Predicting Pro Football Hall of Fame Players Using Machine Learning

Alexander I. Samardzich <asamardz@stanford.edu>

***Abstract*—In this project, four machine learning models were trained to predict the Pro Football Hall of Fame status of professional NFL players. Using a dataset including 5,913 former players (later downsampled to 1,165) with 96 input features, models were trained using logistic regression, SVM with a linear kernel, SVM with a gaussian kernel, and a neural network. At testing, the most successful model was the SVM with a linear kernel, producing a test set F1 score of 92.75%, precision of 91.43%, and recall of 94.12%. The success of the linear kernel is attributed to its avoidance of overfitting by optimizing the margin between both classes while still making accurate predictions. The final output of this project is a list of 28 current and recently retired NFL players who were classified by the model as future Hall of Fame members.**

I. Introduction

THE objective of this project is to build a machine learning model that can accurately determine whether a professional NFL player will be inducted into the Pro Football Hall of Fame (HOF). The models tested in this project accept a vector of career aggregate quantitative statistics for a professional football player and output a prediction of their HOF candidacy. This is an interesting machine learning challenge because there are no set guidelines for what a player must achieve during their career in order to enter the HOF. Instead, a selection committee of 48 media representatives vote on players who have been nominated by fans. The only restriction is that the nominee must have played football for at least five seasons [1]. The subjective nature of the process by which players are chosen for the HOF means that the potential success of any model is constrained by the errors and biases of the selection committee.

Each year, countless sports writers and reporters debate which of their favorite players they believe will make it into the HOF. With the completion of this project, there will hopefully be another data point as to why that special player should be remembered as one the best.

II. Related Work

The task of using machine learning to identify exceptionally talented players in sports is not a new one. Young et al. (2008) built models to predict HOF candidacy for professional baseball players, achieving a 98% accuracy with an artificial neural network, 6% higher than with their logistic regression model [2]. They believe one factor in the success of the neural network is its ability to fit complex, non-linear datasets. Freiman (2010) also found success with neural networks and random decision forests in predicting baseball HOF candidacy, though they suffered from high variance [3]. The superiority of neural networks over logistic regression was not found to be the case in this project, potentially due to differences in degree of non-linearity between baseball and football statistics.

One important difference between the task of classifying baseball players and the task of classifying football players is that the statistics used to measure performance for each position in football tend to differ much more meaningfully than those in baseball. The batting proficiency of each baseball player is measured by essentially the same statistics regardless of position through the use of common metrics such as batting average and on base percentage (somewhat different for pitchers). In contrast, offensive proficiency for different positions in football is measured through disparate metrics, for example receiving yards for wide receivers vs. rushing yards for running backs. This is to say that classifying exceptional football players in a single model can be a challenging task because there are more dimensions on which a player can be above average than with baseball.

While less effort has been invested in identifying HOF players in football than in baseball, comparable work at the team level has been done to try to determine what factors are correlated with teams that make the playoffs in a given season. Fokoue et al. (2013) filtered through numerous NFL statistics to identify which features were strongest predictors of playoff-caliber teams [4]. Interestingly, they found that after removing features that had limited discriminating power, they were left with 13 offensive statistics that were strong predictors of whether a team made a playoff appearance. While their research was biased towards current trends in football as they only explored the 2006 to 2010 seasons, their emphasis on the importance of offense is echoed in the findings of this project. Whether it is simply because offensive statistics are more heavily tracked or because there is larger spread in talent across offensive players allowing for easier identification of superstars, classifying offensive players was consistently easier throughout this project. This can be seen in the final predictions of future HOF members at the end of this report where 17 offensive players were identified compared to only 11 defensive ones.

III. Dataset

In order to construct a successful machine learning model, sufficiently informative player statics will need to be used as features to distinguish HOF quality players. All data used in this report was collected from Pro-Football-Reference.com, a subcomponent of the large online sports database Sports-Reference.com [5]. A data scraper was written in python to pull down statistics from the website for all NFL players at once [6]. Originally, all NFL players since 1920 were included in the dataset with career aggregate statistics acting as features in categories such as passing, receiving, rushing, tackles, interceptions, playoff games, and more. Each player has a binary classification for whether they are a member of the HOF. Before the data was fed to each model, certain players were excluded. All active players and all players who retired later than the end of the 2013 season were removed because a player must be retired for 5 years before they can be considered for the HOF; these players technically have no classification. Next, all players who played fewer than five seasons of professional football were excluded because they are ineligible for nomination to the HOF. Additionally, individuals who played in the NFL, but were inducted into the HOF for their work as a coach or owner, not for their performance as a player, were removed (for example Tony Dungy).

Finally, after close inspection of false negatives (HOF members who were not predicted) during initial testing, it became clear that all models were having difficulty classifying certain positions, particularly offensive linemen. In retrospect, this is not entirely surprising as there are essentially no position specific statistics for centers, guards, and tackles that accurately represent their skills and contributions to the game. Impact and skill of offensive linemen could be measured indirectly by looking at statistics such as sack rate and time to throw for their quarterback or yards before contact for their running backs. Unfortunately, these statistics would require stat lines for each NFL game played, which is outside of the scope of this project. Instead, all offensive linemen were excluded from the dataset. After the final filtering, there were 5,913 players with 96 features in the dataset. In total, 233 Hall of Famers were included, representing 3.94% of total datapoints. The data was shuffled and then split into sets of 70% training, 15% cross validation, and 15% test. Finally, the data was normalized by the mean and standard deviation of the training set.

IV. Methods

The first model tested in this project was logistic regression because it is typically successful on classification problems with a modest number of features and can be implemented quickly. The standard binary cross-entropy loss function with regularization parameter λ, as seen below in Equation 1, was used to train the model, where m is the number of training examples and n is the number of features. The sigmoid function was used for prediction function h_θ, as seen below in Equation 2.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}log(h_\theta(x^{(i)})) + (1-y^{(i)})log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \quad (1)$$

$$h_\theta(x^{(i)}) = \frac{1}{1 + e^{-\theta^T x}} \tag{2}$$

A threshold of 0.5 was used to round the predictions into 0 or 1 classifications. Within MATLAB, the *fminunc* function was used to minimize cost. Because the dataset is skewed and the focus of the project is on the individuals who make the HOF (true positives), accuracy is not the best measure of performance. Instead, F1 score, as seen below in Equation 3, was used to optimize hyperparameters of each model through testing on the validation set.

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{3}$$

The second model tested utilized Support-Vector Machines (SVM) with a linear kernel. SVM models are very effective for classification problems with limited features and a dataset of intermediate size (below 50,000); they are also often able to produce superior decision boundaries over logistic regression through use of kernel functions. The *fitcsvm* function within MATLAB was used to train the model with inputs for BoxConstraint and KernelScale that acted as regularization parameters for the loss function and for controlling margin, respectively. SVM models minimize the same cost function as logistic regression, the only difference being the leading constants before each term and the hypothesis function. The *predict* function was then used to obtain predictions for the trained model. After optimizing the regularization parameters, a gaussian kernel was then tested to try to achieve a superior decision boundary, less vulnerable to overfitting.

The final model tested in this project is a standard neural network. The network had an input layer of 96 nodes, a hidden layer of 50 nodes, and an output layer of one node representing the HOF prediction. Weights within the network were randomly initialized with a mean of zero and range of 0.12 and the sigmoid activation function was applied after each layer. The same cross-entropy loss function with regularization as seen in logistic regression was minimized using the *fmincg* function in MATLAB. While the advantages of neural networks over logistic regression are typically seen on datasets with significantly more features and examples than is the case in this project, a neural network was still included in the event that non-linearities in the dataset were better captured through the use sigmoid activation functions.

V. Experimental Results and Discussion

In total, four models were trained in this project to predict HOF chances for NFL players using: logistic regression, SVM with a linear kernel, SVM with a gaussian kernel, and a neural network. Regularization parameters for each model were tuned using a windowing process to identify which values led to the highest F1 score on the validation set. In the example of selecting λ for logistic regression, powers of 10 were tested from 0.001 to 1,000 to begin narrowing in on an optimal value. A similar process was used for the BoxConstraint and

		Full Dataset				**With Downsampling**			
		Logistic Reg.	*SVM Linear*	*SVM Gaussian*	*Neural Network*	*Logistic Reg.*	*SVM Linear*	*SVM Gaussian*	*Neural Network*
Hyper-param.	λ	2.0			3.0	3.5			2.5
	BoxConst.		1.00	3000			0.05	1500	
	KernelSc.		2.50	300			1.00	100	
Training Set	*Accuracy*	98.77%	98.86%	98.72%	98.86%	97.55%	97.67%	98.02%	97.67%
	Precision	90.67%	90.91%	91.67%	90.91%	95.63%	95.09%	95.73%	96.23%
	Recall	78.61%	80.92%	76.30%	80.92%	92.17%	93.37%	94.58%	92.17%
	F1	84.21%	85.63%	83.28%	85.63%	93.87%	94.22%	95.15%	94.15%
Validation Set	*Accuracy*	98.20%	98.08%	97.97%	98.08%	95.40%	94.25%	95.40%	95.40%
	Precision	78.57%	74.19%	73.33%	75.86%	87.88%	82.86%	87.88%	87.88%
	Recall	68.75%	71.88%	68.75%	68.75%	87.88%	87.88%	87.88%	87.88%
	F1	73.33%	73.02%	70.97%	72.13%	87.88%	85.29%	87.88%	87.88%

TABLE I
HYPERPARAMETERS AND PERFORMANCE METRICS BY MODEL

KernelScale hyperparameters in the SVM models.

After hyperparameter tuning and model training, F1 scores on the validation set were similarly low for all the models, suggesting that higher performance would need to be driven by improvements to the dataset they all shared. One clear weakness of the dataset was the large class imbalance. To alleviate this, downsampling was employed, and negative examples were randomly removed until HOF members represented 20% of the dataset (233 positive examples and 932 negative examples). The value of 20% was chosen to alleviate some of the class imbalance while still providing the model with enough examples of above average players who did not make the HOF in order to learn a robust decision boundary. After downsampling, performance significantly improved across all models, as seen above in Table 1. The only metric to decrease with downsampling was accuracy, but that is because the full dataset includes significantly more negative examples, padding the statistic.

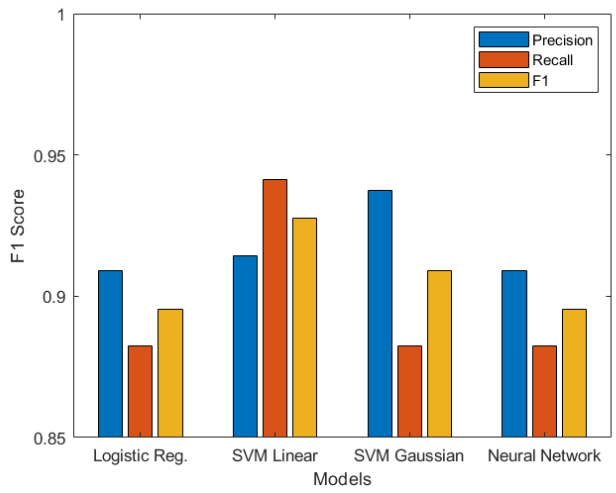

Fig. 1. Test Set Performance by Model

With downsampling in place, performance of each model was then compared using the test set, as seen above in Figure 1. Performance was relatively similar across models, with the SVM model utilizing a linear kernel coming out slightly ahead of the other three. The success of the linear kernel over logistic regression is because SVM models have the additional regularization parameter KernelScale, which controls margin and helps to avoid overfitting to the training and validation sets. Regarding the SVM model utilizing a gaussian kernel and the neural network, it is less clear why the linear kernel was superior. The sigmoid activation functions in the neural network and the use of a gaussian kernel allow the two models to draw more complex decision boundaries, better fitting non-linear datasets. Its possible that the dataset used here was sufficiently linear that this advantage was lost for the two models and that the way in which the data was randomly shuffled happened to favor the linear kernel during testing.

It is worth noting that all models suffered from high variance, even with the use of regularization, as highlighted by the 6-8% decrease in F1 score from the training set to validation set. Unfortunately, due to the nature of this project, many of the common techniques to reduce overfitting were not feasible. For example, it is not possible to obtain additional training examples as all NFL HOF players are currently included in the dataset. Additionally, it is difficult to reduce the number of features without losing relevant information because there is a large variety of player positions that rely on completely different statistics from one another to measure performance (e.g. quarterback vs. cornerback). Over time, however, the model will improve as more and more players enter the HOF. This includes updating the status of fringe players such as Ken Stabler, who was inducted over 30 years after his playing career ended [7].

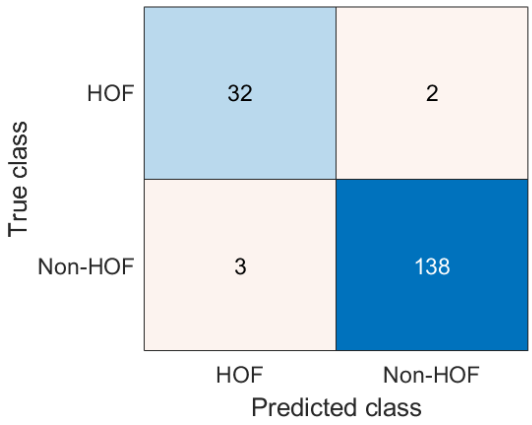

Fig. 2. Confusion Matrix for Linear Kernel SVM Model on Test Set

In the end, the SVM model with a linear kernel pro-

duced the highest test set F1 score of 92.75%, with a precision of 91.43%, and recall of 94.12%. The confusion matrix for the test set, as seen in Figure 2 above, highlights the success of the model, as only 5 examples are mislabeled. The misclassification of those examples is likely due to the subjective manner by which players enter the HOF. The fact that players are elected to the HOF by a selection committee and not through a deterministic process means that the data itself is, in a sense, not perfectly labeled. This was highlighted throughout the error analysis phase of model training, where many of the false positives in the training set were players whose exclusion from the HOF are commonly cited by sports writers and bloggers as great oversights, for example Cliff Branch [8].

VI. Conclusion and Future Work

Four machine learning models were trained in this project to classify the Hall of Fame status of NFL players. Upon completion, the model utilizing SVM with a linear kernel produced the highest test set performance with a F1 score of 92.75%, precision of 91.43%, and recall of 94.12%. Of the 175 examples in the test set, there were only 3 false positives and 2 false negatives. The success of the linear kernel is likely due to its ability to avoid overfitting the training and validation sets through use of the KernelScale parameter and the potential linear behavior of many of the features in the dataset. Using the linear kernel SVM model, Table 2 to the right was compiled, representing predictions for future HOF members among players who are currently active or have retired within the past 4 seasons. The table is a testament to the success of the model as most of the players included are considered guaranteed future HOF members, while the remaining names are at least in the running or common topics of debate [9].

While this project was successful in building an initial HOF classifier for NFL players, there is further work that could be done to improve performance. With additional team members and computational resources, it would be interesting to collect game-by-game statistics and generate measures for offensive linemen performance in order to allow for their classification in the models.

Player	Position
Adam Vinatieri	K
Charles Woodson	CB
Peyton Manning	QB
John Abraham	DE
Shane Lechler	P
Tom Brady	QB
Drew Brees	QB
Reggie Wayne	WR
Dwight Freeney	DE
Julius Peppers	DE
Andre Johnson	WR
Antonio Gates	TE
Jason Witten	TE
Terrell Suggs	DE
Troy Polamalu	DB
Ben Roethlisberger	QB
Jared Allen	DE
Jason Peters	DL
Larry Fitzgerald	WR
Philip Rivers	QB
Aaron Rodgers	QB
DeMarcus Ware	DE
Frank Gore	RB
Darrelle Revis	CB
Marshawn Lynch	RB
LeSean McCoy	RB
Rob Gronkowski	TE
J.J. Watt	DE

TABLE II
Predictions for Future Hall of Fame Members Among Active and Recently Retired Players Sorted by Draft Date

VII. GitHub Repository

The python scraper, MATLAB models, and raw data used in this project can be found at https://github.com/asamardzich/Predicting-Football-Hall-of-Fame-Members.

VIII. References

[1] Profootballhof.com. (2019). Becoming a Hall of Famer Hall of Famers — Pro Football HOF Official Site. [online] Available at: https://www.profootballhof.com/heroes-of-the-game/becoming-a-hall-of-famer/ [Accessed 2 Mar. 2019].

[2] Young, William A., William S. Holland, and Gary R. Weckman. "Determining HOF status for major league baseball using an artificial neural network." Journal of quantitative analysis in sports 4.4 (2008).

[3] Freiman, Michael H. "Using random forests and simulated annealing to predict probabilities

of election to the baseball HOF." Journal of Quantitative Analysis in Sports 6.2 (2010).

[4] Fokoue, Ernest, and Dan Foehrenbach. "A Statistical Data Mining Approach to Determining the Factors that Distinguish Championship Caliber Teams in the National Football League." (2013).

[5] Page, F. (2019). Pro Football Statistics and History — Pro-Football-Reference.com. [online] Pro-Football-Reference.com. Available at: https://www.pro-football-reference.com/ [Accessed 2 Mar. 2019].

[6] Nazrul, Syed Sadat, and Syed Sadat Nazrul. Web Scraping HTML Tables with Python. Towards Data Science, Towards Data Science, 25 July 2018, towardsdatascience.com/web-scraping-html-tables-with-python-c9baba21059.

[7] Brown, Daniel, and Daniel Brown. The Late Ken Stabler Inducted into Pro Football Hall of Fame. The Mercury News, The Mercury News, 12 Aug. 2016, www.mercurynews.com/2016/08/06/the-late-ken-stabler-inducted-into-pro-football-hall-of-fame/.

[8] Weiss, Brad. Cliff Branch Rated as the Raiders Biggest Hall of Fame Snub of All-Time. Just Blog Baby, FanSided, 3 Aug. 2018, justblogbaby.com/2018/08/03/cliff-branch-rated-raiders-biggest-hall-fame-snub-time/.

[9] Deciding Top Hall of Fame Debates: Revis, Rivers, Eli, Beast Mode, More. ESPN, ESPN Internet Ventures, 2 Aug. 2018, www.espn.com/nfl/story/_/id/24230258/deciding-future-pro-football-hall-fame-debates-eli-manning-darrelle-revis-more.