Machine Learning Final Project

Team Upsilon
Claire Duchene, Jessy Sun, George Wei, and Michael Uhrig

Raymond A. Mason School of Business
WILLIAM & MARY

The Business Background

Raymond A. Mason
School of Business
WILLIAM & MARY

- The “business” problem we addressed is a longstanding issue within sports industry: determining the future value of players, an organization’s most important assets which also the hardest to value

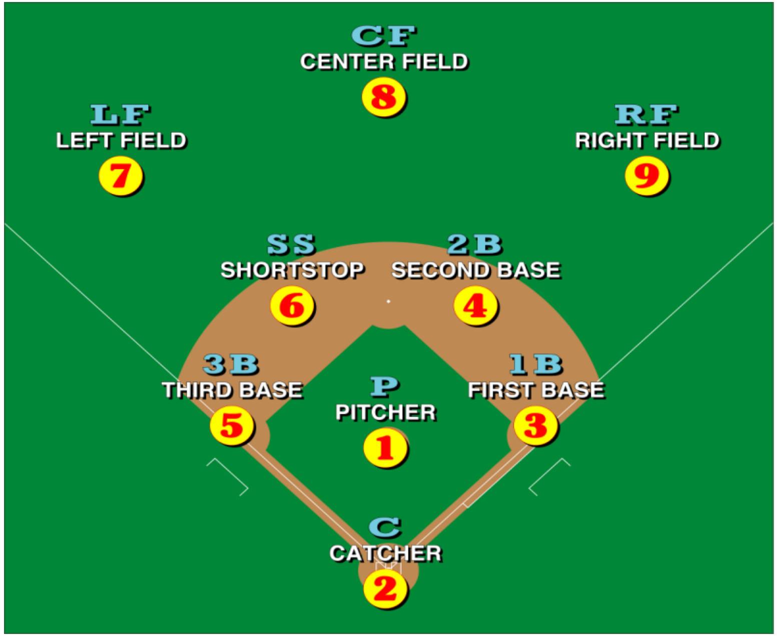
- This problem is important because if baseball clubs could better understand what statistical indicators drive future performance in players, they could have a better chance of winning and a better chance of increasing profitability, which in turn has been shown to further drive wins

Raymond A. Mason School of Business
WILLIAM & MARY

The Data

Sean 'Lahman' Baseball Database

Lahman-package	Sean Lahman's Baseball Database
AllstarFull	AllstarFull table
Appearances	Appearances table
AwardsManagers	AwardsManagers table
AwardsPlayers	AwardsPlayers table
AwardsShareManagers	AwardsShareManagers table
AwardsSharePlayers	AwardsSharePlayers table
Batting	Batting table
battingLabels	Variable Labels
BattingPost	BattingPost table
battingStats	Calculate additional batting statistics
CollegePlaying	CollegePlaying table
Fielding	Fielding table
fieldingLabels	Variable Labels
FieldingOF	FieldingOF table
FieldingPost	FieldingPost data
HallOfFame	Hall of Fame Voting Data
Label	Extract the Label for a Variable
Lahman	Sean Lahman's Baseball Database
LahmanData	Lahman Datasets
Managers	Managers table
ManagersHalf	ManagersHalf table
Master	Master table
Pitching	Pitching table
pitchingLabels	Variable Labels
PitchingPost	PitchingPost table
playerInfo	Lookup Information for Players and Teams
Salaries	Salaries table
Schools	Schools table
SeriesPost	SeriesPost table
teamInfo	Lookup Information for Players and Teams
Teams	Teams table
TeamsFranchises	TeamFranchises table
TeamsHalf	TeamsHalf table



The Aim of Project

- **Make predictions?**
  - Instead of simply make predictions, we hope to use our project to find out which model is the most suitable one to predict athletic performance, which represents a wide range of industries that show the promising applications of model selection
- **An old chinese saying: 授人以鱼不如授人以渔**
  - Giving a man a fish is not better than teaching him to fish
  - Hopefully, our model selection process will at least make some small contributions to the model selection and athlete performance prediction

The Process

Raymond A. Mason School of Business
WILLIAM & MARY

- **Step 1 - Step 2: Data Collection and Cleansing**
  - We gathered all the data from the Lahman database library in R, which ranges from years 1871-2016, imputed the missing value by function named “knn-Imputation”
  - Introduce K-fold cross validation to split our data into 10 equal pieces
- **Step 3 - Step 6: Model Building and Error Rate Collection**
  - Use logistic regression, LDA, QDA and KNN method respectively to build the model
  - Introduce the sum of type one error and false discovery proportion as the sum error rate to evaluate the accuracy of our model
- **Step 7: Model Comparison and Visualization**
  - Use line chart and boxplot to compare error rate and accuracy of four different models

Step 1: Data Collection

Raymond A. Mason School of Business
WILLIAM & MARY

- **We chose an existing dataset about baseball players in R package:**
  - http://www.seanlahman.com/baseball-archive/statistics/

What's our Pros?

- Very extensive (Over 100,000 data points)
- Contains cumulative statistics for each season
- Up to date
- Preloaded into R & Very well documented

On the other hand:

- Missing a lot of data
- Disjoint batting/fielding tables
- Missing zone rating data

Step 2: Data Cleansing

Raymond A. Mason School of Business
WILLIAM & MARY

- **Grouping Players batting and fielding statistics**
- **Summing each player's cumulative statistics over their careers**
- **Selecting and generating batting/fielding rates**
  - Hit percentage
  - Walk percentage
  - Fielding percentage
  - Position dummy variables
  - Slugging percentage
  - Home run percentage
  - Put-outs and assists per out
- **Imputing Data using Knn-Imputation**
- **Divide into groups using K-fold cross validation**

Step 2: Data Cleansing

- A general view of data cleansing: Executed by Alteryx
- Our datasets after data cleansing: Dimension 5000*16

D9 | 2B

	ID	name	year	POS	totalyears	Batpct	Slugpct	Walkpct	HRpct	fldpct	POAperInn	POS1	POS2	POS3	POS4	HOF
2	1	aardsda01	2004	P	9	0	0	0	0	0.93023	0.0396	0	0	0	1	FALSE
3	2	aaronha01	1954	OF	23	0.305	0.57037	0.11339	0.0611	0.98202	0.100302246	0	0	1	0	TRUE
4	3	aaronto01	1962	1B	7	0.22881	0.34004	0.0911	0.0138	0.98485	0.220951792	1	0	0	0	FALSE
5	4	aasedo01	1977	P	13	0	0	0	0	0.93953	0.0607	0	0	0	1	FALSE
6	5	abadan01	2001	1B	2	0.11111	0.11111	0.11111	0	0.97436	0.275362319	1	0	0	0	FALSE
7	6	abadfe01	2010	P	7	0.11111	0.11111	0	0	0.95	0.0462	0	0	0	1	FALSE
8	7	abadijo01	1875	1B	1	0.22449	0.22449	0	0	0.91034	0.200463173	1	0	0	0	FALSE
9	8	abbated01	1897	2B	9	0.25361	0.35348	0.0949	0.00361	0.93085	0.179910975	1	0	0	0	FALSE
10	9	abbeybe01	1892	P	5	0.16889	0.23556	0.0933	0	0.87283	0.066620665	0	0	0	1	FALSE

Step 2: Data Cleansing

Raymond A. Mason School of Business
WILLIAM & MARY

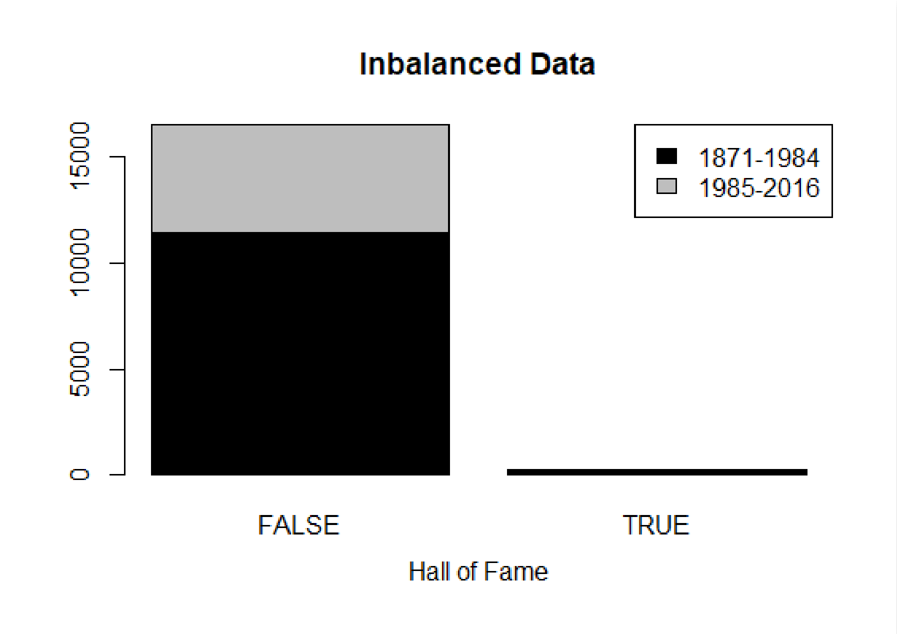
	FALSE	TRUE
FALSE	16461	15
TRUE	267	4

- ○ **Type One Error Rate: 267/271**
- ○ **False Discovery Proportion: 15/19**

- **Type one and false discovery proportion instead of overall error rate**
  - Players are an athletics organization most valuable assets, but their future values are hard to predict
  - We want to focus on the players who are able to join the hall of fame instead of those bench players

Step 2: Data Cleansing

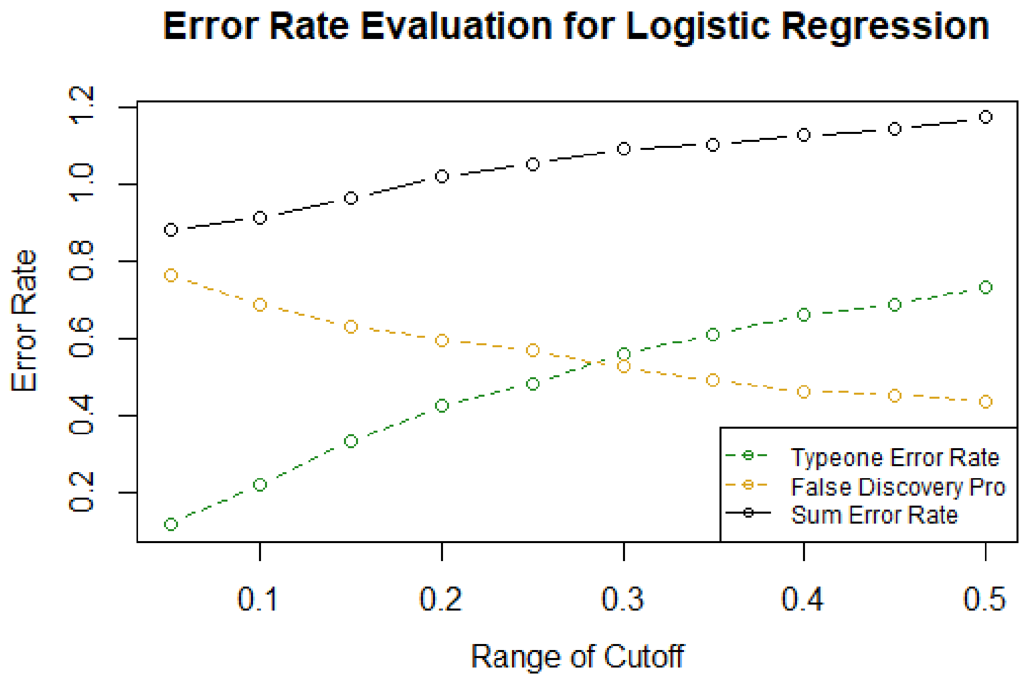
Raymond A. Mason School of Business
WILLIAM & MARY



- **Dealing with imbalanced data: too many players out of the hall of fame**
  - We deleted players before 1985 who aren’t in the hall of fame, then the proportion of null and alternate hypothesis closes
  - We sacrifice some rigorousness in exchange for accuracy of our model, kind of like the trade-off between variance and bias

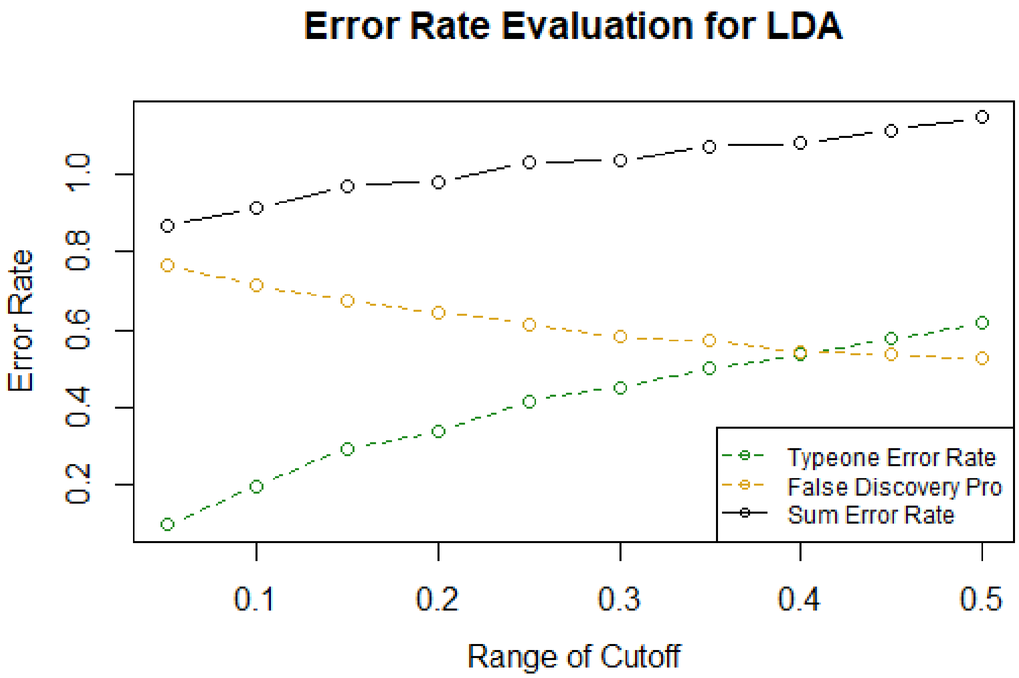
Step 3: Logistic Regression

Raymond A. Mason School of Business
WILLIAM & MARY



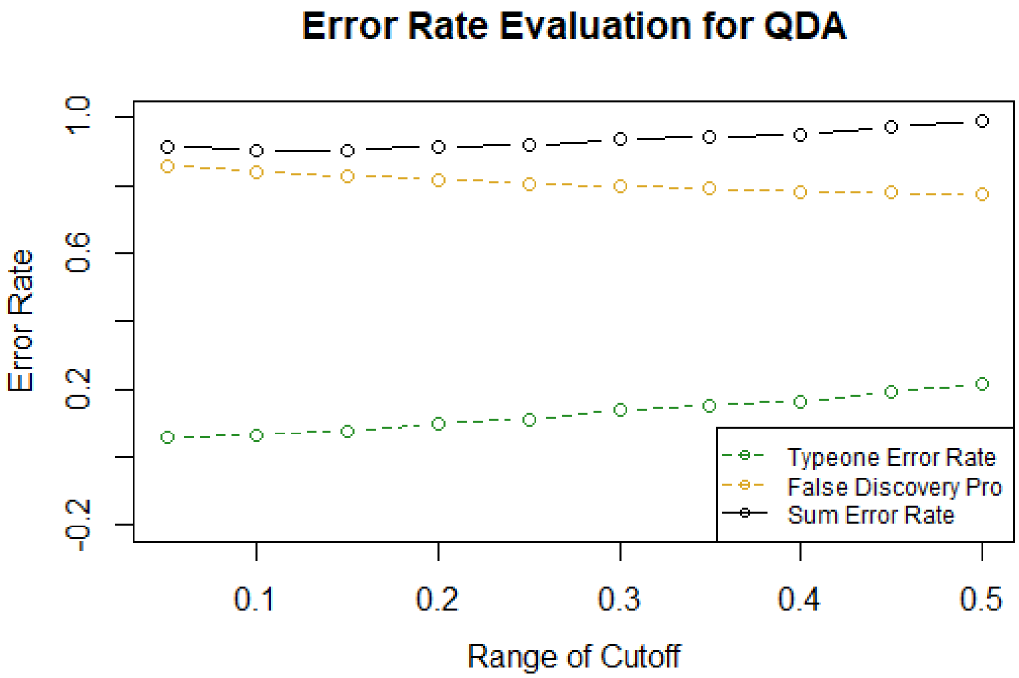
- We decreased the cutoff from 0.5 to 0.05
- Type one error rate greatly decreased following the decreasing cutoff
- However, false discovery proportion increased by almost the same level
- We still get a slightly lower sum error rate by decreasing cutoff

Step 4: LDA Model



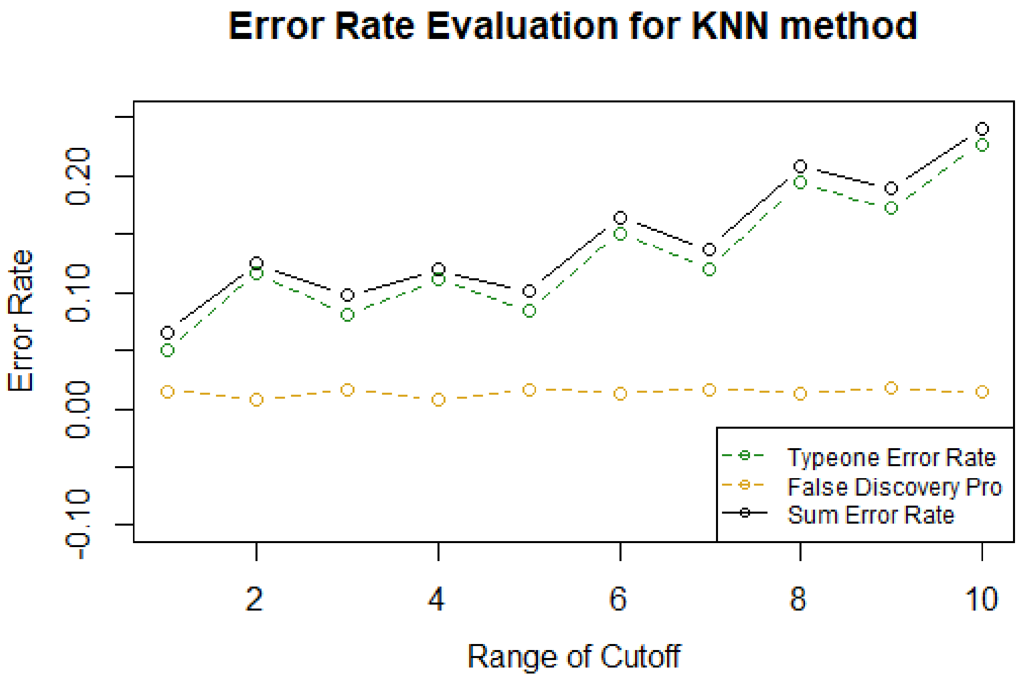
- We decreased the cutoff from 0.5 to 0.05
- Probably the very same story with logistic model
- However, the overall error rate increased, look at the maximum value of y-axis(1.0 to 1.2)
- Still, it is useful to decrease the cutoff point

Step 5: QDA Model



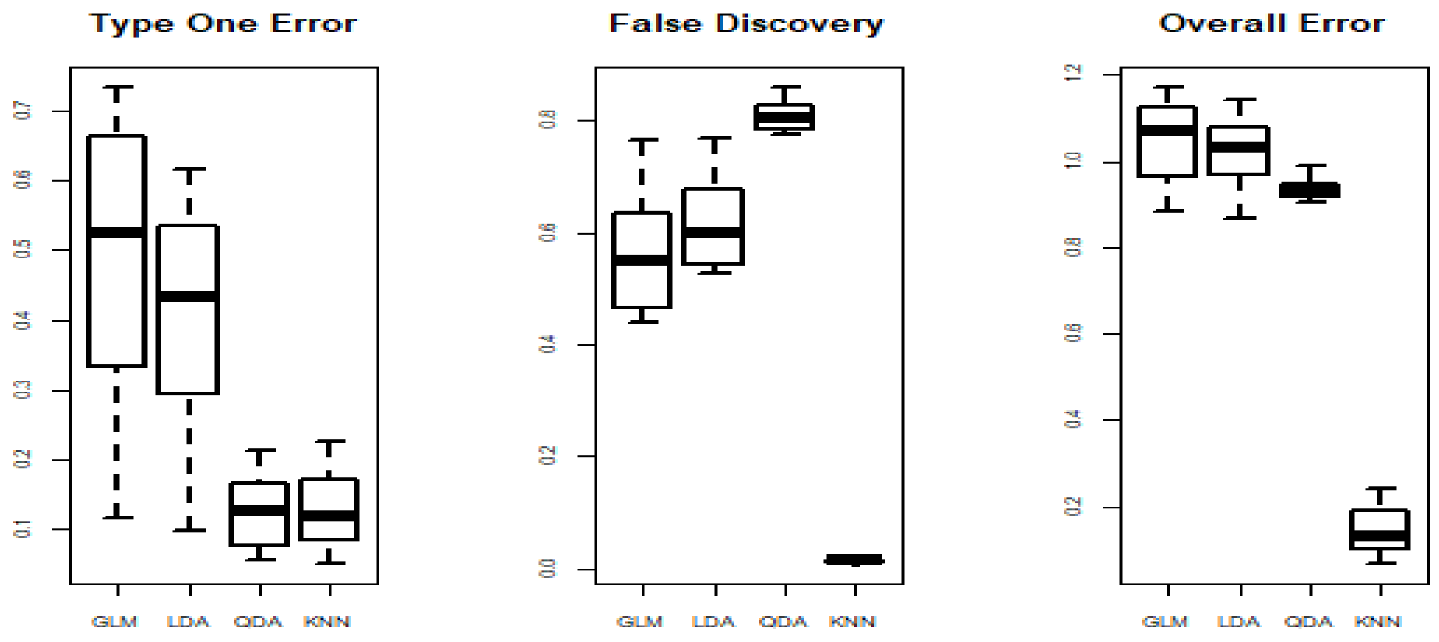
- We decreased the cutoff from 0.5 to 0.05
- However, QDA model seems have no interest in our cutoff adjustment, like this:

Step 6: KNN Method



- We increased the k value from 2 to 11
- Type one error gradually increased following the increasing k value, which is not surprising
- However, false discovery proportion is stable at almost the same level
- We still get a slightly lower sum error rate by decreasing k value

Step 7: Model Comparison

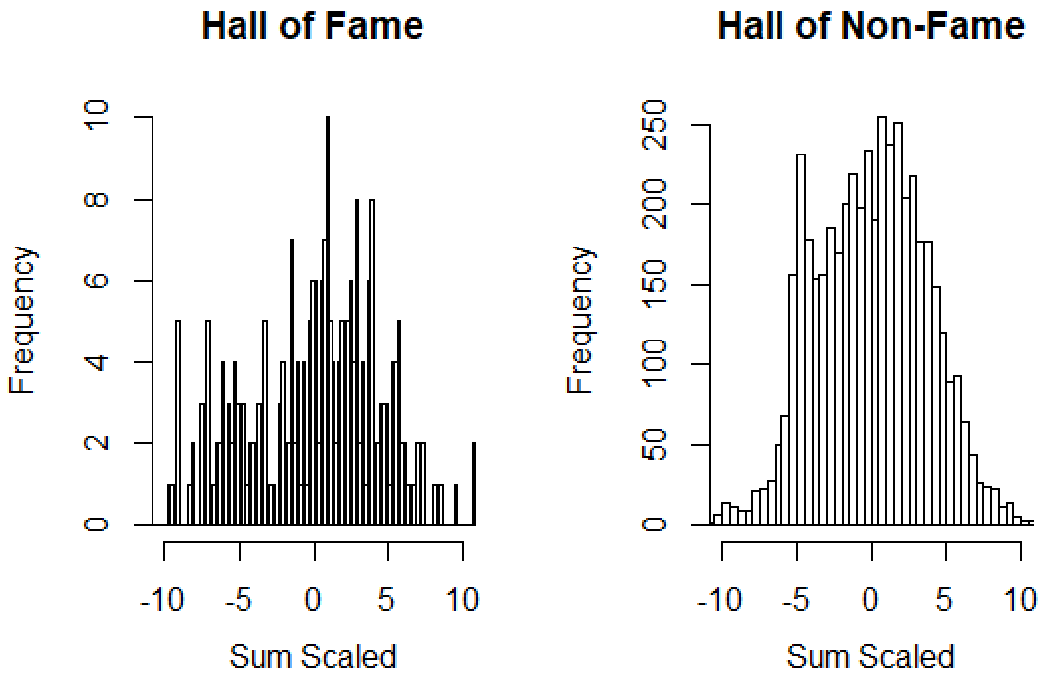


Interpretation and Results

Raymond A. Mason School of Business
WILLIAM & MARY

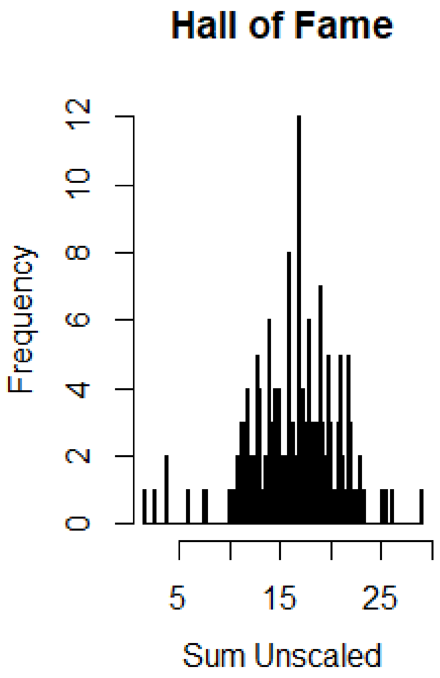
- **Why KNN wins? (Chapter 4.5 Scenario 5 & 6)**
    - Recall the assumption we made in LDA and QDA model: a normal distribution and related parameters. Are they truly exist?
    - This is why the seemingly imprecise model, KNN method, performs better than others
- **The value of non-parametric models: model comparison**
    - In the case of player performance, the decision boundary is highly non-linear, which makes logistic regression and LDA. QDA makes quadratic decision boundaries instead of linear boundaries, so it performs better
    - However, QDA is still not enough in a non-parametric environment or a boundary with more complicated non-linear function, much more flexible KNN method can be superior

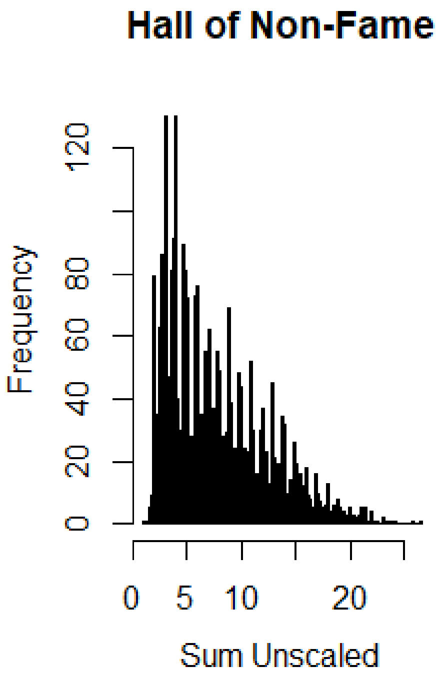
Interpretation and Results



Complicated Boundaries?

Interpretation and Results





Complicated Boundaries!

Conclusions and Takeaways

Raymond A. Mason School of Business
WILLIAM & MARY

- **Implications on Model Selection**
  - The KNN model proved the best in predicting whether a player would make it to the Hall of Fame. Therefore, think a non-parametric model may be superior in analyzing athletic performance
- **In the future...**
  - We would like to expand the range of the KNN method. For example, we would like to use all star data (to predict whether a player is expected to be selected in all star team)
  - We would also like to look into predicting salary data using regression of this data

Merry Christmas!

Team Upsilon

Claire Duchene, Jessy Sun, George Wei, and Michael Uhrig