Group Members: Jacob Lindell and Patrick Mackintosh

Lt Col Gaupp

M3

Milestone 2

**1. Overview**

There are many factors that go into predicting the play an offensive team is going to call in the National Football League (NFL). Defensive coordinators have been seeing the significance of football analytics, and teams like the Broncos have started to use analysts to help with both offensive and defensive play calling (Wagner-McGough). Factors like yardage, down and opponent all could play a large role in the play a team will call, and while at one point it was left to a team’s staff to decide what defense to call, there could be value in an analysis of a individual play data over a season. The goal of this project is to predict the type of play the offense is going to run, and the location of that play, based on the current conditions of the team. With this data, we will be able to inform the Defensive Coordinator on what type of play the offense will most likely run.

**2.** **Data and Data Processing**

The data that we will be using was given by Major Pietz during the OR 495 class last semester for the purpose of determining if teams that run more than 20 times per game are more likely to win. The data includes all of the information collected during a given play; the offensive team, down, yardage, play choice (run or pass), play selected (run right, deep pass), player used, etc. We will need to organize all of the plays by team, and figure out a team’s strengths and weaknesses on both sides of the field in order to best select the next play that should be called. The data is given in Excel and is compatible with Stata. We have data from the 2013 season and 2104 season, but so far the analysis has focused on the earlier year’s data.

The type of data we have acquired is cross-sectional data. While there is a time variable present in our data, we do not see it significant to test how variables differ over time. We are looking at many individual subjects, in this case offensive plays from the regular NFL season, therefore not making time-series or panel data applicable.

The data has 33,398 rows, meaning 33,398 different offensive plays recorded taken from 32 teams playing 16 games each. This spans over 58 variables such as team, opponent, down and field location, and totals to 1,930,299 data points. However, exploring our data we have come to the conclusion that about 22 of the variables will most likely be explanatory in our model, which are listed in the Appendix A. The many variables can be cut down further with more exploration. Another 4 are related to the dependent variable, as they have to do with what the team’s play call is and where the ball is going.

Our data is very complete, especially for its size. Of the almost 2 million data points only 7,693 are not filled in, which is less than 1% of the data. However it appears the missing data is not absent, but represent an action not happening. For example, if there isn’t a fumble on the play there is no data point in the fumble column. This is relevant for other variables like interceptions, receiver, or pass location, which would not have values if the play wasn’t a pass. To deal with this problem we have created dummy variables with a zero when an action didn’t happen and a 1 when it did.

To prepare our data for Stata we initially took out the variables which were not seen as viable in our model. The variables taken out were either seen as unrelated, such as date, or not relevant. Our model is going to predict the outcome of the play so variables like yards gained or the intended receiver aren’t applicable as they happen once the play has already happened, meaning they are dependent. Our model is going to predict the outcome of the play based on what the defense sees before the play, making those variable unusable. Once the data was taken down to the 29 variables important variables, it was imported into Stata. Then to take care of each qualitative measure, for example team name, a dummy variable will be crafted.

**3.** **Preliminary Analysis**

Seeing as we are predicting binary outcomes we are going to use a logit/probit model, depending on which model predicts more accurately. Also as our data set is larger than anything we have worked with before the tidying and preparing the data for analysis has been the primary concern and taken the bulk of our time.

Looking at our data further we have tweaked our objective. The way the data is structured it would be best to predict the type of play the offense is going to run given their current conditions such as yardage, down, and team, rather than to choice the next play as we stated in milestone 1. It is going to be very valuable for the Defensive Coordinator to have an analytical idea of what the offense will call next, and choose the defense’s play accordingly.

The next two steps in analyzing our data would be to determine if there is any heteroscedasticity throughout our variables, and then to determine which variables would be most significant to include. There are too many variables to feasibly check each for heteroscedasticity by hand so we need to research some tools to make sure the data is sound. On top of that, since this data is so large, there is some free software such as WEKA or RStudio to expedite the process of checking which variables are significant. Once we have more information we can start tackling problems like omitted variable bias if they are present.

Our model has some significant assumptions; the biggest being that the play location predicted is dependent on the current conditions and has nothing to do with past plays. This is not ideal as it is fair to assume a team that has passed three times in a row most likely will run, but the assumption is to keep the project within the scope of our abilities. However, it is fair to assume that variables like down or yardage can do a fairly good job accounting for how past plays can impact the current situation.

Finally, looking at further issues, we have discussed that the most significant issue is the size of the data set. This will make it difficult to determine significance and heteroscedasticity. Also categorical variables are rampant throughout the data so many dummy variables will have to be introduced to use within our model. Some variables, such as down, are significant when looking at it as a quantitative number, but deciding whether to classify it as qualitative or not is also an issue present. One final issue that could be present is multicollinearity. Many of our variables are related to each other, and while Stata is able to throw out the ones with perfect collinearity, we have to remain vigilant in understanding our variables. It is very reasonable that many of our factors, like Touch Down count and score, could be too related to both be included in the model even though score could also include field goals.

The next step in our analysis is to determine which variables can do a good job predicting play calls. We have initially found down, To Go, team score, opponent score, Time under, yard line, touchdown count, play percent of goal, first down count, goal to go, success count, absolute score differential to be significant in terms of what the offensive play will be, and there metrics can be seen in Appendix B. Next steps would be to explore the rest of our explanatory variables and to use that to create a concise model to predict play calls. Our model will most likely be iterative, first having a prediction on play type(run or pass) and then play location, so further research will be taken on how that can be accomplished. We then can, by taking our model and predicting on other data, say the 2014 season data, calculate the accuracy of our model and determine its validity.

Works Cited

Burke, Brian. "GB-SEA Analysis." *Advanced Football Analytics*. N.p., 19 Jan. 2015. Web.

Wagner-McGough, Sean. "Broncos' Director of Analytics Will Help Gary Kubiak with Game Decisions." *CBSSports.com*. N.p., n.d. Web. 27 Aug. 2015.

**Appendix**

1. **Variable Present in Data**



\*Highlighted variables were originally chosen by us to include in our analysis, the bottom row represents the outcomes we wish to predict

1. **Initial Logit Regression**



Documentation

None