Predicting NFL Rushing Touchdowns with Random Forest Regression

**April 11, 2024**

# Research Question

To what extent do the following predictor variables affect the total number of rushing touchdowns: age, games started, rushing attempts, rushing yards, first down runs, longest rush, yards per attempt, yards per game, and number of fumbles?

# Hypothesis

* These variables (age, games started, rushing attempts, rushing yards, first down runs, longest rush, yards per attempt, yards per game, and number of fumbles) statistically significantly affect the number of rushing touchdowns

# Random Forest Regression Algorithm

In order to tackle this problem, I used a random forest regression algorithm in R. The data was properly cleaned before analysis and split into training and testing data sets on an 80-20 ratio. An initial model was created using the training data which was then used against the testing data in order to predict the number of rushing touchdowns. After which, several calculations were made to determine how accurate the model was, followed by calculating which of our predictor variables are the most statistically important for our predictions.

# 

# Findings

## Tuning and the initial model

Using cross validation to determine the best value for mtry, an initial random forest regression model was made using 500 trees and a value of 5 for mtry. The initial model resulted in an MSE of approx. 1.58 with 77.72% of the variance explained, indicating a good fit.

1. Making our Predictions

Using a testing data set , we used the initial model to then predict the number of rushing touchdowns. By doing so, we calculated an MSE of approximately 1.59, which means that, on average, the model's predictions were within 1.59 touchdowns of the actual value. This combined with an R-squared value of 76.85% suggests that the model was a relatively good fit for making predictions.

1. Determining which Variables are Most Statistically Significant

The importance of the initial model was derived to determine exactly which of the predictor variables were the most statistically significant. It was determined that the number of rushing first downs (r1D) was the most statistically significant, followed by rushing yards (rYds) and rushing attempts (rAtt) with the number of fumbles being the least statistically significant. In the context of professional football, this makes sense as a running backs ability to sustain drives and keep the chains moving would suggest an ability to score more points thus leading to more rushing touchdowns.

# 

# Limitations

* One limitation of this analysis is that the model is outputting continuous variables (or decimals). In a real football game, you cannot score fractions of a touchdown. Additional rounding is something that may be required in future analyses, depending on the ultimate goal.
* Another limitation is that the algorithm itself is heavily reliant on your computer's memory, often leading to slow loading times.

Proposed Actions

* Since the ability to get first downs is the strongest predictor variable in scoring rushing touchdowns, it is recommended that teams focus on their ability to sustain their drives. This can be done with focusing on the play calling as well as offensive line blocking to give your running back the best chance of obtaining first downs in non-passing situations (3rd and inches, for example).

Benefits

* Player evaluation metrics can be improved, especially when you consider that first downs, yards, and attempts are the most significant predictors
* Since first downs are the most significant predictor, teams can game plan more strategically to maximize the number of first downs in a game, either by calling better rushing plays or improving offensive line blocking