Analysis of Performance and Predictive Modeling in the NFL Regular Season*

Xiyou Wang

2024-04-04

Contents

1 Introduction

The quarterback position in the National Football League (NFL) is often considered the most pivotal role. The ability to predict the performance of the quarterback can provide valuable insights for teams and coaches to help them develop game plans. The purpose of this report is to analyze the performance of the quarterback during the 2023 NFL regular season and to use a linear regression model to predict the expected passing increase (EPA). We will evaluate the predictive ability of this model to provide guidance for decision-making for the remainder of the season.

*Code and data are available at: https://github.com/wxywxy666/NFL-analysis.

2 Methodology

2.1 Data Collection

The analysis was performed in R code (R Core Team 2024), using data from the NFLverse package (Carl et al. 2023), which compiles extensive statistics on NFL games and players. The data consists of various performance metrics for quarterbacks, including completions, attempts, passing yards, touchdowns, and interceptions, along with the `passing_epa`, which is a key measure of a player's contribution to the team's scoring chances.

2.2 Data Preparation

Packages parsnip (Kuhn and Vaughan 2024), dplyr (Wickham et al. 2023), tidymodels (Kuhn and Wickham 2020) and yardstick (Kuhn, Vaughan, and Hvitfeldt 2024) are used to clean the raw dataset.

The cleaned dataset was filtered to isolate the performance metrics for quarterbacks during regular season games. A subset of the data from weeks 1 to 9 of the 2023 season was used to train the regression model, while the remainder of the data from previous seasons served as the testing set to validate the model's forecasts.

3 Results

3.1 Visualization of Predictions

3.1.1 Actual vs. Predicted Passing EPA

Figure 1 shows the comparison of actual and predicted values by EPA values using a scatter plot.The blue dotted line represents the line of perfect prediction, where the predicted values and actual values match exactly.The density of points along the line indicates that the model's accuracy is quite high, with many predictions being very close to actual performance.However, the significant dispersion of points above and below the line indicates that the model's accuracy is volatile.For example, predictions with lower EPA values (both positive and negative) seem to be more volatile than those with higher EPA values.

This volatility can be caused by a variety of factors, such as a nonlinear relationship between the predictor and EPA that the linear regression model fails to capture, or the influence of outliers due to abnormal performance or atypical play.It may also be that important predictors or interactions are missing from the model, leading to a failure to fit more complex patterns in the data.

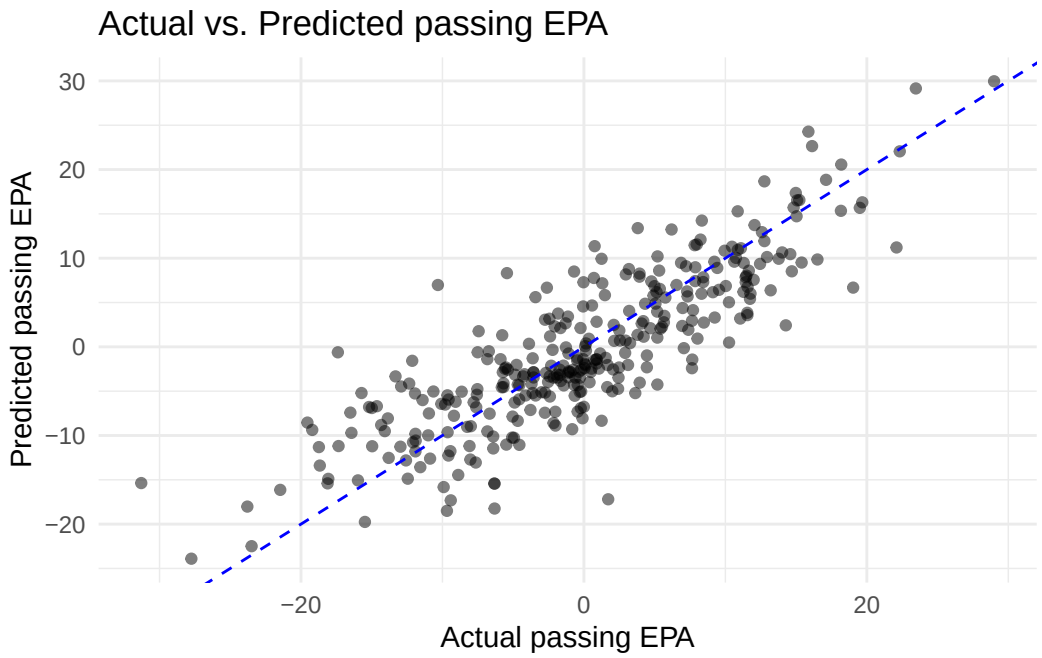

Figure 1: Scatterplot of Actual vs. Predicted Passing EPA

3.1.2 Residual Analysis

Figure 2 shows the relationship between the errors in the model's predictions and the EPA values that actually passed. The errors (the difference between the predicted) and actual values are scattered around the zero point of the horizontal dotted line, indicating no errors. Ideally, the errors should be distributed randomly, without apparent patterns, meaning that the model's errors are not systematic. Although the distribution appears random, some points have larger errors, indicating larger prediction errors. This may indicate a limitation of the model in capturing the full complexity of quarterback performance or specific game instances that are not covered by typical statistical measures

3.2 Model Predictive Results

The model's predictive ability, as shown by the actual and predicted values in the figure, is generally consistent with expected results, especially around the median of the EPA values that passed. The error plot shows that although the model is not biased in the prediction process, there is still room for improvement, especially for outliers.

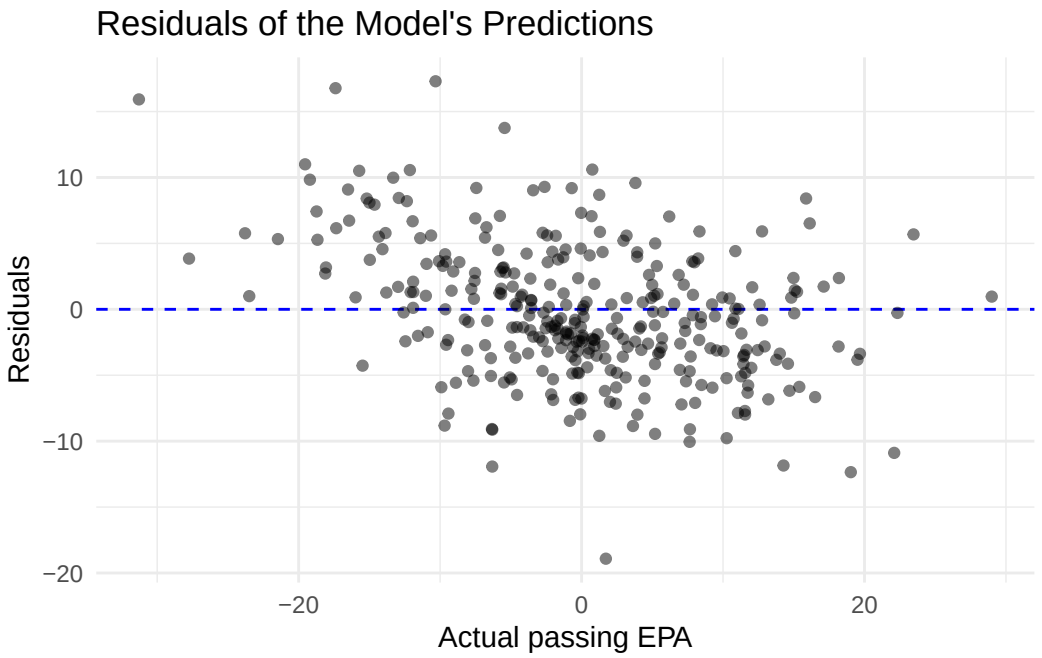

Figure 2: Residuals Plot for Actual vs. Predicted Passing EPA

Reference

Carl, Sebastian, Ben Baldwin, Lee Sharpe, Tan Ho, and John Edwards. 2023. *Nflverse: Easily Install and Load the 'Nflverse'*. https://nflverse.nflverse.com/.

Kuhn, Max, and Davis Vaughan. 2024. *Parsnip: A Common API to Modeling and Analysis Functions*. https://github.com/tidymodels/parsnip.

Kuhn, Max, Davis Vaughan, and Emil Hvitfeldt. 2024. *Yardstick: Tidy Characterizations of Model Performance*. https://github.com/tidymodels/yardstick.

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. 2024. *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.