Predictive Analysis of NFL Hall of Fame Chances

Michael Blankenship and Swapnav Deka

Rice University

INTRODUCTION

What is the goal of this project?

-To predict which NFL players will be inducted into the Hall of Fame in the near future.

How do we achieve our goal?

Acquire data from Kaggle.com and www.pro-football-reference.com

Explore the data to find out what qualifies players to be in the HOF

Create a model that can predict the likelihood that a player will be in the HOF

Run our model on current NFL Players, and players that retired after 2014

OBJECTIVES

- Select an extensive source of NFL Player data and retrieve it.
- Perform extensive Exploratory Data Analysis and data cleaning.
- Use sklearn's test-train-split utility to randomly split up the data and perform cross validation with different combinations of testing and training data.
- Employ Lasso because it can automatically select features based on whether or not they improve the fit of the model.
- Gradient boosting can fit a model to the data and the residuals and create a new model by continually inserting more models that correct the errors of the previous model.
- Decision trees also allow a nice way to visualize splits based on features.

DATA ACQUISITION

The initial dataset was scraped from Pro Football Reference and uploaded to Kaggle.

– We used a number of statistics regarding all three phases of the game: offense, defense, special teams

We consulted other sources: www.pro-football-reference.com

– We obtained Hall of Fame status and merged it with the initial dataset

DATA CLEANING

Only keep players from categories if

– QB career pass yds > 2000 (one season of avg QB passes)

– RB career rush yds > 500 (one season of avg RB runs)

– WR career receive yds > 500 (one season of avg WR catches)

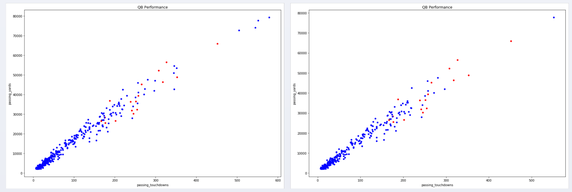
Cut players from training data who are not yet eligible (are still playing 2015 and on)

Cut players less than 5 years in league since no hall of famer has had shorter career

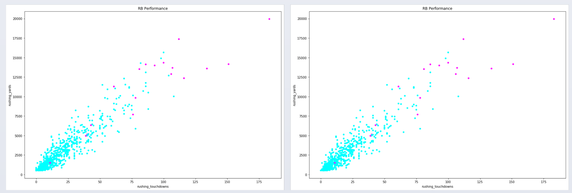
DATA VISUALIZATION

HOF Status by Position Performance: With Current Players v. Without

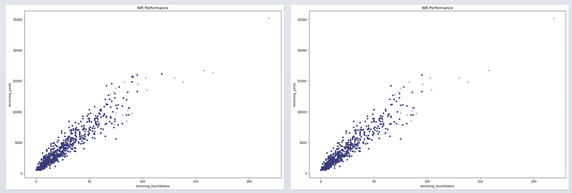
Quarterbacks



Running Backs



Wide Receivers



MODELING PREDICTORS

Response Variable: HOF Status

Predictor Variables:

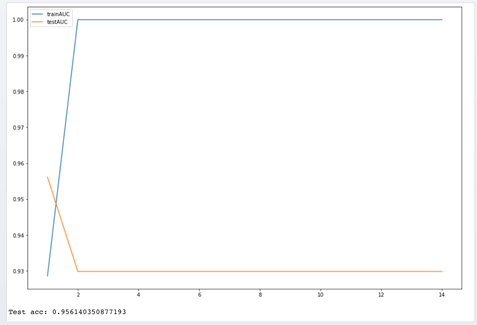
passing_attempts
passing_completions
passing_interceptions
passing_rating
passing_sacks
passing_sacks_yards_lost
passing_touchdowns
passing_yards
receiving_receptions
receiving_targets
receiving_touchdowns
receiving_yards
rushing_attempts
rushing_touchdowns
rushing_yards

MODELING SELECTION

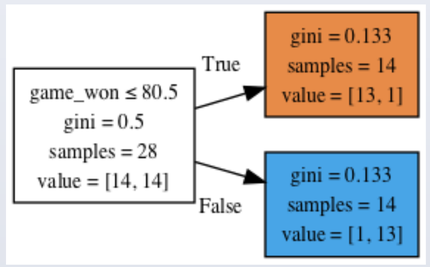
- After evaluating the tools we have available and determining which tools match well with our data and question, we decided to use logistic regression for classifying whether a player will be a Hall of Famer.
- Specifically, we were particularly interested in the L1 (LASSO) regression because of its sparse predictor results.

DECISION TREE ANALYSIS

Accuracy over Epoch Training: Quarterbacks



Decision Tree Diagram: Quarterbacks



RESULTS

LASSO Linear Regression

LASSO Logistic Regression	QB	RB	WR
Specficity:	0.9122807018	0.921875	0.84375
Sensitivity:	1	1	1
Test set acc:	0.9166666667	0.9242424242	0.8484848485
Test AUC:	1	1	0.9296875

Feed Forward Neural Networks

Feed Forward NN	QB	RB	WR
Test loss:	0.8059048057	0.8059048176	0.8059048057
Test accuracy:	0.9500000079	0.9499999881	0.9500000079

Decision Tree

Decision Tree	QB
Test accuracy:	0.9561403509

PREDICTION FOR CLASS OF 2019

In the end, we used the WR Lasso model to predict whether the 2019 WR nominees (there are only three) will be inducted into the Hall of Fame. Our predictions are:

- Isaac Bruce is predicted to be inducted in the Hall of Fame
- Torry Holt is predicted to be inducted in the Hall of Fame
- Hines Ward is predicted to NOT be inducted in the Hall of Fame

ANALYSIS

We tried various methodologies to build a predictive model for the NFL Hall of Fame Class of 2019.

Lasso Logistic Regression

– With the many parameters above, we knew a great place to start with analysis would be a Lasso logistic regression model. Our goal is to predict Hall of Famers from their career stats, and because of this binary decision, logistic regression is a great place to start as it also would highlight which parameters are actually important.

Feed Forward Neural Nets (FFNN)

– We found the FFNNs incredibly quirky. We found that as long as the networks were of some significant width (256 or 512 or more), variations in architecture (think telescoping etc) and depth didn't seem to make a significant difference. We decided to stick with the 512 to 256 telescoping architecture as this tended to be the most consistent performing architecture.

Decision Tree (Quarterback only)

– Our decision tree models were also incredibly inconsistent. Sometimes, the most accurate max depth would be 10 and sometimes 2. However, we did find that the best models had a very low decision tree depth and the one we liked the most was very simple. Did the Quarterback win more 80 games in their career. As simple as it sounds, we got a test accuracy of 96%.

REFERENCES

- https://www.kaggle.com/zynicide/nfl-football-player-stats
- Pro Football Reference

ACKNOWLEDGEMENTS

- Devika for the opportunity to take on this project
- COMP 340 classmates for support throughout the semester