Regression Analyisis of ‘Home-Team Advantage’ in the National Football League

Christopher Hom, Joanna Rashid, Lili Teister, Stephen Yu

04/27/2022

Introduction

Is “home team advantage” a statistically significant phenomenon in the National Football League (NFL)? If so, what variables in the data can explain this phenomenon? Linear and logistic regression are used to explore the relationship of various game data features on game outcomes. We attempt to control for variables that impact the expected result of a game and identify if there remains any additional advantage for the home team.

The phenomena of home-field-advantage is well documented (Swartz, T. B. et al., Jamieson, J. P.) Review of Cleveland, et al., Swartz, T. B., et al. contributed to rigorous selection and computation of control variables. However, there is evidence that the home-field advantage effect is shrinking (Kilgore, A., & Greenberg, N.), even more so in the age of the pandemic where some games are played with no fans in the stands (Ponzo, M., et al., Mccarrick, D et al.).

For this analysis, multiple data sets have been joined to produce both factor and continuous variables related to a game’s time, location, and conditions as well as a team’s injuries, amount of rest, distance traveled, and relative ability. These variables are regressed on the spread of the score where negative values indicate a home team loss. Additionally, we have implemented a logistic regression model wherein the same explanatory variables are used to explain the log-likelihood of a home-team win. A random forest model is also implemented to explore relative variable importance. . # Analysis

Required packages

```r
#install.packages("tidyverse")
#install.packages()
#install.packages("devtools")
#devtools::install_github(repo = "maksimhorowitz/nflscrapR")
#install.packages("nflreadr")
#install.packages("car")
```

```
## Warning: package 'nflreadr' was built under R version 4.1.2
```

Data

Games dataset

The games dataset is the main dataset that will be used for analysis. It contains a row for each NFL games from 1999 to 2021, but only seasons from 2010 to 2019 will be included in the analysis. Also, pre-season games are not included in the analysis.

First, various attributes were selected to use as potential predicting or response variables.

One of those variables is `moneyline`, a betting wager, which can be converted with the following formula into the win probability based on bets placed before the game. `Spread_line` is another useful variable based on betting odds that is used to even out two uneven teams. Both of these variables are driven by advanced analytics and actual wagering, so they are complex and *do* account for any believed effect of the home team advantage. According to sports news outlets, generally, this accounts for as many as 3 but more realistically 1-2 spread points.

Distance between NFL cities

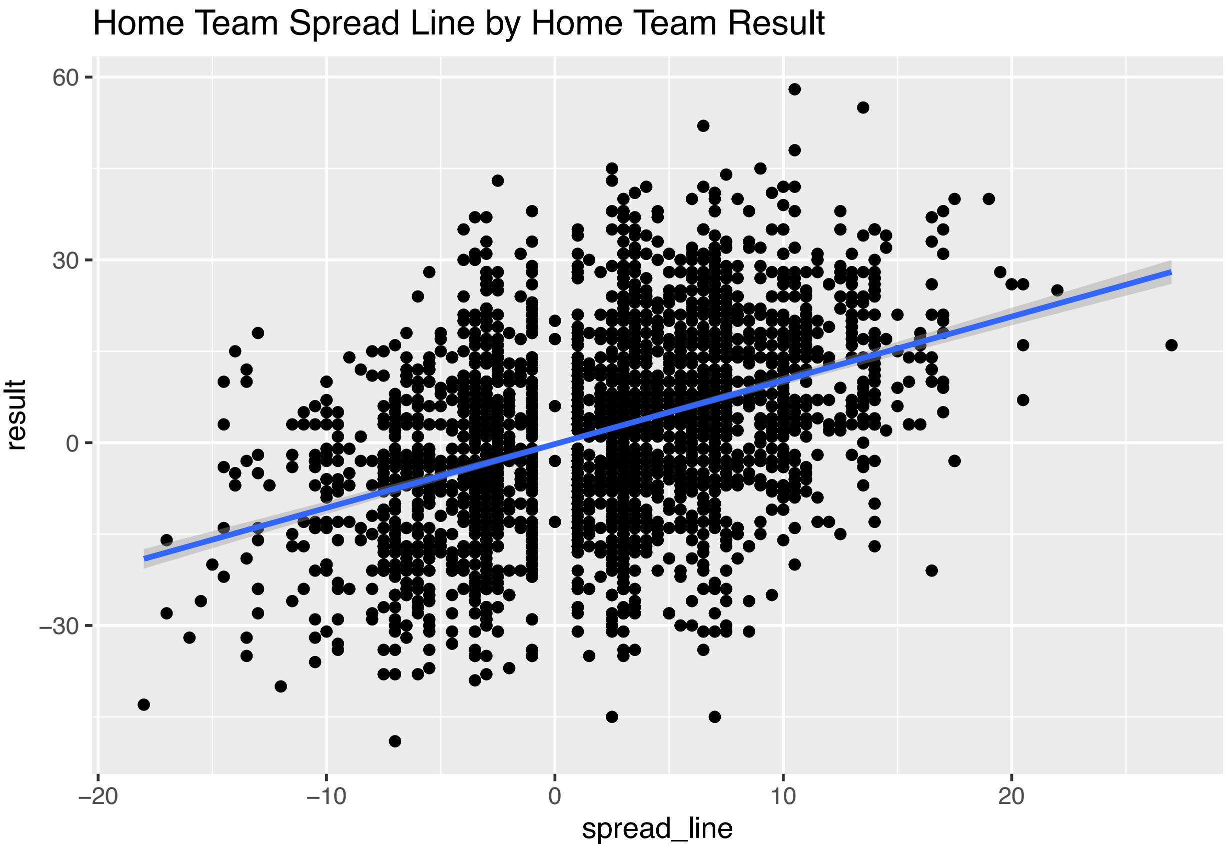
This matrix of distances in miles will be use to determine the distances traveled (in miles) by the away team. Our hypothesis is that this may contribute to the higher rate loss for the away team.

Model Analysis

Analysis 1. Evaluating the spread line

First, as a gut check, a regression analysis will be run to evaluate the accuracy of `spread_line` as a predictor of the actual result of the game. Based on the plot below, there does appear to be a strong linear relationship.

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Using `result` as the response variable and `spread_line` as the predicting variable, we would expect the regression forumula to indicate that the result plus the spread is zero without an intercept, meaning the coefficient of `spread_line` should be 1.

```
##
## Call:
## lm(formula = result ~ spread_line + 0, data = games_select)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.239  -8.585  -0.361   7.804  47.142
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## spread_line  1.03412    0.03672   28.16   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 13.25 on 3223 degrees of freedom
## Multiple R-squared:  0.1975, Adjusted R-squared:  0.1973
## F-statistic: 793.2 on 1 and 3223 DF,  p-value: < 2.2e-16
```

```
## [1] "Is coeff = 1? p-value: 0.352853653933552"
##                2.5 %   97.5 %
## spread_line 0.962127 1.106111
```

Indeed, the coefficient for `spread_line` is statistically significant, the overall regression is statistically significant, and there are no obvious violations of regression assumptions. At an alpha level of 0.05, the coefficient for `spread_line` is **not** statistically significantly equal to 1, but the confidence interval shows that it is likely between 0.96 and 1.12. It is also not statistically significantly less than or greater than 1 either.

It is reasonable to conclude that spread_line is a fairly accurate predictor of the result of a game, but the R-squared value is not particularly high. It would not be relevant as a controlling variable for evaluating home team advantage under the assumption that it takes into account home team advantage, which would mean controlling for it would remove the ability to identify home team advantage individually in our modeling.

Analysis 2. Home Field vs Neutral Field

Every year, there are a few regular season games that are played in a neutral location, usually overseas. The following regression analysis will evaluate if this factor (game played at home) is associated with a decrease in the result. The super bowl is also played in a neutral location, but this is excluded from the analysis because it is not a regular season game. Additionally, only seasons in which there actually were games of this type are included.

```
##
## Call:
## lm(formula = result ~ location, data = games_reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.018  -9.018   0.982   7.982  55.982
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.0177     0.2664   7.575 4.71e-14 ***
## locationNeutral  -4.1528     2.4333  -1.707    0.088 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.71 on 3086 degrees of freedom
## Multiple R-squared:  0.000943,   Adjusted R-squared:  0.0006192
## F-statistic: 2.913 on 1 and 3086 DF,  p-value: 0.08799
```

```
## # A tibble: 2 x 2
##   location games
##   <fct>    <int>
## 1 Home      3051
## 2 Neutral     37
```

The results of this regression analysis can be interpreted as comparing the average result at a game played at the “home” team’s home field versus a game played at a neutral location. The negative coefficient for `locationNeutral` supports an association between playing a team at a neutral location and a decrease in the overall result of the game. The regression coefficient and the overall regression are significant at a 0.05 alpha level, but not at a 0.01 alpha level. There are only 31 games (~1%) that are played in a neutral location however. Additionally the R-squared for this model is extremely low. In summary, this model indicates that a more thorough analysis is needed.

To perform a more thorough analysis, additional predicting and controlling variables should be analyzed, with a focus on the regular season games that *are* played at a home location.

Analysis 3: Detailed Game-Level Predictors

For this analysis, the data will be reformatted to a structure that has the following key variables: `team`, `opponent`, `home` which will be a dummy variable set to 1 if the "team" is the home team, `team_score`, the end score of the "team", `opponent_score`, the end score of the "opponent", and `result`, the team_score minus the opponent_score. Each game will appear twice, once for each team and home and away. Other variables will be kept and assessed as predictor variables for additional analysis. Only regular season games will be considered.

Additionally, a controlling variable `point_diff` will be computed. This variable will be the weighted average point differential (team score - opponent score) for the "team" for that season, weighted ~50% in favor of the prior three games if there are 3 prior games. For the first 3 games of each season, the overall season point differential will be used, but a dummy variable will be added to indicate that this value is computed differently.

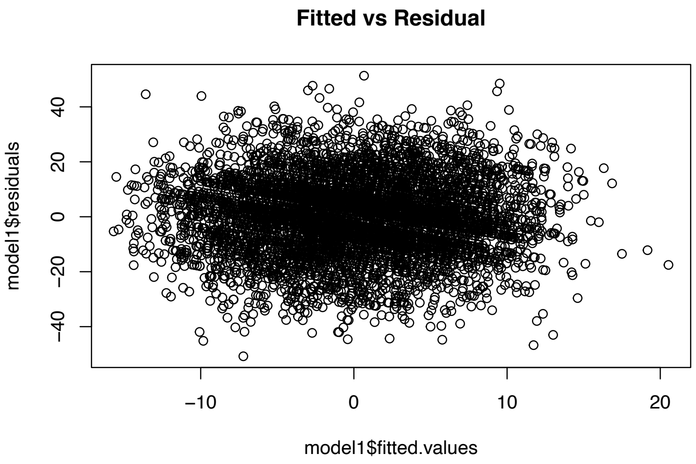
```
## ‘summarise()‘ has grouped output by ’season’. You can override using the
## ‘.groups‘ argument.
```

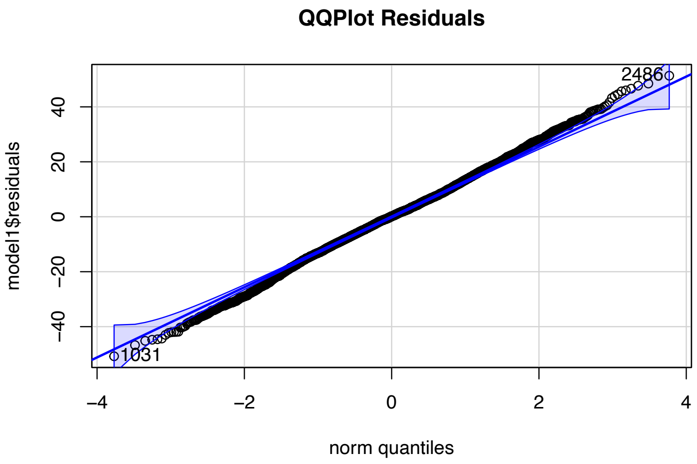
```
## # A tibble: 6 x 23
## # Groups:   season, team [1]
##   season  week gameday             weekday gametime traveled team  opponent
##    <int> <int> <dttm>              <chr>   <chr>       <int> <chr> <chr>
## 1   2010     1 2010-09-12 00:00:00 Sunday  16:15        1517 ARI   STL
## 2   2010     2 2010-09-19 00:00:00 Sunday  13:00        1868 ARI   ATL
## 3   2010     3 2010-09-26 00:00:00 Sunday  16:15         745 ARI   OAK
## 4   2010     4 2010-10-03 00:00:00 Sunday  16:15         358 ARI   SD
## 5   2010     5 2010-10-10 00:00:00 Sunday  16:05        1548 ARI   NO
## 6   2010     7 2010-10-24 00:00:00 Sunday  16:05        1513 ARI   SEA
## # ... with 15 more variables: home <dbl>, team_score <int>,
## #   opponent_score <int>, result <int>, team_rest <int>, opponent_rest <int>,
## #   roof <chr>, surface <chr>, temp <dbl>, wind <int>, team_game_number <int>,
## #   weighted_point_diff <dbl>, first_three_game <dbl>, primetime_game <dbl>,
## #   outdoor_game <dbl>
```

Simple Model The first linear regression model below will use `result` as the response variable and just `weighted_point_diff` and `home` as predicting variables to start.

```
##
## Call:
## lm(formula = result ~ home + weighted_point_diff, data = team_games)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.784  -8.707   0.165   8.531  51.337
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.0798     0.2480  -8.386   <2e-16 ***
## home                  4.1498     0.3507  11.832   <2e-16 ***
## weighted_point_diff   0.7840     0.0264  29.694   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 13.7 on 6099 degrees of freedom
## Multiple R-squared:  0.1426, Adjusted R-squared:  0.1423
## F-statistic: 507.1 on 2 and 6099 DF,  p-value: < 2.2e-16
```

```
## [1] 2486 1031
```





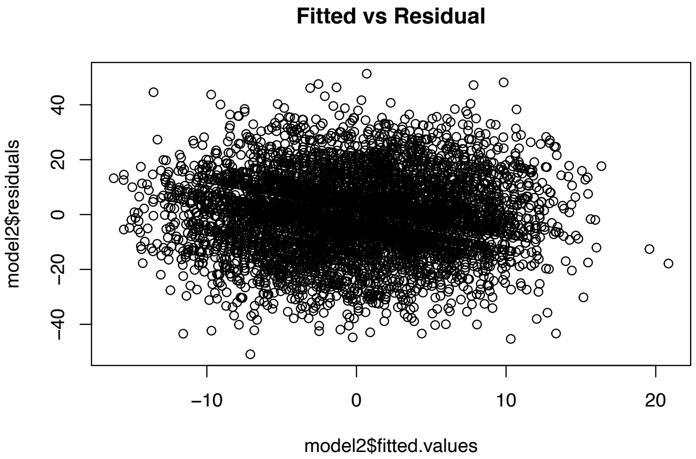
The model is statistically significant overall, and the coefficient for home team is 4.65 and statistically significant, meaning that a team's score played at home is associated with an increase of 4.65 points, holding the weighted point differential constant. The model does not seem to violate normality and constant variance assumptions badly, although there might be a slight tail to the residuals and a possible negative trend in variance. The R-squared value is relatively low, so adding additional predicting variables may help. However, the direction of the result is promising in confirming the home team advantage hypothesis at a high level.

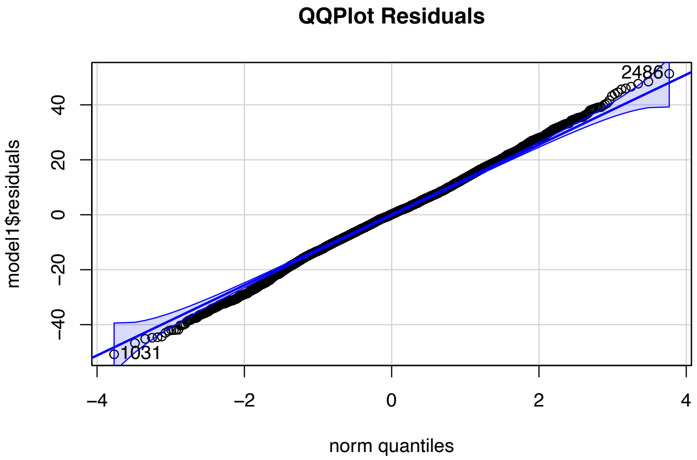
Note: versions of this model removing the first three games or adding an interaction term for the weighted point differential and first three games did not significantly alter the results of this model.

Adding Predicting Variables

```
##
## Call:
## lm(formula = result ~ home + weighted_point_diff + team_rest +
##     opponent_rest + primetime_game + outdoor_game + temp, data = team_games)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.906  -8.743   0.153   8.545  51.313
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.352010   0.900166  -2.613 0.009001 **
## home                 4.176994   0.351232  11.892  < 2e-16 ***
## weighted_point_diff  0.794601   0.026567  29.909  < 2e-16 ***
## team_rest            0.108921   0.089452   1.218 0.223404
## opponent_rest       -0.047464   0.089467  -0.531 0.595766
## primetime_game      -1.803615   0.529212  -3.408 0.000658 ***
## outdoor_game        -0.212524   0.600447  -0.354 0.723393
## temp                 0.005122   0.008519   0.601 0.547698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.69 on 6094 degrees of freedom
## Multiple R-squared:  0.1445, Adjusted R-squared:  0.1435
## F-statistic:   147 on 7 and 6094 DF,  p-value: < 2.2e-16
```

```
## [1] 2486 1031
```





The R-squared improved with the addition of several other variables, and the `home` variable is still statistically significant. —–

Adding injury data: for this, we collected data on reported injuries, filtered out players who either never started or had questionable report statuses. The regression variable will be number of injured players on the home team and the away team.

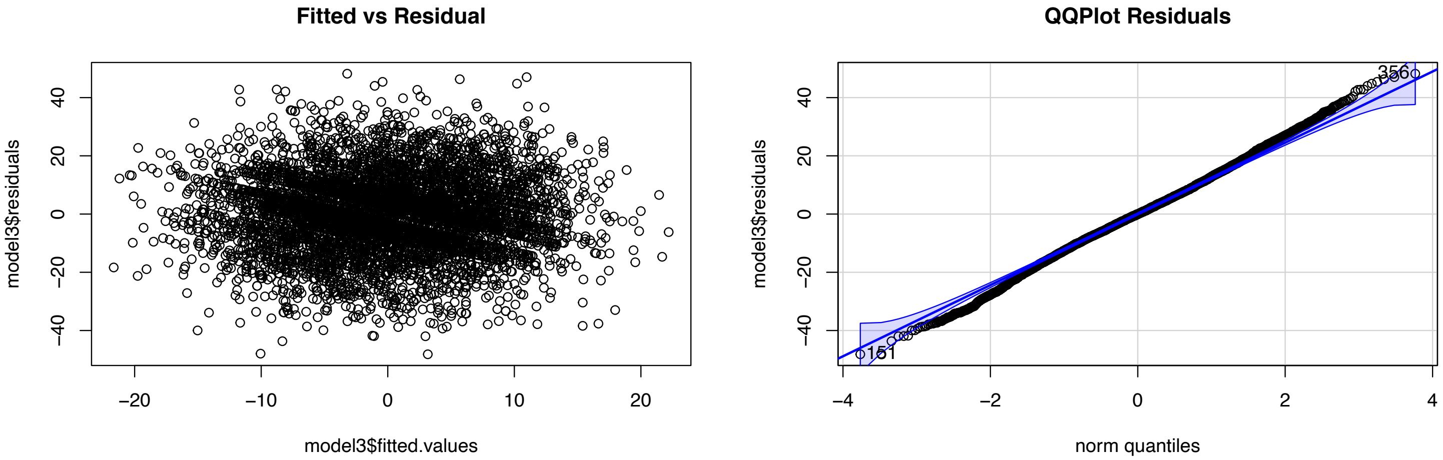
Adding standings: using standings for current season would be "cheating" since it would include information about whether the team won or lost the game, but using the win rate for the previous season would be reasonable here. The limitation is that rosters do change from season to season.

Fit the regression model again adding in all the new variables including: distance traveled, number of injured players on each team, and the win rate of each team in the previous season.

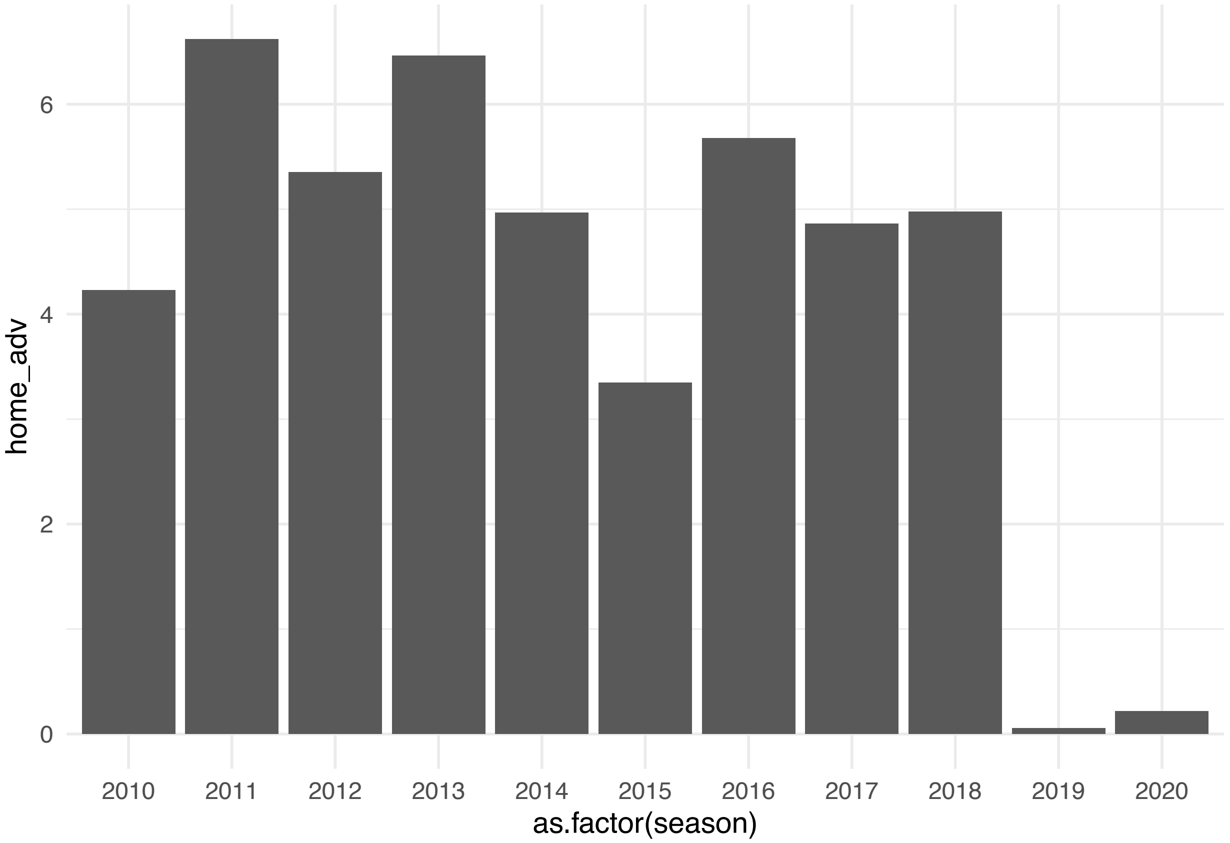
```
##
## Call:
## lm(formula = result ~ home + weighted_point_diff + team_rest +
##     opponent_rest + primetime_game + outdoor_game + temp + n_injured_home +
##     n_injured_away + home_wr_prev + away_wr_prev + as.factor(team) +
##     as.factor(opponent) + traveled, data = df_combine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.163  -8.194  -0.043   8.299  48.205
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.750e-04  1.827e+00   0.000  0.99979
## home                    4.137e+00  3.434e-01  12.048  < 2e-16 ***
## weighted_point_diff     6.944e-01  3.058e-02  22.709  < 2e-16 ***
## team_rest               1.008e-01  8.794e-02   1.147  0.25154
##  [ reached getOption("max.print") -- omitted 77 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.27 on 5927 degrees of freedom
##   (94 observations deleted due to missingness)
## Multiple R-squared:  0.2037, Adjusted R-squared:  0.193
## F-statistic: 18.96 on 80 and 5927 DF,  p-value: < 2.2e-16
```

From this model, it looks like the most relevant variables are which teams are playing each other, home, weighted_point_diff, number of injured players, and previous season win rates.

```
## [1] 356 151
```



Another thing we were interested in looking at is how the home field advantage has changed over time, here we run regression models using the same variables as shown above, except we filter the data by year and then graph the coefficient of the `home` variable. It looks like the home field advantage has been significant from 2010-2018, on average yielding the home team 3-6 point differential advantage. However it shrank significantly in 2019 and 2020, though the home team is still very, very slightly advantaged.



Logistic Regression

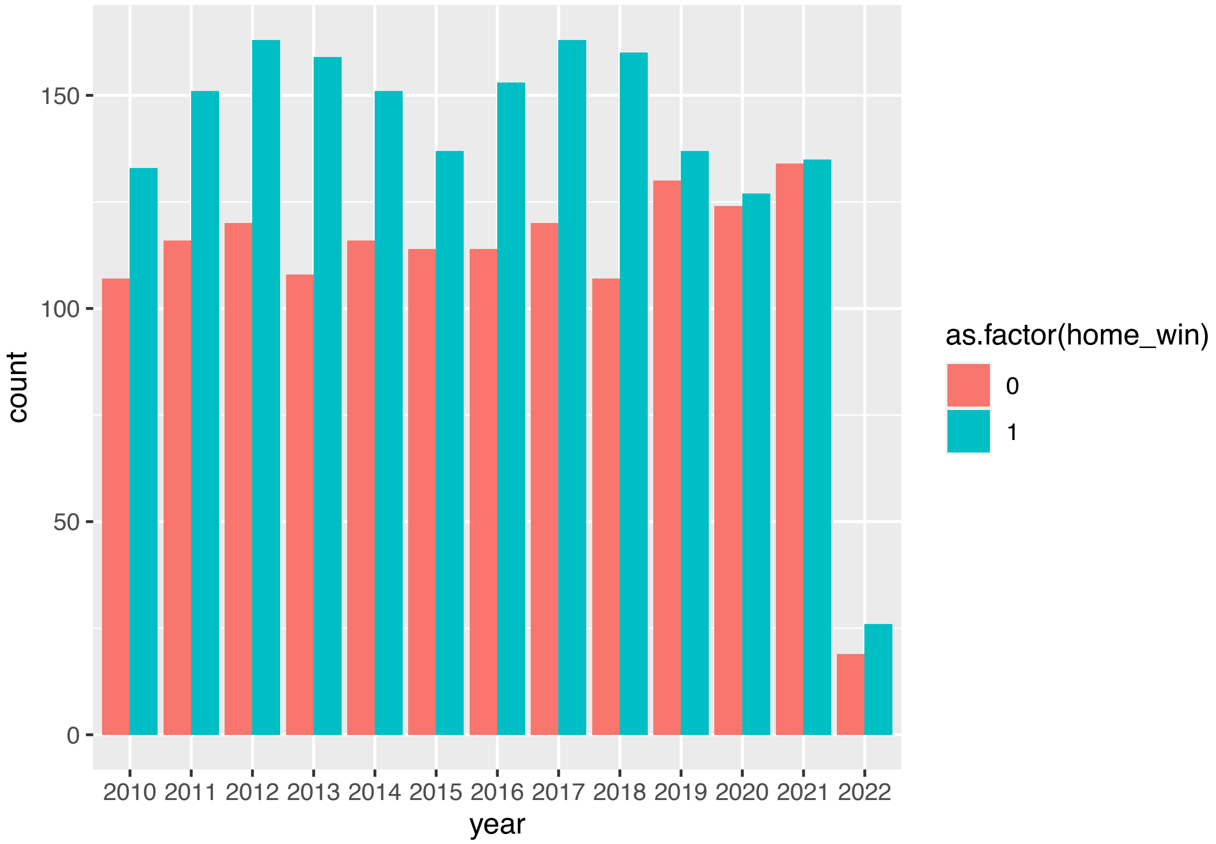
This model uses 'home-win' as binary response variable, where a home-team win = 1 and a home-team loss = 0. The explanatory variables are the same as above. The goal here is to examine if there are any other variables which influence whether the home team is more or less likely to win.

```
##
## Call:
## glm(formula = home_win ~ home + weighted_point_diff + team_rest +
##     opponent_rest + primetime_game + outdoor_game + temp + n_injured_home +
##     n_injured_away + home_wr_prev + away_wr_prev + as.factor(team) +
##     as.factor(opponent) + traveled, family = binomial(link = "logit"),
##     data = df_combine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2817  -1.0322  -0.3953   1.0277   2.2850
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -3.319e-02  2.979e-01  -0.111 0.911274
## home                    5.819e-01  5.653e-02  10.294  < 2e-16 ***
## weighted_point_diff     9.136e-02  5.226e-03  17.480  < 2e-16 ***
## team_rest               2.173e-02  1.447e-02   1.502 0.133082
##  [ reached getOption("max.print") -- omitted 77 rows ]
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8328.8  on 6007  degrees of freedom
## Residual deviance: 7349.5  on 5927  degrees of freedom
##   (94 observations deleted due to missingness)
## AIC: 7511.5
##
## Number of Fisher Scoring iterations: 4
```

Like the linear models, the only statistically significant variables in the logistic model are control variables.

Effect of “fans in stands”

None of the explanatory variables explored have proved to have relationship with home-team wins. However, there is still an anomaly in the data that can be analyzed. The phenomenon of home-team advantage appears to be declining in recent years. Anecdotally, home-team advantage is often attributed to the psycho-social effect of the cheers from fan when playing at home. The pandemic provides a useful natural experiment to test this hypothesis. In 2020 and 2021, games were played with no fans in the stands.



Here we have created a binary variable for games played with no fans in the stands and use it as an explanatory variable in a logistic regression.

```
##
## Call:
## glm(formula = home_win ~ pandemic, family = binomial(link = "logit"),
##     data = df_combine)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.175  -1.175  -1.174   1.180   1.181
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.005718   0.026736  -0.214    0.831
## pandemic    -0.002187   0.092844  -0.024    0.981
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8459.1  on 6101  degrees of freedom
## Residual deviance: 8459.1  on 6100  degrees of freedom
## AIC: 8463.1
##
## Number of Fisher Scoring iterations: 3
```

```
##
## Call:
## glm(formula = home_win ~ home + weighted_point_diff + team_rest +
##     opponent_rest + primetime_game + outdoor_game + temp + n_injured_home +
##     n_injured_away + home_wr_prev + away_wr_prev + as.factor(team) +
##     as.factor(opponent) + traveled + pandemic, family = binomial(link = "logit"),
##     data = df_combine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2791  -1.0320  -0.3951   1.0290   2.2703
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -4.359e-02  2.989e-01  -0.146 0.884062
## home                    5.819e-01  5.653e-02  10.294  < 2e-16 ***
## weighted_point_diff     9.141e-02  5.228e-03  17.485  < 2e-16 ***
## team_rest               2.178e-02  1.447e-02   1.505 0.132248
##  [ reached getOption("max.print") -- omitted 78 rows ]
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8328.8  on 6007  degrees of freedom
## Residual deviance: 7349.3  on 5926  degrees of freedom
##   (94 observations deleted due to missingness)
## AIC: 7513.3
##
## Number of Fisher Scoring iterations: 4
```

This regression suggest that the log-odds of a home-team win are reduced by -.257 when no fans are in the stands giving credence to the assertion that the support of fans improves the teams’ performance. However, when we include this variable in a regression with all other explanatory variables, ‘pandemic’ is not significant.

Random Forest

Running a random forest regression model will achieve optimal predictability however, this comes at the cost of gaining insights from our data. We will create a “black box” if you will and will be difficult to explain specific effects and determine why an outcome is the way it is.

```
##
## Call:
##  randomForest(formula = home_win ~ season + weekday + gametime +      away_team + home_team + overtime + home
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##         OOB estimate of  error rate: 37.75%
## Confusion matrix:
##     0    1 class.error
## 0 471  958   0.6703989
## 1 259 1536   0.1442897
```

```
##
## Call:
##  randomForest(formula = home_win ~ season + weekday + gametime +      away_team + home_team + overtime + home
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 36.29%
## Confusion matrix:
##     0    1 class.error
## 0 775  654   0.4576627
## 1 516 1279   0.2874652
```

```
##
## Call:
##  randomForest(formula = home_win ~ season + weekday + gametime +      away_team + home_team + overtime + home
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 37.56%
## Confusion matrix:
##     0    1 class.error
## 0 761  668   0.4674598
## 1 543 1252   0.3025070
```

```
##
## Call:
##  randomForest(formula = home_win ~ season + weekday + gametime +      away_team + home_team + overtime + home
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 7
##
##         OOB estimate of  error rate: 38.49%
## Confusion matrix:
##     0    1 class.error
## 0 762  667   0.4667600
## 1 574 1221   0.3197772
```

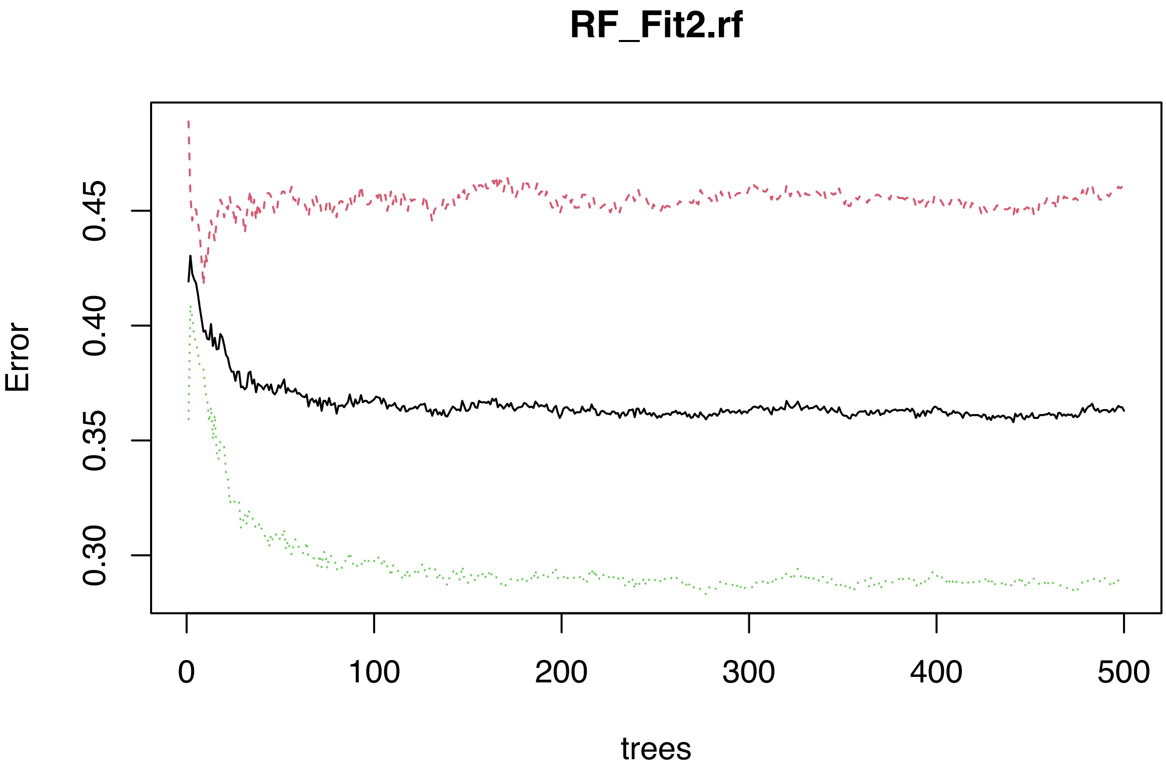
```
##
## Call:
##  randomForest(formula = home_win ~ season + weekday + gametime +      away_team + home_team + overtime + home
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 13
##
##         OOB estimate of  error rate: 37.84%
## Confusion matrix:
##     0    1 class.error
## 0 759  670   0.4688593
## 1 550 1245   0.3064067
```

The best mtry (number of predictors sampled for spliting at each node) was mtry = 2 with an OOB estimate of 36.29

```
##
## Call:
##  randomForest(formula = home_win ~ season + weekday + gametime +      away_team + home_team + overtime + home
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 36.29%
## Confusion matrix:
##     0    1 class.error
## 0 775  654   0.4576627
## 1 516 1279   0.2874652
```

```
##       OOB
## 0.3629032
```

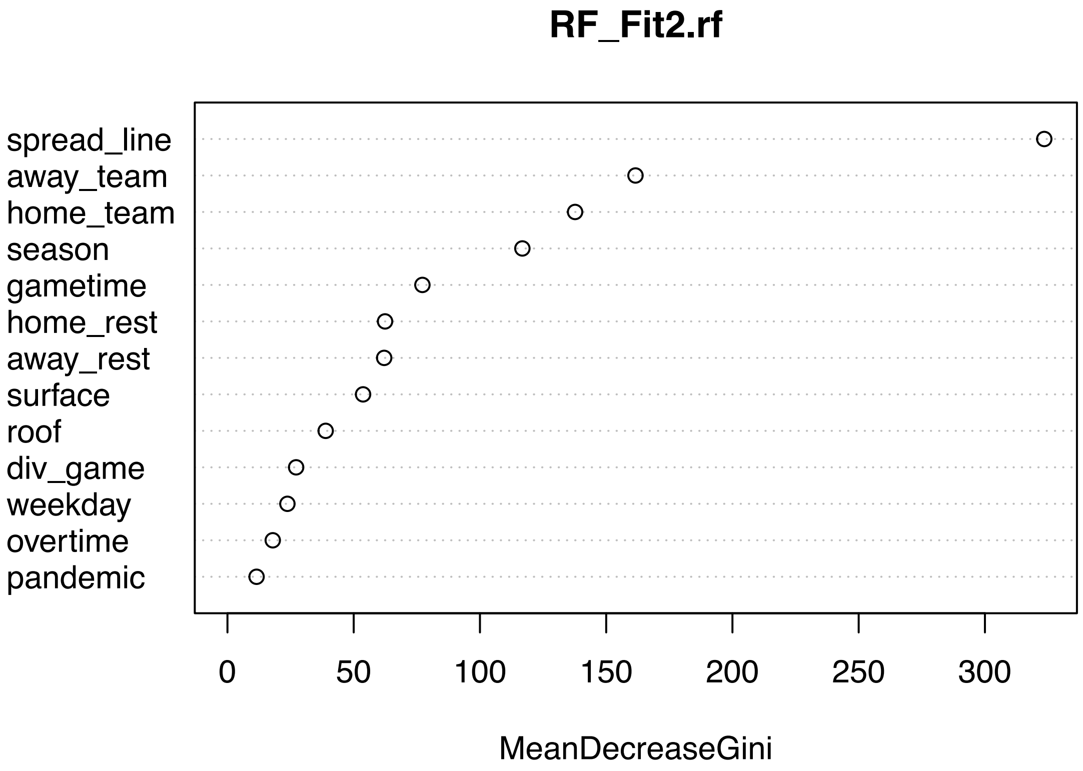
Plotting the model will help us visualize the OOB error rate (black line) as trees are averaged across. This will show us that our error rate stabilizes with around 75 trees and slowly decreases therefore after.



let's try to determine feature importance

```
##                    0           1 MeanDecreaseAccuracy MeanDecreaseGini
## season     0.5147323  2.48739650            2.1180056        116.79734
## weekday   -0.4185777  1.84517307            1.4475372         23.74979
## gametime   1.2919803  1.10319873            1.7787278         77.20774
## away_team  3.6682887  1.35466897            3.5269521        161.59639
## home_team  1.8117437  4.91351288            5.2758136        137.67613
## overtime   3.5452904  0.13802377            2.5438429         17.94921
## home_rest  2.5692296  1.22519487            2.9250988         62.41413
## away_rest  0.6968676  1.30871995            1.5785382         62.06560
## div_game  -1.5521825 -0.37947476           -1.4206282         27.19864
## roof       4.8232386  5.34202829            8.1856312         38.89189
## surface    4.1836580  2.53674663            4.9346132         53.70602
## pandemic  -0.9577515 -0.08407703           -0.7758792         11.53091
## spread_line 60.3373649 53.10487333           70.1230434        323.49724
```

Visualizing the importance of features against accuracy, we notice that spread_line is of most importance in determining if home team wins or not, remember this is our control variable to determine if the team is good or not.



The OOB estimate of error rate is 36.29%, which is only slightly better than random guessing… not too great.

```
##
## Call:
##  randomForest(formula = home_win ~ season + weekday + gametime +      away_team + home_team + overtime + home
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 36.29%
## Confusion matrix:
##     0    1 class.error
## 0 775  654   0.4576627
## 1 516 1279   0.2874652
```

Results

Linear Models

```
##    Model        R2    Adj_R2 Num_Vars      MSE     SSE      MAE
## 1 Model1 0.1425707 0.1422895        2 187.5485 1144421 10.69114
## 2 Model2 0.1444697 0.1434869        7 187.1332 1141887 10.68059
## 3 Model3 0.2037454 0.1929979       80 216.4366 1320696 11.43552
```

Logistic Model

```
##                 Model      AIC Deviance Num_Vars
## 1      Logistic Model 7511.460 7349.460       80
## 2      Pandemic Model 8463.115 8459.115        1
## 3 Pandemic Full Model 7513.281 7349.281       81
```

Random Forest Model

```
##    Model       OOB
## 1  mtry1 0.3774814
## 2  mtry2 0.3629032
## 3  mtry4 0.3756203
## 4  mtry7 0.3849256
## 5 mtry13 0.3784119
```

Conclusion

Predicting the outcome of NFL games, whether regressing on the point differential or using logistic regression on win or loss turned out to be not so trivial of a task. None of the linear or logistic regression models tested had the greatest fit.Using a Random Forest to obtain a model for highest prediction accuracy was a good idea early on in the project, however we found that the accuracy of a random forest model, which we expected to have a large advantage over regression models, was not significant enough. The out of box error we arrived at for random forest was about 36%. This means that the model is not much better than a random guess at predicting home wins. The model error gives our variable importance analysis results some ambiguity. Ideally, we would have achieve a much higher accuracy rate thus giving us more confidence in the features that are significant/important toward home-wins.

In terms of whether or not home field advantage is significant, our findings agree with existing literature which suggest that it is. We were also able to observe that the home field advantage has a sharp drop off in 2019 and 2020. In 2020, this can at least partially attributed to the lack of fans in the stands. Other relevant factors which seemed to have statistically relevant effect on win/loss in regression models besides home field advantage were number of injured players on each team, the performance of each team in the previous season, and the weighted point differential of each team in their last three games.

References

Data Sources

2021 NFL Game Data. (n.d.). Retrieved from http://www.habitatring.com/

NFL Football Stadiums - Quest for 31. (n.d.). Retrieved from http://www.nflfootballstadiums.com/

Nflverse. (Sharpe, Lee). Nfldata/rosters.csv at master · nflverse/nfldata. Retrieved from https://github.com/nflverse/nfldata/blob/master/data/rosters.csv

Literature

Cleveland, T. (2021, September 14). Numbers that matter for predicting NFL win totals: Sharp Football. Retrieved from https://www.sharpfootballanalysis.com/betting/numbers-that-matter-for-predicting-nfl-win-totals-part-one/

Jamieson, J. P. (2010). The Home Field Advantage in Athletics: A Meta-Analysis. Journal of Applied Social Psychology, 40(7), 1819-1848. doi:10.1111/j.1559-1816.2010.00641.x

Kilgore, A., & Greenberg, N. (2022, January 15). Analysis | NFL home-field advantage was endangered before the pandemic. Now it's almost extinct. Retrieved from https://www.washingtonpost.com/sports/2022/01/14/nfl-home-field-advantage-pandemic/

Mccarrick, D., Bilalic, M., Neave, N., & Wolfson, S. (2021). Home advantage during the COVID-19 pandemic: Analyses of European football leagues. Psychology of Sport and Exercise, 56, 102013. doi:10.1016/j.psychsport.2021.102013

Ponzo, M., & Scoppa, V. (2014). Does the Home Advantage Depend on Crowd Support? Evidence from Same-Stadium Derbies. SSRN Electronic Journal. doi:10.2139/ssrn.2426859

Swartz, T. B., & Arce, A. (2014). New Insights Involving the Home Team Advantage. International Journal of Sports Science & Coaching, 9(4), 681-692. doi:10.1260/1747-9541.9.4.681