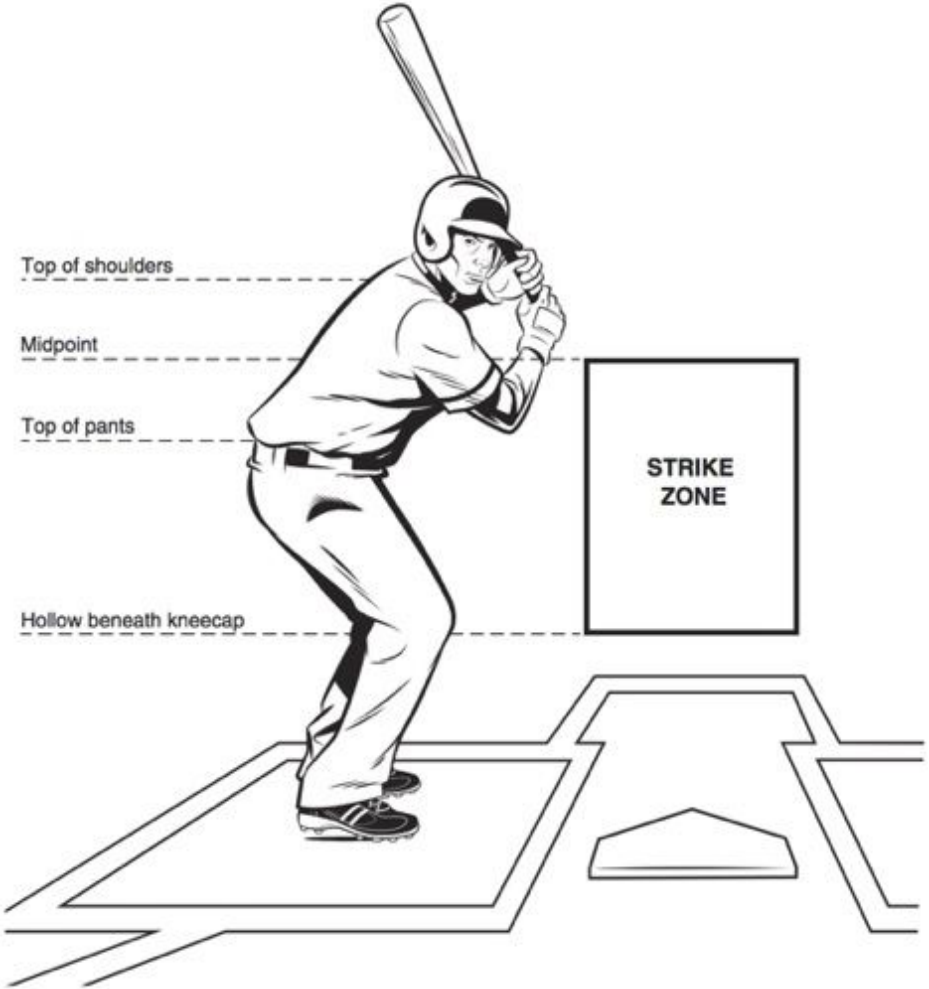
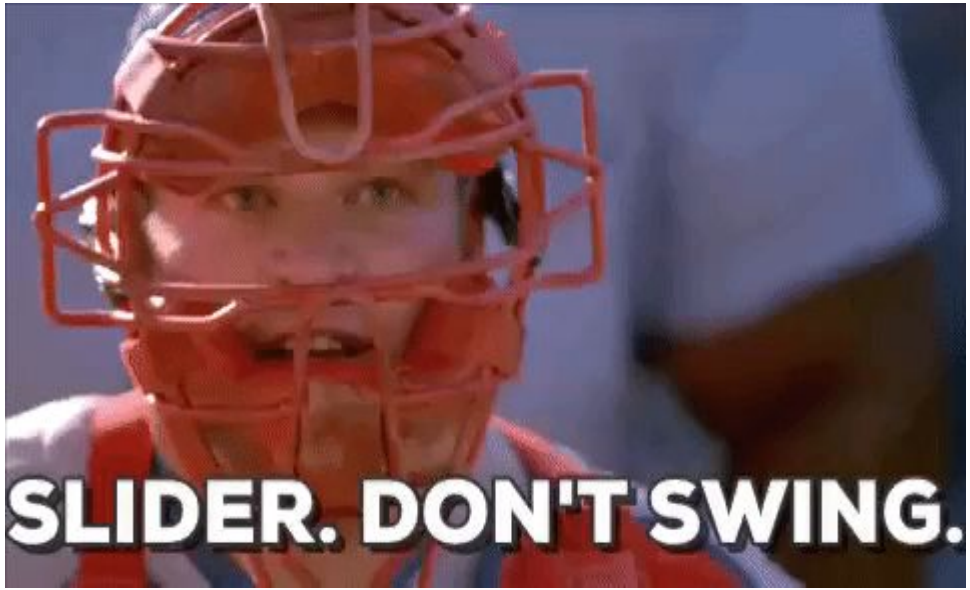
Swing or Take?
Predicting MLB Strikes

Michael Jordan
February 12, 2020



Background

- **Strike Zone**: Area over home plate between batter's knees and midpoint of torso



Challenge

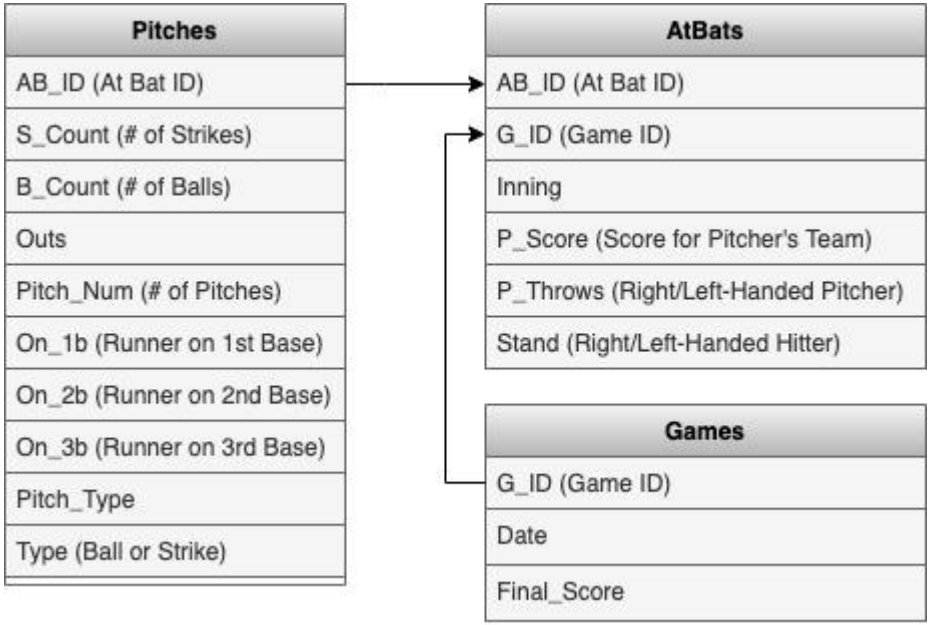
- Pitches within the strike zone are easier to hit
- Swing at strikes, don’t swing at balls..

Objectives

- Classification model to predict strikes for counts < 2 strikes
  - Optimize chance of hitter making contact
- Precision as success metric
- Feature importance

Methodology

Data



Tools

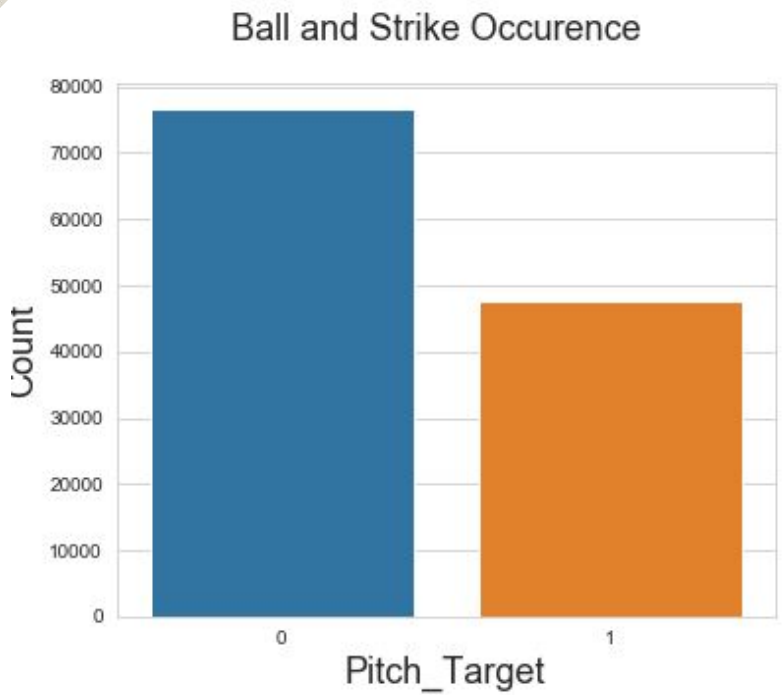
- Google Cloud
- PostgreSQL
- Tableau

Features

1. Strike/Ball Count
2. Number of Outs
3. “Fatigue Factor”
4. Runners on Base
5. Score (Pitcher’s Team)
6. Pitch Type
7. Pitcher/Hitter Position

Target

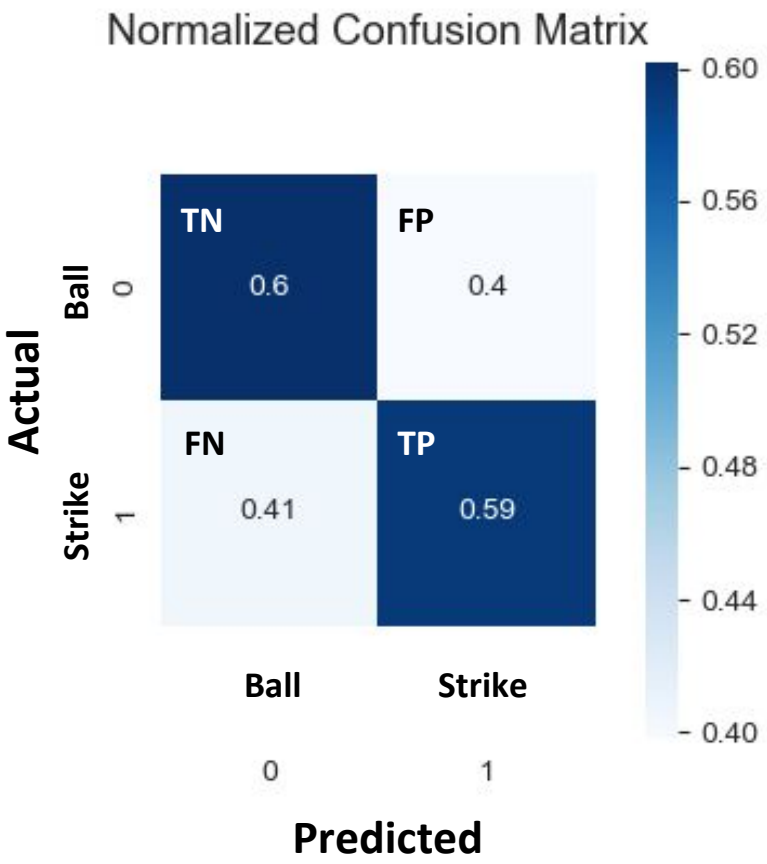
38% Strikes



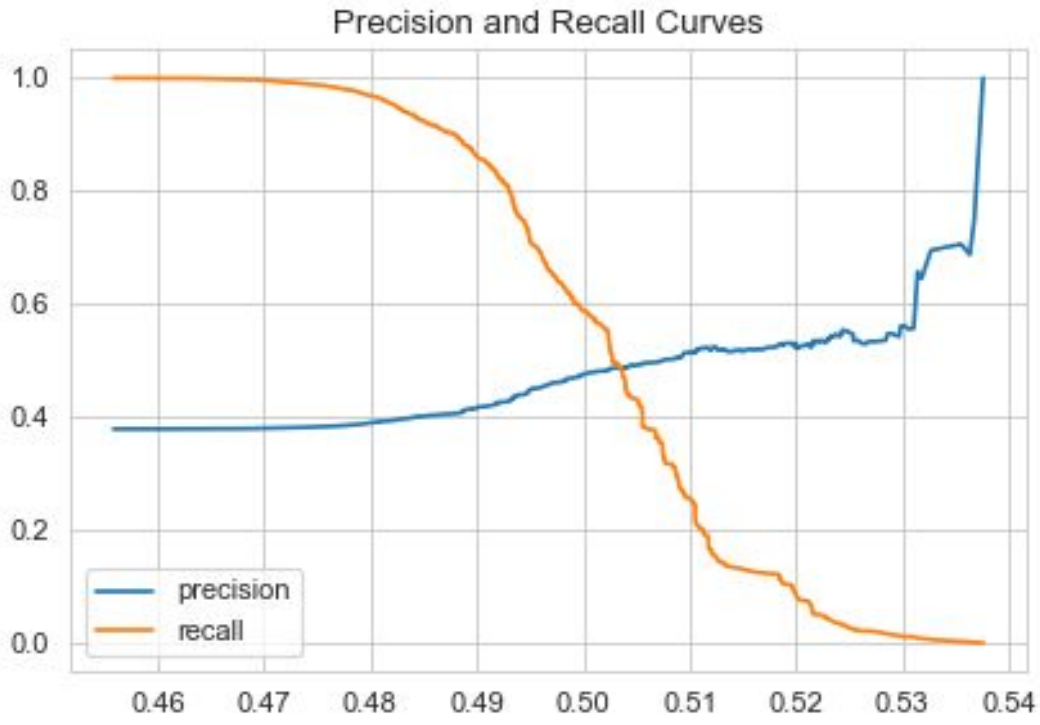
Model Selection

	Model	Train_Accuracy	Test_Accuracy	Precision
1	KNN GSCV	0.563	0.563	0.423
0	Logistic Regression GSCV	0.583	0.581	0.472
3	Decision Tree GSCV	0.585	0.584	0.473
2	Naive Bayes	0.590	0.588	0.476
4	Random Forest RS	0.594	0.599	0.485

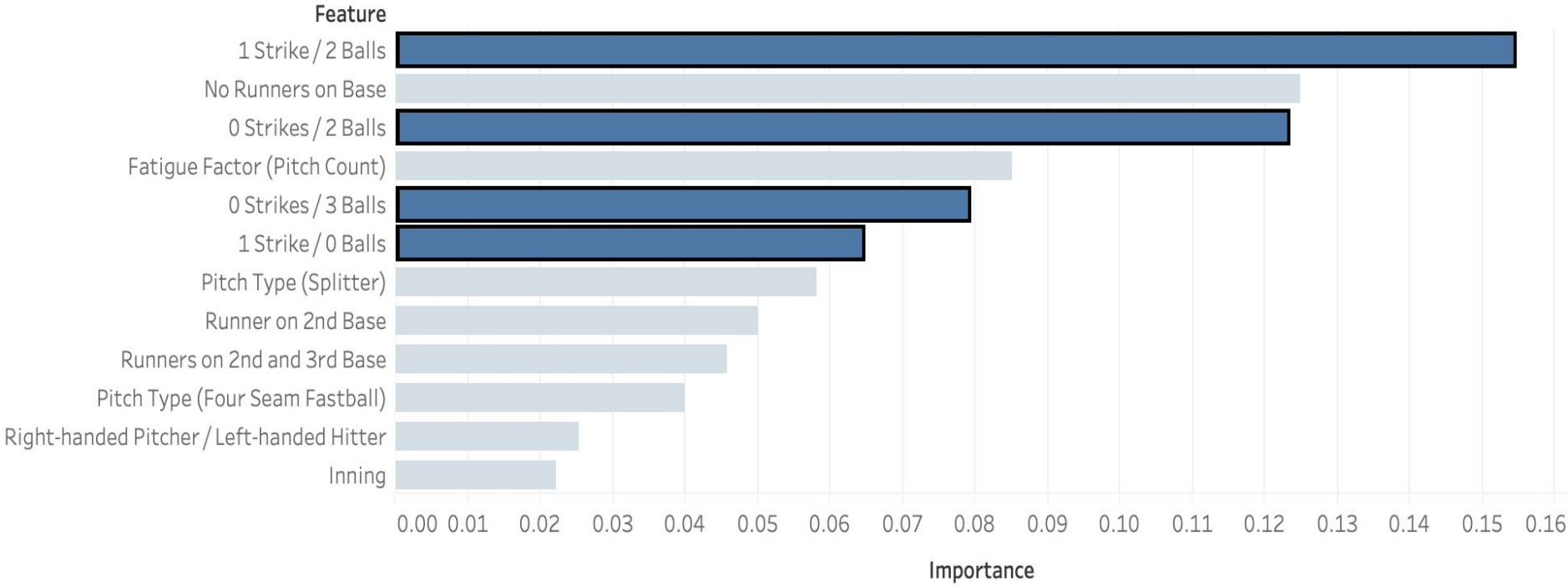
	Model	Holdout Accuracy	Holdout Precision
0	Random Forest RS	0.598	0.475


Normalized Confusion Matrix
TN
0.6
FP
0.4
FN
0.41
TP
0.59
Ball
Strike
0
1
Actual
Ball
Strike
0
1
Predicted
0.60
0.56
0.52
0.48
0.44
0.40

Precision/Recall Curve



Top 12 Features



Conclusions

- Hitters should pay attention to:
  - # of balls/strikes in the count
  - Runners on base
  - Total pitch count (fatigue)

Future Work

- Xgboost
- Look at subset of pitchers/players
- Explore distributions of features further
- Make model usable:
  - Flask App for interactive predictions


THANK YOU

Appendix

- Kaggle. MLB Pitch Data. (2015 - 2018). [Data file]. Retrieved from https://www.kaggle.com/pschale/mlb-pitch-data-20152018#games.csv