Holly Erickson

DSC 630 Course Project

Milestone 1

Introduction

Background

I work in the baseball field and a lot of analytical work tends to center around a pitcher's performance, and how the intrinsic characteristics of a pitch affect the outcome. I have been studying insights for pitchers but would like to explore insights for hitters as well. This means I will be focusing on the variables in the data that relate to the hitter, as well as some characteristics of a pitch that a hitter could identify.

Problem Statement

Predict a hitter's outcome based on things a hitter is controlling, such as launch angle and exit speed, to see if we can determine a "sweet spot" for achieving various outcomes based on the profile of a pitch that was thrown.

Scope

I believe I will be using statcast data from MLB games for the years 2008 - 2019. The statcast system currently uses Trackman, a radar measuring device that uses a single black square elevated behind home plate (usually mounted hanging off of the second deck or some place like that). It measures the flight path of the ball, including spin and

movement, and it also measures exit velocity, exit spin, and the first 300 ft or so of the exit flight path.

Document Overview

This document contains a discussion of preliminary requirements for the project and the plan to deploy and evaluate my model. I will outline the results I think might be possible, as well as my plan to manage this project.

Preliminary Requirements

The MLB data is required, either the statcast Trackman data which I have previously described and is preferred, or the PITCHf/x data would also be ok. PITCHf/x is an optical measuring device that uses three high-speed cameras mounted in various places around the stadium to triangulate the path of the ball. From the flight path, the PITCHf/x system is able to go back and calculate things like movement and spin rate, but they are a calculated estimation, not measured. I am using some data that I received from a friend in the business and I'm not sure where he got the data from. I will need to clarify this with him.

For technical requirements, I am using my gaming laptop, an MSI GS 65 Stealth. It has 6 cores, and the 9th Gen Intel® Core™ i7 processor has 10% performance gains over the previous generation. It has 32 GB of Ram.

If it is taking too long for my models to converge, I can try using less years of MLB data. I arbitrarily chose 2008 - 2019 because that is what was sent to me, but I could use data from 2010 onward, for instance.

Technical Approach

Analysis

I will describe what decisions and actions I took to address data quality problems. I will also need to consider any transformations of the data made for cleaning purposes and their possible impact on the analysis results.

I will do some initial exploratory data analysis, such as the distribution of key attributes as well as relationships between pairs or small numbers of attributes. I will potentially define new features and determine which key features to use in my model.

My analysis will consider several models so that I can evaluate and choose the one with the best fit for the data. I will select the specific modeling techniques for each model, and generate a procedure to test the model's quality and validity. I will assess the model's predicted results and evaluate the outcomes against the true target values.

Requirement Development

The initial work to be done will include selecting the models to use and finetuning the parameter settings. I will split my data into three sets. 1. Training set. 2. Cross-validation

set to fine-tune my model. 3. Test set to evaluate the final performance of my model. I could also use a one-vs-rest technique for hyperparameter tuning, which would allow me to combine the training and cross-validation sets.

Model Deployment

In the deployment stage, you'll typically take your evaluation results and determine a strategy for their deployment. If a general procedure has been identified to create the relevant model(s), this procedure is documented for later deployment. This will also include a plan to maintain the model going forward as new data is made available to train your model on. If this was a project for a company (or in my case baseball team), this is where predictive analytics helps to improve the operational side by having an impact in the real world.

Testing and Evaluation

I will use error rates for predicted versus actual target variables as quality measures for my models. Therefore, I will separate the dataset into train and test sets, build the model on the train set, and estimate its quality on the separate test set.

My plan is to summarize the results by listing the qualities of my generated models (e.g.in terms of accuracy) and rank their quality in relation to each other. I will revise parameter settings and tune them for the next modeling run in order to find the best model. I will document all such revisions and assessments.

Expected Results

This is tough. The success criteria for a successful outcome to the project in technical terms, for example, a certain level of predictive accuracy would be an uneducated guess on my part. I do expect that I will be able to identify the model which performs the best out of several options, as well as finding an improved performance through the use of hyperparameter tuning.

Management Approach

Project Plan

By week 5 I need to complete the Project Milestone 2. Between now and then I will need to put together the final format of my paper:

- Abstract
- Intro/background of the problem
- Methods
- Results
- Discussion/conclusion
- Acknowledgments
- References

While some things will change between week 5 and the end of the semester, I should be able to complete some of the intro/background of the problem, methods, preliminary results, and discussion.

By week 9, I will complete Milestone 3. I will prepare a deck of slides that describes the steps of the analysis up to that point. I will discuss the issues, challenges encountered during the work, and future plans to complete the analysis. The intermediate results of the project will also be presented.

Finally, by week 12 I will complete Milestone 4 - the Final Project Paper and Presentation, which describes the results of the analytics project and use of the concepts and methods taught throughout the course. My presentation will be recorded at 15-20 minutes in length. My final paper will include an executive summary, describing the conclusions of the data analysis to a non-technical audience. The technical report will include an intro/background of the problem, methods, results, conclusion, and references.

Project Risk

The events that might delay the project are difficulty getting models to converge in a timely manner. My contingency plan to reduce this risk includes reducing the size of the dataset (utilizing fewer years of MLB data), or potentially reducing the complexity of my models.