Decoding the Pre-Snap Puzzle: Analyzing NFL Offensive Strategies to Predict Play Outcomes

Khanh Khuat (ltk30)

I. INTRODUCTION

An NFL offense has, on average, just 40 seconds to plan and execute a play—40 seconds filled with strategy, quick substitutions, and positioning. In that short span, more than 20 players' movements and alignments could influence the yardage gained or even determine the play's success. With thousands of data points generated every game, even subtle pre-snap actions have the potential to uncover powerful trends that can define the outcome of a season.

Understanding how pre-snap behaviors, such as offensive formations, player positioning, and motion, impact the yards gained post-snap has significant implications for both teams and the broader NFL community. Accurate predictions on play outcomes can improve play calling, defensive anticipation, and in-game adjustments. This knowledge is especially valuable given the substantial financial and strategic stakes in the NFL, where even slight advantages can contribute to a team's success over a season. The growing field of sports analytics has seen increasing interest in applying machine learning models to predict outcomes from play-level data, yet relatively few studies have rigorously focused on pre-snap behavior, leaving a gap that this research seeks to address.

Several studies have explored related aspects of NFL play analysis, such as pass and rush tendencies or the impact of player speed on outcomes. For instance, Yurko, Ventura, and Horowitz (2019) utilized machine learning to predict play types, demonstrating that offensive formations and player movements before the snap offer predictive insights for both offensive and defensive outcomes. While these studies underscore the predictive power of player and team behavior, they primarily focus on post-snap events, leaving pre-snap strategies relatively unexplored.

In this project, I aim to address this gap by analyzing the influence of pre-snap behaviors on yards gained, using the 2025 NFL Big Data Bowl dataset. The central research question driving this analysis is: How do pre-snap behaviors and offensive formations impact the total yards gained on a play? This question is further divided into sub-questions exploring the influence of down and distance, offensive formation, quarterback time to throw, pre-snap motion, and play-action on play outcomes. By focusing on these factors, this study seeks to provide actionable insights into how early indicators in a play can enhance offensive strategy and inform defensive anticipation. The findings of this analysis will not only contribute to the evolving field of sports analytics but may also assist NFL coaches and analysts in optimizing game strategies based on pre-snap behavior.

II. METHODOLOGY

The dataset used in this research is derived from the NFL Big Data Bowl 2025 competition, hosted on Kaggle, which focuses on pre-snap behavior to predict and understand NFL team and player tendencies. This competition aims to generate actionable insights by leveraging Next Gen Stats player tracking data, which captures detailed spatial and temporal movements of players before the snap of the ball. The dataset serves as a resource for advancing sports analytics, particularly in exploring how pre-snap behaviors influence post-snap outcomes.

1. Data Collection Setting

1.1. Who collected the data and why?

The dataset was originally collected by the National Football League (NFL) through its partnership with Next Gen Stats, a program that utilizes sensors embedded in players' equipment to track their movements during games. The purpose of this data collection is to enhance player and team performance analysis, provide richer insights for fans, and develop metrics that could be used by NFL teams and broadcasters. This specific dataset was curated to enable participants in the NFL Big Data Bowl to generate novel metrics and actionable insights into the relationships between pre-snap behaviors and post-snap outcomes.

1.2. What variables are included in the dataset?

The dataset includes variables capturing positional data for all 22 players on the field, offensive and defensive formations, motion indicators, player speeds, distances, and contextual game information such as down, distance, and score differential. Our primary response variable is **yardsGained**, representing the total yards gained on a single play. We initially considered a wide array of pre-snap and contextual variables. For the final model, we conducted correlation analyses and ANOVAs to select a set of 15 top features. These final predictor variables include:

- **prePenaltyYardsGained**: Yards gained prior to any penalty enforcement.
- **expectedPointsAdded**: Estimated contribution of the play to the team's scoring expectation.
- **penaltyYards_x**: Penalty yardage associated with the play.
- **yardlineNumber**: The offense's yard line at the start of the play.
- **passResult**: Outcome of a pass (complete, incomplete, interception, etc.).
- **visitorTeamWinProbilityAdded** and **homeTeamWinProbabilityAdded**: Change in each team's win probability due to the play situation.
- **expectedPoints**: Expected points for the possession given the current situation.
- **timeToThrow**: Quarterback's time from snap to pass release.

- **preSnapVisitorTeamWinProbability**, **preSnapHomeTeamWinProbability**: Both teams' win probabilities before the snap.
- **visitorFinalScore** and **preSnapHomeScore**: Visitor's final score and home team's pre-snap score.
- **defensiveTeam** and **homeTeamAbbr**: Team identifiers providing context for offensive/defensive matchups.

Some data cleaning steps included handling missing values via forward-filling player coordinates and recoding variables for clarity. Plays with unrealistic values were removed. This also ensured that any variable representing proprietary codes or contextual measures were defined for interpretability. Ultimately, these procedures standardize the data for robust regression modeling.

2. Where and when was the data collected?

The data was collected during the 2024 NFL season using player tracking technologies deployed across all NFL stadiums. The dataset was subsequently released on Kaggle as part of the NFL Big Data Bowl 2025 competition. A citation for the dataset is as follows:
Michael Lopez, Thompson Bliss, Ally Blake, Paul Mooney, and Addison Howard. NFL Big Data Bowl 2025. [1]

3. Data Cleaning Process

Specific data cleaning steps:

1. Handling Missing Data: Missing values in positional or tracking data were imputed using forward-fill techniques to preserve the continuity of player movement data.
2. Variable Encoding: The response variable "Yards Gained" was encoded into a binary variable for success (e.g., gaining 5 or more yards) to simplify certain analyses. This decision aligns with the goal of assessing play efficiency based on pre-snap behaviors.
3. Outlier Removal: Plays with unrealistic values (e.g., exceedingly high player speeds) were identified and removed to prevent skewed analysis.
4. Data Transformation: Positional data were converted from raw coordinates into relative distances from the line of scrimmage to standardize across different game contexts.

4. Dataset Preparation

The original dataset contained approximately 60 million observations across 122 variables. Due to the computational limitations of handling such a large dataset, a **random sampling approach** was employed to create a manageable subset of the data for analysis. A **5% sample** was extracted, resulting in a working dataset with **approximately 3 million observations** while maintaining the distribution and diversity of the original data.

5. Feature Selection

Given the complexity and dimensionality of the dataset, **stepwise feature selection using AIC (Akaike Information Criterion)** was applied to identify the most relevant predictors for the response variable yardsGained. This approach iteratively adds or removes variables from the model based on their contribution to reducing the AIC score, ensuring an optimal trade-off between model fit and complexity.

6. Model Training

Following the feature selection process, I ultimately fitted a multiple linear regression model in R using the top 15 selected variables. Key aspects of the model training process included:

- **Response Variable:**
  yardsGained (the total yards gained on a play).

- **Predictors:**
  The 15 predictors chosen after feature selection included both numeric and categorical variables (e.g., *prePenaltyYardsGained*, *expectedPointsAdded*, *timeToThrow*, *passResult*, *defensiveTeam*). Each predictor was carefully encoded and, where necessary, scaled or transformed to meet model assumptions.

- **Data Subsampling:**
  To ensure computational efficiency without compromising representativeness, approximately 5% of the full dataset was randomly sampled. This subsample maintained the distributional characteristics of the original dataset, providing a manageable yet robust dataset for training.

- **Model Fitting and Evaluation in R:**
  The multiple linear regression model was fit using standard R functions (e.g., `lm()`), and summary statistics (t-tests, F-statistics) were assessed to evaluate predictor significance and overall model fit. We examined the adjusted R^2 to gauge the proportion of variance explained by the chosen features and performed residual diagnostics to check assumptions of normality, homoscedasticity, and linearity.

This final model allowed for a more interpretable understanding of how pre-snap and contextual factors influence yards gained, leveraging a carefully selected subset of informative predictors.

III. RESULTS

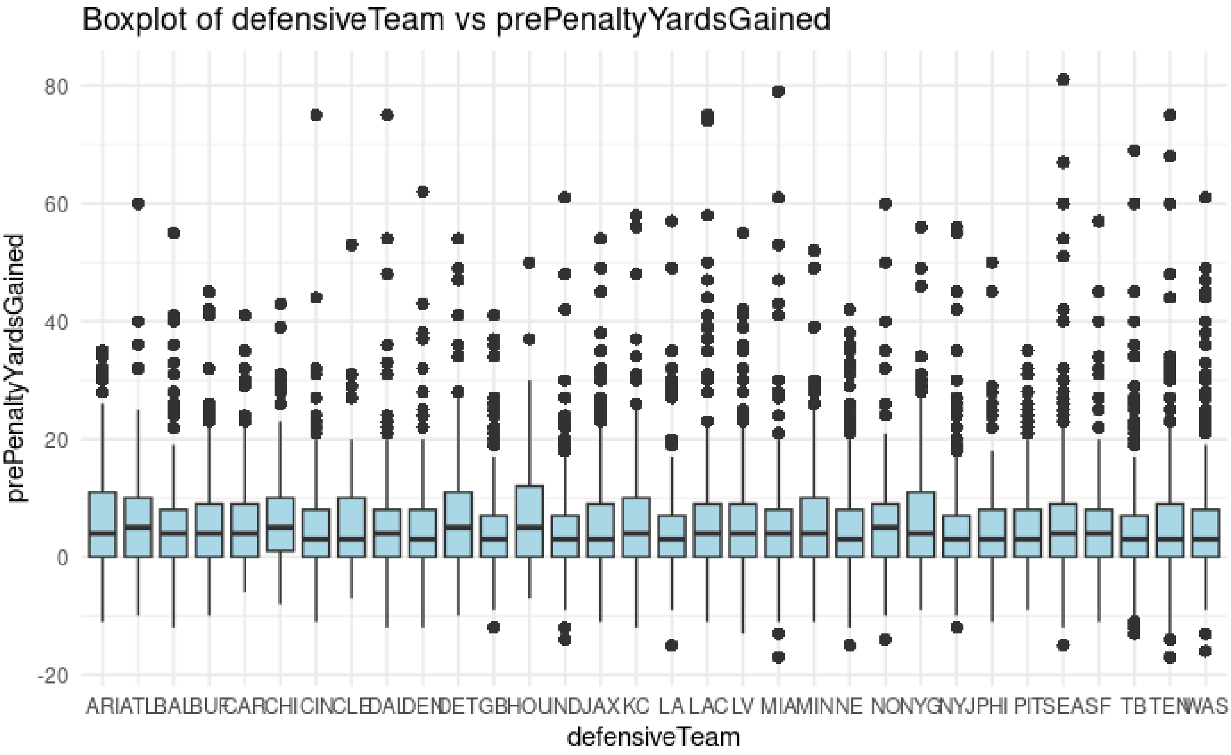
Exploratory Data Analysis

Initial exploratory data analysis revealed that prePenaltyYardsGained is strongly related to the final yardsGained. Correlation tests (e.g., Pearson's correlation) and ANOVAs were performed to assess the relationships between candidate predictors and yards gained. Key findings from the correlation tests include:

- expectedPointsAdded showed a strong positive correlation with yardsGained (r ≈ 0.74, p < 2.2e-16).
- yardlineNumber had a smaller positive correlation (r ≈ 0.086, p < 2.2e-16).
- Some variables, such as penaltyYards_x and timeToThrow, had modest correlations (r ≈ 0.012 and r ≈ 0.012 respectively, p < 2.2e-16).

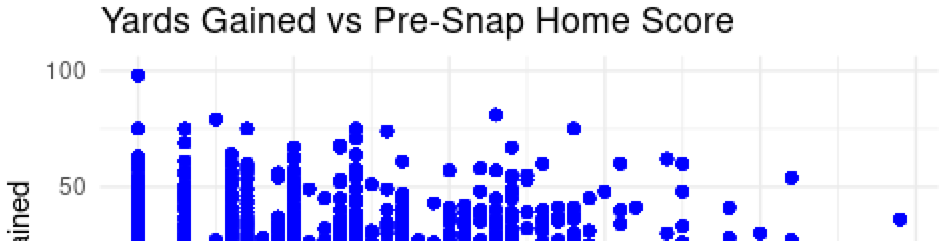
ANOVA results for categorical predictors were also significant:

- passResult and defensiveTeam showed strong evidence of group differences in yards gained (p < 2.2e-16), confirming their relevance in the final model.



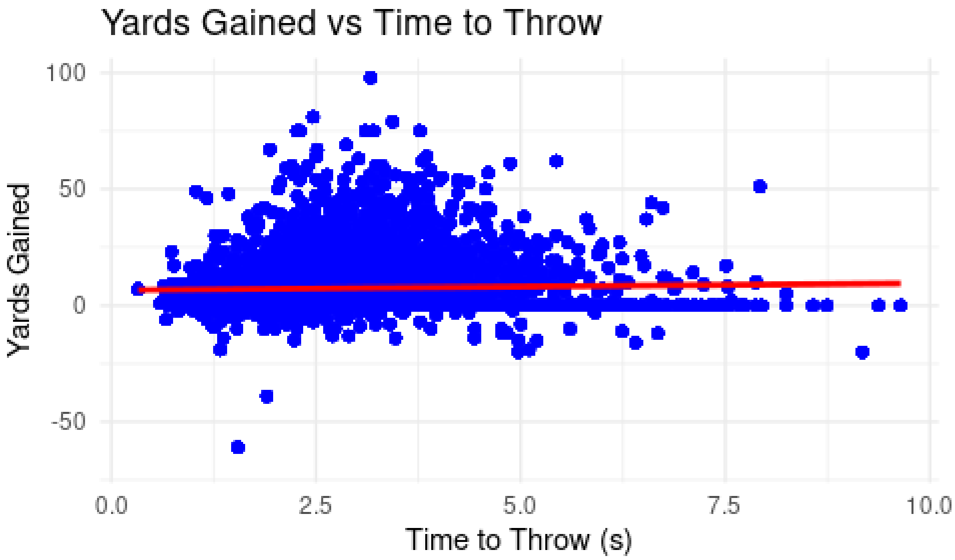
These exploratory visuals highlight varying distributions and relationships, underscoring the complexity of predicting yards gained from pre-snap variables.

1. **Yards Gained vs. Pre-Snap Home Score**



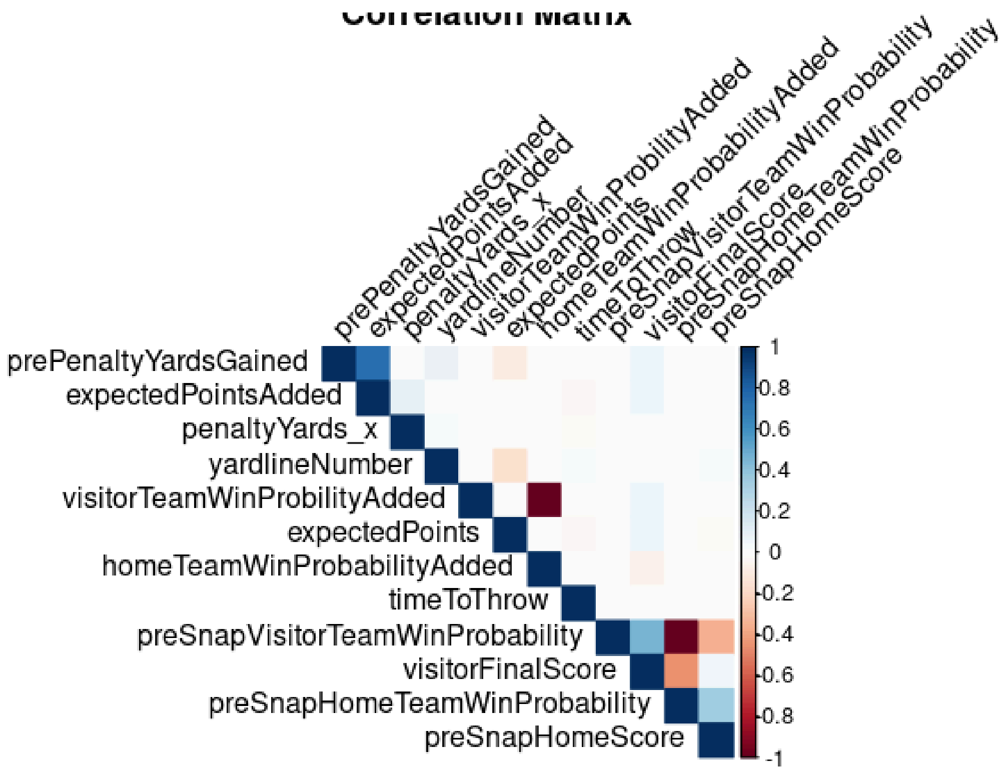
This plot shows the relationship between the pre-snap score of the home team and the yards gained. A slight trend is observed, but variability remains high.

2. **Yards Gained vs. Time to Throw**



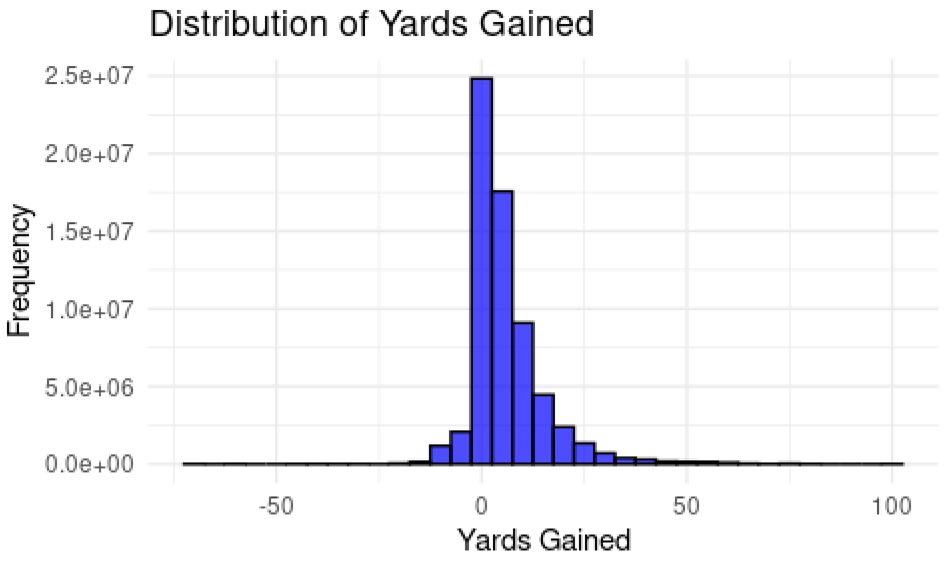
This graph highlights the relationship between quarterback time to throw and yards gained. Plays with a longer time to throw show a marginal increase in yardage gained.

3. **Correlation Matrix**



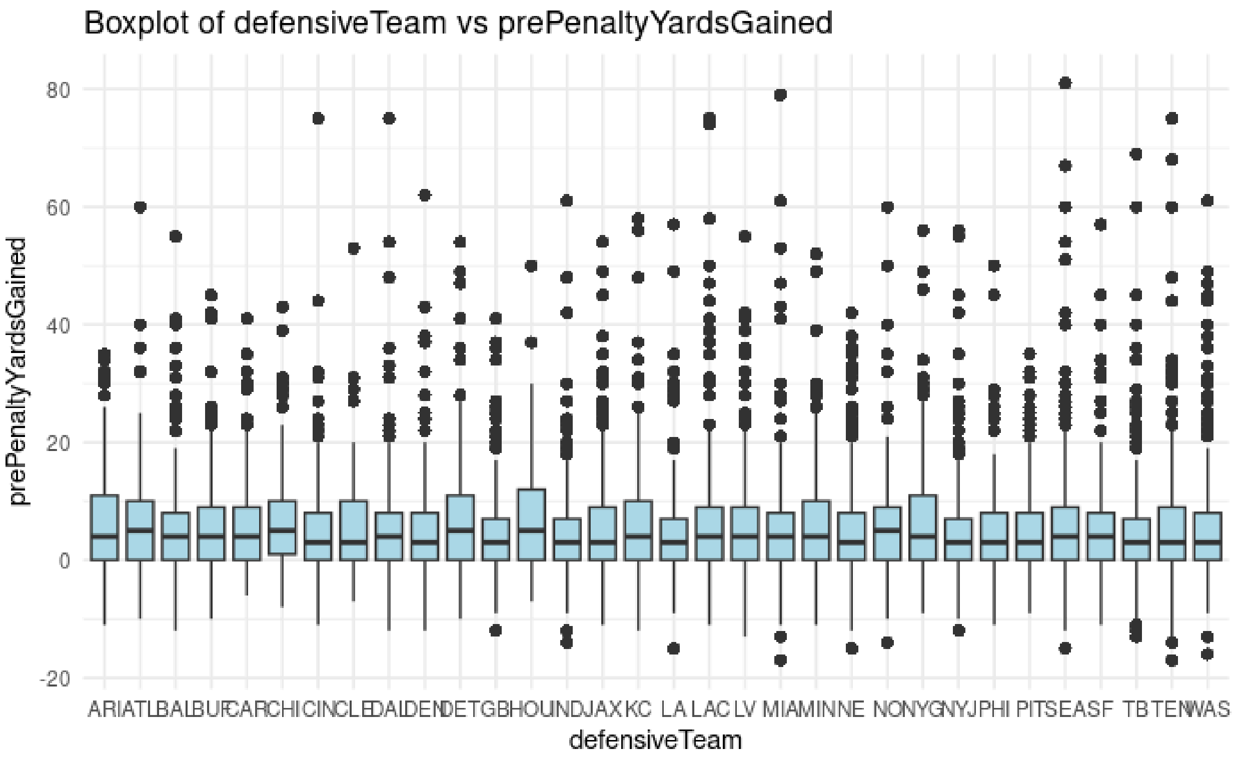
This correlation matrix reveals a notably strong positive relationship between **prePenaltyYardsGained** and **expectedPointsAdded**, suggesting that the yards gained before penalties strongly influence a team's expected scoring potential.

4. Distribution of Yards Gained



The histogram demonstrates a skewed distribution, with most plays yielding between -5 and 10 yards.

5. Boxplot of DefensiveTeam vs. prePenaltyYardsGained



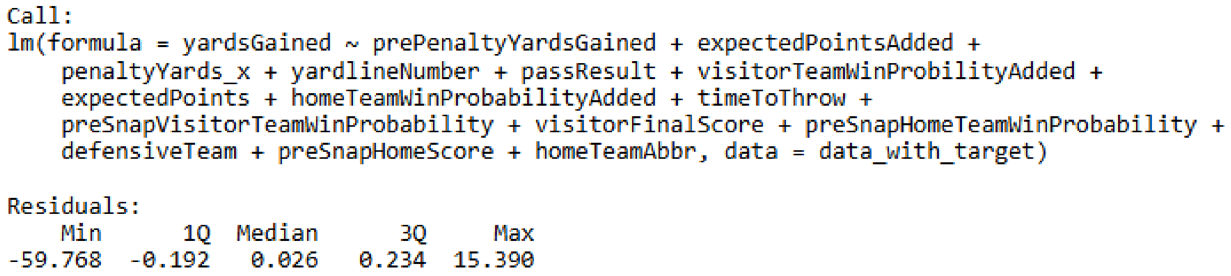
Despite some outliers, the distribution of pre-penalty yardage gained appears broadly similar across defensive teams, suggesting no single defense consistently concedes significantly more or fewer yards before penalties are enforced.

Model Overview

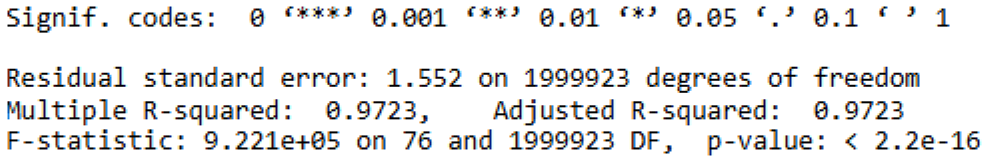
To analyze the relationship between pre-snap variables and the total yards gained during NFL plays, we developed a multiple linear regression model with the following formula:

yardsGained ~ prePenaltyYardsGained + expectedPointsAdded + penaltyYards_x + yardlineNumber + passResult + visitorTeamWinProbilityAdded + expectedPoints + homeTeamWinProbabilityAdded + timeToThrow + preSnapVisitorTeamWinProbability + visitorFinalScore + preSnapHomeTeamWinProbability + defensiveTeam + preSnapHomeScore + homeTeamAbbr, data = data_with_target

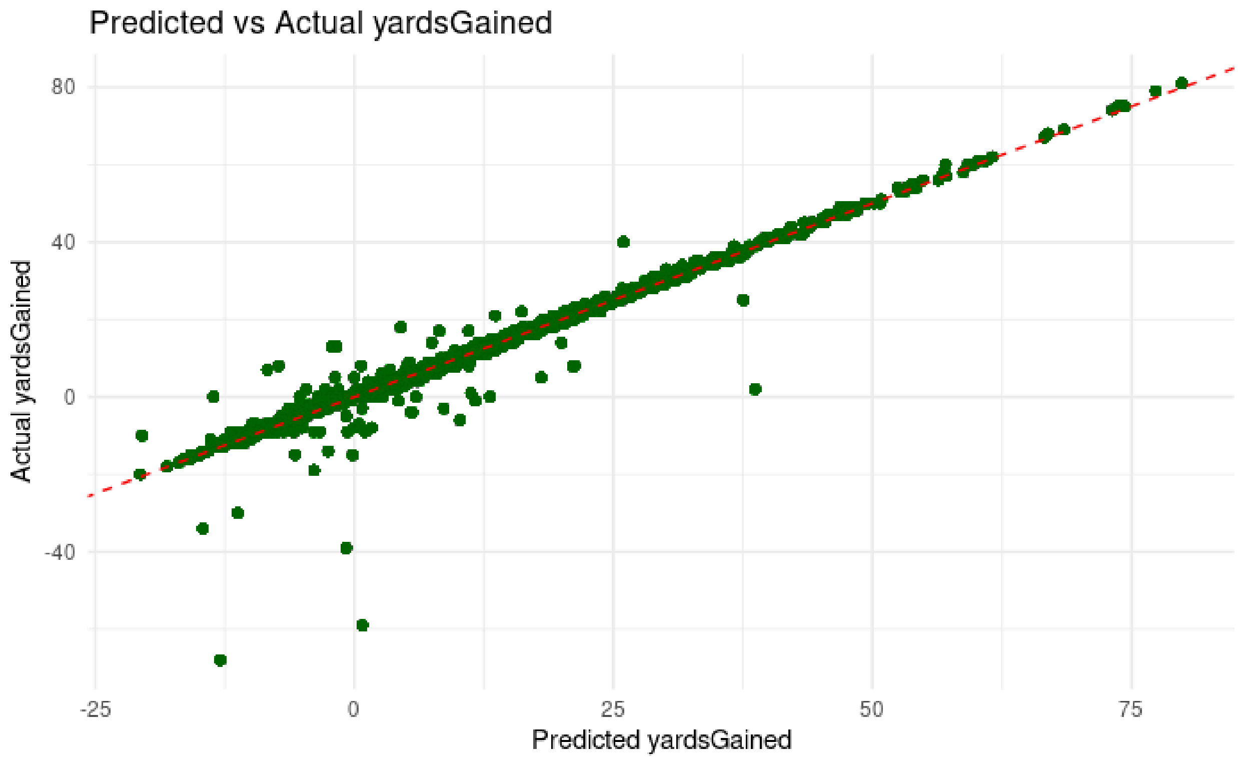
The final fitted model shows an extremely high R^2 (0.9723) and a similarly high adjusted R^2, indicating that the chosen predictors explain a large proportion of the variance in yardsGained. The overall F-statistic is highly significant ($p < 2.2e-16$), confirming the model's predictive strength. Notably, **prePenaltyYardsGained** emerges as a dominant predictor, suggesting that the initial yardage context largely determines the final yards gained on a play.



```
Call:
lm(formula = yardsGained ~ prePenaltyYardsGained + expectedPointsAdded + 
    penaltyYards_x + yardlineNumber + passResult + visitorTeamWinProbilityAdded + 
    expectedPoints + homeTeamWinProbabilityAdded + timeToThrow + 
    preSnapVisitorTeamWinProbability + visitorFinalScore + preSnapHomeTeamWinProbability + 
    defensiveTeam + preSnapHomeScore + homeTeamAbbr, data = data_with_target)

Residuals:
    Min      1Q  Median      3Q     Max 
-59.768  -0.192   0.026   0.234  15.390 
```



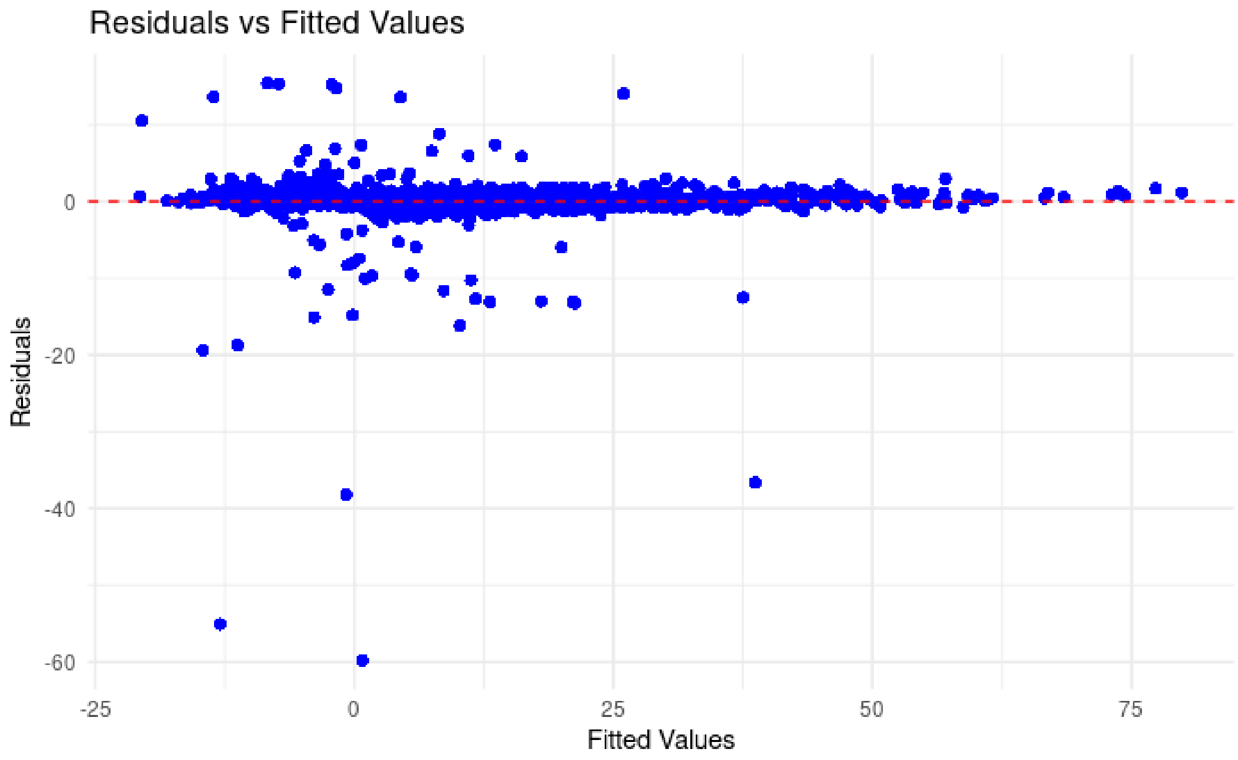
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.552 on 1999923 degrees of freedom
Multiple R-squared:  0.9723,	Adjusted R-squared:  0.9723 
F-statistic: 9.221e+05 on 76 and 1999923 DF,  p-value: < 2.2e-16
```



Model Assumption Checks

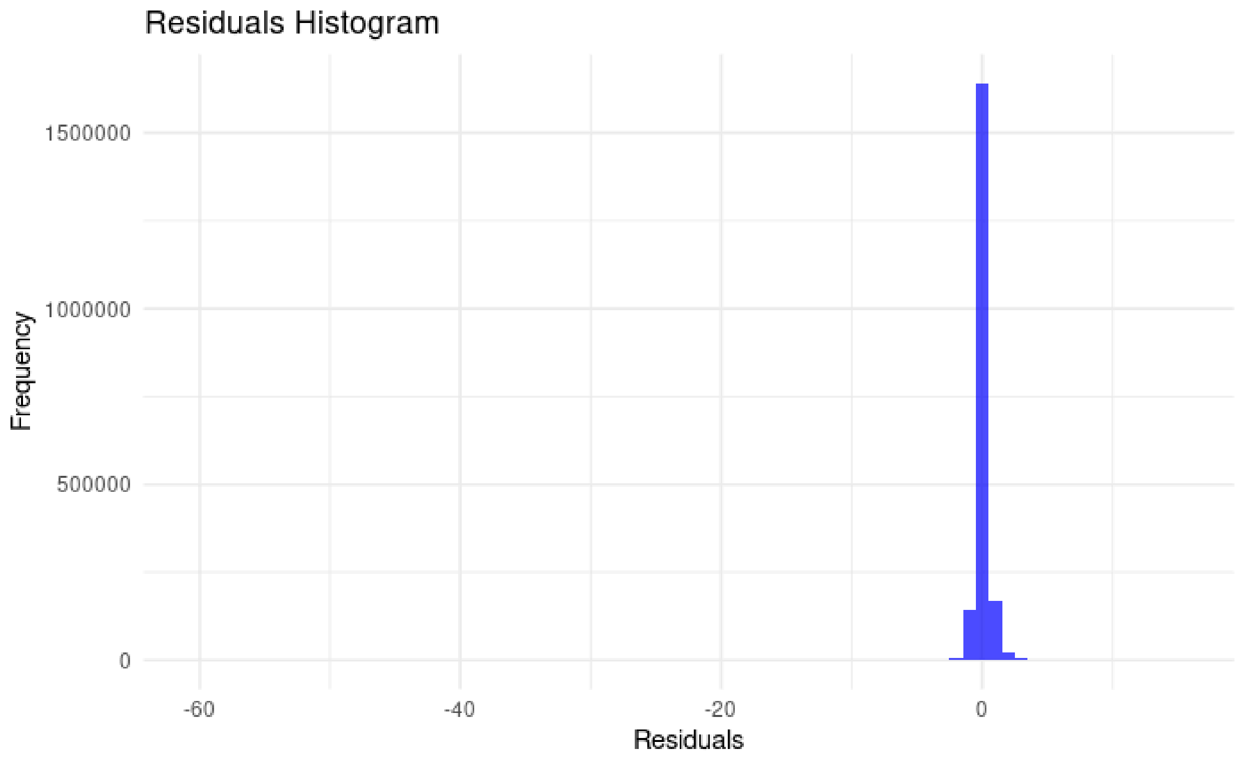
Three key assumptions of linear regression—normality of residuals, linearity, and constant variance—were examined:

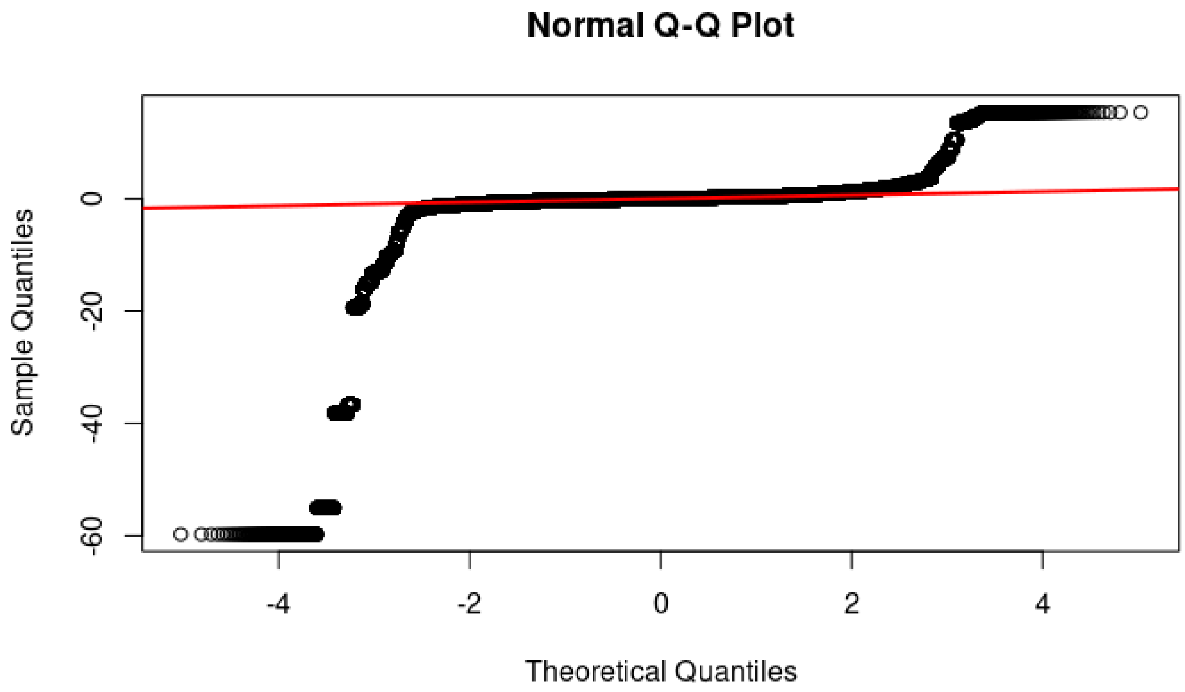
1. **Residuals vs. Fitted Values**:



The residuals should ideally be randomly scattered around zero. Our residual plot shows that while most residuals cluster near zero, there are occasional extreme values. The large sample size makes small deviations visible, but no severe pattern suggests major violations of linearity.

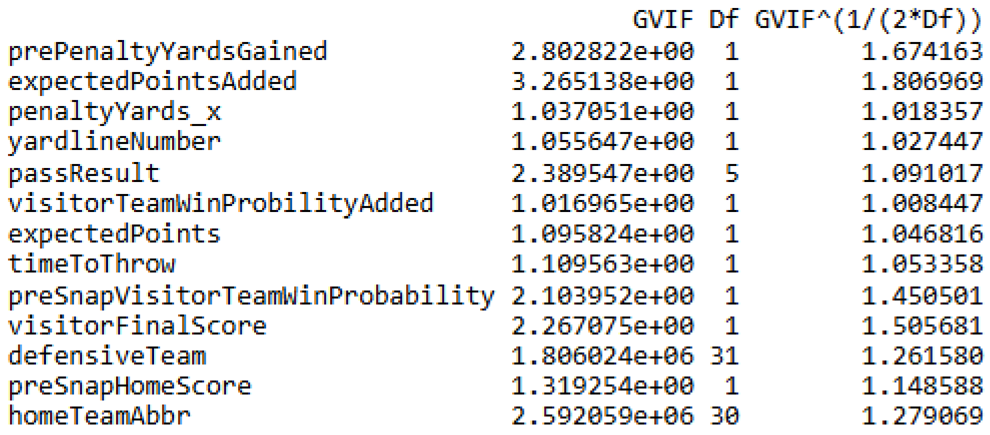
2. **Normality of Residuals**:



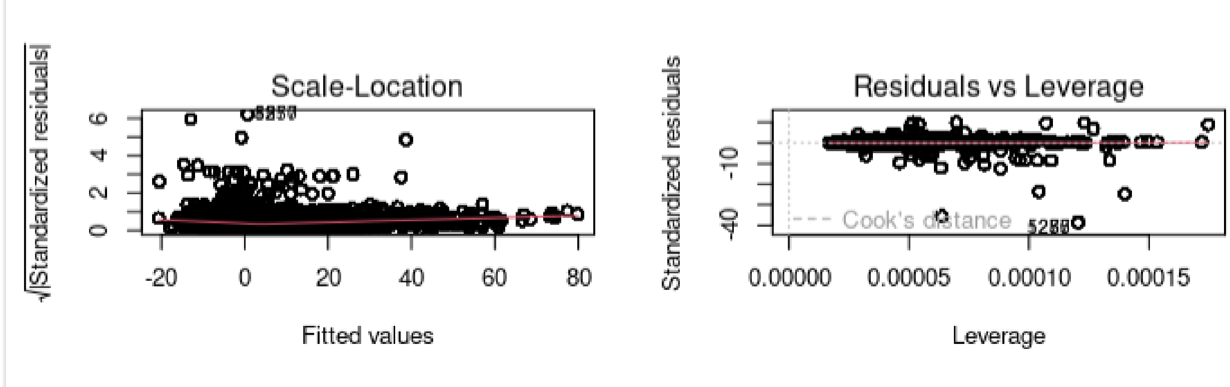


A Q-Q plot indicates that residuals deviate from the normal line in both tails. Given the massive sample size, even slight deviations are magnified. Although residuals are not perfectly normal, the model's predictive power remains strong. Such deviations are not uncommon in large, complex datasets.

3. **Constant Variance (Homoscedasticity)**:



```
                                      GVIF Df GVIF^(1/(2*Df))
prePenaltyYardsGained         2.802822e+00  1        1.674163
expectedPointsAdded           3.265138e+00  1        1.806969
penaltyYards_x                1.037051e+00  1        1.018357
yardlineNumber                1.055647e+00  1        1.027447
passResult                    2.389547e+00  5        1.091017
visitorTeamWinProbilityAdded  1.016965e+00  1        1.008447
expectedPoints                1.095824e+00  1        1.046816
timeToThrow                   1.109563e+00  1        1.053358
preSnapVisitorTeamWinProbability 2.103952e+00  1        1.450501
visitorFinalScore             2.267075e+00  1        1.505681
defensiveTeam                 1.806024e+06 31        1.261580
preSnapHomeScore              1.319254e+00  1        1.148588
homeTeamAbbr                  2.592059e+06 30        1.279069
```



The Scale-Location and Residuals vs. Leverage plots suggest slight heteroskedasticity, but not enough to severely undermine the model's conclusions. Most points fall within a reasonable range, and any patterns are subtle.

These diagnostic checks affirm that while the model is not perfect—no real-world data model is—it provides a robust approximation.

IV. DISCUSSION

Key Findings

Our analysis reveals several important insights:

- **Dominant Role of Pre-Snap Yardage Indicators**: The variable **prePenaltyYardsGained** is a near one-to-one predictor of yardsGained, suggesting that initial field position and penalty contexts strongly determine play outcomes. This intuitive finding underscores that where a team stands before the snap matters significantly.

- **Impact of Pre-Snap Expectations**: Variables related to expected points, visitor/home win probability, and team alignments (defensiveTeam, homeTeamAbbr) also influence yards gained, but to a lesser extent. Offensive strategies, reflected in timeToThrow and passResult, show measurable but smaller contributions once initial conditions are accounted for.
- **High Predictive Performance, but Complexity Remains**: The adjusted R^2 of 0.9723 suggests an unusually strong fit, which likely arises because one key predictor (prePenaltyYardsGained) is highly correlated with the response. While this leads to excellent in-sample prediction, it also indicates that other variables primarily provide incremental improvements beyond what prePenaltyYardsGained already explains.

Limitations and Challenges

1. **Data Limitations**: The dataset lacks defensive alignment variables and weather conditions, which are critical to predicting play outcomes.
2. **Complexity of Football**: The variability in outcomes highlights the multifactorial nature of football plays, where interactions between players play a significant role.

Recommendations for Future Analysis

- Incorporate defensive alignment and environmental factors to improve predictive accuracy.
- Explore advanced machine learning techniques, such as neural networks, for capturing complex relationships.

V. CONCLUSION

This research highlights how key pre-snap variables can predict play outcomes with notable accuracy. Offensive initial conditions—specifically prePenaltyYardsGained—emerge as paramount. While additional variables like expectedPointsAdded, passResult, and team-specific indicators contribute to fine-tuning the prediction, the initial yardage context sets a strong baseline.

These findings provide a valuable starting point for teams seeking to optimize play-calling. Even slight insights from pre-snap alignments and conditions can inform strategic decisions and defensive anticipation. Future work should integrate richer contextual features and explore more advanced models, ultimately aiming to capture the intricate dance between offense, defense, and the countless situational nuances that define NFL football.

REFERENCES

[1] Michael Lopez, Thompson Bliss, Ally Blake, Paul Mooney, and Addison Howard. NFL Big Data Bowl 2025. https://kaggle.com/competitions/nfl-big-data-bowl-2025, 2024. Kaggle.

[2] Yurko, Ronald, Samuel Ventura, and Maksim Horowitz. "nflWAR: a reproducible method for offensive player evaluation in football." *Journal of Quantitative Analysis in Sports* 15.3 (2019): 163-183.