The Efficacy of NFL Defenses Against Passing Plays

Mohamad Quteifan, Data Scientist, M.S. in Data Science

Kayla Thompson, Data Scientist, M.S. in Data Science

Gloria Moore, Data Scientist, M.S. in Data Science

Bellevue University

Authors

Mohamad Quteifan, Department of Science and Technology, Bellevue University

Kayla Thompson, Department of Science and Technology, Bellevue University

Gloria Moore, Department of Science and Technology, Bellevue University

Brett Werner, Department of Science and Technology, Bellevue University

**Executive Summary**

Global Sport Market has increased its total revenues in about 35% in the last 15 years. Total revenue of all 32 National Football League (NFL) teams has risen from about 4 billion U.S. dollars in 2001 to over 15 billion U.S. dollars in 2019, the highest figure to date [1]. The uses of data analysis and statistics in sports helps coaches, players, fans, advertising industry, etc., it helps not only to win games but to improve players performances, prevent injuries, fun for fans, and so on.

The NFL posted a set of data from past games to find relationships between variables and identify patterns to help to predict games outcomes. This project attempt to generate a model that use selected features to predict the result of a play, identified by yard gaining or no yard gaining. Using 3 different classification models, logistic regression, K-nearest neighbor, and Random Forest we will predict the outcome of a play, in our study the selected model was Random Forest Classifier based on accuracy, recall and f1 results. Getting a prediction accuracy of 0.8938 the 50 selected features can give us the prediction of yards gained or no yards gained.

**Introduction**

In 2002 Oakland Athletics baseball team made people realize the serious effect the use of data analytics could have on the success of a team, the first team in NFL to lead the data analytics technique was the Philadelphia Eagles. Beginning in 2014, Eagles head coach Doug Pederson made it clear that all decisions made by the organization were going to be informed by analytics. Ryan Paganetti started in the Eagles’ analytics department in 2014 [2]. Legendary football coach Paul “Bear” Bryant famously said, “Offense sells tickets. Defense wins championships”, if this is true, predicting what is going to be the yards gained using a specific defense scheme and other variables available for the defense coach, would be a great advantage for the team.

Since the inception of the National Football League (NFL) in 1920, defensive coordinators have been aggressively seeking any advantage over opposing offenses. The desire for an advantage has led to numerous reports of cheating, none more notable than deflate-gate. The controversial event gave the New England Patriots defense a huge advantage, given they deflated the footballs during the practices leading up to the American Football Conference (AFC) Championship Game against the Indianapolis Colts in 2014 [3]. The New England Patriots were disciplined by the NFL resulting in the case going to the Supreme Court. The Patriots were found guilty and as a result, lost two draft selections in the 2016 NFL draft, fined $1 Million, and Tom Brady, the Quarterback, was suspended for four games for his involvement in the scandal [3]. Despite this scandal, there remains a need for defensive coordinators to have an ethical means of gaining an advantage.

The NFL team dedicates their time to set defenses strategies according with a list of possible offense plays, based on this concept, prediction of 0-yard gains taking in consideration features like, type of play, yards to gain, offense formation, and others, could give the coaches an insight of how to play the next game.

In the following work we will use Exploratory Data Analysis to better understanding of our data set, we calculate correlation to determine possible relationship of the variables, and the last section will be dedicated to generation of the model options and model selection, to present to any defense coach that will want to stablish prediction of the results of a play based on the selected variables. The data and the techniques that will be implemented to clean the data will be outlined along with the steps we will use to build the model. We will discuss how we intend to evaluate our model’s efficacy and any risks associated with this model. As this is just a preliminary proposal, all of this is subject to change based on the results we observe through our work.

**Problem Statement**

With this project, we aim to create a model to aid defensive coordinators in their game planning. The proposed project will give the defensive coordinators an advantage by using data science modeling techniques to provide the defensive coordinators insights into which of their schemes will best defend against their opponent’s offense. This method will not only be more ethical but also will ultimately prove more effective than deflate-gate.

**Preliminary Requirement**

In this study we will be looking at the data from the Kaggle NFL challenge [4] – specifically the plays data file, which is stored in a Comma Separated Value (CSV) format. Our first step will be cleaning up the column names and removing columns that provide unnecessary information. We will be doing this to simplify the data and will keep only the columns that can potentially be used for our model. For example, we will be removing the “playDescription” column because that information is represented elsewhere in a format better suited to be input into a model. The goal of the model is to determine the outcome of a play based on several factors to ultimately determine which defensive schemes are best suited to defend against offensive plays. Two particularly important columns we need to use to solve this problem will be the “personnelO” and “personnelD” columns. These columns, however, are not in a suitable format for analysis. Thus, we will be creating a code that will turn each combination of personnel for each column into a single digit entry. For example, the first entry in “personnelD” will be encoded as 1 which will translate to meaning four defensive linemen, two linebackers and five defensive backs. Once this is completed, our data will be considered clean enough to begin analysis. The columns that remain are described in the next section.

**Variable Description**

The data file we are working with contains many, many variables. In this section, we will give a brief description of each variable and the types of changes that will be made to each one.

* PlayId: The identification of the play.
  + Type of variable: Numeric
  + Changes:
    - Name Change: playid -to- play.
    - Notes: Other changes have not been determined yet.
* YardsToGo: The number of yards to go to get a first down.
  + Type of variable: Numeric
  + Changes:
    - Name Change: yardsToGo -to- yardstogo (lowercase).
    - Notes: Other possible changes to the variable have not been determined yet.
* OffenseFormation: The formation of the offense. This is the formation that the offense is in.
  + Type of variable: Categorical
  + Changes:
    - Name change: OffenseFormation -to- Oformation
    - Type change: Changing to numeric variable to represent each offensive formation.
    - Notes: The variable will more than likely be removed from the data and not used for the analysis.
* PersonnelO: The amount of skill players and their respectable positions.
  + Type of variable: Numeric and Text
  + Changes:
    - Name change: PersonnelO –to- Otype
    - Type change: changing it to be completely numeric and having numeric values represent the types of skill players that are in the play. 1 would represent a 2-1-2 formation (2 Running backs, 1 Tight Ends and 2 Wide Receivers).
    - Notes: Not set on completely going with a single value representing the offensive personnel on the field, might go with 212 to represent the personnel. Once EDA is conducted we will know the best method to implement.
* DefindersInTheBox: The defenders close or near the line of scrimmage. The more players near the line of scrimmage the more likely that the defense is running a blitz package. It could also be a fake blitz package but that is not our concern at the moment.
  + Type of variable: Numeric
  + Changes:
    - Name change: DefendersInTheBox —to— DBox
    - Type change: None
    - Notes: Including the defenders in the box will enhance the model by allowing for the analysis of blitz packages.
* NumberOfPassRushers: The number of rushing defenders. 5 defenders blitzing or more is considered a blitz.
  + Type of variable: Numeric
  + Changes:
    - Name change: numberOfPassRushers —to— Rushers
    - Type change: None
    - Notes: There needs to be more analysis before deciding on including rushers variable in the model.
* PersonnelD: The defensive personnel, the number of the defenders and their respectable positions.
  + Type of variable: Numeric and Text
  + Changes:
    - Name change: PersonnelD —to— Dtype
    - Type change: This is another one that is going to require a lot of attention. The changes are going to be similar to PersonnelO, where the defense personnel is going to be represented by a singular value.
    - Notes: There are a few more analysis to go through before making the decision on the type of changes to make for a more effective implementation.
* PlayResult: The result of the play, essentially the yards gained on the play.
  + Type of variable: Numeric
  + Changes:
    - Name change: PlayResult —to— Result
    - Type change: None
    - Notes: The variable is one of the most important in determine the success of the play. If a play results in anything less than 10 it is deemed successful for the defense and anything more than 10 yards is deemed successful for the offense. We are analyzing to determine what would be considered a successful or unsuccessful defensive play.

**Technical Approach**

The research for the study will be done within Juypter Interactive Notebooks, an integrated development environment (IDE) for Python programming language, leaving very few technical requirements beyond what we will accomplish inside of the notebook. Before beginning our analysis, it is important to split our data into a test set and a training set so that we can ensure we are not data snooping or tainting our work in any way. To begin our analysis, we will begin with some basic exploratory data analysis. The goal of this analysis will be to learn what the main characteristics of our data set it. We will use several visualizations and descriptive statistics to determine what our data set is comprised of. After this, we are anticipating that we will likely need to balance our data set. If this proves true, we plan to utilize Synthetic Minority Oversampling Technique (SMOTE) to balance our data set. SMOTE will oversample our minority class in our data by creating synthetic observations based on characteristics of our actual data observations [5].

Once this is complete, we can proceed with our feature selection. The main method we plan to use for feature selection will be a calculation of the correlation between each variable and our desired output variable. We plan to use the Pandas library to calculate these correlations and produce a correlation visualization. We intentionally kept some columns in our data set that we do not intend to use but will ultimately determine their fate through our feature selection. The idea behind this is letting the data analysis speak for itself while also ensuring that every decision is backed by the results of our project.

Moving forward, the features we select will be used to build our model. Several models will be created to move forward with the model that best accomplishes the goal of the project. We plan to build several models including logistic regression, random forest and k-nearest neighbors. We will create several iterations of each model to ensure we are selecting the best model parameters to set up each model to solve the problem as accurately as possible. Once these are built, we will evaluate these models using a confusion matrix and ultimately this will allow us to decide what model is best suited to aid in solving the problem statement.

**Expected Results**

We expect that the result of our project will be a model that will indicate the result of a play based on the offensive and defensive schemes. For this project, we consider a defensive scheme successful if the yards gained by the offense is zero or a negative value. Our main objective is that this tool can be used by a defensive coordinator for an NFL team to run several defensive schemes against the same offensive scheme to determine which is most effective for that offense. The coordinator could then repeat this process to have a set defensive scheme for any offensive scheme they may encounter during a game. This work will be completed over twelve weeks leading up to a final model.

There is minimal risk associated with this project, as any negative impacts of using our model could result in the loss of a game but this would not result in any injury or other harm to any other person. Risk of harm with use of the model does not increase compared to without use of this model.

**References**:

[1] Statista. (2020). *Total revenue of all National Football League teams from 2001 to 2019* <https://www.statista.com/statistics/193457/total-league-revenue-of-the-nfl-since-2005/>

[2] Aubrey, J. (2020, June 9). *The Future of NFL Data Analytics* <https://www.samford.edu/sports-analytics/fans/2020/The-Future-of-NFL-Data-Analytics>)

[3] Loyola, K. (2020, September 16). *The true story behind Tom Brady and the Deflategate scandal*. Bolavip US. <https://us.bolavip.com/nfl/the-true-story-behind-tom-brady-and-the-deflategate-scandal-20200915-0014.html>

[4] NFL Big Data Bowl 2021. (n.d.). Retrieved December 12, 2020, from <https://www.kaggle.com/c/nfl-big-data-bowl-2021/rules>

[5] Brownlee, J. (2020, August 20). SMOTE for Imbalanced Classification with Python. Retrieved December 12, 2020, from <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>