Major League Baseball Team Season Winning Percentage Breakdown

William Mai

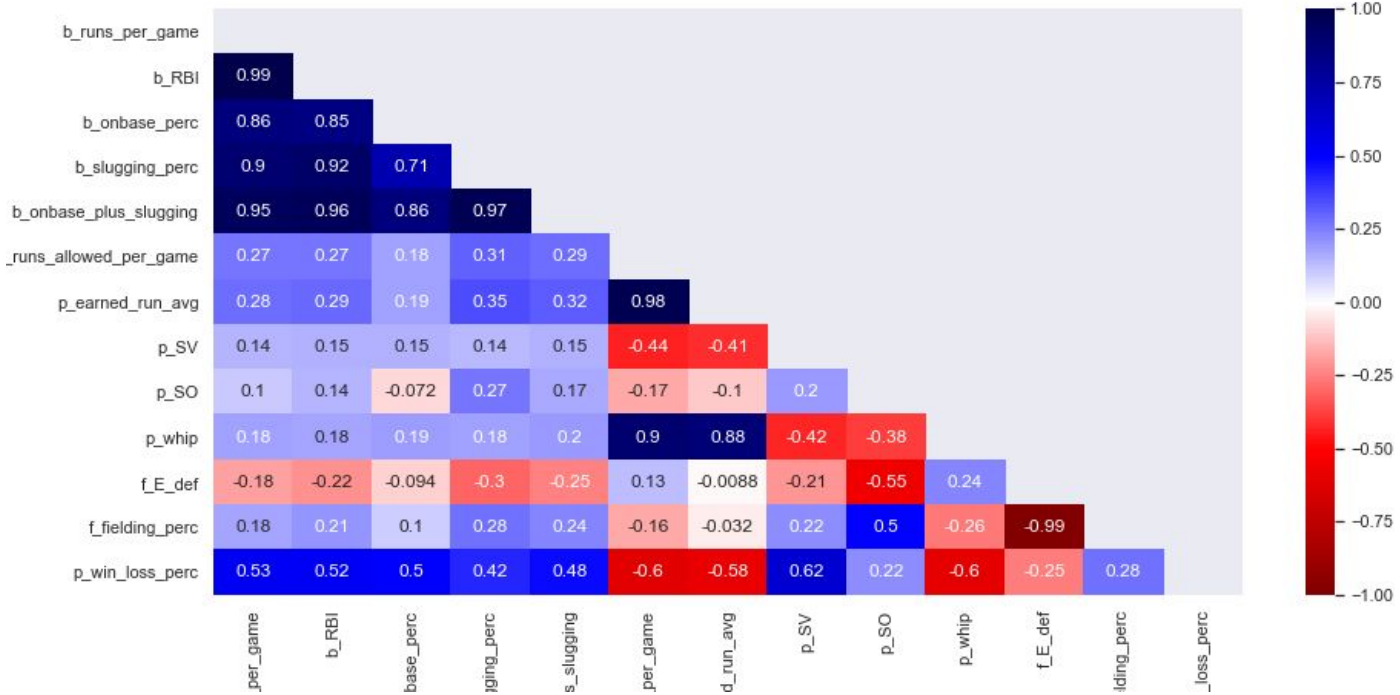
Introduction

- There are around 80 different statistics recorded for each MLB teams
- Statistics are categorized into three groups
  - Batting
  - Pitching
  - Fielding
- What are the main drivers of a team's win rate?
- MLB Teams might be able to use the result to analyze their winning percentage while improving their underperforming statistics

Methodology

- MLB Data: Yearly data from 1980 to 2020 with roughly 30 teams each year
- 1981 (strike), 1994 (strike), 1995 (strike), and 2020 (covid) data are removed due to shorter seasons
- Matrics
    - Winning percentage: wins / (wins + losses)
    - Runs per game: runs scored per game as the team on offense
    - Earned run average: 9 x earned runs / innings pitched as the team on defense
    - Saves: saves as the team on defense
    - Errors: errors committed as the team on defense

Results – Selected Features Correlation Plot

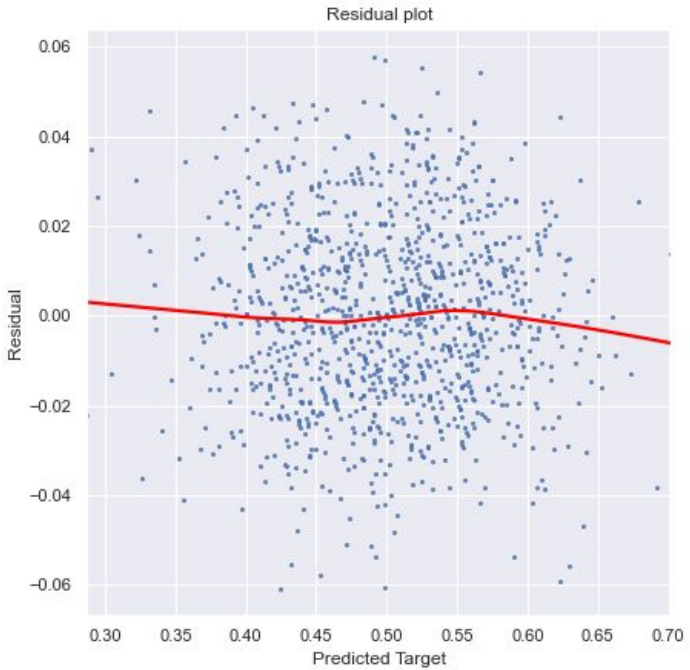


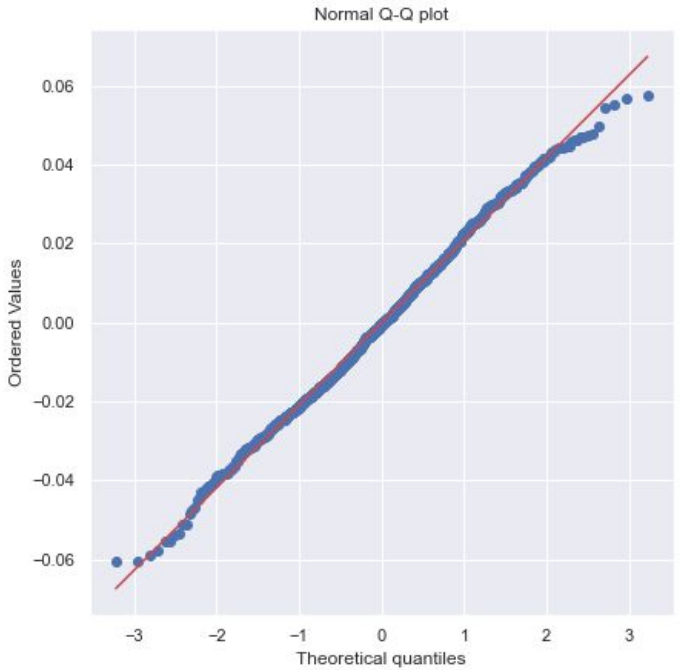
Results – Final Model

$$\begin{aligned} \textit{Win \%} = 0.400 &+ 0.0902 \times \textit{Runs per game} \\ &- 0.0866 \times \textit{Earned run average} \\ &+ 0.00208 \times \textit{Saves} \\ &- 0.000286 \times \textit{Errors} \end{aligned}$$

$$R^2 = 0.911$$

Results – Residual Analysis



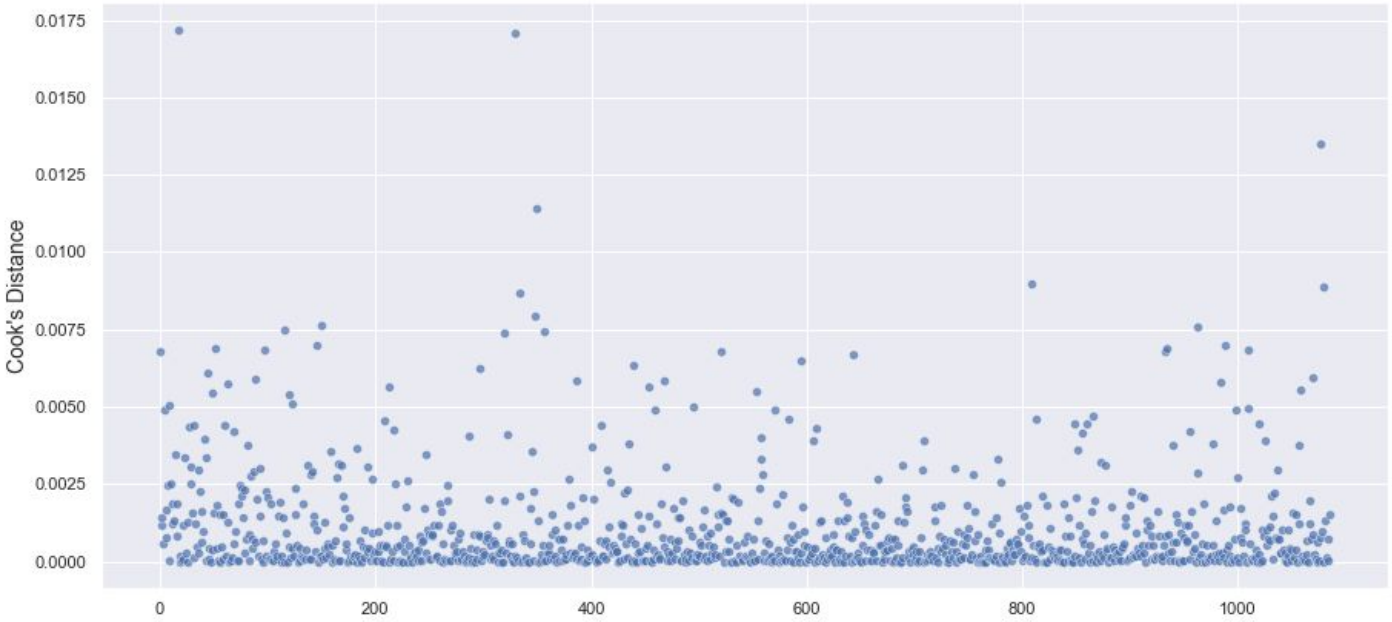


Conclusions

- The ability to score runs is the main batting feature that positively contributes to the target
- Not allowing an opponent to score runs is equally important as the ability to score runs
- Having strong relief pitchers would benefit the win rate as well
- Although MLB teams generally have very low numbers of errors, it could negatively impact the win rate of a team if it has too many errors

Appendix

Estimate the Influence of All Data Points



Results – train / test scores of different models

	12 features	7 features	7 features Ridge	7 features Lasso	4 features	4 features Ridge
Train Score	0.913	0.912	0.910	0.853	0.909	0.909
Test Score			0.922	0.858	0.922	0.922
Note	multicollinearity	multicollinearity	multicollinearity	multicollinearity		