Swapnav Deka and Michael Blankenship

Dr. Subramanian

COMP 340

02 December 2018

Predictive Analysis of NFL Hall of Fame Chances

Introduction

The NFL Hall of Fame is the most prestigious distinction that a player may receive after an illustrious career. While Super Bowls and MVPs are valuable awards, nothing can compare to being enshrined in Canton. Some players come into the league with sky-high expectations, while others come out of nowhere to claim their spot in history. The endless debates of who is deserving are timeless, but they raise the interesting question of whether or not we can accurately predict if an NFL player will become a Hall of Famer? We decided to investigate this question.

The Data

By both the casual fan and the NFL community as a whole, Pro Football Reference is considered the most complete public source of NFL player stats available online. This dataset contains every NFL player in their database going back to the 1940s until December 2017. This includes over 25,000 players, a total of over 1,000,000 football games, and 273 Hall of Famers (there are now 318, but this dataset ended in Dec 2017). The data we used was separated into three different JSON files. The first file contains player profile data (position played, height, weight, etc), the second file contains individual game data for players (passing yards for a game, tackles in a game, etc), and a third JSON which indicates which players have been inducted into the Hall of Fame. Below we have listed all of the fields from the profile and game datasets.

Player Profile Fields:

- *Player ID*: The assigned ID for the player.
- *Name*: The player's full name.

- *Position*: The position the player played abbreviated to two characters. If the player played more than one position, the position field will be a comma-separated list of positions (i.e. "hb,qb").
- *Height*: The height of the player in feet and inches. The data format is -. So 6-5 would be six feet and five inches tall.
- *Weight*: The weight of the player in pounds.
- *Current Team*: The three-letter code of the team the player plays for. This is null if they are not currently active.
- *Birth Date*: The day, month, and year the player was born. This is null if unknown.
- *Birth Place*: The city, state or city, country the player was born in. This is null if unknown.
- *Death Date*: The day, month, and year the player died. This is null if they are still alive.
- *College*: The name of the college they played football at. This is null if they did not play football in college.
- *High School*: the city, state or city, country the player went to high school. This is null if the player didn't go to high school or if the school is unknown.
- *Draft Team*: The three letter code of the team that drafted the player. This is null if the player was not drafted.
- *Draft Position*: The draft position number the player was taken. Again, null if the player was not drafted.
- *Draft Round*: The round of the draft the player was drafted in. Null if the player was not drafted.
- *Draft Position*: The position the player was drafted at as a two-letter code. Null if the player was not drafted.
- *Draft Year*: The year the player was drafted. Null if the player was not drafted.
- *Current Salary Cap Hit*: The player's current salary hit for their current team. Null if the player is not currently active on a team.
- *Hall of Fame Induction Year*: The year the player was inducted into the NFL Hall of Fame. Null if the player has not been inducted into the HOF yet.

Game Stats Fields:

Game Info:

- *Player ID*: The assigned ID for the player.

- *Year*: The year the game took place.
- *Date*: The date the game took place.
- *Game Number*: The number of the game when all games in a season are numbered sequentially.
- *Age*: The age of the player when the game was played. This is in the format -. So 22-344 would be 22 years and 344 days old.
- *Team*: The three-letter code of the team the player played for.
- *Game Location*: One of H, A, or N. H=Home, A=Away, and N=Neutral.
- *Opponent*: The three-letter code of the team the game was played against.
- *Player Team Score*: The score of the team the player played for.
- *Opponent Score*: The score of the team the player played against. You can use this field and the last field to determine if the player's team won.

Passing Stats:

- *Passing Attempts*: The number of passes thrown by the player.
- *Passing Completions*: The number of completions thrown by the player.
- *Passing Yards*: The number of passing yards thrown by the player.
- *Passing Rating*: The NFL passer rating for the player in that game.
- *Passing Touchdowns*: The number of passing touchdowns the player threw.
- *Passing Interceptions*: The number of interceptions the player threw.
- *Passing Sacks*: The number of times the player was sacked.
- *Passing Sacks Yards Lost*: The cumulative yards lost from the player being sacked.

Rushing Stats:

- *Rushing Attempts*: The number of times the the player attempted a rush.
- *Rushing Yards*: The number of yards the player rushed for.
- *Rushing Touchdowns*: The number of touchdowns the player rushed for.

Receiving Stats:

- *Receiving Targets*: The number of times the player was thrown to.
- *Receiving Receptions*: The number of times the player caught a pass thrown to them.
- *Receiving Yards*: The number of yards the player gained through receiving.

- *Receiving Touchdowns*: The number of touchdowns scored through receiving.

Kick/Punt Return Stats:

- *Kick Return Attempts*: The number of times the player attempted to return a kick.
- *Kick Return Yards*: The cumulative number of yards the player returned kicks for.
- *Kick Return Touchdowns*: The number of touchdowns the player scored through kick returns.
- *Punt Return Attempts*: The number of times the player attempted to return a punt.
- *Punt Return Yards*: The cumulative number of yards the player returned punts for.
- *Punt Return Touchdowns*: The number of touchdowns the player scored through punt returns.

Kick/Punt Stats:

- *Point After Attempts*: The number of PAs the player attempted kicking.
- *Point After Makes*: The number of PAs the player made.
- *Field Goal Attempts*: The number of field goals the player attempted.
- *Field Goal Makes*: The number of field goals the player made.

Defense Stats:

- *Sacks*: The number of sacks the player got.
- *Tackles*: The number of tackles the player got.
- *Tackle Assists*: The number of tackles the player assisted on.
- *Interceptions*: The number of times the player intercepted the ball.
- *Interception Yards*: The number of yards the player gained after interceptions.
- *Interception Touchdowns*: The number of touchdowns the player scored after interceptions.
- *Safeties*: The number of safeties the player caused.

Note that if there are games missing in the season for a player (i.e. the player has logs for games 1, 2, 3, 5, 6,...), then they didn't play in game 4 because of injury, suspension, etc.

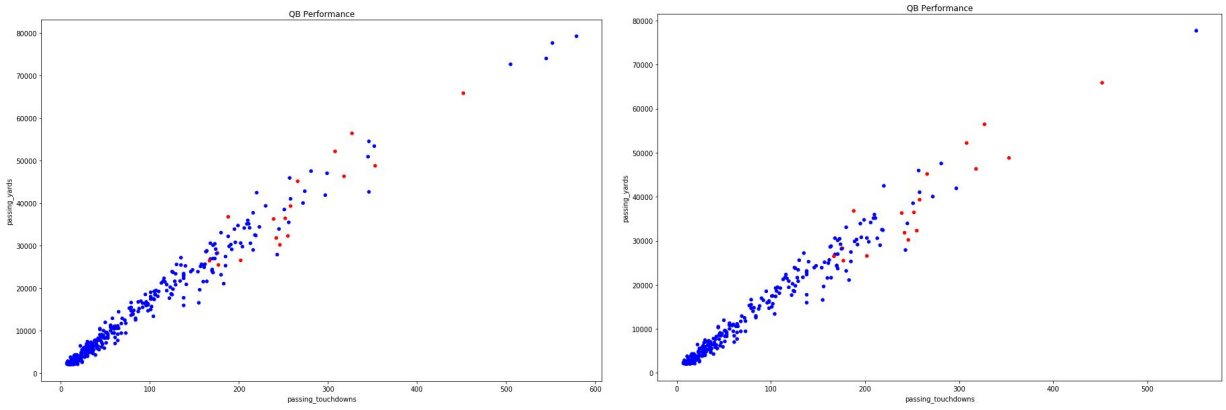
-

Cleaning the Data

Data Cleaning was essential to produce an accurate analysis. For example, we would not want to include players who only participated on practice squads or didn't play a full season. Thus, we filtered out players who had spent less than 5 years playing in the league.

We would also need to account for players who played special roles or did not have traditional statistics. In order to combat this, we decided to only build prediction models for three popular and easy to measure positions (Quarterback, Running Back, and Wide Receiver).

One step in our data exploration, which proved to be crucial, was to plot key player stats against each other. Doing so, reinforced a rather obvious insight that Hall of Famers tend to be outliers in terms of raw performance output. However, this step did reveal something unexpected. We realized there were outlier players who appeared they should be in the Hall of Fame according to their high performance but were not yet inducted. After further investigation, we figured out that these players were indeed ineligible for the Hall of Fame because they had played within the past 5 years. One is only eligible for the Hall of Fame after being retired for 5 years. Thus, we removed all players who had played within 5 years of the last game played in this dataset. As you can see below, the left chart includes the mentioned players (peek the top right corner) and the cleaned data excludes them (and thus appears to be make more sense).



(Left) QB performance measurements plotted against each other; Note: Hall of Famers highlighted in red.

(Right) Same as above, with players that have played within the past 5 years removed.

A Biased Dataset

After creating prediction models with the traditional training and test data split, we realized our models were great at predicting who would not be a Hall of Famer, but only as good

as random at predicting who would be a Hall of Famer (HoFer). This becomes obvious after some thought. There thousands and thousands of NFL players, yet only a couple hundred have been inducted into the Hall of Fame. This means any model which predicts a “Not Hall of Famer” outcome most of the time will be an decent model at predicting non-HoFers, but not good at predicting actual HoFers. After discovering this downfall, we implemented correctly weighted training and test sets for each of the models. We used 90% of the Hall of Famers in the training set and a corresponding absolute number of nonHoFers in the training set. For example, if there are 10,000 total players and 100 HoFers, then we would use 90 HoFers and 90 nonHoFers in the training set. Then, in the test set, we tested on the same proportion of HoFers to nonHoFers. The test set in this example would consist of the 10 HoFers not used in the training set and then 1,000 randomly sampled nonHoFers.

Our results saw a huge immediate improvement. We saw our false negative rate plummet (this is good), and saw our false positive increase slightly (this is bad). Overall, the tradeoff increased our models predictions accuracies and AUCs significantly.

Model 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.

In regards to the risk of overfitting our data, we started with a linear logistic regression model. We used the aggregate data to create a model that could predict a player’s Hall of Fame status. To visualize our results, we generated multiple visualizations including scatter plots of position-by-position statistics and colored indicators of HoF status.

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%.

Results

LASSO Logistic Regression	**QB**	**RB**	**WR**
Specificity:	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 NN:	**QB**	**RB**	**WR**
Test loss:	0.8059048057	0.8059048176	0.8059048057
Test accuracy:	0.9500000079	0.9499999881	0.9500000079

Of the three positions we chose to look at (QB, WR, RB), the WR position had the most 2019 HoF nominees at a whopping total of 3 candidates. Thus, we decided to make predictions on whether

these three candidates will be inducted into the Hall of Fame. In the end, we used this WR Lasso model to predict whether the 2019 WR nominees 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

We won't know if these predictions are accurate until the 2019 class of HoF'ers has been announced.