Forecasting Quarterback Performance: A Statistical Analysis of Passing EPA in the NFL*

Dingning Li

2024-04-03

Table of contents

Introduction

In the professional football, the ability to anticipate a player's impact on upcoming games is invaluable. This paper will predict NFL quaterbacks' passing EPA for the remaining weeks of the 2023 regular season using a dataset provided by nflverse(citation). With Week 9 marking the season's halfway point, our analysis aims to create a model that can accurately predict how quarterbacks will contribute to their teams' chances of scoring, thereby affecting their likelihood of winning future games. This involves a construction of patterns from existing data, avoiding the inclusion of future information that could invalidate our predictions. We will describe the models and methodologies used to approach this task, pursuing the highest accuracy in our predictive analysis.

*Code and data are available at: https://github.com/iamldn2002/prediction

Data

The nflverse project(citation) provides a comprehensive collection of NFL data, including detailed player statistics, team information, and game outcomes. It's designed to be an accessible resource for analysts, researchers, and fans interested in conducting their own analyses of the NFL. The data is collected from publicly available sources and aggregated, typically accessible through R packages designed for data analysis. This makes nflverse an invaluable tool for creating predictive models and conducting in-depth analyses of player performances and team strategies within the NFL

- passing_epa: Expected Points Added (EPA) is a commonly used advanced statistic in football. This stat measures how well a team performs compared to their expectation on a play-by-play basis.
- player_id: a unique identifier assigned to each player
- recent_team: represents the team that the player was most recently a part of.
- week: In the NFL regular season, which spans 18 weeks. It indicates the specific week of the season.
- attempts: represents the number of passing attempts made during a game.

Model

Model set up

I chose the regression model to predict associations between dependent variable "passing_epa" and dependent variables. The regression model is summarized with its corresponding variables and structure. The model reflects a statistical approach to quantifying the relationship between a quarterback's passing EPA and various factors that could influence it.

$$Y = \beta_0 + \beta_1 T + \beta_2 W + \beta_3 A + \varepsilon \tag{1}$$

- Y represents the dependent variable `passing_epa`.
- T represents the independent variable "recent_team". Note that if `recent_team` is a categorical variable with multiple levels, you would need to expand this into multiple terms, $\beta_{1i} T_i$, where i indexes the levels of the categorical variable.
- W represents the independent variable "week".
- A represents the independent variable "attempts".
- $\beta_0, \beta_1, \beta_2, \beta_3$ are the coefficients for the corresponding independent variables.

- ε represents the error term.

Model Evaluation

This part provides a common process in predictive modeling and machine learning known as model evaluation.

```
      RMSE   Rsquared        MAE
10.1008052  0.1180982  8.6032440
```

Results

RMSE: 10.1008052. This is the square root of the average of the squared differences between the predicted and actual values. It's a measure of the model's accuracy, with lower values indicating a better fit. A RMSE of 10.1 suggests that, on average, the model's predictions are about 10.1 units away from the actual data points.

R-squared: 0.1180982. This metric reflects the proportion of variance in the dependent variable that's explained by the model. An R-squared value of approximately 0.118 suggests that the model explains about 11.8% of the variability in the test data's passing EPA. This is generally considered a low value, indicating that the model might not be capturing all the factors that affect the passing EPA or that there is a lot of inherent variability in the data that the model cannot account for.

MAE: 8.6032440. The Mean Absolute Error is the average of the absolute differences between the predictions and the actual values. This value tells us that, on average, the model's predictions are off by about 8.6 units. Compared to the RMSE, the MAE is not as sensitive to outliers, as it does not square the errors.

Discussion

These results indicate that the predictive accuracy of the model is moderate to low. The RMSE and MAE suggest that there is a significant average error in the model's predictions. Moreover, the R-squared value being quite low implies that the model's explanatory power is limited, and there may be other factors not included in the model that are influencing the outcome variable. These results would typically lead a consideration of investigating more complex models, additional features, or alternative methods to improve the model's accuracy and explanatory power. There are many other factors in the database that affect epa, such as the oponent team, and for more accurate predictions, it is necessary to take into account the factors that affect the dependent variable and build a more accurate model.