 # NFL Plays Analytics - Milestone Report 1

Predicting offensive football plays

The NFL is the most profitable major sports league in the US with over $12 billion in revenue last season. However, NFL **coaches and general managers** are woefully behind many other leagues in their utilization of advanced analytics to **enhance in-game decisions** and player acquisitions. With this project I would seek to give teams a leg up on the competition by building a model to **predict the opponent's next offensive play**, or at least narrow down the possibilities. Thus, coaches of NFL teams are the directed client. With these results computed historically and in real-time, **coaches** could call better defensive plays, adjust their personnel, and improve their strategy - giving them a greater likelihood of hindering the opponent's success.

The main open-source **dataset** can be found here, nfldb; it is a Postgresql relational database containing play-by-play, game scores, player stats, schedules, and much more from the 2009-2016 seasons and playoffs. There are 8 tables in total in the database. The database includes a small module to retrieve in-game data in real-time via the NFL gamecenter API, as well. Additionally, I'll seek to merge this play-by-play dataset, dating back to 2002 providing an additional 7 NFL seasons.

Ideally, I would design a **mutliclass classification model** to predict whether the next offensive play will be a run left, run middle, run right, pass left, pass middle, or pass right. This could easily be converted to binary classification: run vs pass, if necessary. Admittedly, such modeling is a challenging task with many contributing factors, including but not limited to time left on the game clock, down and distance, game score, coaching staff turnover year to year, player personnel changes game to game due to injury or trades, and fluctuations in defensive tendencies.

 # Data Wrangling (see plays_wrangle.ipynb)

I explored the various tables with SQL queries to better understand the scope of the data available, and determined that I would begin wrangling data from 3 of the tables - plays, games, and drives table which shared game and drive keys useful for merging later. For initial programming and processing ease, I focused on the 2016 Regular season data for every team. To obtain labels for play type and direction, I searched each plays description for pass vs rush and left vs middle vs right, using string manipulation via regular expressions. This wrangling was used to produce my target variable. In the process, several plays and play types with insufficient data or outliers

were excluded, including kicks, punts, two-point conversions, and quarterback scrambles.

Score tuples, game clock quarter and time, and variety of other categorical variables with string representations were split converted to integers, category codes, and/or time floats as necessary. I shifted statistics forward for each play to prevent leaking future information into the model for a given play. Then, I accumulated these stats separately across drive possessions (collections of plays) and games for each team and game. I zero-filled the resulting missing data for the beginning of drives and forward-filled stats for accumulated game stats.

Following processing plays, I merged this dataframe with general game and drive data that was queried from the database. These merges provided additional data about home vs away teams, day of the week, week of the season, and summary stats from previous play drives. With the current wrangling setup, ~32,000 plays and ~200 features emerged which extrapolates to ~320,000 usable plays across the entire database. The features are likely to grow too, as I incorporate player-specific stats and other potentially important indicator variables.

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

After wrangling the NFL plays data into a suitable form and exploring patterns in the data (see plays_explore_plots.ipynb), 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. Accordingly, 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 will be addressed in future analyses through feature selection.

Future Analyses

The observed 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 conducted a greedy feature selection based on step-wise 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). This analysis revealed that a higher number of features provide meaningful, predictive information when logistic regression is first trained with a binary play type target: pass vs. rush, followed by a separate training with a play direction target, left vs. middle vs. right. Accordingly, I will split the play target into two separate targets: play type and play direction. Future analyses will focus on first predicting pass vs. rush then direction in sequential implementations

In these future explorations, I will train and test a variety of algorithms covering naive bayes, random forests, SVM, gradient boosting, and deep learning neural networks, using 10-fold cross validation on the most recent season of the data. The sequential nature of the plays is extremely important. Thus, I aim to test an RNN with LSTM and an attentional component to not only capture the most recent plays but also plays from previous drives and games in similar situations. Resulting feature importances and model performances will guide further feature engineering and model selection and refinement, which will be trained and tested on incrementally more data.

As a result of this project, I will provide a collection of python code scripts, informative visualizations of the features and model performance, a paper/blog describing the rationale and its effectiveness, and an automated model to predict NFL teams' next offensive play in real-time.