NFL Quarterback Passing EPA Forecasting[*]

Lin Dai

April 2, 2024

1 Working space set up

2 Import data

```r
nfl_2020 <- read_csv(here::here("data/analysis_data.csv"))
```

```
Rows: 646 Columns: 53
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (9): player_id, player_name, player_display_name, position, position_gr...
dbl (44): season, week, completions, attempts, passing_yards, passing_tds, i...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
nfl_2020_train <- nfl_2020 |>
  filter(week<=9)
nfl_2020_test <- nfl_2020 |>
  filter(week>9)
```

3 Model

[*] Code and data available at: https://github.com/Trainingwolf8878/NFL_2020

```
first_model_tidymodels <-
  linear_reg() |>
  set_engine(engine = "lm") |>
  fit(
    passing_epa ~ attempts + passing_yards + completions,
    data = nfl_2020_train
  )

second_model_tidymodels <-
  linear_reg() |>
  set_engine(engine = "lm") |>
  fit(
    passing_epa ~ attempts + passing_yards + completions + week,
    data = nfl_2020_train
  )

modelsummary(list("Model 1" = first_model_tidymodels,
                  "Model 2" = second_model_tidymodels))
```

In this model, we employed linear regression to predict the Passing Expected Points Added (EPA) of NFL players. EPA serves as a metric to gauge the potential impact of each play on a team's score. Therefore, the model aims to predict the EPA of each passing play by analyzing factors such as the number of pass attempts, passing yards, and completions.

The model comprises two versions labeled as "Model 1" and "Model 2". The distinction between them lies in the number of predictor variables included in the model. In "Model 1", we only considered three factors—pass attempts, passing yards, and completions—as predictor variables. Conversely, in "Model 2", we additionally incorporated the week number of the season as a predictor variable.

By training these two models, we obtain predictions for Passing EPA, allowing us to understand the influence of different factors on EPA. Additionally, by comparing the performance of the two models, we can determine which one provides a more accurate and effective explanation of Passing EPA.

4 Predict

5 Model set up

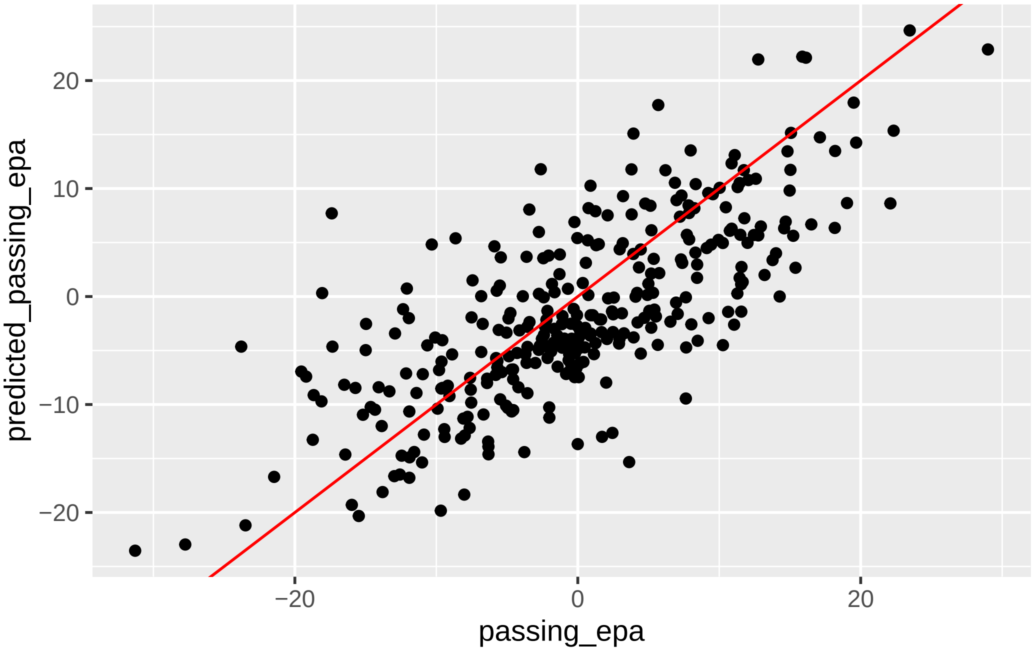
\begin(align) y_i | _i, & Normal(_i,) \ _i &= _0+ 1 x{1i} + 2 x{2i} \end{align}

Table 1: Explanatory models o nfl_2020

	Model 1	Model 2
(Intercept)	−3.076	−2.535
	(0.926)	(1.198)
attempts	−1.315	−1.316
	(0.090)	(0.090)
passing_yards	0.124	0.124
	(0.008)	(0.008)
completions	0.835	0.830
	(0.151)	(0.151)
week		−0.102
		(0.143)
Num.Obs.	318	318
R2	0.599	0.600
R2 Adj.	0.595	0.594
AIC	2105.5	2107.0
BIC	2124.3	2129.6
RMSE	6.53	6.52

where y_i refers to the total expected points added on pass attempts, x_{1i} refers the number of attempts, x_{2i} refers to the expected points added on pass plays, x_{3i} refers to the number of completed passes.

6 Result



7 What did I do

Through the provided code, I accomplished the following tasks:

Firstly, I set up the working environment by loading the necessary R packages, including tidyverse, tidymodels, modelsummary, nflverse, and dplyr. These packages provide the functionalities needed for data manipulation, modeling, and visualization.

Next, I imported the 2020 NFL data and split it into a training set (consisting of data from the first 9 weeks) and a test set (consisting of data from week 10 onwards) for modeling and prediction purposes.

Then, I constructed two linear regression models using the tidymodels package to model the passing EPA (Expected Points Added). The first model considered factors such as the number of attempts, passing yards, and completions, while the second model also incorporated the week of the season. I used the modelsummary package to summarize and compare the results of these two models.

Subsequently, I utilized the second model to make predictions on the test set and merged and arranged the predicted results with the actual data.

Finally, I plotted a scatterplot of the predicted passing EPA against the actual passing EPA and added a red diagonal line to visually compare the relationship between the two.

Through these tasks, I was able to establish and evaluate two models for predicting passing EPA and assess their predictive performance on the test set, providing a foundation for further analysis and interpretation.

8 Conclusion

In this project, we aimed to develop predictive models for the Expected Points Added (EPA) of NFL quarterbacks during the 2023 regular season. By the graph we can see the point is near the red line, which we can say our predict value is near the actual value.