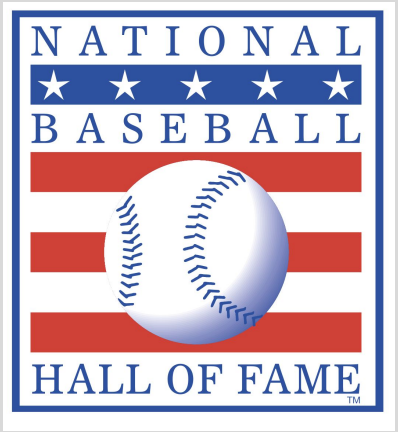
Using Classification Methods to Analyze the Baseball Hall of Fame

Jonathan Eman

Baseball Background

Only **32** non-pitchers have been inducted into the Hall of Fame in their first year of eligibility since 1970

Hank Aaron



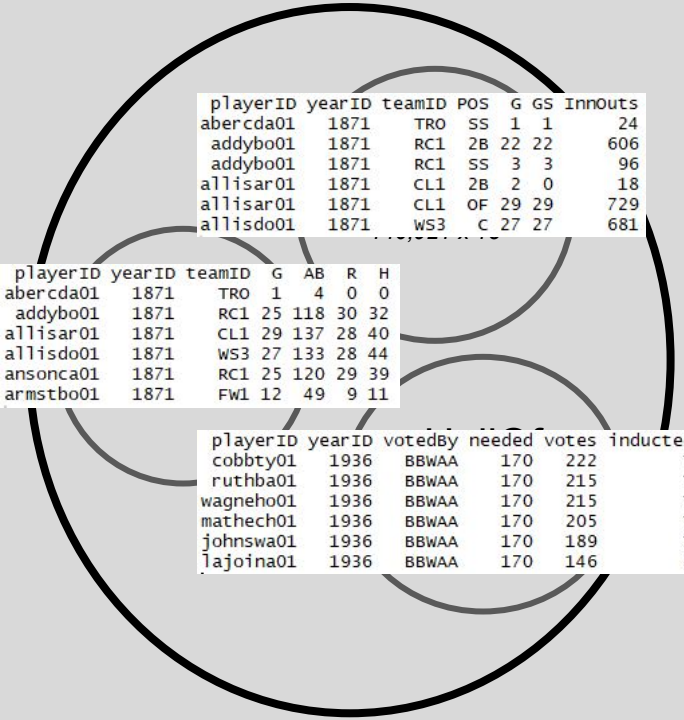
Chipper Jones

Question of Interest

Can we predict whether or not a player will be inducted into the Baseball Hall of Fame (HOF) in their first year of eligibility based on various career batting statistics?

Data Processing & Cleaning

Lahman Package



playerID	yearID	teamID	POS	G	GS	InnOuts
abercda01	1871	TRO	SS	1	1	24
addybo01	1871	RC1	2B	22	22	606
addybo01	1871	RC1	SS	3	3	96
allisar01	1871	CL1	2B	2	0	18
allisar01	1871	CL1	OF	29	29	729
allisdo01	1871	WS3	C	27	27	681

playerID	yearID	teamID	G	AB	R	H
abercda01	1871	TRO	1	4	0	0
addybo01	1871	RC1	25	118	30	32
allisar01	1871	CL1	29	137	28	40
allisdo01	1871	WS3	27	133	28	44
ansonca01	1871	RC1	25	120	29	39
armstbo01	1871	FW1	12	49	9	11

playerID	yearID	votedBy	needed	votes	inducted
cobbty01	1936	BBWAA	170	222	Y
ruthba01	1936	BBWAA	170	215	Y
wagneho01	1936	BBWAA	170	215	Y
mathech01	1936	BBWAA	170	205	Y
johnswa01	1936	BBWAA	170	189	Y
lajoina01	1936	BBWAA	170	146	N

Filters

1. Year >1970
2. Years.played > 9
3. Position ≠ Pitcher
4. VotedBy = Baseball Writers' Association of America

Final dataset

playerID	career.hr	career.rbi	career.h	career.ba
	<int>	<int>	<int>	<dbl>
berrayo~	358	1430	2150	0.285
kinerra~	369	1015	1451	0.279
hodgegi~	370	1274	1921	0.273
slaugen~	169	1304	2383	0.300
mizejo01	359	1337	2011	0.312
reesepe~	126	885	2170	0.269

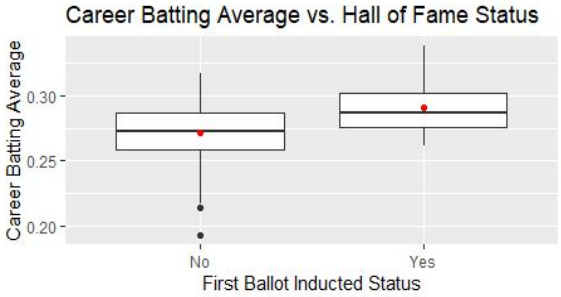
Lahman Baseball Data

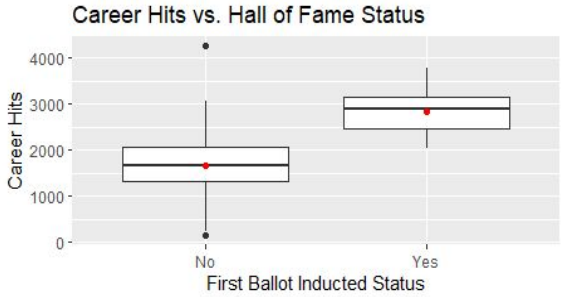
Explanatory Variables Used

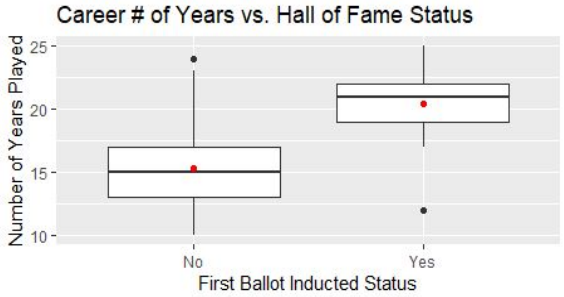
- Career.ba: Batting Average
- Career.h: Hits
- Career.rbi: Runs Batted In
- Career.hr: Homeruns
- Career.tb: Total Bases
- Career.r: Total Runs
- Years.played: Number of years the player competed as a professional

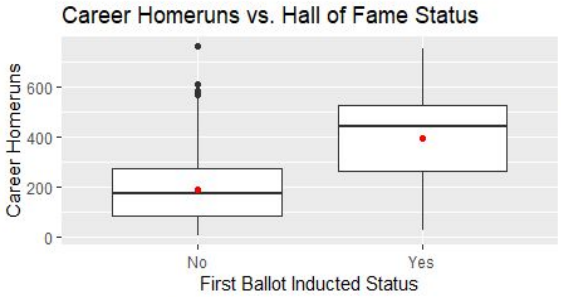
Response Variable

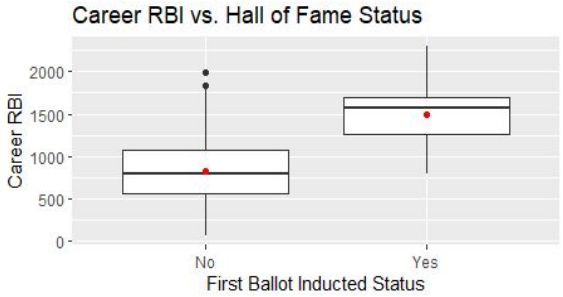
Inducted: “Y” or “N”

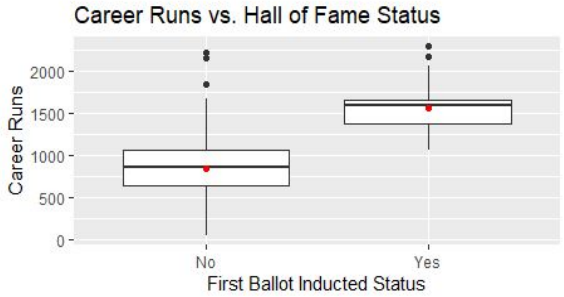

Career Batting Average vs. Hall of Fame Status
Career Batting Average
0.30
0.25
0.20
No
Yes
First Ballot Inducted Status

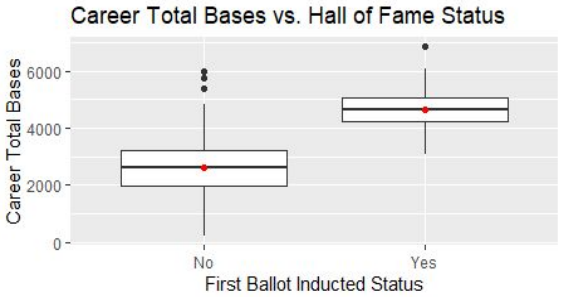

Career Hits vs. Hall of Fame Status
Career Hits
4000
3000
2000
1000
0
No
Yes
First Ballot Inducted Status


Career # of Years vs. Hall of Fame Status
Number of Years Played
25
20
15
10
No
Yes
First Ballot Inducted Status


Career Homeruns vs. Hall of Fame Status
Career Homeruns
600
400
200
0
No
Yes
First Ballot Inducted Status


Career RBI vs. Hall of Fame Status
Career RBI
2000
1500
1000
500
0
No
Yes
First Ballot Inducted Status


Career Runs vs. Hall of Fame Status
Career Runs
2000
1500
1000
500
0
No
Yes
First Ballot Inducted Status


Career Total Bases vs. Hall of Fame Status
Career Total Bases
6000
4000
2000
0
No
Yes
First Ballot Inducted Status

Summary of Modeling Process

<u>Classification Model Building:</u>

- Logistic Regression
  - With 7 quantitative variables
  - With 5 quantitative variables
  - With 4 quantitative variables, 1 categorical
- Linear Discriminant Analysis

<u>Classification Tree Methods:</u>

- Recursive Binary Splitting
  - Pruning
- Bagging
- Random Forests

Addressing Issue of Multicollinearity

Correlation Matrix

Variance Inflation Factor

career.ba	career.h	career.hr	career.r	career.rbi	career.tb
2.495097	88.231426	54.006408	7.695797	24.223646	174.412916

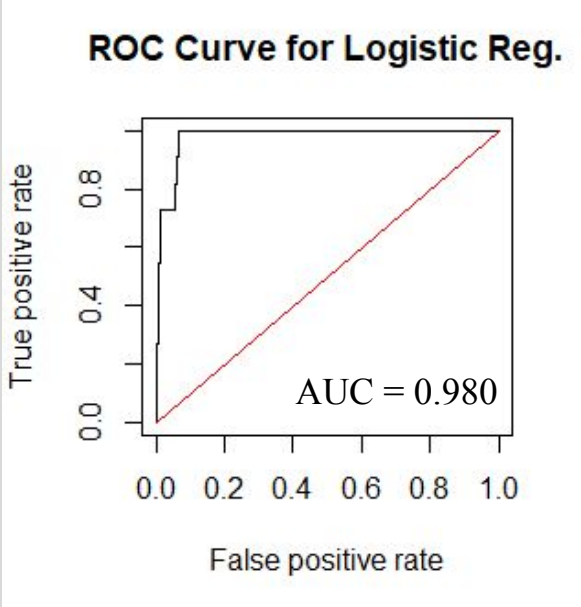
years.played
2.679751

career.ba	career.h	career.hr	career.r	career.rbi	years.played
2.383713	13.284919	19.110770	5.885484	18.486861	2.520309

career.ba	career.h	career.r	career.rbi	years.played
2.375121	5.016877	2.457789	1.425734	2.497270

Logistic Regression Summary

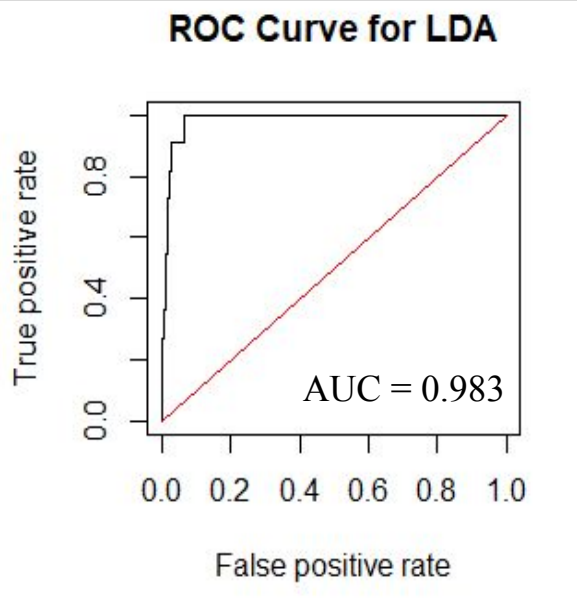
$$p(x) = \frac{e^{-17.70 + 24.64(career.ba) + 0.000814(career.h) + 0.002205(career.r) + 0.002705(career.rbi) + 0.0432(years.played)}}{1+e^{-17.70 + 24.64(career.ba) + 0.000814(career.h) + 0.002205(career.r) + 0.002705(career.rbi) + 0.0432(years.played)}}$$

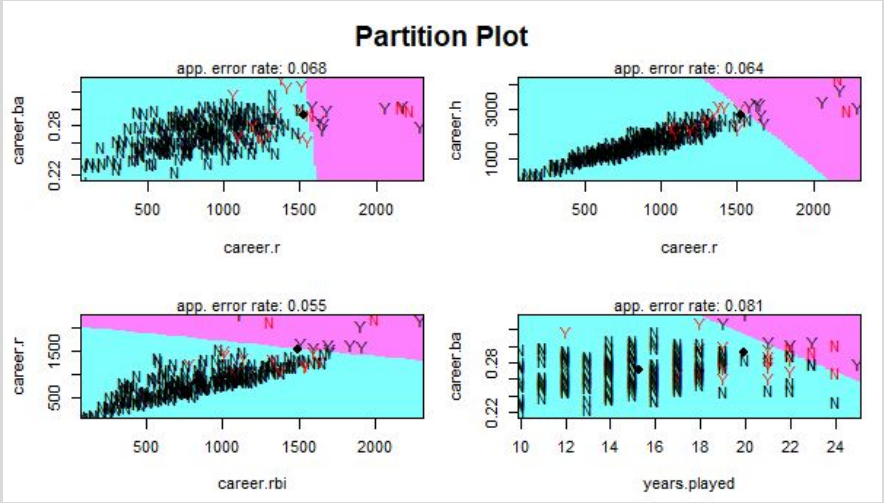


Confusion Matrix for Test Data		
	No - Predicted	Yes - Predicted
No - Actual	223	2
Yes - Actual	4	7

False Positive Rate = 0.88%
False Negative Rate = 36.36%

LDA Summary





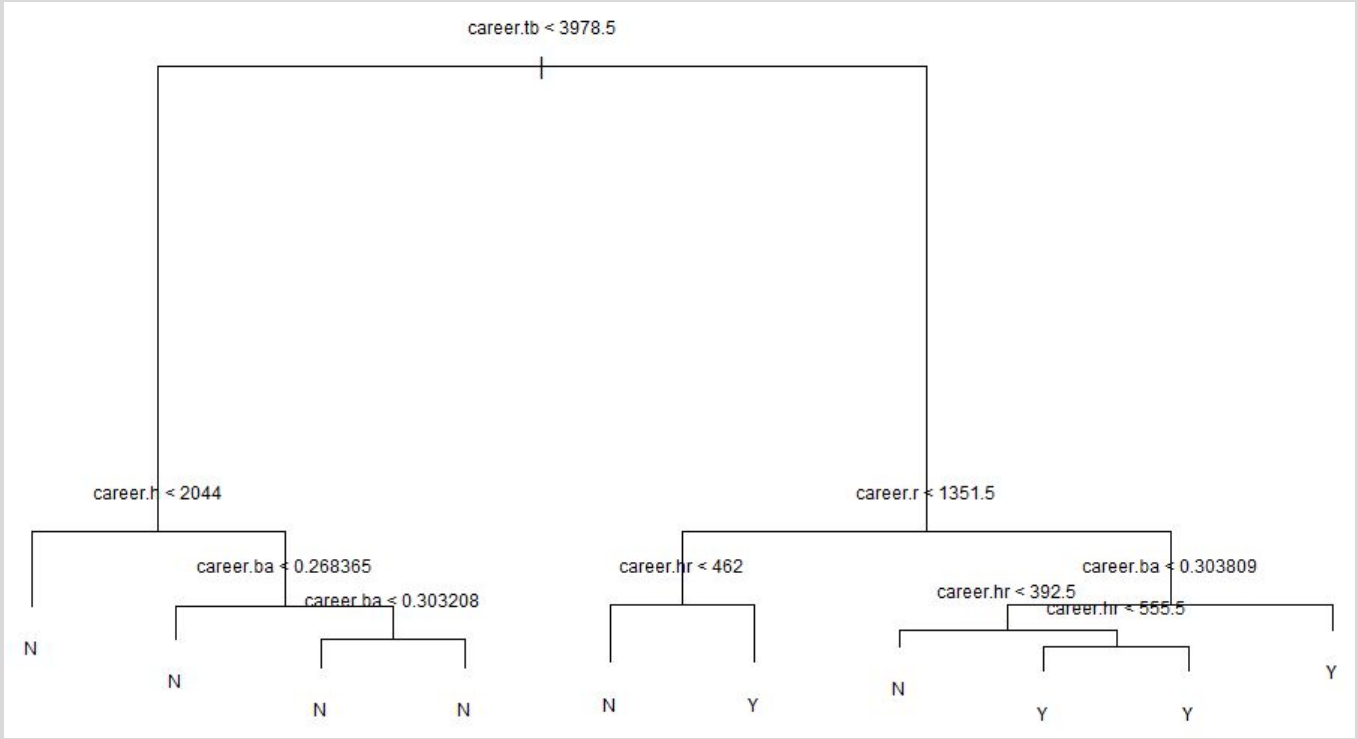
Confusion Matrix for Test Data		
	No - Predicted	Yes - Predicted
No - Actual	221	4
Yes - Actual	3	8

False Positive Rate = 1.78%
False Negative Rate = 27.27%

Comparison of Classification Models

	Logistic Regression	LDA
Overall Test Error Rates from 10-fold CV	**4.3%**	4.7%
False Positive Rate	**0.88%**	1.78%
False Negative Rate	36.36%	**27.27%**
AUC	0.980	**0.983**
Classification for Chipper Jones	Yes: p(X) = 0.684	Yes: p(X) = 0.591
Classification for Derek Jeter	Yes: p(X) = 0.803	Yes: p(X) = 0.822

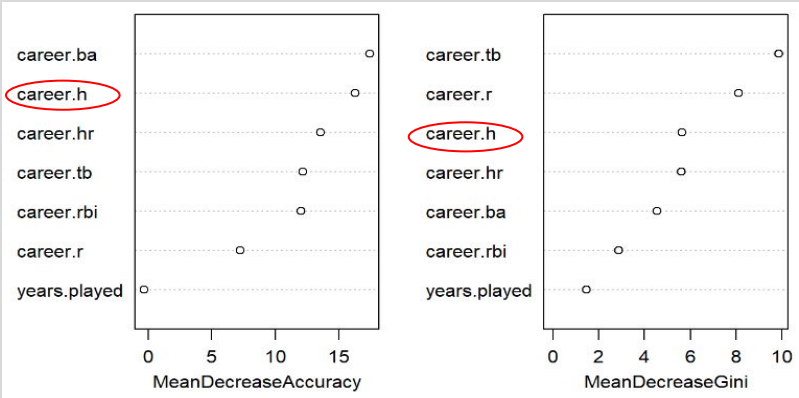
Classification Tree: Recursive Binary Splitting

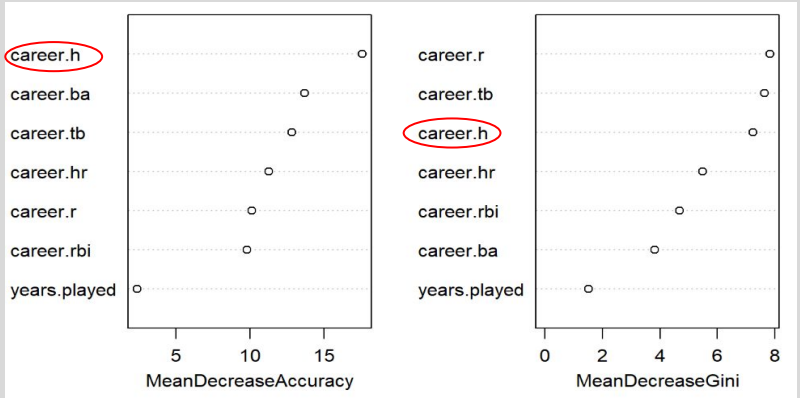


Confusion Matrix for Test Data		
	No - Predicted	Yes - Predicted
No - Actual	218	7
Yes - Actual	4	7

False Positive Rate = 3.11%
False Negative Rate = 36.36%

Tree Improvement: Bagging & Random Forests





Confusion Matrix for Bagging		
	No - Predicted	Yes - Predicted
No - Actual	218	7
Yes - Actual	2	9

False Positive Rate = 3.11%
False Negative Rate = 18.18%

Confusion Matrix for Random Forests		
	No - Predicted	Yes - Predicted
No - Actual	219	6
Yes - Actual	2	9

False Positive Rate = 2.75%
False Negative Rate = 18.18%

Comparison of Tree Methods

	Pruned Classification Tree	Bagging	Random Forest
Overall Error Rates	4.7%	3.8%	**3.4%**
False Positive Rate	3.11%	3.11%	**2.75%**
False Negative Rate	36.36%	**18.18%**	**18.18%**
Most Important Predictors	**Career Total Bases** Career Hits Career Runs	**Career Hits** Career Home Runs Career Total Bases	**Career Hits** Career Batting Average Career Total Bases
Classification for Chipper Jones	Yes	Yes	Yes
Classification for Derek Jeter	Yes	Yes	Yes

Error Analysis

No model was able to accurately classify more than 82% of First Ballot HOF Inductees. ***How can this problem be further addressed?***

Brooks Robinson (Class of 1983)

HR: 268
BA: .267
Hits: 2848
TB: 4270

18x All-Star
Whole Career for Orioles
16x Gold Glove Winner
Highest Career Fielding % for 3B

Joe Morgan (Class of 1990)

HR: 268
BA: .271
Hits: 2517
TB: 3962

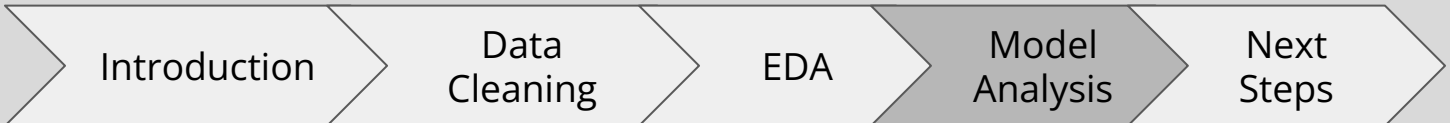
10x All-Star
2x World Series Champ
5x Gold Glove Winner
11th Most SB of All-Time

HOF Average

HR: 397
BA: .290
Hits: 2827
TB: 4663

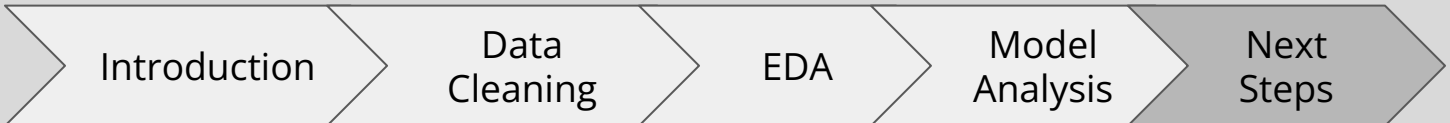
Limitations of Models

- As seen in the previous slide, the models **need to account for more than just batting statistics** to comprehensively reflect a player's entire career
- Interpretability
  - With logistic regression and tree from recursive binary splitting, it is easy to come to a conclusion about 1 particular player
  - LDA, bagging and random forests are more **computationally expensive**
- Modeling assumptions
  - The extent to which the assumption of multivariate normality can be violated is arbitrary
  - Multicollinearity required the removal of important variables (i.e. TB) from logistic and LDA models



Summary of Analysis

- **Random Forests had the lowest error rates** - all methods from logistic regression to decision trees produced very low overall error rates, but **false negative rates were significantly higher**
- Future model development should include **addition of non-batting variables** (i.e. awards, fielding stats, steroid use)
- **Hits, total bases and runs scored** were generally the most influential variables, while **years played and RBIs** were generally the least influential
  - Players with 2,800+ career hits and 4,000+ TBs can generally expect to be inducted into the Hall of Fame
  - A player with a .310 career BA and 250 HRs not necessarily less likely than a player with .310 and 450 HRs
- Our models **accurately predicted that Derek Jeter would be inducted** into the Hall of Fame in his first year of eligibility in 2020



Acknowledgements

- All data analysis done with R programming language
- Data comes from Lahman package
- Data wrangling and visualization done with the tidyverse
- Other packages used for modeling: tree, randomForest, MASS, klaR, ICS, ROCR, boot, ipred