Lab 5 / Final Project

Christopher Hainzl

2023-05-03

Introduction

For as long as I can remember, I have been a huge fan of the NFL (my favorite team being the New York Giants), and always wanted to do an analysis on what factors seem to have the most impact on the probability of an interception happening. Through the years that I have spent watching games, the factors which, to me, seemed to have the greatest effect are:

- The current quarter of the game
- The amount of yards to go before the end zone is reached
- If a timeout was called or not

Collecting & Preparing the Data:

To perform this analysis, I will be using the "Detailed NFL Play-by-Play Data 2009-2018" dataset collected off of Kaggle:

https://www.kaggle.com/datasets/maxhorowitz/nflplaybyplay2009to2016

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
footballdata <- read_csv("C:/Users/hainz/OneDrive/2023 Spring/Applied Statistics (MATH370-01)/Final Pro
```

```
## Warning: One or more parsing issues, see `problems()` for details

## Rows: 449371 Columns: 255

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (74): home_team, away_team, posteam, posteam_type, defteam, side_of_fi...
## dbl  (147): play_id, game_id, yardline_100, quarter_seconds_remaining, half_...
## lgl   (32): lateral_receiver_player_id, lateral_receiver_player_name, latera...
## date   (1): game_date
## time   (1): time
```

```
##
## 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.
```

```
colnames(footballdata)
```

```
##   [1] "play_id"
##   [2] "game_id"
##   [3] "home_team"
##   [4] "away_team"
##   [5] "posteam"
##   [6] "posteam_type"
##   [7] "defteam"
##   [8] "side_of_field"
##   [9] "yardline_100"
##  [10] "game_date"
##  [11] "quarter_seconds_remaining"
##  [12] "half_seconds_remaining"
##  [13] "game_seconds_remaining"
##  [14] "game_half"
##  [15] "quarter_end"
##  [16] "drive"
##  [17] "sp"
##  [18] "qtr"
##  [19] "down"
##  [20] "goal_to_go"
##  [21] "time"
##  [22] "yrdln"
##  [23] "ydstogo"
##  [24] "ydsnet"
##  [25] "desc"
##  [26] "play_type"
##  [27] "yards_gained"
##  [28] "shotgun"
##  [29] "no_huddle"
##  [30] "qb_dropback"
##  [31] "qb_kneel"
##  [32] "qb_spike"
##  [33] "qb_scramble"
##  [34] "pass_length"
##  [35] "pass_location"
##  [36] "air_yards"
##  [37] "yards_after_catch"
##  [38] "run_location"
##  [39] "run_gap"
##  [40] "field_goal_result"
##  [41] "kick_distance"
##  [42] "extra_point_result"
##  [43] "two_point_conv_result"
##  [44] "home_timeouts_remaining"
##  [45] "away_timeouts_remaining"
##  [46] "timeout"
##  [47] "timeout_team"
##  [48] "td_team"
##  [49] "posteam_timeouts_remaining"
##  [50] "defteam_timeouts_remaining"
##  [51] "total_home_score"
##  [52] "total_away_score"
##  [53] "posteam_score"
##  [54] "defteam_score"
##  [55] "score_differential"
##  [56] "posteam_score_post"
##  [57] "defteam_score_post"
##  [58] "score_differential_post"
##  [59] "no_score_prob"
##  [60] "opp_fg_prob"
##  [61] "opp_safety_prob"
##  [62] "opp_td_prob"
##  [63] "fg_prob"
##  [64] "safety_prob"
##  [65] "td_prob"
##  [66] "extra_point_prob"
##  [67] "two_point_conversion_prob"
##  [68] "ep"
##  [69] "epa"
##  [70] "total_home_epa"
##  [71] "total_away_epa"
##  [72] "total_home_rush_epa"
##  [73] "total_away_rush_epa"
##  [74] "total_home_pass_epa"
##  [75] "total_away_pass_epa"
##  [76] "air_epa"
##  [77] "yac_epa"
##  [78] "comp_air_epa"
##  [79] "comp_yac_epa"
##  [80] "total_home_comp_air_epa"
##  [81] "total_away_comp_air_epa"
##  [82] "total_home_comp_yac_epa"
##  [83] "total_away_comp_yac_epa"
##  [84] "total_home_raw_air_epa"
##  [85] "total_away_raw_air_epa"
##  [86] "total_home_raw_yac_epa"
##  [87] "total_away_raw_yac_epa"
##  [88] "wp"
##  [89] "def_wp"
##  [90] "home_wp"
##  [91] "away_wp"
##  [92] "wpa"
##  [93] "home_wp_post"
##  [94] "away_wp_post"
##  [95] "total_home_rush_wpa"
##  [96] "total_away_rush_wpa"
##  [97] "total_home_pass_wpa"
##  [98] "total_away_pass_wpa"
##  [99] "air_wpa"
## [100] "yac_wpa"
## [101] "comp_air_wpa"
## [102] "comp_yac_wpa"
## [103] "total_home_comp_air_wpa"
## [104] "total_away_comp_air_wpa"
## [105] "total_home_comp_yac_wpa"
## [106] "total_away_comp_yac_wpa"
## [107] "total_home_raw_air_wpa"
## [108] "total_away_raw_air_wpa"
## [109] "total_home_raw_yac_wpa"
## [110] "total_away_raw_yac_wpa"
## [111] "punt_blocked"
## [112] "first_down_rush"
## [113] "first_down_pass"
## [114] "first_down_penalty"
## [115] "third_down_converted"
## [116] "third_down_failed"
## [117] "fourth_down_converted"
## [118] "fourth_down_failed"
## [119] "incomplete_pass"
## [120] "interception"
## [121] "punt_inside_twenty"
## [122] "punt_in_endzone"
## [123] "punt_out_of_bounds"
## [124] "punt_downed"
## [125] "punt_fair_catch"
## [126] "kickoff_inside_twenty"
## [127] "kickoff_in_endzone"
## [128] "kickoff_out_of_bounds"
## [129] "kickoff_downed"
## [130] "kickoff_fair_catch"
## [131] "fumble_forced"
## [132] "fumble_not_forced"
## [133] "fumble_out_of_bounds"
## [134] "solo_tackle"
## [135] "safety"
## [136] "penalty"
## [137] "tackled_for_loss"
## [138] "fumble_lost"
## [139] "own_kickoff_recovery"
## [140] "own_kickoff_recovery_td"
## [141] "qb_hit"
## [142] "rush_attempt"
## [143] "pass_attempt"
## [144] "sack"
## [145] "touchdown"
## [146] "pass_touchdown"
## [147] "rush_touchdown"
## [148] "return_touchdown"
## [149] "extra_point_attempt"
## [150] "two_point_attempt"
## [151] "field_goal_attempt"
## [152] "kickoff_attempt"
## [153] "punt_attempt"
## [154] "fumble"
## [155] "complete_pass"
## [156] "assist_tackle"
## [157] "lateral_reception"
## [158] "lateral_rush"
## [159] "lateral_return"
## [160] "lateral_recovery"
## [161] "passer_player_id"
## [162] "passer_player_name"
## [163] "receiver_player_id"
## [164] "receiver_player_name"
## [165] "rusher_player_id"
## [166] "rusher_player_name"
## [167] "lateral_receiver_player_id"
## [168] "lateral_receiver_player_name"
## [169] "lateral_rusher_player_id"
## [170] "lateral_rusher_player_name"
## [171] "lateral_sack_player_id"
## [172] "lateral_sack_player_name"
## [173] "interception_player_id"
## [174] "interception_player_name"
## [175] "lateral_interception_player_id"
## [176] "lateral_interception_player_name"
## [177] "punt_returner_player_id"
## [178] "punt_returner_player_name"
## [179] "lateral_punt_returner_player_id"
## [180] "lateral_punt_returner_player_name"
## [181] "kickoff_returner_player_name"
## [182] "kickoff_returner_player_id"
## [183] "lateral_kickoff_returner_player_id"
## [184] "lateral_kickoff_returner_player_name"
## [185] "punter_player_id"
## [186] "punter_player_name"
## [187] "kicker_player_name"
## [188] "kicker_player_id"
## [189] "own_kickoff_recovery_player_id"
## [190] "own_kickoff_recovery_player_name"
## [191] "blocked_player_id"
## [192] "blocked_player_name"
## [193] "tackle_for_loss_1_player_id"
## [194] "tackle_for_loss_1_player_name"
## [195] "tackle_for_loss_2_player_id"
## [196] "tackle_for_loss_2_player_name"
## [197] "qb_hit_1_player_id"
## [198] "qb_hit_1_player_name"
## [199] "qb_hit_2_player_id"
## [200] "qb_hit_2_player_name"
## [201] "forced_fumble_player_1_team"
## [202] "forced_fumble_player_1_player_id"
## [203] "forced_fumble_player_1_player_name"
## [204] "forced_fumble_player_2_team"
## [205] "forced_fumble_player_2_player_id"
## [206] "forced_fumble_player_2_player_name"
## [207] "solo_tackle_1_team"
## [208] "solo_tackle_2_team"
## [209] "solo_tackle_1_player_id"
## [210] "solo_tackle_2_player_id"
## [211] "solo_tackle_1_player_name"
## [212] "solo_tackle_2_player_name"
## [213] "assist_tackle_1_player_id"
## [214] "assist_tackle_1_player_name"
## [215] "assist_tackle_1_team"
## [216] "assist_tackle_2_player_id"
## [217] "assist_tackle_2_player_name"
## [218] "assist_tackle_2_team"
## [219] "assist_tackle_3_player_id"
## [220] "assist_tackle_3_player_name"
## [221] "assist_tackle_3_team"
## [222] "assist_tackle_4_player_id"
## [223] "assist_tackle_4_player_name"
## [224] "assist_tackle_4_team"
## [225] "pass_defense_1_player_id"
## [226] "pass_defense_1_player_name"
## [227] "pass_defense_2_player_id"
## [228] "pass_defense_2_player_name"
## [229] "fumbled_1_team"
## [230] "fumbled_1_player_id"
## [231] "fumbled_1_player_name"
## [232] "fumbled_2_player_id"
## [233] "fumbled_2_player_name"
## [234] "fumbled_2_team"
## [235] "fumble_recovery_1_team"
## [236] "fumble_recovery_1_yards"
## [237] "fumble_recovery_1_player_id"
## [238] "fumble_recovery_1_player_name"
## [239] "fumble_recovery_2_team"
## [240] "fumble_recovery_2_yards"
## [241] "fumble_recovery_2_player_id"
## [242] "fumble_recovery_2_player_name"
## [243] "return_team"
## [244] "return_yards"
## [245] "penalty_team"
## [246] "penalty_player_id"
## [247] "penalty_player_name"
## [248] "penalty_yards"
## [249] "replay_or_challenge"
## [250] "replay_or_challenge_result"
## [251] "penalty_type"
## [252] "defensive_two_point_attempt"
## [253] "defensive_two_point_conv"
## [254] "defensive_extra_point_attempt"
## [255] "defensive_extra_point_conv"
```

The columns that I want to consider for this analysis are:

- interception = indicating if the pass was intercepted or not (response)
- qtr = indicating the current quarter of the game (predictor #1)
- ydstogo = indicating the amount of yards to go to the end zone (predictor #2)
- timeout = indicating if a timeout was called on a play or not (predictor #3)

```
# Check for missing values
sum(is.na(footballdata$interception))
```

```
## [1] 12874
```

```
sum(is.na(footballdata$qtr))
```

```
## [1] 0
```

```
sum(is.na(footballdata$ydstogo))
```

```
## [1] 0
```

```
sum(is.na(footballdata$timeout))
```

```
## [1] 12874
```

However, there are some missing values in the “timeout” and “interception” columns, which I will have to account for prior to setting up my analysis.

```
footballdata["timeout"][is.na(footballdata["timeout"])] <- 0
footballdata["interception"][is.na(footballdata["interception"])] <- 0
```

```
# Make sure that there are no more missing values in the columns
# that are being considered
sum(is.na(footballdata$interception))
```

```
## [1] 0
```

```
sum(is.na(footballdata$qtr))
```

```
## [1] 0
```

```
sum(is.na(footballdata$ydstogo))
```

```
## [1] 0
```

```
sum(is.na(footballdata$timeout))
```

```
## [1] 0
```

To make it easier for us to keep track, I will be storing my fixed data in a new variable named “footballdata_cleaned”.

```
footballdata_cleaned <- footballdata
```

Analysis

```
unique(footballdata_cleaned$interception)
```

```
## [1] 0 1
```

Since all the values presented in the interception column are either 0 or 1, I will be setting up a logistic regression that predicts if an interception will be made based on the predictor variables listed in the section titled “Collecting & Preparing The Data”. I will also be working with a 95% confidence level when interpreting the model.

```
footballmodel <- glm(interception~qtr+ydstogo+timeout, data = footballdata_cleaned, family = "binomial"
summary(footballmodel)
```

```
##
## Call:
## glm(formula = interception ~ qtr + ydstogo + timeout, family = "binomial",
##     data = footballdata_cleaned)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6177  -0.1580  -0.1404  -0.1208   3.5761
##
## Coefficients:
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -5.408918   0.046632 -115.992  < 2e-16 ***
## qtr          0.119041   0.013179    9.033  < 2e-16 ***
## ydstogo      0.067462   0.002865   23.550  < 2e-16 ***
## timeout     -1.289350   0.173875   -7.415 1.21e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 51579  on 449370  degrees of freedom
## Residual deviance: 50771  on 449367  degrees of freedom
## AIC: 50779
##
## Number of Fisher Scoring iterations: 8
```

Model Parameter Interpretations and Estimates

Starting at a value of approximately e^(-5.41), or 0.004 (B0):

- The probability of an interception increases by a factor of e^(0.119), or approximately 1.13, for each unit increase in the current quarter of the game. (B1)
- The probability of an interception increases by a factor of e^(0.067), or approximately 1.07, for each unit increase in the amount of yards to go in order to reach the end zone. (B2)
- The probability of an interception decreases by a factor of e^(-1.289), or approximately 0.276, if a timeout is called on a play. (B3)

When working with a 95% confidence level, the p-values associated with each of the predictor variables indicate that they are all significant predictors, and none of them should be removed from the model (their p-values are all less than .05). However, whether or not it is safe to conclude that this is true will depend on what the pseudo-r^2 value tells us.

Checking How Well My Model Fits the Data

```
# Use PseudoR2 function to determine the McFadden pseudo-r^2 value for the logistic regression.
PseudoR2(footballmodel)
```

```
##   McFadden
## 0.01567838
```

McFadden says that a pseudo-r^2 value higher than 0.2 indicates that the model fits the data very well (cited from: https://datascience.oneoffcoder.com/psuedo-r-squared-logistic-regression.html). Because the computed pseudo-r^2 value is less than 0.2, I would say that this model is not a good fit for the data.

Conclusions

Based on the computed pseudo-r^2 value, I would say that the model used for this analysis was not good. I also think that the overall analysis was affected by how I had to account for missing values. Therefore, it is not safe to assume that all of the predictor variables being considered for this analysis are significant predictors of whether or not an interception will happen.