Analyzing and Estimating 4th Down Conversion Probability

Nathan Wright
Jayden Cruz Berdecia
Christian Ortiz
Aiden Ellis
Jesus Rodriguez

Overview

With the high stakes of every NFL game, decision-making is crucial to a team's success, with one clutch moment ultimately being the deciding factor in many games. For this research, our group aims to leverage NFL situational snap data, the results of said snaps, and various team grading factors including PFF o-line and d-line grades and team yards per game offensively and defensively in order to estimate the probability of converting a 4th down based on a given scenario.

Research Objectives

-Factors and Data Sources for Model Development:

This study aims to develop a predictive model for assessing conversion probabilities in NFL football games. Key factors and data sources integral to this model include:

-Pro Football Focus (PFF) Grades: Offensive Line (OL) and Defensive Line (DL) Grades: Evaluations from PFF on the performance of offensive and defensive lines throughout the season. These grades provide insights into the blocking and defensive capabilities crucial for successful offensive plays.

-NFLFastR Play-by-Play Data: Detailed play-by-play data sourced from NFLFastR, encompassing a comprehensive range of metrics such as play outcomes, play type (pass or run), field position, and game context (e.g., quarter, time remaining). This data is fundamental for analyzing the dynamics of each play and its impact on conversion success.

-Pro Football Reference (PFR): Seasonal Offensive and Defensive Metrics: Statistical metrics from Pro Football Reference, including seasonal offensive yards per game (OffYPG) and defensive yards per game (DefYPG). These metrics offer foundational data on team performance, aiding in understanding team strengths and weaknesses in various game situations.

Methodological Approach

The research utilizes a structured approach to model development and analysis:

-Data Preparation: Comprehensive selection and cleaning of independent variables (e.g., OL/DL grades, play outcomes) to ensure data quality and consistency.

-Model Training and Evaluation: Implementation of machine learning techniques, including Random Forest and Logistic Regression models, to predict conversion probabilities. Model performance is evaluated using metrics such as accuracy and ROC AUC score.

-Feature Importance and Selection: Utilization of feature selection methods, such as SelectFromModel with Random Forest, to identify the most influential factors affecting conversion success.
-Predictive Function Development: Development of a custom predictive function that integrates average metrics from selected independent variables to provide tailored predictions for specific game scenarios.

Specifications

The Python programming language's usage will be leveraged to build the probability model. Data will be collected from Pro Football Reference and PFF (Pro Football Focus) websites to be ultimately merged with the play-by-play data database provided by NFLFastR for the 2021, 2022, and 2023 seasons. Classification algorithms such as XGBoost and Random forest as well as traditional regression will be used to fit data gathered to find the most accurate prediction method based on the highest accuracy which will be defined as successful binary classification. From there, 5-fold cross-validation will be used to test each possible combination of variables 5 times against varying data samples to determine the most predictive combination of variables. Surface level analysis of probabilities will be conducted in addition to this, providing numerous visuals showing the variance in conversion success rate probability based on the change in the value of different variables tested in the model.

Step-by-Step Research Process

I. Data Preparation

-Imported play-by-play data from NFLFastR from 2021, 2022, 2023 seasons

-Pulled offensive yards per game, defensive yards per game allowed, average offensive line PFF grade, average defensive line PFF grade, and associated rankings for the season for each team in separate CSVs for 2021, 2022, and 2023 seasons

-Imported each CSV into the working database

-Filtered NFL play-by-play data to contain only the desired information for the model:

a. Fourth down plays
    i. Fourth down plays involving only a conversion attempt (pass or run)
b. Deleted numerous columns until left with the following variables:
    i. Team in possession of the ball (posteam)
    ii. Possession team type (posteam_type) home or away
    iii. Defending team (defteam)
    iv. Which team's side of the field the ball is on (side_of_field)
    v. Number of yards from the end zone the team currently is (yardline_100)
    vi. Seconds remaining in the quarter (quarter_seconds_remaining)
    vii. Seconds remaining in the half (half_seconds_remaining)
    viii. Seconds remaining in the game (game_seconds_remaining)
    ix. Which half the game is in (game_half)
    x. Quarter (qtr)
    xi. Down (down)* all downs will be 4 obviously
    xii. Time remaining on the game clock at the time of snap (time)
    xiii. Yards remaining to get a first down (ydstogo)
    xiv. Net yardage gained so far on the drive***
    xv. Type of play run (play_type)
    xvi. Yards gained from play run (yards_gained)
    xvii. Expected points added by the posteam for the given play. (EPA)
    xviii. Estimated win probability for the posteam given the current situation at the start of the given play. (WP)
    xix. Win probability added for the posteam. (WPA)
    xx. Win probability added for the posteam: spread_adjusted model. (Vegas_wpa)
    xxi. Binary variable for whether the fourth down was converted (fourth_down_converted)

xxii. Binary variable for whether the fourth down attempt failed (fourth_down_failed)

xxiii. Binary variable for whether the fourth down was a rush attempt (rush_attempt)

xxiv. Binary variable for whether the fourth down attempt was a pass attempt (pass_attempt)

xxv. Variable for which season the play took place in (season)

xxvi. Away score at end of the game (away_score)

xxvii. Home score at end of the game (home_score)

xxviii. Spread result from the perspective of the home team (result)

xxix. Binary variable for whether it was a divisional game (div_game)

xxx. Variable for whether a roof was in the stadium (roof)

xxxi. Variable for what type of playing surface there was (surface)

xxxii. Float variable for the temperature of outdoor stadium (temp)

xxxiii. Float variable for the degree of wind being dealt with (wind)

xxxiv. Binary variable for whether the conversion was successful (success)

xxxv. Number of defenders in the box for the play taking place (defenders_in_box)

xxxvi. Numeric value indicating the probability for a complete pass based on comparable game situations. (cp)

xxxvii. For a single pass play, this is 1 - cp when the pass was completed or 0 - cp when the pass was incomplete. Analyzed for a whole game or season an indicator for the passer how much over or under expectation his completion percentage was. (cpoe)

-From here, separate previously stated data from step 2 was imported and merged to combine with play-by-play data in a way that the data points in the conversion attempt data corresponded with the possessing and defending teams involved in the given play.

Step-by-Step Research Process (cont.)

II. Independent Variables Selection

A comprehensive list of potential independent variables that may influence conversion success was compiled. This list includes:

-posteam_home: Whether the team in possession is the home team.
-ydstogo: Yards to go for a first down or touchdown.
-yards_to_endzone: Distance to the end zone.
-qtr: Quarter of the game.
-ydsondrive: Yards gained on the current drive.
-wp: Win probability at the time of the play.
-pass_run: Play type (pass or run).
-div_game: Whether the game is a divisional matchup.
-roof_open_air: Game location (roof or open air).
-grass: Surface type (grass or turf).
-quarter_seconds_remaining: Seconds remaining in the current quarter.
-half_seconds_remaining: Seconds remaining in the current half.
-game_seconds_remaining: Seconds remaining in the game.
-off_OffYPG: Offensive yards per game for the team in possession.
-off_OffRank: Offensive rank for the team in possession.
-def_DefYPG: Defensive yards per game for the opposing team.
-def_DefRank: Defensive rank for the opposing team.
-off_PFFOL: Pro Football Focus offensive line grade for the team in possession.
-off_OLRank: Offensive line rank for the team in possession.
-def_PFFDL: Pro Football Focus defensive line grade for the opposing team.
-def_DLRank: Defensive line rank for the opposing team.
-temp: Temperature at the game location.
-wind: Wind speed at the game location.
-defenders_in_box: Number of defenders in the box.
-cp: Completion probability.
-cpoe: Completion probability over expected.

III. Data Cleaning and Imputation

Missing values in the dataset are imputed using the mean of the respective variable, ensuring a complete dataset for analysis.

Step-by-Step Research Process (cont.)

IV. Model Testing, Training, and Feature Selection

-Data Splitting:

The formatted, cleaned, and imputed dataset is split into training and testing sets to facilitate model training and evaluation.

-Random Forest Classifier:

A Random Forest Classifier is trained on the independent variables to assess feature importance and enhance the robustness of the model.

-Feature Standardization:

Units of the independent variables (decimals, percentages, integers) are standardized to improve model performance and ensure consistency in scale.

-Feature Selection:

Important features are selected using the trained Random Forest model to reduce dimensionality and focus on the most influential variables.

V. Model Training and Evaluation

-Logistic Regression:

A Logistic Regression model is trained on the selected features to predict conversion success.

-Model Evaluation:

The Logistic Regression model's performance is evaluated using metrics such as accuracy, ROC AUC score, and a classification report. Cross-validation is then performed to ensure the model's robustness and reliability. Findings from evaluation used in a predictive function.

VI. Implementation of a Predictive Function

-Predictive Function Development:

A custom mathematical function was developed to predict the probability of a successful conversion based on user inputs for yards to go, and the teams on offense and defense. The function computes averages for the other independent variables associated with the specified teams to provide a tailored prediction.

Prediction Model Results

-Full Variable Set:

After completing the data analysis outlined in the preceding steps, a predictive model was developed to forecast the success of a conversion attempt using the values of the independent variables under examination. The resulting weights assigned to each of these independent variables are detailed below:

posteam_home	0.041013
ydstogo	-1.104071
yards_to_endzone	0.874041
qtr	0.538942
ydsondrive	1.502922
wp	0.208747
pass_run	-1.060392
div_game	-0.075648
roof_open_air	0.040943
grass	0.089686
quarter_seconds_remaining	-0.165088
half_seconds_remaining	0.219136
game_seconds_remaining	0.287937
off_OffYPG	-0.041448
off_OffRank	-0.068458
def_DefYPG	-0.370895
def_DefRank	0.334790
off_PFFOL	-0.278593
off_OLRank	-0.287489
def_PFFDL	-0.035333
def_DLRank	0.017843
temp	0.040561
wind	0.054300
defenders_in_box	-0.156111
cp	0.356004
cpoe	3.380223

These variables and their corresponding weights lend to an accuracy in successful prediction of a given 4th down conversion attempt of 85.6%. Listed below is the classification table that resulted from the research:

```
Accuracy: 0.8559670781893004
ROC AUC Score: 0.8557795905802679
Classification Report:
            precision    recall  f1-score   support

       0.0       0.83      0.85      0.84       219
       1.0       0.88      0.86      0.87       267

   accuracy                          0.86       486
  macro avg       0.85      0.86      0.85       486
weighted avg       0.86      0.86      0.86       486
```

The ROC AUC score is the area under the ROC curve. It sums up how well a model can produce relative scores to discriminate between positive or negative instances across all classification thresholds.

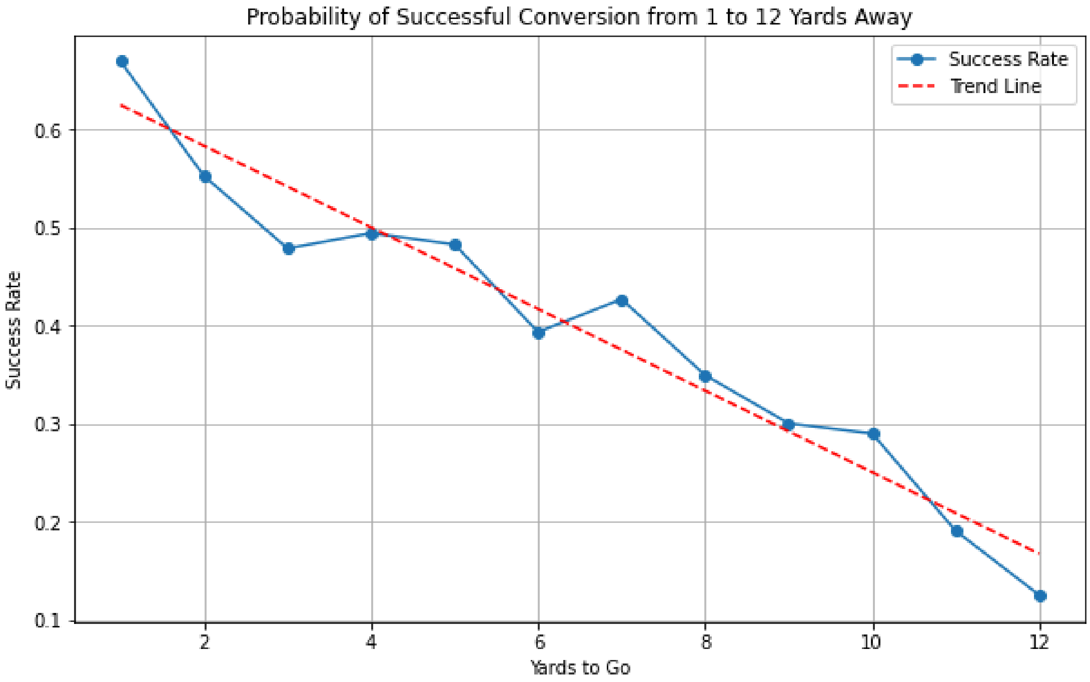
The ROC AUC score ranges from 0 to 1, where 0.5 indicates random guessing, and 1 indicates perfect performance.

- Cross-Validation Results:

After completing the initial modeling, as stated previously, 5-fold cross-validation was used to find the most accurate combination of variables that can predict the outcome of a conversion attempt given the values of the variables included within the model. Five-fold cross-validation is a technique used to assess the performance and generalizability of a predictive model. It involves splitting the dataset into five equal parts (folds). In each iteration, four folds are used for training the model, and the remaining fifth fold is used for testing. This process is repeated five times, with each fold used exactly once as the test set. The results from each iteration are averaged to provide an overall estimate of model performance, helping to mitigate the risk of overfitting and providing a more reliable measure of how the model is likely to perform on unseen data. Listed below are the variables and their associated weights that resulted from this process as well as the accuracy scores for each iteration of this variable set and the mean accuracy:

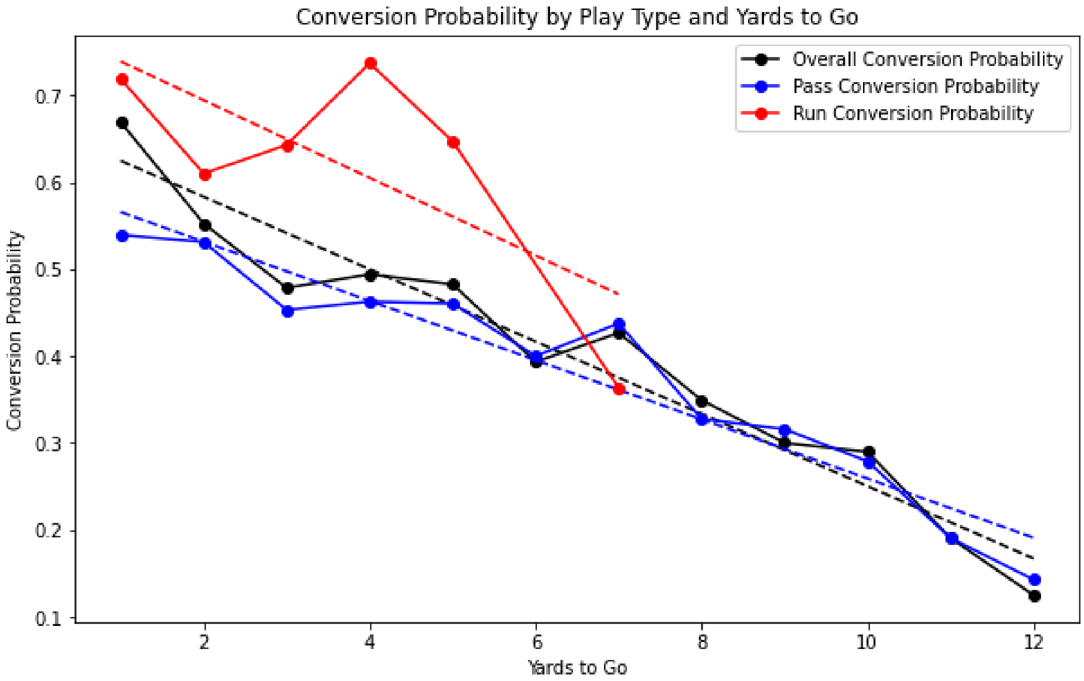
```
ydstogo    -1.456882
yards_to_endzone     0.917689
ydsondrive     1.507169
cpoe     2.504830
cross_val_scores [0.85089974 0.89460154 0.88174807 0.87371134 0.84536082]
Mean cross-validation accuracy: 0.8692643044549863
```

Probability Analysis Findings



Yards to Go	Success Rate
1	66.91%
2	55.23%
3	47.85%
4	49.40%
5	48.25%
6	39.33%
7	42.67%
8	34.92%
9	30.00%
10	28.97%
11	19.05%
12	12.50%

Based on the sample of 2021 to 2023 play-by-play data, a -4.16% change in probability per additional yard to go on a conversion attempt was evidenced as visualized in the graph above with the success rates at each yardage point being displayed in the table.



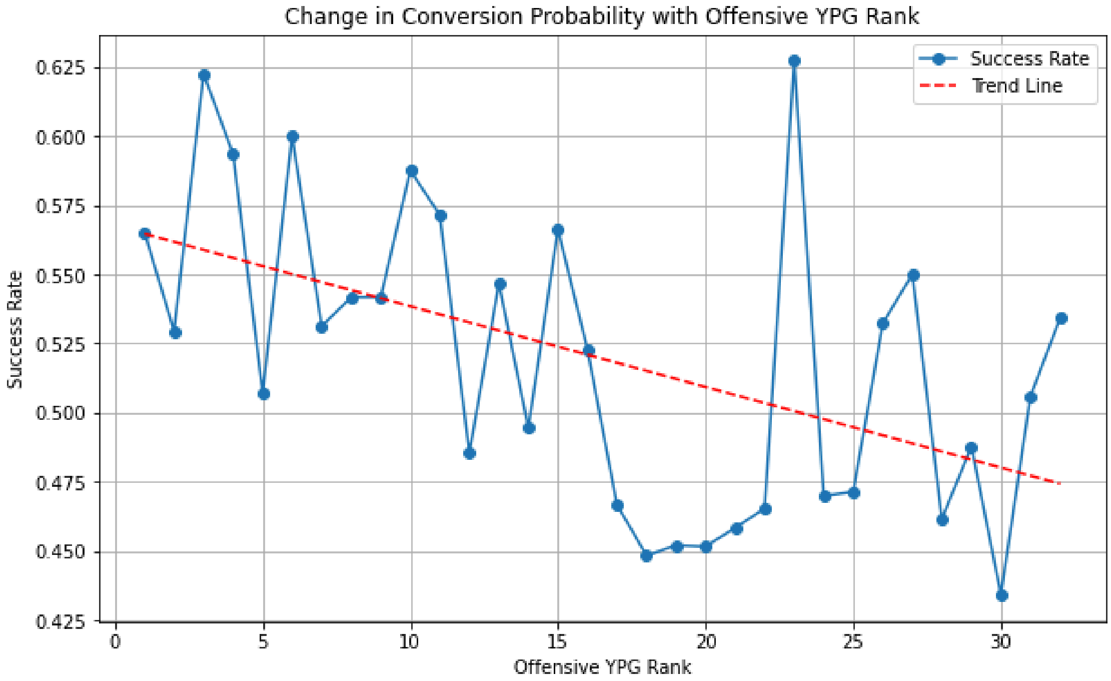
Yards to Go	Overall Success Rate	Passing Success Rate	Running Success Rate
1	66.91%	53.93%	71.88%
2	55.23%	53.11%	61.00%
3	47.85%	45.30%	64.29%
4	49.40%	46.26%	73.68%
5	48.25%	46.03%	64.71%
6	39.33%	40%	n/a
7	42.67%	43.67%	36.36%
8	34.92%	32.67%	n/a
9	30.00%	31.58%	n/a
10	28.97%	27.88%	n/a
11	19.05%	19.05%	n/a

12	12.50%	14.29%	n/a

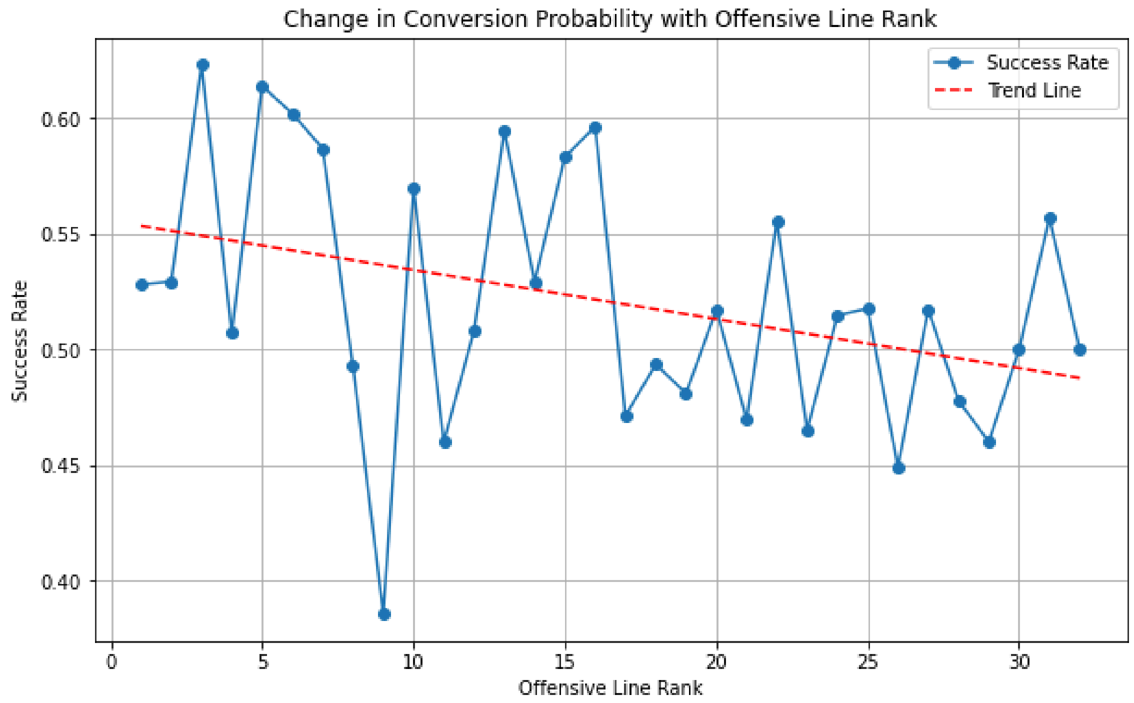
minimum 10 conversions attempts by play type at each yardage point

Based on the sample of 2021 to 2023 play-by-play data:

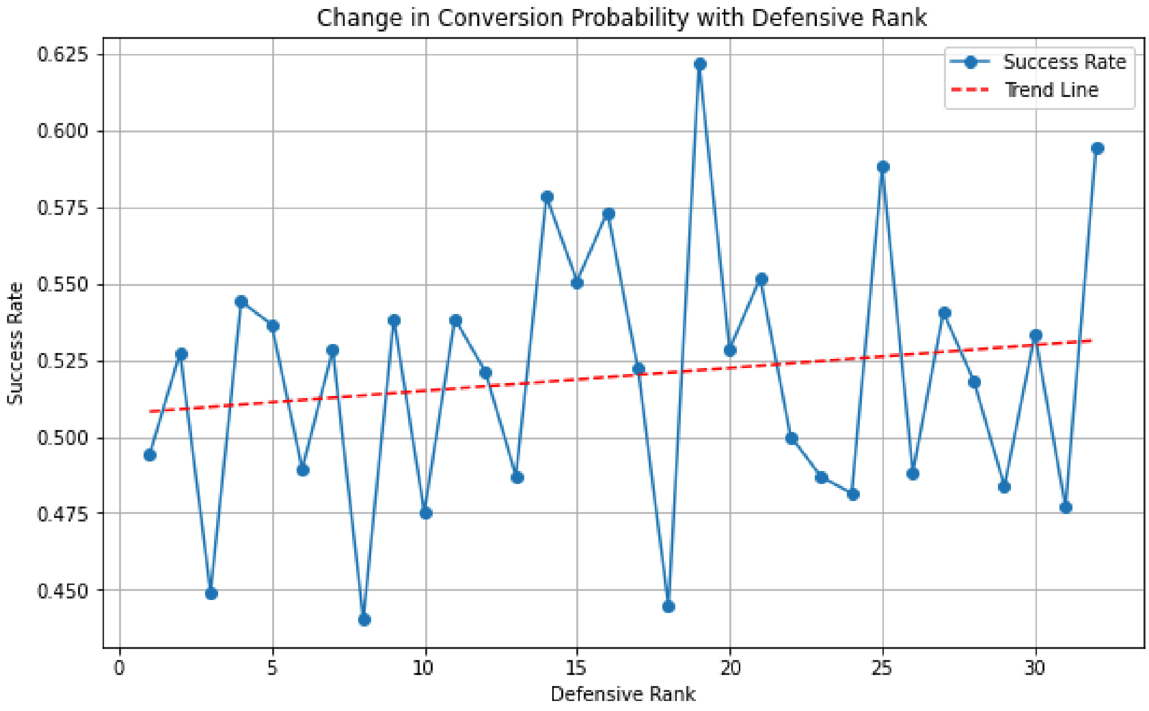
- The change in success rate of conversion attempts involving a pass attempt with each additional yard of distance to go was -3.41%.
- The change in success rate of conversion attempts involving a run attempt with each additional yard of distance to go was -4.47%.



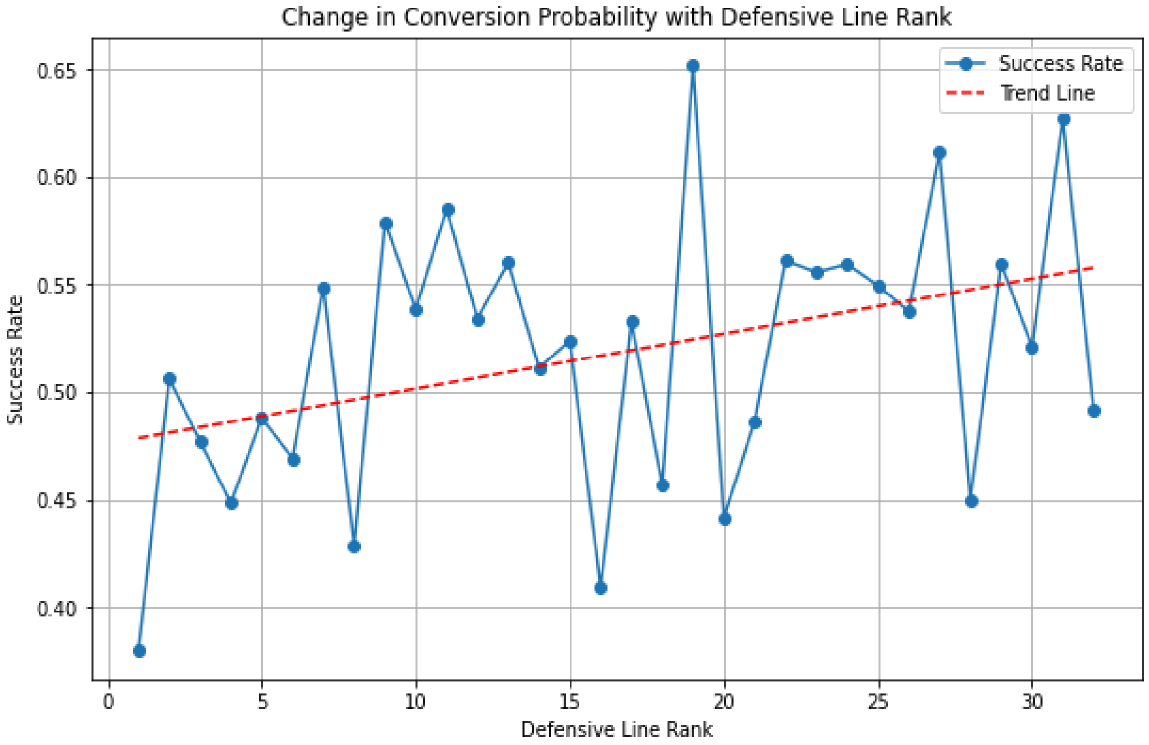
Based on the sample of 2021 to 2023 play by play data, a -0.29% change in probability for each additional decrease in ranking slot from 1st in the league to 32nd for the offensive unit's yardage per game output was evidenced as visualized in the graph above.



Based on the sample of 2021 to 2023 play-by-play data, a -0.21% change in probability for each additional decrease in ranking slot from 1st in the league to 32nd for the possessing team's offensive line was evidenced as visualized in the graph above.

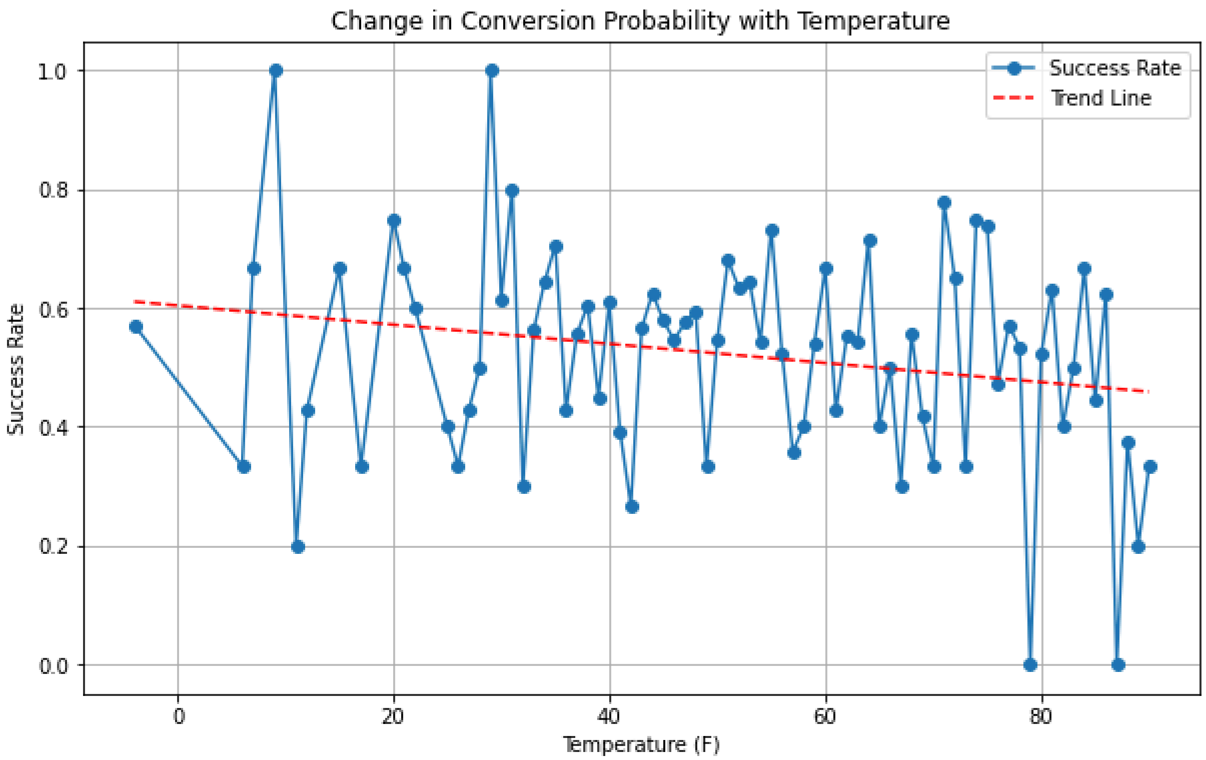


Based on the sample of 2021 to 2023 play-by-play data, a 0.07% increase in probability for each additional decrease in ranking slot of the defense being faced from 1st in the league to 32nd for team defensive yards per game allowed was evidenced as visualized in the graph above.

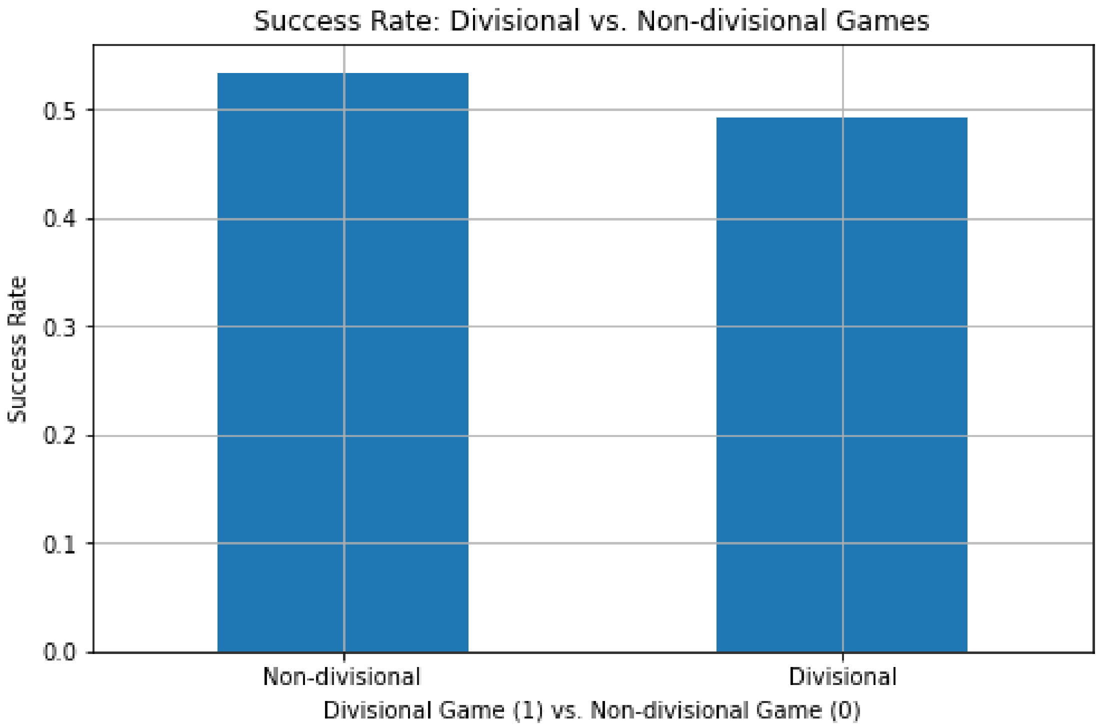


Based on the sample of 2021 to 2023 play-by-play data, a 0.26% increase in conversion probability for each additional decrease in ranking slot from 1st in the league to 32nd for opposing defensive line rank was evidenced as visualized in the graph above.

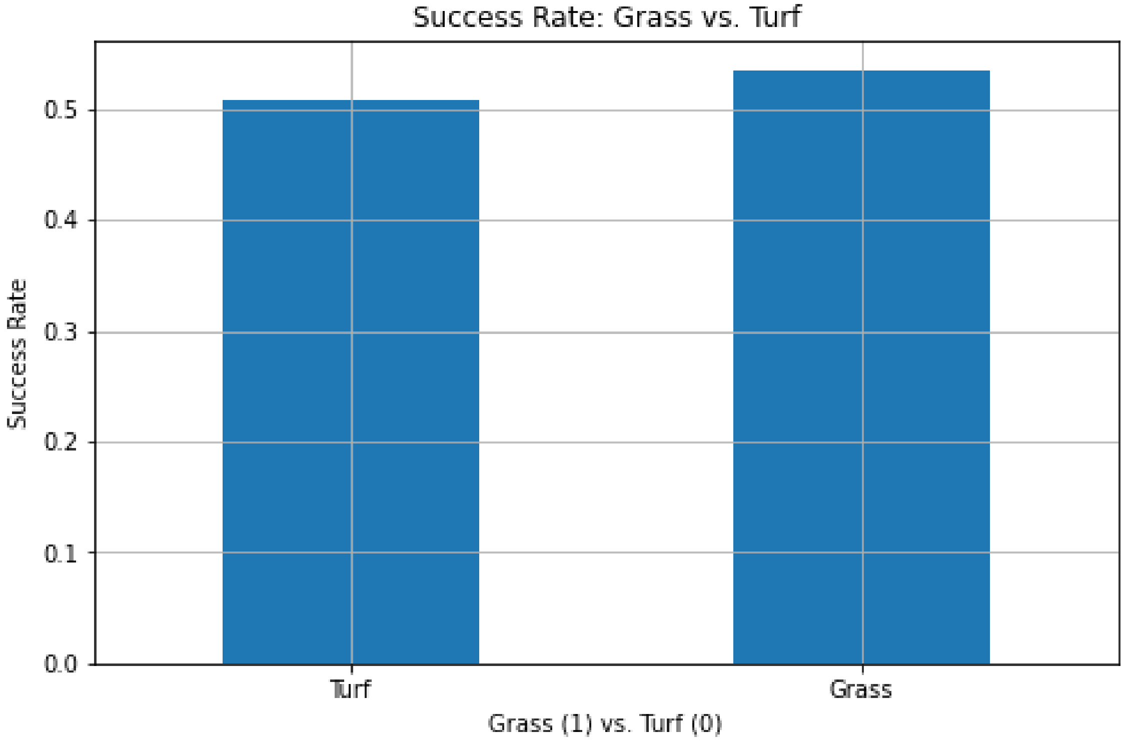
When comparing the opposing defensive line slope to the offensive line slope, it is ever so slightly more important to beef up your defensive line to stop someone else from converting versus improving one's own offensive line to increase the possessing team's odds. Overall, this means on a very minimal basis, a great defensive line unit beats a great offensive line unit based on the previous three seasons of NFL football.



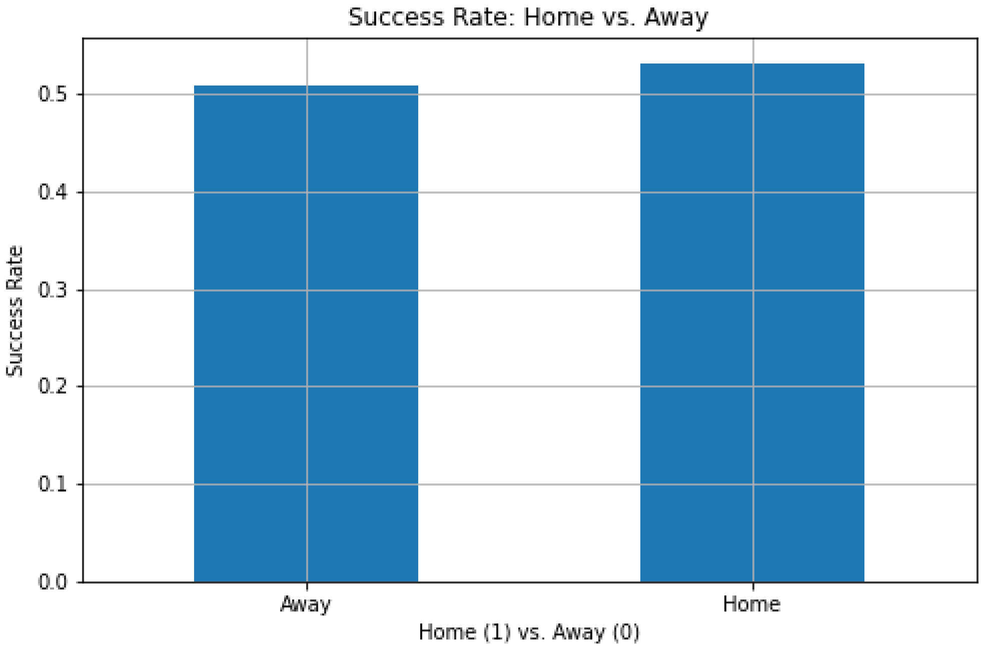
Based on the sample of 2021 to 2023 play-by-play data, a -0.16% change in probability with each degree Fahrenheit increase in temperature was evidenced as visualized in the graph above. This is potentially due to the defending team's players being slower to react to plays in colder temperatures, lending the advantage towards the offense. Additionally, as evidenced in the probabilities at each unit in yards to go, running lends itself to a higher success rate. Expanding on this, run plays are more frequent in colder temperatures, providing an additional explanation for this trend.



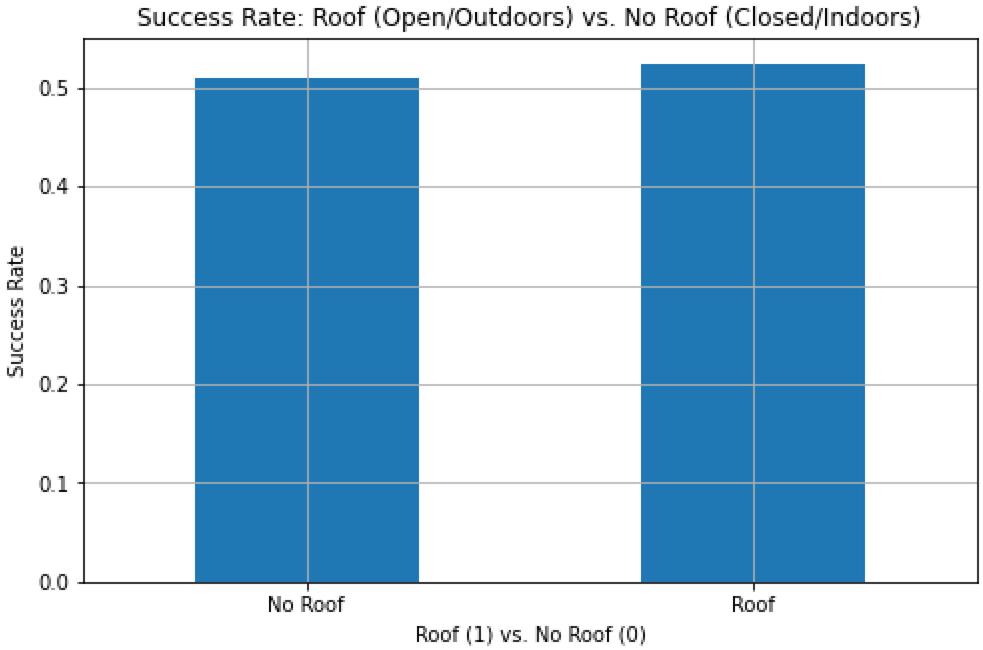
All other variables held constant, playing in a non-divisional game was associated with a 4.10% higher probability of success (53.41%) compared to playing in a divisional game (49.30%), evidencing the advantage of facing different opponents outside of divisional play. Additionally, it shows the increased difficulties in strategic planning and success when facing a given opponent an additional time in a given season.



All other variables held constant, playing on grass was associated with a 2.59% higher probability of success (53.49%) compared to playing on turf (50.88%), evidencing the advantage of natural grass surfaces.



All other variables held constant, playing as the home team was associated with a 2.19% higher probability of success (53.10%) compared to playing as the away team (50.91%), evidencing the magnitude of possessing home-field advantage.



All other variables held constant, playing in a stadium with a roof was associated with a 1.43% higher probability of success (52.40%) compared to playing in a stadium without a roof (50.98%), evidencing the advantage of covered stadiums as they possess a controlled, stable environment.

Conclusion and Future Directions

-Practical Applications:

The research undertaken in this study culminates in the development of a predictive model designed to estimate the probability of converting a 4th down based on specific game scenarios. This model has immediate practical applications for NFL teams and coaching staff, providing a data-driven tool to enhance decision-making processes during crucial moments of the game. By incorporating situational snap data, team grading metrics, and historical play outcomes, the model offers a comprehensive analysis that can be used to inform strategic choices on the field.

-Future Possibilities:

While the current model focuses on predicting 4th down conversions, the framework established in this research paves the way for broader applications. Future enhancements could include:

- Expanded Decision Analysis: Integrating additional decision-making factors such as the probability and expected points added (EPA) of punting or attempting a field goal, thereby creating a holistic decision-making tool for 4th downs.
- In-Game Adaptation: Developing real-time applications that update probabilities based on live game data, enabling dynamic decision-making support for coaches.
- Player-Specific Analysis: Incorporating player-specific performance metrics to refine predictions based on the individuals involved in each play.
- Generalizing to Other Scenarios: Extending the model to predict outcomes for other critical situations in the game, such as 2-point conversion attempts or key 3rd down plays.

-Final Thoughts:

The predictive model developed through this research represents a significant advancement in the use of data analytics for NFL game strategy. By leveraging comprehensive datasets and sophisticated algorithms, the model not only improves the accuracy of 4th down conversion predictions but also provides a foundation for future innovations in sports analytics. As the landscape of professional sports continues to evolve with the integration of technology and data, the insights gained from this research have the potential to shape the future of game strategy and decision-making.

Bibliography/Inspirational Sources

- ➔ https://www.nflfastr.com/index.html
- ➔ https://www.pro-football-reference.com/
- ➔ https://www.pff.com/
- ➔ https://www.sportingnews.com/us/nfl/news/nfl-fourth-down-conversion-chart-rate-by-distance/vofkeub6xwms6imajxqkfipp *
- ➔ https://www.espn.com/nfl/story/_/id/33059528/nfl-game-management-cheat-sheet-punt-go-kick-field-goal-fourth-downs-plus-2-point-conversion-recommendations *
- ➔ https://www.bruinsportsanalytics.com/post/4th-down-model *
- ➔ https://sites.google.com/view/microeconometricswithr/blog/4th-down *

*asterisk denotes source only used for inspiration, with no hard references or methodology used in project research