Exploring NFL Rushing Plays Using Linear Models

Matt Bundas, AST 550 Fall 2020

Abstract

In this project, statistics regarding the sport of American Football were explored. This project's goal is to determine features which can be used to predict the outcome of rushing plays, specifically the number of yards gained. This project uses one dataset, which contains information on tens of thousands of individual rushing plays, and has roughly 50 features. Using this dataset, my goal was to test and determine which features can be used to predict the number of yards gained on a play. Linear models were constructed using JAGS in R, attempting to fit a linear model with a handful of features to predict the number of yards gained in rushing plays. Features were chosen based on general understanding of the sport of football, how well the features performed in the model, and general curiosity. The final models make use of 5 features measured before a play started, including the number of defenders in the box, number of blockers on offense, weight of the ball carrier, whether the team on offense was winning, and the distance the offense needed for a first down. Two models were thoroughly tested models, which for the most part showed similar results. They suggest that the number defenders in the box has a negative and perhaps most significant relationship, winning, weight of the ball carrier and number of blockers have no significant relationship, and the distance to first down has a positive relation.

Introduction

The sport of American Football is not a simple one, and one I have not formally researched previously. It is complex in nature, involves dozens of players, and often described as situational. Despite this, statistics play a major role in the way the sport is played at all levels, with coaches and players placing a heavier and heavier weight on new age statistics the community presents. When enough data is looked at, trends can be found and insights can be made. This project is of course nowhere near the cutting edge of sports statistics, but has been able to provide some insight into the sport of football. In this work, I make use of statistical techniques to explore what might go into a team's success or failure on a given play, but before getting into that, I will give a summary of the sport of football to hopefully provide some context for the statistical side of things.

In its most basic sense, the sport of american football involves two teams made up of dozens of players each, where at any one given time there are 11 players on the field for each team and is played on a 100-yard grass field. One of these teams has the football, and their goal is to try to move the football from one end of the field to the other, they are called the offense. The other team's, the defense's, goal is to try to stop the offense from moving the football down the field by tackling the person with the ball. The offense attempts to move the football down the field using a collection of individual plays. Plays are discrete instances of action and movement,

where the offense tries to move the ball forward, and the defense tries to halt that. There are two types of plays, passing plays and rushing plays. In passing plays the offense throws the ball in the air to someone down the field to advance the football. In rushing plays, the offense hands the ball to one of their players who holds onto it, and attempts to run down the field toward the end-zone to score points, which the defense is defending.

Before a play begins, there is a period of time where the two teams figure out what they think is the best way for them to be positioned, and have an agreed upon plan between their teammates for what they'd like to do. This period of time is where a lot of strategy comes in for both teams. The offense tries to come up with a plan which will give them the best chance of moving the ball down the field, and the defense among other things, tries to guess the offense's plan and put themselves in a position where they can stop the offense from advancing.

There are a lot of strategies each team can implement involving their coordinated plan to execute when the play begins. However, this project concerns the decisions and strategy the teams take before the play begins, and also the circumstances of the game itself before the play begins. In particular, how does the offense line up? How does the defense line up? Who does the offense choose to hand the ball off to? These are all ideas which are explored, determining the decisions or strategies which have an impact, what type of impact, on the outcome of the play, and which don't have an impact at all. Although team formations are different from play to play, Figure 1 shows a simple example of what this might look like.

The bulk of research presented involves the use of linear models, where an equation consisting of constants and user-chosen features is fit to predict the number of yards gained on a play. The results of this fitting are examined and insights are made. In the remainder of this paper, the methods used to carry out this project are shown, results presented and analysed, and finally discussed and concluded.

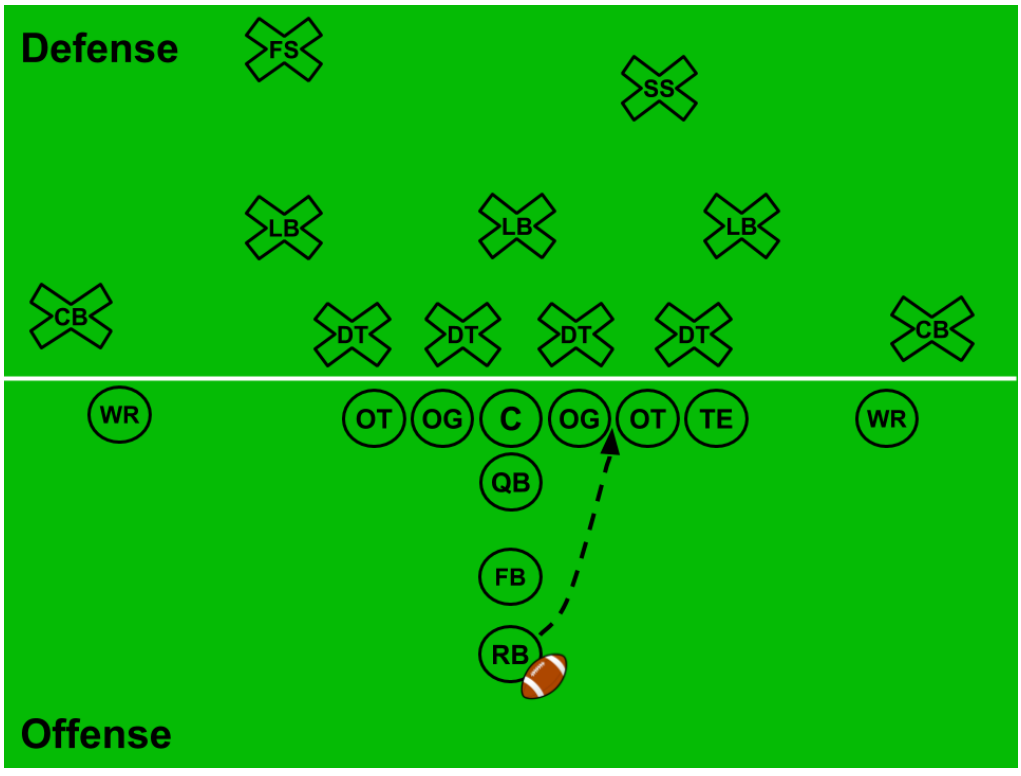

Figure 1
Sample formation of offense and defense before a rushing play begins, separated by white line. Defense is denoted by X's, while the offense is denoted by O's. Player who is handed the ball is denoted using a football image whose path is denoted by a dashed line.

Methods

Before models were able to be formed and tests were constructed, the data from the dataset needed to be processed. As mentioned earlier, the dataset used in this project was originally meant for a Kaggle competition. At its core, the dataset provides information about individual rushing plays. The dataset has information before a given play begins regarding all 22 players involved including their position on the field and attributes like their weight, height. Also provided is team-wide information regarding the formations the two teams took before the play began, as well as game-wide information concerning which team was winning the game, the down, time remaining in the game, weather etc. In terms of the action-side of the play, the dataset provides two items, the direction the rushing play took, either left or right, and the feature this project is all about, the number of yards gained on the play.

While all of this information is great, it was more than needed for the type of analysis performed in this project. The dataset was reduced down to only include the row detailing the player who received the ball on the play. This allowed for focus on the attributes of the rusher, while still having access to the information provided in every row concerning the team-wide and game-wide information. Also, although over 30,000 rushing plays were included in the dataset, for sake of computation time and computer health, a random sample of 10,000 plays was taken and used.

With all of this data, much time was spent selecting and testing features to use in the project and extracting those features from the dataset. Some features are originally found in the dataset while others required feature engineering. At the base of my process was general knowledge about the sport of football, and hypotheses of what makes for a successful or unsuccessful play. Three of the used features came from this line of thinking, defenders in the box, number of blockers and whether the team on offense is winning. Two other features, weight of the ball carrier and distance to first down were not anticipated to be features used in the model, but after testing several easy to extract features they proved to show interesting results.

Features

Yards Gained

The number of yards gained on a play is what this project is all about, and what is trying to be understood. Many factors and situations play a role in the yards gained on a play, which are explored in this project. The yards gained is a measure of how far the offense is able to move the football during a play relative to where the ball was when the play started. This value can be negative if the ball carrier is tackled behind where the offense started, zero if they didn't really

move anywhere, or positive if the ball carrier advanced the football. Typical values for yards gained range from -2-7 yards, with the majority being 3-5 yards. This feature was given directly in the dataset and was easily extracted but required normalization efforts, since linear models behave best when dealing with normally distributed data. The original distribution of yards is shown in Figure 2, where it is roughly normally distributed around 4 yards, but has very long tails. This is the case because while most plays result in just a handful of yards, on very successful plays the offense can go dozens of yards although they are relatively rare. In an attempt to normalize this distribution, cutoffs were made to only include values between -5 and 25 in an attempt to mitigate the tails. The distribution of yards was also made positive by adding the absolute value of the minimum yards gained plus 1, and finally the square root of the distribution was taken resulting in the distribution shown in Figure 2**.** This was the distribution used when creating linear models.

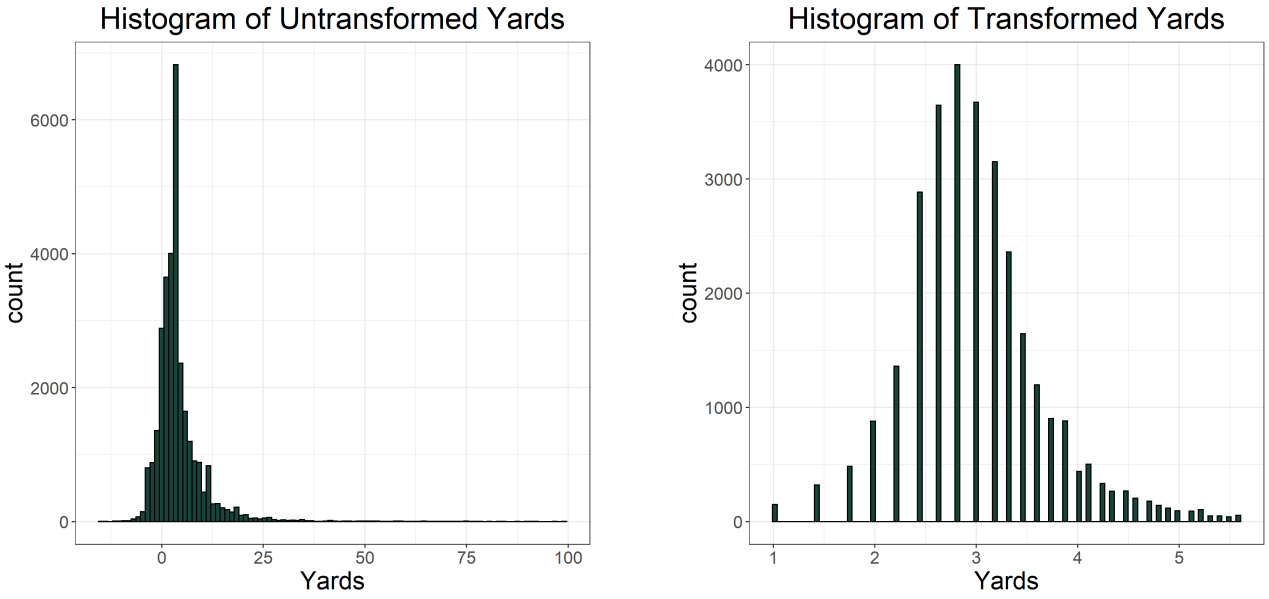

Figure 2

Histograms showing the distribution of response variable, Yards gained on individual plays. Left histogram shows the original distribution. Right histogram shows distribution after being processed, attempting normalization.

Defenders in the Box

The defenders in the box feature attempts to capture how the defense decides to position its players on the field before the play begins. In a simplified sense, the number of defenders in the box refers to the number of players the defense places in a good position to stop a rushing-type play. These defenders are close to the origin of the play, and in a prime position to track down the offensive ball carrier. See Figure 3 for a visual representation. With regards to defenders in the box, the number of defenders the defense decides to put in the box pre-play is widely accepted as a major decision the defense needs to make. With more defenders in the box, it is

thought that the defense has a better chance of preventing a successful rushing play from the offense. This value was given directly in the dataset and required no engineering to extract.

Number of Blockers

On the other side of the ball from the defense, the number of blockers feature attempts to capture how the offense decides to position its players on the field before the play begins. Since only one player can have the ball at a given time, the role of the rest of the offense is to push and block the players on defense away from the ball carrier so he can run down the field. The position which an offensive player starts on a play dictates whether they will be able to block effectively for a rushing play. Offensive players who are a part of the main group in the center of the offensive formation are considered to be in a good position to block for a rusher, which is what the number of blockers feature is. It is widely considered that the more blockers the offense puts in this position before the play starts the better chance they have at a successful play outcome. The line of thinking here is that if the offense can block enough defenders from the ball-carrier that they should have an easier time advancing the ball. See Figure 3 for a visual representation of this. The number of blockers feature was not given directly in the original dataset and was extracted from the dataset. For this project the number of blockers was considered to be the number of offensive lineman, tight ends and additional running backs (not counting the ball-carrier) lined up before the play begins. The number of blockers feature was extracted from the dataset using the OffensePersonnel column, which provides a string containing the information of how many players in each position lined up for the offense in the format “# OL, # RB, # TE, # WR”. For each play, this string was parsed and totals calculated, summed, and one was subtracted because it is assumed that one of these players in the string were the ball carrier and could not block.

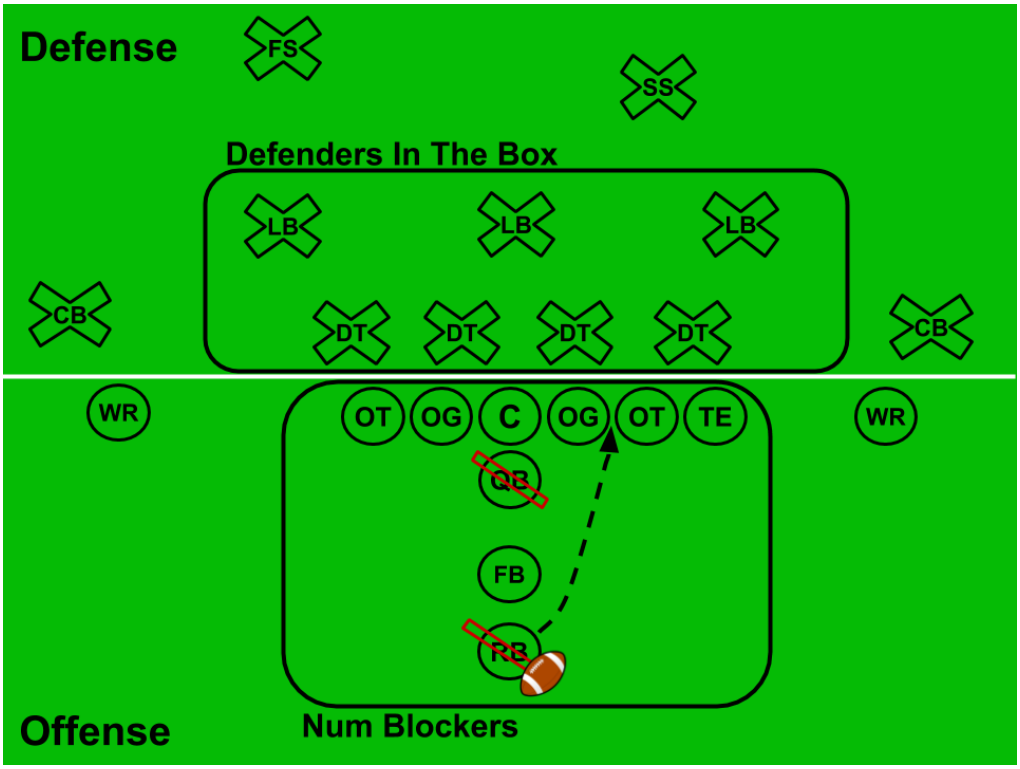

Figure 3
Sample play denoting players counted in number of defenders in the box and the number of blockers calculations.

Winning

With regards to using winning as a feature in linear models, it is widely accepted that the winning team usually is more successful at running the football than the losing team, making it desirable to test with the linear models. It was thought that this feature could serve as a control for the other features, where the outcome of the play might be biased towards the winning team, so incorporating in the linear model would allow for winning to be controlled when considering the other features. This feature was not directly in the dataset, but was easily extracted after considering whether the offense was the home team or away team, then looking at the scores for the home and away teams. If the team on offense was winning, this is represented with a value of 1. If the score was tied or the team on offense was losing, this is represented with a value of 0. In 36% of plays the team on offense is winning, in the remaining 64% of plays the score is tied or the offense is losing.

Weight of Ball-Carrier

This feature is as expected, the weight of the ball carrier measured in pounds. Since size and athleticism of players play such a role in the sport of football, this feature was used as part of the model. However, it was not clear to me what kind of impact this feature would have on the play outcome. It could be argued that a heavier weighted ball carrier would be harder to tackle and run further. However, it could also be argued that a lighter ball carrier would be able to run faster and gain more yards. Weight of the ball carrier was directly included in the dataset and required no feature engineering. The distribution of weight is approximately normal centered around 217 pounds.

Distance to First Down

While on offense, the team gets a set of 4 of what are called downs. These are essentially the number of plays the team gets to advance the football 10 yards from where they received their first down. If the offense makes it the 10 yards in the 4 downs, they get a fresh set of downs and start over again at first down with 10 yards to go, giving them a renewed opportunity to advance the football. Besides trying to advance the football all the way to the other team's endzone, this is the offense's main goal, since if they don't get a first down after a 4th down, they have to give the ball up to the team currently on defense. This feature then, the distance to first down, is before the play starts, how many yards the offense needs to receive a first down. Distance to first down was included in the original dataset and required no feature engineering or extraction. In the vast majority of plays the distance to first down is 10 yards, with 92% of plays having a distance to first down of 10 yards or less.

Linear Models

Linear models are a great tool to use when trying to assess which features can be used as predictors for a response variable, and to what degree the features may have an impact. In simple terms, a linear function is created involving an intercept parameter as well as a number of features with scaling parameters. These parameters are attempted to be fit/estimated using a dataset to minimize the difference between the model's predicted values and the actual value of the response variable. After the model has been fit, the distribution of the estimated parameters can be examined and conclusions can be made. If parameter distributions are centered around 0, it can be argued that the corresponding feature is not significant in predicting the response. If predicted parameter distributions do not contain 0 or can otherwise be shown significant, it can be argued that the corresponding feature is in fact significant in predicting the response.

Model Form

Simple models were constructed in the general form of y ~ B0 + B1x1 + B2x2 … Bnxn, where Bi are the scaling parameters which can be thought of as slopes, xi are values from various features used in the model such as Winning, and B0 is an additive constant or intercept. In this model, y is the response variable, the yards gained. Values for xi are from individual plays and can be continuous such as for player weight, or can be binary such as for winning. Each feature described above were used in some form in the two models examined in this project.

Model 1
Model 1 takes the following form yards ~ B0 + B1*DefendersInTheBox + B2*Winning + B3*Distance + B4*PlayerWeight + B5*NumBlockers. This is a simple model used to explore which features presented in the model are significant in predicting the yards gained or lost on a play.

Model 2
Model 2 takes a similar form as model 1 : yards ~ B0 + B1(NumBlockers - DefendersInTheBox) + B2*Winning + B3*Distance + B4*PlayerWeight. This is again a simple model but the difference between Model 1 is the term being multiplied by B1. In this case, the model is examining the difference in number of blockers and defenders in the box, to see if the difference between the two is significant, rather than their raw values.

Feature Scaling/Standardization

To increase the effectiveness of the linear models, all continuous features including the yards gained were scaled to take a form which can be directly comparable. This allowed the feature values to take a similar range when actually fit in the model, so that something like weight measured in pounds and averaging around 200 doesn't get treated as different as a value like yards to first down which averages something like 5-10. The form of this standardization is

indicated in Figure 4. In this method, each value for a given feature has the mean of the feature value subtracted from it, and then is divided by the standard deviation.

- $x_i = \frac{x_i - \bar{x}}{sd_x}$
- $y_i = \frac{y_i - \bar{y}}{sd_y}$

Figure 4
Shows the method used to standardize continuous features.[3]

Model Running

Linear models in this project were constructed using MCMC methods of JAGS in the programming language R. The essentials for running a linear model in JAGS is the form of the model presented above, initial conditions/priors for parameters to be fit, and then the data used to fit the model. Each parameter was initialized to have a normal, uninformative prior to allow the model to explore the optimal distribution without bias. Data was fed to the model from the dataset after feature extraction, normalization and standardization. For both Model1 and Model2, 6 MCMC chains each with 10,000 iterations after a burn-in period of 2,000 iterations to allow the model to move past any artifacts present directly after model initialization.

Assessment of Model Validity

To assess the effectiveness and accuracy of the linear models, several tests can be constructed and evaluated based on expectations. Depending on the form of the model, model hyperparameters, correlation between features, and many other factors models have plenty of room to go wrong. Correlation between values in MCMC chains is an important thing to control for when it comes to the validity of the model. A high correlation between values can lead to invalid results. To assess the correlation of the models used in this project, autocorrelation values were calculated for each parameter with different lookbacks or lags. The goal value for these which allows for suggestion of little or no correlation is 0, where values around or higher than 0.1 indicate high correlation. Similarly, the number of effective samples from the MCMC linear model can be evaluated to test model validity. In a perfect model, the effective sample size tends towards the number of MCMC iterations multiplied by the number of chains used. In a poor model, this effective sample size is not near that number.. To assess whether the MCMC model has been run for enough iterations to be effective, the Gelman Rubin Convergence diagnostic value can be calculated, indicating how much the distribution of parameter values might change if the model continued to run. A model which has run long enough to converge will have a value of 1, and not require any more iterations. These three statistics, autocorrelation,

effective sample size, and Gelman Rubin diagnostic can and were all calculated using R functions found in the coda package.

Assessment of Model Parameter Results

To assess and make conclusions about the final estimation of parameter values from the simple linear models, the distributions of their values are examined and tested. As indicated earlier, the goal of these assessments is to see which features in the model play an important role in predicting the response variable and which features do not play an important role. If the parameter does play an important role, it is needed to assess what type of relationship that feature has with the response variable, either negative or positive. A negative relationship means that as the feature value increases, the response variable decreases. For a positive relationship, as the feature value increases, the response variable increases. The outputs of the linear models are roughly normal distributions of the estimate of these feature parameter scaling values, giving insight into these questions. These normal distributions are examined to determine the relationship the feature has with the response.

A purely statistical way to make conclusions about the parameter values is to create confidence or credible intervals indicating the range of values which you can be confident the parameter will take. This is typically done with a 90-95% confidence, that is you can assume 90-95% confidence that the parameter takes a value between the given range. In these intervals if 0 is not contained in the interval, that is the predicted range is either all positive or all negative, you can assume the feature value plays a role in predicting the response with the given positive or negative relationship. If the interval contains 0, you can not make a strong argument that the feature is important in the model and has no significant relationship with the predictor.

A graphical, quantitative approach to making conclusions about the parameter values is to create histograms of the parameter distributions. If the distribution is centered around 0, you might expect the feature is not important in the model. If the distribution is mostly or completely contained on one side of the other of 0, you can make the argument that the feature is significant.

Results

Results were all obtained using R through the graphical and statistical approaches described above. Overall, the MCMC models themselves performed well but they were not perfect. I believe they are valid and lead to accurate conclusions but results would likely be more precise if the model was executed in a completely ideal way. Results relating to the predicted distribution of parameter values from the models are meaningful, showing features which have no impact, a negative impact and positive impact on yards gained in a rushing play.

MCMC Model Analytics

The MCMC models were evaluated using the methods described in the Methods section, and are shown in the various figures below. Both model 1 and model 2 performed in a very similar manner, so only model 1 will be discussed, but the same claims can be made about model 2. The statistical analytics for the MCMC models indicate that the models perform well, but not in an ideal manner. Figure 5 shows significant autocorrelation with beta parameters 0,1,2 and 5, but not totally catastrophic autocorrelation. High autocorrelation values are present when looking at Lag 1, but are not present in the other metrics with larger Lag. In an ideal MCMC model, we would like to see autocorrelation values less than 0.1 no matter the lag, but in this project's case we see autocorrelation above 0.35 for several parameters. However when looking at lag 5 the highest autocorrelation observed is 0.015, which is acceptable. This significant correlation is likely why a non-perfect effective sample size is also measured shown in Figure 5. For each parameter, we would hope for effective sample sizes of roughly 60,000 as 6 Markov chains each with 10,000 iterations were used. Instead, for beta parameters 0,1,2 and 5 we see effective sample sizes between 20,000 and 30,000, although other parameters show a good effective sample size. This again is not ideal, but not catastrophic either as the effective sample size is still significant compared to the expected. Finally, the gelman rubin diagnostics shown in Figure 5 are all 1, indicating that our models were run for a long enough duration to converge. Because the chains were able to converge around the same value, this indicates validity in our model.

beta0	28135.574143488
beta1	24315.3473057277
beta2	27637.8335542998
beta3	50984.1379327309
beta4	58180.0183492472
beta5	25094.2610080176
sigma	59692.010020745
tau	59687.5348642128

Potential scale reduction factors:

	Point est.	Upper C.I.
beta0	1	1
beta1	1	1
beta2	1	1
beta3	1	1
beta4	1	1
beta5	1	1
sigma	1	1
tau	1	1

	beta0	beta1	beta2	beta3	beta4	beta5	sigma	tau
Lag 0	1.000000000	1.0000000000	1.000000000	1.0000000000	1.000000000	1.000000000	1.000000000	1.000000000
Lag 1	0.357196335	0.4259503938	0.367817012	0.0698940172	0.021282945	0.399131419	-0.002889549	-0.002919629
Lag 5	0.007613026	0.0138595295	0.004310701	0.0043936536	-0.001161046	0.014808253	-0.002506047	-0.002454289
Lag 10	-0.002091319	0.0008169144	-0.004531788	-0.0006546427	-0.001937407	-0.002111731	-0.004046905	-0.004118993
Lag 50	-0.007941936	0.0016641063	-0.004264180	0.0027232855	0.004338264	-0.002290092	0.003217879	0.003308968

Figure 5

Statistics regarding the performance of the MCMC models themselves. Similar results were found for both models, so only one copy is included. Upper left shows effective sample size, upper right shows Gelman Rubin Diagnostics, lower shows autocorrelation with different lags.

Feature Results From Models

Making use of the methods described in the Methods section, results were found and conclusions were able to be made regarding the features used in the model. Both models boast similar results, although model 1 may give a stronger argument than model 2 for some parameters or visa versa. Histograms and credible intervals are presented in Figures 6-9. A summary of the feature results can be found in Table 1 and Table 2.

Model 1

Defenders In The Box

Predicted by Model1, the feature which seems to have the largest connection with the number of yards gained on a play, is the Defenders In The Box. This is shown both in the histograms and credible intervals. The histogram shows the entire predicted distribution of beta parameter values corresponding to the Defenders In The Box feature well left of zero. This indicates that the number of Defenders In The Box has a negative relationship with yards gained on a play. Similar abstractions can be made with the credible interval, which ranges from -.18 to -.12. This credible interval is well negative and is not close to containing 0 indicating a strong negative relationship.

Winning

The Winning feature interestingly is not indicated to be significant in this model, as the predicted beta parameters value corresponding to Winning is centered around 0. Similarly for its credible interval, ranging from -0.05 to 0.04, 0 comfortable fits in this credible interval.

Distance to First Down

The distance to first down is suggested to have a positive relation with yards gained on a play, that is the more yards needed for a first down the more yards gained on a play are predicted. This conclusion is demonstrated by the histogram with the entire beta parameter predicted distribution being to the right of 0, and credible interval does not contain 0, and relatively is not close to.

Player Weight

With regards to the weight feature, it is hard to make a solid argument that weight has an impact on the outcome of a play as its predicted beta parameter value distribution histogram contains many values near 0, to the left of zero, and to the right of zero. Similarly for the credible interval, it ranges from -0.037 to 0.008, thus containing 0 and indicating no strong relationship.

Number of Blockers

The number of blockers feature is similar to the winning feature in that this model does not predict a relationship between the number of blockers on offense and the yards gained on the play. It's predicted beta value parameter distribution is centered around 0 demonstrated by both the histogram and credible interval.

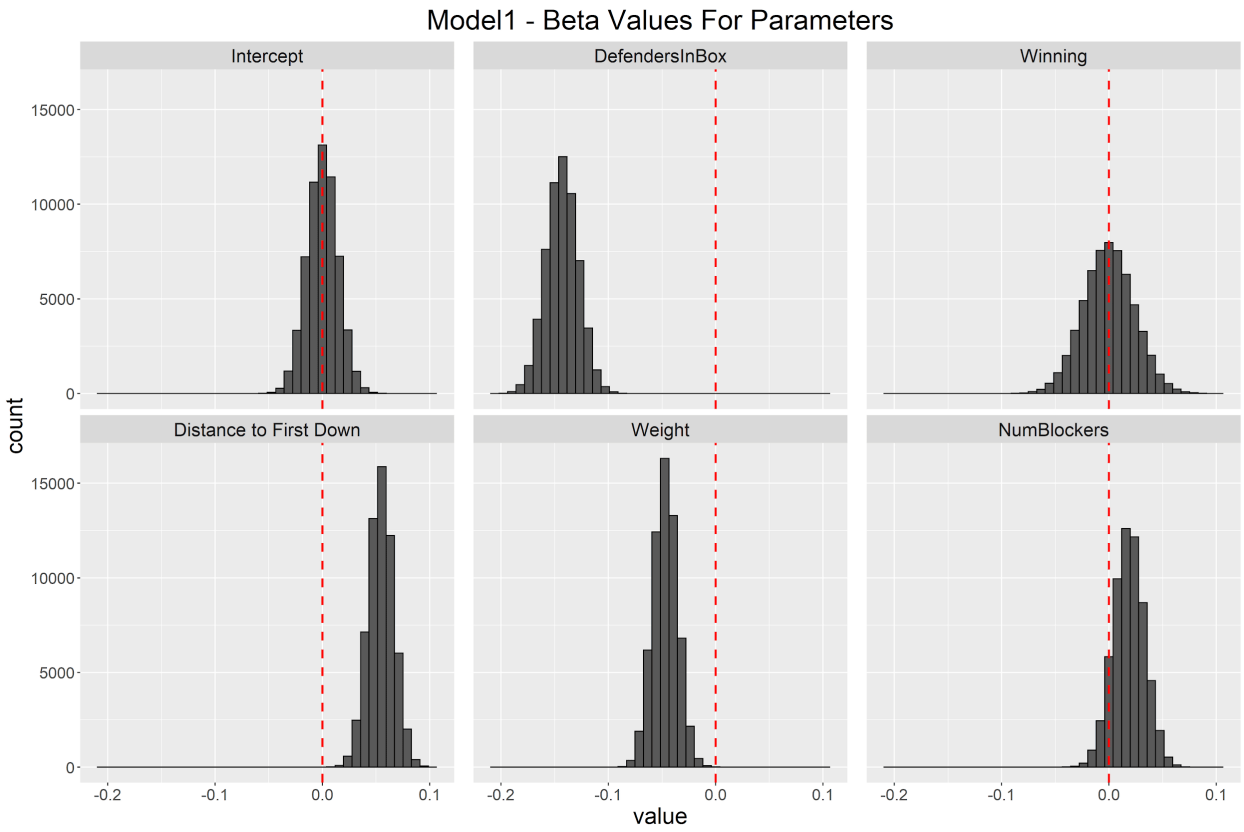

Figure 6
Histograms of resulting beta value parameter distributions from Model 1. X axis shows the value of the parameter. Y Axis shows the number of observations in the bin. Vertical red line denotes the zero point.

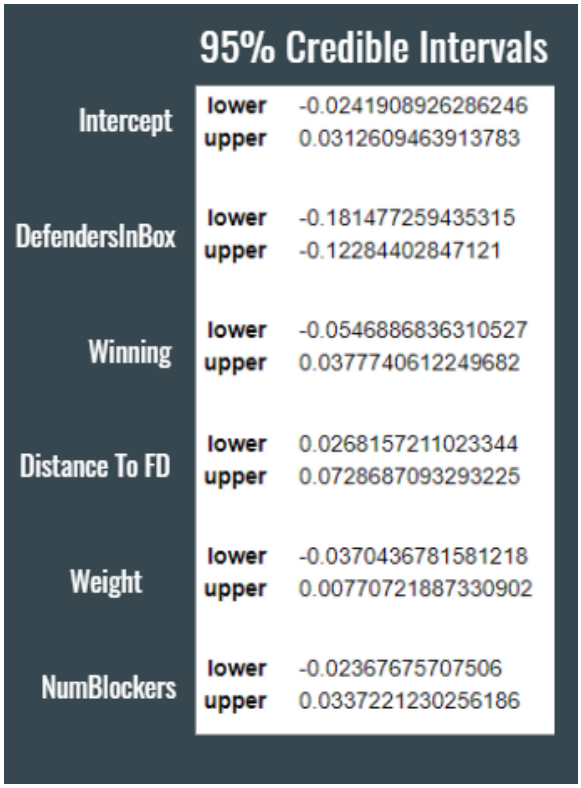

Figure 7.
95% credible intervals for scaling beta parameters for each feature in Model 1.

Model 2

(Number of Blockers - Defenders In The Box)

In model 2, the feature made up of the difference between number of blockers and number of defenders in the box is suggested to be a strong predictor of play outcome. This conclusion can be demonstrated by the histogram of the beta parameter distribution, as its values are all comfortably to the right of zero. Similarly with its confidence interval, running from 0.04 to 0.1, this comfortably contains only positive values and is not close to containing 0. This indicates a positive relationship between play outcome and the value calculated by finding the difference between the number of blockers and defenders in the box features.

Winning

In model 2, Winning is demonstrated to have no relationship with yards gained. If an argument had to be made, it would be that Winning has a negative relationship with play outcome, but it would not be strong. It's beta value parameter predicted distribution contains mostly values to the left of zero in the histogram, and the credible interval is mostly made up of negative values. However, a strong conclusion cannot be made, as the results are mixed with the credible interval containing zero and histogram being divided by zero.

Distance to First Down

Model2 shows the same results as model1, a negative relationship.

Player Weight

Model2 indicates a meaningful negative relationship between ball-carrier weight and outcome of the play. Looking at the histogram, nearly all predicted beta parameter values are to the left of zero, and it's credible interval contains all negative values.

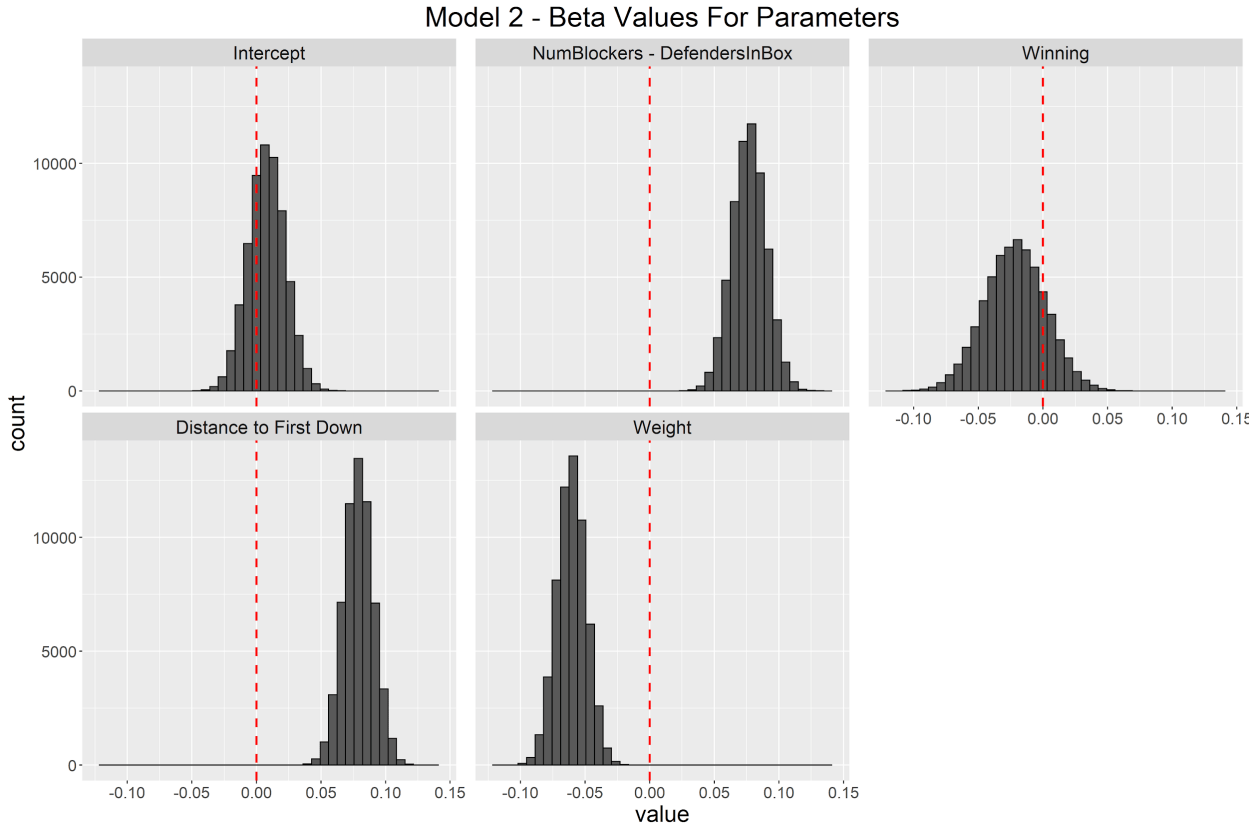

Figure 8
Histograms of resulting beta value parameter distributions from Model 2. X axis shows the value of the parameter. Y Axis shows the number of observations in the bin. Vertical red line denotes the zero point.

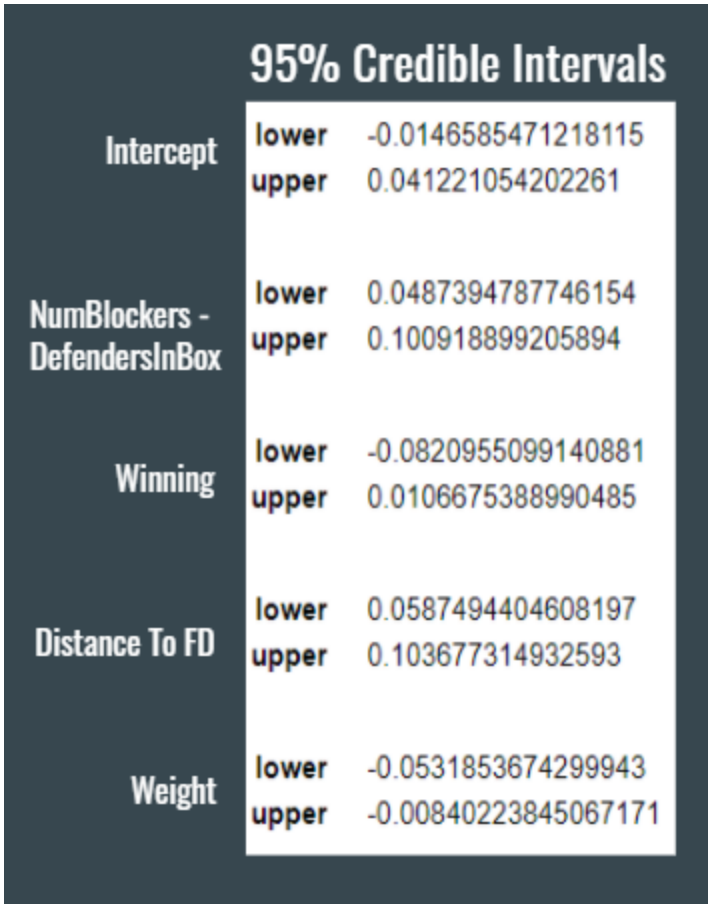

Figure 9
95% credible intervals for scaling beta parameters for each feature in Model 1.

Table of Findings Model 1

Defenders In The Box	Winning	Distance to First Down	Weight	NumBlockers
Strong Negative	No Relation	Strong Positive	No Relation	No Relation

Table 1
Table of findings of model 1 regarding feature relationship with yards gained on rushing plays.

Table of Findings Model 2

(Number Blockers - Defenders In Box)	Winning	Distance to First Down	Weight

Strong Positive	No relation	Strong Positive	Negative

Table 2
Table of findings of model 2 regarding feature relationship with yards gained on rushing plays.

Conclusion/Discussion

In this project, it was able to be shown when considering a handful of features, which of them have a relationship with the number of yards gained or lost and of what type of relationship. The number of defenders in the box before a play starts appears to have a negative relationship on yards gained. The offense winning at the time before the play begins, and weight of the ball carrier are indicated to have no or a slight negative relationship. Distance to first down is indicated to have a strong positive relationship. The number of blockers is concluded to have no relationship with the outcome of a rushing play.

Results of the models lead to the conclusion that the most important variable of the ones tested is the number of defenders in the box. This is especially the case, because it is a decision which can be made before the play snaps and is completely in control by players and coaches. The distance to first down seems to have a relatively similar strength relationship as the defenders in the box, however this is not a simple decision a coach can make, it is a result of previous plays success. The remaining features, number of blockers, winning and weight either are concluded to have no relation or a weak relation, and are probably not all that worth considering if you are a coach or player, as indicated by models in this project. The results concerning the number of blockers and whether the offense is winning are surprising. I would expect that these would both have a strong positive relationship with play outcome. It would be sound to believe that if the offense has more players to block for the rusher, the rusher would have an easier time moving the ball. With regards to winning, I would expect that the winning team would have an easier time rushing the ball as it is reasonable to say that the winning team is likely either better or playing better before the play begins. However the models presented do not make these arguments, they suggest an insignificant relationship.

These surprising conclusions could be a result of some limitations and issues of the study which are worth discussing. Perhaps the most important idea to account for is that the sport of football is incredibly situational. The same rules or way of thinking do not apply for every play as there is a lot of strategy involved on both offense and defense given the situation. However with the sample size used in this project, I expect that it is large enough to extract conclusions about trends and the conclusions do hold some weight. Also worth considering is that these models are not all encompassing. They don't contain every feature which is possible. Controlling for other features could allow for features to be expressed differently and have more decisive conclusions. However, I do believe in the strength of the features which I presented, as they account for both game-wide situations, play-by-play basis situations and formations of both teams.

Also when creating these models, to help massage the yards gained distribution to normality, the sample set was bounded by the number of yards gained to remove the long tails present, leaving just plays with outcomes between -5 and 25 yards. Perhaps if a different method of normalization was used, or these values were able to be incorporated into the model they would be significant enough to lead to different conclusions. Another factor worth considering is that the models presented did not behave perfectly, there were deviations from perfect autocorrelation and sample size, which could indicate something deeper going wrong with the model. Another limitation of this project is that it essentially considers rushing plays in a vacuum without considering passing plays. It would be interesting if I continued to research the subject to somehow account for passing plays either through incorporating related features in the models or created new models altogether.

References

[1] NFL Big Data Bowl. (2020). Retrieved December 09, 2020, from https://www.kaggle.com/c/nfl-big-data-bowl-2020

[2] Trainor, P. (2020). *GLMs in JAGS*. Lecture.

[3] Trainor, P. (2020). *Lecture10NB* [Ipynb].

[4] Trainor, P. (2020). *Markov Chain Monte Carlo*. Lecture.