

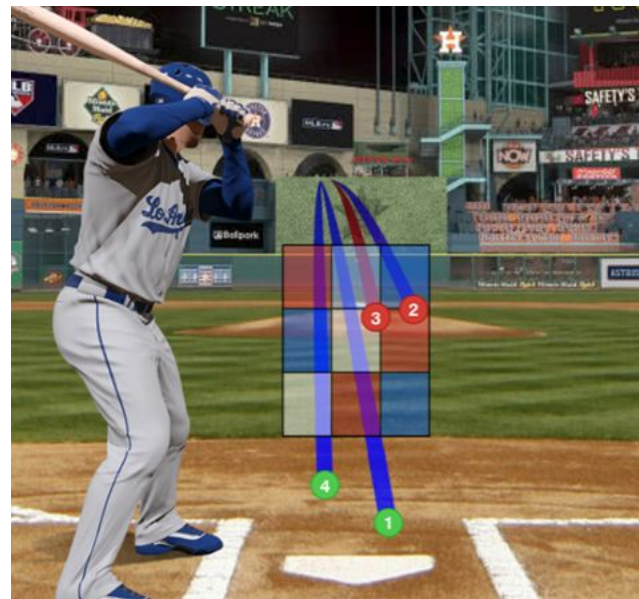
Baseball Pitching Predictor

Sprint 2

Yu Song

Overall of the Project

- Predicting pitch type and location.
- Use RNN model to capture the pattern
- Improve batting performance



Introduction to the dataset


Game data from 2015 to 2019 season, pitch by pitch in sequence.

- 3.5 million pitches during the time span

Target Variable

- Zone : 1-9 (strike), 11-14 (ball)
- Pitch type (fast ,break, off-speed , other)

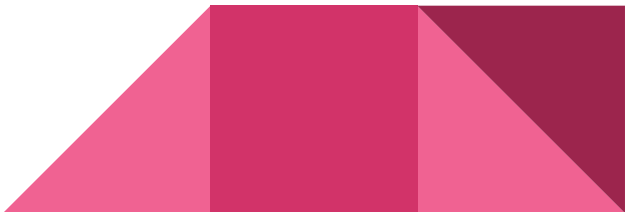
Features

- Pitcher/batter info
 - Outcome of each at-bat
 - In-game information
- 

Data Cleaning

- Removing pitcher's entire game script if data is missing.
- Remove unnecessary columns

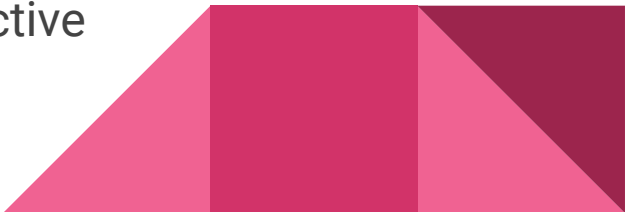
Feature Engineering

- Pitch Count
 - Previous info: at-bat result, pitch type, pitch result etc.
 - Count/Run Difference
 - Weather
- 

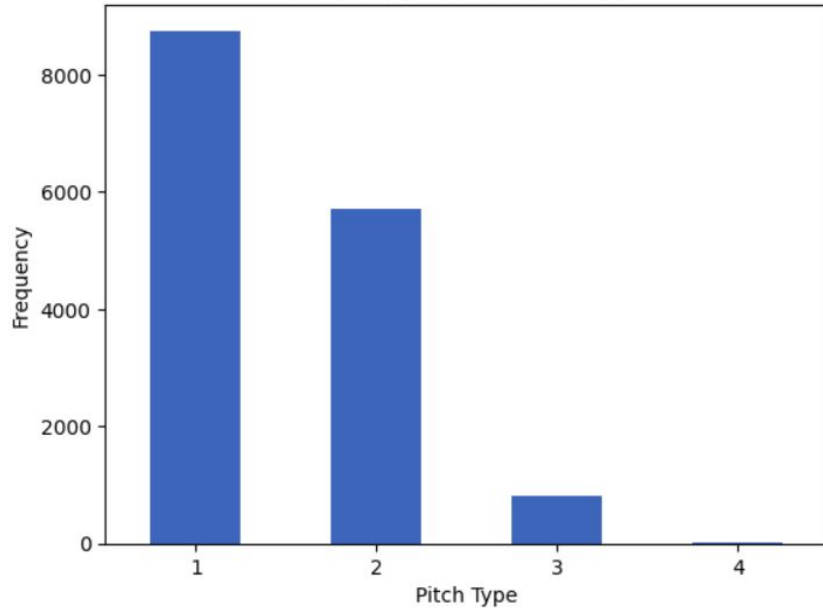
Baseline Model - RNN on Justin Valender

- Capture pitch sequence by splitting dataframe into multiple 3-row sequential dataframes
- RNN layer then separate layers for two target variables
- Softmax layer as output layer to predict 13 classes for zone and 4 classes for pitch type

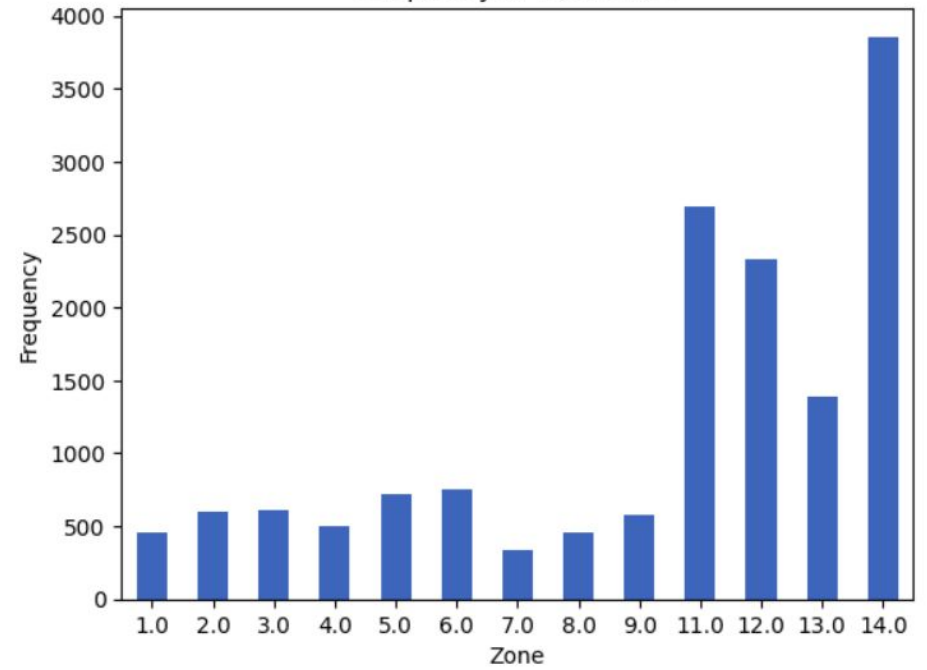
Issue:

- Imbalanced class for both target variables
 - Prediction shows identical vectors in tensor, no predictive power
- 

Frequency of Each Pitch Type



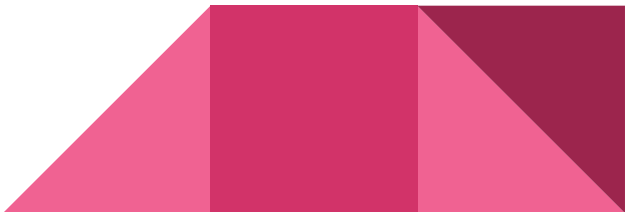
Frequency of Each Zone



Baseline Model - Softmax Regression on Justin Valender

- Predicting multiple classes in a logistics regression model.
- Use the previous pitch/event result info as “recurrent” features.

Issue:

- No predictive power on pitch zone
 - 55% accuracy on pitch type, but only two types in the prediction
 - Imbalanced class
- 

Next Step

- Fine tuning and determining the appropriate model by accuracy and loss
- Deal with class imbalance
- Productizing





Thank you!