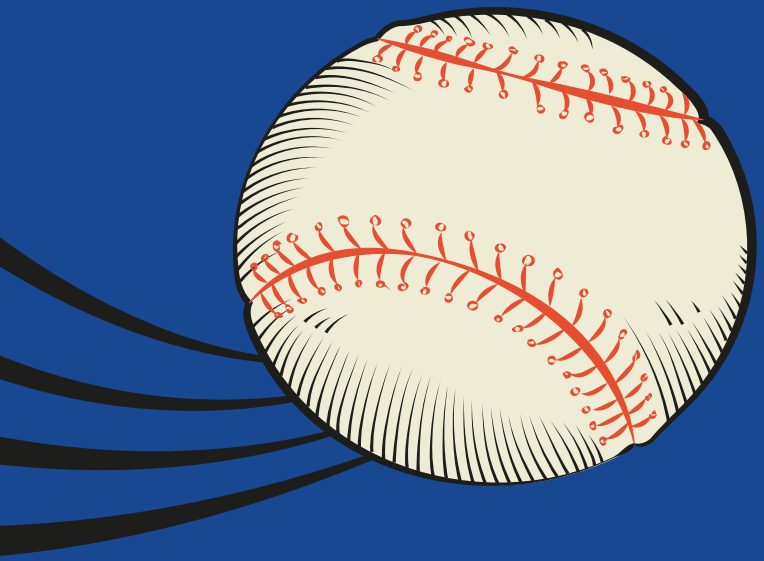


Baseball Hall of Fame Prediction

Alan Wong, Jeffrey Gutierrez, Nelson Duong,
Peter DePaul, Yuji Kusuyama

What is Hall of Fame?

The Baseball Hall of Fame in Cooperstown, New York, honors players, managers, umpires, and executives for their significant contributions to baseball.

Candidates, eligible five years after retirement, are voted on by the Baseball Writers' Association of America, with induction requiring at least 75% of the votes. The Hall features plaques and a museum highlighting baseball history and memorabilia.

Goal:

Using the available datasets, we want to design, refine, and implement a predictive model that can accurately assess whether a given player will be inducted to the Hall of Fame based on baseball related statistics.

Hypothesis:

There is a statistically significant correlation between a baseball player's career statistics and their likelihood of being inducted into the Hall of Fame, for both batters and pitchers.

Process:

Data Cleaning + EDA

We will clean, validate and explore the data at hand and to better understand how to use it to achieve our goal.

Model Refinement

We ranked our models using assessment metrics and refined until we've reached an optimal point.

Model Building

We preprocessed our data and split into both train and test data so we can build out our candidate models.

Model Deployment

We conclude by deploying our model, interpreting our findings, discussing our findings, and recommendations.

Data Cleaning & EDA

Data Cleaning

- Utilized Lahman package for R. Database of 27 baseball data tables.
- Utilized 7 of the data tables within database to create 4 total .csv files for analysis.
- Implemented several changes for final tables used
  - Subsetted Batters and Pitchers
  - Subsetted Batters (by primary position) and Pitchers (Starter/Reliever)
  - Calculated career batting and pitching stats (for both Hall of Fame and Active players)
  - Combined Hall of Fame voting results with Hall of Fame Career Stats
  - Calculated Career Awards and added them to Batting and Pitching datasets

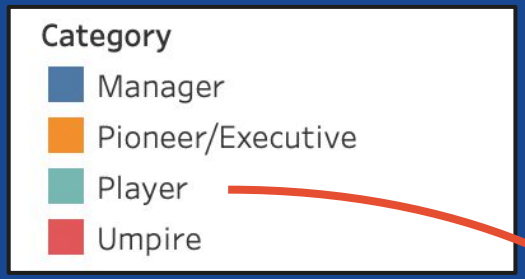
Our Data

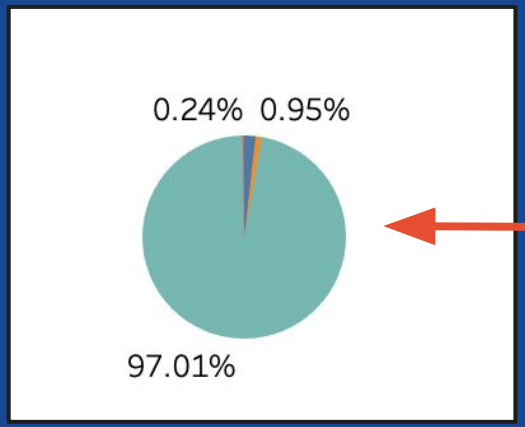
active_batting	**50** variables	**179** observations
hof_batting	**50** variables	**826** observations
active_pitching	**44** variables	**160** observations
hof_pitching	**44** variables	**432** observations

Batting : "playerID", "G", "AB", "R", "H", "X2B", "X3B", "HR", "RBI", "BB", "SO", etc

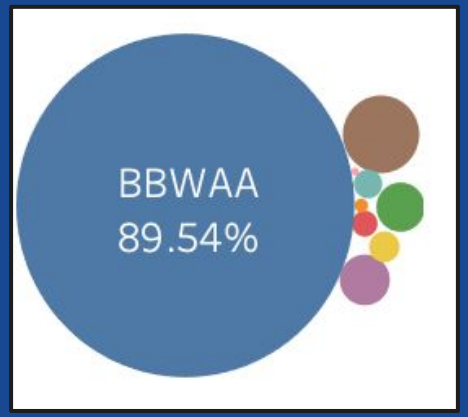
Pitching : "playerID", "POS", "W", "L", "G", "GS", "CG", "SHO","H", "ER", "HR", etc.

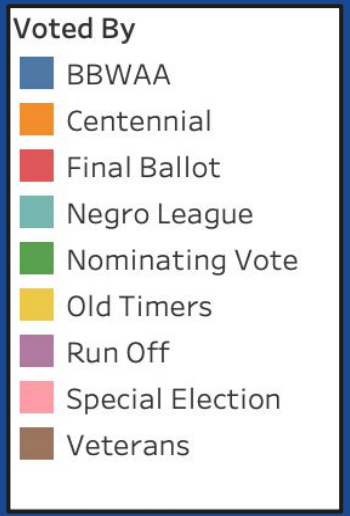
Hall of Fame by Category



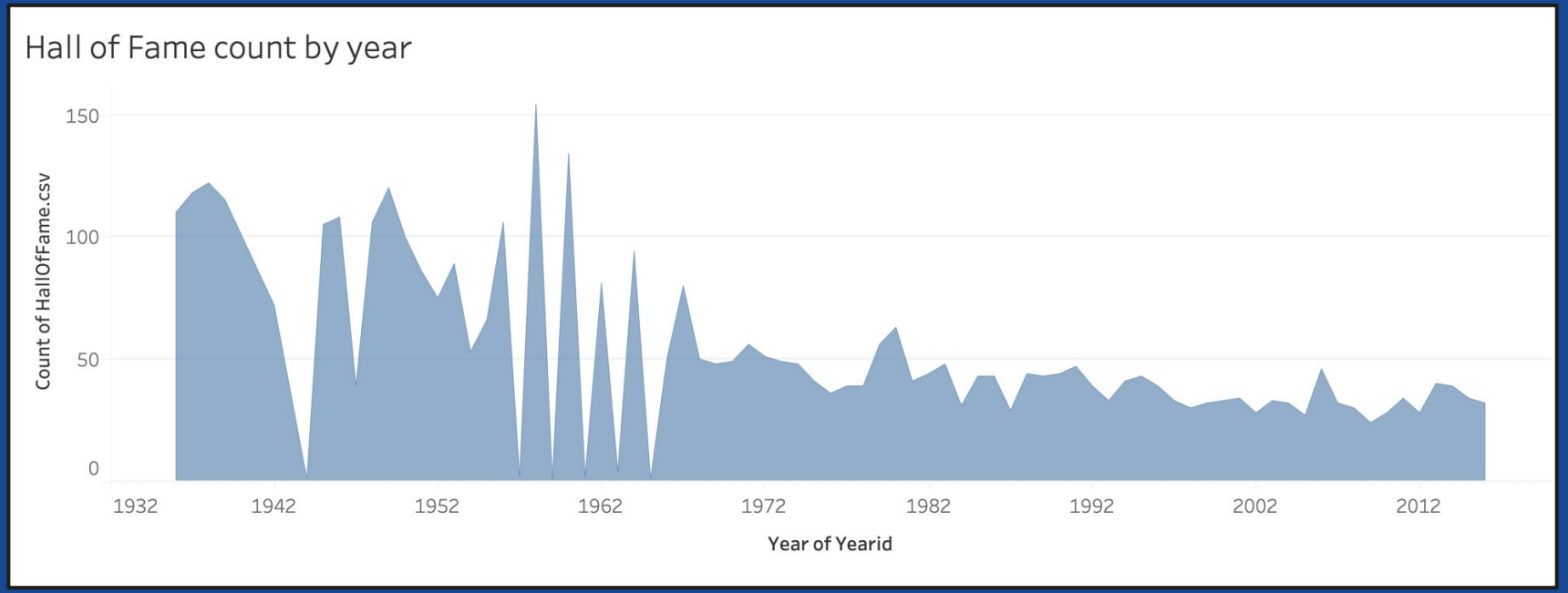


Who voted the most?

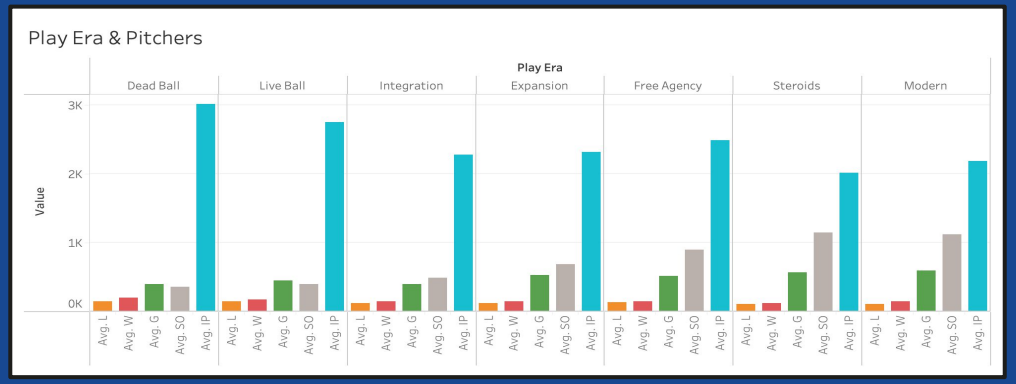


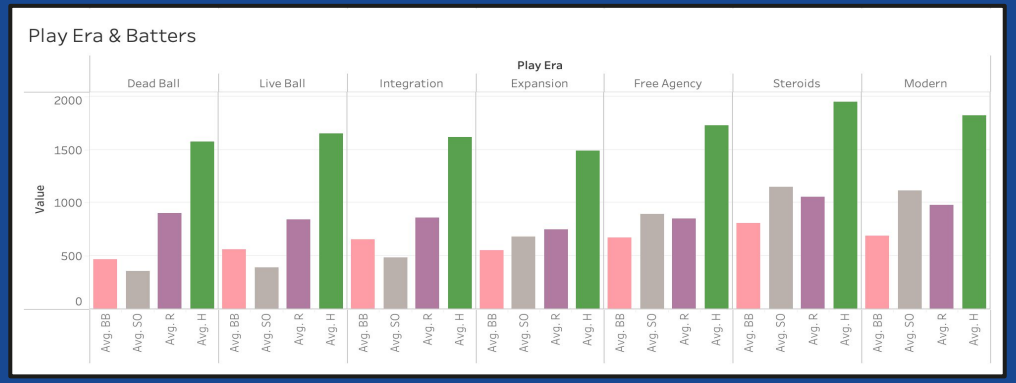


HOF Distribution

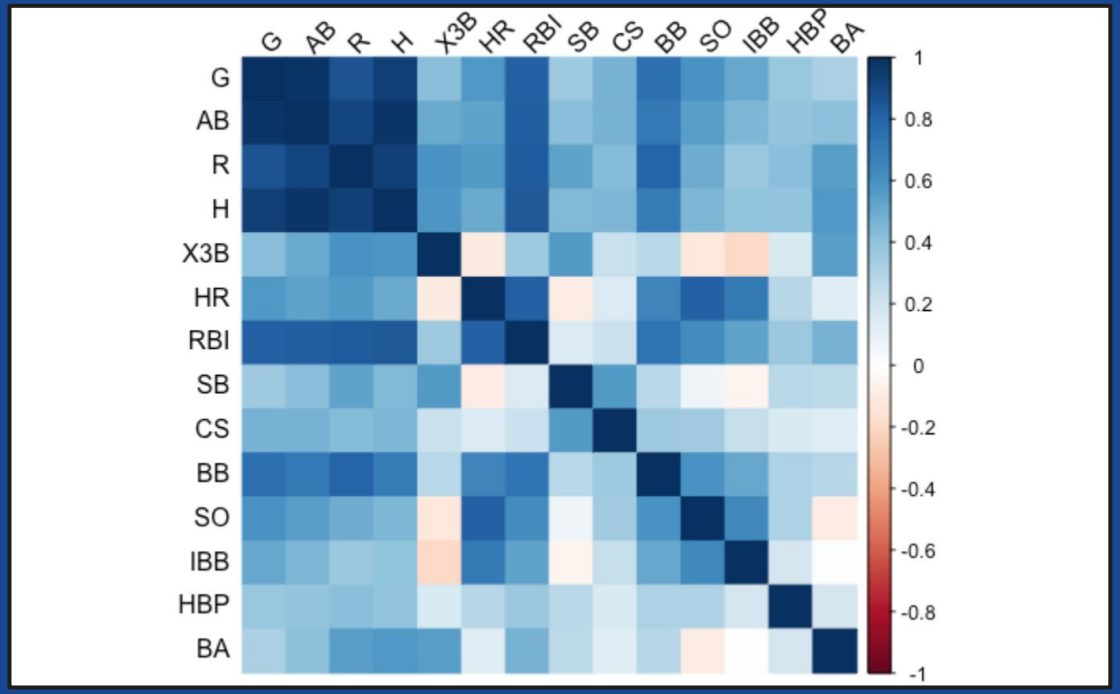


By Play Era and Player Type





active _batting



Model Building, Refinement, and Deployment

Feature Engineering

- Added several different variables to the data set
- Primarily indicator variables (Yes or No) for Career Milestones
  - 3,000 Career Hits (H)
  - 500 career Home Runs (HR)
  - 2,500 Games Played (G)
  - 1,500 Runs Batted In (RBI)
  - 1,500 Runs Scored ®
  - 300 Wins (pitching)
  - 3,000 Strikeouts (SO - pitching)
  - 300 Saves (SV - pitching)
- Added several rate statistics from the existing data
  - Batting Average (BA)
  - On Base Percentage (OBP)
  - On Base plus Slugging (OPS)

Models:

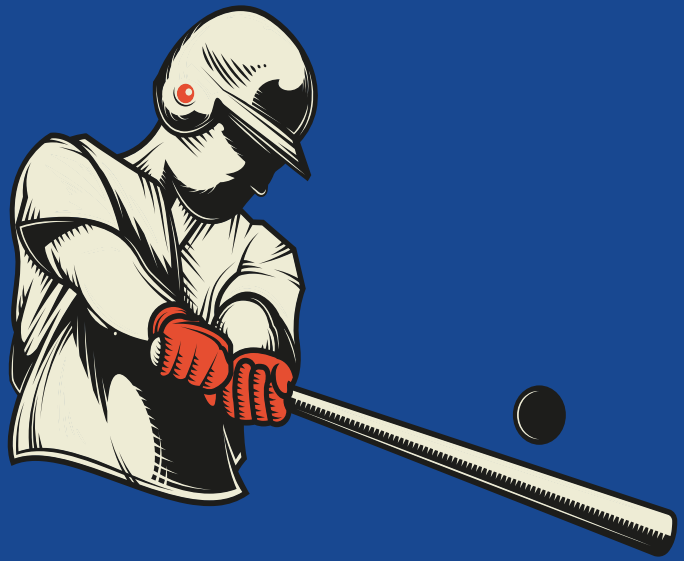
- We implemented 2 models on each of the datasets

Random Forest For Batters

Boosted Tree w/ XGBoost For Batters

Random Forest For Pitchers

Boosted Tree w/ XGBoost For Pitchers



Batting

Random Forest Results

Cross Validation Results:

- Mean accuracy = 0.878
- Standard Error = 0.00852

Results of Confusion Matrix

- Accuracy = 0.9084
- Sensitivity = 0.6977 (True Positive Rate)
- Specificity = 0.9519 (True Negative Rate)

```
> rf_conf_mat
Confusion Matrix and Statistics


     0   1
  0 198  13
  1  10  30

               Accuracy : 0.9084
                 95% CI : (0.8657, 0.941)
    No Information Rate : 0.8287
    P-Value [Acc > NIR] : 0.0002355

                  Kappa : 0.6681

 Mcnemar's Test P-Value : 0.6766573

            Sensitivity : 0.6977
            Specificity : 0.9519
         Pos Pred Value : 0.7500
         Neg Pred Value : 0.9384
             Prevalence : 0.1713
         Detection Rate : 0.1195
   Detection Prevalence : 0.1594
      Balanced Accuracy : 0.8248

       'Positive' Class : 1
```

Boosted Tree Results

Cross Validation Results:

- Mean accuracy = 0.880
- Standard Error = 0.00870

Results of Confusion Matrix

- Accuracy = 0.9044
- Sensitivity = 0.7209 (True Positive Rate)
- Specificity = 0.9423 (True Negative Rate)

```
> boost_conf_mat
Confusion Matrix and Statistics


     0   1
  0 196  12
  1  12  31

               Accuracy : 0.9044
                 95% CI : (0.8611, 0.9378)
    No Information Rate : 0.8287
    P-Value [Acc > NIR] : 0.0004819

                  Kappa : 0.6632

 Mcnemar's Test P-Value : 1.0000000

            Sensitivity : 0.7209
            Specificity : 0.9423
         Pos Pred Value : 0.7209
         Neg Pred Value : 0.9423
             Prevalence : 0.1713
         Detection Rate : 0.1235
   Detection Prevalence : 0.1713
      Balanced Accuracy : 0.8316

       'Positive' Class : 1
```

Incorrect Batter Predictions Random Forest

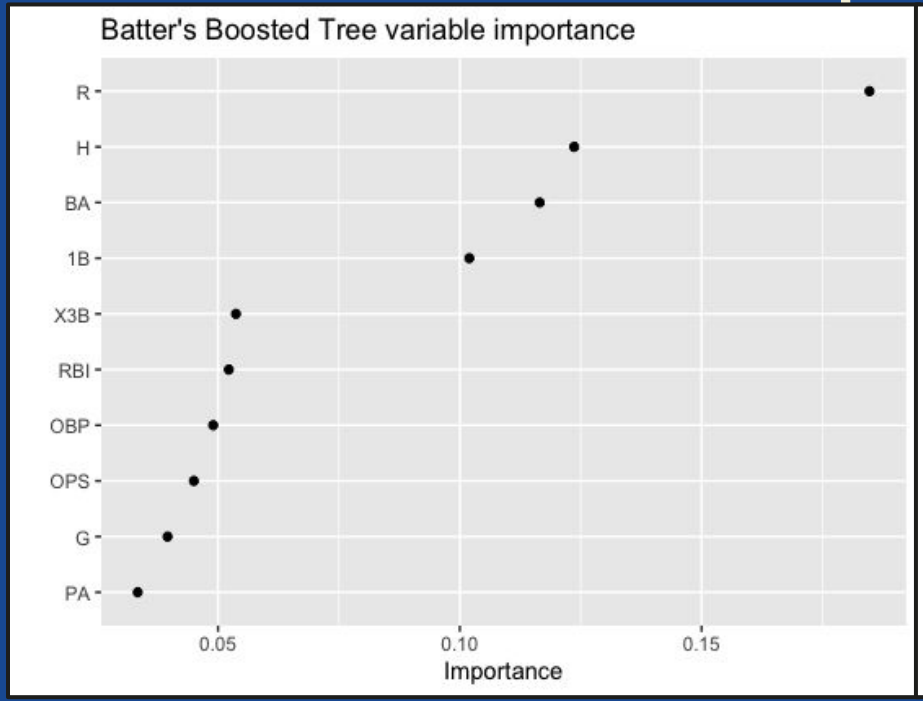
	name	WAR	POS	inducted	play_era	BA	OBP	SLG	OPS	HR	RBI	`SO%`	`BB%`	`BB:SO`	Range
	<chr>	<dbl>	<chr>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Jeff Bagwell	79.9	1B	1	Steroids	0.297	0.408	0.540	0.948	449	1529	0.165	0.149	0.899	8.95
2	Pete Rose	79.6	LF	0	Free Agency	0.303	0.375	0.409	0.784	160	1314	0.0723	0.0991	1.37	4.42
3	Johnny Bench	75.1	C	1	Free Agency	0.267	0.342	0.476	0.817	389	1376	0.148	0.103	0.697	5.40
4	Kenny Lofton	68.4	CF	0	Modern	0.299	0.372	0.423	0.794	130	781	0.111	0.103	0.930	2.38
5	Ryne Sandberg	67.9	2B	1	Steroids	0.285	0.344	0.452	0.795	282	1061	0.136	0.0823	0.604	4.87
6	Jackie Robinson	63.8	2B	1	Integration	0.311	0.409	0.474	0.883	137	734	0.0511	0.130	2.54	4.87
7	Lou Boudreau	63.2	SS	1	Integration	0.295	0.380	0.415	0.795	68	789	0.0451	0.116	2.58	4.95
8	Sammy Sosa	58.6	RF	0	Steroids	0.273	0.344	0.534	0.878	609	1667	0.233	0.0940	0.403	1.98
9	Johnny Damon	56.3	CF	0	Modern	0.284	0.352	0.433	0.785	235	1139	0.116	0.0924	0.798	1.93
10	Jeff Kent	55.5	2B	0	Modern	0.290	0.356	0.500	0.855	377	1518	0.160	0.0841	0.526	4.67

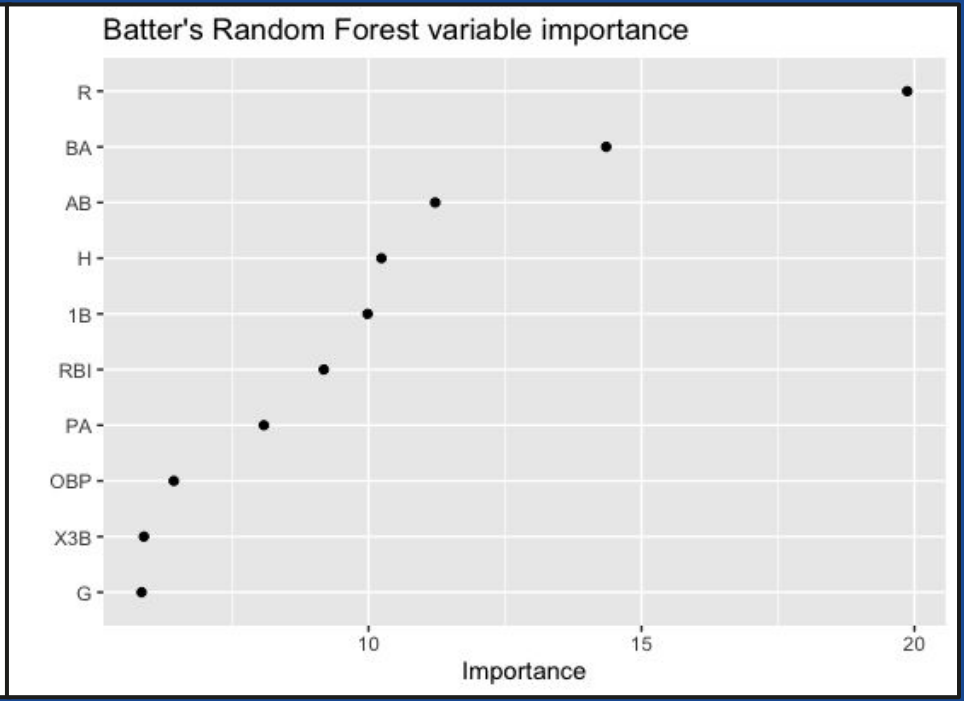
Active Batter Predictions

	Name	WAR	BA	OBP	SLG	OPS	HR	RBI	`SO%`	`BB%`	`BB:SO`	Range
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Albert Pujols	102.	0.296	0.374	0.544	0.918	703	2218	0.108	0.105	0.978	6.32
2	Adrian Beltre	93.5	0.286	0.339	0.480	0.819	477	1707	0.143	0.0700	0.490	2.52
3	Carlos Beltran	70.1	0.279	0.350	0.486	0.837	435	1587	0.163	0.0984	0.604	2.05
4	Robinson Cano	68.1	0.301	0.351	0.488	0.839	335	1306	0.127	0.0650	0.511	4.43
5	Miguel Cabrera	67.3	0.308	0.384	0.524	0.908	507	1847	0.178	0.107	0.604	4.57
6	Ichiro Suzuki	60	0.311	0.355	0.402	0.757	117	780	0.101	0.0606	0.599	1.94

	name	WAR	BA	OBP	SLG	OPS	HR	RBI	`SO%`	`BB%`	`BB:SO`	Range
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Mike Trout	85.2	0.303	0.415	0.587	1.00	350	896	0.220	0.149	0.679	2.30

Batter's Model Feature Importance





Pitching

Random Forest Results

Cross Validation Results:

- Mean accuracy = 0.917
- Standard Error = 0.0134

Results of Confusion Matrix

- Accuracy = 0.8779
- Sensitivity = 0.4737 (True Positive Rate)
- Specificity = 0.9464 (True Negative Rate)

```
Confusion Matrix and Statistics


    0   1
0 106  10
1   6   9

               Accuracy : 0.8779
                 95% CI : (0.8092, 0.9285)
    No Information Rate : 0.855
    P-Value [Acc > NIR] : 0.2737

                  Kappa : 0.4604

 Mcnemar's Test P-Value : 0.4533

            Sensitivity : 0.4737
            Specificity : 0.9464
         Pos Pred Value : 0.6000
         Neg Pred Value : 0.9138
             Prevalence : 0.1450
         Detection Rate : 0.0687
   Detection Prevalence : 0.1145
      Balanced Accuracy : 0.7101

       'Positive' Class : 1
```

Boosted Tree Results

Cross Validation Results:

- Mean accuracy = 0.910
- Standard Error = 0.0116

Results of Confusion Matrix

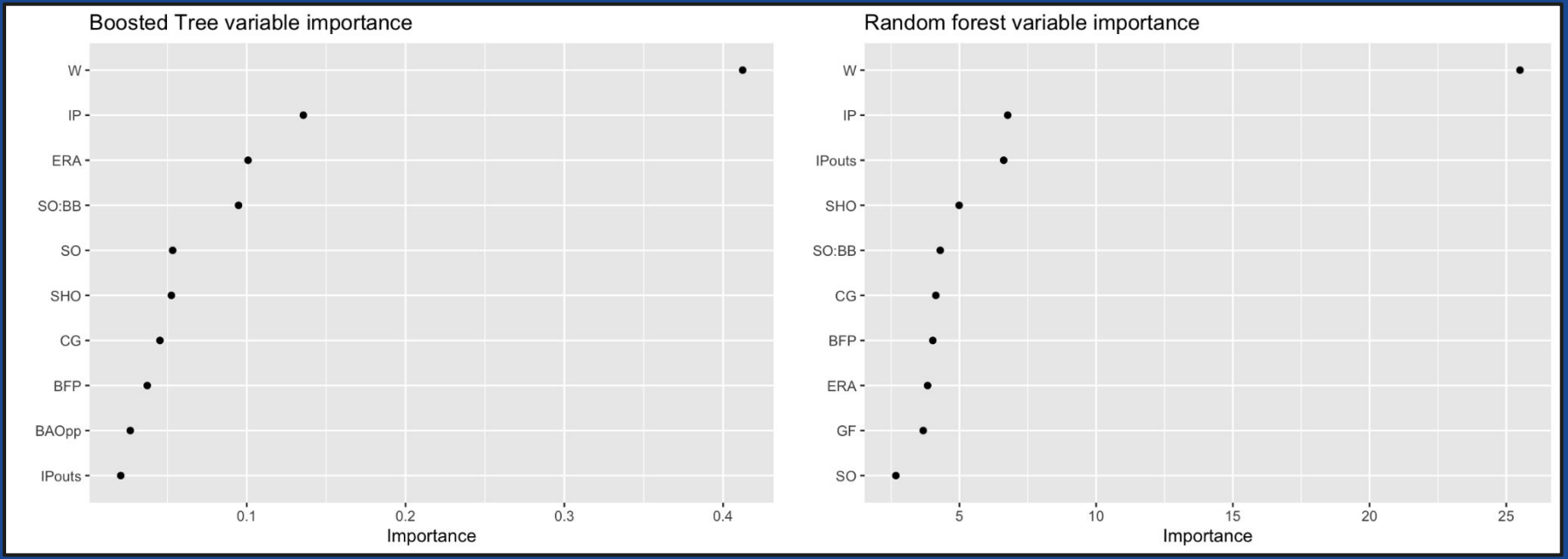
- Accuracy = 0.9084
- Sensitivity = 0.52632 (True Positive Rate)
- Specificity = 0.97321 (True Negative Rate)

```
> boost_conf_mat
Confusion Matrix and Statistics


    0   1
 0 109   9
 1   3  10

               Accuracy : 0.9084
                 95% CI : (0.8455, 0.9518)
    No Information Rate : 0.855
    P-Value [Acc > NIR] : 0.04735

                  Kappa : 0.5749

 Mcnemar's Test P-Value : 0.14891

            Sensitivity : 0.52632
            Specificity : 0.97321
         Pos Pred Value : 0.76923
         Neg Pred Value : 0.92373
             Prevalence : 0.14504
         Detection Rate : 0.07634
   Detection Prevalence : 0.09924
      Balanced Accuracy : 0.74977

       'Positive' Class : 1
```

Active Pitcher Predictions

	name	POS	WAR	W	IP	ERA	BAOpp	SO	`SO%`	`BB%`
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Justin Verlander	SP	80.9	244	3163.	3.24	0.206	3198	0.248	0.0683
2	Clayton Kershaw	SP	79.9	197	2581.	2.48	0.192	2807	0.276	0.0619
3	Zack Greinke	SP	77.5	223	3247	3.42	0.230	2882	0.217	0.0556
4	Max Scherzer	SP	75	201	2682.	3.11	0.201	3193	0.295	0.0647
5	CC Sabathia	SP	62.3	251	3577.	3.74	0.227	3093	0.206	0.0733
6	Bartolo Colon	SP	46.2	247	3461.	4.12	0.245	2535	0.173	0.0647

Pitcher's Model Feature Importance



Conclusion

Thanks!!