How to Play Baseball Better

Data Source & Methodology

Data Source: MLB Game Logs (1946-2016) sourced from Retrosheet

Methodology:

- Target Variable: Match Outcome (Win / Loss)
- 70:30 Train-Test Split
- 5-fold Cross Validation
- Logistic Regression
- Decision Trees
- Random Forest
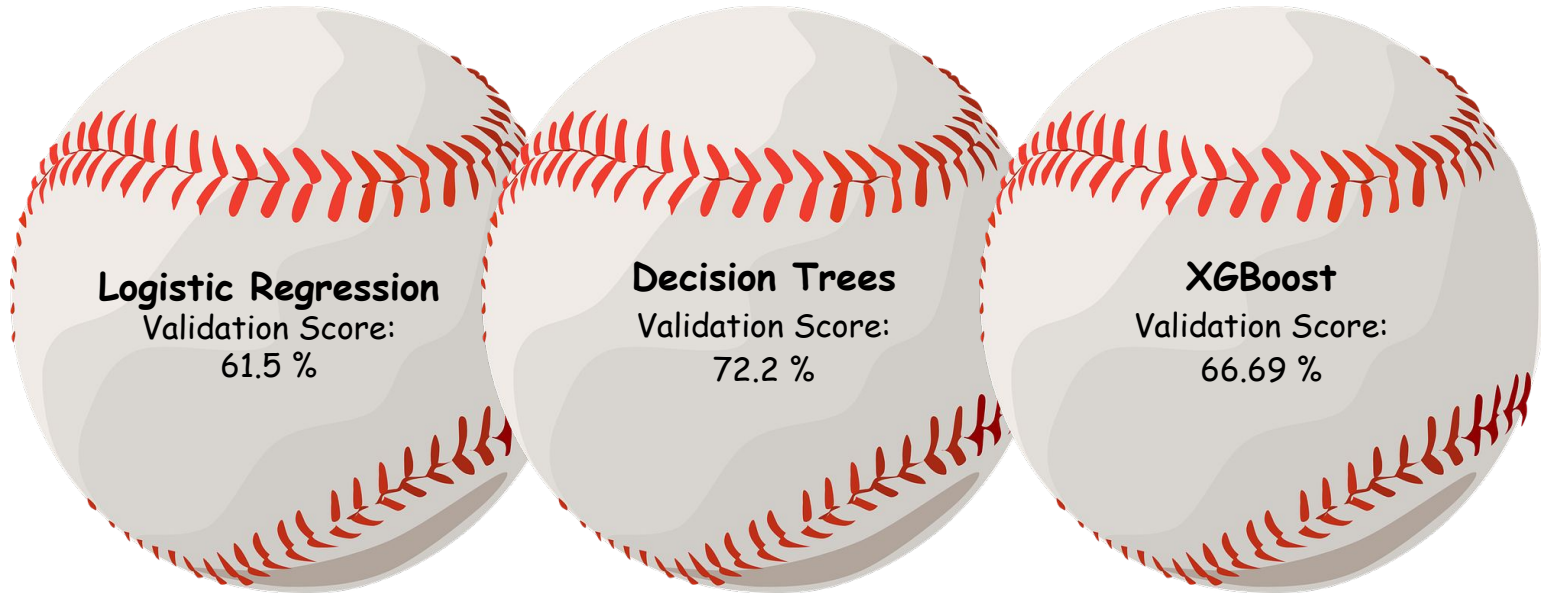- XGBoost
- Comparison by ROC-AUC scores

Data Limitations

- We focused on single day games that had a non-draw outcome.
- Protested or interrupted games have been put aside as well.
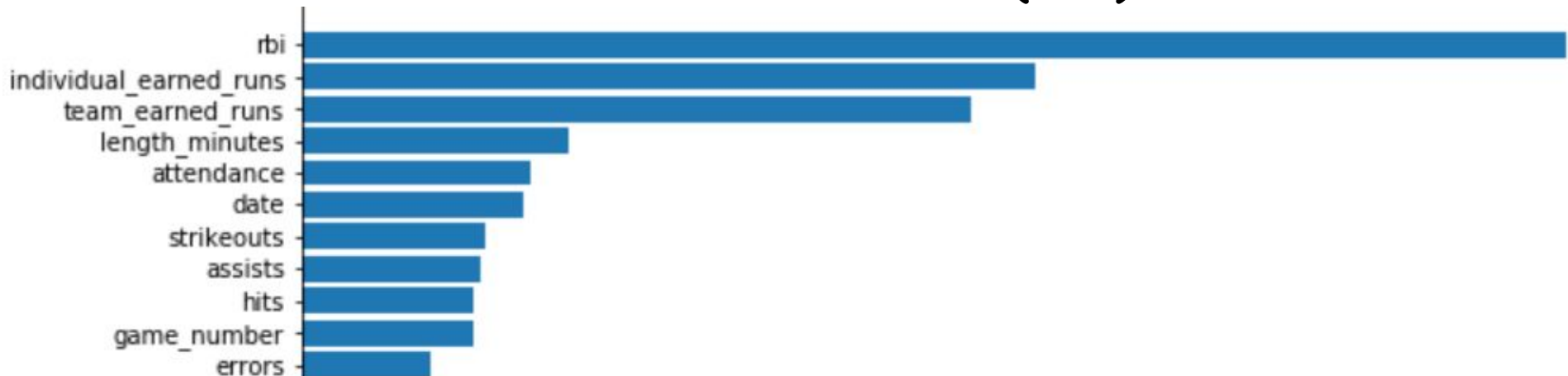- Two decades were missing from the dataset: 70s and 90s.

Comparison Between Models

Validation score for the Random Forest model: 74.2 %

Logistic Regression
Validation Score:
61.5 %

Decision Trees
Validation Score:
72.2 %

XGBoost
Validation Score:
66.69 %

Feature Selection (RF)

Actionable Insights

1. Invest in top pitchers for defensive plays to reduce opponent's RBI and hits.
2. Employ the most consistent batter to improve RBIs.
3. Do not focus on hitting ambitious strikes (doubles, triples, homeruns), it's more effective to just get the bat on the ball as frequent as possible.
4. Approaching the end of the inning, baserunners should focus on taking risk, to reduce number of players left on base so they could complete a run.

Thank You !
Questions ?