Nathan Marzion

Deliverable One

COSC 5610

8 October 2020

The data set that I am using is a play-by-play data set for the 2020 NFL season, which is updated after each game. It has every play from every game, and each play is recorded with a description of what happened along with variables for many different things and almost any stat possible. These include what down it is (1st, 2nd, 3rd, 4th), what type of play was performed (run or pass), how many yards were gained, how many yards a pass traveled in the air, how many yards were gained after the catch, whether or not the pass was complete, which players were involved in the play, where the play took place on the field (what yard line), and much more. A stat that I am going to be using a lot in my analysis is EPA, which stands for Estimated Points Added. I provide a link to a description for EPA in my reference list. Essentially, the stat measures how successful a play is, and it takes into account the context of the play and the expectation of how successful the play should be before it happens. Each play that has a positive EPA is considered a "successful" play for the offense. Thus, there is a "success" column in the data that has either a 0 or 1 value, depending on if the EPA of the play is positive or not (positive = 1, not = 0). This is where I can do the bulk of my analysis, as I can look at many factors and what seems to be significant in creating a "successful" play. There is also information about how drives end (field goal, touchdown, no points) which will be helpful in analyzing what creates a scoring drive.

There are many types of questions I can then ask that center around what contributes to success. Does passing more often seem to be a significant regressor for team success? How do the chances of being successful change based on play type? Or based on down and distance? Does making deep, difficult passes seem to have a bigger impact on success than attempting short and easy passes, or is it the opposite? How does a team's overall chances of success change based on 1st down play calling, and is that significant? How much more likely is a team to be successful when they are throwing the ball in the red zone compared to running? What happens to the chances of success when a quarterback throws the ball past the first down line compared to short of the first down line? How much do the chances of success change based on which players are involved? How much do the chances of scoring on a drive change based on the team's initial starting field position? I can also look at what contributes to passes being completed, since there is a categorical variable for whether a pass is completed on the play or not (yes = 1, no = 0). How much do the chances of completing a pass change based on distance of the pass? How much does it change when a certain quarterback throws to a certain receiver? Which players seem to be significantly increasing the chances of a completed pass? Does a completed pass always result in

a “successful” play? These are just some examples of many types of questions that I can ask to evaluate significant factors that create success.

The original data set has about 300+ columns and 7000+ rows currently (through week 4 of the 2020 season), which is far too big and messy to work with. I will be filtering it down to only run and pass plays (the plays that I actually want to analyze) to get rid of kickoffs, extra points, punts, and other plays that are not relevant. I will also eliminate many of the columns that I don’t need and filter it down to approximately 20-25 columns at most. I will mainly just keep the columns that I mentioned previously (description, play type, down, distance, location, players involved offensively, EPA, success, yards gained, air yards, yards after the catch, whether the pass was complete). This smaller, condensed version of the data set will be a lot easier to work with. In references, I also provide a link to a description for work that is done with these play-by-play data sets to further illustrate the types of variables that are included in the data set and analysis that can be done. There is also an R package associated with this data that I provide a link to. It updates the play-by-play repository with clean and tidy data sets, and provides models for some of the stats.

References

Alok Pattani. 2012. Expected points and EPA explained. (September 2012). Retrieved October 7, 2020 from https://www.espn.com/nfl/story/_/id/8379024/nfl-explaining-expected-points-metric

Ben Baldwin. A beginner's guide to nflfastR. Retrieved October 8, 2020 from https://mrcaseb.github.io/nflfastR/articles/beginners_guide.html

Ben Baldwin. An R package to quickly obtain clean and tidy NFL play by play data. Retrieved October 8, 2020 from https://mrcaseb.github.io/nflfastR/

guga31bb. 2020. guga31bb/nflfastR-data. (October 2020). Retrieved October 3, 2020 from https://github.com/guga31bb/nflfastR-data/blob/master/data/play_by_play_2020.zip