

# *The Slot-ification of Baseball*

*Predicting the next pitch in  
America's Pastime*

*Joe Buzzelli*



*May 28, 2020*





## The Slot-ification of Baseball: A study in predicting the next pitch

- Goal is to predict the next pitch type a Major League Baseball (MLB) pitcher will throw given the current scenario in a game that performs better than simply guessing the pitcher's most frequent pitch
- Applications of this study include sports gambling applications, in-game coaching strategy, and/or any other gamification of the MLB





- This study includes data that would be on any scorecard
- Any MLB scoreboard provides good summary of this project's data

3



Batter metrics

Pitcher metrics

Game situation



Data for this study includes almost 600,000 observations

- Data collected for this study derive from MLB's PITCHf/x stats from **2019**
  - Every pitch thrown by a pitcher with at least 750 pitches
  - Every **batter's statistics** as of the beginning of each game
- Features created from the initial data include:
  - **Running pitch totals** for each pitcher per **at-bat, game, batter, and season**
  - **Running batter's statistics** per the beginning of each game
  - **Clustering** of all MLB batters into **four groups** included in the modeling

595,412 Pitches



373 Pitchers



990 Batters

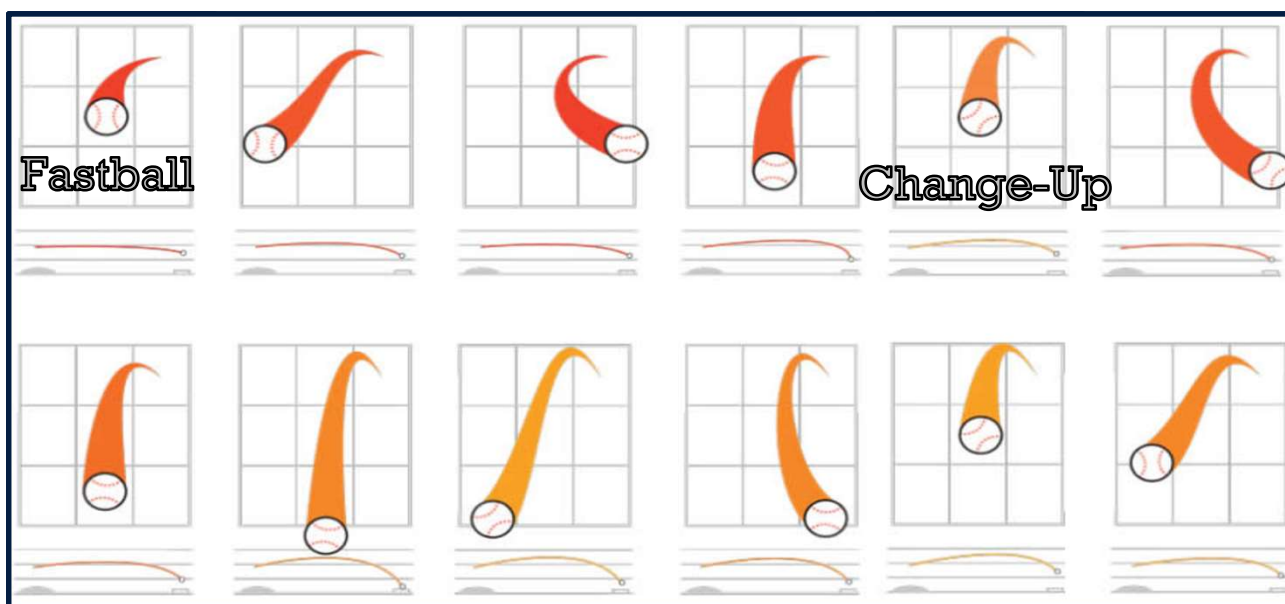






## Pitch types focused on three groups

- The pitch types in this experiment were grouped as either fastballs, change-ups, or movement pitches.





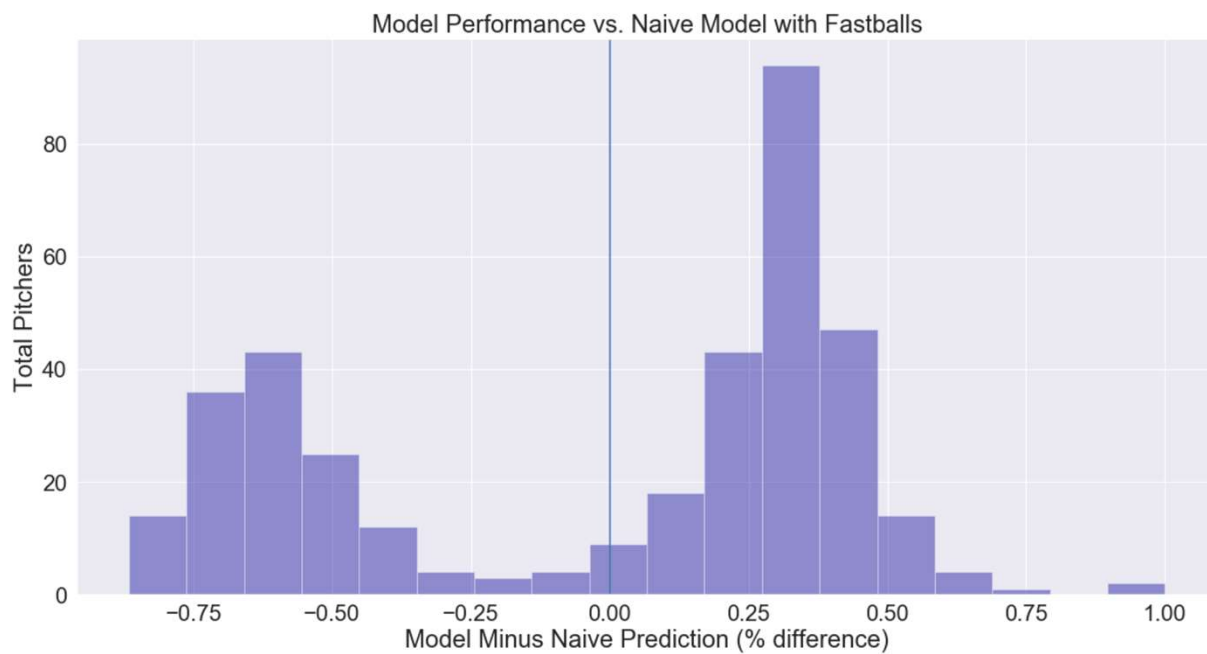
The model outperforms naïve approach for fastballs and offspeed

- By subtracting naïve predictions from the model's predictions, the model outperforms picking the most frequent pitch type thrown
- Movement pitches include seven pitch types and higher variability is expected given the parameters of this experiment



## Model performs well with fastballs

- Model predicts pitches better for 229 or 373 (**61.4%**)

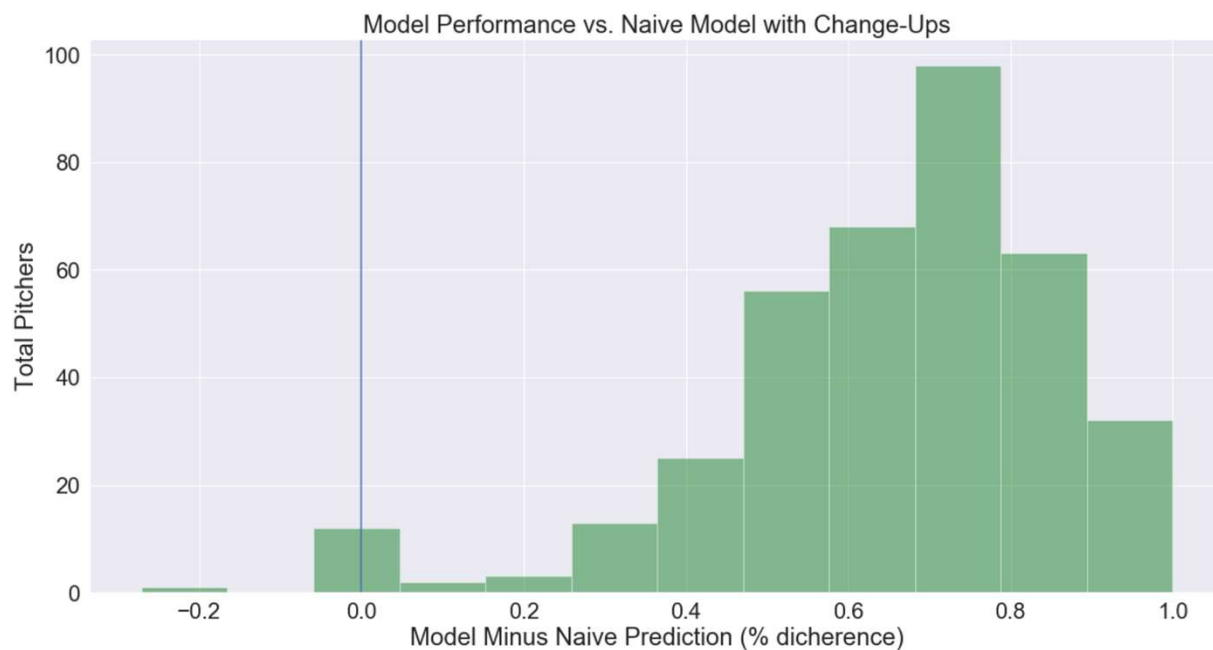






## Model does even better with change-ups

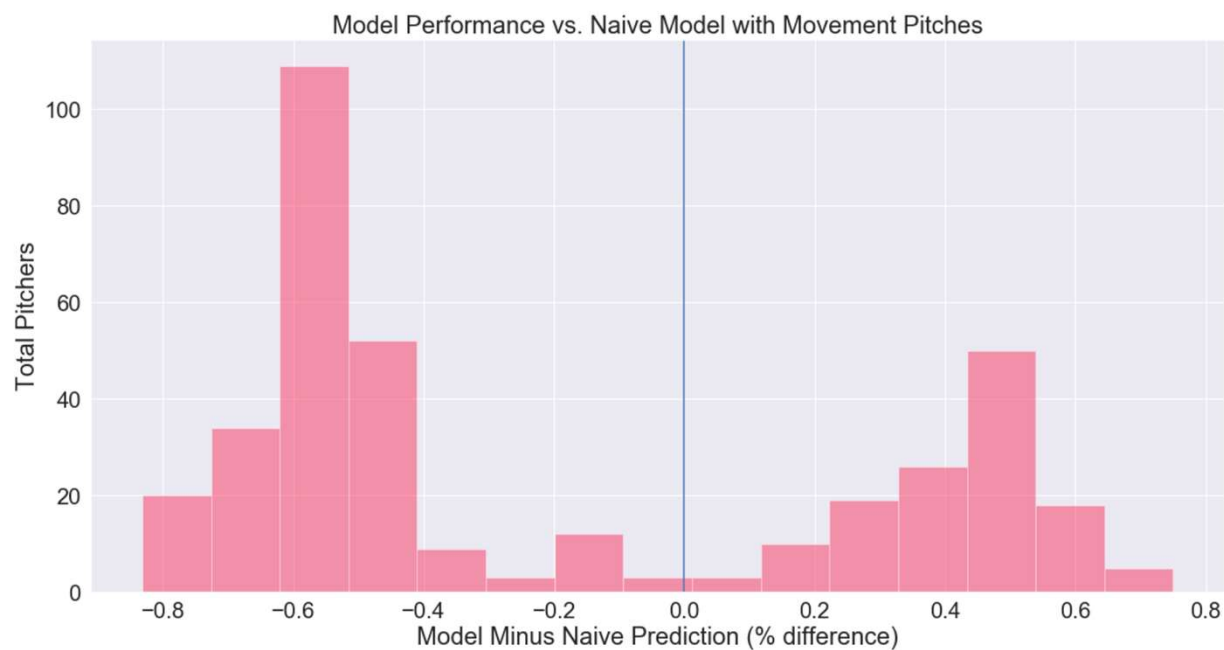
- Model predicts pitches better for 362 or 373 (**97.1%**)





## Model performs worse with movement pitches

- Model predicts pitches better for 131 or 373 (**35.1%**)





## Final model performance... it's complicated

- The final model's performance...depends
- Predicting the pitch from a MLB pitcher is difficult
  - Some pitchers purposefully attempt to disrupt their own tendencies
  - Other pitchers just throw one pitch
- Additional analysis is required across 373 pitchers as to why the model performs better than some pitchers than others





## Future improvements

- Investigating a cluster analysis on pitchers to explore correlations between grouping and appropriate models
- Model on a pitcher/batter basis rather than modeling in the aggregate
  - Will yield 59,952 models
- Create new features to include the prior pitch types and most frequent pitch type per situation for every pitcher/batter combination
- More research into how MLB pitchers strategize specifically as it relates to pitch selection and how often these strategies are changed

A close-up photograph of a brown leather baseball glove lying on a green grassy field. A baseball is nestled inside the glove. In the background, several other baseballs are scattered on the grass, slightly out of focus.

*Thank you for your time...*

*I'll be happy to ...*

*Field any questions*

*Joe Buzzelli*

*[linkedin.com/in/joebuzzelli/](https://www.linkedin.com/in/joebuzzelli/)*


# Appendix

//





## The Slot-ification of the MLB increases transactions by 24,980%

	Traditional Gambling	In-Running	Slot-ification
Description	Based on the final result of the game (winner or total score)	Based on the outcome of each hitter's at bat	Based on the type of each pitch in a game
Transactions per MLB season	162 games	10,692 at bats per season (33 per game per team)	40,630 pitches per season (3.8 pitches per at bat)
Percent increase from baseline	-%	6,500%	24,980%
Potential transactions (not to scale)			



## MLB's PITCHf/x provides exhaustive data for Slot-ification

- Data elements sourced from Baseball Savant at [baseballsavant.mlb.com](https://baseballsavant.mlb.com)

Data Element	Transformation	Data Element	Transformation
Pitch type (target)	Ordinal	Inning	MinMax Scaling
Batter stance (left/right)	Categorical	Outs during at bat	MinMax Scaling
Infield alignment	Categorical	Batter's whiffs	MinMax Scaling
Outfield alignment	Categorical	Batter's swings	MinMax Scaling
Balls and strikes	Categorical	Batter's takes	MinMax Scaling
Runners on base	Categorical	Batter's strikeouts	MinMax Scaling
Nationals (home/away)	Categorical	Batter's walks	MinMax Scaling
Batter's batting average	Already standardized	Batter's singles	MinMax Scaling
Batter's slugging percentage	MinMax Scaling	Batter's doubles	MinMax Scaling
Batter's isolated power	MinMax Scaling	Batters triples	MinMax Scaling
Batter's BA on balls in play	Already standardized	Batters homeruns	MinMax Scaling
Pitch number (per at bat)	MinMax Scaling	Batter's contact types	MinMax Scaling
Pitch number (per batter per game)	MinMax Scaling	Batter's RBIs	MinMax Scaling
Pitch number (per batter per season)	MinMax Scaling	Batter's sacrifices	MinMax Scaling



## The Random Forest performed best in initial testing

- The table below outlines the models assessed in this analysis to include default results, cross validation (CV) scores

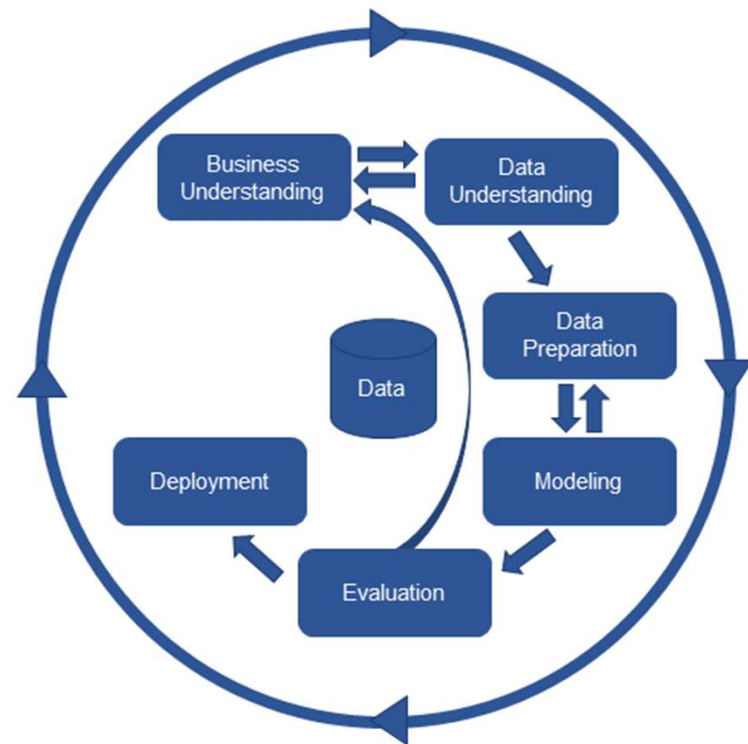
Model	Default Accuracy	CV Accuracy
Dummy	33.3%	n/a
Linear Regression	42.5%	n/a
Decision Tree Classifier	99.9%	45.2%
Random Forest Classifier	99.9%	54.2%
Gradient Boosting Classifier	46.6%	54.1%
Ada Boost Decision Tree Classifier	99.9%	53.2%
Support Vector Classifier	33.8%	38.0%
MLP Classifier	65.4%	51.9%





The experiment process followed the CRISP-DM framework

- In the spirit of the cross industry process for data mining (CRISP-DM) framework, this experiment included several iterations
  - The selection and tuning of different models
  - Just Max Scherzer's pitches for 2019
  - Scherzer's pitches from 2015 to 2019
  - All pitchers from 2015-2019
  - Different standardization techniques for batters and other statistics
  - Different feature engineering techniques
  - Different pitch type groupings





## Confusion matrix for final, tuned Random Forest model

