Forecasting NFL Quarterback Performance: A Mid-Season Analysis*

Pu Yuan

March 30, 2024

This report aims to predict how well NFL quarterbacks will play in the second half of the 2023 season. I used a straightforward method to fill in missing data in my statistics by replacing them with average values. Then, I created a model that looks at how different factors, like a quarterback's past performance, might predict future performance. My basic model gives me an initial look at what to expect from the quarterbacks in the upcoming games. While my approach is simple, it sets the stage for more detailed analysis in the future, offering teams a starting point for improving their strategies.

1 Introduction

Halfway through the 2023 NFL season, the analysis of quarterback performance is crucial for understanding and forecasting future outcomes. The passing Expected Points Added (EPA) serves as a significant indicator of a quarterback's effectiveness and contribution to the team's offensive efforts. This essay explores the development of a predictive model to forecast passing EPA for each NFL team's quarterbacks for the remaining games of the season.

The rest of the paper is organized as follows: The Section 2 represents the data. The Section 3 shows how modeling works. The Section 4 presents the results. The Section 5 Gives insight of the research.

We use the statistical programming language `R` (R Core Team 2023). We also made use of the following `R` packages: `readr` (Wickham, Hester, and Bryan 2024), `tidyverse` (Wickham, Vaughan, and Girlich 2024), `dplyr` (Wickham et al. 2023), `caret` (Kuhn and Max 2008).

[*] Code and data are available at: https://github.com/scottyuan6/prediction.git

2 Data

In the dataset, I have a wealth of quarterback statistics from the current 2023 NFL season up to Week 9. However, I encounter a common issue in data analysis: missing values in several crucial columns. To address this, I apply a simple imputation strategy, filling in missing numeric data with the mean of each respective column. This method maintains the structure of the data and allows me to proceed without discarding valuable information. Although this approach assumes that the missing data is randomly distributed and that the mean is a reasonable estimate for the missing values, it is a widely accepted practice for initial analyses and is straightforward to implement.

3 Model

With a complete dataset, I proceed to enhance the model's predictive capabilities through feature engineering. I create new variables that reflect potential influences on passing EPA, such as pass completion ratios. Using these features, I prepare to build the predictive model.

For the modeling phase, I employ a linear regression approach as a starting point, due to its simplicity and interpretability. A linear regression model serves as a useful baseline and can provide quick insights into the relationships between features and the target variable, passing EPA.

I utilize the train function from the caret package, allowing for a streamlined modeling process that includes built-in cross-validation. This function simplifies the task of model selection and tuning, providing a robust framework for evaluating model performance.

4 Result

Upon training our linear model, we assess its performance using the built-in summary function. This function reveals the significance of each feature in predicting passing EPA and the overall fit of the model. For example, the model's coefficients for completion ratio and other features provide insights into their relative importance.

5 Discussion

The simplicity of the imputation approach and the transparency of the linear regression model provide me with a clear, although preliminary, picture of quarterback performance influences. My results highlight key factors that can be targeted for improvement and suggest strategies teams might employ for the remainder of the season.

However, it's important to recognize the limitations of mean imputation and linear modeling. Mean imputation does not account for the potential structure in missing data, which could lead to biased estimates. Similarly, linear regression assumes a linear relationship between features and the target variable, which may not always hold in complex sports scenarios. Future work could explore more sophisticated imputation methods and advanced models such as ensemble methods or neural networks for improved prediction accuracy.

By refining my techniques and expanding the dataset with real-time data and more advanced metrics, I can further enhance the predictive power of our models, providing teams with actionable insights that can inform strategy and decision-making in the highly competitive environment of the NFL.

References

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

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://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data.* https://readr.tidyverse.org.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data.* https://tidyr.tidyverse.org.