Predicting Baseball Game Outcomes

Brendan Whitney

Monday, December 10th

Data

- Data from 2018 MLB Season
- Consists of 2431 games
- Baseball Statistics from baseball-reference.com
  - Game Location
  - Runs Scored
  - Runs Allowed
  - Wins
  - Losses
  - Streak
- Weather Data from Global Historical Climatology Network
  - Maximum Temperature
  - Minimum Temperature
  - Precipitation

Calculated Data

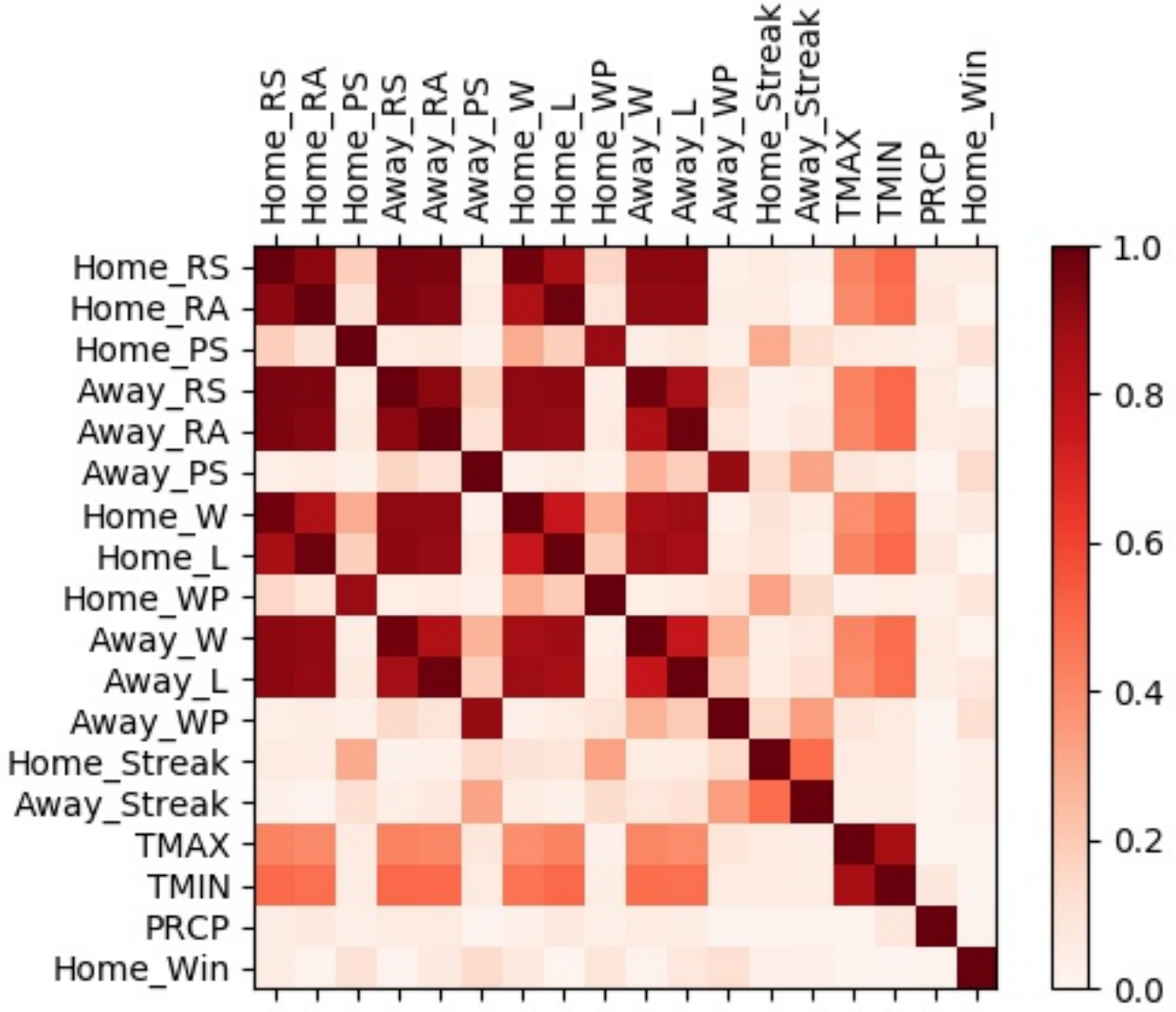
- Pythagorean Score

$$\frac{RS^{\gamma}}{RS^{\gamma} + RA^{\gamma}} \quad \text{for} \quad \gamma = 1.79$$

- Win Percentage

$$\frac{W}{W + L}$$

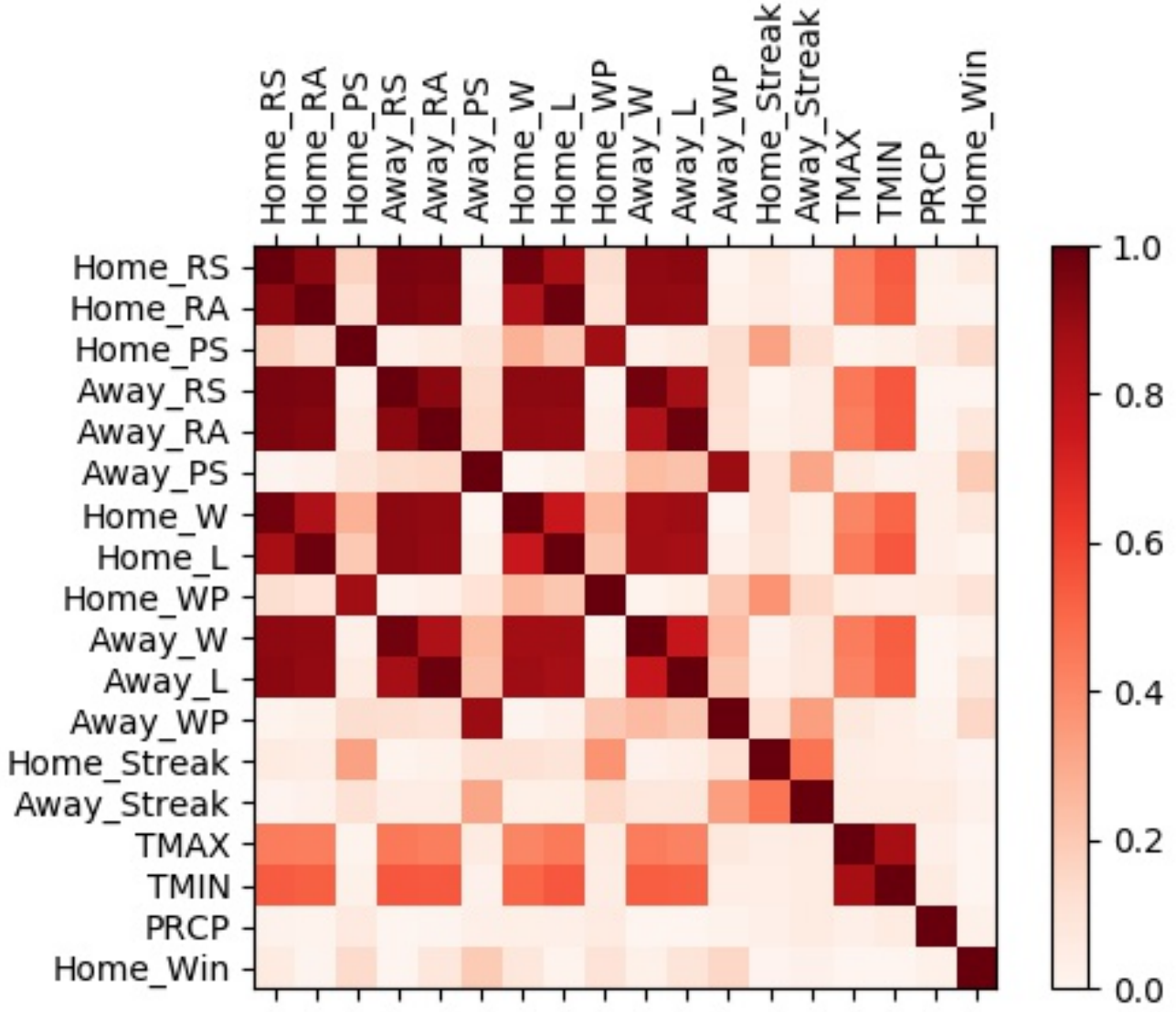
Correlations



Sampling Seasons

- Sampled seasons to minimize correlations
- Randomly chose 1 game from each series

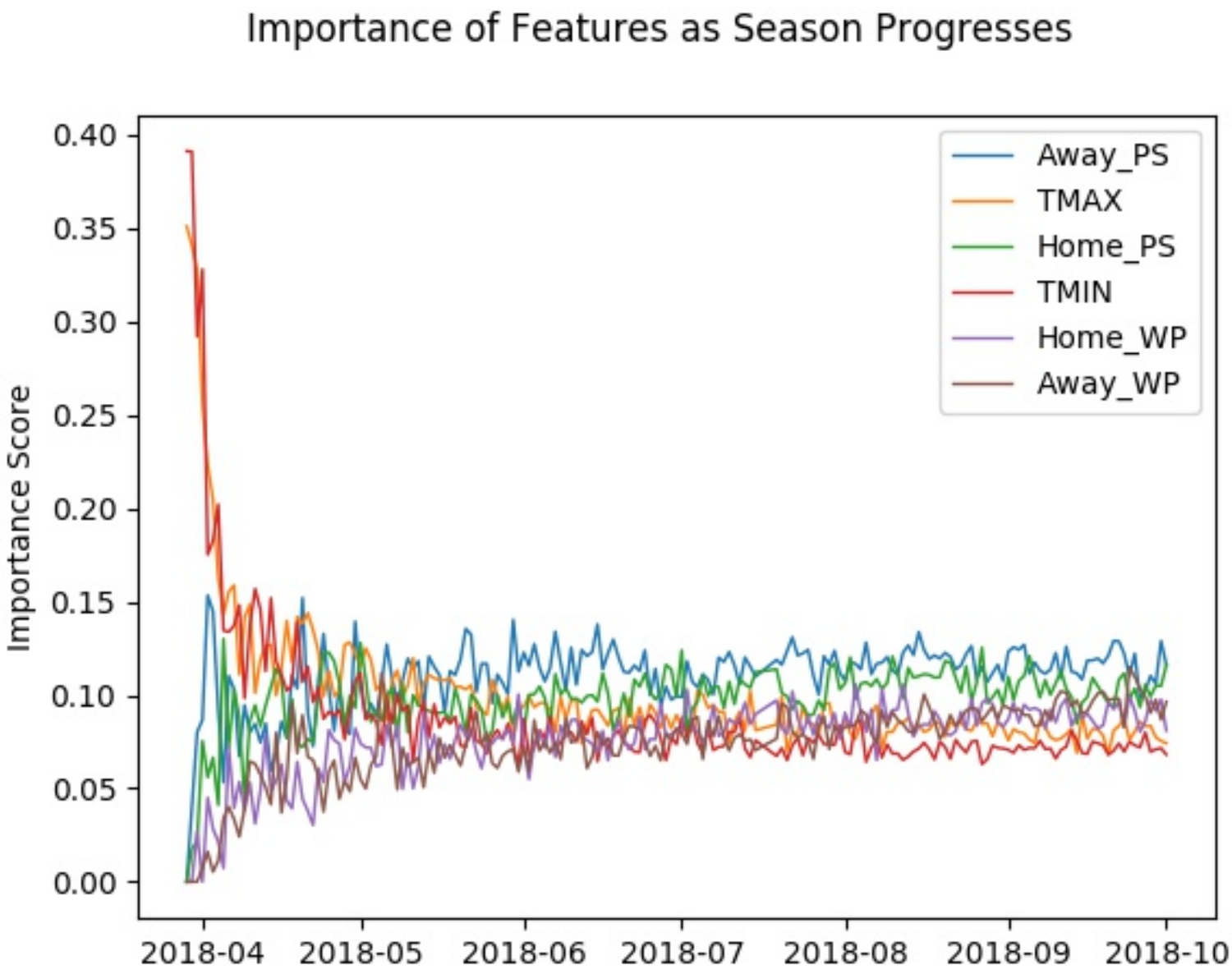
Sampling Correlations



Feature Selection

- Used a Random Forest Regressor for feature selection
- Found six key features for regression
  - Away_PS
  - TMAX
  - Home_PS
  - TMIN
  - Home_WP
  - Away_WP

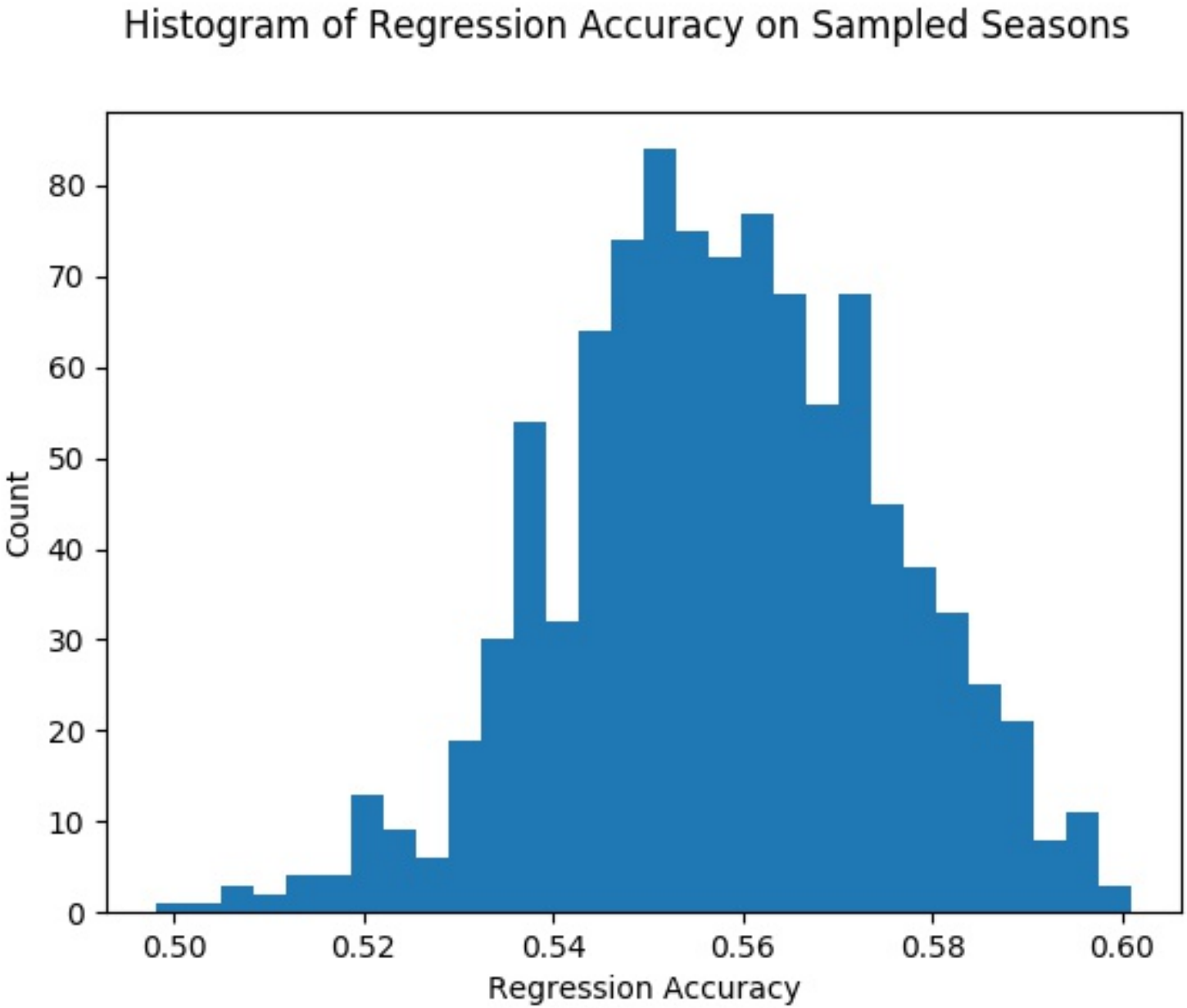
Feature Selection



Regression

- 56.58% accuracy for the full 2018 season
- For 1000 sampled seasons:
  - 55.77% mean accuracy
  - 1.7% standard deviation

Regression


Histogram of Regression Accuracy on Sampled Seasons
Count
0
10
20
30
40
50
60
70
80
0.50
0.52
0.54
0.56
0.58
0.60
Regression Accuracy