NFL Play Predictor

A binary classifier - Rush vs. Pass

Author: Philip Ramirez

Most work done in sports analytics is related to game scores, outcomes, or player performance forecasting, but modeling game time decisions is a complex - yet potentially fruitful - task. From the perspective of opposing teams, coaches, and defensive coordinators having a reliable, data driven prediction of the next play would be invaluable. Furthermore, knowing which salient features determine game time decisions can ultimately influence defensive preparation as well as in-game strategy.

Data

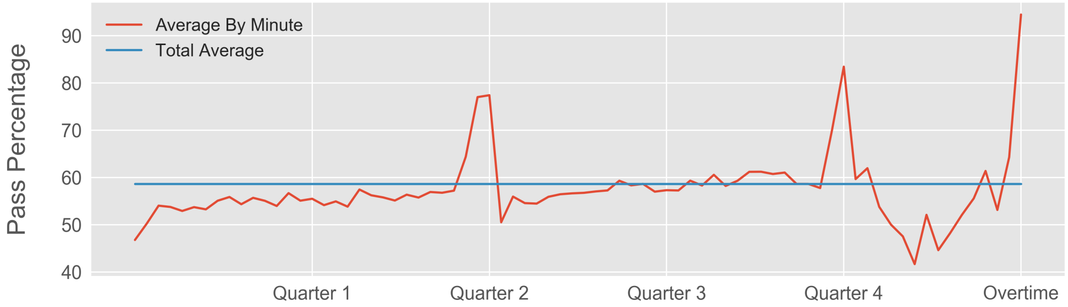
- Full play by play from 2013 through 2016. Provided by NFLSavant.com
- Weather and coach data scraped from NFLWeather.com and Pro-Football-Reference.com.
- Test set (data from season year 2017) scraped from Pro-Football-Reference.com

Model

The model feature set includes 31 features ranging from game situations to stadium conditions. The target is ***pass***. Every pass play attempt is given a value of ***1***. Every rush attempt is given a value of 0.
Three classifiers were attempted and evaluated.

	Accuracy	**Precision**	**Recall**
Logistic Regression	0.71	0.71	0.71
Random Forest classifier	0.69	0.69	0.69
Gradient Boosting Classifier	0.73	0.73	0.73

After grid searching and evaluating models with the use of AWS clusters, the Gradient Boosting Classifier seemed to perform best consistently. The five most important features, in order, include yards until first down, time left in half, field position, score differential, and formation.



I've evaluated the Gradient Boosting Classifier's performance against a baseline naive model that always selects pass (the majority class). These final model scores

- Baseline accuracy: 58%
- GB Classifier accuracy: 74%
- GB Classifier precision, accuracy, f1 score: 74%

**** When compared to the baseline accuracy of 58%, the Gradient Boosting Classifier is 30% more accurate.***