ATLS 5214 Baseball Data Project Idea

Wil Boshell

1 Introduction

For this course, I want to make my project be on a topic that interests me. I have been a baseball fan since I taught myself the game listening to radio broadcasts when I was 6 years old. The numbers aspect of the game has always appealed to me and I would love to take a much deeper dive into the plethora of available data. Baseball data can be used for a variety of purposes. I would personally like to do player valuation and game/season projections. While some models for this already exist and could be implemented very quickly, it would be valuable to explore the data ourselves. However, existing metrics (such as the ones on Fangraphs) are a very helpful starting place and some could be incorporated into the final model. Machine Learning algorithms will be very useful for this entire project. Developing and tweaking the model on our own is likely enough to provide work for the entire semester.

My first thought is to gather all MLB data from the 2000-19 seasons to explore, though the additions of PITCHf/x in 2007 and its replacement Statcast in 2015 both add valuable insight missing from previous years. Statcast especially provides a lot of potential data to explore and develop our model with. If possible, minor league data from AAA and AA could be helpful to create projections for younger players. I hope to update the predictions at least after each game once the regular season starts on March 26, so the ability to feed data into the model and update it quickly is important. However, the model can always be tweaked after this time. Multiple models could also be used from the outset and the first part of the season can help determine which model is better. It would also be good if we could put some projections on a website once we get the model working. I personally plan to continue examining the models for the entire season, even after the semester ends because this is very interesting to me.

2 Preliminary Timeline

2.1 Immediately

Find a team and begin working out logistics. We should also begin developing a basic model. It also might be useful to see if any Spring Training data is useful for predicting regular season performance. While most spring training stats do not carry over once the regular season begins, it is worth exploring to see if there is any indicator of a breakout or step back.

2.2 February 21: Spring Training Begins

Only one game is on the 21st, most teams start on the 22nd. Especially early on, the major league players will not play much and there is a level of competition issue from the stats we do get. However, this is still a good time to test a data collection system while we work on our models. Additionally, the HackCU Hackathon is February 22-23, which might make a good time to work on some of this project.

2.3 March 26: Regular Season Begins

All 30 teams will begin the regular season on March 26, so it is very important to have our data collection system working perfectly by then. It is also important to have at least one model make predictions before this. There will be games every day, so we can see how well the predictions are working as we go along.

2.4 April 29: Last Day of Class

I do not know when exactly the presentations are, but 5 weeks of real data is enough to perform a preliminary analysis of the projections, although small sample size nuances still exist.

3 Conclusion

This is a very preliminary outline of the project as there are a number of different ways to determine player value, which in turn, will be used to create team performance projections. I am very interested to see where this can go!