BASEBALL STATS MODELING

EXAMINING PLAYER AND TEAM PERFORMANCE USING MACHINE LEARNING

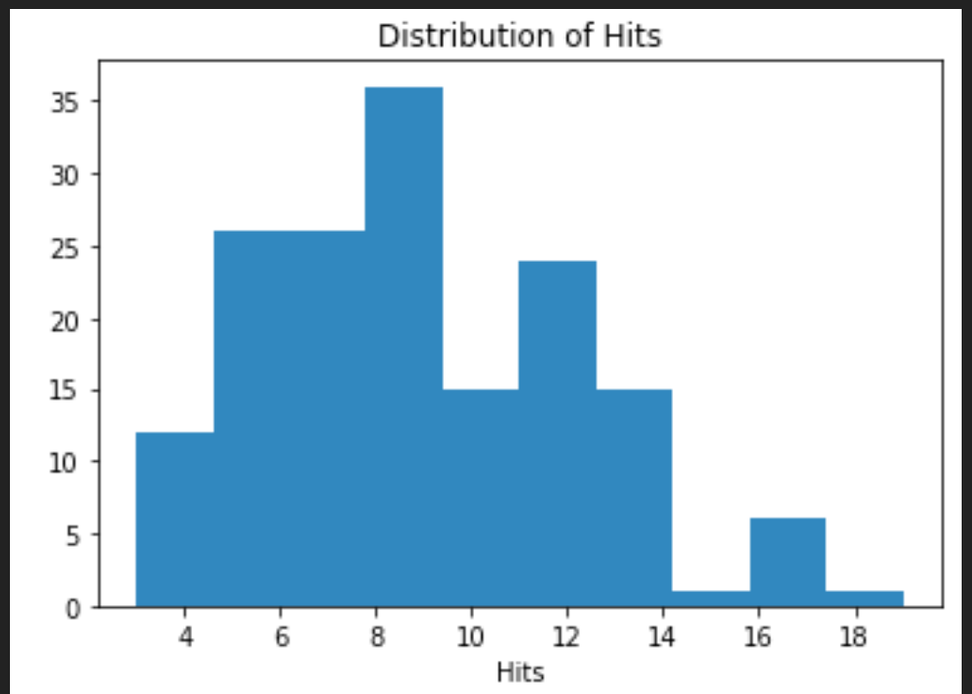
OVERVIEW

- Detailed exploratory analysis of data
- Part 1: Examining team performance and determining the feasibility of predicting baseball game outcomes
- Part 2: Using Machine Learning to evaluate player Wins-Above-Replacement totals for the 2019 season

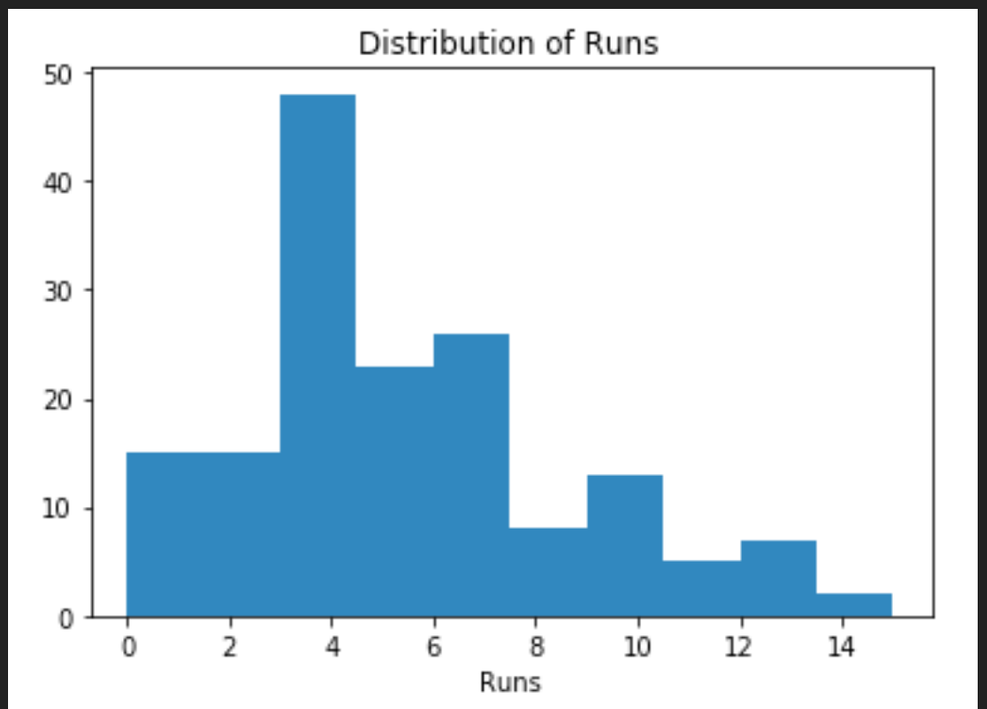
ANALYSIS OF DATA

- Baseball is an incredible keeper of records, splits are available on virtually every stat possible to record.
- Trends in hits and runs, the primary factors of scoring ranked as the most important for team success. However cross referencing 9-player fielding versus 1 player batting was extremely difficult.
- For individual player performance in offense versus their peers, WAR was identified as the overall best predictor of player value, traditional offensive number as well as fielding metrics were identified as valuable.

Distributions of Hits and Runs, 2019 Atlanta Braves



Mean Hits/Game: 5.277777777777778



Mean Runs/Game: 5.277777777777778

MACHINE LEARNING APPLIED TO BASEBALL RESULTS

- Extreme number of factors to account for, environmental conditions were dropped as well as head-to-head and pitching performance as the value versus a particular hitter is near-impossible to predict.
- Correlation in results is strongest between hits and runs scored of the team, trends in certain months identified and winning streaks established.
- Highest correlation columns were included in the final data

Results of predicting model efficiency of predicting game reults

```
Rslt          0.598765
PA           38.901235
AB           34.320988
R             5.277778
H             8.839506
2B            1.709877
3B            0.179012
HR            1.537037
RBI           5.086420
BB            3.820988
IBB           0.240741
SO            9.055556
ROE           0.314815
SB            0.549383
BA            0.261333
SLG           0.451302
OPS           0.793025
hit_bins      1.969136
dtype: float64
```

```
Rslt          0.714286
PA           40.285714
AB           35.750000
R             6.678571
H             9.892857
2B            2.285714
3B            0.214286
HR            2.000000
RBI           6.464286
BB            3.571429
IBB           0.178571
SO            8.928571
ROE           0.321429
SB            0.500000
BA            0.261750
SLG           0.451750
OPS           0.787786
hit_bins      1.928571
dtype: float64
```

```
lr.fit(x_train, y_train)
predictions = lr.predict(x_test)

# Determine mean absolute error
mae = mean_absolute_error(y_test, predictions)

# Print `mean absolute error'
print(mae)
print(predictions)
```

executed in 21ms, finished 09:18:09 2020-06-29

```
0.3922009257092688
[ 0.63880199  0.92926195  0.5543073   0.46710779  0.83530545  0.49487821
  0.82332197  0.60255316  0.24451183  0.16953551  0.81868119 -0.24077974
  1.1066538   0.48950692  0.17629844  0.36114219  0.89343522  0.64548067
  1.48430622  1.33328704  0.37423865  0.82605636  0.46110203  0.14906376
  1.07885472  0.73101191  0.38620497 -0.82707437  0.22169401  0.79054386
  1.18161317  0.55069677  0.97755823  0.3867639   0.68690898  0.7968166
  0.79825573  0.57832276  0.66558962  0.73504176]
```

As seen here, the overall mean error

Achieved was 39%, not great.

This can be attributed to a number of things.

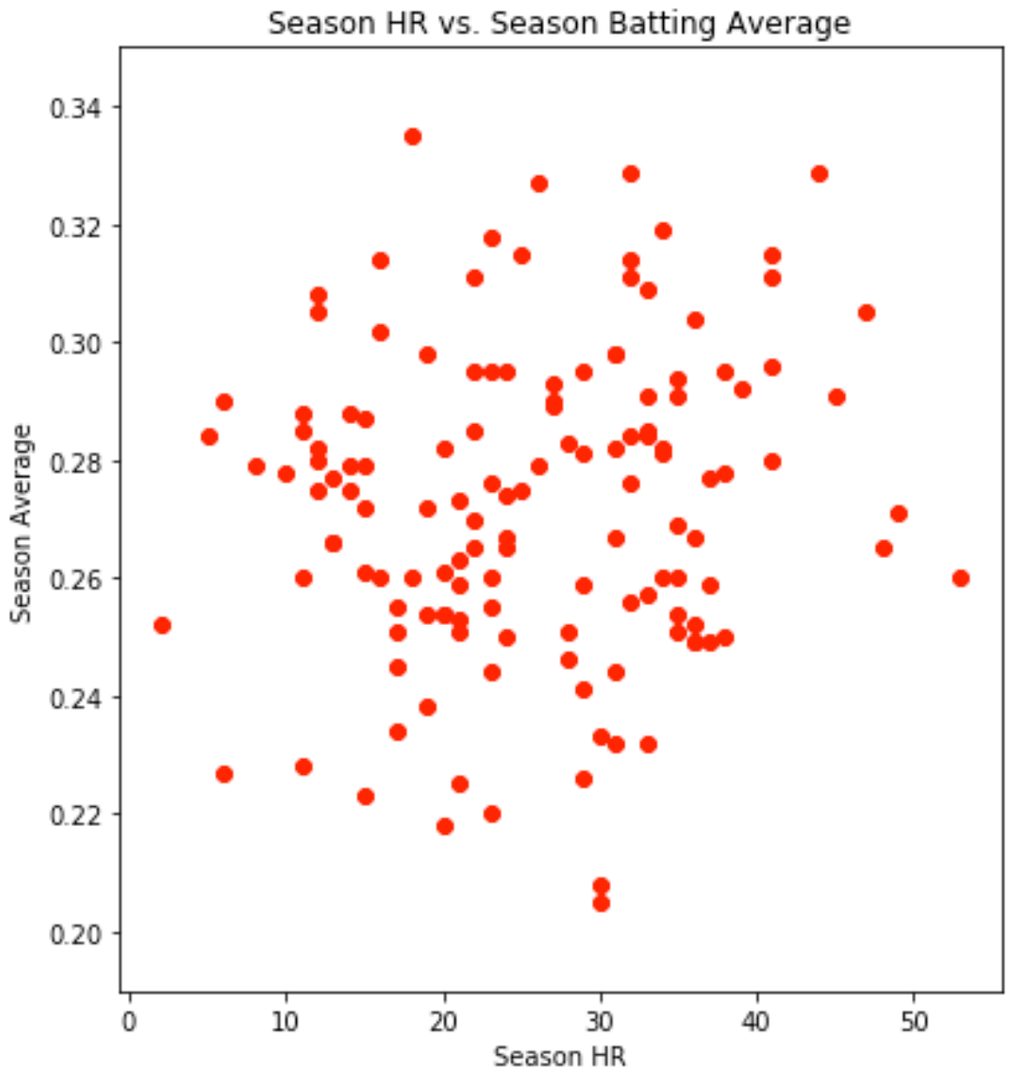
1. We are only examining total team stats.
2. No environmental or pitching statistics.

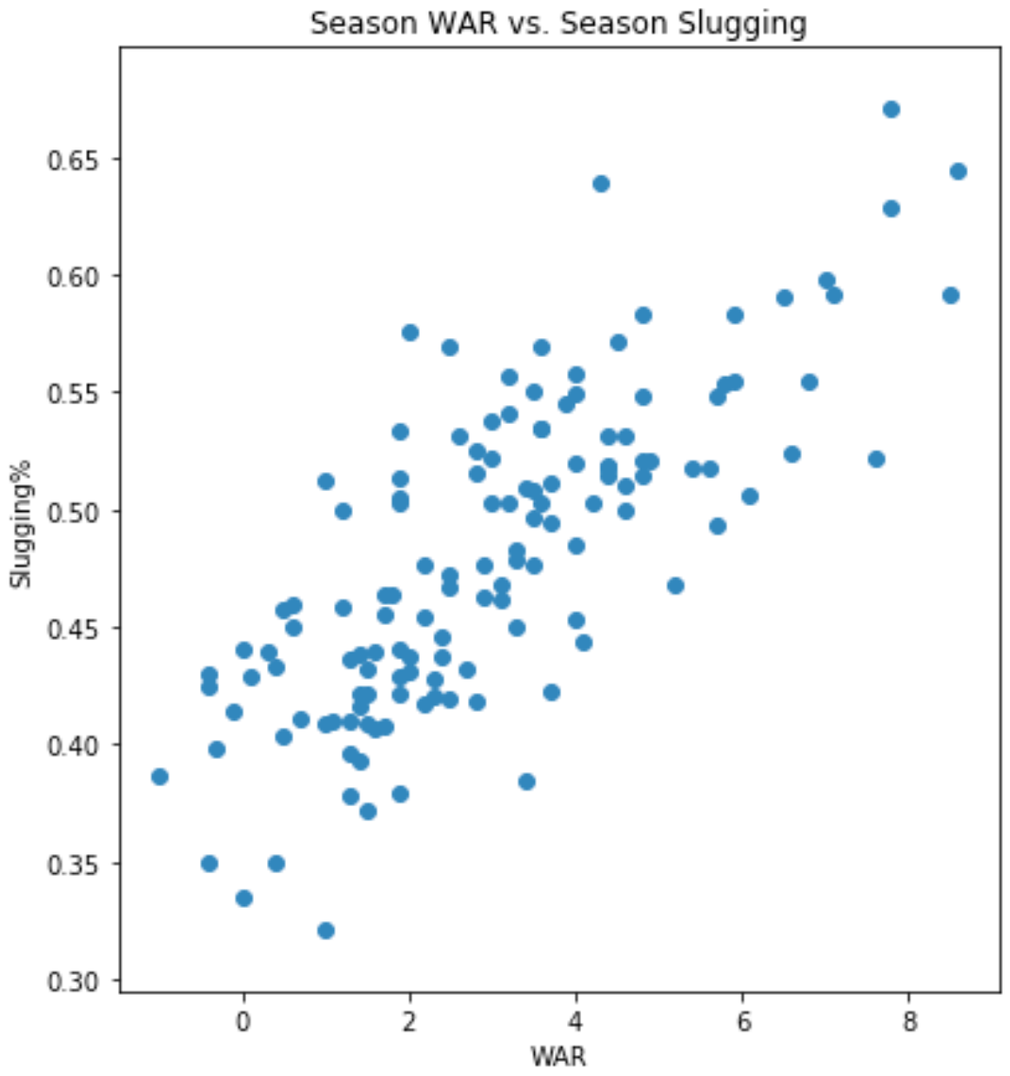
	PA	AB	R	H	2B	3B	HR	RBI	BB	IBB	SO	ROE	SB	BA	SLG	OPS
0	36	30	4	7	1	0	1	4	6	0	9	0	1	0.233	0.367	0.728
1	40	36	6	10	2	1	2	6	4	0	9	0	1	0.258	0.470	0.825
2	36	29	1	4	0	0	0	0	6	0	8	0	1	0.221	0.368	0.708
3	43	36	8	12	3	0	2	6	7	1	9	2	0	0.252	0.427	0.795
4	39	30	6	8	1	0	2	6	8	0	8	0	1	0.255	0.441	0.817

BUILDING MODEL TO PREDICT INDIVIDUAL PLAYER PERFORMANCE

- Wins-above replacement identified as primary stat.
- Created model to predict positive war values.
- Model still need refining with additional statistics. High performers are considered outliers and model has difficulty determining their productivity as an outlier.

Stats distribution






```
0    67
1    17
2    14
3     2
4     1
Name: WAR, dtype: int64
-----------------------------------
0    0.663366
1    0.168317
2    0.138614
3    0.019802
4    0.009901
Name: WAR, dtype: float64
```

CONCLUSIONS

- Overall baseball is an extremely complex sport to analyze statistically through modeling. 1 vs 9 for offense and defense is near impossible to represent mathematically.
- Statistical trends are evident however creating accurate models for those trends requires extensive turning.
- Top performers in baseball are statistical outliers, and even single at-bats or single game performances can increase or decrease a players value significantly, would like to figure out the ‘TrueSkill’ library and add weighted player metrics moving forward.

THANK YOU.

Will VanDerKloot