The Efficacy of NFL Defenses Against Passing Plays

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 with XGBoost optimizer , 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 f1 score on test of 0.90. With this model, defense coaches can introduce their data and get predictions that would be accurate of the result of play.

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, based on yards gained or no yards gained on a play. This method will not only be more ethical but also will ultimately prove more effective than deflate-gate.

Method

The project has been developed following the CRISP-DM process with the exception of the deployment stage, first part already presented, Business Understanding was made using NFL data science prior usage information, understanding technicality of the game, and other resources of research to understand our data and problem. Data understanding and processing stage started on the data obtained from the Kaggle NFL challenge [4] – specifically the plays data file, which is stored in a Comma Separated Value (CSV) format.

First step after collection of the data, included, 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 the model.

Feature engineering was made creating new categories for PersonnelO and PersonnelD, eliminating the categories considered special plays, these ones are categories with less than 10 observations. Observations with missing values were dropped to not affect posterior analysis.

EDA was made on the selected variables to identify patterns and relationships, histogram of our target variable:

Following graph shows the distribution of the outcome variable gotten after classifying negative to 0 yards gained on the "playResult" column with value 0, and 1 and up yards gained as value 1

From the EDA analysis on our data understanding and cleaning stage on the CRISP-DM process, the key points learned:

- No Running Plays
- Shotgun most common play
- Most common offensive formation: 1 RB, 1 TE, 3 WR, 2nd most common: 1 RB, 2 TE, 2 WR 4
- There usually 6-7 defenders in the box but this varies -- 4 to 8 is rather consistent.
- Usually, 4 defenders rush the passer
- Most common defensive formation: 4 DL, 2 LB, 5 DB, with 3 DL, 3 LB, 5 DB as the runner up defense formations deviate more than offensive formations
- Defense usually wins the battle with the offense on most plays, the most

common play result ==0 2.

- More completed passes than incomplete.
- Net Yards Gained on complete passes are between 1 and 15yards

From the correlation analysis not evident relationship were identify, higher correlation with our outcome variable was “epa”

Features used in the models are:

Modeling and evaluation stage was done using 3 models with our data, Logisitc Regression with XGBoost to optimize the model, K-Nearest Neighbor and Random Forest, these models were selected based on the classifying nature of our problem. Test set and Validation set were define with a proportion of 70/30. Random Forest was the model selected, evaluation and selection of the model was made based on F1 scores

Results

The results obtained from our models were: Logistic regression with accuracy of f1 score of 0.90, applying XGBoost to the Logistic Model the f1 score improves to 0.9013, Random Forest accuracy: 0.89 an f1 of 0.905. Selecting Random Forest model option based on F1 score and accuracy, the results as follow:

Confusion Matrix of our Selected Model:

The Logistic Regression model with XGBoost identified the most important predictors as, epa, yardlineNumber, yardstogo, quarter and down, these variables are always available for coaches on a game, these variables were use in all the models, and coach can use them when they need the predictions, personnelO and personnelD values were used as well in the model, and this information is also available to coaches on the team. The model selected will give predictions of play result based on yards gained or no yards gained with approximately 0.90 accuracy

Conclusions

Using of data science on sports is more and more popular, teams need to know what is the next effective move to do when the face an opponent, the previous work was made with the purpose of offer an option to coaches to know what would be the result of the next move, using features that are available before, during and after a game, the Random Forest model created will present predictions with accurate values, giving the extra help the coaches needs to identify if their strategies will allow the opponent to gain yards or not.

Logistic regression model optimized with XGBoost randomized search gave good results with a f1 of 0.901 better than the simple logistic regression model, just a few points bellow the Random Forest, what is important to note in this model, is that the logistic regression with this optimizer is a fast model to run.

Teams and coaches can be used the model proposed in any time and against any team that is next to play with, the model proposed does not take in consideration specific defense team or offense team, as well as game location or play identification, what make the model very general and versatile to apply for any team and/or game.

Acknowledgments

We would like to our professor for his big help on the feature engineering section and for his

comments on clarifying the predictions to be made

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/

APENDIX

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.