Proposal: Predicting Pitch Outcomes for Major League Pitchers in 2024

Author: Sara Clokey

University of Connecticut

Department of Statistics

Date: 10-28-2024

Introduction: Over the last 25 years, baseball has become increasingly obsessed with sabermetrics, or “the search for objective knowledge about baseball” (Costa, Huber, and Saccoman (2019)). This process started with engineering advanced statistics that expanded on player performance more than just their batting average. One example is the creation of weighted on-base average (wOBA) (Tango, Lichtman, and Dolphin (2007)), designed to measure a player’s overall offensive contributions per plate appearance through linear weights. Statistics like wOBA have evolved into machine learning and motion-analysis techniques for player performance prediction and injury prevention (Mizels and Chalmers (2022)). Sensor data has also risen in importance; by 2015, all MLB stadiums added sensors to track ball movement and player positioning (Healey (2017)). This sensor data allows for the analysis of pitch outcomes that will occur in my project, as the variables in the data exist accurately due to these technological advancements. In attempting to build on the rich history of sabermetrics, prediction models, and analysis within baseball, I will analyze pitch factors to determine whether the outcome is positive or negative for the pitcher. Analysis of this kind bridges the gap between advanced statistics and prediction in baseball, which is the direction of the general data science field as well. Baseball analysis is ideal for predictive models, with little proxy data, while improving training strategy, coaching decisions, and player development.

Specific Aims: I am interested in what factors contribute to the outcome of a pitch. Every MLB season, some pitchers seem automatic, able to fool the best hitters. Some pitchers seem to lose their ‘stuff,’ unable to perform in big moments. What factors contribute to pitch outcomes? What traits should pitchers strengthen to improve their pitch results? In data science, this results in logistic regression of several explanatory variables versus a binary variable of ‘pitch_outcome’; is there a statistically significant relationship between ‘pitch_outcome’ and sabermetrics like release extension (height of pitch release) or spin rate? How can we predict ‘pitch_outcome’ using the variables featured within the dataset? Which variables are most effective at predicting ‘pitch_outcome’? These statistical inquiries will help determine which factors contribute to pitch outcomes in an MLB game.

Data: I will use data from the 2025 Connecticut Sports Analytics Symposium (CSAS) Data Challenge. This data includes 346,250 plate appearances during the first half of the 2024 MLB season (4/2/2024 to 6/30/2024). The data features pitch-level information from Baseball Savant, MLB’s storage site for sensor data. There are 94 variables in the CSAS dataset. Among these variables, those of interest to this project include ‘pitch_type’, ‘release_speed’, ‘events’, ‘zone’, ‘p_throws’, ‘type’, ‘balls’, ‘strikes’, ‘pfx_x’, ‘pfx_z’, ‘outs_when_up’, ‘inning’, ‘release_spin’, release_extension’, ‘at_bat_number’, ‘pitch_number’, and ‘spin_axis’. These variables deal primarily with the shape and nature of each pitch and the result of the play/pitch, which may help predict the binary ‘pitching_outcome’ variable with a pitcher-focused lens.

Research Design and Methods: I will use the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani (1996)) to focus on model selection, overfitting mitigation, and interpretability with proper variable preparation. For LASSO regression, I will create a binary variable, ‘pitch_outcome’, with two outputs: 1 if positive, 0 if negative. I am interested in pitching analytics and will, therefore, consider positive and negative outcomes from a pitcher’s perspective. I will use the ‘events’ and ‘type’ variables to determine pitch outcome; ‘events’ denotes the result of the at-bat, and ‘type’ denotes the result if the at-bat is not over. For an output of 0, I will include all negative pitch outcomes for the pitcher from the ‘type’ and ‘events’ variables. For an output of 1, I will include all positive pitch outcomes for the pitcher from the same variables. LASSO regression then requires creating testing and training sets from the data, fitting the model to predict ‘pitch_outcome’ based on other features, and including tuning parameters using the ‘scikit-learn’ package. Model validation will involve finding the confusion matrix, accuracy, precision, recall, F1 Score, and AUC. The ‘scipy.stats’ package will work to conduct t-tests and Chi-squared tests to explore relationships between variables, including the binary ‘pitch_outcome’ variable.

Discussion: The most challenging part will be the risk of overfitting the model with several explanatory variables under consideration. Once the LASSO model is constructed, I will use its variable selection to run another regression and weed out unnecessary variables as a counteraction to overfitting. There is also a possibility of class imbalance in the binary variable; this can be mitigated by using the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. (2002)). The limitations of this work include my categorization of outcomes as ‘good’ or ‘bad’; I made these delineations based on my baseball experience, but others may categorize these features differently to varying results. This work is also limited to the first half of the 2024 season, which may impact generalizability given baseball has existed for over 100 years. If something unexpected happens, I have a fall-back dataset from the UConn Baseball team that would permit similar statistical analysis as outlined in this proposal.

Appendix

Referenced MLB data and 'pitch_outcome' binary variable:

```python
import os
import pandas as pd
import pyarrow as pa
file_path = ('data/statcast_pitch_swing_data_20240402_20240630.arrow')
table = pa.ipc.open_file(file_path).read_all()
df = table.to_pandas()

bad_events = ['single', 'walk', 'home_run', 'double', 'field_error',
              'hit_by_pitch', 'catcher_interf', 'triple', 'sac_fly',
              'sac_bunt', 'stolen_base_2b']
good_events = ['strikeout','field_out','force_out',
                'grounded_into_double_play','double_play','fielders_choice',
                'caught_stealing_home','fielders_choice_out',
                'caught_stealing_2b','strikeout_double_play',
                'caught_stealing_3b','other_out',
                'pickoff_caught_stealing_home','pickoff_caught_stealing_3b',
                'pickoff_3b','sac_fly_double_play','pickoff_1b','triple_play']
def classify_pitch_outcome(row):
    if row['events'] in bad_events or row['type'] == 'B':
        return '0'
    elif row['events'] in good_events or row['type'] == 'S':
        return '1'
    else:
        return 'None'
df['pitch_outcome'] = df.apply(classify_pitch_outcome, axis=1)
pd.set_option('display.max_columns', 5)
print(df.head())
```

```
  pitch_type   game_date  ...  swing_length  pitch_outcome
0         FF  2024-04-02  ...           NaN              0
1         CH  2024-04-02  ...           NaN              1
2         SI  2024-04-02  ...           NaN              1
3         SI  2024-04-02  ...           NaN              0
4         FF  2024-04-02  ...           NaN              1

[5 rows x 95 columns]
```

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–57.

Costa, Gabriel, Michael Huber, and John Saccoman. 2019. *Understanding Sabermetrics: An Introduction to the Science of Baseball Statistics*. McFarland & Company.

Healey, Glenn. 2017. "The New Moneyball: How Ballpark Sensors Are Changing Baseball." *Proceedings of the IEEE* 105 (11): 1999–2002.

Mizels, Joshua, and Peter Chalmers. 2022. "Current State of Data and Analytics Research in Baseball." *Current Reviews in Musculoskeletal Medicine* 15 (4): 283–90.

Tango, Tom, Mitchel Lichtman, and Andrew Dolphin. 2007. *The Book: Playing the Percentages in Baseball*. Potomac Books.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the LASSO." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.