Pass or Run: Predicting Play Types in the NFL

Jonathan Steinberg

June 2021

1 Introduction

There is an old saying in sports; offense wins games, defence wins championships. This statement perfectly describes my favorite football team, The Pittsburgh Steelers. Year after year, their high-powered offense allows them to outscore their opponents and win most of their regular season games. However, during the playoffs, their lackluster defense gives up too many points to the opposing team, preventing them from winning championships. One of the major challenges that an NFL defence is faced with is anticipating the play of the opposing team and aligning their players accordingly. Often, this is a matter of preparation and knowing the other team's tendencies. However, what if there was a way to predict the opposing team's play based on statistics rather than intuition?

Using statistical modelling and data analysis techniques, this project will attempt to predict the play type of the Steelers' opponents. Characteristics prior to each play will be measured using play by play data and will be used to predict whether the team on offense will pass or run. Don't get me wrong, NFL coaches already account for all of this information to decide the optimal defensive formation to stop the opposing team's play. However, if there was a model that objectively assessed the likelihood of the opposing team passing or running, it can serve as a gut-check to all coaching decisions made on the sideline. After all, if the Steelers had access to such a model, perhaps they would have a couple more championships to their name.

2 Data Description

2.1 Data Preprocessing

The dataset used in this study contains 255 variables and 449,371 observations of all NFL play by play data for seasons between 2009 and 2018 inclusive. The columns of the dataset were condensed by only keeping variables that are measured prior to a play occurring. First,

Figure 2: Data Dictionary

play_id	*ID of the play*
game_id	*ID of the game*
home_team	*Team that is playing at their home field*
away_team	*Team that is not playing at their home field*
posteam	*Team that is currently on offense*
posteam_type	*Specifies home or away for the team currently on offense*
defteam	*Team that is currently on defense*
side_of_field	*Specifies if team on offense is located on their own side or their opponents side of the field*
yardline_100	*Specifies yard line from which the play is occuring*
game_date	*Date that the game is being played on*
quarter_seconds_remaining	*Seconds remaining in the quarter*
half_seconds_remaining	*Seconds remaining in the half*
game_seconds_remaining	*Seconds remaining in the game*
game_half	*The half that the game is currently in*
qtr	*The quarter that the game is currently in*
down	*Attempt that team on offense is on to get a first down*
ydstogo	*Yards needed for a first down*
shotgun	*Specifies if quarterback is directly behind the centre or a few yards away from the centre (shotgun)*
no_huddle	*Specifies if a team does not huddle prior to the play*
posteam_timeouts_remaining	*Amount of timeouts that team on offense has left*
defteam_timeouts_remaining	*Amount of timeouts that team on defence has left*
posteam_score	*Current score of team of offense*
defteam_score	*Current score of team of defence*
score_differential	*Difference in score between team on offense and team on defence*
play_type	*Pass or run*

all observations from the 2018 season were removed to later be used a test data. Then, all observations were eliminated except those in which the Steelers were the team on defense and the team on offense was an opponent of the Steelers during the 2018 season. This was done since the goal of the model is to predict the plays of opposing teams that the Steelers will face during the 2018 season. In addition, all play types that were not a pass or a run were eliminated, as well as all plays occurring on 4th downs or kickoffs. Although predicting punts, field goals, or whether a team will attempt a pass or run on 4th down would prove invaluable NFL coaches, this analysis would be better suited in a separate model given the distinct difference between special teams and a standard offense. The dataset was further cleaned by ensuring that all team abbreviations were the same for teams that relocated between 2009 and 2018. Once the preprocessing stage was complete, the final dataset contained 28 variables and 5,372 observations.

2.2 Variable Distributions and Relationships

The distributions of continuous variables in the dataset were explored through histograms. Of particular interest, the yards to go variable had a mean of 8.8 yards, standard deviation of 3.9 yards, and was positively skewed (Figure 1, Appendix A). The vast majority of yards to go was 10 yards, since once a first down is obtained, the next first down is always 10 yards away. Anything under 10 yards is also common, as teams usually gain yards in small chunks before achieving a first down. However, sometimes players can be tackled for a loss of yards behind the line of scrimmage, causing the yards to go to increase beyond 10 yards. In addition, there are certain anomalies where yards to go increases dramatically due to a fumble or a bad snap by the centre, causing the distribution as a whole to be skewed to the right. However, these observations were not removed; sometimes there are unusual plays that happen in a game, and I wanted the model to capture these outliers. Plays in which outliers are present may even result in higher prediction accuracy. In the case of yards to go being over 10, the model may more accurately predict a passing play.

While most of the variables in the dataset are quantitative, most are not continuous, making it irrelevant to analyze their distributions through histograms or box plots. However, bar plots offer key insights to understanding these variables further. The possession team variable illustrated that Baltimore, Cleveland, and Cincinnati played against the Steelers more than double the amount of the next closest team (Figure 2, Appendix A). This is because these three teams are in the same division as the Steelers. Divisional opponents play against each other twice a year, teams within the same conference play each other once every two years on average, and teams in opposite conferences play against each other every four years. This is why teams like Kansas City, Oakland, and New England played against the Steelers more than teams like Atlanta, Carolina, and New Orleans between 2009 and 2017.

3 Model Selection Methodology

3.1 Variable Selection

The methodology for selecting the variables for the model began by conducting a random forest with all predictors to assess variable importance. First, a random forest model with a cp of 1e-7 was ran. This model was intentionally over-fitted to find the optimal cp value, which was subsequently used in another random forest model to assess variable importance.

As seen in Figure 1 of Appendix B, the shotgun variable was found to be the most important variable in terms of MSE and node purity; removing shotgun from the model would cause both MSE and node purity to almost double. This makes sense since a shotgun formation often has no running back in the backfield, making it very indicative of a passing play. Other important variables included down, and yards to go; removing these variables would have increased MSE and node purity by an average of approximately 60% and 67% respectively. This also makes sense, since teams will tend to pass if they have fewer attempts or a longer distance to obtain a first down. Score differential as well as quarter, half, and game second remaining were also found to be important variables. An incomplete pass stops the clock from running while any running play that remains in bounds does not. Therefore, the losing team will tend to pass as the game progresses in an attempt to preserve time.

Next, a Principal Component Analysis (PCA) was conducted to assess collinearity between predictors. While, quarter and half seconds remaining exhibited collinearity, game seconds remaining was far removed from them and all other variables as well (Figure 2, Appendix B). Game seconds remaining is also inherently valuable as it takes into account other variables such as game half, quarter, as well as the seconds remaining in the half and quarter. Possession team timeouts and defending team timeouts were also found to be highly related as they are often a function of time; both teams tend to use their timeouts the more the game progresses, especially in the last two minutes of each half as a time management

tactic.

Utilizing the knowledge from the variable importance and PCA plots, the following variables were selected; opposing team, game seconds remaining, down, yards to go, shotgun, and score differential. While opposing teams was not found to be particularly important, it was included given that the goal of the model is to predict the play types for the Steelers' 2018 opponents. It is also important to note that these variables are not all-encompassing; other variables that were not selected are key factors in determining whether a team will pass or run. However, to maintain parsimony and limit over-fitting , the number of selected variables was limited to six.

3.2 Model Selection

The boosting algorithm was selected to run the model with the chosen variables. Boosting is an ideal method as it is one of the most powerful predictive algorithms that exists; many simple classification trees are created, each one learning from the previous tree. The boosting model iterated through 10,000 trees, each containing four internal nodes. The boosting algorithm was instructed to follow a Bernoulli distribution, since the response variable play type is a discrete random variable. The boosted model was then used to predict probabilities of play types using 2018 data. For each play, if the probability of a pass is above 0.5, the model predicts the play type will be a pass. Conversely, if the probability of a pass is under 0.5, the model predicts the play type will be a run.

4 Results

The total play type prediction accuracy of the model was found to be 70.9%. This suggests that the model would have predicted 647 plays correctly out of the 913 total defensive plays during the Steelers' 2018 season. Specifically, 71.7% of all pass plays and 68.1% of all run plays were predicted correctly. Passing plays are expected to have a higher prediction

accuracy than running plays because the shotgun formation is a strong indicator of the ensuing play being a pass. In the shotgun formation, the quarterback is a few yards away from the line of scrimmage, usually with no running back next to or behind him. However, a normal formation usually consists of the quarterback being closer to the line of scrimmage with the running back behind him. Teams can easily pass or run out of a normal formation unlike the shotgun formation, causing the play type to be more unpredictable. In the Steelers' 2018 season, 94.1% of shotgun plays were passes, while only 60.1% of plays were passes from a normal formation.

Another important finding is that the model's total prediction accuracy decreases to 57.3% for plays in which the probability of a pass is between 0.6 and 0.4. Therefore, the probability threshold was increased; a probability of over 0.6 now corresponds to pass, while a probability under indicates a run. The plays with probabilities between 0.6 and 0.4 were ignored. The accuracy of the model increased by 2.5% with this new probability threshold. The prediction accuracy of the model continued to increase as the probability thresholds were expanded further (Figure 1, Appendix C). It is important to note that the prediction accuracy for running plays decreases from the 0.7 threshold to the 0.8 threshold, however this may be due there being only 71 running plays that corresponded to the 0.8 threshold.

Play type prediction accuracy was also calculated for each type of opponent. For all play types and for passing plays, the model was most successful at predicting play types for opposing teams in their own division. This was expected since teams play their divisional opponent twice during each season and are therefore more familiar with their style of play. Surprisingly, the model had a higher accuracy rate for NFC teams than AFC teams, despite the Steelers playing AFC teams more frequently. In terms of running plays, the model was most successful at predicting the play type for AFC teams, as opposed to NFC teams or divisional opponents. However, this difference is only marginally, and may be due to the unpredictability that is associated with non-shotgun formations. Team accuracy's were also

calculated for each probability interval (Figure 2, Appendix C).

5 Managerial Implications

In conclusion, a boosting model was ran to predict the play types of the Steelers' opponents during the 2018 season. The total play type prediction accuracy of the model was found to be 70.9%, meaning that 647 out of the 913 plays were predicted correctly. The managerial implications for this model are obvious; coaches of the Steelers can use this model to predict the play types of their opponents and ultimately achieve more defensive success. Furthermore, this model can be replicated for any team in the NFL, as there are other defences in the NFL that are far worse than that of the Steelers. One caveat is that the NFL prohibits the use of any data analytics on the sideline, however there has been push-back in recent years by teams against this outdated rule. While this still stands, teams can still utilize this model while watching film of their opponent in the week prior to the match-up, and simulate these scenarios with their defence during practice. If a thorough predictive model does not prevent the Steelers defence from giving up 30points in a playoff game, than I don't know what will.

6 Appendix

6.1 Appendix A: Dataset Description

Figure 1: Histogram of Yards to Go

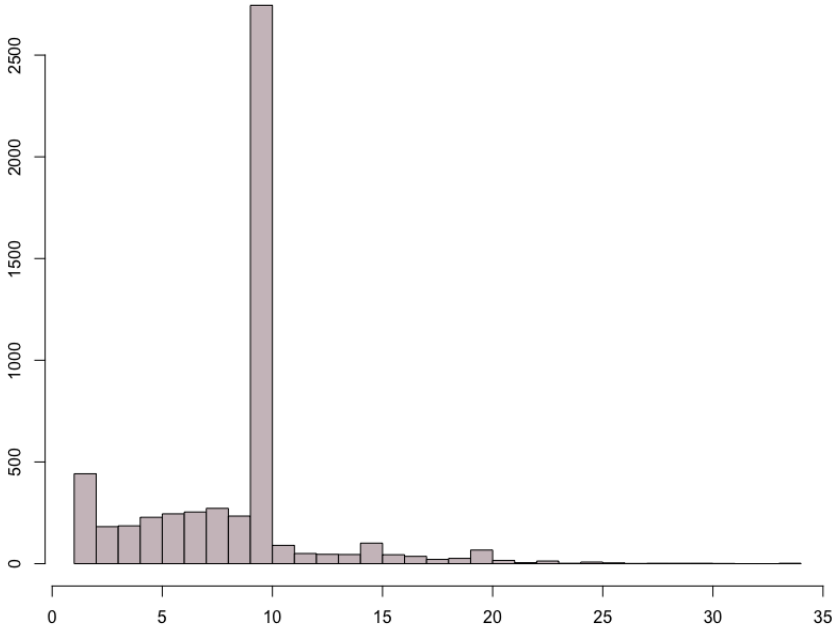
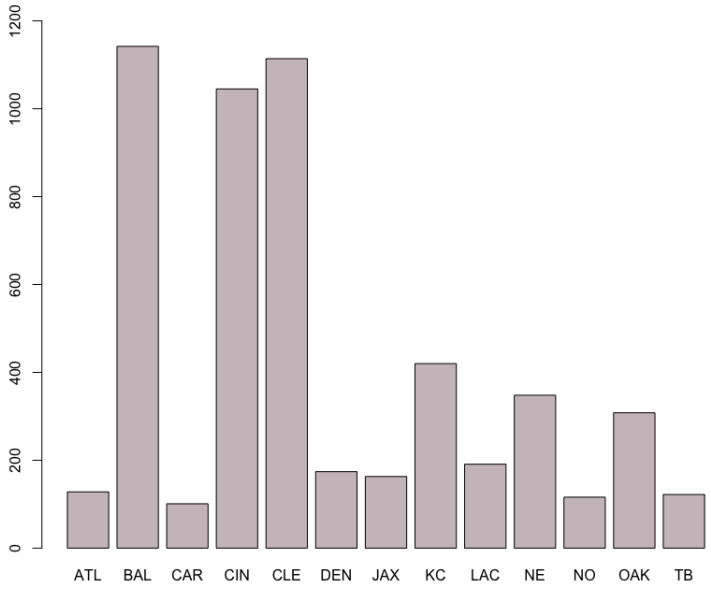


Figure 2: Bar Plot of Possession Teams



6.2 Appendix B: Model Selection

Figure 1: Variable Importance

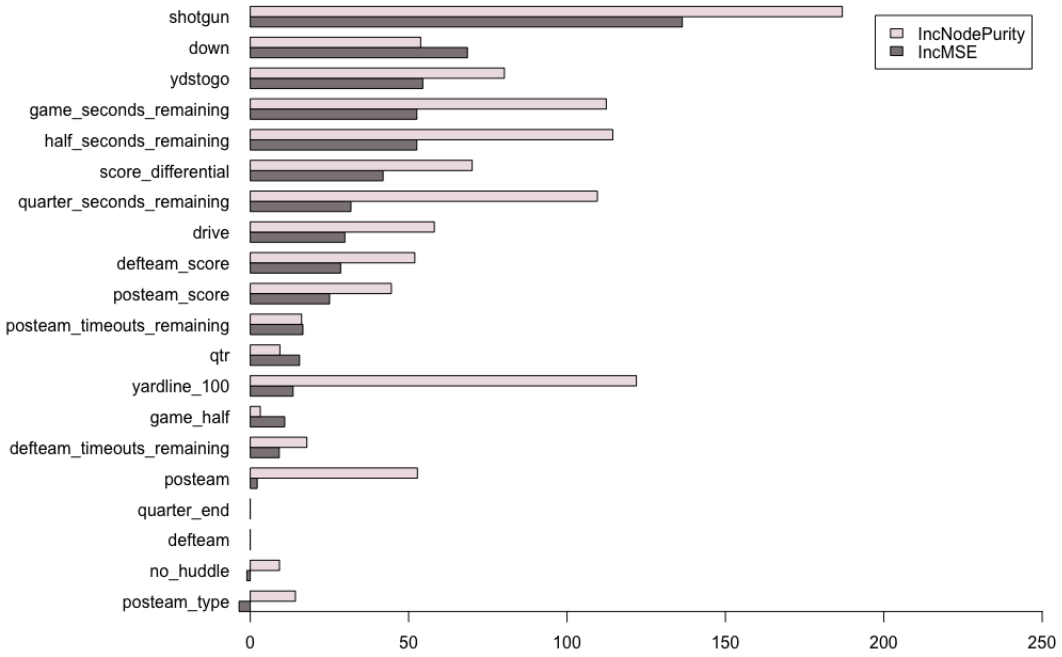
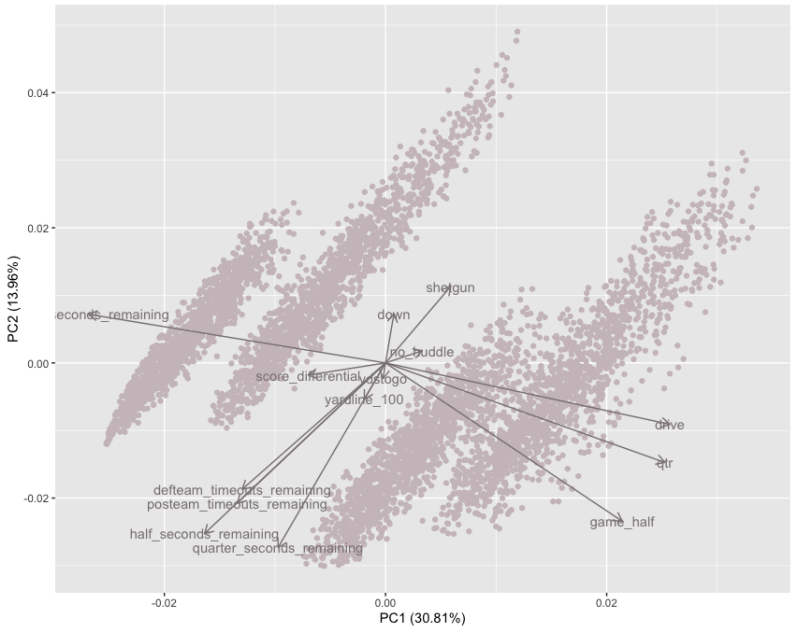


Figure 2: Principal Component Analysis



6.3 Appendix C: Results

Figure 1: Play Type Prediction Accuracy with Different Probability Thresholds

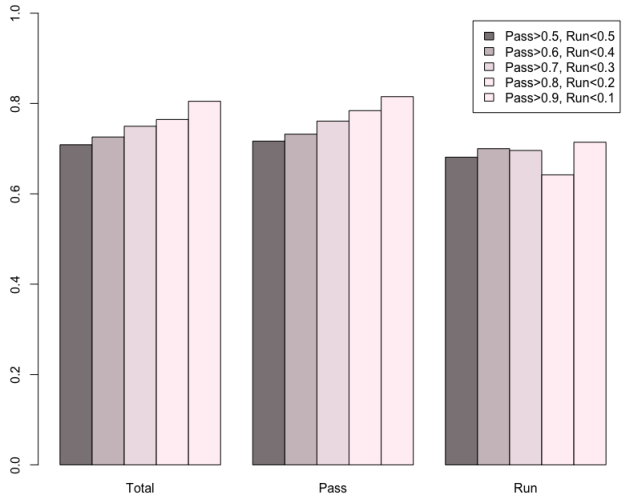


Figure 2: Play Type Prediction Accuracy by Opponent