Forecasting Quarterback Efficiency with Passing EPA for the Latter Half of the NFL 2023 Season*

Jiwon Choi

March 28, 2024

This study introduces a novel approach to forecasting quarterback efficiency in the NFL by leveraging passing Expected Points Added (EPA) through logistic regression. Recognizing the pivotal role quarterbacks play in influencing game outcomes, we sought to predict their performance with respect to EPA - a metric that encapsulates the value added on a per-play basis. By transforming passing EPA into a binary variable—above or below the median—we employed logistic regression to model quarterback efficiency, drawing on data from the first nine weeks of the 2023 NFL season. The methodology not only highlights significant performance indicators but also offers strategic insights for teams. Despite inherent limitations, our model demonstrates a robust framework for predicting quarterback performance, with implications for enhancing team strategies and understanding game dynamics.

1 Introduction

The quarterback position in American football is often heralded as the most crucial within the sport, with a quarterback's performance significantly influencing the outcome of a game. Given this importance, accurate prediction of quarterback efficiency is vital for team strategy and game planning. This study aims to forecast quarterback efficiency for the latter half of the NFL 2023 season using a statistical approach grounded in passing Expected Points Added (EPA).

Passing EPA is a sophisticated metric that quantifies the contribution of each passing play to the team's scoring potential, adjusting for the context of the play. Our research employs

[*] Code and data are available at: https://github.com/jwonc4602/nfl-prediction.

logistic regression, a predictive modeling technique well-suited for binary outcomes, to predict whether a quarterback's performance in terms of passing EPA will be above or below the season's median. This binary classification facilitates a clear, actionable insight into quarterback performance expectations.

By analyzing quarterback statistics from the first half of the 2023 NFL season, this study applies logistic regression to project future performances. The choice of logistic regression was driven by its effectiveness in handling binary data and its capacity to provide probabilistic outcomes, offering a nuanced view of quarterback efficiency that is both informative and practical for team decision-making.

This paper outlines the methodology for constructing and applying the logistic regression model, discusses the implications of our findings, and considers the model's limitations and potential areas for future research. Through this analysis, we contribute to the broader field of sports analytics by providing a model that can predict quarterback efficiency with notable accuracy, offering valuable insights for teams, analysts, and enthusiasts of the NFL.

2 Data

In this project, I sourced the dataset through the 'nfl_data_py' Python API (cooperdff 2023), focusing on the 2023 NFL season. The data collection and analysis were performed using Python (Van Rossum and Drake 2009), with the support of libraries such as pandas (McKinney et al. 2010) and numpy (Harris et al. 2020).

The specific variables chosen for analysis were 'player_id', 'season', 'week', 'season_type', 'passing_epa', 'completions', 'attempts', 'passing_yards', 'passing_tds', and 'interceptions'. To forecast the second half of the NFL 2023 season, I cleaned the dataset, retaining only the first 9 weeks of data and excluding any week beyond week 9. (see Figure 1)

	player_id	season	week	season_type	passing_epa	completions	attempts	passing_yards	passi
0	00-0023459	2023	1	REG	-2.031960	0	1	0.0	0
1	00-0024243	2023	4	REG	0.000000	0	0	0.0	0
2	00-0024243	2023	7	REG	0.000000	0	0	0.0	0
3	00-0026498	2023	1	REG	20.679981	24	38	334.0	0
4	00-0026498	2023	2	REG	-5.089193	34	55	307.0	1
5	00-0026498	2023	3	REG	-8.404790	18	33	269.0	1
6	00-0026498	2023	4	REG	11.374351	27	40	319.0	1

Figure 1: Sample of Cleaned NFL Data

3 Model

For our research, we implement a forecasting model utilizing logistic regression, a technique well-suited for predicting binary outcomes. Our objective is to forecast a quarterback's passing EPA (Expected Points Added) performance, categorizing it as above or below the median value. This process involves two main steps: constructing a logistic regression model with a selected dataset and then applying this model for prediction.

Initially, we select a dataset for developing our model, in this case, quarterback statistics from the 2023 NFL season. These statistics include completions, attempts, passing yards, touchdowns, and interceptions, which serve as our predictive features. The target variable, passing_epa, is transformed into a binary outcome based on whether the EPA is above or below its median value within the dataset. This binary formulation is crucial for applying logistic regression, which is inherently designed for dichotomous outcomes.

The logistic regression model is defined as follows:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{completions} + \beta_2 x_{attempts} + \beta_3 x_{passing_yards} + \beta_4 x_{passing_tds} + \beta_5 x_{interceptions} \quad (1)$$

In equation 1, $\hat{p}$ represents the probability of a quarterback's performance being above the median EPA, and each β denotes a coefficient estimated through logistic regression analysis. The features chosen—completions, attempts, passing yards, touchdowns, and interceptions—are critical indicators of a quarterback's performance, offering a comprehensive view of their ability to contribute positively to the game's outcome.

Following the development of the logistic regression model, we employ a Python script for the analysis. The train_model function is central to this process, facilitating the model training using the scikit-learn library (Pedregosa et al. 2011). This library provides a robust platform for implementing logistic regression and evaluating the model's performance based on accuracy.

The final step involves applying our model to new or unseen data, allowing us to forecast quarterbacks' performances in terms of EPA. By assessing the accuracy of our predictions, we can gauge the model's efficacy and refine it for improved future forecasts.

Our use of logistic regression, especially within the framework of sports analytics, underscores its versatility and effectiveness in predicting binary outcomes. However, it's crucial to acknowledge the model's limitations, including its dependence on the quality and relevance of the input features. Additionally, while logistic regression can offer insights into the likelihood of specific outcomes, it does not account for the intricacies of game dynamics or individual player conditions.

3.1 Discussion

In our investigation, we endeavored to forecast quarterback efficiency through passing EPA (Expected Points Added) for the second half of the NFL 2023 season. The methodology centers on logistic regression, a statistical technique adept at handling binary outcomes. The essence of our approach was to categorize quarterbacks' passing EPA performances into binary classifications: performances above or below the median EPA value. This dichotomy allows us to employ logistic regression effectively, providing a probabilistic understanding of a quarterback's efficiency relative to a defined benchmark.

The logistic regression model is formulated to predict the likelihood ($\hat{p}$) of a quarterback's passing EPA being above the median, considering key performance indicators such as completions, attempts, passing yards, touchdowns, and interceptions. These variables were meticulously chosen based on their established influence on a game's scoring dynamics, thereby acting as a proxy for quarterback efficiency.

A pivotal aspect of our analysis was the binary transformation of the `passing_epa` variable, which was essential for adapting the dataset to the logistic regression framework. This approach, while simplifying the outcome into a binary variable, allows for a nuanced interpretation of quarterback performances across different game contexts and against varying defenses.

The construction of our logistic regression model was facilitated by the use of Python's scikit-learn library (Pedregosa et al. 2011), a choice motivated by its robustness and ease of use for implementing and evaluating statistical models. Our analysis process involved not only the training of the model using historical data from the first half of the 2023 season but also the validation of its predictive power on unseen data.

The prediction of quarterback efficiency using passing EPA (Expected Points Added) through logistic regression directly relates to understanding and forecasting how effectively a quarterback contributes to their team's scoring opportunities. In the context of NFL analytics, `passing_epa` is a sophisticated metric that quantifies the value a player adds to their team's expected point tally on a per-play basis, adjusted for the situation of the play (e.g., down, distance, field position).

3.1.1 Relationship Between Predictions and Quarterback Efficiency

1. **Quantifying Impact**: By predicting whether a quarterback's performance (in terms of passing EPA) will be above or below the median, we're essentially gauging their impact on the game. Quarterbacks predicted to perform above the median are viewed as having a more significant positive impact on their team's scoring chances.

2. **Performance Benchmarks**: The median passing EPA serves as a benchmark for efficiency. Predicting a quarterback's performance relative to this benchmark allows teams

and analysts to understand how a quarterback's performance compares to league-wide standards.

3. **Decision-Making and Strategy**: The predictions inform coaching staff and management about potential quarterback performance, aiding in strategic decisions such as play-calling, game planning, and personnel adjustments. For example, a quarterback forecasted to perform well against certain defenses might influence game strategies to exploit these matchups.
4. **Evaluating Key Metrics**: The model considers various performance metrics (completions, attempts, passing yards, touchdowns, interceptions) that are intrinsically related to a quarterback's efficiency. Predictions based on these metrics offer insights into which aspects of a quarterback's game are most likely to contribute to efficient scoring opportunities.
5. **Binary Outcome Limitations**: While the binary prediction (above or below median EPA) provides a simplified view of quarterback efficiency, it highlights the significance of `passing_epa` as an evaluative tool. However, it does not capture the magnitude of efficiency or the specific value a quarterback adds beyond the median level.

3.1.2 Implications

The use of logistic regression to forecast quarterback efficiency via `passing_epa` reflects an analytical approach to understanding the multifaceted role of quarterbacks in American football. The predictions not only serve as an evaluative benchmark but also as a strategic tool for enhancing team performance. By focusing on the probability of a quarterback exceeding a median performance threshold, teams can better understand potential game outcomes and adjust their strategies accordingly.

However, it's important to recognize the model's scope and the complexity of football dynamics. Quarterback efficiency, while critical, is just one piece of the puzzle. Team success depends on a multitude of factors, including defense, special teams, and coaching strategies. Future models might integrate these additional dimensions to offer a more holistic view of game outcomes and player contributions.

3.2 Results

Upon applying the logistic regression model to the test dataset, we observed a commendable level of accuracy in forecasting quarterback performances in terms of passing EPA. The model's predictions, delineated as probabilities, provide insights into the likelihood of a quarterback's efficiency surpassing the median benchmark in forthcoming games.

Our findings underscore the significance of key performance metrics in forecasting quarterback efficiency. For instance, quarterbacks with higher completion rates and passing yards tend to have a higher probability of exceeding the median passing EPA, highlighting the importance of accuracy and offensive play in determining quarterback performance.

However, it’s crucial to acknowledge the limitations inherent in our approach. While logistic regression offers a structured method for prediction, the binary outcome does not encapsulate the magnitude of a quarterback’s efficiency—merely the probability of it being above a certain threshold. Furthermore, external factors such as team strategy, opponent strength, and environmental conditions, which could significantly impact game outcomes, are not directly accounted for in the model.

Despite these limitations, our model presents a viable framework for forecasting quarterback performance in the NFL, offering valuable insights for coaches, analysts, and fans alike. Future research could explore more complex models or incorporate additional variables to enhance predictive accuracy and provide a more comprehensive analysis of quarterback efficiency.

References

cooperdff. 2023. "nfl-data-py: Python Library for Interacting with NFL Data." https://pypi.org/project/nfl-data-py/.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

McKinney, Wes et al. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, 445:51–56. Austin, TX.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (Oct): 2825–30.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace.