Predicting the Spread, Total Points, and Passing Yards of NFL Games: Weeks 7-11

Author: Truongan Nguyen

Sports Analytics 538

October 15, 2025

Section 1: Data

1.1 Data Sources

Our primary dataset was obtained from **NFLfastR**, an open-source play-by-play database for the NFL. We first accessed the data using the NFLfastR R package, which calls the public API and returns structured play-level data for each game and season. This dataset alone included variables such as play type, overall yards gained, success indicators, expected points added (EPA), and other situational details (down, distance, and field positions).

We supplemented this with a **Kaggle** dataset containing game final scores and team records (NFL Game Data: Scores & Plays). This outcome-level information complemented the play-by-play data and allowed us to calculate variables such as pre-game records.

During the brainstorming phase, we then explored external predictors such as weather data (via NFLfast R) and player injury data (Kaggle). While these offered potential insights, both were ultimately excluded from the finalized dataset due to certain challenges with predictive reliability and integration.

1.2 Data Cleaning

Each dataset underwent a cleaning process to ensure consistency before merging.

NFLfastR data cleaning included:

- Updating outdated team names (e.g., Oakland Raiders → Las Vegas Raiders).
- Filtering out preseason and Pro Bowl games.
- Dropping variables that are unnecessary (e.g., coach name, player IDs) and removing fully missing observations.
- Replacing missing field goals and PAT % with 0.

Kaggle scores dataset cleaning included:

- Updating outdated team names (e.g., Redskins → Commanders).
- Filtering out preseason and Pro Bowl games.
- Creating a pre-game record variable, since the dataset only provided final scores.

1.3 Dataset Integration

To merge multiple sources, we then constructed a unique identifier called game_id , which is formatted as Season_Week_Away_Home (e.g., 2023_01_ARI_WAS).

game_id = paste(Season, week_num, away_abbr, home_abbr, sep = "_")

This key was added to both the cleaned NFLfastR and Kaggle datasets, which were then merged into a comprehensive table called games.

1.4 Variable Engineering

We engineered several variables to better analyze team performance and play style. These variable descriptions are broken up into their relation to either offense or defense.

Additionally, these statistics were engineered by restricting the dataset to the first six weeks and then calculating these metrics at a team level.

Offensive Metrics:

- EPA per play: The mean of the expected points added

$$epa\_per\_play = \frac{1}{N}\sum_{i=1}^{N} EPA_i$$

- Success rate: The mean of the _success_ variable, a binary 0/1, providing an average of how often a play had positive EPA

$$success\_rate = \frac{1}{N}\sum_{i=1}^{N} successs_i$$

- Pass rate: The percentage of plays that the offensive team passed the ball

$$pass\_rate = \frac{\#Pass\ Plays}{\#Pass\ Plays + \#Rush\ Plays}$$

- Rush rate: The percentage of plays that the offensive team rushed the ball

$$rush\_rate = \frac{\#Rush\ Plays}{\#Pass\ Plays + \#Rush\ Plays}$$

Defensive Metrics:

- Defensive EPA per play:
    - The mean of the expected points added, the higher the *_def_epa_per_play_*, the worse the defense.
- Defensive success rate:
    - The mean of *_success_*, representing the percentage of plays that the offense succeeded. Larger percentages show defensive failure.

To strengthen predictive power, we engineered additional variables across five categories: difference-based, lagged, rolling average, play style & pace, and win percentage variables.

1. Difference-Based Features:
We captured relative advantages by subtracting away team statistics from home team statistics.

- *diff_passing_yards = home_passing_yards - away_passing_yards*
- *diff_rushing_yards = home_rushing_yards - away_rushing_yards*

2. Lagged Features:
These lagged metrics summarize each team's most recent performance trends and efficiency, identifying predictive patterns for the model.

- *home_lag_points_scored = total points the home team scored in the previous game*
- *away_lag_points_scored = total points the away team scored in the previous game*

3. Rolling Averages:
We considered that 3-game rolling averages could be helpful when predicting Week 7, but realized the other weeks may be more difficult because we won't have the last 3 games before entering our predictions. Therefore, these lagged predictors were not included in the final data set.

- 3-game scoring trends:
    - *home_ma_points_scored_3* = Computed the 3-game rolling average of points scored by the home team to capture short-term offensive trends.
    - *away_ma_points_scored_3* = computed the 3-game rolling average of points scored by the away team to reflect recent scoring form.
- EPA per play avg:
    - *home_epa_per_play3* = took the mean of offensive EPA per play over the last three games for the home team to measure offensive efficiency.
    - *away_epa_per_play3* = took the mean of offensive EPA per play over the last three games for the away team to summarize recent effectiveness.

4. Play Style and Pace Metrics:
These variables capture differences in how teams approach their offensive strategy and the tempo at which they play.

- *home_pass_rate, away_pass_rate* - fraction of pass attempts in recent games.
- *home_plays, away_plays* - total offensive plays run in prior games, reflecting pace and possession volume.

5. Win Percentage:
These variables were created for our initial basic model that was based solely on win percentage.

$$Home\ Win\ Percentage\ = \frac{Home\ Wins}{(Home\ Wins + Home\ Losses)}$$

$$Away\ Win\ Percentage\ = \frac{Away\ Wins}{(Away\ Wins + Away\ Losses)}$$

However, as will be seen later in the discussion of the models, most of these predictors that were engineered were not used. During our process of selecting predictors, these 'innovative' variables were not chosen by our methods of selection, such as stepwise regression. The significant predictors are discussed further in Sections 2-4.

1.5 Additional Datasets & Variables

In addition to these core variables, we looked at the potential of incorporating more creative predictors, such as game-day weather conditions and injury data, to capture external factors that could influence performance or scoring outcomes.

Weather-related data:
To account for how weather conditions might influence scoring or passing outcomes, we collected weather-related data such as temperature, wind speed, and roof status (open or closed) using the NFLfastR package in R. However, we encountered difficulties predicting future weather for Weeks 7–11, as forecasts would quickly become outdated, limiting the usefulness of these variables in our models.

Injury-related data:
Because player availability, especially for key positions like quarterback or wide receiver, can significantly impact game outcomes, we decided to collect an injury dataset from Kaggle. We

used this data to create variables like *home_injuries*, *away_injuries*, *home_qb_out*, and *away_qb_out*. However, after trying to implement it in our models, we found that historical injury data was difficult to apply due to frequent roster changes and evolving player roles between seasons. We also realized we wouldn't have access to injury reports to make predictions for Weeks 8–11.

Ultimately, due to multiple challenges we encountered with the weather and injury data, we found that integrating this information became too complex to implement effectively within our timeline.

1.6 Train-Test Split

The data was split into training and testing sections based on the year. Our analysis was done on data from seasons 2017-2024. We have two distinct errors, an internal cross-validation error on the 2017-2023 training set and an external error on the 2024 season. The final model was then retrained on the entire 2017-2024 data set to maximize prediction power.

1.7 Current Season Data Collection

Finally, we pulled data from the first six weeks of the 2025 season using the NFLfastR package, which automatically updates week by week. The averages from the first six weeks in each metric were added as additional columns. This data was later used in our models as possible predictors.

Section 2: Methodology for Spread

2.1 Objective

The goal was to predict the point spread, defined as:

$$Spread = Home\,Score - Away\,Score)$$

A positive predicted spread indicates a home win and a negative spread predicts an away win. Our objective was to model and predict this difference using past team performance and efficiency metrics.

2.2 Model Selection Process

We started off with a simple linear regression model using only home and away win percentages as predictors:

$$Spread = \beta_0 + \beta_1 (Home\,Win\,\%) + \beta_2 (Away\,Win\,\%)$$

Win percentage reflects overall team strength and efficiency. The model achieved a Mean Absolute Deviation (MAD) of 10.11 in cross-validation and 9.65 on the 2024 test set. Both predictors were statistically significant while the intercept ($\beta_0 = 2.27$) was not, suggesting a home-field advantage of roughly 2 points.

Estimated coefficients were +11 for home win % and -13 for away win %, meaning a 10-point increase in the home team's win rate % raises the expected spread by 1.1 points, while a similar increase for the away team decreases by 1.3 points.

Next, we created a model that highlighted the relative differences in team strength by engineering difference features between home and away metrics. Since spread shows performance gaps, we compared passing, rushing, and receiving yards and incorporated moving averages and one week lags for points scored, yards gained, and EPA per play to capture recent team trends.

Key predictors:

- Performance: *diff_passing_yards*, *diff_rushing_yards*, *diff_receiving_yards*
- Moving Average: *home_ma_points_scored_3, away_ma_points_scored_3*
- Lagged: *home_lag_points_scored, away_lag_points_scored*
- Efficiency: *home_epa_per_play, away_epa_per_play* (used in our final model)

Model Evaluation:

The model produced a MAD = 1.71 and R^2 = 0.9795, which indicated overfitting since it was trained and tested on known outcomes. Moving average and lagged features, while significant, were not viable for future predictions beyond Week 7. Features like EPA per play, however, remained useful in later models.

On our next model attempt, we excluded the use of moving averages and lagged variables and tested multiple approaches: linear regression (stepwise), GLMs, regularized regressions (Elastic Net, Ridge, Lasso), Random Forest, and XGBoost. Variable selection started with a pool of over 200 features using both automatic techniques (stepwise and feature importance) and manual review of correlations and contextual factors (e.g., stadium, day of week, rest days, win-loss record).

We also tested different rolling game windows (3-, 4-, 6-week averages), finding that using the full six weeks reduced MAD on both cross-validation and test sets. This suggests that using a larger sample of recent games smooths out week-to-week variability in NFL outcomes. Outliers were removed using boxplots to prevent extreme values from skewing results, ensuring model stability.

2.3 Best Model Description

After testing several models, we selected a linear regression as our final approach for spread because it produced similar MAD values to more complex models while remaining interpretable and practical.

The final model includes 5 predictors:

- ***Home average spread:*** Average point spread for the home team in previous weeks
- ***Away average spread:*** Average point spread for the away team in previous weeks
- ***Away average spread line:*** Average betting line (spread) set by bookmakers for the away team in previous weeks

- ***Home average EPA per play:*** Average Expected Points Added (EPA) per play for the home team in previous weeks
- ***Away average EPA per play:*** Average Expected Points Added (EPA) per play for the away team in previous weeks

Model Evaluation:
The model produced a mean MAD = 9.94 on cross-validation and MAD = 9.65 on the 2024 test set. While the MAD for the 2024 season was comparable to the baseline, this model was selected due to its superior R^2 value (0.14 vs. 0.08 for the baseline). This represents a notable improvement, despite the R^2 still being low. The 2024 season was relatively predictable, exhibiting consistently lower errors than other seasons. It likely contained more outliers, and therefore, we put greater emphasis on the cross-validation results, rather than single-season test performance.

Regression Results:

Variable	***Estimate***	***Std. Error***	***t value***	***p-value***	***Significance***
(Intercept)	*0.2825*	*0.8149*	*0.347*	*0.72904*	*–*
home_avg_result	*0.2526*	*0.1142*	*2.212*	*0.02743*	*
away_avg_result	*-0.3119*	*0.1107*	*-2.818*	*0.00503*	**
away_avg_spread_line	*0.6590*	*0.2936*	*2.244*	*0.02529*	*
home_avg_epa_per_play	*25.5995*	*9.6165*	*2.662*	*0.00803*	**
away_avg_epa_per_play	*-25.9224*	*9.6792*	*-2.678*	*0.00766*	**

The model indicates that all predictors except the intercept are statistically significant at the 5% level.

- ***home_avg_result (0.2526):*** Positive coefficient indicates that higher average results for the home team increase the expected spread in their favor.

- ***away_avg_result (-0.3119):*** Negative coefficient shows that higher average results for the away team reduce the expected spread for the home team.
- ***away_avg_spread_line (0.6590):*** The positive coefficient indicates that the average spread line for the away team increases the expected spread of the home team. This conclusion aligns with betting markets.
- ***home_avg_epa_per_play (25.5995):*** Strong positive effect indicates that more efficient offensive performance by the home team increases the expected spread.
- ***away_avg_epa_per_play (-25.9224):*** Strong negative effect shows that more efficient offensive performance by the away team significantly reduces the expected spread for the home team.

The model's low R^2 of 0.1446 indicates that about 14% of the variance in game results is explained by these five variables. The positive and negative coefficients align intuitively with team performance metrics: stronger home teams increase the expected spread, while stronger away teams decrease it.

2.4 Summary

We compared our Week 7 predictions against the Vegas odds, widely believed as a key indicator of outcome. The results were extremely close, with only a few showing strong differentiation. The proximity of our predictions to the industry standard provided us with confidence in their accuracy.

Section 3: Methodology for Total

3.1 Objective

Our objective was to predict the total game points , defined as:

$$Total = Home\ Score + Away\ Score$$

This variable reflects overall scoring and offensive strength for both teams.

3.2 Model Selection Process

We started off with a baseline linear regression using home and away win percentages:

$$Total = \beta_0 + \beta_1 (Home\ Win\ \%) + \beta_2 (Away\ Win\ \%)$$

This model produced a MAD ≈ 15, serving as the benchmark for future models.

Next, we engineered features capturing team offensive efficiency and play style, including passing and rushing statistics. We also tested moving averages for offensive and defensive performance to show short term trends.

Key predictors in the model:

- Performance Variables: *home_passing_yards, away_passing_yards, home_rushing_yards, away_rushing_yards*
- Moving Average Variables: *home_ma_points_scored_3, away_ma_points_scored_3, home_ma_points_allowed_3, away_points_allowed_3*

- Efficiency Variables: *home_epa_per_play, away_epa_per_play*

The model produced MAD = 6.18 and R^2 = 0.7255, but it overfit due to training and testing on known outcomes. Moving averages were predictive but not usable for games past Week 7, and features like EPA were kept for the final model.

3.3 Best Model Description

After testing several models, we selected a linear regression as our final approach for spread because it produced similar MAD values to more complex models while remaining interpretable and practical. This model uses the squared average points for home and away teams to predict total points.

The final best predictive model for total is a linear model incorporating the following predictors:

- ***Home average points scored ^2****:* Square of the home team's average points scored in previous games, capturing potential nonlinear effects.
- ***Away average points scored ^ 2****:* Square of the away team's average points scored in previous games, capturing potential nonlinear effects.

Model Evaluation:

The model produced a mean MAD = 10.72 on cross-validation and MAD = 8.00 on the 2024 test set.

Regression Results:

Variable	***Estimate***	***Std. Error***	***t value***	***p-value***	***Significance***
(Intercept)	*36.2582*	*1.8437*	*19.666*	*<2e-16*	***
I(home_avg_points_for^2)	*0.008994*	*0.002294*	*3.921*	*9.9e-05*	***
I(away_avg_points_for^2)	*0.007535*	*0.002278*	*3.307*	*0.001*	**

The model indicates that all predictors, including the intercept, are statistically significant at the 5% level.

- ***Intercept (36.2582):*** The intercept represents the baseline expected total points when both squared averages are zero.
- ***I(home_avg_points_for^2) (0.008994):*** The positive coefficient indicates that higher squared home points are associated with higher total points, suggesting the effect of strong offensive performance grows slightly more than linearly.

- ***I(away_avg_points_for^2) (0.007535):*** Similarly, higher squared away points contribute positively to total points, though the effect is slightly smaller.

The model has a low R^2 of 0.0466, explaining about 4.7% of the variance in total points. Despite the statistical significance of both predictors, the overall predictive power is limited. Moreover, given the positive squared coefficients, we know that this model cannot accurately capture games with low total scores, as it inherently predicts higher totals even when the underlying averages are small.

3.4 Summary

The total score predictions were compared against the Vegas betting lines. Due to games on average falling between 40-50 points, our predictions were similar to the proposed line. Larger predicted scores of 47+ did correspond to higher-scoring Vegas predictions, which signifies that our model has predictive power.

Section 4: Methodology for Pass

4.1 Objective

The objective was to predict the total passing yards, defined as:

$$Passing = (Home\ Passing\ Yards + Away\ Passing\ Yards)$$

This shows team passing performance and part of their offensive efficiency.

4.2 Model Selection Process

We started off with a baseline linear regression using only home and away average passing yards:

$$Passing = \beta_0 + \beta_1(Average\ Home\ Passing\ Yards) + \beta_2(Average\ Away\ Passing\ Yards)$$

Next, we explored features capturing relative passing tendencies and short term trends.

Key predictors:

- Performance Variables: *home_passing_yards, away_passing_yards, home_rushing_yards, away_rushing_yards*
- Lagged Variables: *home_lag_passing_yards, away_lag_passing_yards*
- Moving Average Variables: *home_ma_passing_yards_3, away_ma_passing_yards_3*
- Efficiency Variables: *diff_pass_rate, epa_per_play* (used in the final model)

Model Evaluation:

The model had MAD = 51.29 and R^2 = 0.696, but overfit due to using known outcomes. Moving averages and lags were predictive but not usable beyond Week 7. In the end, the baseline linear model was used for its simplicity, since the more complex approaches did not improve predictive performance.

4.3 Best Model Description

After testing several models, we selected a linear regression as our final approach for spread because it produced similar MAD values to more complex models while remaining interpretable and practical. This baseline linear model predicts total passing yards based on the previous passing performance of the home and away teams.

The final best predictive model for spread is a linear model incorporating the following predictors:

- ***Home average passing yards:*** Average passing yards gained per game by the home team in previous weeks.
- ***Away average passing yards:*** Average passing yards gained per game by the away team in previous weeks.

Model Evaluation:
The model produced a mean MAD = 83.4 on cross-validation and MAD = 84.1 on the 2024 test set.

Model Results:

Variable	***Estimate***	***Std. Error***	***t value***	***p-value***	***Significance***
(Intercept)	*267.8866*	*39.6729*	*6.752*	*4.2e-11*	***
home_avg_passing_yards	*0.4879*	*0.1156*	*4.222*	*2.9e-05*	***
away_avg_passing_yards	*0.3820*	*0.1108*	*3.449*	*0.000612*	***

The model indicates that all predictors, including the intercept, are statistically significant at the 5% level.

- ***Intercept (267.8866):*** The intercept represents the baseline expected passing yards when both teams' past averages are zero. Though in practice, this metric is mainly a reference point.
- ***home_avg_passing_yards (0.4879):*** A positive coefficient indicates that higher past passing yards by the home team increase the expected total passing yards in the current game.

- ***away_avg_passing_yards (0.3820):*** Higher past passing yards by the away team contribute positively to the expected total passing yards.

The model's R^2 is low (0.0609), meaning that only about 6% of the variance in total passing yards is explained by these two variables. Despite statistical significance, the predictive

power of this simple baseline is limited and reflects the high variability of passing performance in NFL games.

4.4 Summary

Total passing yard predictions from betting organizations were not available for all games. Therefore, we were unable to compare our predictions. However, with a MAD of approximately 84, we believe our predictions will be within one drive of the actual total passing yards.

Conclusion

The goal of this project was to build three models to predict key metrics in football: spread, total points, and total passing yards. Through the overall process of building, cleaning, and integrating various datasets, our team was able to design models that truly captured the valuable elements of NFL team performance and overall efficiency. While our first baseline models offer simple, but informative benchmarks, the engineering of difference-based, efficiency, and performance variables provide vivid insights into game outcomes. Across our spread, total, and passing yard predictions, our findings highlight the challenge of forecasting NFL games: despite the statistically significant predictors, overall explanatory power remained relatively low. Our predictive capacity was further limited by the project requirement to submit forecasts for Weeks 8-11 without access to the preceding weeks' data. Therefore, the team's "dream" model involving the features discussed in the data section was infeasible and not reproducible for later weeks.

Still, the comparison of our predictions to Vegas betting lines suggests that our models hold relevance, especially when approximating spreads and totals. The inclusion of variables such as EPA per play, average results, and squared scoring metrics allowed us to move beyond win percentage alone and better capture offensive and defensive tendencies. The predictive limitations of our passing yards model emphasized that there is variability within each game. This limitation reinforced the general need for larger samples and more specific player-level data in future work.

Ultimately, we believe this project demonstrated both the potential and difficulty of predictive sports analytics. NFL outcomes are influenced by many factors, like injuries, weather, and play-calling decisions, that cannot be fully captured in historical data. Despite these challenges, our models showed meaningful alignment with the industry benchmarks and provided valuable lessons in feature engineering, model evaluation, and predictive trade-offs. Our analysis emphasizes the complexity of sports predictions and showcases how data-driven analysis brings clarity and organization to unpredictable events.