MGT 6203 Group Project

Phase 3: Final Report
Date 04/27/2022

Chris Hom, Liliann Teister, Joanna Rashid, Stephen Yu

Georgia Tech

Introduction

- Home field advantage is well documented phenomenon in which teams playing on their home field benefit from a competitive advantage
- The goal of our project is to explore the influence of home field advantage in the National Football League (NFL)
  - Is this a significant factor that can influence which team wins?
  - What factors, if any, influence the effect size of home field advantage?
- There is some evidence that the home field advantage is shrinking
  - Is this supported by data and if so, can we find any explanations?

Approach

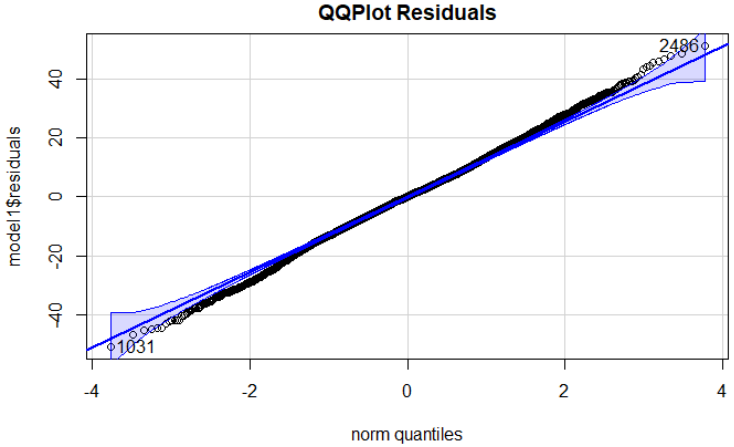
- Dataset of NFL games from 2010-2020 was the primary dataset used - attributes in this dataset include but are not limited to: teams, score, location, gametime, when the game was played, playoffs vs regular season, overtime, stadium conditions, and more
- Secondary datasets used:
  - NFL standings data 2009-2020
  - NFL player injuries and roster data
  - Distances between NFL cities
- Data was cleaned and combined, we then ran various linear and logistic regression models and looked at the effect size of home field advantage
- We also used variable selection methods and random forest models to estimate the importance of variables in determining a home win

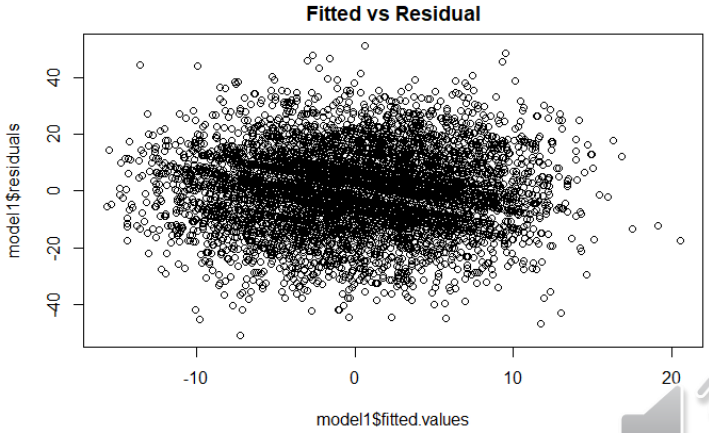
Linear Regression Model 1

Model 1

model1 <- lm(result ~ home + weighted_point_diff, data=team_games)

- The model was ran with result - the point differential between the two teams as the response and home field advantage as well as weighted pointed differential for the last 3 games played as the predictor variables
- 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 (control for team strength) 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.

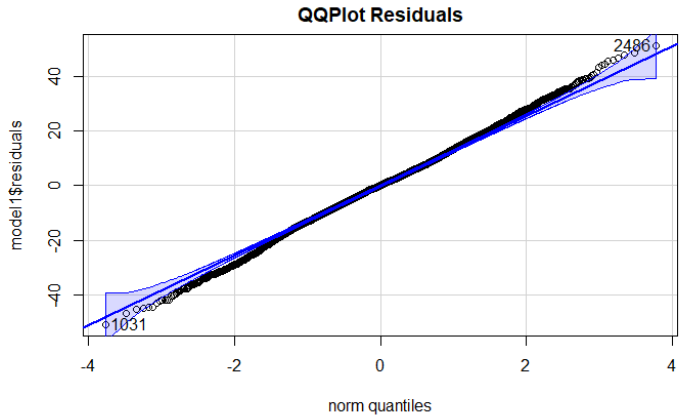


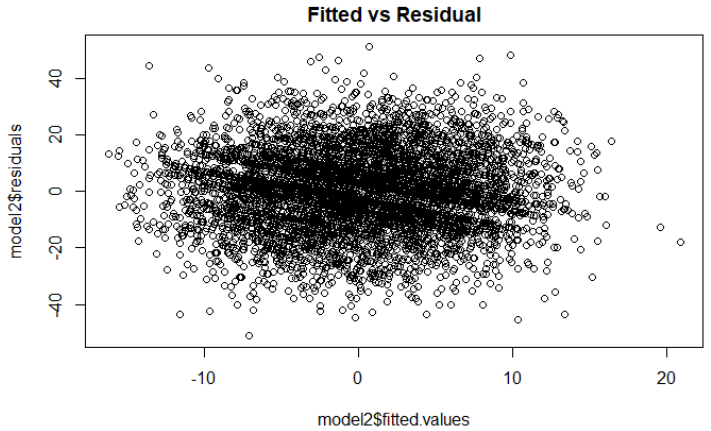


Linear Regression Model 2

model2 <- lm(result ~ home + weighted_point_diff + team_rest + opponent_rest + primetime_game + outdoor_game + temp, data=team_games)

- Model 2 includes more more variables than model 1, additional variables include:
  - Rest time of each team
  - Outdoor or indoor game
  - Temperature on game day
- Adding more variables shrank the effect of home field advantage, yielding 4.18 point advantage on average when other variables are held constant
- The residuals in this model look similar to model 1 with a slight tail on QQ plot, but are not obviously abnormal

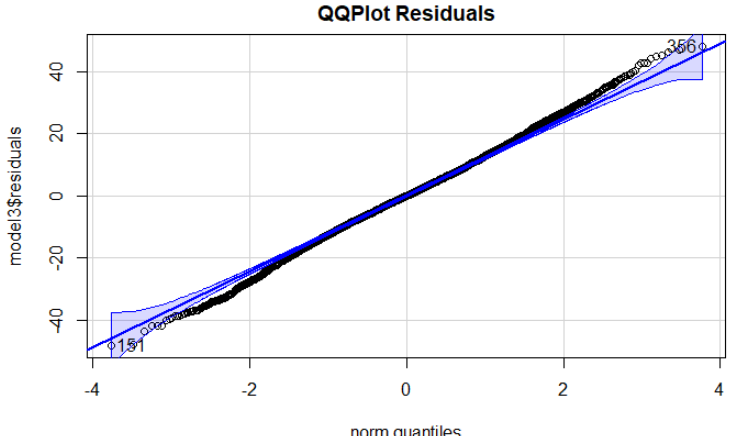


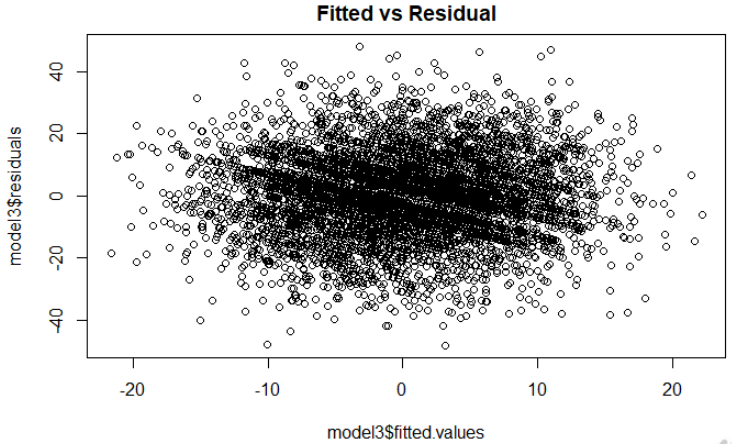


Linear Regression Model 3

model3 <- lm(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)

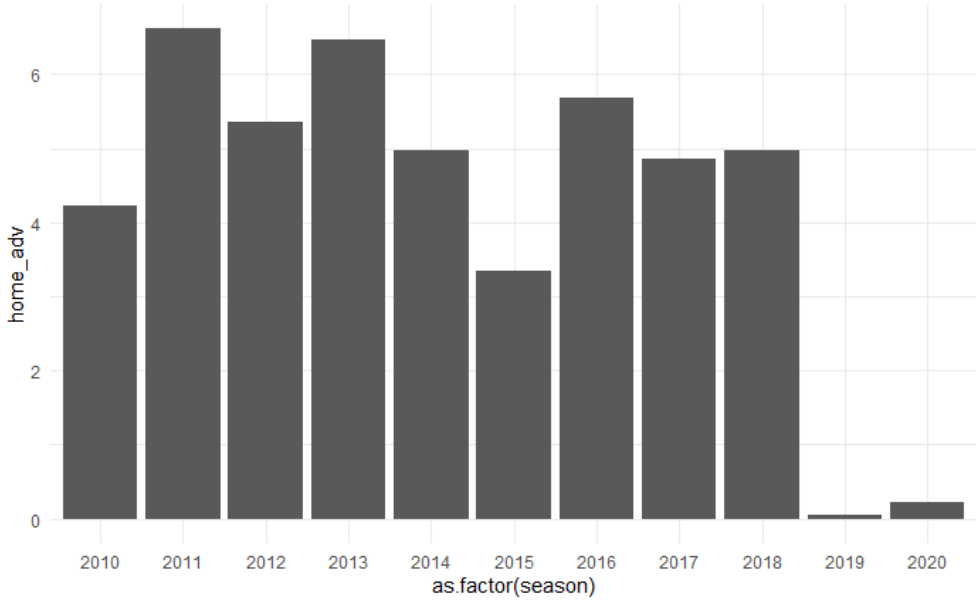
- Model 3 includes even more variables than model 2:
  - Number of injured player on each team
  - Win rate of each team in the previous season
  - Which team is home/away one hot encoded
- Home field advantage coefficient does not significantly change from model 2 to 3, though R^2 improves slightly
- No obvious issues with residuals observed in model 3





Home field advantage has been declining

- Linear regression was ran using point differential as the response variable, and predictors included the same ones as linear regression model 3 on each NFL season
- From 2010-2018, home field advantage yielded approximately 4-6 point advantage on average
- In 2019 and 2020, this advantage dropped off steeply to near 0



Summary of Linear Regression Models

Model <chr>	R2 <dbl>	Adj_R2 <dbl>	Num_Vars <dbl>	MSE <dbl>	SSE <dbl>	MAE <dbl>
Model1	0.1425707	0.1422895	2	187.5485	1144421	10.69114
Model2	0.1444697	0.1434869	7	187.1332	1141887	10.68059
Model3	0.2037454	0.1929979	80	216.4366	1320696	11.43552

3 rows

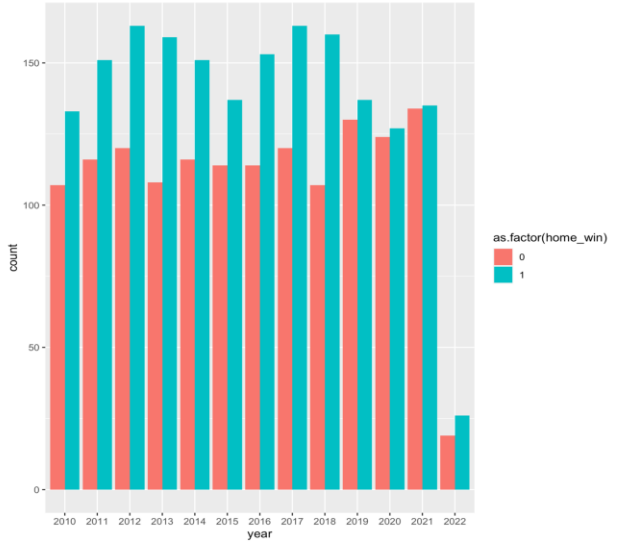
- For the 3 models tested, model 3 performed the best with 80 variables (mostly due to one hot encoded teams) by R^2 and adjusted R^2
- However, by error metrics (MSE - mean squared error, SSE - sum of squared errors, MAE - mean absolute error), model 3 actually performs slightly worse than the other 2
- Generally, none of the models had great predictive power as even in the best case, only ~20% of variance is explained by the predictors

Georgia Tech

Logistic Regression Models

- logistic_model <- glm(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, data = df_combine

- In search of a better model, we regressed the same variables on 'home-win', which is a binary variable where a home-team win =1 and a home-team loss = 0.
- In this model too, only control variables were statistically significant.
- 'Pandemic model' regresses a binary variable for games played with no fans in stands which had a coefficient of -0.002 at a level of 95% significance, indicating fans in the stands is slightly associated with home-team wins.
- When the pandemic variable is includes with all other predictor variables ('Pandemic Full Model') no explanatory variables were found to be significant.



Model <chr>	AIC <dbl>	Deviance <dbl>	Num_Vars <dbl>
Logistic Model	7511.460	7349.460	80
Pandemic Model	8463.115	8459.115	1
Pandemic Full Model	7513.281	7349.281	81

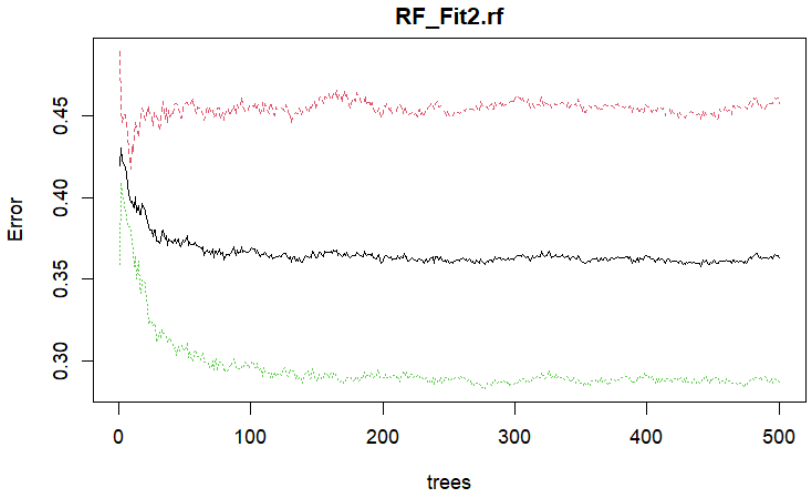
Georgia Tech

Random Forest

Approach:
Centered parameter tuning around ntree = 500. This decision was made in response to research done (ref: Philipp Probst & Anne-Laure Boulesteix "To tune or not to tune the number of trees in random forest?")

Ran multiple models with an adjusted mtry over various values (1, 2, 4, 7, 13)

Optimal Model:
Mtry = 2 @ an OOB estimate of error: 36.29%



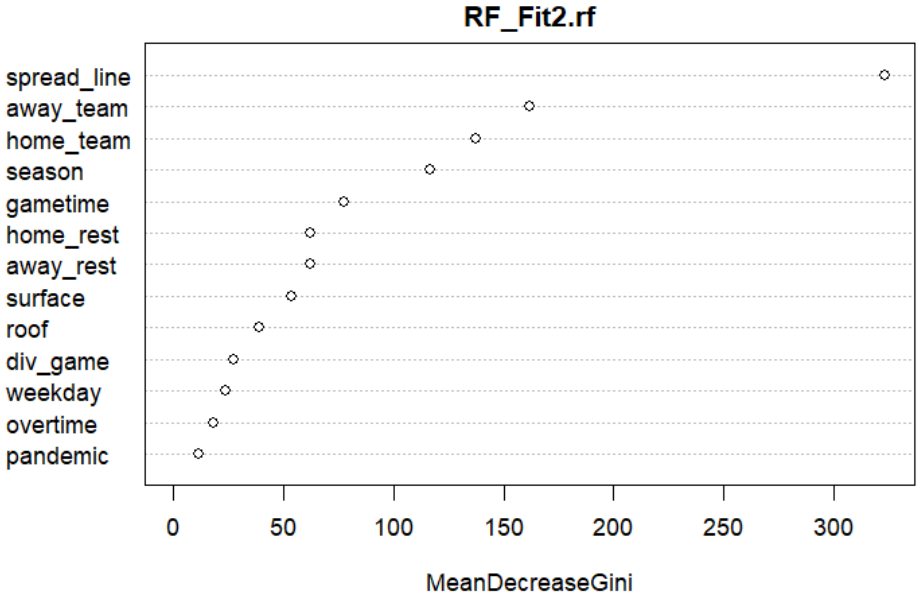
Model <chr>	OOB <dbl>
mtry1	0.3774814
mtry2	0.3629032
mtry4	0.3756203
mtry7	0.3849256
mtry13	0.3784119

Georgia Tech

Random Forest

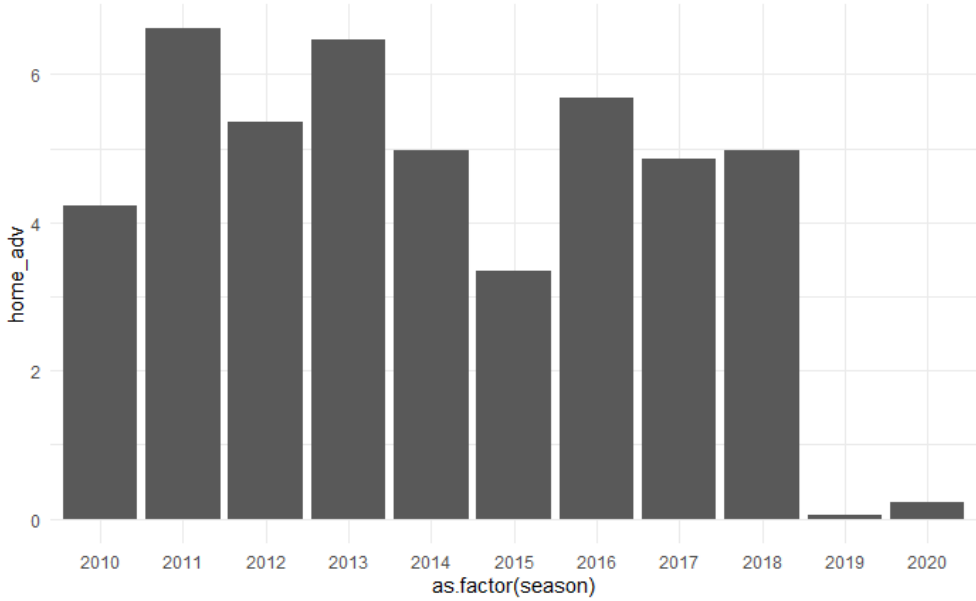
Takeaways:

- Random Forest accuracy was not as strong as initially expected.
- 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 prefaces our variable importance analysis results with some ambiguity. Ideally, we would have achieved a much higher accuracy rate thus giving us more confidence in the features that are significant/important toward home-wins



Home field advantage has been declining

- Linear regression was ran using point differential as the response variable, and predictors included the same ones as linear regression model 3 on each NFL season
- From 2010-2018, home field advantage yielded approximately 4-6 point advantage on average
- In 2019 and 2020, this advantage dropped off steeply to near 0



Georgia Tech

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 was an 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. Ideally, we would have wanted a much higher accuracy rate thus giving us more confidence in the features that the model shows as more 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/Sources

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

References:

- Philipp Probst & Anne-Laure Boulesteix "To tune or not to tune the number of trees in random forest?"
- 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

Georgia Tech