Machine Learning in American Football

Exploratory Research into the Modeling of In-Game Decision Making using Logistic Regression, Decision Trees, and Random Forests

Joseph Wysocki
Joseph.T.Wysocki@gmail.com

Introduction

When it comes to the development and advancement of data science and analytics, American football has always seemed to lag significantly behind the other four major sports. This, however, is in no way caused by a lack of available data. With companies such as Pro Football Focus (PFF), Sportradar, and Zebra logging meticulous detail on every snap, there has never been a time where more data existed. The problem arises in how to best utilize this information.

In this brief writeup, we will explore one simple question: can we predict whether a team will run or pass on any given play? While it may be small in scope, this question can lead us to great revelations in how we utilize data for in-game analysis. On top of that, we can create massive value for any coach if we can reliably and accurately anticipate the actions of any opponent.

Scope, Data, and Introductory Analysis

For the sake of exploring this question, we will limit ourselves to analyzing one specific team (Florida State) over the course of one year (2016). With this set of limitations, we are assuming that the decision patterns related to play calling are unique to each head coach. Therefore, our model will likely be more accurate if it is built around one specific individual.

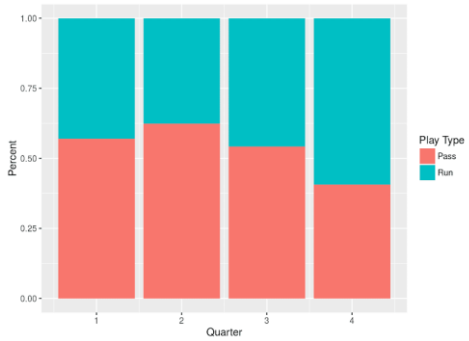
The data used here was collected by Pro Football Focus. It encompasses roughly 160 unique data points for over 700 offensive snaps ran by Florida State throughout the season. A lot of the recorded data for each snap is relatively useless and it is important that we only utilize the variables that are likely to be strong predictors of play calls.

To do this, we can explore the data visually using graphs. The table below charts out several different variables and how they relate to the overall run/pass distribution for Florida State throughout the entire season. By analyzing these graphs, we can gain a better understanding as to which characteristics lead to the highest variability amongst play type. It is important to be sure that, when analyzing these variables, we only account for those that are readily apparent on a pre-snap basis. Any potential play descriptor that only becomes available after the ball is snapped is of no use to us.

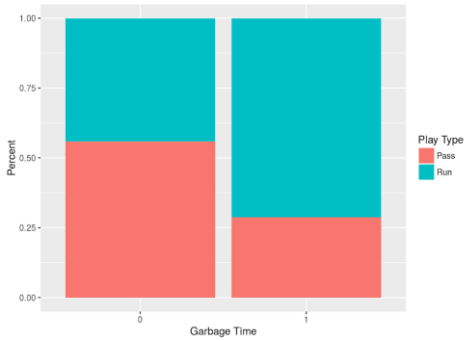
Based on the results shown in the graphs, it seems that we have a good understanding of which variables we will want to include in our model and those that we would like to exclude as they likely have little to no predictive value. As for useful variables, we will use the following: Garbage Time, Distance, Shotgun, TE Alignment, Off Personnel, and Off Formation Group. The others (Quarter, Hash, and Drive Play) will not be used going further.

Play Type Indicator Analysis

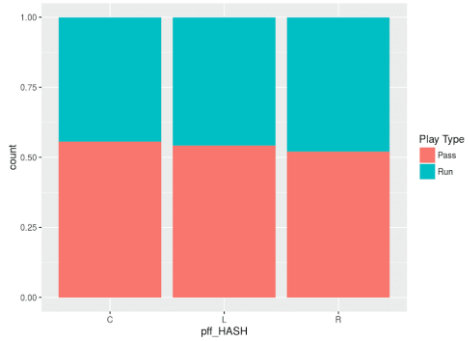
Run/Pass Splits by Quarter



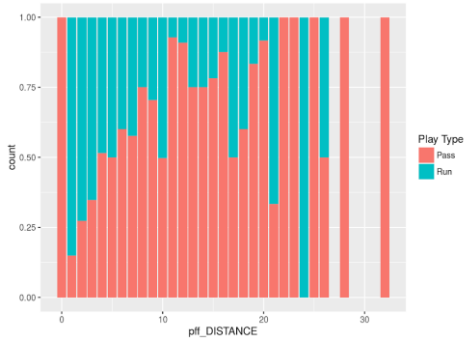
Run/Pass Splits by Score Differential



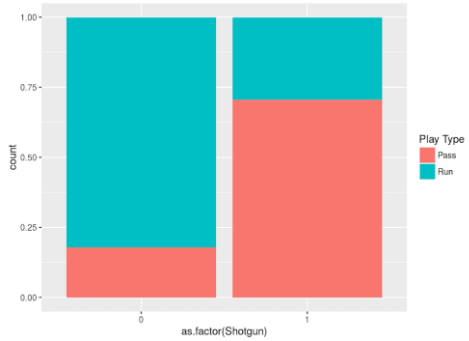
Run/Pass Splits by Field Position



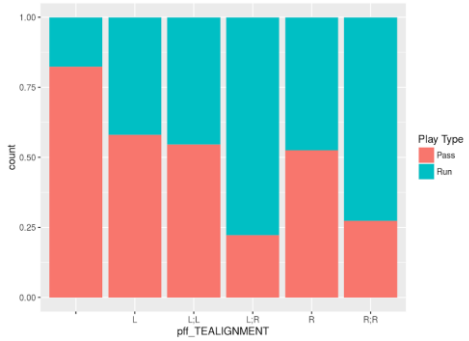
Run/Pass Splits by Distance to Go



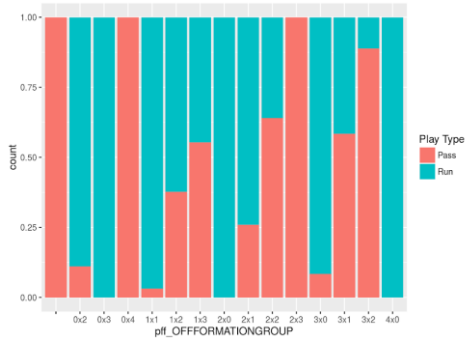
Run/Pass Splits by QB Alignment



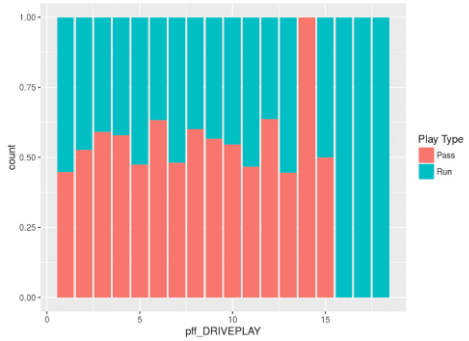
Run/Pass Splits by TE Alignment



Run/Pass Splits by Formation



Run/Pass Splits by Drive Play



Prediction Part One: Logistic Regression

Our ultimate goal is to predict whether a team will run or pass any given play. In our first attempt to model this, we can use logistic regression. Here, the output of our model will be a value from 0 to 1 and it will represent the probability that Florida State will run the ball. For prediction purposes, we will assume that any play with a run probability greater than 50% will be a run and all other plays will be passes.

With this method, over the sample size of 758 plays, our model came out to be 78% accurate – a very promising result. Additionally, this model only used five variables: Down, Distance, Shotgun, Off Form Group, and TE Alignment. Even with little pre-snap information, we were able to effectively anticipate the play calls of Florida State on 78% of all snaps.

Prediction Part Two: Decision Trees

While we made some impressive progress with logistic regression, perhaps we can make even bigger strides with machine learning. Here, we will utilize decision trees to handle the modeling and predictions for us.

For our tree, we will only use four variables: Quarter, Down, Distance, and Shotgun. On our initial training data, this model was 74.5% accurate when predicting play calls. However, on the test data set, it achieved an accuracy of 80.2% - slightly more than two whole percentage points better than our logistic regression model with one less variable.

Prediction Part Three: Random Forest

Finally, we will dig a bit deeper into our machine learning capabilities by utilizing random forest models. These models are quite a bit more complex and they are also significantly less transparent. Because of this, it is difficult to elaborate on their intricacies. When tested, our random forest model achieved an accuracy of around 79% with results reaching as high as 81%.

Conclusions

For predicting, we used three different modeling techniques - logistic regression, decision trees, and random forests. The best prediction accuracy we saw came from the random forest model. However, the other two techniques produced nearly identical results.

Overall, the goal of this process was to create a model that has a practical use within the course of a football game. With this in mind, we see the decision tree as being the most practical. First, it is relatively easy to understand conceptually. It could be presented to someone with no statistics/mathematics background and they would likely be able to understand the logic behind it. Additionally, our tree used only 4 variables, all of which would be known well in advance of

the play. Because of this, it would be easy to input them into the model before the play, get a prediction, and still have time for that prediction to be useful.

While our models may never make it to the coach's booth on game day, they do have a bright potential when it comes to pre-game analysis. By creating, testing, and re-running these models, we can quickly identify tendencies and tells that may go unnoticed during the typical film review session.

The goal is never to replace the self-evaluation done by coaches. Instead, the true goal is to leverage our large sets of data and immense computing abilities to gain an edge over opponents. By utilizing such technology, we can add immense amount of value to the already complex operations of modern day American football teams.