A Data Driven Look at NFL Passing Strategy

Lazar Gueorguiev, Radin Marinov

Abstract

This project aims to analyze passing-play data released by the NFL for the 2018 regular season. By doing so, teams can optimize a variety of pre-snap factors in order to lead to a favorable outcome. We focus on optimizing for expected points added on the play (EPA), by training a number of machine learning models with pre-snap data to fit for the EPA. This allows us to analyze our highest performing models to see which features are most important in determining the EPA. While one model had the best performance, most of them were similar enough, which allowed us to examine the overlap between which features were most important to each model.

Keywords: passing, NFL, data analysis, football strategy

1 Introduction

From movies such as "Moneyball" and "ESPN: Enhanced," we have learned that data analytics are playing a big role in major sport organizations. An article published earlier this year by Dell explains how the NFL is just starting to tap into the potential of data analytics (Hayhurts, 2020). According to the article, the NFL started collecting and publicly making available data they collected through RFID chips in shoulder pads in 2014. Their goal with this project is to empower teams to use this data for whatever reasons they see fit, whether that be to make changes with their plays or strength and conditioning coaching.

Football is a game of strategy such as chess, with designed plays and players fulfilling different roles on the field. With this new data becoming available, the problem that we want to address is how to identify the positively impactful strategies that lead to winning outcomes for teams. More specifically given our dataset, we wish to identify the offensive decisions (formation, personnel) given game environment (down, distance, score, etc) and defensive decisions (formation, personnel, etc) that lead to positive gains for the offense on passing plays. Given that play calling is a large part of an offensive strategy, we believe that this is an important challenge that data analytics may be able to help solve.

2 Prior Literature

With sports being an important part of many people's lives, open source machine modeling tools becoming increasingly available, and huge amounts of data being released by the NFL, there are various academic projects in the same realm as this one. Like any sport however, the NFL has a lot of complexity surrounding its sport which leaves room for a lot of different analysis to be done. From the research that we have conducted, most academic papers and analyses focus on improving the performance of the modeling itself by experimenting with different models and deep learning techniques.

One such analysis seeks to predict yardage outcomes of plays using image data in combination with deep learning models (Taylor, 2020). The main goal of this analysis, as summarized in the paper's abstract, is to increase the performance of previously used machine learning models in predicting yardage outcomes. Another analysis done by Pablo Bosch at Vrije University seeks to find the best performing models for predicting a game's winner (Bosch, 2018). Within his analysis, Bosch uses a wide variety of models, both supervised and unsupervised learning as well as deep learning. It's clear that the focus of his analysis is model performance. Lastly, an analysis conducted by Rory P. Bunker and Fadi Thabath seeks to develop an Artificial Neural Network for sport result prediction for a variety of different sports (Bunker and Thabath, 2017). Once again, the focus of the analysis is on the model itself, its application in previous works, and how it can be applied in the future.

While we aim to have a model that is high performing, that is only part of our overarching goal. We also want our results to be interpretable, so that teams can make pre-snap decisions in order to maximize their goal. For example, gaining the maximum amount of yardage from a play may not be the team's goal, based on the type of offense formation they are using. Thus, our project aims to dive deeper into the analysis of passing-play decisions, rather than solely focusing on optimizing a model for the purpose of predicting a certain target.

3 Project Description

3.1 Data

Our data was retrieved from kaggle.com as part of Kaggle's "NFL Big Data Bowl 2021" competition. The dataset contains player tracking, play, game, and player level information for all possible passing plays during the 2018 NFL regular season. Here, passing plays are plays where a pass was thrown, the quarterback was sacked, or any one of five different penalties was called (defensive pass interference, offensive pass interference, defensive holding, illegal contact, or roughing the passer). Lineman data is not provided for any play. We focused on play data, for which there are roughly 19,000 plays with 27 variables associated with each play. These variables describe game situation (score, time remaining, etc), offensive choices (formation, personnel), defensive choices (personnel), and play outcomes.

We decided that the play outcome that we would focus on would be EPA - Estimated Points Added. EPA is a statistic that is calculated by comparing the difference in the down, distance, and field position before and after a play was run. We decided that we would focus on EPA as our target outcome instead of another viable metric, yards gained, because not all yards are equal. In short, a 3 yard gain is much more valuable on 3rd and 1 than it is on 1st and 10, and a 10 yard gain is much more valuable when the offense is 20 yards away from the opponent's end zone than when they are 50 yards away.

3.2 Data Analysis

Prior to building any machine learning models, we wanted to get a better understanding of the data on our own, so we performed various aggregation functions, such as mean or count, on our play data aggregated by combinations of variables such as offensive formation, offensive personnel, and defensive personnel. Our goal with these aggregations was to see if there was any trend or relationship between the combination of variables, their values, and EPA.

3.3 Modeling

After preliminary data analysis, we wanted to utilize machine learning methods to explore if there were other, hidden connections between variables that were not able to be easily seen manually. In addition, we also believed that we may be able to confirm/rebuke any conclusions we made from our prior data analysis. For our models, we would split up our data into a training set and a testing set. We would train the models on the training set, and evaluate their performance on the testing set. We would tune our models for best performance on the testing set, and then interpret the models afterward. We decided on three classes of models: linear regression, tree-based, and neural networks. With linear regression, we can interpret results as coefficients for each variable, where a positive coefficient is good for EPA, and a negative coefficient is bad for EPA. With tree-based models, we can interpret relative importance of variables (how does a variable show up in the tree model). With neural networks, we seeked to maximize performance, but their interpretability is very low, so these models would be used to see the limit of performance on our data.

4 User Testing/Iterating

4.1 Data Cleaning

Prior to beginning any analysis, we wanted to make sure that our data only contained plays that were relevant to our goal. Thus, we made the decision to remove several types of data: passing plays that had some penalty called on either team, passing plays where the play type was unknown, and plays that were not run out of an offensive formation (e.g. passes during a play fake on special teams). We also removed plays that were missing data. This resulted in a dataset of 17935 plays from the original 19239.

4.2 Feature Engineering

An important aspect of our project was creating new features from existing ones that could be used both as inputs or outputs in our model. This iteration on top of the existing data that we obtained from kaggle allowed us to both clean up some existing data and examine new features or targets we felt could have a greater impact on the interpretability of the passing play. For example, the kaggle dataset included a feature called “yardsToGo” which is the distance needed for a first down. We created a new feature called “longYardsToGo” which was a boolean feature that was set to ‘true’ if there were seven or more yards needed for a first down. We added this feature in order to try and see how other features were affected, like the type of offense formation used, in response to having a large number of yards to a first down. An example of a target we created was another boolean variable called “touchDown” which was set to “true” if the passing play concluded with a touchdown for the offensive team. This would allow us to analyze all the plays that resulted in a touchdown and see if we can find anything interesting about the pre-snap features.

4.3 Data Analysis

We conducted our data analysis on offensive formation, offensive personnel, and defensive personnel. To do so, we wrote code that would aggregate plays by each variable and give the mean EPA as well as the count of plays that occurred under that combination of formation/personnel. We wrote the code to generate these aggregations automatically, and manually interpreted the results.

4.4 Model Evaluation

The performance of the initial six models we trained was evaluated by calculating the root mean square error of each model on a testing set of data that it had not encountered. Root mean squared error (RMSE) allowed us to see the deviation of the models’ predictions from their respective regressions. By calculating the RMSE on a testing set, we were able to objectively determine the performance of each model because they would not have previously fit to the testing data.

4.5 Linear Models

For our linear models, we began with a linear regression on EPA. Our model input variables were based on game situation, offensive decisions, and defensive decisions. The input variables that we used can be found in Appendix A. After training the first linear model, we decided to try to iterate on it by incorporating some feature selections as well, which is where we attempt to separate the non-important variables out of the model. To do this, we used Lasso regression, which is a linear regression that penalizes weights which do not improve the model and shrinks their value toward 0. To identify the best Lasso model (by finding the optimal lambda parameter that is the model’s input), we tested the model on a range of lambda values between 0 and 1 and selected the best performing value (0.01).

4.6 Tree Models

We used three different tree models in our analysis: Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression. We fit each model with our testing set of data and evaluated them by the same process discussed in Section 4.4. In a sense, Random Forest Regression and Gradient Boosting Regression iterate on top of the Decision Tree Regression model in their own different ways. We decided to include these models in our analysis instead of just using the Decision Tree model because we hypothesized that they would be higher performing models.

4.7 Neural Model

Our neural network was a multi-layer perceptron model. For this model, we decided to tune the model for activation function between each layer, the learning rate, the number of layers in the model, and the number of neurons in each layer. To do so, we ran models for each possible combination of activation functions, layers, and number of neurons. The specific combinations can be found in appendix B.

5 Findings

5.1 Analysis Results

From our analysis, we found that aggregating plays by multiple variables left the resulting aggregations as mostly small samples of the dataset overall, so for that reason we instead focused on aggregations by one variable only. We first looked at offensive formations, for which there were seven of them: Shotgun, Singleback, Empty, I-form, Pistol, Jumbo, and Wildcat. We found that Shotgun was the most common formation, accounting for 66% of passing plays, but it resulted in a negative EPA on average. In addition, the Empty formation (12% of plays) also resulted in negative EPA on average. Conversely, formations which feature a more significant threat of a run had positive EPA on average; these formations are Singleback, I-form, and Pistol, who account for 15%, 5%, and 1% of passing plays, respectively. The formation with the highest EPA on average was the Jumbo formation, which only accounted for 0.2% of plays. We believe that this is because the Jumbo formation (which involves bringing an extra offensive lineman) is only used in situations where the offense is very close to scoring a touchdown, thus giving plays a high EPA. A table of these results can be found in Appendix C.

We also looked at offensive personnel. Typically on the football field, on offense there are five offensive lineman, one quarterback, and five other skill position players that can vary. From our data, 1 running back, 1 tight end, and 3 wide receivers was the most common personnel (71% of plays), but this resulted in a negative EPA on average. The next two most common personnel groupings actually both had a positive EPA on average: swapping a wide receiver for a second tight end (14% of plays) or swapping a wide receiver for a second running back (6%). A table of these results can be found in Appendix D.

5.2 Model Evaluation & Interpretation

As stated previously, we decided to evaluate our models by calculating the root mean square error of each model on a testing set of data that it had not encountered. All of our models had a similar RMSE, which made us believe that there was no model that was much better at prediction than the others. Thus, we elected to interpret the results of our linear and tree based models, as our neural network did not perform much better than either and also does not offer any interpretability. A table of RMSE for each model can be found in Appendix E. Our linear regression model with Lasso returned some results that made sense intuitively, which made us trust the validity of the model. Variables such as yards to go, down, and number of pass rushers all had negative coefficients, meaning that the higher their values were, the worse the EPA for a play would be. In addition, we also found that other model coefficients were similar to the work that we had done in the initial data analysis; formations such as Singleback, I-form, and Pistol were given positive coefficients, indicating that they were associated with positive gains in EPA.

5.3 Feature Importance & Linear Coefficients

One of the driving goals for this analysis was to have results that were interpretable, so that offensive passing plays can be made through the results of our analysis. The Linear Regression models assign coefficients to each feature, so looking at the coefficient of a feature tells us how important it is in determining the EPA and whether it affects it negatively, or positively. Graphs for the linear coefficients of the Linear Regression models can be found in Appendix G, and the relationship between feature number and feature name can be found in Appendix H. According to these graphs, the most significant features for the Linear Regression model are having a shotgun, or singleback offense formation (decrease in EPA), and having a traditional drop back (increase in EPA). The most significant features for the Linear Regression with Lasso model are those that have a negative impact on the EPA, which are: the number of yards needed for a first down, and the down that the offensive team is at. Each of the three tree models have an attribute called "feature_importance" which lets us interpret the results similarly to the way linear coefficients were. Graphs of the feature importance for these models can also be found in Appendix G. The Gradient Boosting Regression model had the same results as our Linear Regression with Lasso model. The Random Forest Regression model had the time at which the play was conducted, and the distance to the end zone as its most important features. We did not take into account the Decision Tree Regression model as its performance was significantly worse than the other models.

6 Discussion

From our initial data analysis, we believe that we have shown an insight that offensive formations which feature a larger threat of a running play are more conducive to successful passing plays than formations which do not feature this threat. The reason that we believe that this is true is because of the average EPA by formation. For example, when we compare the Shotgun formation to the Pistol formation, they are both similar in that the quarterback begins the play a few yards behind the center. However, in Shotgun formation, a running back is typically immediately to the left or right of the quarterback; in Pistol formation, the running back is 2-3 yards further behind the quarterback. In the Pistol formation, this distance away from the quarterback allows the running back to build up speed and momentum as they would take a handoff on a running play, whereas in Shotgun formation, the running back takes the handoff from a standstill position. When defenses see a Pistol formation, they are likely more prepared for a running play than they would be if the offense were in Shotgun, and thus less prepared for a pass. Being able to keep the defense from knowing that a passing play is coming seems to be at least one reason why the EPA for a passing play out of Pistol formation is higher than out of Shotgun. This comparison can be made as well for the Singleback and I-formation, which feature a running back several yards away from the ball in the backfield, with the Empty formation, which does not feature any running back, and thus very little running threat. Both Singleback and I-form have positive EPA on average, whereas Empty formation has negative EPA. When comparing these formations, we also know that with Singleback and I-form, a play action can be called (this is when a running play hand off is faked, thus attempting to fool the defense during the play), whereas with Empty set, it cannot.

This takeaway is also confirmed by the analysis done on offensive personnel. Substituting personnel meant for passing (wide receivers) for personnel meant for running and blocking (running backs and tight ends) both led to an increase in average EPA. Our models also seem to confirm this takeaway, by giving positive coefficients to these formations with a larger threat of a run. Thus, making the play more difficult for the defense to determine whether it is a run or a pass seems to be conducive to passing play success, and we have found that strategywise, this can be done by offensive formation decisions, as well as offensive personnel decisions.

7 Limitations and Future Work

There are some limitations that we faced as we progressed with our project. One very large limitation is that we did not/ were not able to account for specific player talent when assessing plays and their outcomes. Each player on the football field has their own talent level, ranging from Hall of Famer to bench warmer, so being able to separate this variable for each position would have most likely been beneficial for us in looking at how offensive formations and personnel decisions affect plays. For example, a team with three very good receivers and no good running backs or tight ends may be better off keeping their talented players on the field instead of pursuing formations or personnel that would leave them on the bench.

In addition, we only had access to one season of data, so it would have been beneficial for a larger sample size to remove noise from the dataset. This was a problem when we tried to aggregate data using multiple fields; in most cases, these aggregations only represented extremely small percentages of plays.

Another limitation that we believe exists with the technology of RFID chips is that they may not be accessible for everyone depending on the sport as well as where they would be placed. In the future, we would like to see more consideration for the use of RFID chips with the idea of inclusion for all. An example of technology tracking data generated by players would be SpokeSense, which is a fitness tracker for wheelchair athletes.

Finally, the largest limitation of our project is that we did not have access to running play data; we only had access to passing play data. Given our findings that we believe that making a play more difficult for the defense to determine whether it is a run or a pass seems to be conducive to passing play success, it would have been ideal to see if prior running success may influence future passing success (or vice versa).

In future work, we would like to analyze all offensive plays, not just passes, in order to determine the optimal strategies for successful offense. We also would like to be able to use the RFID chip technology to a greater potential (assessing exact player position and movement on the field as a cartesian graph).

8 Conclusion

Overall, we are very pleased with the results that we found from our project. With many people saying that the NFL is becoming a pass-heavy or pass-oriented league, we believe that our results show that this may be jumping the gun, and the running game and the threat of it still has impacts as well. Making an offensive play more difficult for the defense to determine whether it is a run or a pass seems to be conducive to passing play success, and we have found that strategywise, this can be done by offensive formation decisions, as well as offensive personnel decisions.

References

Bosch, Pablo. (2018, February). Predicting the winner of NFL-games using Machine and Deep Learning. *Vrije University*. Retrieved from https://www.math.vu.nl/~sbhulai/papers/paper-bosch.pdf

Bunker, Rory P. and Thabtah, Fadi. (2017, September). A machine learning framework for sport result prediction. https://doi.org/10.1016/j.aci.2017.09.005

Hayhurst, Chris. (2020, January 7). Big Hit: The NFL Turns to Data Analytics. *Dell Technologies*. Retrieved from https://www.delltechnologies.com/en-us/perspectives/big-hit-the-nfl-turns-to-data-analytics/

Taylor, Cameron. (2020). Deep Learning for In-Game NFL Predictions. *Stanford University.* Retrieved from http://cs230.stanford.edu/projects_winter_2020/reports/32263160.pdf

Link to Code: https://github.com/radinmarinov/NFL-Passing-Final-Project

Appendix

A. Model Input Variables

Our model input variables were based on game situation, offensive decisions, and defensive decisions. Some of the variables had to be converted to dummy variables for modeling. The variables are:

'quarter', 'down', 'yardsToGo', 'defendersInTheBox', 'numberOfPassRushers', 'absoluteYardlineNumber', 'longYardsToGo', 'redzone', 'numRBoffense', 'numTEoffense', 'numWRoffense', 'numOLoffense', 'numDLdefense', 'numLBdefense', 'numDBdefense', 'gameClockSecsQuarter', 'gameClockSecsOverall', 'lastTwoMinutes', 'offenseScore', 'defenseScore', 'offenseFormation_EMPTY', 'offenseFormation_I_FORM', 'offenseFormation_JUMBO', 'offenseFormation_PISTOL', 'offenseFormation_SHOTGUN', 'offenseFormation_SINGLEBACK', 'offenseFormation_WILDCAT', 'typeDropback_DESIGNED_ROLLOUT_LEFT', 'typeDropback_DESIGNED_ROLLOUT_RIGHT', 'typeDropback_SCRAMBLE', 'typeDropback_SCRAMBLE_ROLLOUT_LEFT', 'typeDropback_SCRAMBLE_ROLLOUT_RIGHT', 'typeDropback_TRADITIONAL', 'typeDropback_UNKNOWN'

B. Neural Network Parameters

Activation Functions:

‘identity’, ‘logistic’, ‘tanh’, ‘relu’

Hidden layers

2,3,4

Neurons per layer

2, 5, 10, 100

Learning Rate

0.0001, 0.001, 0.01, 0.1

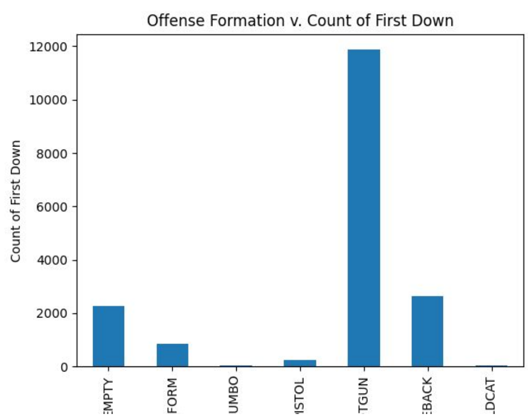
C. Formation Analysis

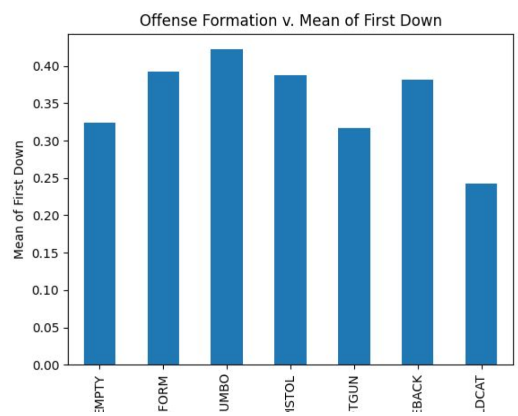
Formation	Average EPA	PCT of Plays Run
SHOTGUN	-0.0788406	66.2
SINGLEBACK	0.163521	14.62
EMPTY	-0.0569778	12.61
I_FORM	0.251433	4.83
PISTOL	0.200709	1.31
JUMBO	0.464762	0.25
WILDCAT	-0.159156	0.18

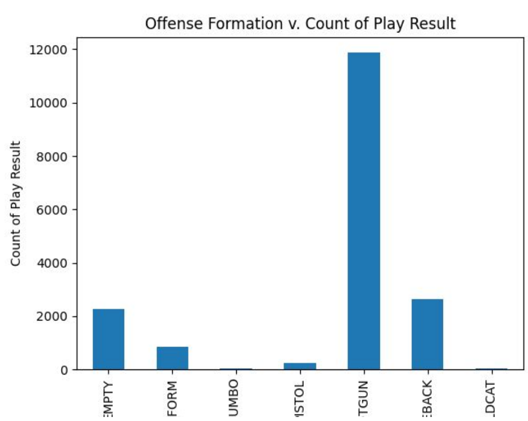
D. Offensive Personnel Analysis

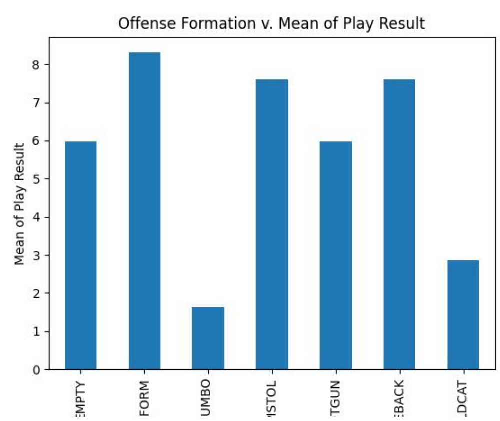
numRBoffense	numTEoffense	numWRoffense	Average EPA	PCT of Plays Run
1	1	3	-0.0716761	71.2
1	2	2	0.144216	14.47
2	1	2	0.102368	5.73

E. Relationships Between Inputs and Outputs from Analysis







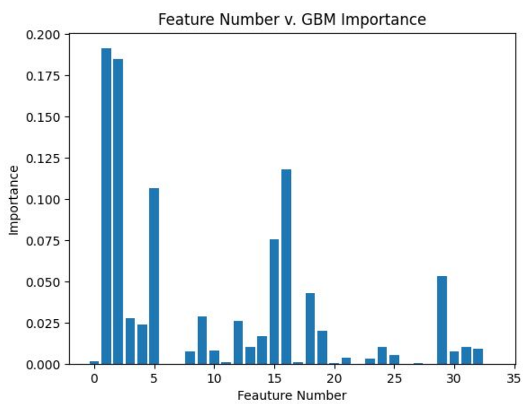


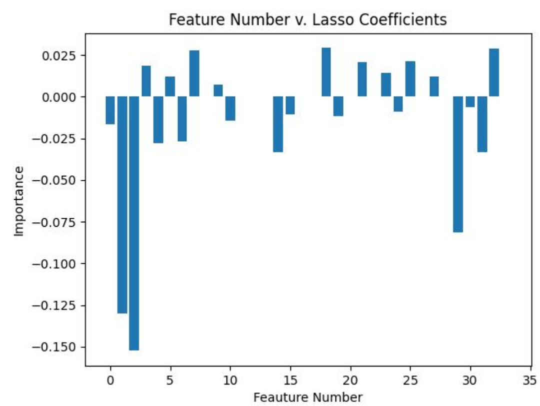
F. Model RMSE

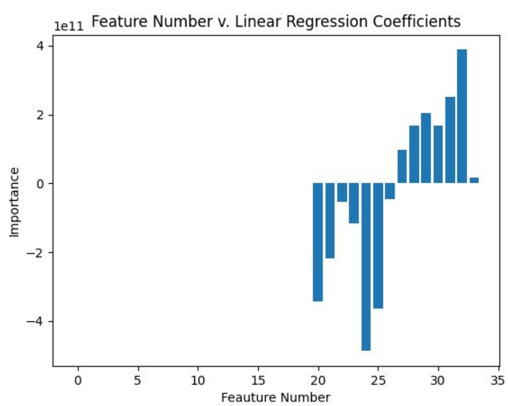
Model	RMSE
Decision Tree	2.304

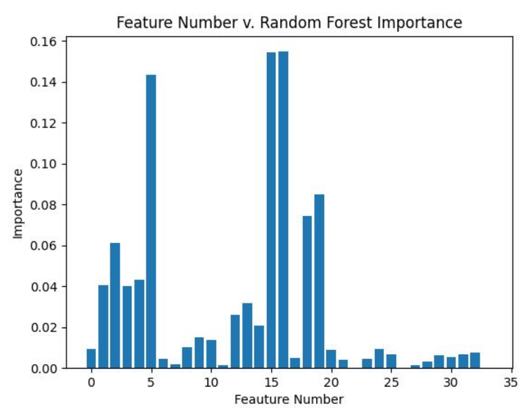
Random Forest	1.594
Gradient Boosting	1.552
Multi-Layer Perceptron	1.546
Linear Regression	1.549
Linear Regression with Lasso	1.548

G. Feature Importance and Linear Coefficients









H. Relation of Feature Number to Feature Name

```
0 quarter
1 down
2 yardsToGo
3 defendersInTheBox
4 numberOfPassRushers
5 absoluteYardlineNumber
6 longYardsToGo
7 redzone
8 numRBoffense
9 numTEoffense
10 numWRoffense
11 numOLoffense
12 numDLdefense
13 numLBdefense
14 numDBdefense
15 gameClockSecsQuarter
16 gameClockSecsOverall
17 lastTwoMinutes
18 offenseScore
19 defenseScore
20 offenseFormation_EMPTY
21 offenseFormation_I_FORM
22 offenseFormation_JUMBO
23 offenseFormation_PISTOL
24 offenseFormation_SHOTGUN
25 offenseFormation_SINGLEBACK
26 offenseFormation_WILDCAT
27 typeDropback_DESIGNED_ROLLOUT_LEFT
28 typeDropback_DESIGNED_ROLLOUT_RIGHT
29 typeDropback_SCRAMBLE
30 typeDropback_SCRAMBLE_ROLLOUT_LEFT
31 typeDropback_SCRAMBLE_ROLLOUT_RIGHT
32 typeDropback_TRADITIONAL
33 typeDropback_UNKNOWN
```