The best method in forecasting NFL statistics on passing epa*

Sean Liu

March 29, 2024

This report shows an approach to forecasting NFL quarterback performance. offering valuable perspectives for coaches and analysts. In order to predict the future team success.

Table of contents

1 Introduction

In the National Football League (NFL), understanding and predicting quarterback performance an important aspect of game strategy and player evaluation. The Passing Expected Points Added (passing_epa) stands out as an important metrics that captures a quarterback's contribution to the team's scoring potential on each play. This report uses predictive modeling, to forecast NFL quarterbacks' passing EPA based on their performance in the first nine weeks of the 2023 regular season.

The remainder of this paper is structured as follows. Section 2

*Code and data are available at: https://github.com/Sluuu/Essay12-Prediction

2 Data

Our data is from the package `nflverse` Carl et al. (2023). The package includes the data of the National Football League. As the data contains too many information, I have cleaned the data that filter player statistics for the 2023 NFL season. It specifically targets quarterback (QB) statistics from the regular season (REG) up to week 9. And I will be forecasting the model of passing_epa for week 10-18. Inside my data, I have chosen 10 variables that I think would be important to the study.

- `player_name`: It shows the name of the player
- `recent_team`: It shows the most recent team the player plays in
- `week`: It shows the game week number
- `passing_yards`: It shows the yards gained on pass plays
- `passing_tds`: It shows the number of passing touchdowns
- `interceptions`: It shows the number of interceptions thrown
- `attempts`: It shows the number of pass attempts by the NFL
- `completions`: It shows the number of completed passes
- `passing_epa`: It shows the total expected points added on pass attempts and sacks
- `completion_rate`: This is a new variable I created, it is calculated by the completions divided by attempts. Which shows the passes completion rates.

After filtering my data set, there are 335 observations. However, it also contains some missing values, which I will be dropping/eliminating them. As there weren't a lot of missing values. Which concluded that there are 316 observations and 10 variables in my data-set.

```
 player_name        recent_team             week      passing_yards
 Length:316         Length:316         Min.   :1.00   Min.   : -1.0
 Class :character   Class :character   1st Qu.:3.00   1st Qu.:149.0
 Mode  :character   Mode  :character   Median :5.00   Median :217.5
                                       Mean   :4.94   Mean   :204.4
                                       3rd Qu.:7.00   3rd Qu.:272.0
                                       Max.   :9.00   Max.   :470.0
  passing_tds    interceptions       attempts      completions
 Min.   :0.000   Min.   :0.0000   Min.   : 1.00   Min.   : 0.00
 1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:24.00   1st Qu.:15.00
 Median :1.000   Median :0.0000   Median :31.00   Median :20.00
 Mean   :1.158   Mean   :0.6772   Mean   :29.29   Mean   :19.09
 3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:38.00   3rd Qu.:25.00
 Max.   :5.000   Max.   :4.0000   Max.   :58.00   Max.   :40.00
  passing_epa        completion_rate
 Min.   :-32.76908   Min.   :0.0000
 1st Qu.: -7.33045   1st Qu.:0.5781
 Median :  0.03159   Median :0.6582
```

```
Mean   : -0.37558   Mean   :0.6449
3rd Qu.:  5.79043   3rd Qu.:0.7227
Max.   : 28.41896   Max.   :1.0000
```

In the summary statistics above, it shows the summary of my dataset. There are two types of variables, numerical and categorical. In Numerical variables, such as `passing_yards`, `week`, `passing_tds`, `interceptions`, `passing_attempts`, `passing_completions`, `passing_epa`, and `completion_rate`. It summarize Min: the minimum value, 1st Qu: The first quartile (25th percentile), Median: The median (50th percentile), Mean: The average of the values. 3rd Qu: The third quartile (75th percentile), and Max: The maximum value. These statistics provide insights into the distribution of each numerical variable. Which helps me to understand the performance of quarterbacks in terms of passing yards, touchdowns, interceptions, etc., until week 9 of the 2023 regular season. The next variable is the categorical variables, which are `player_name` and `recent_team`.

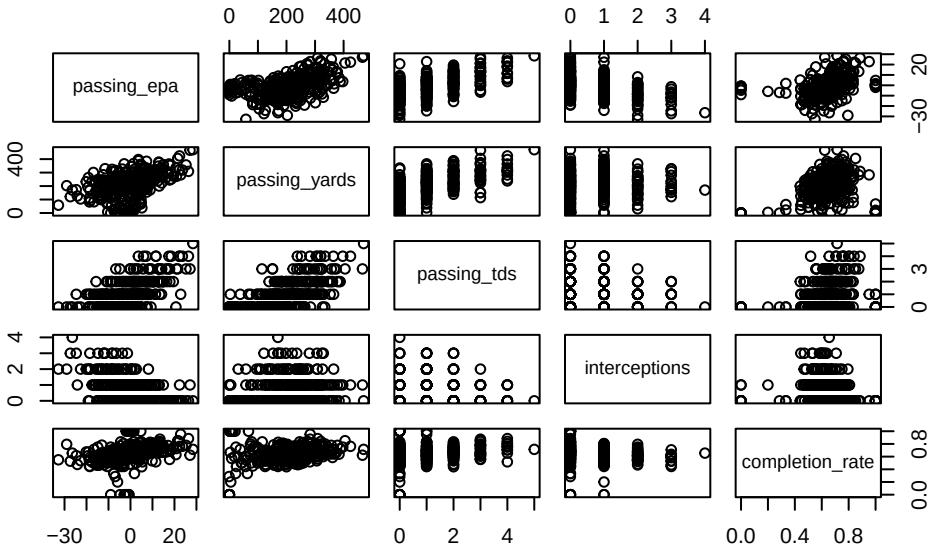

Figure 1: Scatterplot matrix of each variables

From Figure 1, it shows the variables included in the scatterplot matrix. Each variable will be plotted against every other variable in the matrix. Each row represents one of the variables with the oreder, `passing_epa`, `passing_yards`, `passing_tds`, `interceptions`, and `completion_rate`. Inside each cell of the matrix, where the variable corresponding to the row is plotted on the y-axis and the variable corresponding to the column is plotted on the x-axis. This graph helps us to visualize the relationships between the variables in my dataset.

3 Model

The model and the data were constructed in R (R Core Team 2023) with the help of the following packages, `nflverse` package of Carl et al. (2023), `tidyr` package of Wickham, Vaughan, and Girlich (2024), `randomForest` package of Liaw and Wiener (2002), `dplyr` package of Wickham et al. (2023), `lubridate` package of Grolemund and Wickham (2011), `caret` package of Kuhn and Max (2008), and 'readr" Wickham, Hester, and Bryan (2024).

For my model, I will first split my data into 2 sets, training and testing sets. It is separated by variable `week` and 80% of the data will be allocated into the training set. Then I will use the random Forest model to predict the variable `passing_eba` based on all other except for `player_name`, `recent_team`, and `week`. The code sets up a predictive model where the target variable is passing_epa, and it uses the quarterback's performance data as predictors.

```
     RMSE  Rsquared       MAE
6.0804345 0.6402223 4.6547220
```

I have used the postResample function from the `caret` Kuhn and Max (2008) package, which provides a summary of the predictive model. By comparing the predicted `passing_epa` from your model and the actual `passing_epa` values from the test dataset. The output has three key metrics: RMSE (Root Mean Squared Error), R-squared (Rsquared), and MAE (Mean Absolute Error).

The RMSE equals to 5.9451548: This is the square root of the average of the squares of the errors. RMSE is a measure of the model's prediction error, with lower values indicating a better fit. In this study, RMSE of 5.9451548 means that, on average, the model's predictions are about 5.9451548 units away from the actual values on the `passing_epa` scale.

The R-squared equals to 0.6583139: This measure shows how close the data are to the fitted regression line. R-squared ranges from 0 to 1, where 0 means that the model explains none of the variability of the response data around its mean, and 1 means that it explains all the variability. An R-squared of 0.6583139 suggests that approximately 65.83% of the variation in passing_epa can be explained by the model's predictors.

The MAE equals to 4.5219004: This is the average of the absolute errors. Like RMSE, MAE gives an idea of how wrong the model's predictions are, a lower MAE indicates a better model. An MAE of 4.5219004 means that, on average, the model's predictions are about 4.5219004 units away from the actual value.

4 Conclusion

In the analysis of NFL quarterback performance, focusing on `passing_epa`. I chose the Random Forest model due to its robustness and its ability to handle non-linear relationships. After training and testing the model on the first nine weeks of the 2023 NFL season data, I found the Random Forest model to provide great predictive accuracy.

The model's prediction results, as showed by metrics such as RMSE, MAE, and R-squared, was found to be compelling. An RMSE of 5.945 and an MAE of 4.522, coupled with an R-squared value of 0.658, indicate a model that provides reliable and accurate forecasts of passing EPA.

The Random Forest model's ability to generalize well from our training data to the test data suggests it can reliably forecast future quarterback performance in the regular season's remaining weeks.

References

Carl, Sebastian, Ben Baldwin, Lee Sharpe, Tan Ho, and John Edwards. 2023. *Nflverse: Easily Install and Load the 'Nflverse'*. https://CRAN.R-project.org/package=nflverse.

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.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

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://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.