Is the MLB Entering an Era of Speed Over Power?

Ivy Maynard – STAT 228 Final Project

My Research

Inspiration:

- I've observed an increased focus in sports media on fast players and stolen base records.
- Ronald Acuna Jr., Bobby Whitt Jr., and Jarren Duran

Question:

- Do fast baserunners get into the heads of opposing teams, causing an increased chance of winning?
- Ultimately, should teams change their priorities from power hitters to fast baserunners?

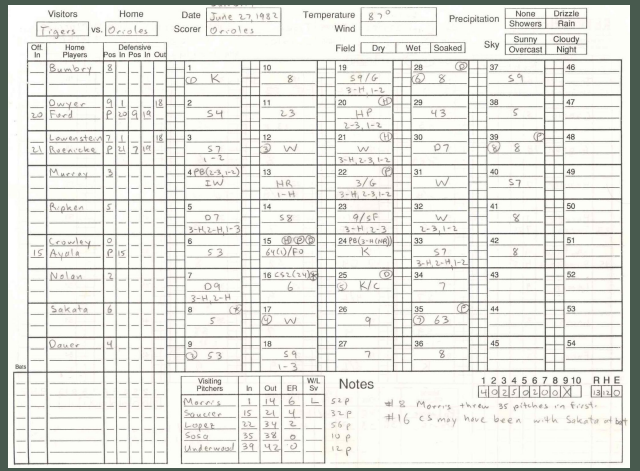
Analysis:

- I decided to look into whether or not a stolen base or a triple in the first inning increases a teams chance of winning.

My Data

Retrosheet:

- A nonprofit baseball statistics organization that created a collection of all public play-by-play and game data for every MLB+ game dating back to the 1800s
- R package "Retrosheet" allows you to import and scrape data directly from the website to R



Data Wrangling

Steps:

- Used function get_my_retrosheet to scrape play-by-play data for every MLB game in 2024
- Filtered my data to only include the first inning of each game
- Created columns to identify stolen bases and triples (1 for true and 0 for false)
- Grouped and summarized to find total T and SB in the inning
- Used get_retrosheet to find game data for each game of 2024
- Created column to determine win status using ifelse statements
- Used pivot_longer to create one row for each team involved in the game
- Joined data frames using game ID and selected relevant columns

GAMEID	team	SB	T	Win
ANA202404050	0	0	0	W
ANA202404050	1	0	0	L
ANA202404060	0	0	0	L
ANA202404060	1	0	0	W
ANA202404070	0	0	0	W
ANA202404070	1	0	0	L
ANA202404080	0	0	0	L
ANA202404080	1	1	1	W
ANA202404090	0	0	0	W
ANA202404090	1	0	0	L
ANA202404100	0	0	0	W
ANA202404100	1	0	0	L
ANA202404220	0	1	0	W
ANA202404220	1	0	0	L
ANA202404230	0	0	0	L
ANA202404230	1	0	0	W

1 to 16 of 4,858 entries, 5 total columns

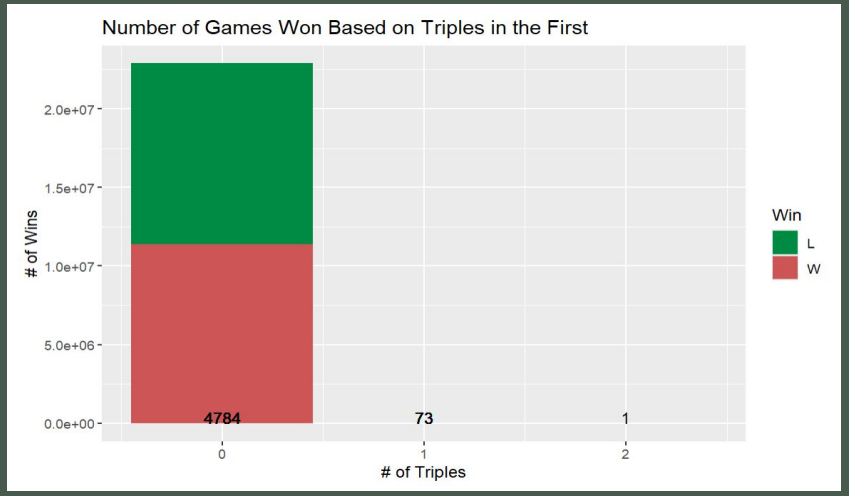
Visualization

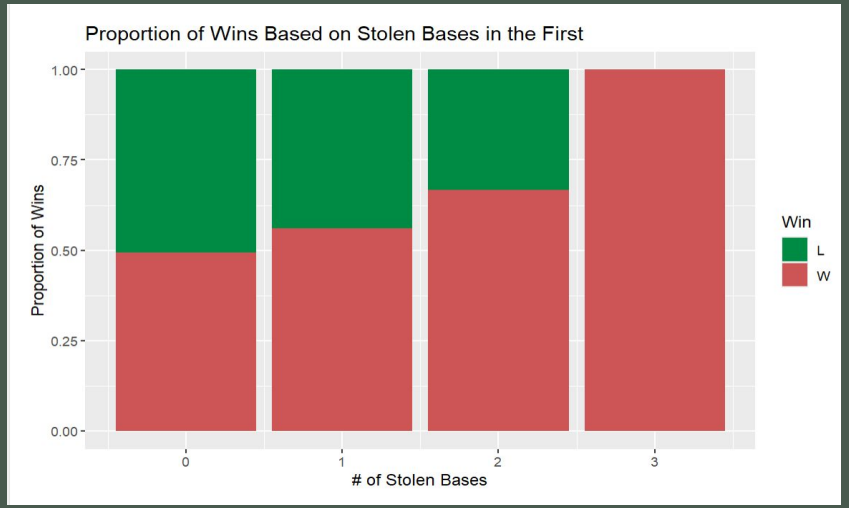
A bar plot showing games in 2024 that were wins vs losses based on # of SB in the 1st

- Stacked vs filled bar charts show the effects of disparities in count
- Many more games without a SB or T in the first

A bar plot showing games in 2024 that were wins vs losses based on # of T in the 1st

- Stacked vs filled
- Only 1 game had 2 triples in the 1st which leads to unexpected filled bar chart





Modeling

Steps:

- Split data into train (80%) and test(20%)
- Fit a logistic regression model
- Evaluate my model
  - Find the predicted probability of wins for test and training data
  - Create a confusion matrix
  - Determine F1 score
- Cross validate with data from 2023
  - Repeat steps

```
Call:
glm(formula = Win ~ SB + T + SB:T, family = "binomial", data = data_train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.02586    0.03359  -0.770   0.4414
SB           0.21682    0.10464   2.072   0.0383 *
T            0.23858    0.29562   0.807   0.4196
SB:T         1.76769    1.09916   1.608   0.1078
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Model:
Win = 0.217(SB) + 0.239(T) + 1.768(SB:T) - 0.026

F1 Scores:

2024 Training Data: 0.177	2023 Training Data: 0.171
2024 Test Data: 0.200	2023 Test Data: 0.161

```
Confusion Matrix and Statistics

          Reference
Prediction    L    W
         L 1783 1739
         W  161  205
```

Conclusion

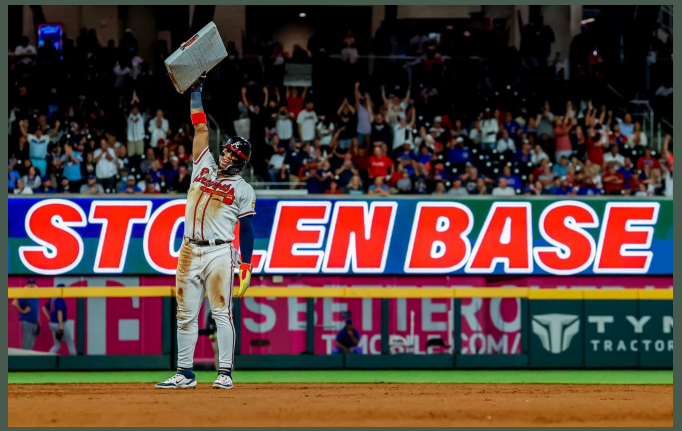
What does the model tell us?

- F1 score is on a scale of 0-1 with 1 being the most accurate and 0 being the least
- The low F1 values shows us that our model is not much better at predicting wins than a random model
- The P value of SB is below the level of significance of .05 while T is not
- Suggest that there is a correlation between wins and SB in the first inning but no correlation between wins and T in the first inning

Where can we go from here?

Lots of follow-up analysis can be done to see if we can find a better way to predict runs:

- Look at SB and T in all 9 innings
- Remove triples from the model
- Look at other speed metrics like sprint speed, 90 ft/s, doubles on shallow hits, etc.



What does this say about the MLB?

This model does not provide concrete evidence to suggest that hitting more triples in the first inning leads to an increased win probability, but it does suggest that stolen bases can impact the outcome. Therefore, teams may benefit from acquiring players who steal more bases as well as training the top of their batting order to steal early in the game. More analysis would need to be done.

Citations and References

Retrosheet:

- Data: https://www.retrosheet.org/
- Function: https://cran.r-project.org/web/packages/retrosheet/index.html

Get_my_retrosheet function:

- Author: Jim Albert
- GitHub: https://gist.github.com/bayesball/e7d56e5edf31d7a71c6bf40cc1dfe743
- https://baseballwithr.wordpress.com/2024/03/27/retrosheet-package-and-comparing-count-rates/

Packages:

- tidyverse
- retrosheet
- openintro
- caret
- pROC
- rpart
- rpart.plot