

A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is light green. They are positioned diagonally, with the blue one partially covering the green one.

# Predictive Analytics for NFL Defenses

Mohamad Quteifan  
Kayla Thompson  
Gloria Prada Moore

# Exploratory Data Analysis

Questions:

1. Introduction to Research
2. Data Frame analysis
3. Feature analysis
4. Challenges of the Research
5. Questions about the data -- Research



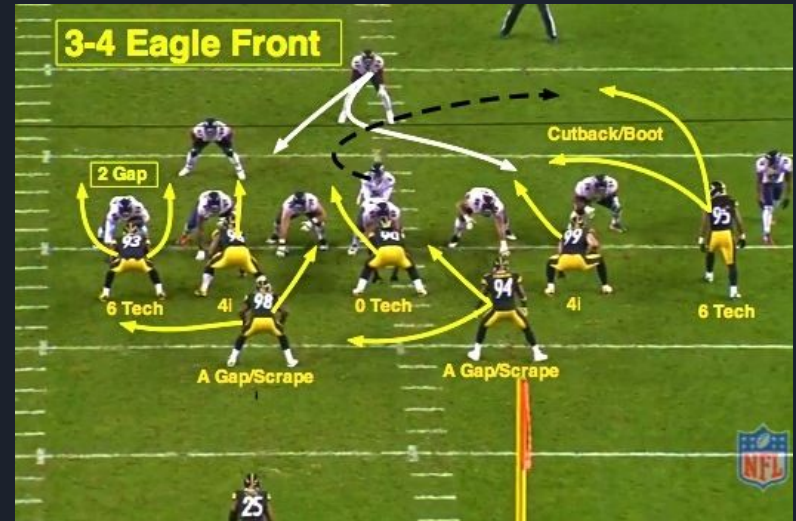
# Introduction-- Beat the Offense before the play

Purpose -- How great would it be to be able to effectively defend the offense before the play even starts. Realistically speaking that would require a lot of last second decision making based on the offensive formation. The implementation of data science could enhance the process!

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.

Model Concept:

- Create a model that can effectively determine the best defensive formation based on the offensive formation.





# Data Frame Analysis

Data: Plays.csv found on the keggel website.

The data frame consists of 27 features and 19239 rows.

Features: 'gameId', 'playId', 'playDescription', 'quarter', 'down', 'yardsToGo', 'possessionTeam', 'playType', 'yardlineSide', 'yardlineNumber', 'offenseFormation', 'personnelO', 'defendersInTheBox', 'numberOfPassRushers', 'personnelD', 'typeDropback', 'preSnapVisitorScore', 'preSnapHomeScore', 'gameClock', 'absoluteYardlineNumber', 'penaltyCodes', 'penaltyJerseyNumbers', 'passResult', 'offensePlayResult', 'playResult', 'epa', 'isDefensivePI'

The top features in the study:

1. PersonnelO: formation of the offense
2. PersonnelD: formation of the defense
3. PlayResult: Result of the passing play
4. DefendersInTheBox: amount of defenders near the line of scrimmage

\*\* The entire data frame consists only of passing plays !

# Data Frame Analysis -- Complications

The data consists of all passing plays. One thing that caused some confusion is “Unknown play type” but after conducting some EDA it was determined that the plays were all passing plays “unknown” just meant the play ended in a sack.

The good thing about the data is that we did not have to conduct any data imputing -- there were a limited number of missing values and most the missing values were from features that are not going to be used.





# Data Frame Analysis --Key Takeaways

1. No Running Plays
2. Shotgun most common play
3. 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.
5. Usually 4 defenders rush the passer.
6. 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.
7. Defense usually wins the battle with the offense on most plays, the most common play result ==0 .

# Feature Analysis -- PersonnelO

The Offensive formation of the play

Most common formation: 1 RB, 1 TE, and 3 WR.

This is comes to no surprise to anyone-- this represents a shotgun formation and that is clearly the most common passing formation.

The issue with the variable is the lack of variation -- low distribution the 2nd most common is a distant second.

This will be one of the independent variables(X) for the models were running



# Feature Analysis-- PersonnelD

PersonnelD is the formation of the defense. The good thing about this feature is that there is high variation, unlike the PersonnelO.

The most common formation: 4 DL, 2 LB and 5 DBs. This makes sense because this is the default formation to guard against shotgun formation. A somewhat close second is 4 DL, 3 LB and 4 DBs. In my opinion this formation is used when its an obvious running down-- the offense is in shotgun but they're 3rd and 1.

This is the other independent variable in our model!



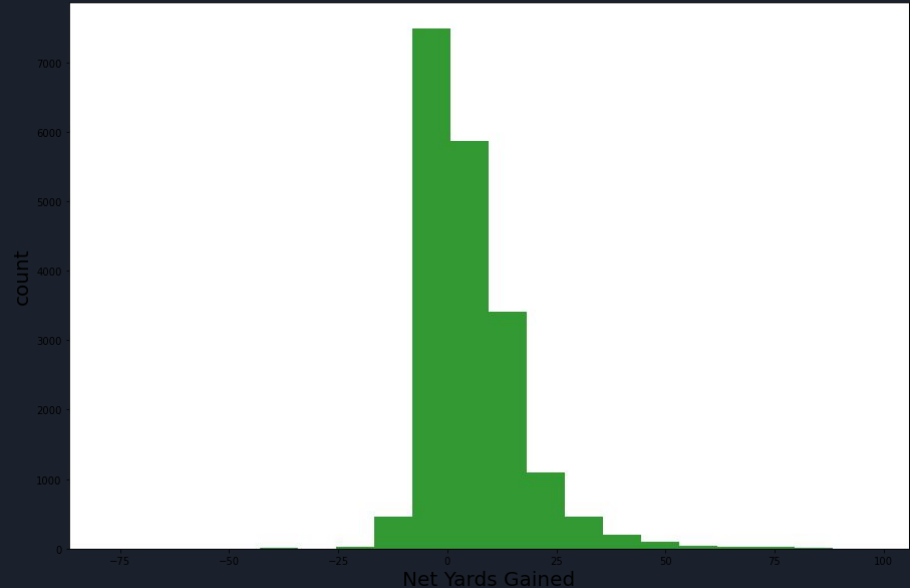



# Feature Analysis-- PlayResult

This is the result of the play. The most common play result is a net gain of around 6 yards. This means majority of the plays are positive plays for the offense and negative plays for the defense.

Any play that gains yards is considered a successful play in our model. This would be the dependent variable!

Play Result Histogram



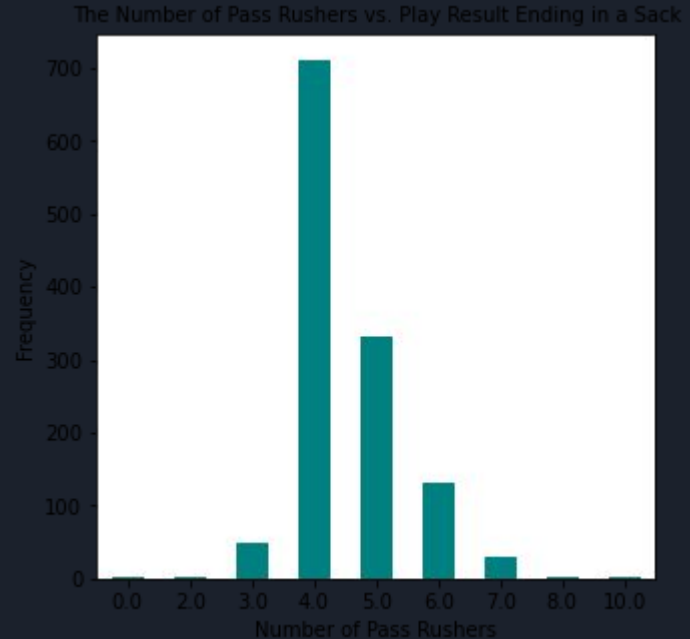



Do shotgun plays == 5 or more DBs defensive formation?  
Defenders in the box == more pass rushers? what about sacks?

The first question-- Yes they usually mean that the defense is set in a 5 DB formation.

Defenders in box == more pass rushers?

- Actually that's not true, many defenses put defenders near the line to trick the offense then go in a cover 3. So more defenders in the box does not mean blitz or sack. There are actually more sacks when a defense does not have more than 4 rushers!





# Challenges of the Research

The biggest challenge of the research is the conducting feature engineering. The implementation of feature engineer is challenging but it will enhance our model. The current plan is to implement feature engineering to the offense and defense (formation) features. This will lead to higher model effectiveness.



# One Hot Encoding Categorical Variables

- In order to run our models and complete our analysis we need to One Hot Encode our categorical variables.
- One Hot Encoding is where the categorical variable is removed and a new binary variable is added for each unique category within each variable.
- If an observation falls into the category a 1 is placed in that category - if not a 0 is placed.

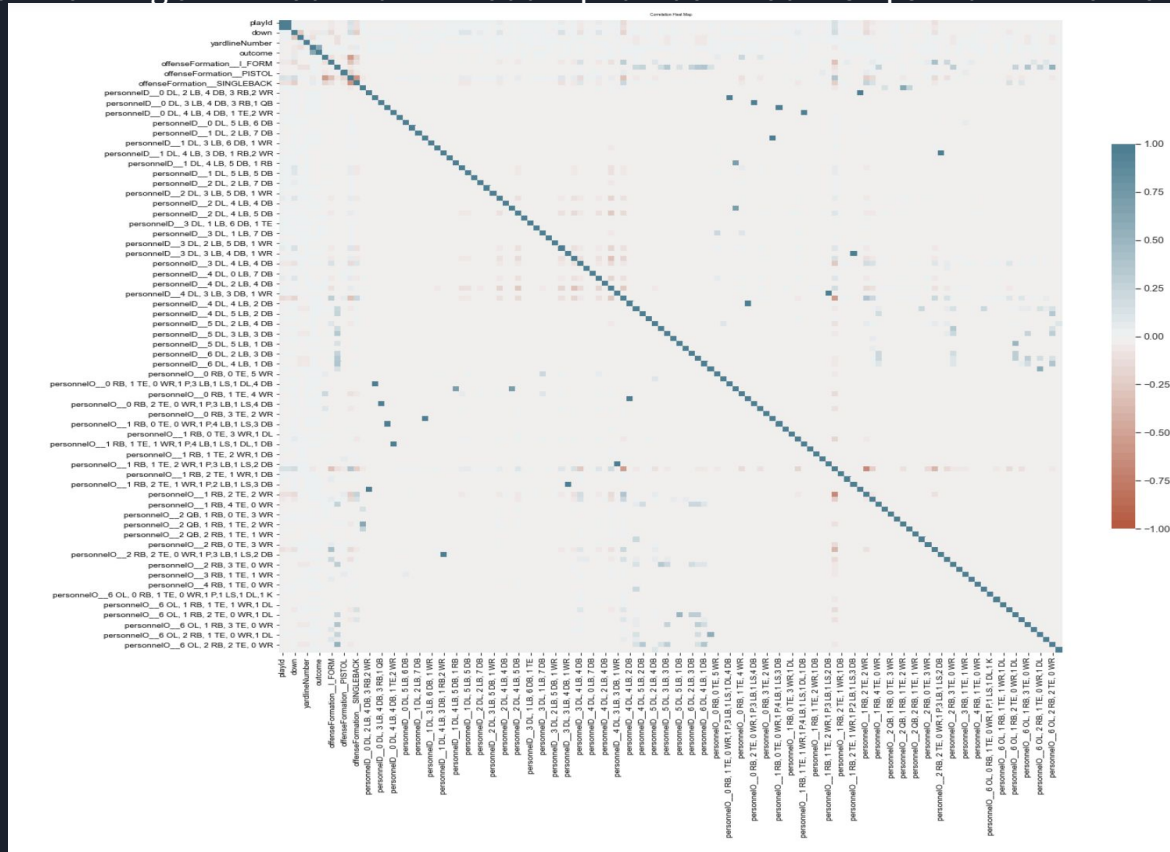


# Continuing Data Clean Up

- Next steps include:
  - Dropping Missing Values
  - Dropping Unnecessary Columns
- Used the dropna and drop methods in order to accomplish these tasks.

# Feature Selection - Correlation

We are running a correlation to look at the potential relationships between our variables.





# Feature Selection

- A Chi-Square test is used to finalize the features we move forward with.
- We used the Sklearn library to select the Chi features.
- This resulted in 6 selected features.
  - playId
  - down
  - yardsToGo
  - yardlineNumber
  - personnelD\_5 DL, 3 LB, 3 DB
  - personnelO\_2 RB, 3 TE, 0 WR
- **Based on this, we decided to move forward with our dataset as is. As our analysis continues to evolve we may choose to drop some columns from our dataset but this is a very preliminary result.**

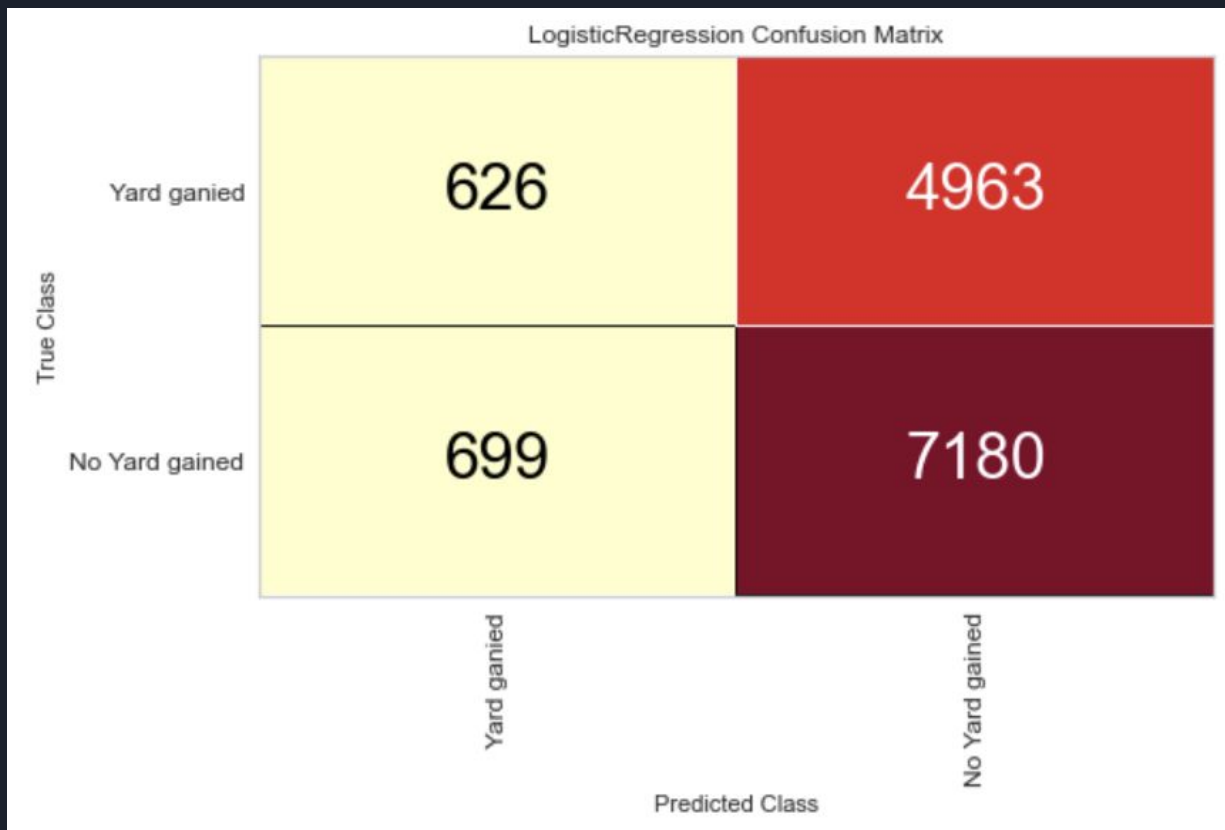


# Logistic Regression Model

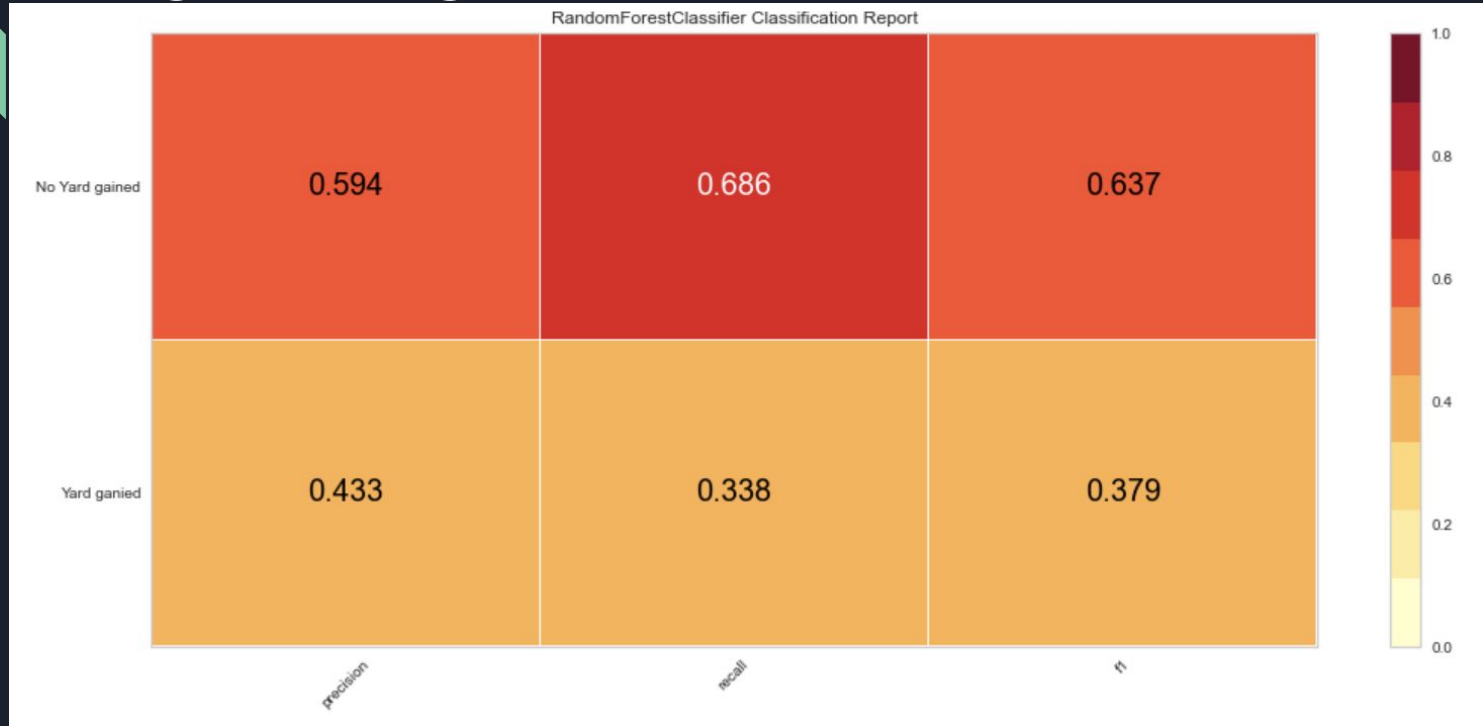
- The first model we tested was a logistic regression.
- We used the sklearn library to train this model.
- After training the model, a confusion matrix and a classification report is used to gauge the performance of the model.



# Logistic Regression Results



# Logistic Regression Results

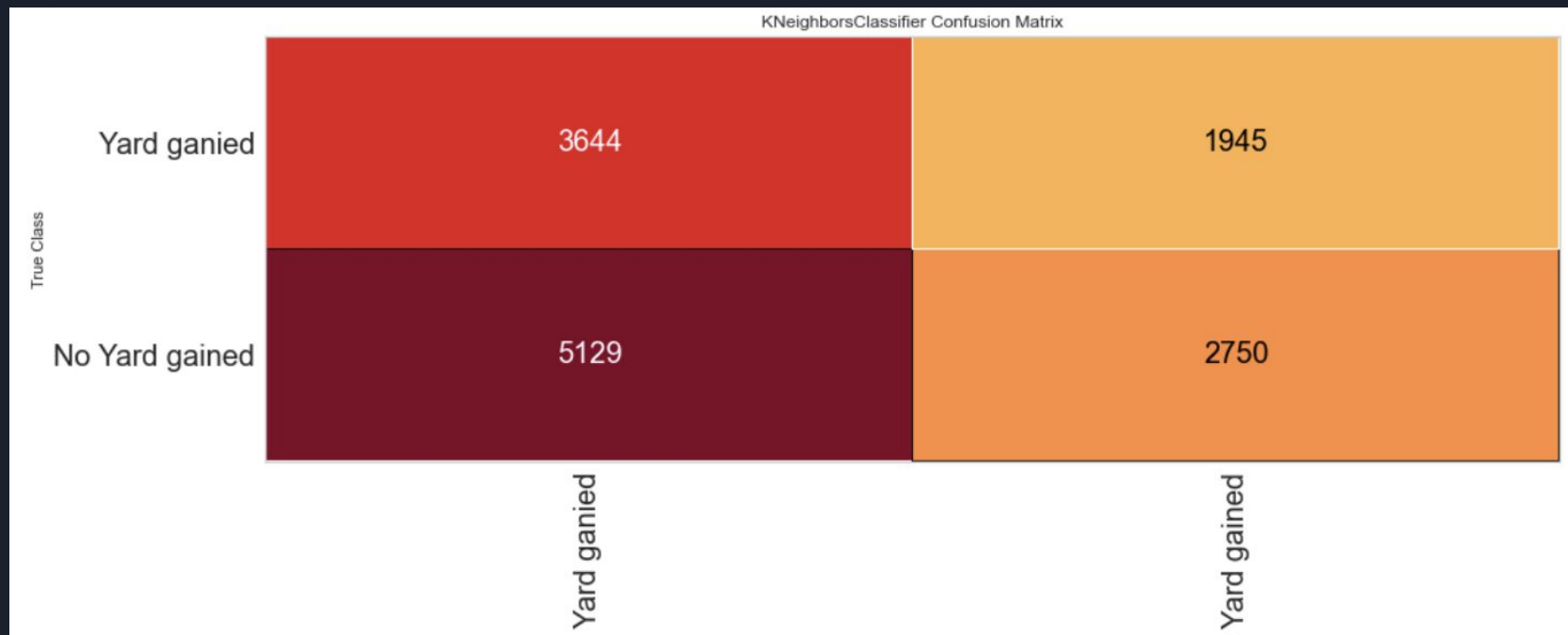




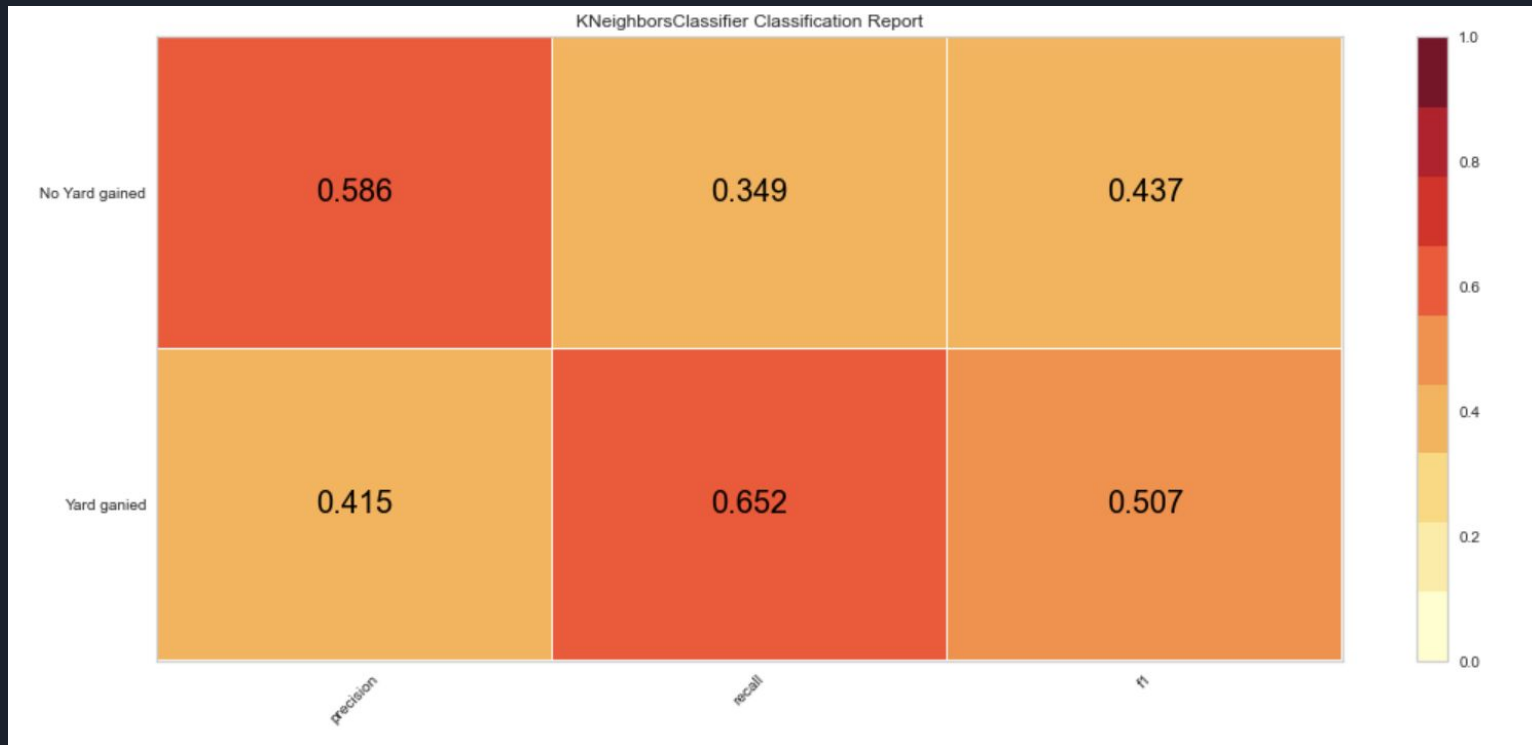
# K-Nearest Neighbors Classifier

- The second model we tested was a K-Nearest Neighbors Classifier.
- The sklearn library was used to train this model.
- After training the model, a confusion matrix and a classification report is used to gauge the performance of the model.

# K-Nearest Neighbors Results



# K-Nearest Neighbors Results

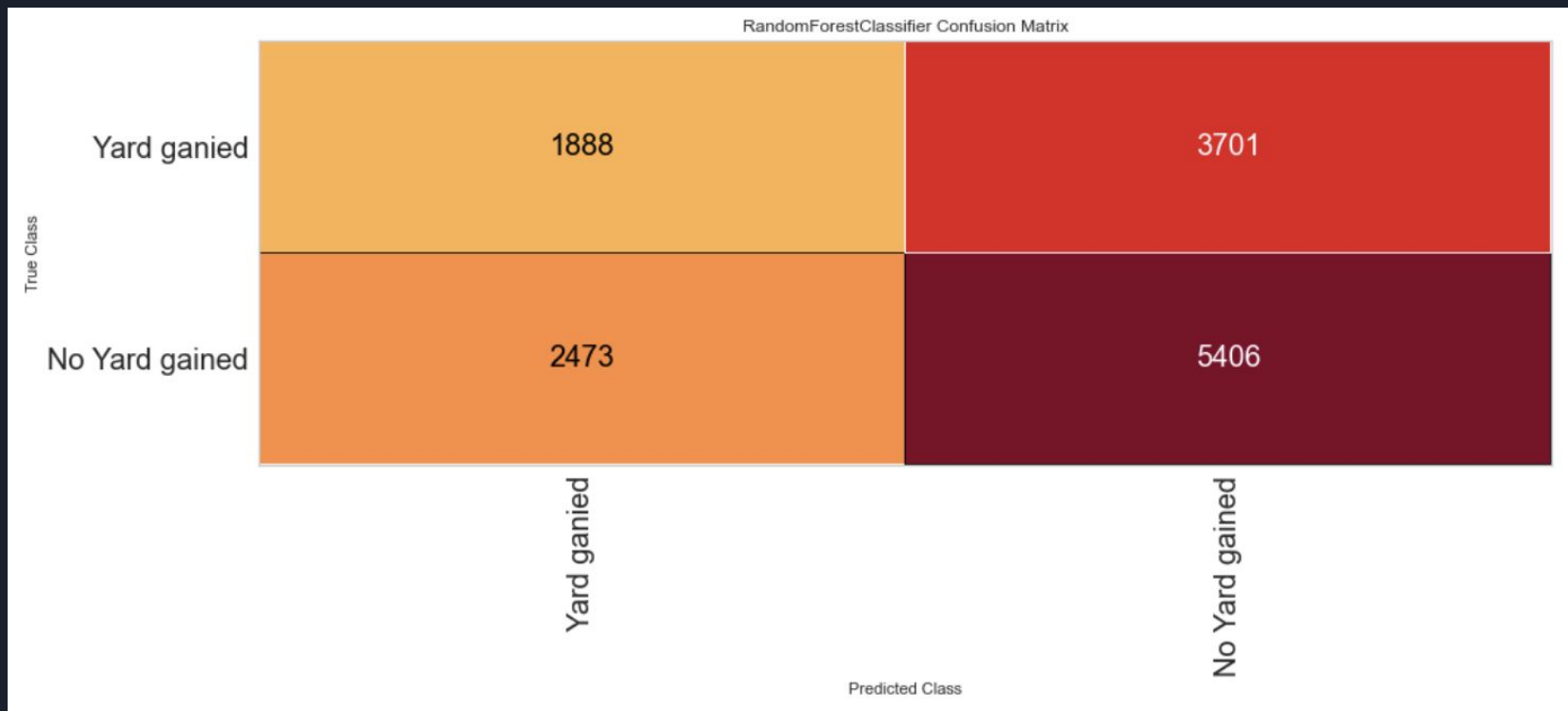




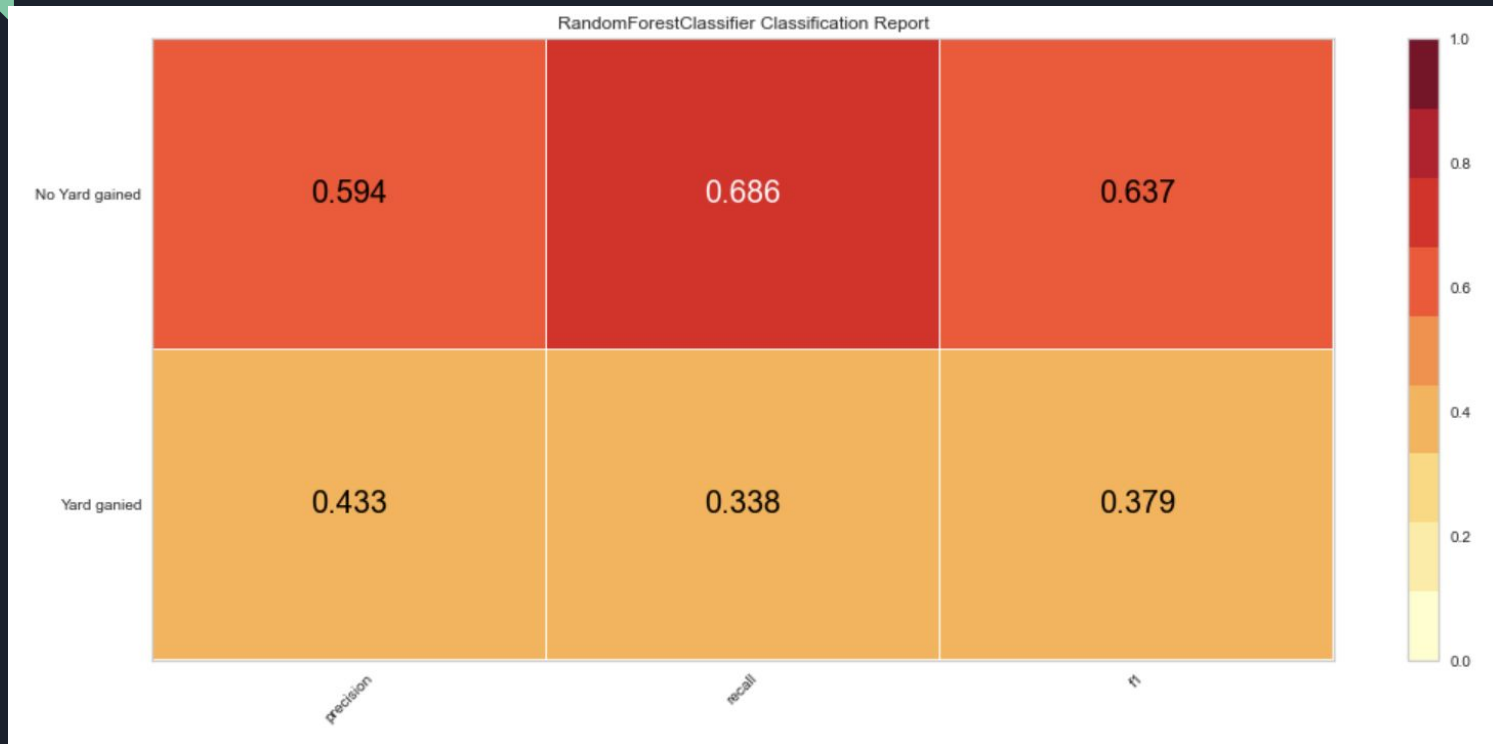
# Random Forest Classifier

- The third model we tested was a Random Forest Classifier
- The sklearn library was used to train this model.
- After training the model, a confusion matrix and a classification report is used to gauge the performance of the model.

# Random Forest Results



# Random Forest Results







# Challenges and Plan for Future Analysis

In this project we have faced several challenges, from the understanding of the NFL technicalities, feature analysis, selection and engineering, as well as model selection, this last one, will get more defined in the incoming weeks of work.

Our feature analysis or EDA, has been focused on graph analysis, correlation and common statistics as mean, standard deviation, and quartiles, each feature has been analyzed and plotted individually and some analysis like correlation analysis has been made as well. For feature selection Chi-Square was used, these features were used to feed our models:

- playId
- down
- yardsToGo
- yardlineNumber
- personnelD\_5 DL, 3 LB, 3 DB
- personnelO\_2 RB, 3 TE, 0 WR



# Challenges and Plan for Future Analysis

In this case, we have models that give us results for those PersonnelO formations. Other Features like offenseformation needs to be considered, and some feature engineering needs to be made to make our model more realistic approach and make it generalizable.

In the incoming weeks are working on:

1. Feature Engineering focused on defining categories for offenseformation, numberofpassrushers, separated by number-position
2. Features like playID needs to be analyze and converted to catgeories and not int
3. Checking distribution of Play Resul of our data set and compare it with distribution of the predicted Play result of our models are other task that we are integrating to out project to help us to understand and select a specific model
4. EDA using correlations between features to eliminate possible multicollinearity that some features might present



# Conclusions

Our preliminary conclusions indicate that none of our models are currently working all that great. As we continue to optimize our features and our models we anticipate that we will start to get better results. At this time our Logistic Regression Model is working the best.