Table of Contents

1 Introduction

Is Terrell Owens going to be in the Pro Football Hall of Fame? Will Rob Gronkowski get in when his career is over? These are the types of questions that are frequently debated among football enthusiasts.

The Hall has a 48-person selection committee that decides who is worthy of getting in. It is this committee's job to induct the best players that have ever played the game. **The goal of this project is to predict who will be inducted into the Hall of Fame in the near future.**

To investigate this, we will analyze historical NFL data and find out what factors are influential to being voted into the Hall. We will then create a model that is able to predict the likelihood that a player will be inducted. Using this model, we will examine recent NFL players, and predict who is likely to be inducted in the near future.

2 Project & Model Restrictions

We restrict our project to only include players with the position of Quarterback, Running Back, or Wide Receiver. It was initially our goal to include all offensive positions other than linemen. However, our dataset has an unexpected limitation: it does not tell us what position the player is. We can still identify Quarterbacks and Running Backs, but we're not able to distinguish a difference between Wide Receivers and Tight Ends. This is a problem because Tight Ends should not be held to the same standard as Wide Receivers, as their statistics are often lower. For this reason, we exclude Tight Ends from our analysis.

We restrict our training model to include only players whose career started after 1960, and ended before 2006. The game of football has evolved a lot over the years, and those who played the game in 1920 were playing a very different game from the one we watch today. For our analysis, it's important that we are comparing players who played a similar style of game. Given that the 1960s brought the advent of the

Super Bowl, we believe this time period is an appropriate cutoff point for our earliest players.

Also, because so few players are inducted each year, the average wait time to be inducted is about 12 years (after retirement). For this reason, we will use only players who retired in 2005 or earlier to train our model.

3 Data Collection and Cleaning

3.1 Data Acquisition

Acquire Initial Dataset

The core source of our data comes from the Kaggle dataset *NFL Statistics*, which has individual player statistics, organized by year. The dataset is separated into multiple tables, so we start by collecting the three tables we want: Passing, Rushing, and Receiving. We then merge these three tables together, so that we have one table with all of our information. The merge creates null values for columns that were not shared, so we replace all of the null values with 0.

Merge in Supplemental Data

To increase our number of relevant features, we bring in more data from another source, *www.pro-football-reference.com*. Here we get 6 supplemental tables, and merge them with our initial dataset. The tables are: Hall of Famers, Super Bowls, Game-Winning Drives, Playoff Game-Winning Drives, MVPs, Super Bowl MVPs.

3.2 Data Cleaning

Remove Text

Some of the number fields have text mixed in with them. This includes comma separators, and the letter 'T' where there is a tie. We remove both of these text characters. Also, number fields have the text '--' to indicate no value. We replace all of

these instances with 0. With our number fields in correct form, we adjust columns to be type *integer* and *float*, as appropriate.

Fix Position Column: Remove Nulls and Reformat

The only shared column that had null values was *Position*. To fill these nulls, we create a function that is able to determine the player's position. The function looks at passing yards, rushing yards, and receiving yards, and determines which is the highest. It then assigns the player a position of quarterback (QB), running back (RB), or wide receiver (WR). Rather than assign the actual text, it assigns it a number 0, 1, or 2. We run the function on all values in the *Position* column.

4 Data Pre-Processing

4.1 Create New Columns

Create First and Last Year Columns

Next, we create two new columns. One for the first year, and one for the last year of the player's career. This will allow us to filter the data by year later on.

Adjust for Inflation

Over the years, the game of football has changed, and many player statistics are much higher now than they used to be. Because of this, we create a function that will calculate the inflation rate for each year, and apply the rate to the original value, thus giving us an inflation-adjusted metric. We apply this function to the columns: *Passing Yards, TD Passes, Rushing Yards, Rushing TDs, Receiving Yards, Receiving TDs*. The full inflation calculation is listed in the Appendix.

Combine Features

We have fields that are highly related and so we need to create a few interaction variables. This will help us reduce the total number of features, and simplify our model.

The following is a list of the variables created. These can also be found in the Appendix.

- ***SB*** - Total Superbowls (Super Bowl Wins + Super Bowl Losses)
- ***RRYd*** – Receiving Yards + Rushing Yards
- ***RRTD*** - Receiving Touchdowns + Rushing Touchdowns

4.2 Transform the Data

Transform to Career Stats

We need to transform our data from yearly stats, to career stats. We accomplish this by pivoting the data. We use all of our non-number fields (along with our new First and Last Year columns) as the index, and the result is a table with each player's career stats.

Recalculate Ratio Columns

Some of our columns are ratios of other fields (ex: *Passing Yards Per Game, Completion Percentage, Yards Per Carry*). When we pivot our data to into career statistics, these ratios become the average of each year's ratio. It is more accurate for us to recalculate these ratios, since we have all the necessary information in the other columns. We recalculate these columns.

4.3 Lasso (L1 Regularization)

A problem we have with our dataset is that many of the variables are correlated with each other. For example, we have *Rushing Attempts, Rushing Yards, Rushing Touchdowns,* and *Rushes over 20 yards*. If we have too many similar features, it will create noise, and hurt our model's performance.

Our solution to this problem is to use the Lasso method to perform feature selection. We standardize our features, and then run the Lasso method with a few different values for alpha. We choose to use alpha at .01, because it gets rid of most of the noise, but still keeps our best performing features.

alpha	# of features with coefficient > 0
1	0
0.1	0
0.01	10
0.001	24
0.0001	40

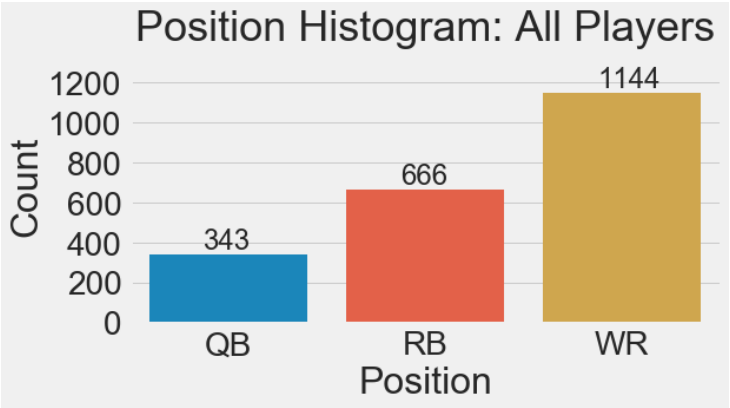
5 Data Exploration

5.1 Introduction to the Cleaned Dataset

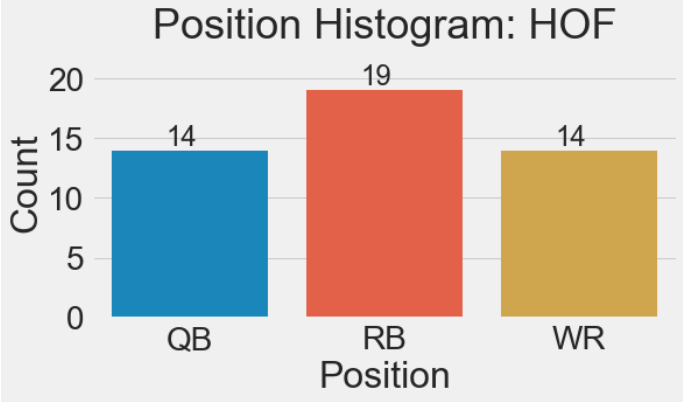
Our cleaned dataset consists of 2,153 rows and 14 columns. Each row is an individual player, and the columns contain general information and career statistics. The statistics we have are in regards to passing, rushing, and receiving. We also have the number of Super Bowls played in, and the number of MVP (Most Valuable Player) awards won. The full description of each column is listed in the Appendix.

Of the 2,132 players, only 47 of them are in the Hall of Fame (2.1%). This is a very imbalanced dataset, which may create some challenges for us when we create our model. We'll revisit this problem later in the Modeling section.

The positions in our dataset are quarterback, running back, and wide receiver. A typical football offense utilizes 1 quarterback, 1-2 running backs, and 3-4 receivers. Let's look at how many players of each position our dataset has.



We can see that our dataset's distribution of positions lines up with a typical offense pretty closely. Let's also look at the distribution of positions for Hall of Fame players, and see if it is similar.



The position distribution of Hall of Famers is much more even than the one we saw for all players. Even though quarterbacks represent only about 16% of all players in the dataset, they represent 31% of the Hall of Famers. We assume this is due to the high importance of the Quarterback position. Quarterbacks are often the leader of the team, and tend to get most of the credit for a team's success.

We know there are differences between each position in the game of football. The responsibilities are different, as are the metrics for success. Because of this, we'll look at each position individually as we explore what it is that qualifies a player to be in the Hall of Fame.

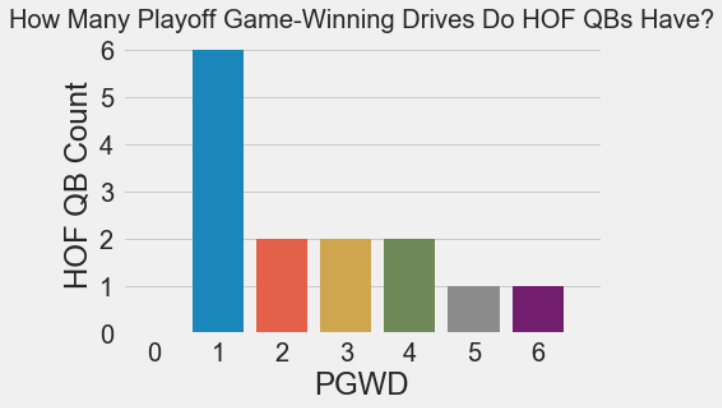
5.2 Quarterbacks (QB)

What makes a Hall of Fame (HOF) quarterback? We can begin to answer this question by separating our data to only quarterbacks, and then running a correlation test with our *HOF* variable. This is a list of our features and their correlation coefficient with *HOF*.

Variable	Corr Coeff
HOF	1
PGWD	0.757597
TD Passes adj	0.687025
SB MVP	0.622009
MVP	0.550945
RRTD	0.544044
Rushing Yards adj	0.527474
SB	0.525957
RRYd	0.524371
Passing Yards Per Game	0.436127
Receiving Yards Per Game	0.235078

We see that *Playoff Game-Winning Drives* has the highest correlation. This is followed by passing touchdowns, and then by both MVP awards. Lower down we have rushing statistics, Super Bowl appearances, and finally some per game statistics.

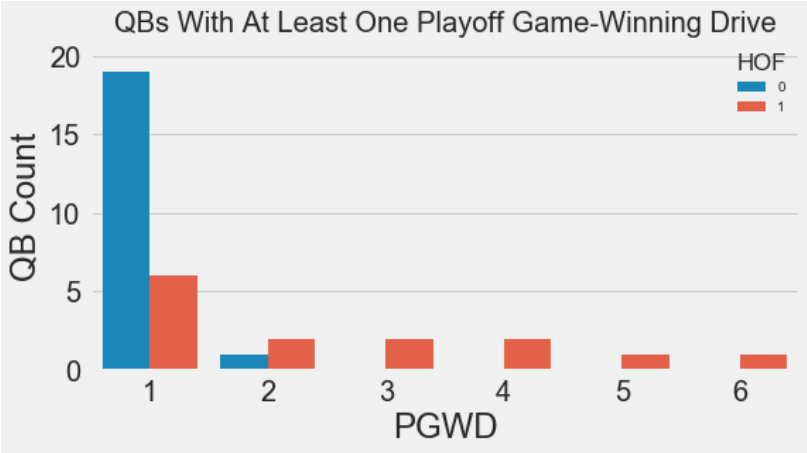
Since it's at the top of the list, let's look at *Playoff Game-Winning Drives*. Of our 14 HOF quarterbacks, how many playoff game-winning drives does each one have?



We can see that every one of our HOF quarterbacks has *at least* one PGWD. So, if you want to be a HOF quarterback, you better lead your team to a late victory in the playoffs at least once.

On the flip side, having a PGWD does not necessarily get you into the HOF. There are 19 quarterbacks with one PGWD who are not in the HOF. There is also one quarterback with two PGWDs who is not a Hall of Famer (Joe Theismann).

However, every QB in our dataset who has 3 or more PGWD is in the HOF. Here's a plot that shows this. Blue bars represent non-HOF quarterbacks, and red bars indicate those that are in the HOF.



5.3 Running Backs (RB)

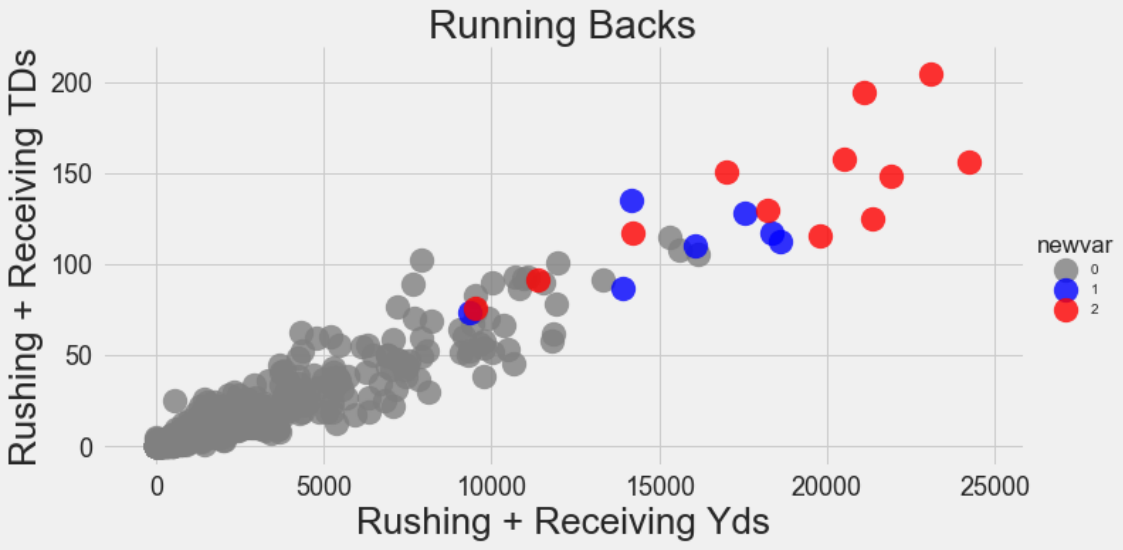
The first area we will look at is MVP awards. There are 13 running backs that have won either a League MVP or a Super Bowl MVP (3 have won both). All of the running backs with one of these awards are in the Hall of Fame, except for one, Ottis Anderson. Given this, we know that it is very likely that these award winners will be in the Hall.

Since 12 out of the 13 MVP award winners are in the HOF, this means there are 7 more running backs in the hall, who have not won an MVP. Let's see how the remaining 7 running backs qualified themselves. The following scatter plot shows the relationship between yardage and touchdowns (rushing and receiving combined). The 7 remaining running backs in are labeled in blue; the full legend is listed below.

Red – HOF Running Backs with at least one MVP (Super Bowl or League MVP)

Blue – HOF Running Backs without an MVP award

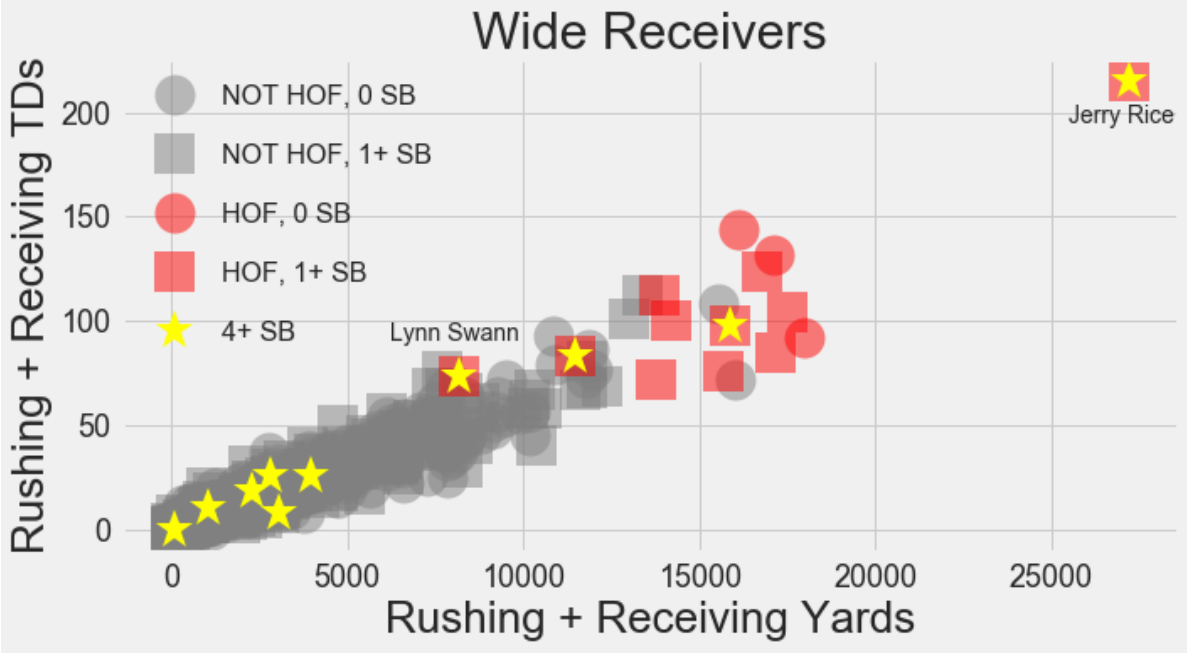
Gray – Non-HOF Running Backs



We can see that in addition to MVP awards, yardage and touchdowns are also influential in whether or not a running back makes the HOF. The lowest valued blue dot, at approximately the point (9000,75), is Gale Sayers. Sayers has no MVP award, and relatively lower statistics compared to other HOF RBs, but this can be attributed to the fact that his career was cut short due to multiple injuries.

5.4 Wide Receivers (WR)

It is very rare for a Wide Receiver to win an MVP award, so it is not the best category to help us identify Hall of Famers. Instead we will look at yards, touchdowns, and Super Bowl appearances. The following scatterplot is the best way to illustrate the relationship between these variables.



In this plot, HOF Wide Receivers are in red. Squares represent someone who has been to at least one Super Bowl. Shapes with a gold star in the middle show a player who has been to 4 or more Super Bowls. This gold star explains why the two red squares who are lagging in yardage and touchdowns are still in the HOF. The shape in the distant top right is Jerry Rice. The red square with the lowest yardage and touchdowns is Lynn Swann, who also won a Super Bowl MVP award.

6 Modeling

6.1 Introduction to Modeling

We will be testing a number of supervised learning algorithms in order to create the best possible classifier. **The algorithms we will test are: Random Forest, Logistic Regression, Support Vector Classifier, and Gradient Boosting**. We will narrow down to the best classifier, tune the parameters, and then retest it with optimal parameters.

Our dataset is very imbalanced, as only 2.16% of our players are in the Hall of Fame. Because of this, we will test a few resampling methods to try and improve our

classifier's ability to predict the positive class. **The resampling methods we will test are: Random Over Sampling, SMOTE, and Random Under Sampler.**

Because our data is so imbalanced, a classifier that predicts "Not Hall of Fame" for every player, would still have 97.84% accuracy. For this reason, we must look at performance metrics other than accuracy rate. We will look at Precision, Recall, and F1 score, and **we will use F1 score to make our ultimate decision**.

6.2 Initial Testing

We begin by separating our data into a training and testing set. We use 25% of our data in the test set, and the remaining 75% in the training set. Our response variable is *HOF*. Our predictor variables are: *SB MVP, RRTD, Receiving Yards Per Game, MVP, Rushing Yards adj, PGWD, TD Passes adj, RRYd, Passing Yards Per Game, SB, SB Win,* and *Position* (see Appendix for definitions).

We test our four classifiers and three re-samplers simultaneously, to make sure we find the best possible combination. Since our dataset is on the smaller side, we run our test 50 times and take the average of each metric. Here are our results, sorted by F1 score.

Classifier	Sampler	Accuracy	F1	Precision	Recall
GradientBoosting Classifier	nosampler	0.993321	0.824689	0.87056	0.803636
	SMOTE	0.990538	0.812408	0.797711	0.842179
	RandomOverSampler	0.991095	0.805656	0.799634	0.82599
RandomForest Classifier	nosampler	0.992022	0.785662	0.905429	0.699957
LogisticRegression	nosampler	0.989239	0.770769	0.844451	0.71984
RandomForest Classifier	RandomOverSampler	0.98961	0.756497	0.780022	0.746907
	SMOTE	0.988312	0.738019	0.791825	0.709014
GradientBoosting Classifier	RandomUnderSampler	0.960297	0.554085	0.401329	0.954762
RandomForest Classifier	RandomUnderSampler	0.956401	0.549563	0.392505	0.993333
LogisticRegression	SMOTE	0.956401	0.486309	0.334287	0.984615
	RandomOverSampler	0.948237	0.420189	0.269592	1
	RandomUnderSampler	0.926345	0.358347	0.224078	0.977381
SVC	SMOTE	0.44026	0.072452	0.037694	1
	RandomUnderSampler	0.187384	0.050136	0.025757	1
	RandomOverSampler	0.97885	0	0	0

	nosampler	0.980519	0	0	0

The top performing classifier is Gradient Boosting, and the best combination is Gradient Boosting with no resampler. The Gradient Boosting classifier was able to take the top 3 spots, so we are comfortable moving forward with it as our algorithm of choice. We will move on to hyper-parameter tuning.

6.3 Hyper-Parameter Tuning

For the Gradient Boosting algorithm, there are 5 hyper-parameters that we will tune. We will use the *GridSearchCV* tool to help us find optimal values. The hyper-parameters, and the values we will test for each one, are as follows.

Hyper-Parameter	Values
'max_depth'	[5, 6, 7, 8],
'max_features'	[0.1, 0.2, 0.3, 0.4],
'min_samples_leaf'	[5, 10, 20, 50],
'min_samples_split'	[0.05, 0.075, 0.1],
'subsample'	[0.5, 0.75, 0.9, 1]

After performing grid-search, we were able to achieve a test accuracy rate of about 99.4%. The optimal parameters are as follows.

Hyper-Parameter	Values
'max_depth'	7
'max_features'	0.2
'min_samples_leaf'	5
'min_samples_split'	0.1
'subsample'	0.9

Now that we have our optimal parameters, we'll retest the model. Just to be sure we have the best possible model, we'll retest each resampling method as well.

Classifier	Sampler	Accuracy	F1	Precision	Recall
GradientBoosting Classifier	RandomOver Sampler	0.994286	0.875867	0.841344	0.922296
	nosampler	0.992764	0.833550	0.875636	0.807443
	SMOTE	0.992393	0.832843	0.765702	0.925740
	RandomUnder Sampler	0.963636	0.560210	0.401982	0.983445

We see an improvement in our F1 score of over 5%! We also see a change that the Random Over Sampler is the best resampling method. Now that we have our optimal model, we can test it on current NFL players, and see who it predicts to be a future Hall of Famer.

7 Future Hall of Famers

We will now take our trained model, and allow it to make predictions on current NFL players. Our previous dataset consisted of all players whose last year of play 2005 or earlier, so we will define current players as anyone whose last year of player was after 2005. We can see that this includes some people who have already been inducted into the Hall of Fame (these players are highlighted in green)

Player	Probability
Brady, Tom	99.92%
Manning, Peyton	99.85%
Tomlinson, LaDainian	99.73%
Ward, Hines	99.52%
Brees, Drew	99.50%
Warner, Kurt	99.27%
Owens, Terrell	99.27%
Manning, Eli	98.66%
Harrison, Marvin	97.98%
Gore, Frank	97.03%
Smith, Steve	95.77%
Favre, Brett	94.82%
Rodgers, Aaron	86.15%
Roethlisberger, Ben	78.78%
Boldin, Anquan	75.91%
James, Edgerrin	72.24%
Forte, Matt	72.17%
Fitzgerald, Larry	68.15%
Alexander, Shaun	66.40%
Jackson, Steven	61.84%
Barber, Tiki	60.74%
Wayne, Reggie	60.44%
Flacco, Joe	55.27%
Peterson, Adrian	47.50%
Hasselbeck, Matt	30.86%
Dillon, Corey	29.51%
Mason, Derrick	27.55%
Taylor, Fred	24.44%
Rivers, Philip	14.00%
Johnson, Chris	11.62%
Charles, Jamaal	11.03%
Testaverde, Vinny	9.45%

The model will predict a player will be in the HOF if their probability is 50% or higher. We can see that the model did well in making its predictions. It was able to correctly predict all four players who have already been inducted into Hall of Fame whose career ended after 2005.

8 Conclusion

In conclusion, we are happy to have created a model that is able to classify a test set with 99.4% accuracy, 84% precision, and 92% recall. We have proven that predicting Hall of Famers is a problem that can be solved using data science and machine learning. While the model performs well, we believe there are a few limitations, and some room for improvement.

8.1 Limitations

The first limitation is that our dataset is small. There is no real solution for this other than time. As more time goes by, we will have more players to train our model on.

The other obvious limitation is that we restricted our analysis to only a few positions. Future study could leverage other data sources to predict Hall of Famers in the many other NFL positions.

8.2 Room for Improvement

We could improve our model by acquiring other relevant data and statistics. A couple of other features that come to mind are the number of Pro Bowl teams and the number of All-Pro teams that a player has been on.

Also, because we used a resampling method, our model has better recall, at the price of worse precision. What this means is that we are less likely to miss any Hall of Famers, but we are more likely to classify a Non-HOF player as HOF. Depending on the goal of the analysis, we could tweak our prediction probability threshold (ex: only predict HOF if probability is over 90%), or possibly not use a resampling method

Appendix

Variables Used in Modeling

adj: means that metric is adjusted for inflation

inflation calculation: $\frac{a}{b} = c$

$a = Mean\ of\ the\ top\ 15\ players\ in\ year\ 2016$

$b = Mean\ of\ the\ top\ 15\ players\ in\ current\ year$

$c = Inflation\ Multiplier\ (this\ number\ is\ multiplied\ by\ the\ original\ value)$

Columns	Data Type	Description
RRTD	float64	Rushing TDs adj + Receiving TDs adj
SB MVP	float64	Super Bowl MVP awards
Receiving Yards Per Game	float64	Receiving Yards adj / Games Played
Rushing Yards adj	float64	Rushing Yards adjusted for inflation
MVP	float64	Most Valuable Player Awards won
PGWD	float64	Playoff Game-Winning Drives
TD Passes adj	float64	TD Passes adjusted for inflation
RRYd	float64	Receiving Yards adj + Rushing Yards adj
Passing Yards Per Game	float64	Passing Yards adj / Games Played
SB	float64	Super Bowl Appearances
Name	object	Player's Name
Player Id	object	Individual Identifier
HOF	int64	Hall of Fame Status 1 – In Hall of Fame 0 – Not in Hall of Fame
Position	int64	0- Quarterback 1- Running Back 2- Wide Receiver

All Original Variables

Columns	Data Type
Player Id	object
Name	object
Position	int64
Completion Percentage	float64
First Down Receptions	float64
First Year	int64
Games Played	float64
HOF	int64
Int Rate	float64
Ints	float64
Last Year	int64
Pass Attempts Per Game	float64
Passer Rating	float64
Passes Attempted	float64
Passes Completed	float64
Passes Longer than 20 Yards	float64
Passes Longer than 40 Yards	float64
Passing Yards Per Attempt	float64
Passing Yards Per Game	float64
Passing Yards adj	float64
Percentage of Rushing First Downs	float64
Percentage of TDs per Attempts	float64
Receiving TDs adj	float64
Receiving Yards Per Game	float64
Receiving Yards adj	float64
Receptions	float64
Receptions Longer than 20 Yards	float64

Receptions Longer than 40 Yards	float64
Rushing Attempts	float64
Rushing Attempts Per Game	float64
Rushing First Downs	float64
Rushing More Than 20 Yards	float64
Rushing More Than 40 Yards	float64
Rushing TDs adj	float64
Rushing Yards Per Game	float64
Rushing Yards adj	float64
SB Loss	float64
SB Win	float64
Sacked Yards Lost	float64
Sacks	float64
TD Passes adj	float64
Yards Per Carry	float64
Yards Per Reception	float64
Year	float64
MVP	float64
SB MVP	float64
GWD	float64
PGWD	float64
SB	float64
RRYd	float64
RRTD	float64