Keep your Eye on the Ball: Understanding the Pitches and Payoffs of the MLB

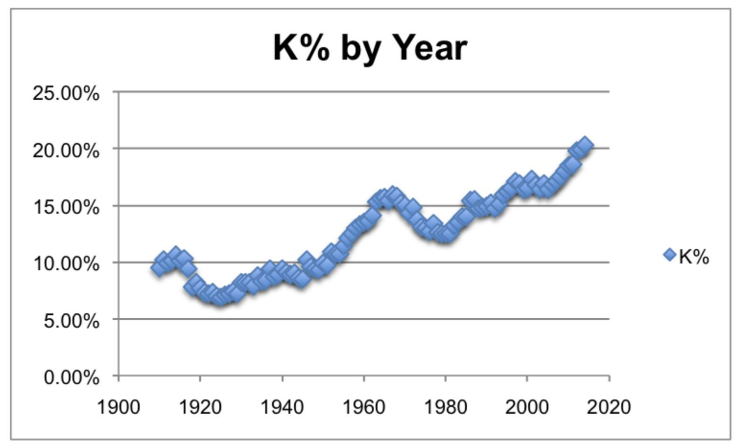
PRINCETON UNIVERSITY

Ami Berman, Ally Dalman, David Major, Uri Schwartz, Princeton University

Abstract

- **Data analytics have been proven to be an effective tool in baseball strategies**
- **Sabermetrics in baseball are a growing field of interest.**
- **Challenge: Can we use machine learning to predict the trajectory of a game given its current situation? How can a manager use this information to win games?**

Background and Related Work



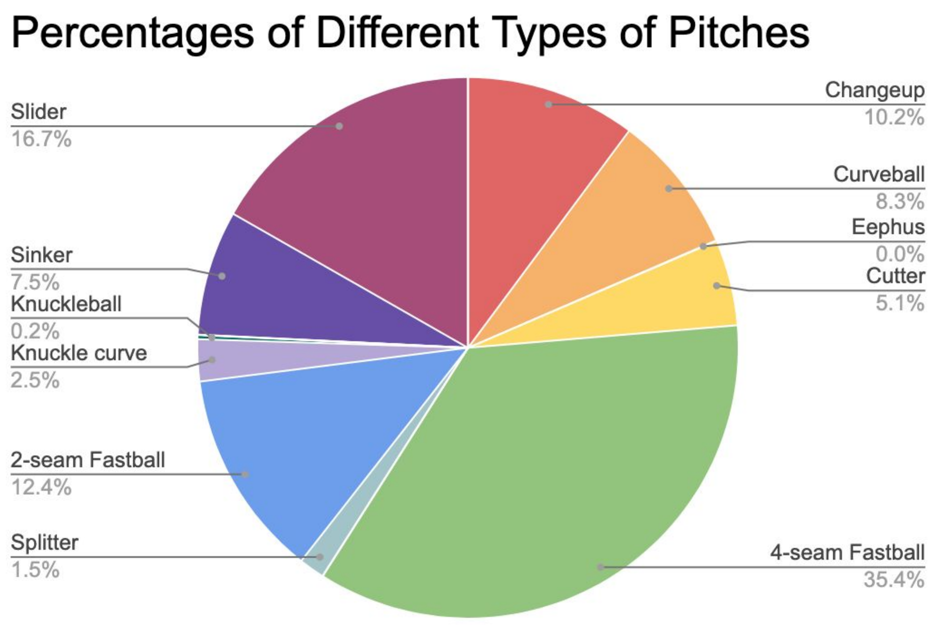
- **How can teams maximize reward for the batters and the pitchers?**
- **There has been a huge growth in data analytics for baseball [3]**
- **Past studies have analyzed specific teams and players, what can we understand more generally?**
- **Teams spend millions on data analytics but can just SciKit-Learn analysis yield useful insights for them?**

Data Processing and Approach

- **MLB Pitch Data from 2015-2018**
- **Pitch data** includes information about speed, location, balls and strikes, and type of pitch
- **At-bat data** includes teams, players, score, inning, outcome, players on base, etc.
- **Game data** includes score, teams, umpires, weather, time, venue, wind, delay, etc
- **Ejection data** includes umpires, player, time, teams, etc.
- **Used unsupervised learning to reveal latent structure of the data and supervised learning to identify correlated features with useful outcomes**

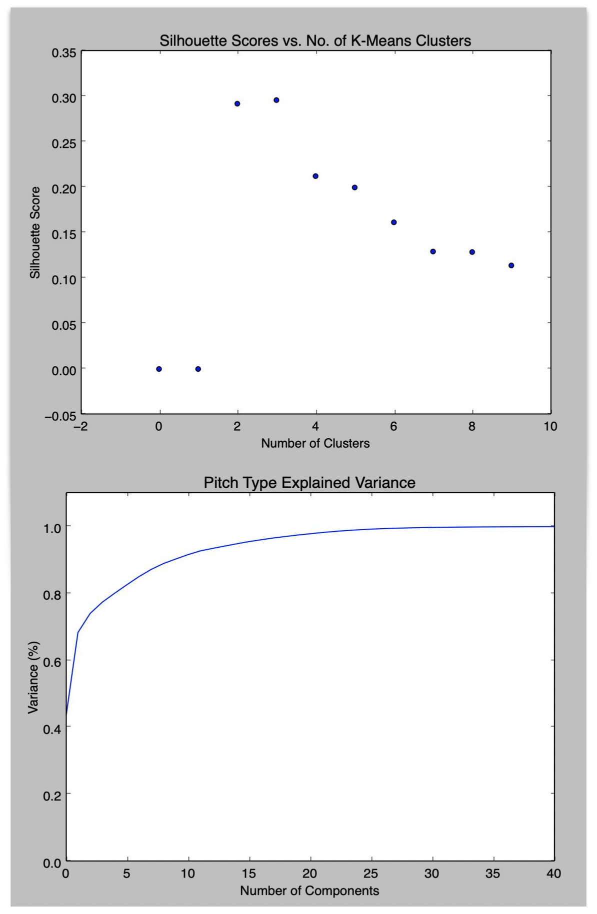
Basic Analysis

- **Not one type of pitch dominates the MLB in terms of frequency**
- **Collectively, different types of fastball form a majority**



- **Players most likely to throw fastball or not throw fastball:** Steven Wright, Clayton Richard, Justin Wilson, Mike Freeman, Chase Anderson, Drew Gagnon, David Hale
- **Players most likely to throw breaking ball or not throw breaking ball:** Ubaldo Jimenez, Sean Doolittle, Alexi Ogando, Robert Stock, Wade Miley, Bud Norris, Zach Britton, Jarred Cosart
- **Umpires most likely to reject players:** Jeremie Rehak, Bob Davidson, John Hirschbeck, Bill Welke, Dale Scott, Mike Everitt, Tom Hallion

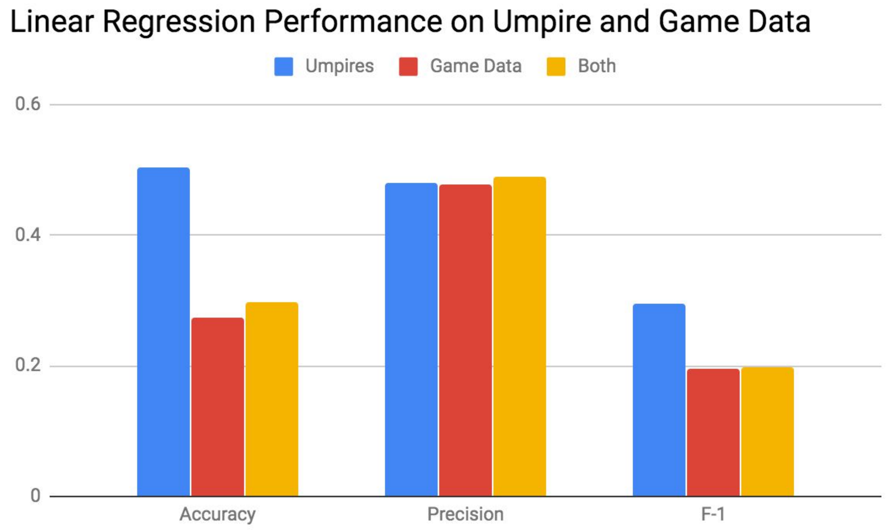
Latent Structure of Pitch Data



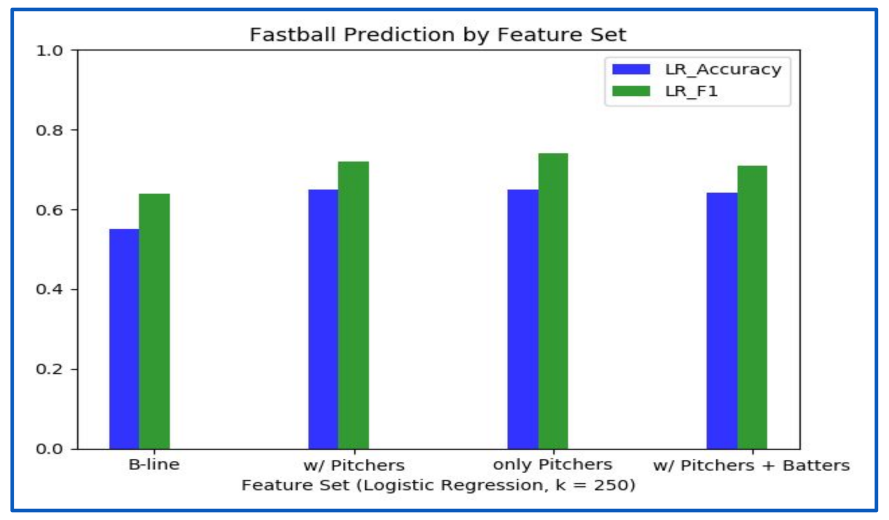
- **K-Means Clustering silhouette score maximized at 0.296 using three clusters**
- **Scree Plot used to determine that with around 20 principal components almost 100% of the total variance of the data is preserved**
- **Factor Analysis identified common trends in at bat situations**

Predicting Ejections

- **Overall, the ejection data was not overly predictive of future ejections**
- **Predicting whether a game will have an ejection based on:**
  - **1) Umpires** of that game
  - **2) Game data** (teams, attendance, length, scores, weather, delay)
  - **3) Combo of the above**



Predicting Pitch Type

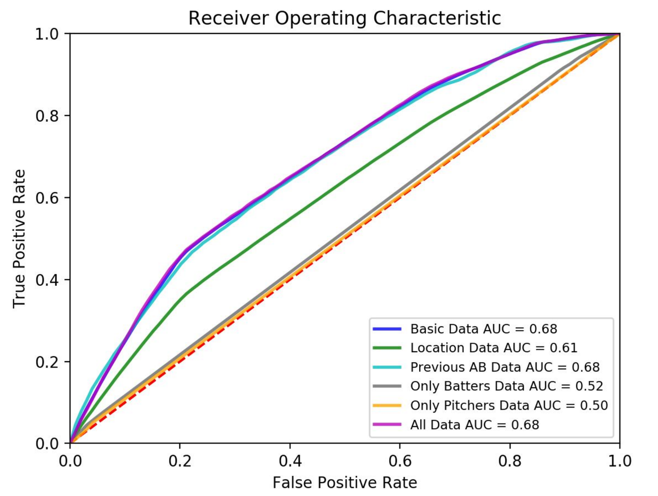


- **Multiple models show that the pitcher is, by far, the best indicator of choice to throw fastball or breaking ball.**
- **Other factors, including identity of the batter, team identity, score, matter very little for pitch type prediction.**
- **We identify the pitchers most predictive of the type of pitch thrown.**

Predicting At-Bat Outcome

We looked at the final pitch in every at bat in order to predict its outcome. We used a Logistic Regression classifier with different features in the input.

	Basic	Location	Previous	Pitchers	Batter	Everything
Accuracy	0.71	0.72	0.72	0.67	0.68	0.71
Precision	0.71	0.72	0.72	0.67	0.69	0.72
Recall	0.97	0.96	0.96	0.98	0.99	0.96
F-1	0.82	0.82	0.82	0.80	0.81	0.82



- **Location and Previous at-bat data slightly improved the model.**
- **Unlike in pitch predictor, the IDs of the players involved were not very predictive.**
- **Most outs are predicted correctly. Some non-outs are predicted incorrectly. This is because of the skewed nature of the data.**

References

1. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
2. Sarris, Eno. "How Has the Value of a Strikeout Changed over the Years?" Sports on Earth. July 01, 2014. Accessed May 09, 2019. http://www.sportsonearth.com/article/82424896/mlb-value-of-a-strikeout-victor-martinez-carlos-lee-placido-polanco.
3. Baumer, Benjamin, and Andrew Zimbalist. 2014. The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball. University of Pennsylvania Press.