NFL Plays Analytics

Predicting offensive football plays

Statistical Inference (for plots of results, see plays_compare_stats_infer .ipynb)

After wrangling the NFL plays data into a suitable form, I performed inferential statistics to examine differences and relationships between and within the feature (predictor) and target (dependent) variables. The dependent variable type is categorical with 6-classes composed of play type and direction: rush vs. pass and left vs. middle vs. right.

Initially, I examined whether play call is dependent on the team's current location on field. I first scaled yardlines from 1-99, where yardline 49-1 on one's own side of the field are 51-99, reflecting the fact that one is farther away from the endzone to be scored on. With a one-way ANOVA comparing mean yardline across the play classes, I found an overall significant difference across plays (F=2.5, p=0.028, η^2=0.0004). Follow-up pairwise Tukey comparisons, found that only the mean yardline of rush middle significantly differed from pass middle and left, indicating rushes up the middle have a lower average yardline or are slightly more likely as a team approaches the 1 yardline. Nonetheless, play types nor directions were dissociated across yardline.

In football, plays are run on four separate downs, where the goal is to gain the desired number of yards in order to get a new 1st down before 4th down. First down begins with 10 yards to gain (or go) and the number of yards an offense gains or loses on each play is subtracted or added to the yards to go. Offenses typically punt the ball to the opponent on 4th down, so as 4th down approaches it becomes more imperative to gain the full yards to go. I tested whether play class depended on the current down. Because both of these variables are categorical I used a chi-square test of independence, which tests if the proportion of play classes is uniformly distributed across downs. I found play was significantly dependent upon down (χ^2=1997.95, p<0.001), with more pass plays called on 3rd down and fewer plays were directed over the middle generally.

Using a one-way ANOVA comparing mean yards to gain with play class, I found that play significantly depended upon yards to gain (F=58.34, p<0.001, η^2=0.009). Importantly, the eta-squared (η^2) value is larger than that for yardline, reflecting yards to go has a bigger impact on play than the current yardline. Follow-up comparisons found significant differences with pass > rush, pass middle > all except pass left, and rush middle < all except rush left with respect to mean yards to go. Thus, pass plays are more likely when the yards to gain is larger. Further, middle passes occurred more frequently with long yards to gain, whereas middle rushes occurred more with short yards to gain. In a subsequent one-way ANOVA, play class was significantly dependent upon the mean score difference between teams (F=166.1, p<0.001, η^2=0.025), with a larger effect size than either above. Follow-up comparisons found significant differences with rush > pass, pass middle < all except pass left, and rush left > all

except rush right with respect to mean score difference. Note score difference scale ranges from negative to positive. Hence, a team is more likely to pass when losing, and middle passes and rushes have a lower mean score difference than their outside counterparts.

I computed a chi-square test to examine if the observed frequency of play types across teams differed from the expected frequencies based on the total play distribution in the NFL. The frequency distribution for each play type significantly depended upon team (pass left: χ^2=90.06, pass middle: χ^2=184.14, pass right: χ^2=54.12, rush left: χ^2=167.55, rush middle: χ^2=497.94, rush right: χ^2=164.30; all *p's*<0.001), indicating teams have different play calling strategies as expected.

To examine the relationships between the features, I computed separate correlation matrices for cumulative game stats, cumulative drive stats, and previous play stats. The resulting r-values revealed a large degree of dependency within cumulative and previous play stats (see notebook plays_compare_stats_infer .ipynb). This colinearity suggests a number of features could be eliminated in order to optimize performance in future testing of machine learning algorithms. To this end, I will conduct a greedy feature selection based on Logistic Regression coefficients obtained from 3-fold cross validation, which provides a subset of features that are maximally predictive of the target variable plays (see notebook plays_greedy_feature_select.ipynb).