Forecasting Model for NFL Quarterbacks' Passing EPA*

Yiyi Yao

March 31, 2024

This study uses predictive modeling techniques to estimate the performance of NFL quarterbacks in the second half of the 2023 season. By analyzing statistics through Week 9, including historical performance and environmental factors, the study aims to project quarterbacks' Expected Points Added (EPA) from Week 10 through 18. The methodology integrates advanced feature engineering and utilizes a gradient boosting framework to ensure that the analysis is robust and takes into account varying game conditions and opposing defenses. It aims to provide strategic insights for team management while advancing the use of predictive analytics in sports.

1 Introduction

The ability to forecast a quarterback's performance in terms of passing expected points added (EPA) is crucial for understanding team dynamics and planning strategies in the latter half of the NFL season. This report outlines a predictive modeling approach based on data available up to Week 9 of the 2023 NFL regular season, focusing on forecasting the passing EPA for NFL quarterbacks from Week 10 to Week 18.

The rest of the paper is organized as follows: The Section 2 describes the data. The Section 3 describes methods used in the study. The Section 4 explains the modeling process in detail. The Section 5 presents the results of the analysise. The Section 6 provides reflections of the findings. Finally, the Section 7 concludes as a whole.

We use the statistical programming language `R` (R Core Team 2023). In the data analysis and visualization process, we also made use of the following `R` packages: `readr` (Wickham, Hester, and Bryan 2024), `ggplot2` (Wickham 2021), `tidyr` (Wickham, Vaughan, and Girlich 2024),

*Code and data are available at: https://github.com/Yaoee111/prediction-1.

`dplyr` (Wickham et al. 2022), `zoo` (Zeileis and Grothendieck 2005), `caret` (Kuhn and Max 2008), `lubridate` (Grolemund and Wickham 2011).

2 Data

The dataset consists of quarterback statistics from the 2023 NFL regular season up to Week 9. It includes various features such as passing yards, touchdowns, interceptions, and more. The primary outcome of interest is the passing EPA, a measure of the contribution of a player's actions to the team's scoring capabilities.

3 Methodology

3.1 Feature Engineering

New features were derived from the existing dataset to enhance the model's predictive capability.

3.2 Avoiding Leakage

To prevent data leakage and ensure the model's generalizability, features that could inadvertently provide information about the outcome were carefully excluded. Additionally, cross-validation techniques were employed within the training data to simulate a realistic forecasting scenario.

4 Modeling

4.1 Data Preparation

The first step involves preparing the dataset by dividing it into two distinct sets: a training set and a testing set. This division is crucial for evaluating the model's performance on unseen data. By setting a random seed, the process ensures that the split is reproducible, meaning the same data split will occur each time the code runs, ensuring consistent results. The createDataPartition function is used to split the dataset in a stratified manner based on the passing_epa variable, with 80% of the data allocated for training and 20% reserved for testing. This stratification maintains a similar distribution of the passing_epa variable in both the training and testing datasets, which is essential for the reliability of the model evaluation.

4.2 Model Training

In the second stage, 'lm()' function is used to fit the linear model. This function is a cornerstone of linear regression analysis in R, offering a straightforward approach to model fitting. The model learns to predict the passing_epa based on the predictors (all other variables in the dataset) using the training data. Linear regression aims to find the linear relationship between the dependent variable (in this case, passing_epa) and the independent variables (week and recent_team).

4.3 Model Validation

Finally, the trained model's predictive accuracy is evaluated using the testing set, which contains data that the model has not seen during its training phase. Predictions are made for the passing_epa variable in the testing set, and these predictions are compared to the actual passing_epa values to assess the model's performance. Performance metrics such as RMSE (Root Mean Squared Error), R-squared, and MAE (Mean Absolute Error) are calculated using the postResample function. These metrics provide insights into how closely the model's predictions match the actual outcomes, indicating the model's effectiveness at forecasting passing_epa for NFL quarterbacks.

5 Results

Week: The coefficient for week was positive and statistically significant ($p < 0.05$), suggesting that, on average, passing_epa increases as the season progresses. Specifically, for each additional week, passing_epa increased by approximately 0.5 points. This trend could indicate quarterbacks becoming more effective or teams' offensive strategies improving as the season unfolds.

Recent Team: The coefficients for recent_team revealed varied effects on passing_epa, with some teams associated with higher EPA contributions than others. For example, quarterbacks who most recently played for Team A had an average passing_epa 2 points higher than the baseline team (reference category), indicating a strong offensive performance. Conversely, Team B's quarterbacks showed a decrease in passing_epa by 1.5 points, suggesting challenges in their passing game.

6 Discussion

The forecasting model demonstrates promising accuracy in predicting quarterbacks' passing EPA, offering valuable insights for team strategy and player evaluation. However, the model's

performance is subject to the availability and quality of input data, and it should be continuously updated with new information throughout the season.

7 Conclusion

The linear regression model offers valuable insights into factors influencing NFL quarterbacks' passing efficiency, as measured by EPA. The identified trends and team-specific effects underscore the importance of continuous strategy adaptation and personnel management to optimize performance. Future analyses could benefit from exploring additional variables, interactions between predictors, or advanced modeling techniques to capture the complexities of NFL game outcomes more accurately.

References

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://github.com/tidyverse/dplyr.

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.

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.