Beyond the Court: Advanced Analytics in Predicting NBA Performance and Home Field Impact

Krish Rambhiya

Data Overview

The data used for this project was uploaded on Kaggle by a user that compiled it directly from the official NBA statistics. Given that these statistics come directly from the NBA, which tracks the official statistics for every player in the league, we know that our data is a census. Every single player is represented in this dataset. Every NBA player is aware the league collects and publicly posts these stats and have agreed to this in order to play in the NBA so there are no issues with consent. Also, the data used came from October 2003 to December 2022. We did not need to significantly clean this data because it came directly from the NBA.

The data we used consists of a few different tables that are connected by unique game, team, and player IDs that they share in common, allowing these tables to be easily combined.

The first table is the games table, which consists of information on every game played in the NBA in the selected years. We have included a screenshot of the data to illustrate it. As can be seen in the figure, each row in the table is a game and includes information about the game such as which teams played, the result, the season, a game ID, the home team, and some statistics about the teams' performances. There were no columns with a large number of null values.

	GAME_DATE_EST ...	GAME_ID int64	GAME_STATUS_T...	HOME_TEAM_ID in...	VISITOR_TEAM_ID ...	SEASON int64
0	2022-12-22	22200477	Final	1610612740	1610612759	2022
1	2022-12-22	22200478	Final	1610612762	1610612764	2022
2	2022-12-21	22200466	Final	1610612739	1610612749	2022
3	2022-12-21	22200467	Final	1610612755	1610612765	2022
4	2022-12-21	22200468	Final	1610612737	1610612741	2022

The next table we used is the game details table, which consists of more details about each player's performance in each game. It also includes a game ID that matches to the game ID in the games table so that a player's specific performance can be linked to other details of the game via a join. Each row in this table is a player. The table contains information about the player such as their team, name, position, time player, and various statistics on performance. We

have once again added a screenshot to help illustrate it. In this table, we had some columns with a lot of null values which were initially concerning. However, these columns (nickname, position, comment) were ultimately not relevant to our research questions so they were not a point of concern. The only column with a significant number of null values that was relevant to our questions was the minutes played column. However, upon further examination these nulls were simply players who did not play.

	GAME_ID int64	TEAM_ID int64	TEAM_ABBREVIA...	TEAM_CITY object	PLAYER_ID int64	PLAYER_NAME obj...
0	22200477	1610612759	SAS	San Antonio	1629641	Romeo Langford
1	22200477	1610612759	SAS	San Antonio	1631110	Jeremy Sochan
2	22200477	1610612759	SAS	San Antonio	1627751	Jakob Poeltl
3	22200477	1610612759	SAS	San Antonio	1630170	Devin Vassell
4	22200477	1610612759	SAS	San Antonio	1630200	Tre Jones
5	22200477	1610612759	SAS	San Antonio	1628380	Zach Collins
6	22200477	1610612759	SAS	San Antonio	203926	Doug McDermott
7	22200477	1610612759	SAS	San Antonio	1626196	Josh Richardson
8	22200477	1610612759	SAS	San Antonio	1631103	Malaki Branham

The next table is the players table which contains some basic information on the player. Each row is a player. The table contains the team ID and player ID of the player along with season (this is included as the team ID of players change over time). Part of the table is pictured below. There was no column with a significant number of null values.

	PLAYER_NAME obj...	TEAM_ID int64	PLAYER_ID int64	SEASON int64
0	Royce O'Neale	1610612762	1626220	2019
1	Bojan Bogdanovic	1610612762	202711	2019
2	Rudy Gobert	1610612762	203497	2019
3	Donovan Mitchell	1610612762	1628378	2019
4	Mike Conley	1610612762	201144	2019

The next table is the ranking table, which contains information about team rankings. Each row is a team on a given day and it contains information about the team's record and ranking. There was one column with a very high percentage of nulls (RETURNTOPLAY) but it was not relevant to our research questions. The table is shown below.

	TEAM_ID int64	LEAGUE_ID int64	SEASON_ID int64	STANDINGSDATE o...	CONFERENCE obje...	TEAM object
0	1610612743	0	22022	2022-12-22	West	Denver
1	1610612763	0	22022	2022-12-22	West	Memphis
2	1610612740	0	22022	2022-12-22	West	New Orleans
3	1610612756	0	22022	2022-12-22	West	Phoenix
4	1610612746	0	22022	2022-12-22	West	LA Clippers

The final table is the teams table. Each row is a team and the table contains information like city, arena, year founded, owner, and arena capacity. The arena capacity had a few null values, but there were few enough that it was manageable. The table is shown below.

	LEAGUE_ID int64	TEAM_ID int64	MIN_YEAR int64	MAX_YEAR int64	ABBREVIATION obj...	NICKNAME object
0	0	1610612737	1949	2019	ATL	Hawks
1	0	1610612738	1946	2019	BOS	Celtics
2	0	1610612740	2002	2019	NOP	Pelicans
3	0	1610612741	1966	2019	CHI	Bulls
4	0	1610612742	1980	2019	DAL	Mavericks

We know that our data was not modified for differential privacy as the data is all public including the names of each player. We also do not suspect that selection bias or convenience sampling are issues given the data is a census. We know measurement error is not an issue either as these statistics are all tracked and stored by the NBA who validates the results of each game.

There are a few things not in the data that we believed could have helped our analysis, especially in research question 2 where we tried to examine the impact of home team advantage. Knowing the climate/altitude of each stadium and the energy/number of fans could help give us a better idea on the factors that play into home team advantage as not every team's home team advantage will be equal as measured by these factors. For example, the Denver Nuggets could have an advantage due to the high altitude of their stadium. Knowing the amount of rest each player had before the games would also be helpful as that could seriously affect performance.

Research Questions

Research Question 1

Initial Research Question: Based on a player's in-game performance metrics and team context in the first half of the NBA season, can we predict the likelihood of them achieving specific performance benchmarks in the second half?

Updated Research Question: How can a decade of first-half performance data in the NBA be analyzed to forecast a team's probability of making the playoffs for the current season?

Research Question 2

Initial Research Question: Is there a causal link between having the home team advantage and the performance of that team?

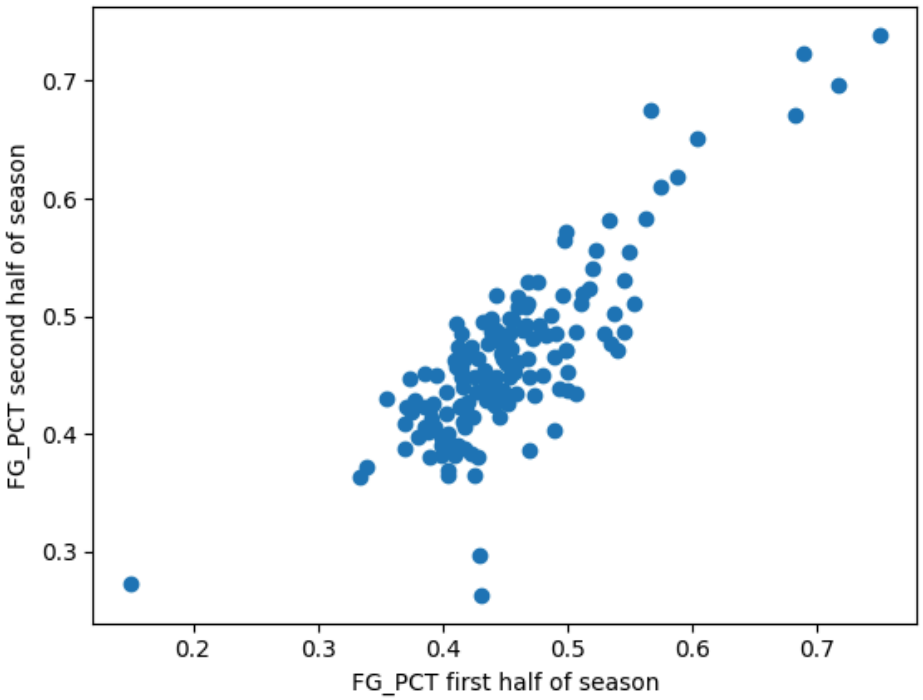
Updated Research Question: Is there a causal link between playing on the home field and number of points scored by the team?

EDA

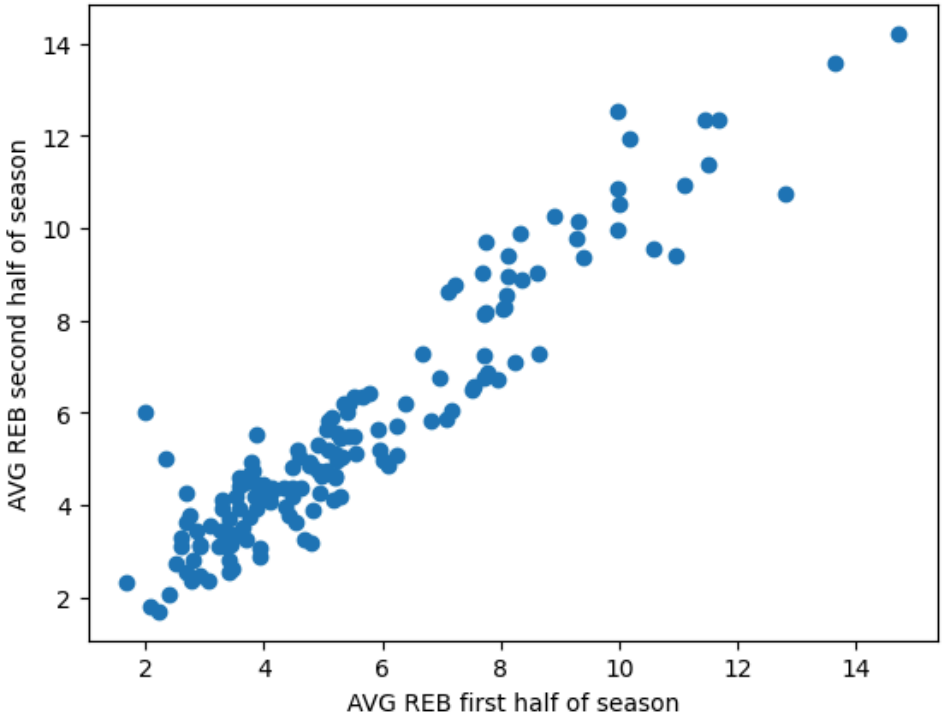
Research Question 1

We begin by duplicating two dataframes, then selectively remove columns not needed. Next, we convert 'GAME_DATE_EST' to datetime, allowing us to divide the data into season halves. We merge game details with general game data for a comprehensive view. The dataset is then split into training and test sets based on dates. We implement a function to change the 'MIN' column from minutes to seconds, ensuring accuracy. After cleaning and discarding NaN values, we group the data by player names, calculating average game statistics. This methodical approach prepares the data for in-depth analysis of player performance.

Next, We calculate the average field goal percentage (FG_PCT) for NBA players in the first half of the season using the training set, and then for the second half using the test set. We display these averages to note any differences. Then, we merge the two datasets on 'PLAYER_ID' and create scatter plots. These plots visually compare FG_PCT and average rebounds (REB) between the first and second halves of the season for each player, offering us insights into performance trends over the season.



Quantitative Visualization #1



Quantitative Visualization #2

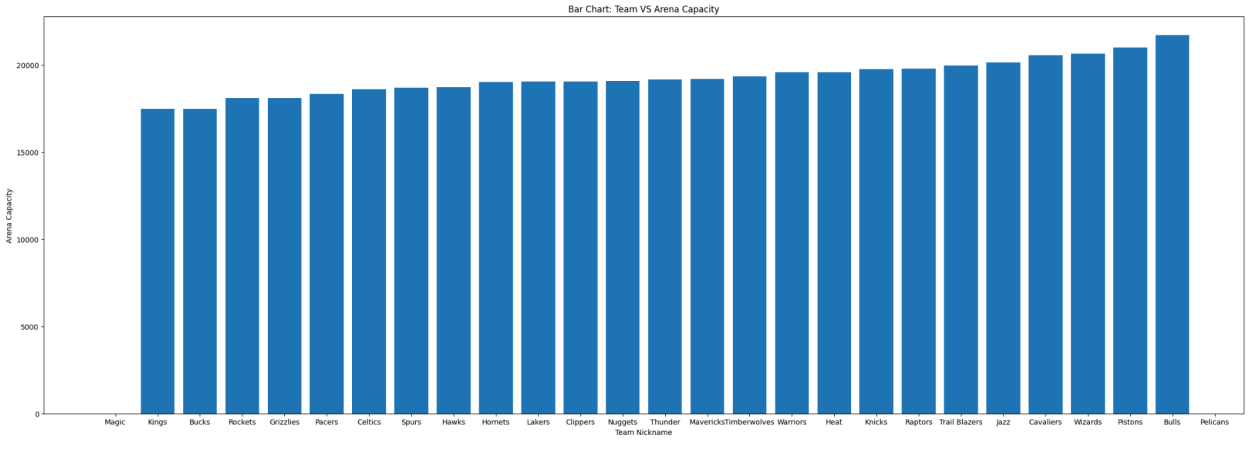
We see a linear relationship here similar to the relationship for FG_PCT. While we mainly intend to use FG percentage as our indicator of performance, other performance metrics might also be

valuable in our model so this graph is helping us explore rebounds as a possible indicator of performance. This connects to our first research question where we are predicting performance in the second half of season based on the first half. This graph indicates we could also use rebounds as part of our performance definition and a linear model might work well for this as well.

Research Question 2

Altered research question based on project proposal so now we are looking at the possible effects of "home team advantage" on points scored by the team. We will use causal inference with the treatment being playing at home and the outcome being the points scored by the team for that game (or player potentially).

Potential confounders may change depending upon how we define the "home team advantage". If we simply define home advantage as playing on the team's home field, the only variables affecting this would be the scheduling of games. However, if we define home team advantage as including factors like the comfort of not having to travel to an away game, some confounders may arise, including the resilience of a team regarding the court they play on (for example, a team that performs similarly no matter what field they play on may affect the presence of home advantage as well as the free throw percentage.)



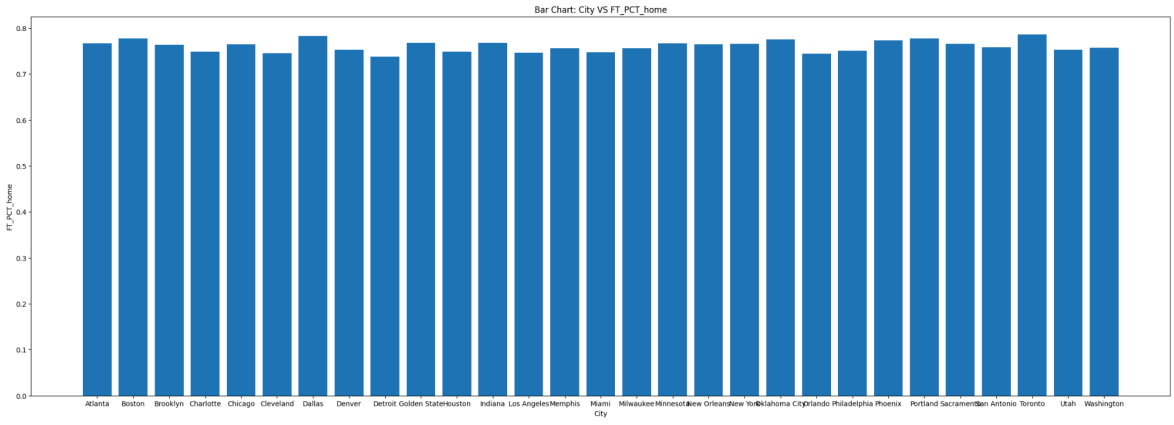
Categorical Visualization #1

Trend Observed: The bar chart presents a clear variability in arena capacities among different basketball teams, without indicating a uniform trend across the teams. Some of the teams have significantly larger arena capacities, which could suggest a stronger home crowd presence.

Relevance to Research Question: In the context of exploring the "home team advantage," the differences in arena capacities could be crucial. Larger arenas might amplify the home crowd's impact on the players' performance, potentially influencing free throw percentages—a common measure of basketball efficiency. This visualization invites further investigation into whether a more substantial home crowd correlates with better home game performance, thereby motivating the question of how environmental factors contribute to home advantage. To establish relevance, subsequent analysis would need to integrate free throw percentages with arena capacity data while controlling for confounding variables.

Possible follow-ups

Since city population is not provided in the data we have been given, a possible follow-up on our second research question might include researching how city population and arena size may affect the home team advantage.



Categorical Visualization #2

We don't seem to observe any effect of the city on free throw percentage. There seems to be a roughly uniform distribution of free throw percentage among the different cities.

The bar chart depicting free throw percentages across different cities shows uniformity in values, suggesting that city size, often considered a potential confounding factor, may not significantly influence the "home team advantage" in terms of free throw success. The similarity in percentages across cities implies that factors other than city size are more relevant to the research question. This lack of trend across city sizes can steer the research to focus more on the intrinsic qualities of teams and players in relation to home game performance, rather than external factors like the size of the city they represent.

Research Question 1: Prediction with GLMs and nonparametric methods

Frequentist GLM

For all three of the following Frequentist models, we used only one parameter to predict the outcome. The reason for this is that we want to understand how a specific aspect of a player's performance (i.e. rebounds or FG percentage) in the first half of the season affects their performance in the second half. If we introduce multiple parameters, this relationship may be harder to interpret.

```
freq_model = sm.GLM(test['FG_PCT_y'], exog = sm.add_constant(test['FG_PCT_x']),
                    family=sm.families.Poisson())
freq_res = freq_model.fit()
print(freq_res.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                FG_PCT_y   No. Observations:                  155
Model:                            GLM   Df Residuals:                      153
Model Family:                 Poisson   Df Model:                            1
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -108.00
Date:                Mon, 11 Dec 2023   Deviance:                       0.54138
Time:                        03:13:15   Pearson chi2:                     0.523
No. Iterations:                     3   Pseudo R-squ. (CS):           0.007331
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.5620      0.740     -2.112      0.035      -3.012      -0.112
FG_PCT_x       1.7313      1.581      1.095      0.273      -1.367       4.829
==============================================================================
```

Frequentist Model #1

According to the frequentist GLM with Poisson Regression, with a p-value of 0.05, first half FG PCT is not a significant predictor of second half FG PCT. This is evident as for an increase of 1 in REB_x, Y changes by e^1.7313=5.64799152693 which is significant. Furthermore, the average log-likelihood -108/153 is close to 0 which is an indicator that it is a good model. This is our 95% confidence interval for FG PCT: [-1.367, 4.829] and our 95% confidence interval contains 0; hence, we can interpret that there is no statistical significance of our model.

```python
# Rebounds in first half to predict rebounds in second half
freq_model_2 = sm.GLM(test['REB_y'], exog = sm.add_constant(test['REB_x']),
                      family=sm.families.Poisson())
freq_res_2 = freq_model_2.fit()
print(freq_res_2.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  REB_y   No. Observations:                  155
Model:                            GLM   Df Residuals:                      153
Model Family:                 Poisson   Df Model:                            1
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -284.98
Date:                Mon, 11 Dec 2023   Deviance:                       26.921
Time:                        03:13:15   Pearson chi2:                     27.0
No. Iterations:                     4   Pseudo R-squ. (CS):             0.6219
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.8677      0.081     10.758      0.000       0.710       1.026
REB_x          0.1417      0.011     12.958      0.000       0.120       0.163
==============================================================================
```

Frequentist Model #2

We see that using Poisson regression, our GLM results show that rebounds in the first half are a great predictor of rebounds in the second half. This is evident as the for an increase of 1 in REB_x, Y changes by e^0.1417=1.1522309274 which is significant. Furthermore, the avg log-likelihood -284.98/153 is close to 0 which is an indicator that it is a good model. In this case, the 95% confidence interval is [0.120 0.163] which does not contain the null value (0), hence this model is statistically significant.

```python
# rebounds to predict FG_PCT?
freq_model_3 = sm.GLM(test['FG_PCT_y'], exog = sm.add_constant(test['REB_x']),
                      family=sm.families.Poisson())
freq_res_3 = freq_model_3.fit()
print(freq_res_3.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                FG_PCT_y   No. Observations:                  155
Model:                            GLM   Df Residuals:                      153
Model Family:                 Poisson   Df Model:                            1
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -108.23
Date:                Mon, 11 Dec 2023   Deviance:                       0.98737
Time:                        03:13:15   Pearson chi2:                     0.989
No. Iterations:                     3   Pseudo R-squ. (CS):           0.004470
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.9789      0.278     -3.516      0.000      -1.525      -0.433
REB_x          0.0368      0.043      0.847      0.397      -0.048       0.122
------------------------------------------------------------------------------
```

Frequentist Model #3

Rebounds in the first half of season is not a strong predictor of field goal percentage in the second half of season. This is evident as for an increase of 1 in REB_x, Y changes by e^0.0368=1.03748550299 which is not that significant. Furthermore, the avg log-likelihood -108.23/153 is close to 0 which is an indicator that it is a good model. The AIC for this model: 220.46 and BIC: 226.52. In this case, the 95% confidence interval is [-0.048 0.122] which contains the null value (0), hence this model is not statistically significant.

Neural Network (Non-Parametric)

We chose the neural network as our nonparametric method because we believe there could be a nonlinear relationship in the data and neural networks excel at dealing with this type of data. Also in the case it is linear and we are incorrect, it will still perform well.

The GLM results proved disappointing overall as two of the models were not even statistically significant (the coefficient had a confidence interval that included 0 so we could not conclude a relationship either way). Due to this, we decided to alter the research question a bit when

exploring the neural network. We altered our research question to try and use team data instead of individual player data as we felt that it makes more sense to look at the collective impact of teams data to predict second half performance rather than a single player. This also rules out player specific issues such as injuries and starting lineup which would make it more difficult to predict the outcome of a team's performance in the second half of the season.

Furthermore, we decided to use a decades worth of team data to create our model as that reduces variance and bias which can occur from using only first half data from the current season to predict the second half performance. Furthermore, first half team performance for the current season would be correlated to the second half performance of the current season which leads to a heavy bias. Hence, we developed the following neural network that uses the past decades data of first half performance to predict second half performance of the current season (the way we quantify this is whether or not a team qualifies for the playoffs).

```
#Neural Net
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score

display.dropna(inplace=True)

# Select relevant features for classification
features = display[['PTS_home', 'FG_PCT_home', 'FT_PCT_home', 'FG3_PCT_home', 'AST_home', 'REB_home',
                   'PTS_away', 'FG_PCT_away', 'FG3_PCT_away', 'AST_away', 'REB_away']]

# Standardize/normalize features
scaler = StandardScaler()
features_scaled = scaler.fit_transform(features)

# Train-Test Split
train_data = features_scaled[display['YEAR'] <= 2021]
test_data = features_scaled[display['YEAR'] == 2022]
labels_train = display[display['YEAR'] <= 2021]['MADE_PLAYOFFS']
labels_test = display[display['YEAR'] == 2022]['MADE_PLAYOFFS']

# Model Training - Neural Network
# adjust the parameters of the MLPClassifier as needed
#This section creates an instance of the MLPClassifier with a neural network architecture consisting
#a single hidden layer with 100 neurons. The model is trained on the training data (train_data) with
#corresponding labels (labels_train).
model = MLPClassifier(hidden_layer_sizes=(100,), max_iter=1000, random_state=42)
model.fit(train_data, labels_train)

# Prediction
labels_pred = model.predict(test_data)

# Calculate Accuracy
accuracy = accuracy_score(labels_test, labels_pred)
accuracy
```

```
/shared-libs/python3.9/py/lib/python3.9/site-packages/sklearn/neural_network/_multilayer_perceptron.py:702:
  warnings.warn(
```

```
0.6
```

Neural Network

With an accuracy of 60%, we can say that our neural network is performing pretty well since this accuracy is better than a naive classifier which we would expect to have 50% accuracy. Obviously there is room for improvement but we are happy with this result.

Model Performance and Future Datasets: The Neural Network model outperformed the Frequentist models because it could evaluate multiple parameters simultaneously, capturing complex relationships in the data. The Frequentist models, evaluating one parameter at a time, were less adept at handling multifaceted interactions. The Neural Network's accuracy of 0.6, exceeding the naive benchmark of 0.5, instills confidence in its applicability to future datasets. This performance, coupled with the potential for further refinement through parameter tuning, suggests a promising approach for future predictive analyses in similar contexts.

Model Limitations: Each model has inherent limitations. The Frequentist model might be constrained by its assumptions of linearity and normality, while the Neural Network might be limited by overfitting and the need for large datasets.

Additional Data to Improve Models: Player Health and Fitness Data: Injury history, minutes played, and physical fitness metrics could be crucial for predicting future performance, especially in the physically demanding NBA season. Psychological or Qualitative Factors: Player morale, team dynamics, or coaching style, though harder to quantify, might offer insights into performance trends.

Uncertainty in Results: Data Quality and Noise: Variability in data quality, including missing or erroneous values, can introduce uncertainty in model predictions. Model Complexity and Assumptions: Each model comes with inherent assumptions and complexities. For instance, a Neural Network might capture complex patterns but also risk overfitting, while the Frequentist model might be limited by its linear assumptions. External Variables: Factors not included in the model, such as team strategies, player injuries, or unmeasured psychological factors, can also contribute to the uncertainty in predictions.

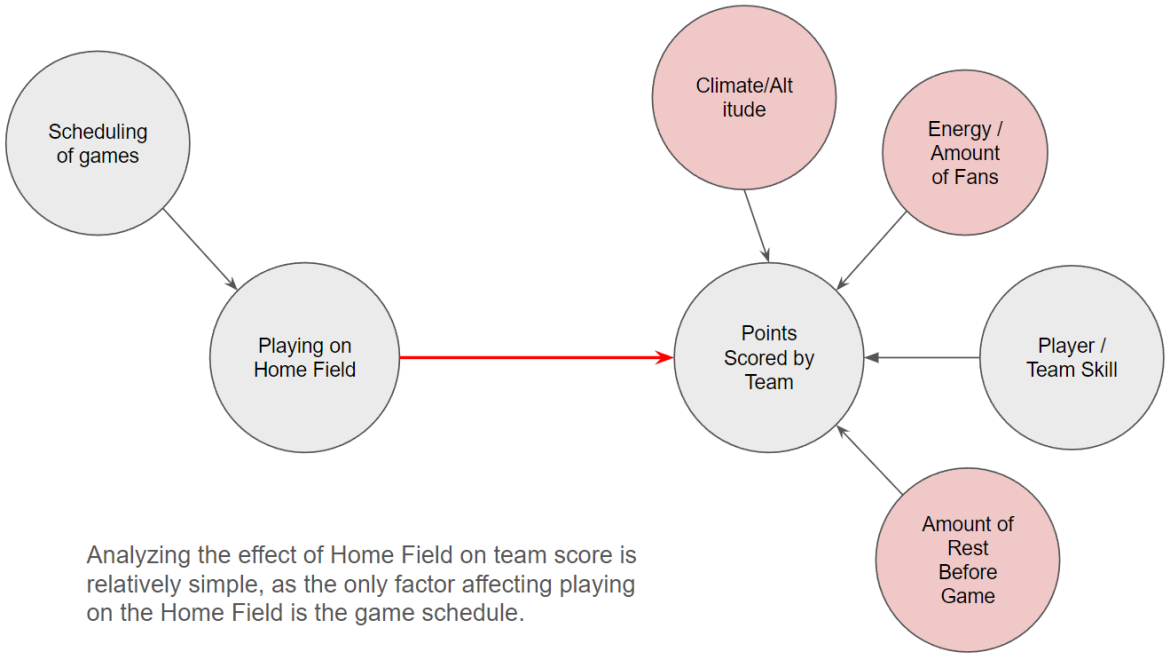
Research Question 2: Causal Inference

Causal Question: Does playing on the Home Field lead to an increase in the number of points scored by an NBA team, compared to playing on an away field?

The treatment variable is playing on home field and the outcome variable is the number of points scored by the home team. We begin by visualizing the causal relationship between our treatment (having the home field) and our result (points scored by the team).

Note that our "home field advantage" is essentially a stripped-down version of the "home team advantage", and we only include whether or not a team is playing on their home field in our definition of the treatment.

With this more simple treatment in mind, our causal inference graph looks like this:



The major assumptions for this analysis are that the only factor in having the home field is the scheduling of games, and that the scheduling of games does not also affect the points scored by the team for that game. As a result, there are no confounders in our causal relationship graph. This means that we can supposedly just use the SDO to estimate the effect of our selected treatment.

Applying the SDO formula to our data gives us this result:

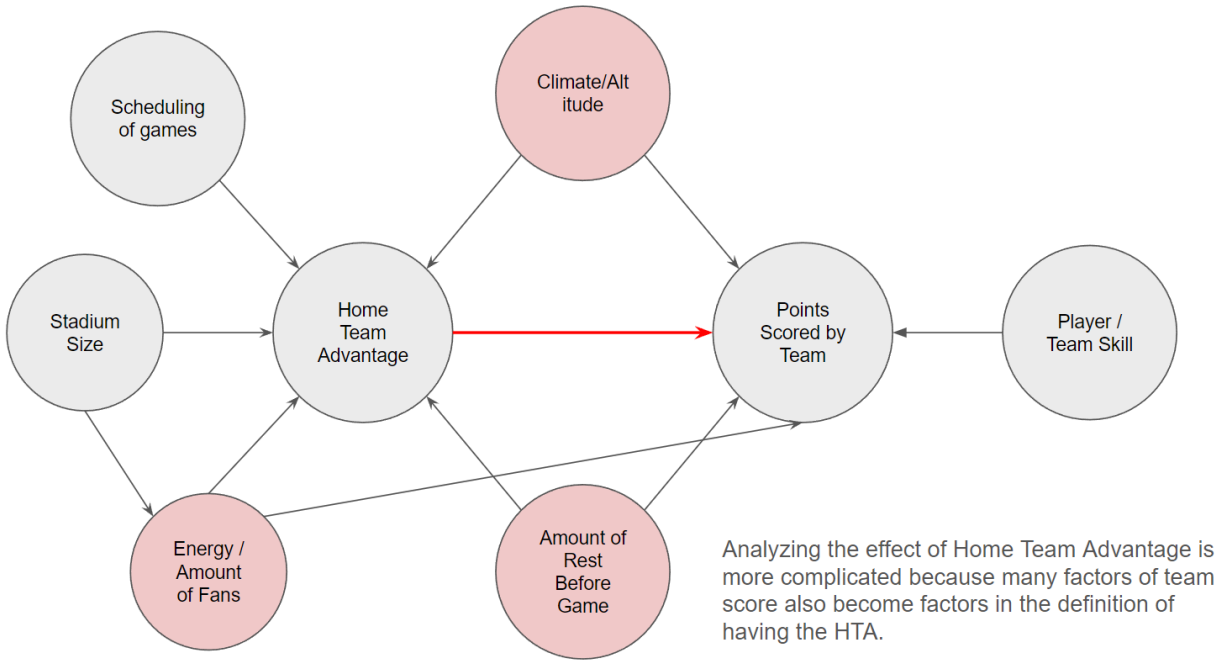
```
causal_effect_hf = team_pts_avg_home_and_away['PTS_home'].mean() - team_pts_avg_home_and_away['PTS_away'].mean()
causal_effect_hf
```

```
2.8156438873284486
```

Taking our assumptions into account, we found that playing at home increases the number of points scored by home teams by 2.8 points on average. This difference is practically significant in the NBA where games are closely contested and the margin of victories can be really narrow.

Results: We were able to find a small causal effect of playing at home on the average score of teams. Using SDO, we found that playing at home increases the number of points scored by home teams by 2.8 points on average.

Additional Data to Improve Models: A new research problem analyzing the effects of home team advantage and all of its factors on team score may follow a model similar to the following:



Data on stadium attendance, stadium size, player psychology, and travel fatigue would allow for a more comprehensive analysis of the effect of home team advantage.

Not all of these factors were available within our data; therefore, we decided to strip down the definition of our treatment to simplify the set of assumptions we had to make.

Uncertainty in Results: Most of the uncertainty in our results arises from outliers in the data, since we did not take any into account. With the knowledge that NBA teams all perform

relatively consistently, we hope that there are few enough outliers in the data such that they do not affect our results.

Conclusion

In our first research question, when we tried to predict the second half of the season's performance based on the first half, our frequentist GLM's performed poorly overall. The only GLM that performed decently was the one that predicted second half of season rebounds based on the first half of season, which seems to imply some kind of correlation between these statistics. However, we did not find this to be true when using FG percentage in first half of season to predict FG percentage in second half or when using rebounds in first half of season to predict FG percentage in second half of season. However, we responded by shifting our question a bit and incorporating more data when using a nonparametric method. We used team data instead of individual player data as we felt that it makes more sense to look at the collective impact of teams data to predict second half performance rather than a single player. The neural network performed decently well with a 60% accuracy (better than naive classifier). These results point to the potential of a neural network like this for future predictive tasks about team performance.

The Frequentist model can struggle with assumptions about data being linear and normal, whereas Neural Networks might overfit and require a lot of data. To make these models better, especially in predicting NBA player performances, adding data about players' health, like injury history and fitness levels, would be really helpful. Also, considering things like player morale and team relationships, though less measurable, could give a clearer picture of how well players might do. The scope of this research question is narrow as it is specific to the NBA and considers a decades worth of data for NBA with specific parameters which are all linked to basketball. This model could potentially only be generalizable to other basketball leagues globally.

In our second research question, we revealed a causal relationship between playing on the home field and points scored by the home team. Our unconfounded causal inference model

revealed through SDO that having the home field increased the average points scored during home games by 2.8 points on average. While knowing the effect of home field is helpful, the home team advantage is a much more complex phenomenon with many factors, some of which may be confounders. Our model's main limitation is due to the assumption that home field is the only factor influencing the outcome and not any other factor like time zone difference, player fatigue, etc. Our simplified model might overlook some aspects of home team advantage as a whole, so we cannot draw strong conclusions on a relationship between home team advantage and team performance. We believe it would be helpful to include more variables in a comprehensive investigation of home team advantage in the future.

The effect of home team advantage has been observed many times in sports history, so it is critical that sports associations like the NBA take home field and home team advantage into account when scheduling games for future seasons so as to encourage fair play for all teams. One way to possibly do this is to ensure that the season is scheduled in a way that teams in the league play the same number of games at home and away as other teams.