The Application of Machine Learning to Predict NFL Running Back Performance

STATS-5405

Giovanni Lunetta[a,*], Sam Lutzel[a,*]

[a] *University of Connecticut, Statistics, 2075 Hillside Road, Storrs, 6269*

Abstract

This study employs an ensemble of machine learning techniques, including Multiple Linear Regression (MLR), Generalized Linear Models (GLM), and XGBoost to enhance predictive analytics in the National Football League (NFL). The research focuses on one of the most critical aspects of football offense - the running game. In order to predict the yards gained by NFL running backs following a handoff, the study analyzes a plethora of player tracking data from the 2017 to 2019 seasons. It integrates a variety of factors such as player positions, orientation, and the situational context of the game, including down, distance, and field position. This multifaceted approach aims to yield highly accurate models that can serve as valuable tools for coaching staffs to optimize play-calling, manage player workloads, and enhance game planning. The predictive insights derived from this research are intended to support teams in deploying their running backs more effectively, leading to potentially improved outcomes on the field. As the NFL continues to evolve with a greater emphasis on analytics, this study seeks to contribute significantly to the field of sports analytics by providing a model that underscores the importance of the running game in a predominantly pass-oriented league.

Keywords: Machine Learning, Predictive Analytics, Sports Analytics

1. Introduction

1.1. Background of the National Football League

The National Football League (NFL) is the most popular professional sports league in the United States, with an estimated 187 million fans. The league consists of 32 teams divided into two conferences, the American Football Conference (AFC) and the National Football Conference (NFC). Each conference is further divided into four divisions, with four teams in each division. The NFL season consists of 17 weeks, with each team playing 16 games and having one bye week. The regular season is followed by a 12-team playoff, with the winner of each conference advancing to the Super Bowl, the league's championship game.

The NFL game strategy primarily consists of two ways to advance the football down the field: running and passing. Running the ball is defined as a play in which the quarterback gives the ball to another player that is located behind them (a handoff or a small toss). The player then attempts to run the ball down the field as far as possible before being tackled to the ground. Passing the ball is defined as a play in which the quarterback throws the ball to another player down the field. The player then attempts to catch the ball and advance it down the field as far as possible before being tackled to the ground. The goal of each team is to score as many points as possible by advancing the ball down the field and into the end zone or by kicking the ball through the field goal posts. The team with the most points at the end of the game wins.

*Corresponding author

Email addresses: giovanni.lunetta@uconn.edu (Giovanni Lunetta), samuel.lutzel@uconn.edu (Sam Lutzel)

1.2. Motivation

The primary goal of this research is to develop cutting-edge predictive models capable of accurately estimating the yards a running back will gain after a handoff during NFL games. This objective is pivotal for formulating advanced offensive strategies and refining player evaluations. The running play is a fundamental aspect of the game that can dictate the tempo, control the clock, and establish physical dominance.

The motivation behind this study stems from the transformative impact that data analytics has had on sports, particularly in the NFL, where the fusion of technology and sports science has begun to redefine how the game is played and understood. In an era where marginal gains are increasingly sought after, the ability to predict the outcome of running plays with high precision can provide teams with a significant competitive advantage. It enables coaches to make informed decisions regarding play selection, player rotations, and game management, especially in critical moments of a match. Additionally, the insights from this study can empower front offices in their scouting and drafting processes by quantifying the expected value a running back adds to their team. Furthermore, there is also a tremendous opportunity to leverage these predictive models on the defensive end of the ball, allowing teams to better anticipate and defend against running plays. With better insights into the opponents running game, defenses can adjust their schemes and personnel to counter the opposing team's offensive strategy, such as by stacking the box or blitzing the quarterback. In addition to the benefits performance analytics provides to the teams, it also helps fans select better fantasy football teams and make more informed betting decisions. This leads to a more engaging and enjoyable experience for the fans, which is critical for the long-term success of the league.

Ultimately, the true beauty of this research lies in its ability to bridge the gap between complex player tracking data and practical on-field strategies. In specific, it's about enhancing the very essence of the game and enriching the broader discourse on sports performance analytics. By pioneering research in this domain, this study is set to propel the analytical capabilities of NFL teams to new heights, providing them with tools that were once considered unimaginable. Ultimately, it contributes to the ongoing evolution of the sport itself, marking a pivotal moment in the history of football and the broader world of sports analytics.

2. Methods

2.1. Data Cleaning and Preprocessing

Data cleaning and preprocessing are critical to ensuring the quality and integrity of machine learning models. The cleaned dataset was obtained by addressing any inconsistencies and handling missing data through imputation techniques. Preprocessing steps included encoding categorical variables using one-hot encoding or label encoding methods and scaling and normalizing numerical variables to ensure they are on comparable scales.

To begin, several variables were removed in order to simulate the beginning of each play. For instance, the dataset include the speed, acceleration, direction, and distance traveled of each player on the field at the beginning of each play. However, this information is not available to the coaching staff prior to the snap of the ball. Therefore, these variables were removed from the datset. Furthermore, some variables may introduce multicollinearity issues since they represent the same information. For instance, the dataset includes the name and id of each player on the field. However, these variables are highly correlated with each other. Therefore, only one of these variables was kept in the dataset.

When the dataset was first obtained, each row represented a single player on the field. However, the goal of this study is to predict the yards gained after a handoff for each play. Therefore, every 22 rows had to be merged into one row (11 players on each side). During this process, any duplicated information/columns were dropped from the dataset. One example of this is the weather during the play. When all 22 rows are merged, it included the weather 22 times, even though weather is only needed once.

In addition to the attempt at removing unnecessary variables and proactively removing potenital multicollienarity issues, variables such as weather, wind speed, and wind direction had multiple levels, some of which represented the same data. For instance, "rainy" and "showers" were considered the same weather type.

There were multiple instances of relationships between inputs on the same variable that existed similar to the example above. Due to this, these weather types were grouped into one category - "rain." This helps to drastically reduce the complexity of the dataset. These variables were also converted to factors for the modeling process.

Variables such as offense personnel had to be converted to dummy variables. The input for each dummy variable was either 0 or 1 depending on whether or not the offense had that personnel on the field. For instance, if the offense had 2 running backs on the field, the input for the "RB" dummy variable would be 2. The input for the "WR" dummy variable would be 0 since there were no wide receivers on the field. This process was repeated for all personnel types.

One additional step that was taken was the randomization of all plays within the dataset. The reason for this randomization is to ensure that each row (play of a game) is independent of one another. Furthermore, the dataset only contains running plays. Therefore, passing plays that occured between the running plays were removed. In other words, all plays are not dependent on the previous outcome. In all, the randomization of the dataset in conjunction with the removal of passing plays ensures that each row is independent of one another.

2.2. Data Exploration

Prior to the development of the predictive models, a comprehensive exploratory data analysis (EDA) was conducted to understand the underlying structure and characteristics of the data. This step involved visualizing the distribution of key variables, identifying patterns and outliers, and exploring correlations and interactions between predictors. Data visualization, through techniques like scatter plots, histograms, and heatmaps, can offer an intuitive understanding of data distributions, correlations, and potential clusters within the dataset. EDA-informed feature selection and engineering strategies highlighted potential predictors that are most informative of yards gained after a handoff.

2.3. Feature Engineering

During preprocessing, we will also undertake feature engineering to create new variables that may have a stronger relationship with the target variable. This could include interaction terms that capture the combined effect of two predictors, polynomial features for capturing non-linear relationships, and domain-specific features that encapsulate strategic elements of the game.

2.4. Model Development and Evaluation

With a clean and prepared dataset, we will then proceed to develop our MLR, GLM, and XGBoost models.

2.5. Multiple Linear Regression (MLR)

Multiple linear regression models predict a continuous repsonse variable using a linear combination of predictors. For our baseline MLR model, we will use the ordinary least squares (OLS) method to estimate the coefficients of our predictor variables. The model is specified as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_j X_{ij} + \epsilon_i$$

where Y_i represents the yards gained after the handoff for the i^{th} observation, X_{ij} is the i^{th} observation on the j^{th} predictor variable (where $j = 1, ..., p$), β_j is the j^{th} coefficient to be estimated corresponding to the proper X_{ij} variable, and ϵ_i is the error term. The j^{th} coefficient, β_j, represents the change in the response variable for a unit change in the X_{ij} predictor variable, while holding all other predictors constant.

The ordinary least squares (OLS) method will be deployed to minimize the sum of the squared differences between the observed and predicted values. The optimization problem can be represented as minimizing the sum of errors squared:

$$\Sigma_{i=0}^{n} \epsilon_i^2 = \Sigma_{i=0}^{n} (Y_i - \hat{Y}_i)^2 = \Sigma_{i=0}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - ... - \beta_j X_{ij})^2$$

Here, $\Sigma_{i=0}^{n}\epsilon_i^2$ represents the sum of the squared residuals, which we aim to minimize. The variable $\hat{Y}_i$ represents the predicted outcome of the model for the i^{th} observation. This approach assumes that the relationship between the independent variables and the dependent variable is linear. To ensure the robustness of our model, we will conduct a series of diagnostic tests:

1. Linearity: We will use scatter plots and residual plots to verify that the relationship between the predictors and the response is linear.
2. Homoscedasticity: We will inspect the residuals to confirm constant variance across all levels of the independent variables. This can be assessed visually using a residual vs. fitted values plot.
3. Independence: The Durbin-Watson test will help in detecting the presence of autocorrelation in the residuals, which should not be present in the data.
4. Normality of Residuals: Normality will be checked using Q-Q plots and statistical tests like the Shapiro-Wilk test.

If any assumptions are violated, we may consider transformations of variables or the use of robust regression techniques.

2.6. Generalized Linear Models (GLM)

GLMs extend the linear model framework to allow for response variables that have error distribution models other than a normal distribution. They are particularly useful when dealing with non-normal response variables, such as count data or binary outcomes. In its general form, a GLM consists of three elements:

1. Random Component: Specifies the probability distribution of the response variable Y, such as normal, binomial, Poisson, or exponential.
2. Systematic Component: A linear predictor $\eta = X\beta$.
3. Link Function: A function g that relates the mean of the response variable $E(Y)$ to the linear predictor η.

The choice of link function is crucial and is typically selected based on the nature of the distribution of the response variable. For instance, a logit link function is used for a binomial distribution, and a log link function is often used for a Poisson distribution.

The likelihood function for a GLM can be written as:

$$L(\beta) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi)$$

where $f(y_i; \theta_i, \phi)$ is the probability function for the i-th observation, θ_i is the parameter of interest (e.g., mean), and ϕ is the dispersion parameter. The goal is to find the values of β that maximize this likelihood function.

2.7. XGBoost

XGBoost stands for eXtreme Gradient Boosting and is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It is a powerful technique that can handle a variety of regression and classification problems. For regression, it can be configured to optimize for different loss functions; the most common for regression being the squared error loss:

$$L(\theta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where y_i are the observed values, and $\hat{y}_i$ are the predicted values.

In XGBoost, each new tree is built to correct the errors made by the previous ones. The algorithm combines weak predictive models to form a strong predictor. The model's complexity is controlled by the regularization term $\Omega(\theta)$ which is a function of the tree structure and the number of leaves. The overall objective function to be minimized is:

$$\text{Obj}(\theta) = L(\theta) + \lambda \sum_k (w_k^2) + \gamma T$$

where w_k represents the leaf weights of the trees, T is the number of leaves, λ is the L2 regularization term on the weights, and γ is the complexity control on the number of leaves. For regression tasks, we can also utilize the quantile loss which is particularly useful for prediction intervals:

$$L_\tau(\theta) = \sum_{i=1}^{n} \left[\tau(y_i - \hat{y}_i)\mathbb{1}_{y_i \geq \hat{y}_i} + (1-\tau)(\hat{y}_i - y_i)\mathbb{1}_{y_i < \hat{y}_i} \right]$$

Here, 1 is an indicator function, and τ is the quantile of interest, allowing us to model different parts of the conditional distribution of the response variable.

XGBoost also provides a feature importance score, which is a metric that quantifies the contribution of each feature to the model's predictive power. This is done by measuring the impact on the model's accuracy each time a feature is used to split the data across all trees.

2.8. Model Performance Metrics

For MLR and GLM, the model's performance will be evaluated using metrics such as the Mean Squared Error (MSE) and the Mean Absolute Error (MAE). For the XGBoost model, along with MSE and MAE, we will assess performance using additional metrics like the R-squared for regression tasks and feature importance scores to understand which variables are most predictive.

2.9. Model Interpretation and Application

The final step will be to interpret the models in the context of NFL games. This will involve translating the statistical outputs into actionable insights for coaches and team strategists, providing recommendations on how to leverage the results for competitive advantage in play-calling and player evaluation.

2.10. Software and Tools

All analyses will be conducted in R, a statistical computing language that provides a wide array of packages for machine learning and data analysis. For MLR and GLM, we will utilize the `stats` package that comes with the R base installation. For our XGBoost model, we will use the `xgboost` package, which is specifically designed for speed and performance. Data manipulation and cleansing will be managed with packages like `dplyr` and `tidyr`, while `ggplot2` will be employed for data visualization to facilitate understanding and interpretation of the data and model outputs. For feature engineering and preprocessing, we will take advantage of `caret` or `recipes`. Hyperparameter tuning can be optimized using the `tune` package, and for cross-validation, the `rsample` package will be employed. The `broom` and `modelr` packages will be useful for tidying model outputs and working with models in a pipeline, respectively. This suite of packages will enable a comprehensive workflow within R for developing, evaluating, and interpreting the predictive models.

2.11. Model Development

To commence our analytical journey, we initiate with Multiple Linear Regression (MLR) as our foundational modeling technique. This approach is not just a stepping stone, but a critical phase in our analysis, offering valuable insights into the relationships between various features of the game and the yards gained. MLR helps discern the linear associations and relative importance of different variables, setting the stage for more complex modeling. Subsequent to this exploratory analysis, we transition to XGBoost, an advanced machine learning technique renowned for its predictive power and efficiency. XGBoost, a gradient boosting framework, is chosen for its ability to handle the dataset's complexity, non-linear relationships, and interactions among variables more adeptly. This shift from MLR to XGBoost embodies our methodological progression, from understanding the foundational relationships in our data to harnessing advanced computational techniques for more accurate and robust predictions in the dynamic and unpredictable context of NFL games.

3. Multiple Linear Regression Model

First, lets start by importing the completely clean and ready to use dataset. This dataset was the result of procedures outlined in the data cleaning and preprocessing section.

```
# Import the dataset
football_data <- read.csv("/Users/samlutz10/Desktop/STAT5405/Final Project/train_ready.csv")

str(football_data)
```

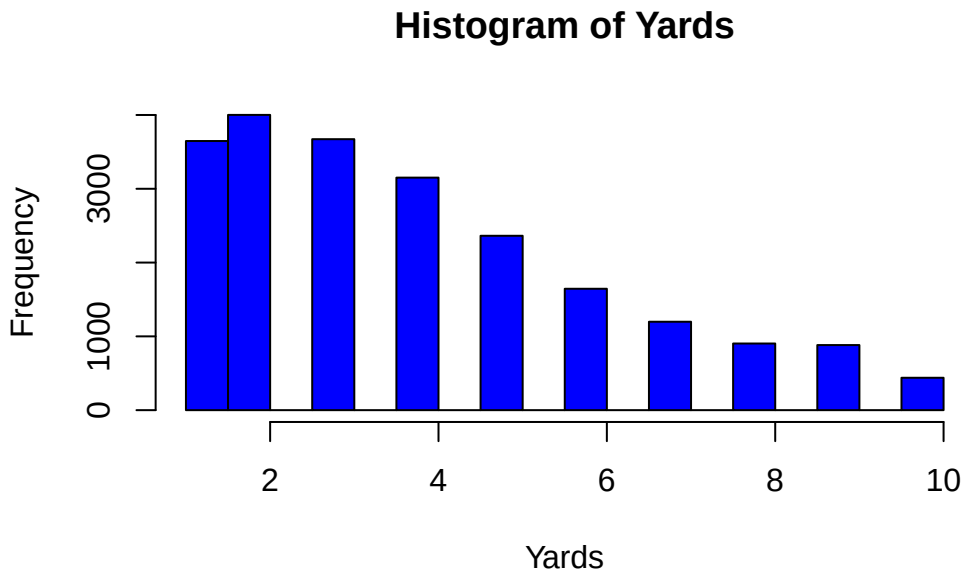
```
'data.frame':   31007 obs. of  21 variables:
 $ Quarter          : int  2 2 4 2 2 1 2 4 4 4 ...
 $ Down             : int  1 1 1 2 1 1 1 1 1 1 ...
 $ Distance         : int  10 10 10 10 10 10 10 10 10 10 ...
 $ OffenseFormation : chr  "SINGLEBACK" "I_FORM" "I_FORM" "SINGLEBACK" ...
 $ RB               : int  2 2 2 1 1 1 1 2 1 ...
 $ TE               : int  1 1 2 1 1 2 1 1 1 1 ...
 $ WR               : int  2 2 1 3 3 2 3 3 1 3 ...
 $ DefendersInTheBox: int  8 7 8 6 7 8 6 7 7 7 ...
 $ Yards            : int  1 12 3 -2 14 4 3 7 7 3 ...
 $ Week             : int  3 9 14 2 2 2 17 11 17 3 ...
 $ GameWeather      : chr  "clear" "overcast" "clear" "clear" ...
 $ Temperature      : int  84 77 39 81 79 79 19 68 71 68 ...
 $ Humidity         : int  25 62 55 45 48 48 36 54 66 42 ...
 $ WindSpeed        : num  5 16 11 5 5 5 12 4 9 13 ...
 $ WindDirection    : chr  "west" "south east" "none" "east" ...
 $ GameHour         : int  15 9 3 3 13 12 2 6 15 12 ...
 $ DL               : int  4 3 4 4 2 4 2 4 4 4 ...
 $ LB               : int  3 4 3 2 4 3 4 2 3 2 ...
 $ BL               : int  4 4 4 5 5 4 5 5 4 5 ...
 $ YardsFromOwnGoal : int  40 84 79 21 86 25 49 25 65 54 ...
 $ ScoreDelta       : int  -7 0 -13 -11 0 7 -11 41 -16 -8 ...
```

Since we are only focused on the most common plays and not breakout runs, we will remove any data of 10 yards or more. This is due to the fact that plays resulting in 10 yards are considered to be just as successful are play resulting in more than 10 yards. In addition, any plays resulting in 0 or negative yards will be removed. This is due to the fact that these plays are considered to be unsuccessful.

```r
# Remove any rows whose yards column is greater than 30
football_data <- football_data[football_data$Yards <= 10 & football_data$Yards > 0, ]
```

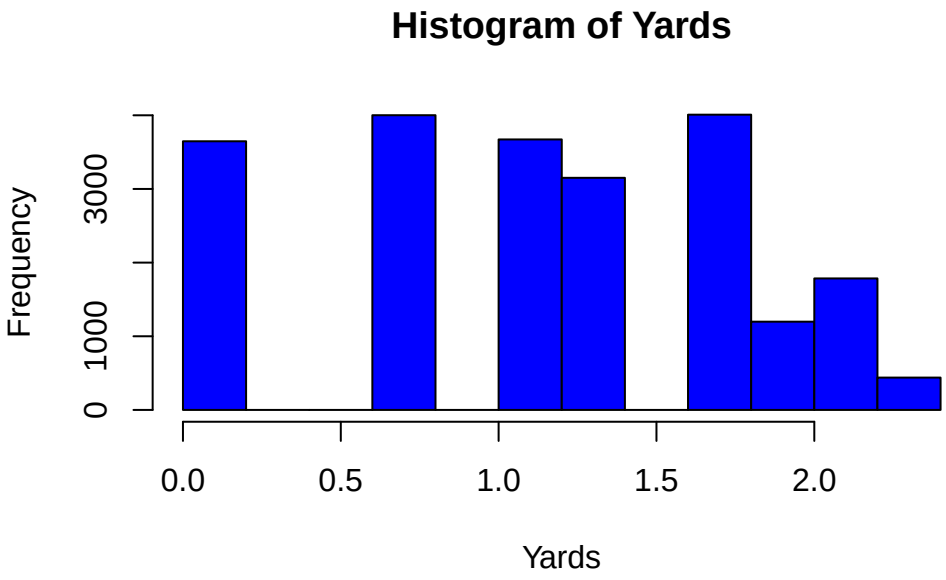
Lets now look at the histogram of Yards to get a better understanding of the distribution of the response variable.

```r
# Histogram of Yards
hist(football_data$Yards, main = "Histogram of Yards", xlab = "Yards", ylab = "Frequency", col = "b
```



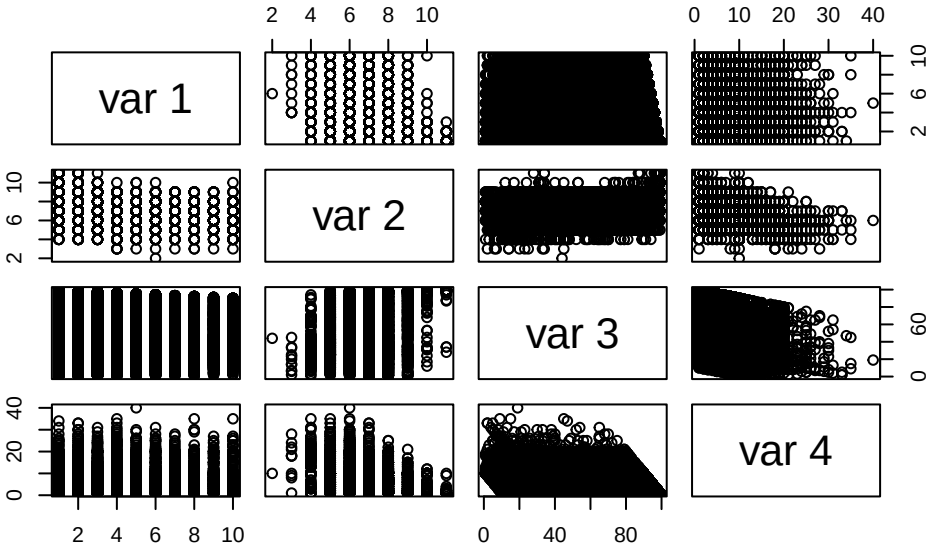
Based on the histogram above, it is apparent that the number of yards is skewed to the right. In order to address this issue, we will take the log of the yards variable. This will help to normalize the distribution of the response variable.

```r
# Histogram of Yards
hist(log(football_data$Yards), main = "Histogram of Yards", xlab = "Yards", ylab = "Frequency", col
```



In addition, lets take a look at some scatterplots of the response variable versus some of the predictors. This will help to gain a better understanding of the relationship between the response variable and the predictors.

```
pairs(cbind(football_data$Yards, football_data$DefendersInTheBox, football_data$YardsFromOwnGoal, f
```



The first model that will be synthesized has only an intercept and no predictors.

```
lm_model_null <- lm(log(Yards) ~ 1, data = football_data)
# summary(lm_model_null)
```

Now the full model using all of the predictors will be created. In this full model, there will be 0 interactions between the predictors.

```r
# Fit the model
lm_model_full <- lm(log(Yards) ~ ., data = football_data)
# summary(lm_model_full)
```

Now lets use the step function to find the best model for the data.

```r
lm_model_step <- step(lm_model_full, direction = "backward")
```

```
Start:  AIC=-17699.97
log(Yards) ~ Quarter + Down + Distance + OffenseFormation + RB +
    TE + WR + DefendersInTheBox + Week + GameWeather + Temperature +
    Humidity + WindSpeed + WindDirection + GameHour + DL + LB +
    BL + YardsFromOwnGoal + ScoreDelta

                    Df Sum of Sq    RSS    AIC
- WindDirection      8     1.415 9504.2 -17713
- GameWeather        4     1.958 9504.8 -17704
- TE                 1     0.001 9502.8 -17702
- LB                 1     0.002 9502.8 -17702
- WindSpeed          1     0.002 9502.8 -17702
- DL                 1     0.003 9502.8 -17702
- Quarter            1     0.029 9502.8 -17702
- BL                 1     0.039 9502.8 -17702
- RB                 1     0.050 9502.9 -17702
- Humidity           1     0.072 9502.9 -17702
- Week               1     0.252 9503.1 -17701
- WR                 1     0.599 9503.4 -17701
<none>                           9502.8 -17700
- ScoreDelta         1     1.144 9504.0 -17699
- GameHour           1     1.598 9504.4 -17698
- Down               1     1.864 9504.7 -17698
- Temperature        1     4.896 9507.7 -17691
- YardsFromOwnGoal   1    25.906 9528.7 -17643
- Distance           1    35.557 9538.4 -17621
- OffenseFormation   8    46.438 9549.2 -17611
- DefendersInTheBox  1    86.488 9589.3 -17506

Step:  AIC=-17712.75
log(Yards) ~ Quarter + Down + Distance + OffenseFormation + RB +
    TE + WR + DefendersInTheBox + Week + GameWeather + Temperature +
    Humidity + WindSpeed + GameHour + DL + LB + BL + YardsFromOwnGoal +
    ScoreDelta

                    Df Sum of Sq    RSS    AIC
- GameWeather        4     1.517 9505.7 -17717
- TE                 1     0.000 9504.2 -17715
- LB                 1     0.001 9504.2 -17715
- DL                 1     0.003 9504.2 -17715
- Humidity           1     0.010 9504.2 -17715
- Quarter            1     0.029 9504.2 -17715
- BL                 1     0.040 9504.3 -17715
```

```
- RB                 1     0.048 9504.3 -17715
- WindSpeed          1     0.129 9504.4 -17714
- Week               1     0.186 9504.4 -17714
- WR                 1     0.618 9504.8 -17713
<none>                           9504.2 -17713
- ScoreDelta         1     1.116 9505.3 -17712
- GameHour           1     1.610 9505.8 -17711
- Down               1     1.826 9506.0 -17711
- Temperature        1     4.576 9508.8 -17704
- YardsFromOwnGoal   1    25.791 9530.0 -17656
- Distance           1    35.502 9539.7 -17634
- OffenseFormation   8    46.662 9550.9 -17623
- DefendersInTheBox  1    86.611 9590.8 -17519

Step:  AIC=-17717.3
log(Yards) ~ Quarter + Down + Distance + OffenseFormation + RB +
    TE + WR + DefendersInTheBox + Week + Temperature + Humidity +
    WindSpeed + GameHour + DL + LB + BL + YardsFromOwnGoal +
    ScoreDelta

                    Df Sum of Sq    RSS    AIC
- TE                 1     0.000 9505.7 -17719
- WindSpeed          1     0.001 9505.7 -17719
- LB                 1     0.002 9505.7 -17719
- DL                 1     0.003 9505.7 -17719
- Quarter            1     0.027 9505.8 -17719
- RB                 1     0.039 9505.8 -17719
- BL                 1     0.042 9505.8 -17719
- Week               1     0.166 9505.9 -17719
- Humidity           1     0.431 9506.2 -17718
- WR                 1     0.648 9506.4 -17718
<none>                           9505.7 -17717
- ScoreDelta         1     1.085 9506.8 -17717
- GameHour           1     1.624 9507.4 -17716
- Down               1     1.800 9507.5 -17715
- Temperature        1     4.231 9510.0 -17710
- YardsFromOwnGoal   1    25.718 9531.5 -17661
- Distance           1    35.409 9541.1 -17639
- OffenseFormation   8    46.637 9552.4 -17628
- DefendersInTheBox  1    86.888 9592.6 -17523

Step:  AIC=-17719.3
log(Yards) ~ Quarter + Down + Distance + OffenseFormation + RB +
    WR + DefendersInTheBox + Week + Temperature + Humidity +
    WindSpeed + GameHour + DL + LB + BL + YardsFromOwnGoal +
    ScoreDelta

                    Df Sum of Sq    RSS    AIC
- WindSpeed          1     0.001 9505.7 -17721
- LB                 1     0.002 9505.7 -17721
- DL                 1     0.003 9505.7 -17721
- Quarter            1     0.027 9505.8 -17721
```

```
- BL                 1     0.042 9505.8 -17721
- RB                 1     0.083 9505.8 -17721
- Week               1     0.167 9505.9 -17721
- Humidity           1     0.433 9506.2 -17720
<none>                           9505.7 -17719
- ScoreDelta         1     1.085 9506.8 -17719
- GameHour           1     1.624 9507.4 -17718
- Down               1     1.801 9507.5 -17717
- WR                 1     2.395 9508.1 -17716
- Temperature        1     4.232 9510.0 -17712
- YardsFromOwnGoal   1    25.728 9531.5 -17663
- Distance           1    35.419 9541.2 -17641
- OffenseFormation   8    48.132 9553.9 -17626
- DefendersInTheBox  1    87.018 9592.8 -17524

Step:  AIC=-17721.3
log(Yards) ~ Quarter + Down + Distance + OffenseFormation + RB +
    WR + DefendersInTheBox + Week + Temperature + Humidity +
    GameHour + DL + LB + BL + YardsFromOwnGoal + ScoreDelta

                    Df Sum of Sq    RSS    AIC
- LB                 1     0.002 9505.7 -17723
- DL                 1     0.003 9505.7 -17723
- Quarter            1     0.027 9505.8 -17723
- BL                 1     0.042 9505.8 -17723
- RB                 1     0.083 9505.8 -17723
- Week               1     0.166 9505.9 -17723
- Humidity           1     0.462 9506.2 -17722
<none>                           9505.7 -17721
- ScoreDelta         1     1.086 9506.8 -17721
- GameHour           1     1.624 9507.4 -17720
- Down               1     1.800 9507.5 -17719
- WR                 1     2.395 9508.1 -17718
- Temperature        1     4.288 9510.0 -17714
- YardsFromOwnGoal   1    25.730 9531.5 -17665
- Distance           1    35.433 9541.2 -17643
- OffenseFormation   8    48.139 9553.9 -17628
- DefendersInTheBox  1    87.022 9592.8 -17526

Step:  AIC=-17723.29
log(Yards) ~ Quarter + Down + Distance + OffenseFormation + RB +
    WR + DefendersInTheBox + Week + Temperature + Humidity +
    GameHour + DL + BL + YardsFromOwnGoal + ScoreDelta

                    Df Sum of Sq    RSS    AIC
- Quarter            1     0.027 9505.8 -17725
- RB                 1     0.083 9505.8 -17725
- DL                 1     0.109 9505.8 -17725
- Week               1     0.166 9505.9 -17725
- Humidity           1     0.462 9506.2 -17724
<none>                           9505.7 -17723
- ScoreDelta         1     1.084 9506.8 -17723
```

```
- GameHour           1     1.627 9507.4 -17722
- Down               1     1.801 9507.5 -17721
- WR                 1     2.394 9508.1 -17720
- BL                 1     3.354 9509.1 -17718
- Temperature        1     4.293 9510.0 -17716
- YardsFromOwnGoal   1    25.750 9531.5 -17667
- Distance           1    35.442 9541.2 -17645
- OffenseFormation   8    48.457 9554.2 -17629
- DefendersInTheBox  1    87.158 9592.9 -17528

Step:  AIC=-17725.23
log(Yards) ~ Down + Distance + OffenseFormation + RB + WR + DefendersInTheBox +
    Week + Temperature + Humidity + GameHour + DL + BL + YardsFromOwnGoal +
    ScoreDelta

                   Df Sum of Sq    RSS    AIC
- RB                1     0.085 9505.9 -17727
- DL                1     0.109 9505.9 -17727
- Week              1     0.168 9505.9 -17727
- Humidity          1     0.464 9506.2 -17726
<none>                        9505.8 -17725
- ScoreDelta        1     1.071 9506.8 -17725
- GameHour          1     1.618 9507.4 -17724
- Down              1     1.819 9507.6 -17723
- WR                1     2.409 9508.2 -17722
- BL                1     3.370 9509.1 -17720
- Temperature       1     4.297 9510.1 -17718
- YardsFromOwnGoal  1    25.738 9531.5 -17669
- Distance          1    35.537 9541.3 -17646
- OffenseFormation  8    48.477 9554.2 -17631
- DefendersInTheBox 1    87.202 9593.0 -17530

Step:  AIC=-17727.04
log(Yards) ~ Down + Distance + OffenseFormation + WR + DefendersInTheBox +
    Week + Temperature + Humidity + GameHour + DL + BL + YardsFromOwnGoal +
    ScoreDelta

                   Df Sum of Sq    RSS    AIC
- DL                1     0.106 9506.0 -17729
- Week              1     0.169 9506.0 -17729
- Humidity          1     0.450 9506.3 -17728
<none>                        9505.9 -17727
- ScoreDelta        1     1.061 9506.9 -17727
- GameHour          1     1.617 9507.5 -17725
- Down              1     1.832 9507.7 -17725
- WR                1     2.333 9508.2 -17724
- BL                1     3.338 9509.2 -17721
- Temperature       1     4.302 9510.2 -17719
- YardsFromOwnGoal  1    25.798 9531.6 -17670
- Distance          1    35.623 9541.5 -17648
- OffenseFormation  8    48.393 9554.2 -17633
- DefendersInTheBox 1    87.237 9593.1 -17532
```

```
Step:  AIC=-17728.8
log(Yards) ~ Down + Distance + OffenseFormation + WR + DefendersInTheBox +
    Week + Temperature + Humidity + GameHour + BL + YardsFromOwnGoal +
    ScoreDelta

                    Df Sum of Sq    RSS    AIC
- Week               1     0.147 9506.1 -17730
- Humidity           1     0.457 9506.4 -17730
<none>                           9506.0 -17729
- ScoreDelta         1     1.039 9507.0 -17728
- GameHour           1     1.618 9507.6 -17727
- Down               1     1.810 9507.8 -17727
- WR                 1     2.313 9508.3 -17726
- BL                 1     3.242 9509.2 -17723
- Temperature        1     4.211 9510.2 -17721
- YardsFromOwnGoal   1    25.771 9531.7 -17672
- Distance           1    35.525 9541.5 -17650
- OffenseFormation   8    48.294 9554.3 -17635
- DefendersInTheBox  1    87.147 9593.1 -17534

Step:  AIC=-17730.46
log(Yards) ~ Down + Distance + OffenseFormation + WR + DefendersInTheBox +
    Temperature + Humidity + GameHour + BL + YardsFromOwnGoal +
    ScoreDelta

                    Df Sum of Sq    RSS    AIC
- Humidity           1     0.410 9506.5 -17732
<none>                           9506.1 -17730
- ScoreDelta         1     1.066 9507.2 -17730
- GameHour           1     1.614 9507.7 -17729
- Down               1     1.798 9507.9 -17728
- WR                 1     2.352 9508.5 -17727
- BL                 1     3.240 9509.3 -17725
- Temperature        1     5.340 9511.4 -17720
- YardsFromOwnGoal   1    25.706 9531.8 -17674
- Distance           1    35.518 9541.6 -17652
- OffenseFormation   8    48.268 9554.4 -17637
- DefendersInTheBox  1    87.446 9593.6 -17534

Step:  AIC=-17731.53
log(Yards) ~ Down + Distance + OffenseFormation + WR + DefendersInTheBox +
    Temperature + GameHour + BL + YardsFromOwnGoal + ScoreDelta

                    Df Sum of Sq    RSS    AIC
<none>                           9506.5 -17732
- ScoreDelta         1     1.005 9507.5 -17731
- GameHour           1     1.594 9508.1 -17730
- Down               1     1.826 9508.3 -17729
- WR                 1     2.308 9508.8 -17728
- BL                 1     3.234 9509.8 -17726
- Temperature        1     4.954 9511.5 -17722
```

```
- YardsFromOwnGoal   1    25.767 9532.3 -17675
- Distance           1    35.677 9542.2 -17652
- OffenseFormation   8    48.495 9555.0 -17638
- DefendersInTheBox  1    87.145 9593.7 -17536
```

```r
summary(lm_model_step)
```

```
Call:
lm(formula = log(Yards) ~ Down + Distance + OffenseFormation +
    WR + DefendersInTheBox + Temperature + GameHour + BL + YardsFromOwnGoal +
    ScoreDelta, data = football_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.70141 -0.45056  0.06292  0.50516  1.57950

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.8300909  0.6691300   2.735 0.006243 **
Down                       0.0161672  0.0079365   2.037 0.041657 *
Distance                   0.0131352  0.0014588   9.004  < 2e-16 ***
OffenseFormationEMPTY      0.0436278  0.6783426   0.064 0.948720
OffenseFormationI_FORM    -0.1140737  0.6635919  -0.172 0.863515
OffenseFormationJUMBO     -0.4506772  0.6645768  -0.678 0.497689
OffenseFormationPISTOL    -0.1008327  0.6639361  -0.152 0.879290
OffenseFormationSHOTGUN   -0.1508435  0.6634845  -0.227 0.820153
OffenseFormationSINGLEBACK -0.1118227  0.6634980  -0.169 0.866164
OffenseFormationUNKNOWN   -0.1587977  0.7417972  -0.214 0.830493
OffenseFormationWILDCAT   -0.0488034  0.6691285  -0.073 0.941858
WR                        -0.0205143  0.0089567  -2.290 0.022009 *
DefendersInTheBox         -0.0962724  0.0068411 -14.073  < 2e-16 ***
Temperature               -0.0008836  0.0002634  -3.355 0.000794 ***
GameHour                   0.0019662  0.0010329   1.904 0.056986 .
BL                         0.0286437  0.0105657   2.711 0.006713 **
YardsFromOwnGoal          -0.0014363  0.0001877  -7.652 2.06e-14 ***
ScoreDelta                -0.0006387  0.0004226  -1.511 0.130711
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6634 on 21604 degrees of freedom
  (279 observations deleted due to missingness)
Multiple R-squared:  0.04789,   Adjusted R-squared:  0.04714
F-statistic: 63.92 on 17 and 21604 DF,  p-value: < 2.2e-16
```

Running this reduced model yields the following results:

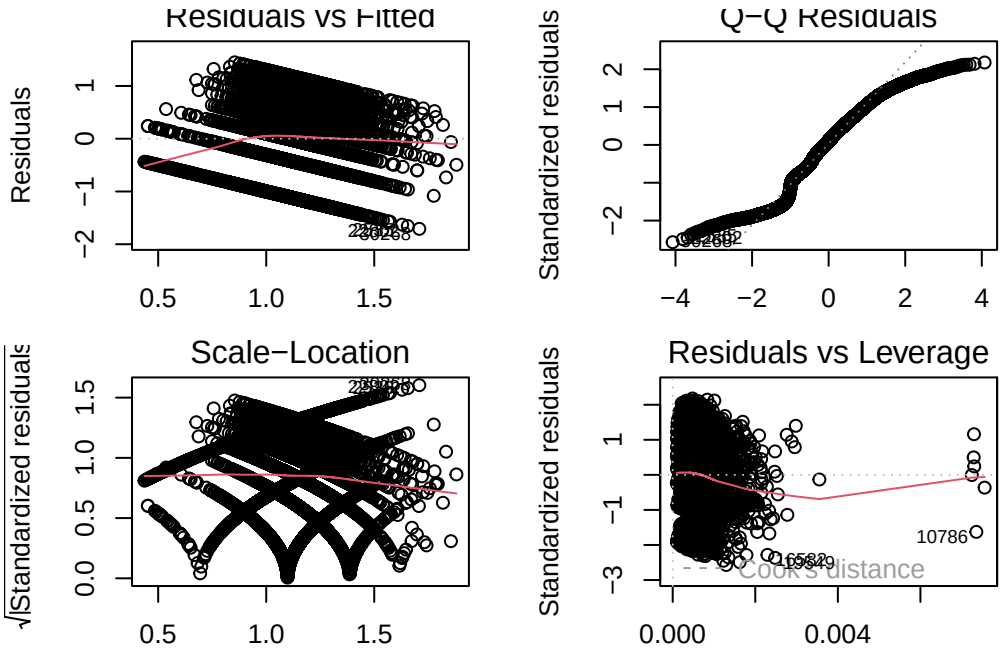
```r
reduced_model <- lm(log(Yards) ~ Distance + RB + WR + DefendersInTheBox^2 + Temperature + GameHour
summary(reduced_model)
```

```
Call:
```

```
lm(formula = log(Yards) ~ Distance + RB + WR + DefendersInTheBox^2 +
    Temperature + GameHour + DL + LB + YardsFromOwnGoal^2, data = football_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.70860 -0.45847  0.06848  0.50619  1.44826

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.120e+00  8.627e-02  24.574  < 2e-16 ***
Distance           1.266e-02  1.275e-03   9.929  < 2e-16 ***
RB                -1.792e-05  1.231e-02  -0.001  0.99884
WR                -1.878e-02  9.028e-03  -2.081  0.03748 *
DefendersInTheBox -1.017e-01  6.702e-03 -15.176  < 2e-16 ***
Temperature       -8.982e-04  2.643e-04  -3.398  0.00068 ***
GameHour           2.158e-03  1.031e-03   2.094  0.03630 *
DL                -3.407e-02  1.100e-02  -3.098  0.00195 **
LB                -3.416e-02  1.075e-02  -3.177  0.00149 **
YardsFromOwnGoal  -1.642e-03  1.868e-04  -8.791  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.665 on 21612 degrees of freedom
  (279 observations deleted due to missingness)
Multiple R-squared:  0.04282,   Adjusted R-squared:  0.04242
F-statistic: 107.4 on 9 and 21612 DF,  p-value: < 2.2e-16
```

By creating a Residuals vs Fitted plot, a Q-Q Residuals plot, a Scale-Location plot, and a Residuals vs Leverage plot, the assumptions of the model can be checked.

```r
par(mfrow = c(2, 2))
old.par = par(mar = c(3, 4, 1, 2))
plot(reduced_model)
```

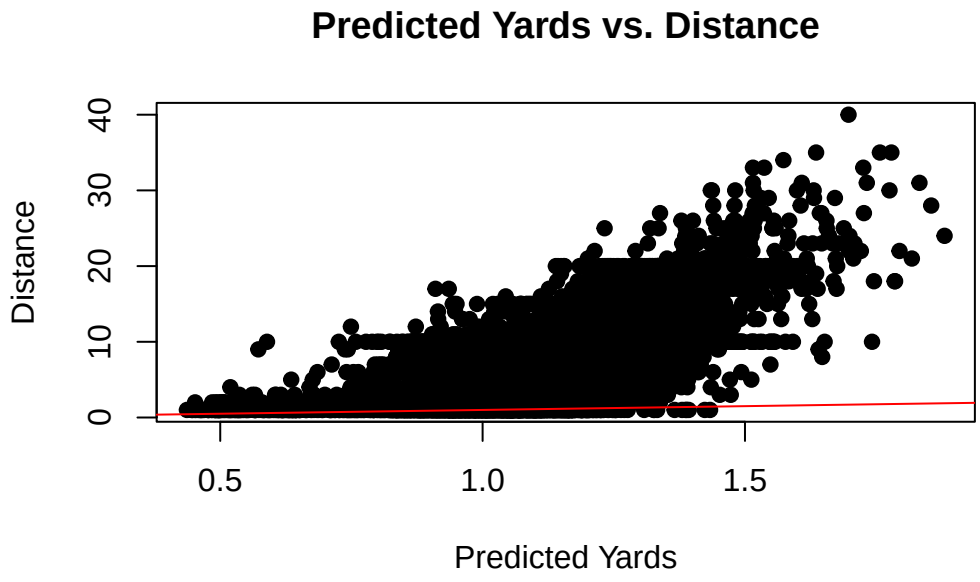


Based on the plots above, the normality assumption is not violated, as the Q-Q Residuals plot shows that the residuals are mainly normally distributed.

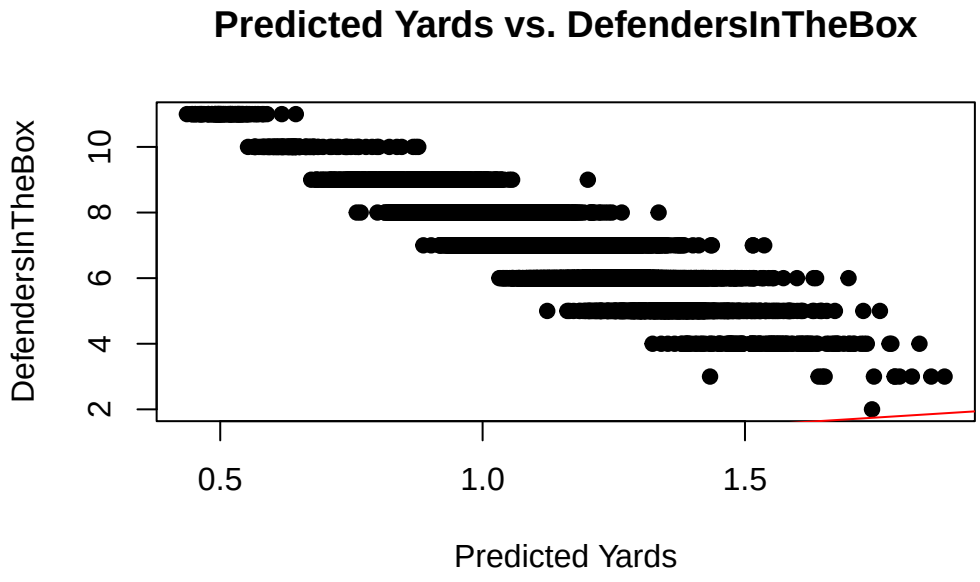
In order to gain a better understanding of how these variables affect the predicted yards, scatter plots of the predicted yards versus each of the predictors will be created. However, the only scatter plots that will be synthesized are those that have the three highest t values: Distance, DefendersInTheBox, and YardsFromOwnGoal.

```
football_data$predicted_yards <- predict(reduced_model, football_data)
```

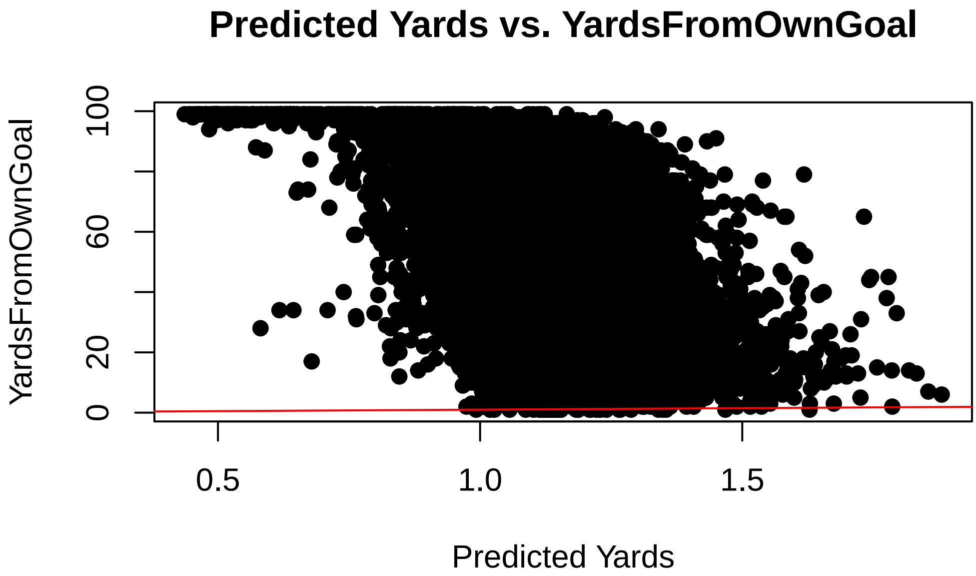
```
# Scatter plot of predicted yards vs. Distance
plot(football_data$predicted_yards, football_data$Distance, main = "Predicted Yards vs. Distance",
abline(0, 1, col = "red")  # Adds a 45-degree line
```




```r
# Scatter plot of predicted yards vs. DefendersInTheBox
plot(football_data$predicted_yards, football_data$DefendersInTheBox, main = "Predicted Yards vs. De
abline(0, 1, col = "red")  # Adds a 45-degree line
```




```r
# Scatter plot of predicted yards vs. YardsFromOwnGoal
plot(football_data$predicted_yards, football_data$YardsFromOwnGoal, main = "Predicted Yards vs. Yar
abline(0, 1, col = "red")  # Adds a 45-degree line
```



4. GLM with Gamma Response

The next model that will be synthesized is a GLM with a gamma response. The gamma distribution is a two-parameter family of continuous probability distributions. The gamma distribution is a generalization of the exponential distribution. The gamma distribution is frequently used to model right-skewed data. The gamma distribution is a two-parameter family of continuous probability distributions. The gamma distribution is a generalization of the exponential distribution. The gamma distribution is frequently used to model right-skewed data. The gamma distribution is a two-parameter family of continuous probability distributions. The gamma distribution is a generalization of the exponential distribution. The gamma distribution is frequently used to model right-skewed data. The gamma distribution is a two-parameter family of continuous probability distributions. The gamma distribution is a generalization of the exponential distribution. The gamma distribution is frequently used to model right-skewed data. The gamma distribution is a two-parameter family of continuous probability distributions. The gamma distribution is a generalization of the exponential distribution. The gamma distribution is frequently used to model right-skewed data. The gamma distribution is a two-parameter family of continuous probability distributions. The gamma distribution is a generalization of the exponential distribution. The gamma distribution is frequently used to model right-skewed data.

Lets start by removing any rows whose yards column is negative. This is because the gamma distribution cannot have negative values.

```
# Remove any rows whose yards column is negative
football_data <- football_data[football_data$Yards >= 0.1, ]
football_data <- football_data[football_data$Yards <= 10, ]
```

```
# Fit the model
glm_model_gamma <- glm(Yards ~ ., data = football_data, family = Gamma(link = "log"))
```

Now lets use the step function to find the best model for the data.

```
#glm_model_step <- step(glm_model_gamma, direction = "backward")
# summary(glm_model_step)
```

Running this reduced model yields the following results:

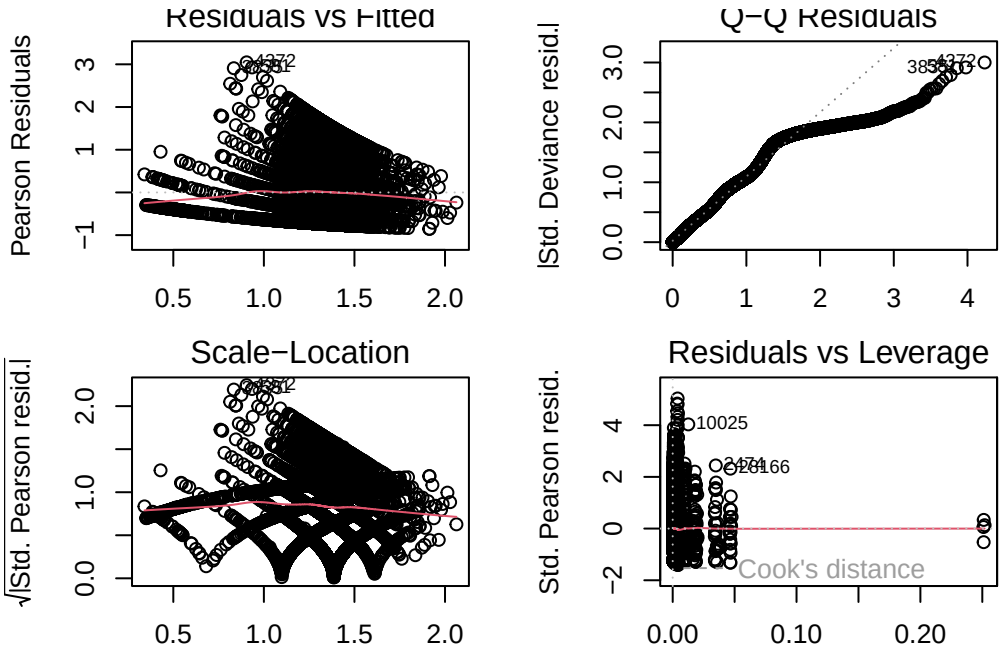
```
reduced_model <- glm(Yards ~ YardsFromOwnGoal + Distance + RB + WR + DefendersInTheBox + BL, data=f
summary(reduced_model)
```

```
Call:
glm(formula = Yards ~ YardsFromOwnGoal + Distance + RB + WR +
    DefendersInTheBox + BL, family = Gamma(link = "log"), data = football_data)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.8518606  0.0767537  24.127  < 2e-16 ***
YardsFromOwnGoal  -0.0015937  0.0001702  -9.365  < 2e-16 ***
Distance           0.0122863  0.0011656  10.541  < 2e-16 ***
RB                 0.0008371  0.0112653   0.074 0.940768
WR                -0.0143714  0.0082703  -1.738 0.082276 .
DefendersInTheBox -0.0945131  0.0061146 -15.457  < 2e-16 ***
BL                 0.0328700  0.0095367   3.447 0.000569 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.3702443)

    Null deviance: 9063.4  on 21621  degrees of freedom
Residual deviance: 8725.7  on 21615  degrees of freedom
  (279 observations deleted due to missingness)
AIC: 92734

Number of Fisher Scoring iterations: 5
```

By creating a Residuals vs Fitted plot, a Q-Q Residuals plot, a Scale-Location plot, and a Residuals vs Leverage plot, the assumptions of the model can be checked.

```
par(mfrow = c(2, 2))
old.par = par(mar = c(3, 4, 1, 2))
plot(glm_model_gamma)
```

```
Warning: not plotting observations with leverage one:
  19377
```



5. XGBoost Model

Moving now into the XGBoost model, the data was split into a training set and a testing set in order to evaluate performance metrics. The training set will be used to train the model, while the testing set will be used to test the model. The training set will be 80% of the data, while the testing set will be 20% of the data and randomly selected.

```
library(fastDummies)
```

```
Thank you for using fastDummies!

To acknowledge our work, please cite the package:

Kaplan, J. & Schlegel, B. (2023). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from C
```

```
# Convert specified factor columns to dummy variables
football_data <- dummy_cols(football_data,
                            select_columns = c("WindDirection", "GameWeather", "OffenseFormation"),
                            remove_selected_columns = TRUE)

# Split the data into a training set and a testing set
set.seed(123457)
train.prop <- 0.80
strats <- football_data$Yards
rr <- split(1:length(strats), strats)
idx <- sort(as.numeric(unlist(sapply(rr,
        function(x) sample(x, length(x)*train.prop)))))
train.set <- football_data[idx, ]
test.set <- football_data[-idx, ]
```

The average of yards in both the training and testing sets are computed to ensure they are about the same. In addition, both of those should be similar to that of the original dataset. This process will help to ensure that the training and testing sets are representative of the entire dataset.

```
# Check that the average of yards is about the same in both the training and testing sets
(ave_train <- mean(train.set$Yards))
```

```
[1] 3.875942
```

```
(ave_test <- mean(test.set$Yards))
```

```
[1] 3.877081
```

```
(ave_all <- mean(football_data$Yards))
```

```
[1] 3.87617
```

Based on the results above, the average yards is about the same in both the training and testing sets. In addition, both of those are similar to that of the original dataset.

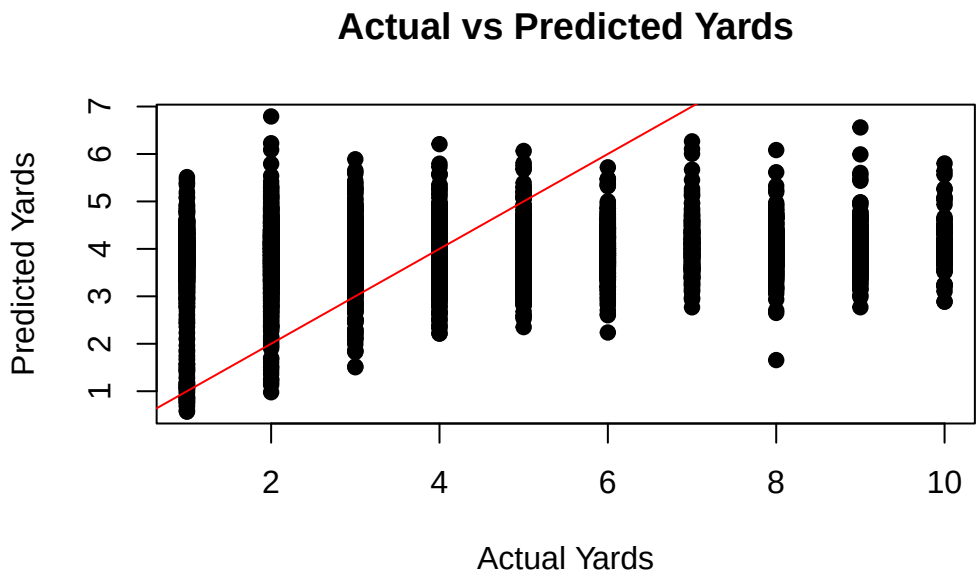
Now that the data has been split into a training set and a testing set, development of the XGBoost model can begin.

```
# Load the xgboost package
library(xgboost)

# Prepare the data for XGBoost
# Convert data to matrix format as xgboost works with the matrix
train_matrix <- as.matrix(train.set[, -which(names(train.set) == "Yards")])
test_matrix <- as.matrix(test.set[, -which(names(test.set) == "Yards")])

# Create DMatrices
dtrain <- xgb.DMatrix(data = train_matrix, label = train.set$Yards)
dtest <- xgb.DMatrix(data = test_matrix, label = test.set$Yards)

# Set XGBoost parameters
params <- list(
    booster = "gbtree",
    objective = "reg:squarederror",  # Objective function for regression
    eta = 0.1,                       # Learning rate
    max_depth = 6,                   # Depth of trees
    subsample = 0.5,                 # Subsampling of the training data
    colsample_bytree = 0.5           # Subsampling of features
)

# Number of boosting rounds
nrounds <- 100

# Train the model
xgb_model <- xgb.train(params = params, data = dtrain, nrounds = nrounds)
```

```r
# Predicting
xgb_predictions <- predict(xgb_model, dtest)

true_values <- test.set$Yards
```

A scatter plot of the actual versus predicted values can provide a clear visual indication of how well the model is performing:

```r
# Scatter plot of actual vs. predicted values
plot(true_values, xgb_predictions, main = "Actual vs Predicted Yards", xlab = "Actual Yards", ylab
abline(0, 1, col = "red")  # Adds a 45-degree line
```



As we can see, the model is performing well, but fails to predict carries over 20 yards. This is likely due to the fact that there are very few carries over 20 yards in the dataset. We can check the summary statistics of yards to confirm this.

```r
summary(football_data$Yards)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   3.876   5.000  10.000
```

75% of the carries 6 yards or less, while 50% of the carries are 3 yards or less. This is likely the reason why the model is failing to predict carries over 20 yards.

Our goal is to determine what plays a team should run on, therefore, because over 75% of our data consists of carries 6 yards or less and our model fails to predict a run of over 20 yards, we will remove all carries over 20 yards from the dataset. This will help to ensure that the model is predicting the majority of the carries in the dataset.

```r
# Remove all carries over 20 yards
football_data <- football_data[football_data$Yards <= 20, ]
```

Now that the data has been cleaned, we can re-run the XGBoost model:

```
# Split the data into a training set and a testing set
set.seed(123457)
train.prop <- 0.80
strats <- football_data$Yards
rr <- split(1:length(strats), strats)
idx <- sort(as.numeric(unlist(sapply(rr,
        function(x) sample(x, length(x)*train.prop)))))
train.set <- football_data[idx, ]
test.set <- football_data[-idx, ]

# Check that the average of yards is about the same in both the training and testing sets
(ave_train <- mean(train.set$Yards))
```

```
[1] 3.875942
```

```
(ave_test <- mean(test.set$Yards))
```

```
[1] 3.877081
```

```
(ave_all <- mean(football_data$Yards))
```

```
[1] 3.87617
```

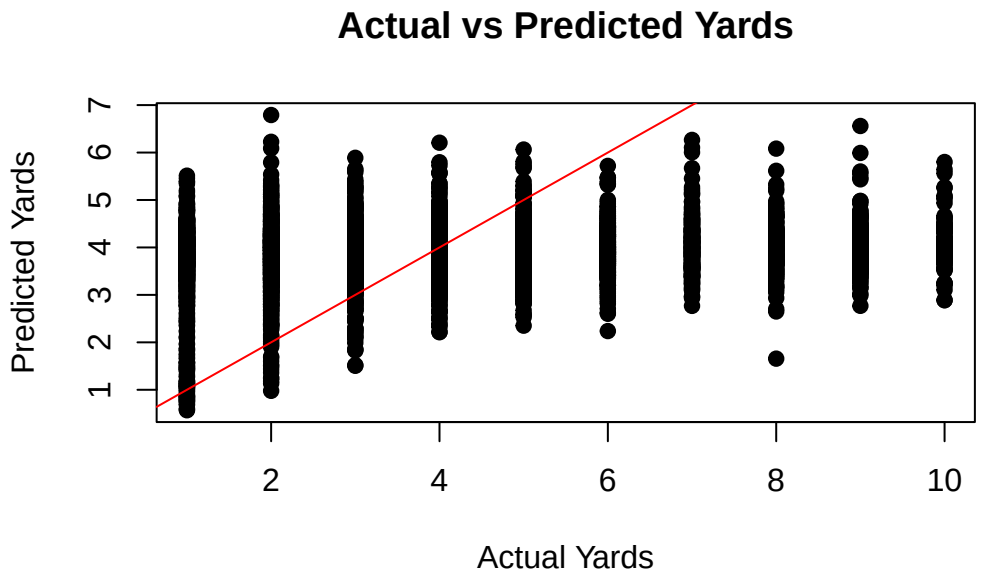
```
# Load the xgboost package
library(xgboost)

# Prepare the data for XGBoost
# Convert data to matrix format as xgboost works with the matrix
train_matrix <- as.matrix(train.set[, -which(names(train.set) == "Yards")])
test_matrix <- as.matrix(test.set[, -which(names(test.set) == "Yards")])

# Create DMatrices
dtrain <- xgb.DMatrix(data = train_matrix, label = train.set$Yards)
dtest <- xgb.DMatrix(data = test_matrix, label = test.set$Yards)

# Set XGBoost parameters
params <- list(
    booster = "gbtree",
    objective = "reg:squarederror",  # Objective function for regression
    eta = 0.1,                       # Learning rate
    max_depth = 6,                   # Depth of trees
    subsample = 0.5,                 # Subsampling of the training data
    colsample_bytree = 0.5           # Subsampling of features
)

# Number of boosting rounds
nrounds <- 100
```

```
# Train the model
xgb_model <- xgb.train(params = params, data = dtrain, nrounds = nrounds)

# Predicting
xgb_predictions <- predict(xgb_model, dtest)

# Evaluate the model
# For example, using Root Mean Squared Error (RMSE)
true_values <- test.set$Yards
rmse <- sqrt(mean((true_values - xgb_predictions)^2))
print(paste("RMSE:", rmse))
```

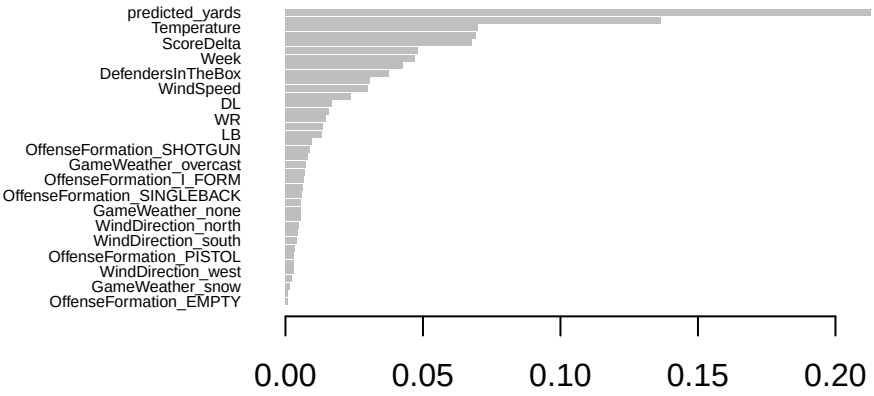
```
[1] "RMSE: 2.35442254039298"
```

Before commenting on the results, lets take a look at some plots to get a better understanding of the model's performance.

```
# Scatter plot of actual vs. predicted values
plot(true_values, xgb_predictions, main = "Actual vs Predicted Yards", xlab = "Actual Yards", ylab
abline(0, 1, col = "red")  # Adds a 45-degree line
```



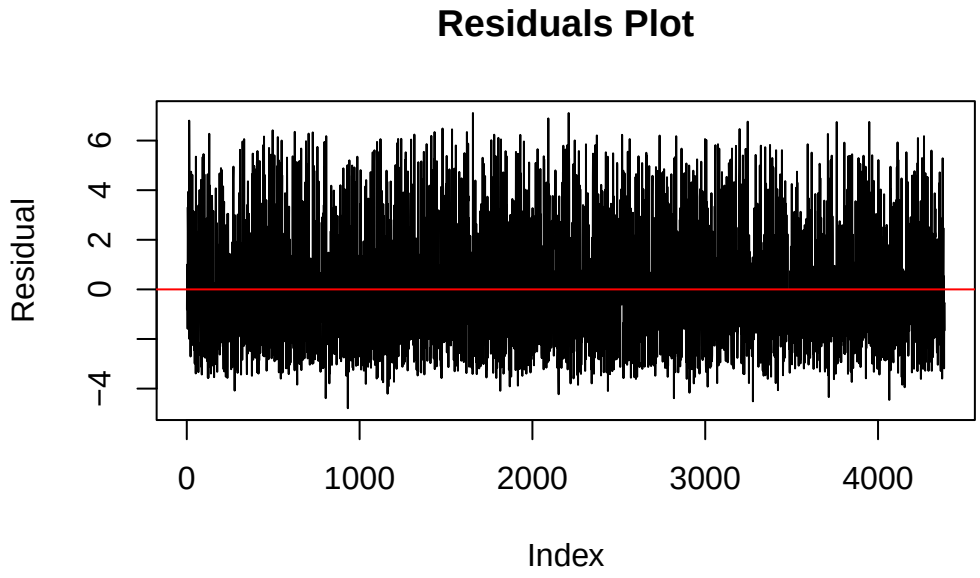

```
# Variable importance plot
importance_matrix <- xgb.importance(feature_names = colnames(train_matrix), model = xgb_model)
xgb.plot.importance(importance_matrix)
```




```
# Mean Absolute Error
mae <- mean(abs(true_values - xgb_predictions))
print(paste("MAE:", mae))
```

```
[1] "MAE: 1.90529614272384"
```

```
# Residuals plot
residuals <- true_values - xgb_predictions
plot(residuals, type = "l", main = "Residuals Plot", xlab = "Index", ylab = "Residual")
abline(h = 0, col = "red")
```



References

6. Bibliography styles

Here are two sample references: **(author?)** [1] **(author?)** [2].

By default, natbib will be used with the `authoryear` style, set in `classoption` variable in YAML. You can sets extra options with `natbiboptions` variable in YAML header. Example

`natbiboptions: longnamesfirst,angle,semicolon`

There are various more specific bibliography styles available at https://support.stmdocs.in/wiki/index.php?title=Model-wise_bibliographic_style_files. To use one of these, add it in the header using, for example, `biblio-style: model1-num-names`.

6.1. Using CSL

If `cite-method` is set to `citeproc` in `elsevier_article()`, then pandoc is used for citations instead of `natbib`. In this case, the `csl` option is used to format the references. By default, this template will provide an appropriate style, but alternative `csl` files are available from https://www.zotero.org/styles?q=elsevier. These can be downloaded and stored locally, or the url can be used as in the example header.

References

[1] R. P. Feynman, F. L. Vernon Jr., The theory of a general quantum system interacting with a linear dissipative system, Annals of Physics 24 (1963) 118–173. doi:10.1016/0003-4916(63)90068-X.

[2] P. A. M. Dirac, The Lorentz transformation and absolute time, Physica 19 (1-–12) (1953) 888–896. doi:10.1016/S0031-8914(53)80099-6.