Precision Forecasting in NFL: An Analytical Approach to Passing EPA Prediction*

Yihang Cai

March 28, 2024

This paper presents an analytical approach to predicting the NFL's Expected Points Added (EPA) on passing plays, a crucial metric in understanding quarterback and offensive performance. By integrating historical player performance data and employing linear regression models, we scrutinized on-field metrics such as completions, passing yards, and touchdowns. The inclusion of a rolling average of past EPA in our models captures the dynamic nature of player performance. Results indicate that this historical context is instrumental in enhancing predictive accuracy, with our second model achieving a lower mean absolute error (MAE) compared to the first."

1 Introduction

Predictive modeling in sports analytics has taken center stage in the realm of performance metrics. Among these, the NFL's Expected Points Added (EPA) on passing plays is a vital measure of quarterback effectiveness and team offensive strength. This study aims to build a predictive model for NFL passing EPA by examining historical player and game data. We statistically analyze relationships between passing EPA and various on-field metrics, emphasizing completions, yards, and turnovers.

Incorporating historical performance through a rolling average of EPA, we aim to capture the changing dynamics of player performance. Using linear regression as our foundational modeling technique, we assess the predictive capability of our model in forecasting EPA outcomes. The paper provides insights into the quantifiable aspects of passing performance, offering a statistical lens through which team strategy may be refined.

The subsequent sections are structured as follows: Section 2 introduces the dataset we use. Section 3 explains the model used to predict passing EPA Section 4 compares the predicted

*Code and data are available at: https://github.com/peachvegetable/sta302-mini-essay-12-NFL

value with the testing dataset that splits from the original dataset. Section 5 discusses potential improvements.

2 Data

The dataset is downloaded using 'nflverse' package Carl et al. (2023), which includes 53 variables and we are selecting 8 of them to predict the passing EPA. We use 'dplyr' Wickham et al. (2023) to select the variables we want i.e. player_id, recent_team, season, week, passing_epa, season_type, completions, interceptions, passing_tds, attempts, sacks, passing_yards that are related with passing EPA. Then we use 'zoo' Zeileis and Grothendieck (2005) to create a column for each player according to their historical records of passing EPA to better predict the value (feature engineering).

3 Model

3.1 Model set-up

Define y_i as the passing epa. Then β_0 is the interception and here the passing epa is normally distributed with a mean μ and a standard deviation σ, where the mean depends on eight parameters β_0, and their attempts, sacks, completions, passing yards, passing touchdowns, and interceptions.

$$
\begin{aligned}
y_i|\mu_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \cdot (\alpha_i^{\text{attempts}} + \alpha_i^{\text{sacks}} + \alpha_i^{\text{completions}} + \alpha_i^{\text{passing\_yards}} + \\
&\quad \alpha_i^{\text{passing\_tds}} + \alpha_i^{\text{interceptions}} + \alpha_i^{\text{passing\_epa\_rolling\_avg}}) \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5)
\end{aligned}
$$

We run the model in R (R Core Team 2023) using the tidymodels package of (Kuhn and Wickham 2020).

Table 1: Predicted value from first model

(a) actual vs predicted

Player ID	Season	Week	Passing EPA	Prediction
00-0026158	2023	13	-7.8	-4.4
00-0026158	2023	14	-1.7	5.3
00-0026158	2023	15	-17.4	-2.2
00-0026158	2023	16	19.0	11.6
00-0026158	2023	17	6.2	14.3
00-0026498	2023	11	-4.6	-3.0
00-0026498	2023	12	13.2	9.9
00-0026498	2023	13	5.3	12.3
00-0026498	2023	14	6.5	7.2
00-0026498	2023	15	7.9	7.0

(b) MAE

MAE
3.41

Table 2: Predicted value from second model

(a) actual vs predicted

Player ID	Season	Week	Passing EPA	Prediction
00-0026158	2023	13	-7.8	-6.1
00-0026158	2023	14	-1.7	3.0
00-0026158	2023	15	-17.4	-4.7
00-0026158	2023	16	19.0	9.5
00-0026158	2023	17	6.2	12.6
00-0026498	2023	11	-4.6	-3.9
00-0026498	2023	12	13.2	9.0
00-0026498	2023	13	5.3	11.6
00-0026498	2023	14	6.5	8.2
00-0026498	2023	15	7.9	8.2

(b) MAE

MAE
3.16

3.2 Model justification

The model's predictive accuracy was assessed using mean absolute error (MAE), revealing the distance between the predicted and actual passing EPA. The first model yielded an MAE of 3.41, indicating the average magnitude of prediction errors. Comparative analyses between actual and predicted passing EPA were conducted for two models. As illustrated in Table 1 and Table 2, these comparisons highlight the model's performance across different weeks of the NFL season for selected players.

The first model, while robust in its explanatory power, exhibited a tendency to underpredict the higher range of passing EPA, as evident from the over-performance instances in Week 16 and Week 17. On the contrary, for the second model, the introduction of the rolling average feature for passing EPA significantly improved the precision, reducing the MAE to 3.16, demonstrating the value of incorporating historical performance trends into predictive analyses.

Therefore, since the second model has smaller error by MAE ($3.16 < 3.41$), the feature engineering of considering the historical records of passing EPA makes the model to predict in more accuracy and precision.

4 Results

Table 3 displays the estimated coefficients for two linear regression models predicting NFL passing EPA. The values outside the parentheses are the point estimates of the coefficients, and the values inside the parentheses are the standard errors, which measure the precision of the coefficient estimates. Intercept: Both models estimate a negative intercept, but the exact value is smaller in the second model. The intercept is the predicted value of passing EPA when all other predictor variables are zero. Completions: The positive coefficients for completions in both models suggest that a higher number of completions is associated with an increased passing EPA. The coefficient is slightly higher in the second model, indicating a marginally greater impact per completion. Passing Yards: These positive coefficients indicate that more passing yards are associated with higher passing EPA. The effect is slightly less in the second model. Passing Touchdowns (tds): A strong positive relationship is shown here, as touchdowns have a significant impact on passing EPA, with each touchdown contributing more in the first model than in the second. Interceptions: Negative coefficients for interceptions imply that they have a detrimental effect on passing EPA, with a slightly less negative impact in the second model. Sacks: The negative coefficients suggest that sacks negatively affect passing EPA, with the impact being somewhat reduced in the second model. Attempts: The negative coefficients indicate that more pass attempts may not necessarily lead to higher EPA, suggesting diminishing returns or inefficiencies. Passing EPA Rolling Average: Present only in the second model, the positive coefficient for the rolling average of passing EPA indicates that recent historical performance is a significant predictor of current performance.

Table 3: Model summary of first and second models

	first	second
(Intercept)	−0.848	−0.608
	(0.582)	(0.547)
completions	0.430	0.451
	(0.096)	(0.090)
passing_yards	0.095	0.079
	(0.006)	(0.006)
passing_tds	1.783	1.502
	(0.282)	(0.267)
interceptions	−3.750	−3.388
	(0.296)	(0.283)
sacks	−1.934	−1.615
	(0.131)	(0.132)
attempts	−0.763	−0.689
	(0.061)	(0.059)
passing_epa_rolling_avg		0.264
		(0.040)
Num.Obs.	318	318
R2	0.850	0.869
R2 Adj.	0.847	0.866
AIC	1798.5	1757.8
BIC	1828.6	1791.6
RMSE	3.99	3.73

5 Discussion

5.1 Interpretation of Results

The study's findings elucidate the intricate dynamics of NFL passing plays. The slight reduction in MAE from the first to the second model underscores the importance of historical context, as players' past performances offer a gauge for future outcomes. The models' comparative analysis affirms the complexity of predicting sports metrics, where even small enhancements in model features can yield notable improvements in accuracy.

5.2 Practical Implications

Beyond theoretical exploration, these findings have practical implications for coaching strategies, betting markets, and player evaluations. The ability to forecast passing EPA with greater accuracy enables teams to make informed decisions, tailor training to individual player profiles, and adapt in-game tactics more dynamically.

5.3 Third discussion point

5.4 Weaknesses and next steps

Despite the strengths of the models, limitations persist. For instance, the exclusion of certain situational variables and defensive metrics may hinder the model's comprehensive interpretative capacity. Future research endeavors should aim to integrate additional layers of data, such as in-game decision-making, player fatigue levels, and real-time defensive adjustments, to enhance the model's predictive scope.

Additionally, exploring non-linear modeling techniques and machine learning algorithms could address the complex interactions within the data that linear models may overlook.

References

Carl, Sebastian, Ben Baldwin, Lee Sharpe, Tan Ho, and John Edwards. 2023. *Nflverse: Easily Install and Load the 'Nflverse'*. https://CRAN.R-project.org/package=nflverse.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles*. https://www.tidymodels.org.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zeileis, Achim, and Gabor Grothendieck. 2005. "Zoo: S3 Infrastructure for Regular and Irregular Time Series." *Journal of Statistical Software* 14 (6): 1–27. https://doi.org/10.18637/jss.v014.i06.