NFL_forecasting*

Yunzhao Li

April 1, 2024

1 Introduction

The National Football League (NFL) offers a dynamic and unpredictable arena for sports analytics, particularly in forecasting the performance of key players such as quarterbacks. This paper presents a predictive analysis focused on the passing_epa (expected points added on pass attempts and sacks) metric for NFL quarterbacks in the second half of the 2023 regular season. By leveraging statistical modeling techniques, we aim to unearth the underlying factors contributing to quarterback performance and offer actionable insights for teams and fantasy football enthusiasts. We use the nflverse package (Carl et al. 2023) from R (R Core Team 2020) to load some statistics for NFL quarterbacks during the regular season.

2 Model

In the quest to forecast `passing_epa`, I opted for a linear regression model. This model allows us to explore the relationship between a quarterback's performance metrics and their ability to contribute positively to the game's score, quantified through `passing_epa`.

2.1 Features Considered:

- **Completions**: A basic yet powerful indicator of a quarterback's accuracy.
- **Passing Yards**: Directly correlates with the offensive advancement in the game.
- **Passing Touchdowns (TDs)**: The pinnacle of a quarterback's job, directly impacting the score.
- **Interceptions**: Mistakes that can turn the tide against a team.

*Code and data are available at: https://github.com/yunzhaol/NFL-forecasting.git.

- **Passing First Downs**: Essential for maintaining possession and extending play opportunities.

These features were not randomly chosen; each represents a critical aspect of a quarterback's game that logically could affect their expected points added. The model's premise is that successful offensive actions (like completions and touchdowns) are likely to increase `passing_epa`, whereas negative outcomes (such as interceptions) could decrease it.

2.2 Model set-up

We define our model to forecast the `passing_epa`, which is the expected points added on pass attempts and sacks for NFL quarterbacks. Our linear regression model is specified as follows:

$$\begin{aligned}\text{Passing\_epa}_i = &\beta_0 + \beta_1 \times \text{Completions}_i + \beta_2 \times \text{Passing\_Yards}_i \\ &+ \beta_3 \times \text{Passing\_TDs}_i + \beta_4 \times \text{Interceptions}_i \\ &+ \beta_5 \times \text{Passing\_First\_Downs}_i + \epsilon_i\end{aligned}$$

where:

- passing_epa_i is the expected points added for the i^{th} quarterback.
- Completions_i is the number of completed passes.
- Passing_Yards_i is the number of yards gained through passing.
- Passing_TDs_i is the number of passing touchdowns.
- Interceptions_i is the number of interceptions thrown.
- $\text{Passing\_First\_Downs}_i$ is the number of first downs achieved through passing.
- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the parameters to be estimated.
- ϵ_i is the error term for the i^{th} observation.

3 Results

Upon analyzing the predictors of passing_epa, the linear regression model provided us with several key insights:

Each additional completion is associated with a slight decrease in passing_epa by approximately 0.89 points, which might reflect a complex interaction with other variables in the model.

Passing yards contribute positively, with each yard adding a small yet significant 0.06 points to the passing_epa.

Touchdowns are highly valuable, with each one contributing an impressive 2.82 points, underscoring the critical importance of scoring plays.

Interceptions are particularly detrimental, reducing passing_epa by approximately 5.50 points, highlighting the costly nature of turnovers.

Finally, each passing first down contributes positively by about 0.91 points, emphasizing the importance of sustaining drives. These coefficients paint a nuanced picture of a quarterback's contribution to the game, from the big wins of touchdowns to the subtle yet essential gains from moving the chains.

4 Model Performance:

- **RMSE (Root Mean Square Error)**: The model achieved an RMSE of 5.8, a number that, while not perfect, suggests a reasonable degree of prediction accuracy for a fledgling analyst's first foray into NFL statistics.
- **R-squared**: With an R^2 of 0.65, the model claims to explain 65% of the variance in `passing_epa`. This figure was a beacon of encouragement, hinting that the features selected did indeed have a significant impact on a quarterback's performance.

5 Process

The code is referenced (Alexander 2024). We used R (R Core Team 2020) and nflverse package (Carl et al. 2023) to complete the whole model. It began with the extraction of quarterback statistics from the `nflverse` package. Focusing on regular season play and narrowing down to quarterbacks, I separated the data into training and testing sets based on the week—ensuring a fair test of the model's predictive power without peeking into the future it was trying to forecast.

6 Conclusion

This exploration into predicting NFL quarterbacks' `passing_epa` for the latter half of the 2023 season creates a well performed linear regression model, with its selected features, provided a solid foundation for understanding the factors that contribute to a quarterback's performance.

References

Alexander, Rohan. 2024. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com.

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

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.