baseball is to score runs. Teams score runs by getting people on base. The plot above shows that OBP can be used to differentiate good and bad teams. We can see that for most of the 2018 season, the Yankees and Red Sox (blue and orange) had better rolling OBP than those of the Orioles and Royals (green and red). The "good" stat criteria is the reason that we didn't include traditional stats such as runs batted in or RBI. Granted, RBI directly represents scoring runs, but it doesn't only account for the skill of the featured hitter. For example, a hitter can have inflated RBI numbers if his teammates who hit ahead of him in the lineup get on base often, or if they're fast runners so they can score from second on a mere single.



The plot above shows that slugging percentage (SLG) can also be a good stat to differentiate teams.

Modeling

We ran a Logistic Regression (LR) and a Random Forest Classifier (RF) on our current dataset. The resulting ROC-AUC scores are as follows:

- LR (train) : 0.595
- LR (test) : 0.532
- RF (train) : 0.923
- RF (test) : 0.536

Moving Forward

Looking at the results above, logistic regression underfits the data. To tend to this, changes (or rather additions) should be made to our dataset. Firstly, we definitely will incorporate games from other seasons. Putting together the data for the 2018 season was a significant coding task, but now we can easily extend that to other

seasons. As mentioned, our dataset currently only includes team hitting stats and stats of the two starting pitchers. We plan on adding stats of relief pitchers. One aspect in regards to relief pitchers that we will have to keep in mind is that they don't pitch in every game. Another change we can make is to include other "types" of data. Thanks to modern technology, baseball has Statcast which keeps track of literal ball movement; not only pitch velocity, but also how much a pitch moves. Maybe pitchers with greater pitch movement are significantly better. We can also look at launch angle and velocity of the bat; it makes sense that hitters who hit the ball more square more often find more success. At the end of the day, we recognize the limits of our project. We most likely won't achieve a "perfect" model, given the complex facets and randomness that are a part of baseball.