Predicting Pro Football Hall of Fame Players Using Machine Learning

Alexander I. Samardzich <asamardz@Stanford.edu>

CS229a - Applied Machine Learning. Stanford University, Winter 2018-2019

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). In total, four models were trained for this task using: logistic regression, SVM with a linear kernel, SVM with a gaussian kernel, and a neural network. 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. After training and hyperparameter optimization, the SVM model with a linear kernel produced the highest test set F1 score of 92.75%, with a precision of 91.43%, and recall of 94.12%.

Data

All data used in this report was collected from Pro-Football-Reference.com [1]. Players were excluded from the dataset if they:

- *Played fewer than 5 seasons in the NFL*
- *Are active or retired later than the 2013*
- *Played offensive line*
- *Were inducted into the HOF as a coach*

After filtering out ineligible players, the dataset contained 5,913 players with 96 quantitative features in categories such as passing, receiving, rushing, tackles, and more. In total, 233 Hall of Famers were included, representing 3.94% of the dataset. The data was split into sets of 70% training, 15% cross validation, and 15% test.

Models

The models trained in this project minimized the cross-entropy loss function below with slight variations to leading constants and regularization parameters depending on the model

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

Logistic Regression: Cost function minimized using *fminunc*

SVM Models: Two models trained using a linear kernel and a gaussian kernel. Included regularization parameters BoxConstraint and KernelScale. Trained using *fitcsvm*

Neural Network: Input layer of 96 nodes, a hidden layer of 50 nodes, and an output layer of one node. Weights randomly initialized with a mean of zero and range of 0.12, with sigmoid activation function applied after each layer. Cost function minimized using *fmincg*

Results and Discussion

Initial error analysis revealed the class imbalance was limiting model performance. Downsampling by removing negatives examples until HOF members represented 20% of the dataset (1,165 total examples) resulted in immediately higher performance across all models as seen in in Table 1. Figure 1 shows test set performance for each model, with the SVM utilizing a linear kernel producing the highest F1 score; the corresponding confusion matrix is show in Figure 2. 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 set. With respect to the gaussian kernel and neural network, it's possible the dataset was sufficiently linear that their advantage in drawing non-linear decision boundaries was irrelevant or that the way in which the data was randomly shuffled happened to favor the linear kernel during testing [2]. Predictions for future HOF members using the linear kernel SVM model are shown in Table 2.

		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 1 – Hyperparameters and Performance Metrics by Model

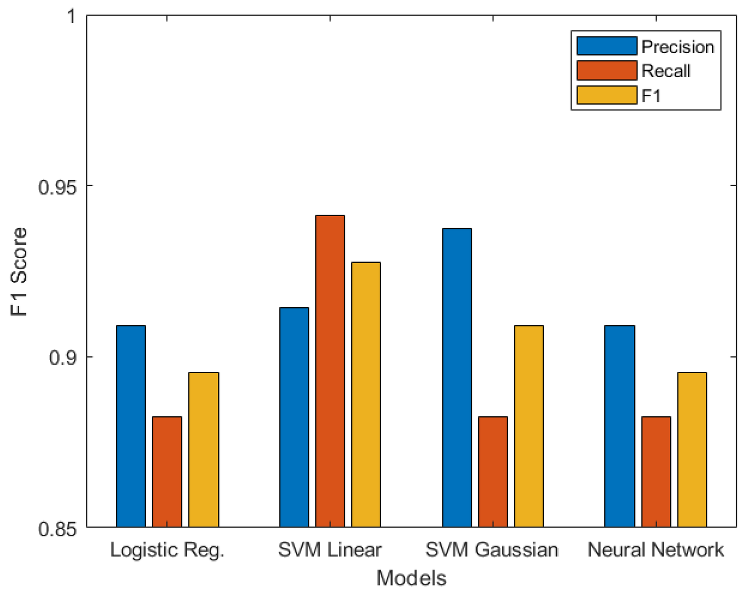

Figure 1 - Test Set Performance by Model

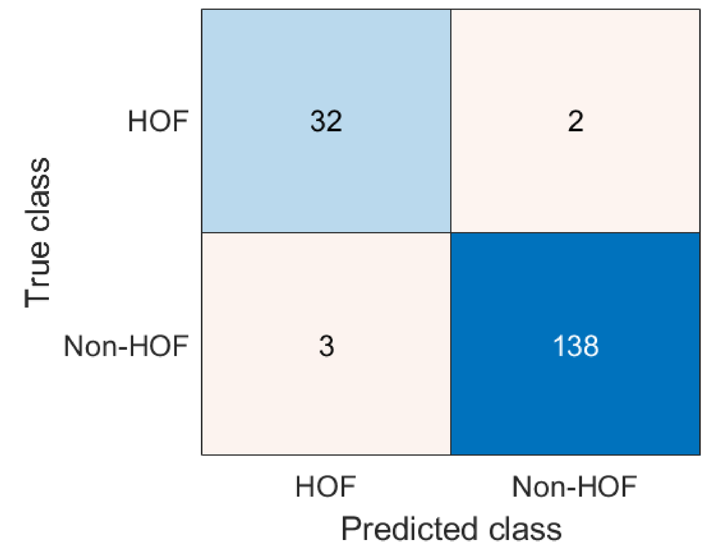

Figure 2 - Confusion Matrix for Linear Kernel SVM Model on Test Set

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 2 - Predictions for Future Hall of Fame Members Among Active and Recently Retired Players

Future Steps

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 by the models.

References

[1] 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].

[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).