A Machine Learning Approach to Home Field Advantage in the NFL

PPOL 564 Final Project

Word count: 2966

Chau Nguyen

December 20, 2020

1 Introduction

In this project, I want to learn what constitutes "Home field advantage" in the National Football League (NFL). Using machine learning techniques and unique game-level data spanning 28 seasons, I aim to answer the question: "Why do home teams tend to win more?"

The report is organized as followed: Section 2 gives an overview of what home field advantage is, Section 3 discusses the data used, Section 4 explains the analysis, Section 5 lays out the results of my findings, and Section 6 ends with further discussions on the topic.

2 Problem Statement and Background

"Home field advantage" is often taken for granted in sports analysis - teams are *assumed* to have a competitive edge when playing at their home stadiums. In a meta-analysis across a variety of sports, game types and error, Jamieson (2010) found that home teams do win more and hypothesized that crowd noise is one the main reasons. As the phenomenal exists across all sports, the NFL is no exception. According to FiveThirtyEight.com, "the NFL home field advantage is real" and "over 57 percent of games are won by home teams" (Hermsmeyer, 2018). Hermsmeyer (2018), on the other hand, does not find that crowd noise affects the "false start"[1] rate for road teams in the league. Similarly, Moskowitz and Wertheim (2012) did not find that crowd noise affects the away team's performance in other sports.

Then the 2020 covid-19 global pandemic took place. If the NFL wanted to have a season at all, teams needed to adhere to local health guidelines. With stadiums having to limit crowd capacity (or not allow fans in the stands whatsoever), pundits and analysts made bold claims that "There's no such thing as home-field advantage in the NFL this season" because "fans, if any, in the stands means less crowd noise to disrupt the road team's offensive flow" (Greenberg, 2020). Contradicting beliefs on what makes "home field advantage" what it is does not change analysts' beliefs that the phenomenon exists.

My theory is that "home field advantage" in the NFL is more about what the *away team* has to do in other to play the game than the home team winning. For instance, the Seattle Seahawks, a West Coast team, is scheduled to travel 28,082 miles in 2020, while the Baltimore Ravens, an East Coast team, is only supposed to fly 6,420 miles to get to their games - this is a significant discrepancy (Breech, 2020). I think that the 3-hour difference in time zones from coast-to-coast traveling can also affect a team's performance, and the more a team has to fly, the less time they get to rest. With this project, I aim shed light on what factor affects "home field advantage" in the NFL using machine learning.

3 Data

3.1 Source

I built several webscrapers using python package BeautifulSoup (Richardson, 2007) to scrape the dataset. Data on stadiums was scraped from Wikipedia.com and data on team records, game schedule and fan attendance was scraped from Pro-Football-Reference.com (PFR).

[1] After the neutral zone has been established (ball is made or declared ready for play), an offensive player may not make a false start, a defensive player may not encroach (initiate contact with a member of the offensive team) or commit a neutral zone infraction, and no player of either team may be offside when the ball is put in play. (NFL Football Operations, 2020b)

3.2 Coverage

The scraped data from PFR covers 28 NFL seasons from 1992 to 2019, totaling 7292 unique games.

3.3 Unit of observation

The unit of observation is the outcome of each individual game.

3.4 Unit of analysis

Game data is reshaped into a dyadic dataset, with each dyad represents a single game. Each member of the dyad represents the one of the two teams in the game and thus distinguishable from the other.

3.5 Variables of interest

Target variable

The target variable is a dichotomous variable *Win* which equals 1 if the first team in the dyad (Team_A) has a positive points differential (which means they have won the game) and equals 0 if the points differential for Team_A is negative.

Feature Variables

The following variables are used to capture the "Home field advantage" of each team::

- *Time_rest_days*: Continuous variable measuring the number of days between two games for Team_A.
- *Miles_traveled*: Continuous variable measuring the distance between the coordinates of Team_A and Team_B's home stadiums. This variable was calculated using the Great-Circle distance formula following the methods described by Houston (2019).
- *Time_diff*: Continuous variable measuring the number of hours in timezone between a team's home stadium and the stadium the game is being played at.
- *Capacity*: Continuous variable measuring the seating capacity for each stadium for each season.

- *Attendance_pct*: Continuous variable calculating the number of fans in attendance as a percent of capacity.
- *Grass*: Dichotomous variable equaling 1 if the stadium the game is being played at has a grass surface and equals 0 if the stadium has an artificial turf surface.
- *Same_surface*: Dichotomous variable equaling 1 if the stadium surface material is the same as the surface material of the home stadium of Team_A and equals 0 if it's not. If Team_A is the home team, this variable most likely equals 1 except for special circumstances which I will go into more details in section 3.7.

The following variables are used to capture features unrelated to "Home field advantage" that can affect the outcome of a game:

- *Season_offense*: Continuous variable measuring Team_A's offensive rating for the season. This variable is PRF's Offensive Simple Rating System (OSRS).
- *Season_defense*: Continuous variable measuring Team_A's defensive rating for the season. This variable is PRF's Defensive Simple Rating System (DSRS).
- *Rivalry*: Ordinal variable to proxy for the level of rivalry between 2 teams and the frequency they play each other [2]. This variable takes value 0 if each team in the dyad belongs to a different conference (AFC and NFC); 1 if both teams are in the same conference but different divisions; and 2 if both teams are in the same conference and division.
- *Week*: Ordinal variable that captures the week of the season. This variable ranges from 0 to 18 for the regular season [3] and contains strings for the Wild Card round, the Divisional round, the Conference Championship, and the Super Bowl.
- *Season*: Continuous variable for the year the each league year so that games played after December 31st of the year would be counted as games in the same season.

[2] In the current NFL model with 32 teams, each team plays 16 regular season games: 6 against the other 3 teams in its division, 4 against another division in its conference, 2 against one team from each of the remaining divisions in its conference, and 4 against a division in the other conference (NFL Football Operations, 2020a)

[3] There are 17 weeks in the NFL regular season - teams play one game each week and get one "bye" in the middle of the season to recuperate. The only exceptions were the 1993 season where an additional "bye" week was built into the schedule, and the 2001 season when games were postponed a week following the attacks on September 11 (Wikipedia contributors, 2020g).

- *Regular*: Dichotomous variable that equals 1 if the game is a regular season game and 0 if it's a playoff game.
- *International*: Dichotomous variable that equals 1 if the game is a played at an overseas stadium and 0 if it's not.

3.6 Potential issues

Tied games

In order to keep the classifying problem simple, I only want games with a Win-Loss outcome. However, during the regular season, tie games can happen in the NFL after one overtime period that lasts 10 minutes (Wikipedia contributors, 2020e). The rule notwithstanding, tied games are very rare in the NFL. There are 12 instances of tie games in the dataset of 7292 games, accounting for 0.16% of the entire dataset spanning 28 seasons. Therefore, I was comfortable dropping tied games from the full dataset before doing the training/ test splits and analysis.

Missingness in ***Time_rest_days*** variable

Missing data due to the lag variables - for most teams, I could get 1991 data. However, teams that existed after 1992 will miss the first year of data.

One of the variables which I hypothesize to play a role in a team's "Home field advantage" is the time between games for a team. Although most teams in the NFL play a game every Sunday, there are a few exceptions to the 7 days of rest: Each week there is a game played on Monday night and one played on Thursday night. For instance, a team that last played on Sunday would only have full days to recover if their next game happens on a Thursday night. Moreover, since each team is awarded a "bye" week during the regular season, a team with 14 days of rest might have competitive advantage over one with 7. Similarly, in the playoffs, the top seeds would get a first round "bye," furthering the advantage.

However, the inclusion of this variable created a synthetic issue with missingness for the first week of each season: teams don't have rest times if they haven't played any prior games. I thought about using the team's final pre-season game to calculate time rest, but there were

2 issues: 1) Pre-season schedules going far back are difficult to find, and 2) Starters often do not play the last pre-season game.

Thus, I opted to impute the missing rest time for Week 1 for every team using the season's longest regular-season time rest. This means that every team would have the same time rest for Week 1 each season, which is the season's longest bye week[4]. This ensures that the first week of data is not dropped unnecessarily from the dataset.

The ever-changing NFL

Over the 28 seasons covered in the dataset, NFL teams have played in 68 different stadiums in 48 cities across 4 countries [5]. Over this span of time, teams have relocated, changed names [6], changed divisions, etc. Not only that, even franchises with long histories would be forced temporarily move due stadium construction or reconstruction. Figure 1 shows the 48 cities where the NFL has held games since 1992. This proved to be the toughest challenge in the data wrangling steps, which I will detail in section 3.7.

[4] A team can have a "bye" longer than 14 days if their last game before the bye was on Thursday night and first game back after the bye was on Monday night, for example.

[5] The United States, Canada, the United Kingdom, Mexico.

[6] There are 38 unique team names in the dataset.

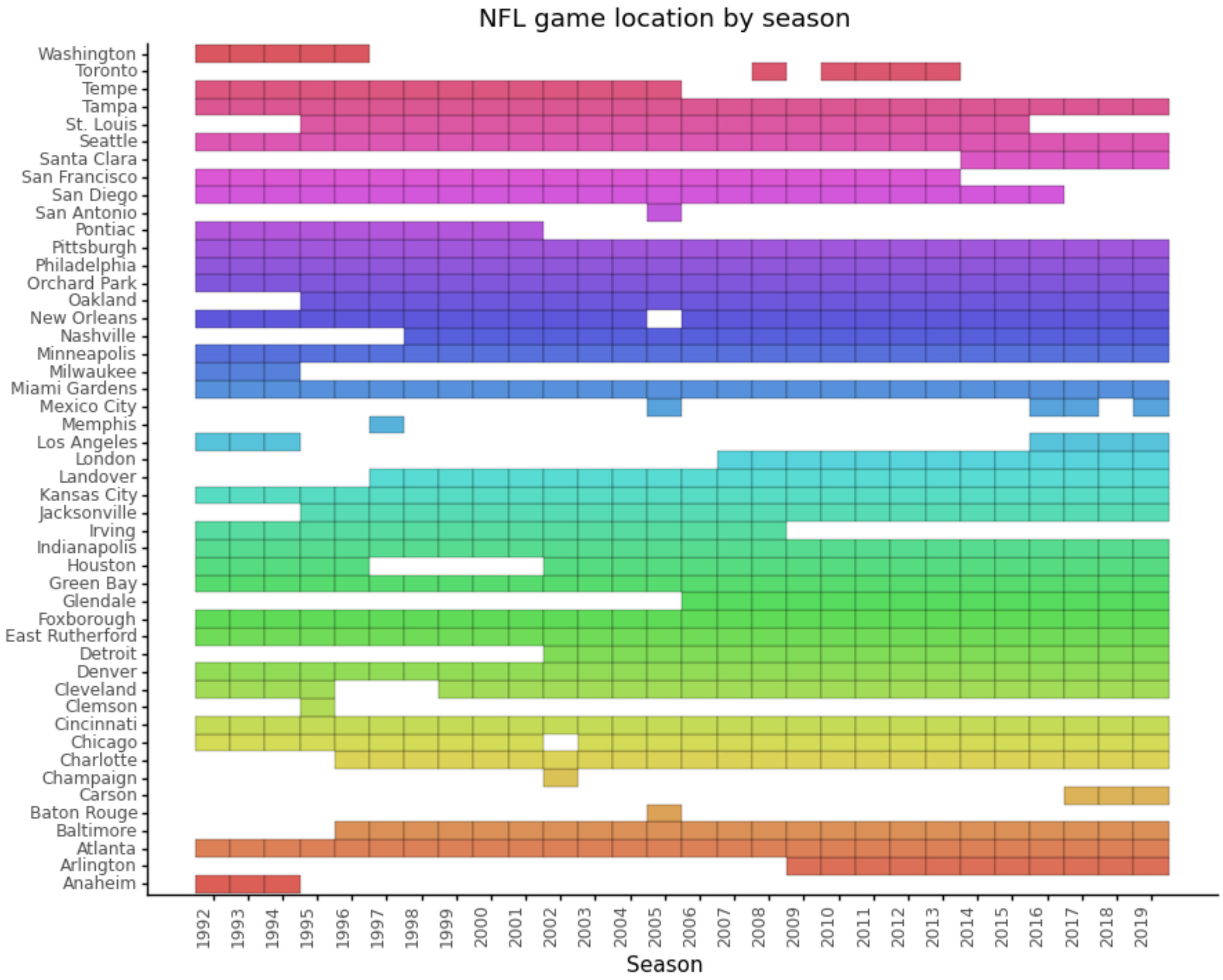


Figure 1: Cities where NFL games were held between the 1992 and 2019 seasons

Outliers

When examining the training data set, I noticed a few outliers in the set of home field advantage feature variables. Figure 2 shows a scatter plot between the *Miles_traveled* *Time_rest_days* variables.

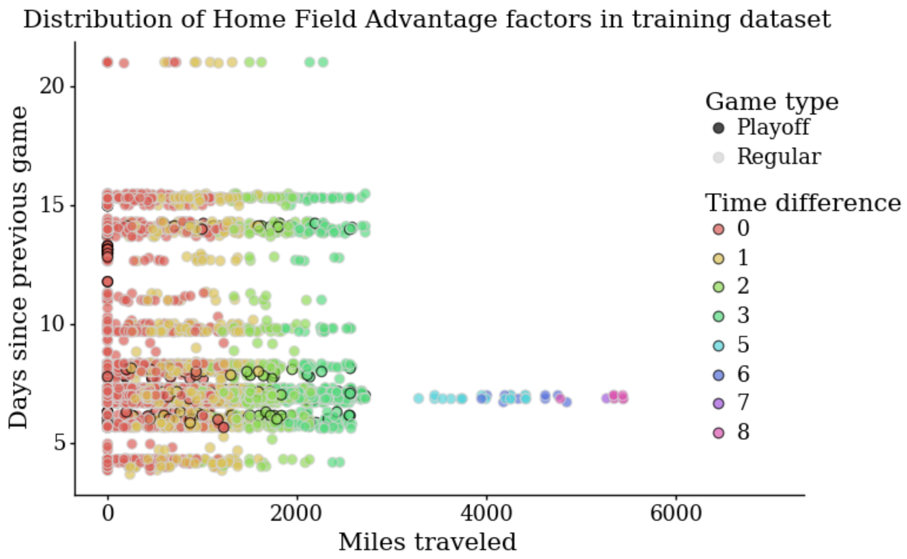


Figure 2: Distribution of Home Field Advantage factors in training dataset

The outliers in Miles traveled can be easily explained - every year since 2007, at least one game is held in London as part of the NFL International Series (Wikipedia contributors, 2020f). International travels also explains the outliers in time zone differences.

The outliers in days between games is a bit more puzzling. Initially, I had assume that these were from top seeds in the playoff who got to skip a first-round bye, but I was wrong. After examining the training data, I learned that the outliers came from the synthetic way I was filling in the missingness for the *Time_rest_days* variable described earlier in section . All the outliers came from 2001. Due to the one-week postponement in week 2 after the 9/11 attack, the longest "bye" a team had that season was 3-weeks long. As I apply the season's longest bye week to every team's time rest for week 1, this outliers multiplied. This is a flaw in the data that I choose to accept.

Rating a team

I use PFR's Offensive Simple Rating System (OSRS) and Defensive Simple Rating System (DSRS) to rate a team's overall quality in the season. Figure 3 shows the distribution of the variables in the training data. These variables measure a team strengths' relative to the average team in the regular season and centered around 0. PFR does not disclose how the

OSRS and DSRS are calculated or normalized - only the overall SRS (Drinen, 2007), so I felt a bit uneasy using them in the model. That being said, they seem to be reasonably normally distributed and would suffice as proxy for a team's strengths and weaknesses.

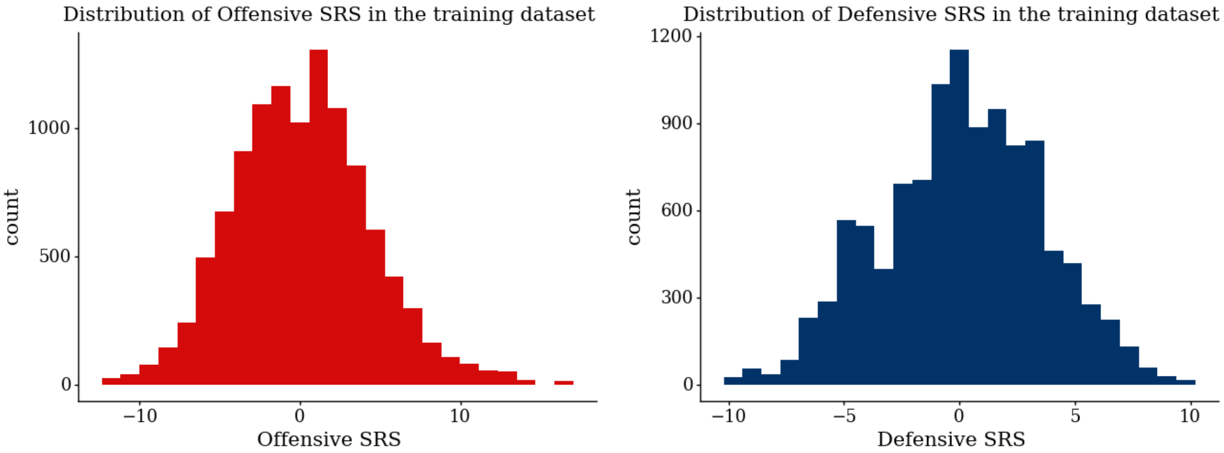


Figure 3: Distribution of Offensive and Defensive SRS in the training data

3.7 Wrangling steps

Scraping and cleaning the scraped data

The most challenging part of data wrangling was to assign the correct stadium location to each of the 7292 games in the dataset. I first downloaded the HTML of the Wikipedia page with information on each time's stadium (Wikipedia contributors, 2020c) using package requests (Reitz, 2011) in Python, then used the package Beautiful Soup to clean up the parsed HTML to get the individual link for each stadium's URL contained in the page. With this list of URLs, I then built scrapers to retrieve the GPS coordinates for each stadium, as well as the seating capacity over the years (if the information exists). I then created a facet plot of every stadium's capacity overtime to check for missingness or strange spikes that the scraper might have missed[7], then fixed them accordingly.

At the same time, I scraped data for each season's schedule, game and attendance record from PFR using python package pandas, then concatenated the yearly tables into a long

[7] For example, the Seating Capacity table for the Green Bay Packers' Lambeau Field (Wikipedia contributors, 2020d) does not have a datapoint for 2011 - therefore I filled it with capacity from the previous season

panel dataframe. The PFR data was quite clean to begin with, thus minimal effort was needed to further clean this data.

Afterwards, I merged the stadium data to each home team in the PFR dataset for every season, taking note of special circumstances where teams had to play at a "home" stadium that's not their main stadium - either due to participation in planned events[8], mid-season stadium moves[9] or due to other unforeseen circumstances[10]. I manually fixed this part of the merged data, using raw data that I had already scraped.

Additionally, since the data on PFR is only for regular season games, I scraped game records, schedule, location and attendance data for playoff games from Wikipedia, using techniques similar to ones described above. Finally, I merged all the data together. This dataset is set up in a wide format where each row contains information for a unique game, along with both teams' scores, ratings, as well as their respective home stadium.

Setting up the dyad

After merging the cleaned into a wide panel, I moved onto transforming it into a dyadic dataset. I created this dataset by making a mirrored copy of the wide data described above, switching the columns with information for one team with the other. Finally, I concatenated the original and mirrored data set together, and dropped the excessive columns and renamed the remaining columns accordingly.

Finally, I calculated the *Miles_traveled* variable using the Great-Circle distance between the GPS coordinates of Team_A's Home stadium and the stadium the game is being played at with methods described by Houston (2019), as well as the time rest each time has between 2 consecutive games on the schedule using pandas' timedelta function. The final dyadic dataset is set up so that 2 consecutive rows make up a dyad representing one game. The team-specific information in each vector (such as miles traveled, time rest, season overall

[8] NFL International Series in the UK and Mexico (Wikipedia contributors, 2020f), Bills Toronto Series in (Wikipedia contributors, 2020b)

[9] For example, in 2005 the St. Louis Rams had 2 Home stadiums - Busch Memorial Stadium for the first 4 home games, and Trans World Dome for the remainder of the season (Wikipedia contributors, 2020a).

[10] For example, the New Orleans Saints played the entire 2005 season in alternate "home" stadiums because the Superdome suffered extensive damage due to Hurricane Katrina (Wikipedia contributors, 2020b).

offensive rating, etc.) contains information for Team_A in that row. This dyadic dataset is the data on which I train my models.

Finally, I split it into a training and a test dataset, with the test dataset containing 25% of the rows in the full dataset.

4 Analysis

After setting up the dyad, I ran the data through a pipeline with 4 classifiers: K-nearest neighbor, Gaussian Naive-Bayesian, Decision Tree, and Random Forest, and set tuning parameters for them. Aside from the built-in MinMax scaler using sklearn API (Buitinck et al., 2013), I did not do further pre-processing of my data, as I think the outliers in the feature variables such as miles traveled are important in understanding the away team's disadvantage.

The model that "won out" in the pipeline is a Random Forest model with a max depth of 7, max features of 7, and 500 estimators. When using this model to predict whether it made correct predictions in the training data itself, it received an Area Under Curve (AUC) score of 0.779 and Accuracy score of 0.707.

The AUC score of the model on the test dataset is 0.705 and the Accuracy score is 0.645. Figure 4 shows the Receiving Operating Characteristic Curve illustrating the model's false positive rate and true positive rate. As a coin flip would have an AUC score of 0.5, so the black-box model does a little better than random, but not impressively so. However, the information we learn from which variable is more important in making the prediction is useful in our understanding of home field advantage, which I will discuss in more details in the next section.

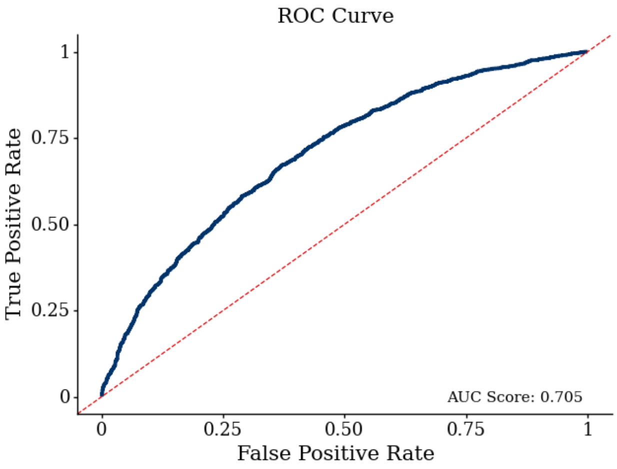


Figure 4: Receiver Operating Characteristic Curve

5 Results

Figure 5 ranks the importance of the feature variables in the Random Forest model. For obvious reasons, a team's ORSR and DSRS are the 2 most important variables in whether it wins a game or not. More interesting is the next 2 important variables: *Miles_traveled* and *Attendance_pct*.

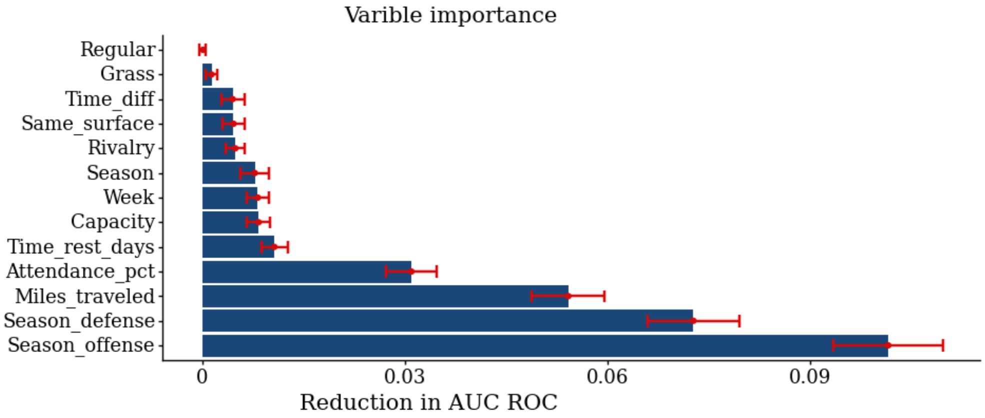


Figure 5: Variable importance

The Partial dependency plots in Figure 6 shows an obvious decrease in probability of winning between a team that had to travel versus one that didn't. Attendance percentage in the stadium seems to play a role in predicting a team's Win as well. Interestingly, the number of days a team gets to rest did not.

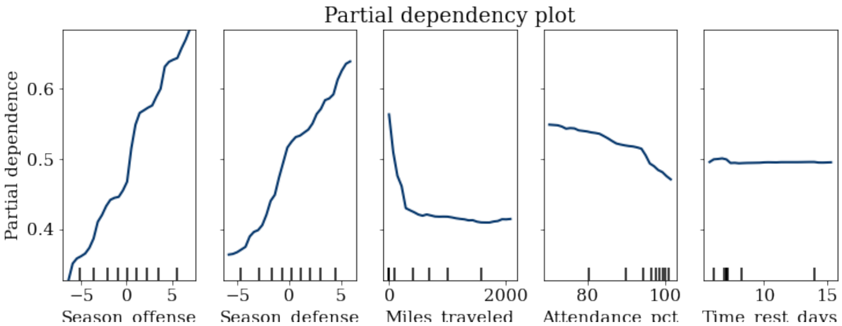
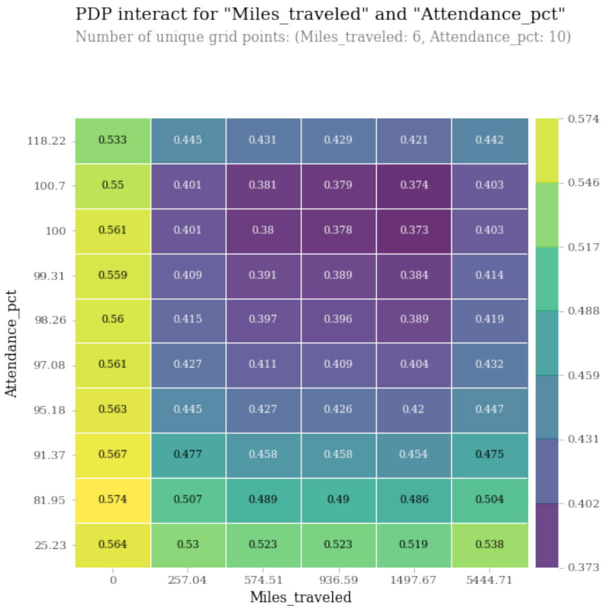


Figure 6: Partial Dependency Plot (PDP)

Figure 7 shows the interaction between *Miles_traveled* and *Attendance_pct* in predicting the outcome. The model predicts that a home team who didn't have to travel (thus *Miles_traveled* = 0) with a packed stadium will win more than teams that had to travel. Interestingly, international travels (the last column in the grid) don't seem to present much of a disadvantage, as both teams had to travel, but domestic long-distance travels do.



	0	257.04	574.51	936.59	1497.67	5444.71
118.22	0.533	0.445	0.431	0.429	0.421	0.442
100.7	0.55	0.401	0.381	0.379	0.374	0.403
100	0.561	0.401	0.38	0.378	0.373	0.403
99.31	0.559	0.409	0.391	0.389	0.384	0.414
98.26	0.56	0.415	0.397	0.396	0.389	0.419
97.08	0.561	0.427	0.411	0.409	0.404	0.432
95.18	0.563	0.445	0.427	0.426	0.42	0.447
91.37	0.567	0.477	0.458	0.458	0.454	0.475
81.95	0.574	0.507	0.489	0.49	0.486	0.504
25.23	0.564	0.53	0.523	0.523	0.519	0.538

Figure 7: PDP interact for Miles traveled and Attendance percentage

Figure 8 shows that distance travel does have a negative impact on a team's win probability. There does seem to be heterogeneity in this - for obvious reasons, because a team's defense and offense matter more to the outcome than distance traveled.

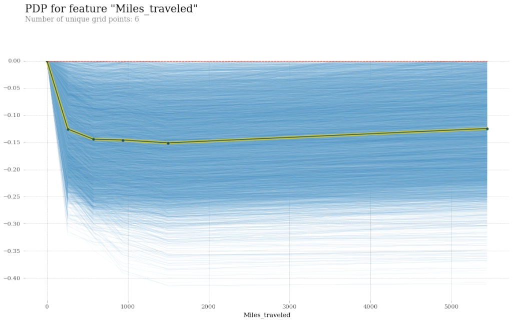


Figure 8: Ice plot for Mile traveled

Because the best predictor came from a Random Forest model - a black-box model that's not easily interpretable, I created a global surrogate model to interpret the result. The decision tree surrogate in Figure 9 has an R^2 of 0.77, which means that the decision tree model explains 77% of the behavior of the Random Forest model. This decision tree helps us understand some of the splittings that went into the black-box model in predicting a team's win.

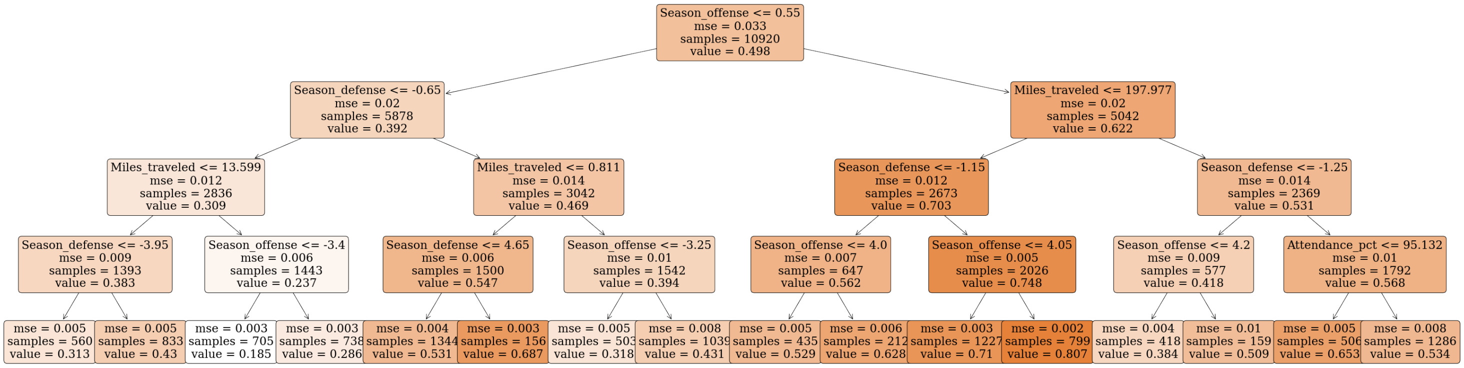


Figure 9: Global Surrogate Model

Out of all the feature variables, distance traveled seem to be the most likely explanation for a team's home field advantage in my model. Home teams win because they did not have to travel, and the more domestic travel a team has to do for a game, the less likely they'll win.

6 Further discussion

Overall, I consider the project a moderate success. Data wrangling took the bulk of the time I spent in this project, but I learned a lot from it. I am happy that I seem to find some indication that distance travel matters more to home field advantage than crowd noise, rest time and timezone difference.

If given more time, I would have liked to tinker more with the pre-processing steps to try to get a better accuracy score. I would also be interested to learn how the model would on games in the 2020 season - as fans attendance doesn't seem to matter as much as distance traveled (those poor Seahawks!)

References

Breech, J. (2020). *2020 NFL schedule: Ravens get huge travel-related advantage, Seahawks will fly the most miles.* https://www.cbssports.com/nfl/news/2020-nfl-schedule-ravens-get-huge-travel-related-advantage-seahawks-will-fly-the-most-miles/.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Drinen, D. (2007). *A very simple ranking system.* https://www.pro-football-reference.com/blog/index4837.html?p=37.

Greenberg, N. (2020). *There?s no such thing as home-field advantage in the NFL this season.* https://www.washingtonpost.com/sports/2020/11/11/home-field-advantage-is-lie/.

Hermsmeyer, J. (2018). *The NFL?s Home-Field Advantage Is Real. But Why?* https://fivethirtyeight.com/features/the-nfls-home-field-advantage-is-real-but-why/.

Houston, P. (2019). *Calculate distance of two locations on Earth using Python.* https://medium.com/@petehouston/calculate-distance-of-two-locations-on-earth-using-python-1501b1944d97.

Jamieson, J. P. (2010). The home field advantage in athletics: A meta-analysis. *Journal of Applied Social Psychology*, 40(7):1819–1848. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.2010.00641.x.

Moskowitz, T. and Wertheim, L. (2012). *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won.* Three Rivers Press.

NFL Football Operations (2020a). *Creating the NFL Schedule.* https://operations.nfl.com/the-game/creating-the-nfl-schedule/.

NFL Football Operations (2020b). *Pre-Snap Fouls.* https://operations.nfl.com/the-rules/nfl-video-rulebook/pre-snap-fouls/#article-2.-false-start.

Reitz, K. (2011). *Requests: HTTP for Humans.* https://requests.readthedocs.io/en/master/.

Richardson, L. (2007). *Beautiful Soup Documentation.* https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

Wikipedia contributors (2020a). *1995 St. Louis Rams season.* https://en.wikipedia.org/wiki/1995_St._Louis_Rams_season, accessed December 20, 2020.

Wikipedia contributors (2020b). *Bills Toronto Series.* https://en.wikipedia.org/wiki/Bills_Toronto_Series, accessed December 20, 2020.

Wikipedia contributors (2020c). *Chronology of home stadiums for current National Football League teams.* https://en.wikipedia.org/wiki/Chronology_of_home_stadiums_for_current_National_Football_League_teams, accessed December 20, 2020.

Wikipedia contributors (2020d). *Lambeau Field.* https://en.wikipedia.org/wiki/Lambeau_Field#Seating_capacity, accessed December 20, 2020.

Wikipedia contributors (2020e). *List of NFL tied games*. https://en.wikipedia.org/wiki/List_of_NFL_tied_games, accessed December 20, 2020.

Wikipedia contributors (2020f). *NFL International Series*. https://en.wikipedia.org/wiki/NFL_International_Series, accessed December 20, 2020.

Wikipedia contributors (2020g). *NFL regular season*. https://en.wikipedia.org/wiki/NFL_regular_season, accessed December 20, 2020.