Swapnav Deka and Michael Schach

Dr. Cox

STAT 413

24 April 2019

Predictive Analysis of NFL Hall of Fame Players

Section 1: Introduction

Every year, a small number of history's greatest football players are inducted into the hall of fame. Journalists reflect on great moments, nostalgia and of course, statistics, from a player's career in order to argue for who has the right to be called a hall-of-famer. Currently, there are 273 players in the hall of fame and a few more will be added this year. We hope to answer the question of who deserves to be in the hall of fame through a statistical analysis. Specifically, we want to find the most accurate algorithm which can predict if a player, based on their career performance, is in the hall of fame.

The hall of fame decisions are an event with national exposure, however, we found only two relevant pieces of work on predicting the problem using machine learning methods. In 2017, ESPN Analytics released an article titled *Artificial Intelligence Predicts Hall Chances for Warner*, *Tomlinson and Davis* (three potential hall of famers). The researchers used a neural network for their model, citing "This kind of analytic model is especially good at handling complex relationships and logic among multiple factors". The author continues to say "Suppose things worked like this: A running back wouldn't make the Hall without at least two All-Pro selections no matter

how many total yards or Pro Bowl selections he has, except if his TDs also exceed some value. Neural networks can recognize those types of patterns in the data. Other kinds of prediction models are typically additive, and would errantly predict a Pro Bowl regular to be inducted without the other requisite qualifications". The author also combines multiple factors into the data they analyze by examining career statistics, length of career, career awards and postseason performance. Furthermore, they clean the data by adjusting yards and touchdowns by the league averages and season length. Although we do not have extra feature variables such as league averages and season lengths, we have access to variables such as length of career and career awards in our own data set.

The other relevant piece of work, by a data scientist named Samuel Binenfeld, examined the effectiveness of various models in tackling this problem. He found that random forests and a gradient boosting classifier performed well for this problem. Additionally, he added several interesting feature variables, specifically, game winning drives and playoff game winning drives which were important in prediction. He was able to find these important feature variables by using lasso regression, a technique previously discussed in class. Unfortunately, our data set did not have these key feature variables that Binenfeld found, but it could be useful in other attempts at this problem.

Overall, the insights from the previous works helped influence our own methodologies and how we approached the problem. In the following section, we will explore the data used and our methods.

Section 2: Materials and methods

The Dataset

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 (https://www.kaggle.com/zynicide/nfl-football-player-stats), scraped from Pro Football Reference, 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. 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. After examining previous research on the problem, we determined it may be useful to add other metrics to the data set, such as career awards, which were missing from the data set. We scraped web pages on Pro Football Reference which contained the number of pro bowl appearances for every NFL player and added that metric to our data set.

The following bullet points are the variables within the data sets that we had access to.

Player Profile Variables

- *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 Info Variables

- *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 Variables

- *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 Variables

- *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 Variables

- *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 Variables

- *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 Variables

- *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.

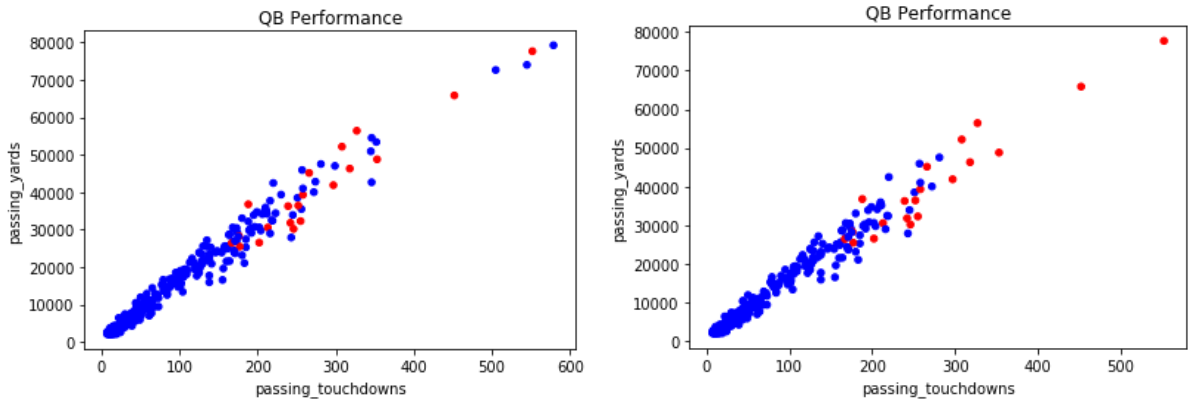
Defense Stats Variables

- *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.

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 the three most popular offensive 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. In the graphs below, Hall of Famers are represented by red dots and non HoFers are represented by blue dots. You can see that by removing ineligible players, the three quarterbacks with incredible statistics were not snubbed from the Hall of Fame; they simply have not been voted on yet.



Many of our feature variables had different scales with wildly different magnitudes. Thus, it was important to normalize the variables to values between 0 and 1, because

the difference in magnitudes would cause some features to be weighted more than others for certain algorithms. After normalizing the variables, not surprisingly, it led to a boost in performance on some algorithms.

Methodology and Algorithms

- Logistic Regression
    - Because of the large number of parameters, we decided a good place to start would be a logistic regression with LASSO regularization. This regression method performs both variable selection and regularization. Our goal was to predict a binary Hall of Fame decision and highlight the importance of specific parameters.
- KNN
    - Like logistic regression, KNN was a relatively straightforward algorithm and a good place to start our investigation. We thought it may also help us determine how clustered or similar hall of fame players were to one another.
- Feed-Forward Neural Networks
    - Feed-forward networks are arranged in layers, with the first layer taking in inputs and the last layer producing outputs. The middle layers have no connection to anything outside and are referred to as hidden layers. Information is constantly “fed-forward” from one layer to the next.
- Decision Trees

    - Decision trees use a tree-like model to portray decisions and their possible consequences. In this case, we identify various splitting conditions to identify the most accurate way to predict a player's Hall of Fame status.
- Random Forest
    - Previous research indicated that random forests might be an algorithm that performs well. It is a collection of decision trees each of which only use a subset of the feature variables.
- SVM
    - SVMs were another algorithm that we discussed in class. They showed promise in this research problem due to previous data visualization which showed the hall of fame players and non hall of fame players were somewhat separable.
- SMOTE (synthetic minority over-sampling)
    - One issue that we noted earlier was the large discrepancy between the number of players in the data set who are in the hall of fame and those who are not. The vast majority of players were not in the hall of fame which caused many models to perform poorly, despite having a high accuracy. In order to alleviate this problem, we looked to a technique called SMOTE. SMOTE generates synthetic samples of the minority group (hall of fame players) and creates a new data set using those samples. In theory, this creates a more balanced data and could improve results.

Section 3: Results

- Logistic Regression

LASSO Logistic Regression	QB	RB	WR
Specificity:	0.947368	0.9375	0.828125
Sensitivity:	1.0	1.0	1.0
Test accuracy:	0.95	0.939394	0.833333
Test AUC:	0.982456	0.992188	0.976563

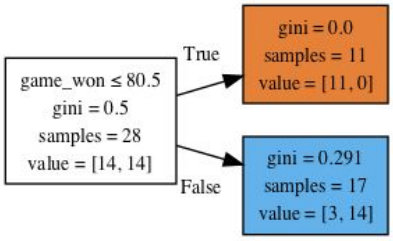
- Feed-Forward Neural Networks

FFNN	QB	RB	WR
Test loss:	0.805905	0.488427	0.488427
Test accuracy:	0.950000	0.969697	0.969697

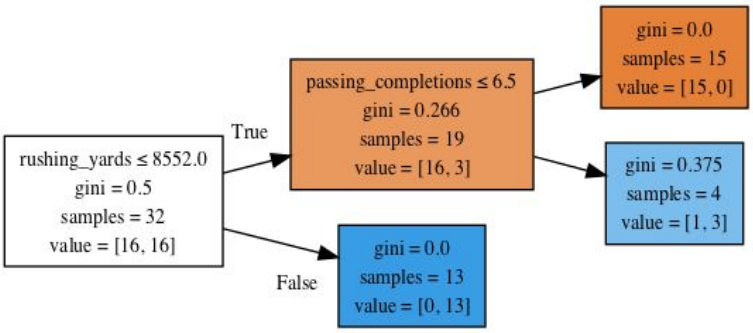
- Decision Trees

Decision Trees	QB	RB	WR
Depth:	1	2	1
Important elements:	Games won	Rushing yards, passing comp	Receiving touchdowns
Test accuracy:	0.745614	0.953125	0.984375

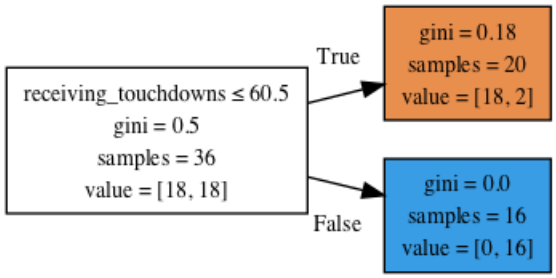
QB Tree



RB Tree



WR Tree



- KNN

KNN	QB	RB	WR
Accuracy:	0.940	0.975	0.971
F1:	0.314	0.499	0.489

- Random Forest

Random Forest	QB	RB	WR
Accuracy:	0.944	0.976	0.969
F1:	0.493	0.592	0.328

- SVM

SVM	QB	RB	WR
Accuracy:	0.937	0.978	0.971
F1:	0.404	0.53	0.374

- SVM using SMOTE

SVM	**QB**	**RB**	**WR**
Accuracy:	0.931	0.964	0.967
F1:	0.714	0.714	0.6

Section 4: Conclusions

The development of our various models to predict hall of famers quickly led to a realization: our prediction models with a traditional random training and test data split 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 a 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.

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. When using the FFNN model, the model predicts that none of the three receivers will be inducted into the Hall of Fame.

The decision trees were surprisingly accurate while incredibly simple. All three decision trees were of no greater than 2 depth. The wide receiver prediction was as simply as relying on the number of receiving touchdowns scored by a player over the course of his career: all three nominees have over 60.5 touchdowns, so the decision tree would predict all three of the nominees to be inducted into the Hall of Fame.

Although previous research had pointed to random forests as being an effective approach to this problem, we found that random forests did not perform well. The accuracy for this method was quite high, with QBs having the lowest test accuracy at 0.944. However, due to the imbalanced data set, the accuracy does not tell the whole story. The F1 score, which uses both precision and recall, was only 0.493. This means that although the forests were classifying the majority class well, it was actually quite

poor at classifying the minority class. This is seen as well with the WR class, which only had a F1 score of 0.328.

KNN was also a poor algorithm for this problem. Once again, although the accuracies were high, the F1 score showed that KNN was poor at classifying the players. The QB category had 0.940 accuracy, but only 0.314 for an F1 score. This could have been predicted from the data visualization. Reflecting on the graphs of QBs, one could see that based on yards thrown and touchdowns, the non hall of famers and hall of famers were mixed together. Only a few outliers in the graph would result in KNN predicting that a certain player would be in the hall of fame.

Regular SVMs were also a poor choice for this problem. They did not perform any better than the KNN or the random forest. Interestingly, the linear kernel performed the best out of all the different kernel types. However, when SMOTE was applied to the data set, and SVM was applied to this resampled training set, it produced far better results on the test set. There was an improvement of 0.31 for the f1 score of the QB classification. While this did produce good results on this given test set, it is questionable whether there would be an improvement on other sets. The model may now be over fit to the synthetic samples and not representative of the actual population. Certainly, if this problem is looked at again by other groups, it may be beneficial to investigate SMOTE at a greater depth for other algorithms.

There are many factors that were not considered in our modeling that could be investigated for future studies. One such instance is that the standards of Hall of Fame status might change over time, but we did not weigh any era differently than any other.

For example, it is often said that the league is more “passer friendly” than ever now, which means that statistics that may have been impressive and Hall of Fame worthy for a quarterback in the past may be considered merely average in the current age. We also considered Hall of Fame induction to be purely binary, i.e. we did not differentiate between Hall of Famers via different ballots. There are individuals in the training set who may not currently be Hall of Famers, but might be inducted into Canton later. Because we do not know if this will be the case, it is possible that we are inaccurately training our models to view certain players as “not worthy” when they may simply be “not worthy yet.”

Although not all of these problems are easily managed, some things would be interesting to consider in the future. We could apply an inflation factor to statistics from older eras to adjust for the difference in play style among different time periods. We could also adjust our dataset to differentiate among players who are inducted into Canton as 1st, 2nd, nth ballot Hall of Famers. This would allow us to predict more specifically on a player based on if they have been on the ballot previously.