NFL Quarterback Hall of Fame Predictions

Adam Yamarik

Abstract

The Hall of Fame is reserved for only the greatest players in the National Football League, and requirements to be inducted are intense. A committee of people use a lengthy process to vote on which players will make the hall of fame based of certain factors. Using logistic regression and a kernelized perceptron, I try to create models which will accurately predict if a player will make the Hall of Fame, or not. Logistic regression will also help find the usefulness of the features, to give better information on which stats appear to be the most important for Hall of Fame voting.

1. Introduction

The Hall of Fame is one of the greatest honors for a player in the National Football league. In addition to this, the Quarterback position is the most important in the entire sport. They not only need the most versatile skillset of any player, but they act as the leaders of the team, as well as the game manager by a play-by-play basis. Needless to say, the company of quarterbacks in the Hall of Fame are well known and have the necessary resume to be inducted. Only 21 quarterbacks since 1955 have been inducted, with only 27 quarterbacks being inducted in modern NFL history. However, the game is consistently changing, and quarterback play has changed significantly since then, and with that, the goalposts for making the hall of fame change as well.

A player may only be inducted into the hall of fame after they have been retired for 5 years or more, and may take multiple ballots before they get inducted.

2. Problem Definition

The inductees to the Hall of Fame are chosen by a committee, who begin by choosing the "senior members", who are inducted with an 80 percent vote. Then, the remaining group in reduced to 15 by the committee, then to 10, and the finally to 5. From there, at least two members must be inducted, with a vote of at least 80 percent, but if deigned by the committee, all 5 members can be inducted. Our goal is to create a model which can be used to accurately predict whether a player, specifically a quarterback, will make the hall of fame. This is regardless of their time in the league, or their ability to be inducted into the hall of fame. The current Hall of Famers are decided on by a committee as explained before, but the guidelines for which they use are not random. The model created should learn those qualities and make a prediction for any player.

This model will mainly be used to get more insight on players who are uncertain. Players like Tom Brady or Aaron Rodgers is almost certainly going to be inducted into the Hall of Fame when they eventually retire. There are also many players who will people will never entertain the thought of them making the Hall. The main focus of this problem will focus on players like Eli Manning, Matthew Stafford, Joe Flacco, etc. These are players with very respectable careers, but there are many arguments among analysts and pundits as to if these players will eventually make the Hall of Fame. This model will attempt to compare these quarterbacks to those already in the Hall of Fame, and predict if those stats are similar enough to those in the Hall of Fame, and make a prediction accordingly.

3. Proposed Method

3.1. Classification method

The classification method that eventually would be used needed to handle the different features used, which may or may not have direct correlation to one another. Since this is the case, I did not believe that a linear separator would be able to accurately separate the data., so the preposed method would be to use a kernelized perceptron to allow for non-linearity. This method, however, will need good selection of the data, as well as cleaning the data, but initially, there is uncertainty as to what features will be useful. Therefore, the plan is to use logistic regression to find a weight vector that will then be used to help the classifier. Since logistic regression is being used, I also plan of using this classifier to compare to the kernelized perceptron.

3.2 Data used and data cleaning

The data chosen is inspired by common sense of which stats are potentially useful for a player to make the Hall of Fame. These stats include yards, completions attempts, touchdowns, interceptions, and championship wins. However, we cannot take the total values for training. The reason why is that this model is going to be used to predict if any player will eventually make the Hall of Fame. So taking a player like Patrick Mahomes, who has only been in the league for 5 years, would never be predicted to make the Hall if we compared it to players who have had fifteen year careers.

Therefore, the data will be modified so that we use stats on a per game basis, with the exception of championships. So the final features that will be used are completions per game, attempts per game, yards per game, touchdowns per game, interception per game, and championships. There is uncertainty as to how useful these stats will specifically be for the model, but a justifiable case could be made that each of these stats are important for consideration for the Hall of Fame.

3.3 Intuition

As mentioned above, I assumed the data would not be linearly separable, so a kernelized perceptron would make sense to use. However, I do not know how important each feature is exactly, and I do not wish for less useful features to hold a large amount of influence when using the kernelized perceptron. The intuition of using logistic regression is that a weight vector is learned, which corresponds to how important each feature is to the model. So if I use logistic regression, I can get a weight vector that will tailor the features so that the most important features take precedence over the others.

Once the model has been trained using the per game data of quarterbacks that are eligible to be inducted into the Hall of Fame, the model should be able to make a decent prediction on any quarterback, with their stats also being scaled down to a per game basis.

4. Experiments

4.1 Testbed

The testbed being used has been introduced previously, so this section will focus of the specifics of the testbed. First, I made the data in a readable form, a couple of txt files for the x and y data. The x data is in the form of, ‘total games played, total completions, total pass attempts, total yards, total touchdowns, total interceptions, total championships, player name’. The name has no use in the model, but it is included to allow for validation of the stats, if desired. The data is read in and placed in numpy arrays to be used. The data is then scaled to the level described previously, where total games no longer have any impact on the model. Then the data is used in logistic regression, with a test size of 25%. The score for the model is then predicted, and a weight vector is gotten.

The kernelized perceptron is then initialized using the same test-train split of the logistic regression model did. The data is then modified by multiplying each feature by its corresponding weight. The kernelized perceptron is using the polynomial kernel provided by Sklearn. Its score is then also found, and the model is ready to be used for predictions.

There are some other specifics to the testbed that will be revealed when discussing the experiments that the testbed was designed for.

4.2 Initial experiments

First, I will discuss what I initially planned for the models to do. The first is the whole point of the project. Given enough data, can the model accurately predict if a player will make the Hall of Fame, or not?

The next experiment is the comparison of the two models. Will the kernelized perceptron be better than the logistic regression model? Or Visa versa.

The last experiment I initially planned for is if the models could predict if any quarterback will make the Hall of Fame, despite their tenure in the league. This part would need to be compared to theoretical discussions, since there is no definitive way of knowing whether a player will make the Hall of Fame or not until they are eligible.

4.3 Other Experiments

There were some questions that I had not initially planned for, that I eventually decided to add in the testbed. This section is a list of these questions.

The first of these experiments was to test to see whether the weights found in the logistic regression model would help the perceptron. Is the score of the kernelized perceptron with the weights better or worse than if the weights were not used?

The other big experiment I wanted to do was to see the cut-off point, so to speak, between the two models. Can we find a quarterback that is predicted by one model to eventually be inducted, while the other does not?

These questions were not initially implemented and were added due to my own curiosity. These questions, along side the first group of questions are all solved using the testbed.

5. Results and Observations

Let’s start with the first group of questions, and dive into why the results are the way that they are. For the purposes of simplifications, we will use the domain of {Does not make the Hall of Fame , Does make the Hall of Fame} = {0, 1}.

5.1 Scores and comparisons

First are the resulting scores of the two models that were used. The logistic regression model had a score of .875. The test set had 48 instances in it, so this gives a split of 42 correct classifications and 6 incorrect classifications. Of the incorrect classifications, there were 5 instances where the model predicted 0 and the actual classifications was 1, and only one instance where the prediction was a 1, but the actual prediction was a 1.

This would lead us to believe that our model is a bit more restrictive than reality. The incorrect predictions are almost all players who in reality made the Hall of Fame, but our model failed to predict that they would make it. We will dive more into why this is in the discussion about the examples on some players.

The kernelized perceptron had a score of .833. This used the same data split as the logistic regression model, so it also used 48 instances. This means that 40 instances were properly classified, and 8 were incorrectly classified. Of these, 7 were predicted as 0 but were a 1, and 1 instance was classified as a 1 but was a 0. Interestingly, of these misclassifications, it includes all the misclassifications of the logistic regression.

Once again, this model seems to be more restrictive from what reality is, being more on the side of classifying a player as missing the Hall of Fame, where they made it in reality.

Comparing these two models, using this split, it seems that the logistic regression model is strictly better. Both models were more restrictive when compared to the test instances, but the kernelized perceptron was more restrictive. The interesting relationship between the two models make sense, since the weights used in the perceptron were found from the logistic regression model, so the similar properties are understandable.

5.2 Weights

The weights found in the logistic regression model obviously hold a lot of importance in both models, so it is important for us to discuss them, with specific values. As a reminder, the order of the weights will be (Completions per game, attempts per game, yards per game, touchdowns per game, interceptions per game, Championships). Lets being with some intuition on these weights, using common knowledge of football.

Completions and attempts have a relationship to one another, but each stat in particular is not all that important. Many attempts or many completions does not necessarily translate to success.

Yards per game would presumably be an important stat for this purpose. Higher yards per game should have a positive correlation to points and winning games, which are very important to making the Hall of Fame.

Touchdowns and interceptions should have opposite effects. Higher touchdown counts should be very important to a players chance to make the Hall of Fame, and higher interceptions should hurt that players chances.

Finally, are the championships, which I believe will be the most important feature of those included. Winning championships is widely believed to be the most important factor for players making the Hall of Fame, so I believe this will have the most weight given to it.

The actual weights of the given split are as follows: (-0.12354417, 0.05243707 , 0.03272731, 1.21452078, -0.03722771, 1.54188867)

The touchdowns and championships are the two weights which match the initial beliefs the most. Touchdowns are important, but championships are the most important feature for making the Hall of Fame.

Attempts and completions are somewhat expected, not having much impact on the prediction. A case could be made that a negative weight for competitions may not be expected, but the weight is close enough to zero that it is close to a non-factor.

The two weights which are a bit unexpected are the yards per game, and interceptions per game. I expected yards to have a fairly significant positive correlation for its weight, but it turned out to be insignificant. Interceptions were expected to have a significant negative correlation to, but it also ended up being a non-factor.

The discrepancies between the different classes does not value these two weights, which is interesting, as it does differ from what common sense would predict. Of course, the model does not have access to this information, so given the specific split that was used, these two features must not have enough differences to find them significant.

These weights can give us a better idea on how the models will make predictions, and can allow us to see why predictions will be made the way they are.

5.3 Prediction examples

Here, we use the two models to make predictions on players who are eligible to be inducted into the hall. We will compare this to the weights, as well as common sense to check whether the models are justified in their predictions.

For reference, we will use the Hall of Fame Monitor (HoF Monitor) score from Pro Football Reference. This score is a value attributed to each quarterback, where the average of quarterbacks who have made the Hall of Fame is

109. We can compare these examples to this value, and also compare them to the different Hall of Fame quarterbacks to get a good idea of if the predictions are valid.

We begin with a couple of players who almost certainly will make the Hall of Fame when they are eligible. The first is Tom Brady, who is considered to be the best quarterback in the history of the NFL and may even be the best single player in history. Presumably, the models will predict the Brady will make the Hall of Fame, and in fact, this is the case. Both models predicted that Brady will make the Hall of Fame. For reference, Brady has a HoF Monitor score on 259.32, comfortably above the average, so the prediction is very much justifiable.

Next is Aaron Rodgers, who while he may not have the same career as Brady, is still considered to be one of the best quarterbacks in the NFL. His HoF Monitor score is 197.26, which would imply he will make the Hall. The models agree with this, with both predicting that Rodgers will make the Hall of Fame.

The next experiment was to try a player who has little chance of making the Hall of Fame. The specific quarterback I chose did not really matter, so I decided on Case Keenum. Keenum has had a decent career in the NFL, even making an NFC Championship game with the Minnesota Vikings. However, his career has certainly not been Hall of Fame worthy, and with a HoF Monitor score of 10.30, it seems that Pro Football Reference agrees. Our models agree as well. Both models do not believe that Keenum will make the Hall, so it appears that they are handling the extremes correctly.

However, not every quarterback will be cut and dry. Some have been rigorously discussed and argued by fans and analysts alike. The first player I wanted to look at Eli Manning, the well tenured quarterback of the New York Giants, who, while being retired, is not eligible to be inducted into the Hall of Fame. Eli has a HoF Monitor score of 85.01, which is close to the average. Additionally, eight Hall of Fame quarterbacks have a score below Eli.

Both of our models predict that Manning will make the Hall of Fame. From the weights we have, we know that championships hold the most importance, and Eli Manning has two Superbowl wins, so it makes sense that the models would predict that Manning would make the Hall.

Next is Matthew Stafford, who was the winning quarterback on the Super Bowl 56, the championship game of the previous NFL season. Stafford was the topic of many talks this off season, arguing if this Super Bowl win is enough to induct him into the Hall of Fame.

Our models predicted that Stafford would make the Hall of Fame, and some would argue that the model will eventually be correct. Stafford's HoF Monitor score in 68.44, with only two quarterbacks below him. So while it is less likely that he will make the Hall of Fame when compared to Manning, it is still well within the realm of possibility. For curiosities sake, I modified Stafford's stats so he had zero Super Bowl wins, and the logistic regression model changed its prediction, which shows how important championships are to the models prediction.

5.4 Other tests

First, lets test to see if the weights found in the logistic regression model are a benefit to the kernelized perceptron. As a reminder, the perceptron used had a score of .833. I made a new kernelized perceptron using the flat values, instead of the weighted ones. The score of this new model was .791. This is an improvement of two instances, which indicates that the weights did improve the performance of the kernelized perceptron, which was excepted.

The next test was to try and find a discrepancy between the two models. During the test of Matthew Stafford, I mentioned that if the championships were changed to 0, the logistic regression model changed its prediction.

However, the perceptron did not change its prediction.

Another example is Lamar Jackson, who won league MVP in his second year in the league. The logistic regression model predicts that Lamar will not make the Hall, but the kernelized perceptron predicts the opposite. What these example show is that the lack of a championship win is very impactful to the logistic regression model, and significantly decreases that players chance of being predicted a 1. This correlation does not necessarily translate to the kernelized perceptron, however. A lack of a championship is significant, but not as much as logistic regression model.

6. Conclusions

Through the many tests done and investigations to the properties of the models, much was learned about the approached problem. However, there were also some issues that presented themselves, and some discrepancies from what would be expected, that we will dive into here.

6.1 Issues

The first issue that I believed caused the most problems was the general lack of data available. There are only 21 quarterbacks from the modern era inducted into the Hall of Fame, and only so many quarterbacks in general. One of the best ways to improve a models performance is to use more data, but with this data, it isn't possible to just make new instances out of thin air. This let the data limited, and some parts of the testbed may not as good as they can be.

The NFL is constantly changing year after year, and standards for players change as well. Offenses and quarterback play in particular are more important than they have been in the past. Even just twenty years ago, teams scoring 50 points in a game was very rare. Presently, it is still uncommon, but much more likely than in the past. This point throughs a wrench into the model a bit, and some people may argue that the standard for making the Hall of Fame changes with this standard.

The last issue is that the inductees to the Hall are limited by year, so a players chance to make it are directly correlated to how strong their class is. If a borderline quarterback is included in a ballot with eight players who are mostly guaranteed to make it, then that borderline will miss the induction on that ballot. This adds an extra level of complexity that was not included in the models, but would also be near impossible to do, since the ballots are newly made each year.

Overall, the actions of the models were all fairly expected. None of the classifications were too farfetched, and the weights found were all pretty understandable. At the end of the day, however, Hall of Fame voting is done by humans, which use certain properties to make their decision. It isn't easy to simulate 'the eye test' for example. Despite this, the models were still able to do a fine job with their predictions, and I will be very interested to see if its predictions on borderline players will end up coming true or not.

References

Bonilla, A. (2019, December 19). *Introducing the PFR Hof monitor*. Sports. Retrieved May 22, 2022, from https://www.sports-reference.com/blog/2019/12/introducing-the-pfr-hof-monitor/

Pro Football QB Hall of Fame Monitor. Pro. (n.d.). Retrieved May 22, 2022, from https://www.pro-football-reference.com/hof/hofm_QB.htm?__hstc=213859787.082fcf40b49be66334f7121acaced370.1653268522466.1653268522466.1653268522466.1&__hssc=213859787.3.1653268522467&__hsfp=2015750177

DiCresce, J. (2020, April 14). *Which QB stats are the most important?*

mfootballanalytics. Retrieved May 25, 2022, from https://mfootballanalytics.com/2020/04/06/which-qb-stats-are-the-most-important/