Zach Austin and Jimmy Brunette

ITAO 40420

9.17.2021

## Classifying the Problem

Adapting to an opponent's next play in a football game is a crucial piece of strategy that can give a team a significant edge on defense. However, some teams have introduced apparently random schemes, audibles, play signals, and other play calling methodologies that deviate from typical football norms; play calling styles have become more erratic and less predictable for an opponent. The goal of this analysis is to present information from models correlating several in-game measures of an offense and their efficiency to the play call outcome of either run or pass. To begin, a simplistic approach will be taken by forming a model for just one college football team's play calling within one game of the season; complexity will then grow after understanding model characteristics, intentions, and output. In order to explore these relationships, we will utilize extensive data on play by play college football information. The initial data has myriad variables included; not only are many of the variables seemingly irrelevant to the goal at hand, but some peripheral college football teams are not of interest, and plays characterized outside of the run or pass outcome largely will be ignored during analysis. Focus will mostly be attributed to variables that have the greatest reliability in the dataset as strong predictors with no presence of missing values, as these are the most easily obtained inputs that are entirely prepared to serve in a model. Several model approaches will be explored, including random foresting, bagging, and XGBoost, to test the approach's ability to learn on this data in a quick, concise, and effective manner. Potential bias may be introduced through variables that are not included in our dataset, such as head coach or opponent defensive players; however, familiarity with the underlying game that represents this data is strong, and biases will either be mitigated in pre-processing or modeling, or will be thoroughly explained in analyses.

## Describing the Dataset

Data on college football play by play results was obtained from the cfbfastR package, a community R package. The initial data had roughly 104,000 observations with 330 descriptive variables per observation. The sheer magnitude of plays in this dataset provides a strong opportunity for analysis on the qualities and metrics of interest. However, with the goal of predicting a binary run or pass outcome, many of the variables in this dataset are conceptually unrelated to the predicted outcome. Variables were therefore prescreened and eliminated based on conceptual relatability to the outcome, and the current form of the tidied dataset has roughly 60 variables. Examples of the anticipated strongest variables in the dataset are as follows:

| | |
|---|---|
| down | Notes if the team is in first, second, third, or fourth down. |
| distance | Distance to achieve a first down. |
| wp_before | Win probability of a team before any given play. |
| score_diff | Difference between scores of teams. |
| clock.minutes and clock.seconds | Time remaining in the game |
| Calculated column on run success throughout the game | Indicates yards per rush up until that point in the game. |
| Calculated column on pass success throughout the game | Two variables that indicate yards per pass and pass completion rate up until that point in the game. |
| goal_to_go | Notes if the team is within ten yards of the end zone. |
| drive_number | What number drive the team is within the game. |

Several other variables revolve around various yard/field, time, drive, and other play number measures that were figured to have some relevance to play choice.

## Project Plan

**Phase One:**

Objective: Data cleanup and preprocessing

Time Frame: Fully prepare data for modeling by Tuesday, September 21.

Description: This phase will involve all of the data transformation prior to building any sort of model on the data. This includes removing unnecessary and unrelated variables, shifting around columns and rows for convenience, filtering on teams that are of interest, calculating columns for new measures, and fixing outcomes to be more readable for a model. Approximately half of this phase is already completed.

**Phase Two:**

Objective: Implement model code and ensure reasonable run time and output for the information

Time Frame: Implement model and fine tune code by Tuesday, September 28th.

Description: This phase revolves around everything modeling: testing the various types of models that we can employ; testing their predictive strength on unseen data; and formatting a favorable output reading in R.

**Phase Three:**

Objective: Compile final report; check with Martin prior to submission.

Time Frame: Present draft to Martin by Friday, October 1st; finalize report by Wednesday, October 6th.

Description: This phase will involve compiling findings and visualizations into a report that outlines approaches, methodologies, and implications of results. To achieve feedback, the plan is to present a working draft to Martin at least 6 days prior to the due date, perhaps to receive feedback that can by incorporated into the final draft.