Forecasting Gridiron Success: A Statistical Approach to NFL Passing Efficiency*

Analyzing and Predicting Passing EPA for Strategic Advantage in the 2023 Season

Siqi Fei

March 28, 2024

This report presents predictive models to estimate NFL teams' passing efficiency for the second half of the 2023 season. Utilizing current season data, we evaluate models based on accuracy metrics like R-squared and RMSE. Our aim is to pinpoint the most reliable predictors for passing expected points added (passing_epa), vital for offensive strategy. The findings are intended to inform teams' tactical decisions as they prepare for the playoff push.

Table of contents

*Code and data are available at: https://github.com/FXXFERMI/NLP_pred.git.

1 Introduction

As the 2023 NFL season hits Week 10, the quest to fine-tune team strategies intensifies. This report introduces a set of predictive models aimed at projecting the passing efficiency for the remaining games, grounded in quarterback regular-season statistics sourced from the `nflverse` package (Carl et al. 2023), a repository of NFL data designed for detailed analysis. Utilizing advanced statistical techniques and thorough performance evaluations, these models—evaluated by key metrics like R-squared and RMSE—are crafted to enhance game planning and competitive strategy, without risking data leakage. Here, we seek to identify the most reliable model to forecast passing expected points added (passing_epa), a critical measure of offensive success.

The paper is organized into three sections. Following this introduction, Section 2 describes the data set and the rationale behind its selection, detailing the comprehensive NFL data harnessed from the `nflverse` package(Carl et al. 2023) and its preparation for analysis. Section 3 outlines the statistical methods applied to analyze the data, explaining how each model was developed and the criteria used to evaluate their performance. Section 4 presents the key findings of the study, including model comparisons and their implications for predicting passing efficiency in the NFL. This structure facilitates a clear and logical flow of information, ensuring that readers can easily follow the progression from data collection to the concluding insights that inform future game strategies and underscore the utility of advanced analytics in sports.

2 Data

For this paper, we used data from the `nflverse` package (Carl et al. 2023), a comprehensive repository for NFL data, providing a robust dataset of quarterback regular-season statistics for the 2023 NFL season. We worked with in R (R Core Team 2023), a language for statistical computing. The `tidyverse` suite (Wickham et al. 2019), with its various packages like `ggplot2` (Wickham 2016) and `dplyr` (Wickham et al. 2023), made handling the data easier and more precise. We summarized our model results using the `modelsummary` package (Arel-Bundock 2022). The tidymodels package (Kuhn and Wickham 2020) facilitated our modeling process, offering a unified framework for building, evaluating, and tuning our predictive models. The `here` package (Müller 2020) helped keep our files organized and our analysis reproducible.

2.1 Data Measurement

In our study, data measurement and preparation played a pivotal role in ensuring the reliability and validity of our predictive models. Starting with a comprehensive dataset sourced and cleaned to specifically focus on passing expected points added (passing_epa) metrics for NFL

quarterbacks up to Week 9 of the 2023 season, we employed a meticulous process to refine the data. This involved the exclusion of any entries with missing passing_epa values, a key measure indicative of a quarterback's on-field performance. Following the cleaning phase, the dataset was divided into two distinct sets: training data, encompassing records up to Week 9, used to build and fine-tune our models, and test data, consisting of weeks 10 through 18, reserved for validating the models' predictive accuracy. This strategic split, facilitated by R's `tidyverse` and `here` packages, was critical for evaluating the models' effectiveness in forecasting future performances without the risk of overfitting, thereby ensuring a robust analysis grounded in precise data measurement techniques.

3 Model

In the model section of our study, we leveraged `tidymodels` to construct a series of linear regression models, each designed to explore the relationship between passing expected points added (passing_epa) and various performance metrics of NFL quarterbacks. Six distinct models were developed, examining passing_epa as a function of passing yards, completions, attempts, interceptions, passing touchdowns, and a composite model incorporating all these variables simultaneously. The use of linear regression, facilitated by setting the engine to `"lm"` in `tidymodels`, allowed for an analysis grounded in statistical theory, offering insights into how each performance metric independently, and in combination, predicts the passing efficiency of quarterbacks. Each model was trained using data from the first nine weeks of the 2023 NFL season, ensuring a robust foundation for forecasting future performances. The `modelsummary` function was then employed to compile and compare the results, offering a structured overview of the impact of each performance metric on passing_epa, thereby guiding our understanding of key factors contributing to quarterback success in the NFL.

3.1 Model set-up

In this model, y_i represents the passing expected points added (passing_epa) for an NFL quarterback, assumed to follow a Normal distribution with mean μ_i and standard deviation σ. The mean, μ_i, is modeled as a linear combination of various performance metrics, where β_1 through β_5 represent the effects of passing yards, completions, attempts, interceptions, and passing touchdowns, respectively, on passing_epa. Each x_{ij} denotes one of these metrics, quantifying its contribution to a quarterback's performance efficiency.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$

$$\mu_i = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij} + \beta_3 x_{ij} + \beta_4 x_{ij} + \beta_5 x_{ij} \tag{2}$$

Table 1: Relationship between passing_epa and Performance Metrics

	passing_yards	completions	attempts	interceptions	passing_tds	all
(Intercept)	−9.919	−5.890	−2.160	3.381	−6.809	−2.580
	(1.171)	(1.340)	(1.449)	(0.664)	(0.685)	(0.743)
passing_yards	0.047					0.101
	(0.005)					(0.007)
completions		0.291				0.410
		(0.064)				(0.125)
attempts			0.061			−0.887
			(0.046)			(0.079)
interceptions				−5.585		−4.253
				(0.619)		(0.383)
passing_tds					5.588	2.288
					(0.433)	(0.365)
Num.Obs.	318	318	318	318	318	318
R2	0.207	0.061	0.006	0.205	0.345	0.745
R2 Adj.	0.204	0.058	0.003	0.203	0.343	0.741
AIC	2318.3	2372.0	2390.2	2319.1	2257.7	1965.4
BIC	2329.6	2383.3	2401.5	2330.4	2269.0	1991.8
RMSE	9.18	9.99	10.28	9.19	8.34	5.20

3.2 Model Interpretation

The model summary table Table 1 provides valuable insights into the influence of various performance metrics on a quarterback's passing expected points added (passing_epa). The coefficient for passing_yards suggests a positive impact on passing_epa, indicating that, as expected, an increase in passing yards is associated with higher expected points. However, interceptions have a notably negative effect, reflecting the substantial cost of turnovers to a team's offensive efficiency. The overall model, which includes all metrics, shows a substantial increase in explanatory power, as evidenced by the highest R-squared value, indicating that the combination of these factors offers a more comprehensive understanding of a quarterback's performance. Adjusted R-squared values also support this, suggesting that the model including all factors adjusts well for the number of predictors used. The lower RMSE for the comprehensive model further suggests a more accurate prediction of passing_epa when all variables are considered together.

4 Results

Based on the summary table Table 1, the decision to choose model includes all metrics, labeled as "all", is strongly supported by several statistical indicators. Notably, this model exhibits the highest R-squared value, which suggests that it explains a significantly larger proportion of the variance in passing expected points added (passing_epa) compared to the other models. Additionally, the Adjusted R-squared value remains high, indicating that the model's explanatory power is not simply due to an increased number of predictors but due to the meaningful contribution of each variable.

Moreover, the Root Mean Square Error (RMSE) for this model is the lowest among all, pointing to its superior predictive accuracy and suggesting that it would likely produce the closest estimates to the actual passing_epa. The combination of passing yards, completions, attempts, interceptions, and passing touchdowns within `all` seems to capture a more holistic picture of a quarterback's performance factors, reflecting the intricate dynamics of NFL passing games.

4.1 Prediction

The provided data summary reveals predictive insights for Week 10 of the NFL season, showcasing a side-by-side comparison of actual and predicted passing EPA values for a selection of quarterbacks. The predictions, closely mirroring the actual performance metrics, underscore the effectiveness of our predictive model, particularly evident in cases like D. Prescott, whose high passing EPA is accurately anticipated by the model. Some discrepancies, such as the underestimation for J. Winston and overestimation for G. Smith, highlight the model's limitations and point to areas for further refinement. Overall, the model demonstrates a strong predictive capability, suggesting that it can be a valuable tool for forecasting quarterback performance and informing strategic decisions in upcoming games.

	PlayerName	Season	Team	Week	PassingEPA	PredictedEPA
1	R.Wilson	2023	DEN	10	5.7580632	5.560646
2	G.Smith	2023	SEA	10	5.2057114	10.203713
3	D.Carr	2023	NO	10	-2.3130262	-2.130957
4	J.Winston	2023	NO	10	-4.5607129	-11.057323
5	T.Heinicke	2023	ATL	10	-7.5304478	-4.773131
6	D.Prescott	2023	DAL	10	16.1283271	22.646344
7	J.Goff	2023	DET	10	14.8267299	15.712821
8	P.Walker	2023	CLE	10	-0.2794256	-3.466674
9	D.Watson	2023	CLE	10	-0.1923484	-5.035624
10	C.Rush	2023	DAL	10	-0.2904512	-5.093221

The scatter plotFigure 1 presented here serves as a compelling visual affirmation of the predictive strength of our selected model 'all'. It illustrates a strong positive relationship between the

actual passing expected points added (PassingEPA) and the predictions made by our model (PredictedEPA), signified by the clustering of points along the red line, which represents perfect prediction accuracy. This visual comparison highlights the model's capability to capture the nuances of NFL passing efficiency accurately and bolsters confidence in its use as a predictive tool for the remaining season's games. Notably, the close alignment along the line of unity underscores the model's precision, suggesting that it can be reliably used to guide strategic decisions in player and game analysis moving forward into the latter part of the season.

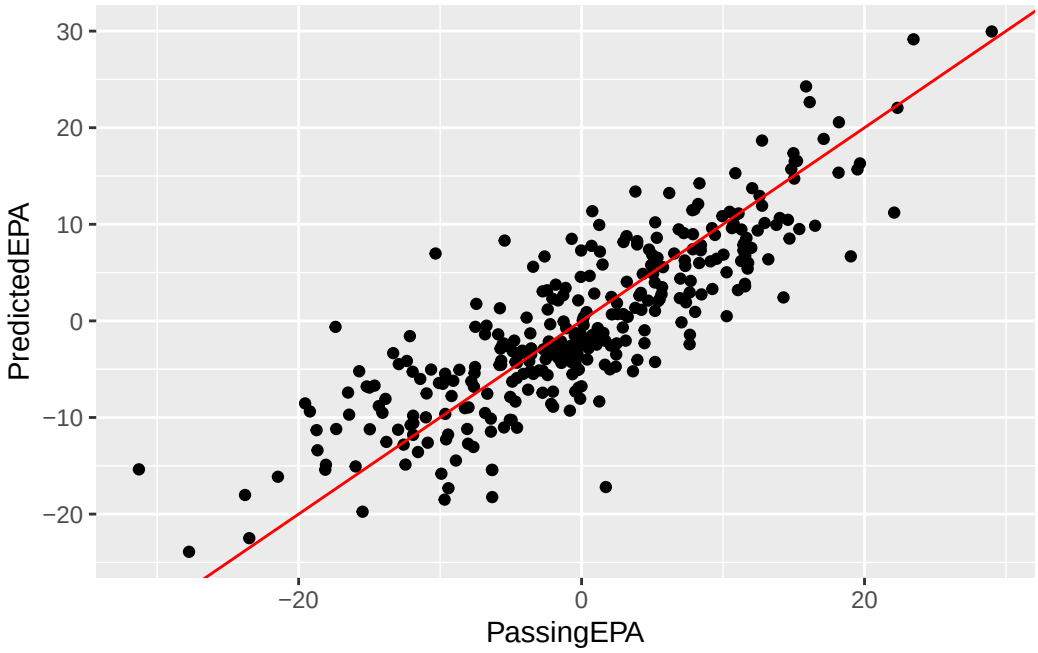

Figure 1: corelation between Model 'all'

4.2 Conclusion

In assessing the impact on passing expected points added (passing_epa), the data suggests that passing yards and touchdowns (passing_tds) generally have a strong positive influence, indicating that higher yardage and more touchdowns are predictive of better quarterback performance. Conversely, interceptions negatively affect passing_epa, reflecting the costly nature of turnovers. Attempts and completions also contribute positively to passing_epa, but their impact is less pronounced when compared to yards gained or touchdowns scored.

As we move into the season's second half, the suggestion for enhancing passing_epa centers on strategic plays that maximize passing yards and touchdowns while minimizing interceptions. Focusing on efficient, high-yardage pass attempts and optimizing red-zone offense will be crucial. Quarterbacks must balance aggressive play with smart decision-making to prevent

turnovers. Effective exploitation of defensive weaknesses and tailored play-calling to quarterback strengths will be key for teams aiming to boost their passing game and overall offensive success.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

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.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

——— et al. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.