Project KOBE: Research Question

Analysts commonly agree on a phenomenon known as the home-court advantage. It is thought that the psychological impact of playing in front of one's own team gives players of the home team a significant advantage—largely because of the comfort associated with one's own home-town.

The first part of our question is: do NBA teams perform better at home than away?
We'll be answering this question using a hypothesis test, repeated for each of the 30 NBA teams. While the distribution we are using will remain the same (we'll talk more about it in sections below), the observed statistic for each team will vary.

At a higher level, performance will be evaluated based on wins. Specifically, the number of games at home that result in wins will be compared to the number of games away that result in wins (percent of home games that are wins vs percent of away games that are wins).

The second focus of our research will be to utilize various derived metrics in order to predict the metric that captures the proportion of wins at home minus the proportion of wins away.

These utilized metrics will be derived from the EDA process below, with some examples potentially including: field goals made, turnovers, rebounds, and assists per game (amongst others).

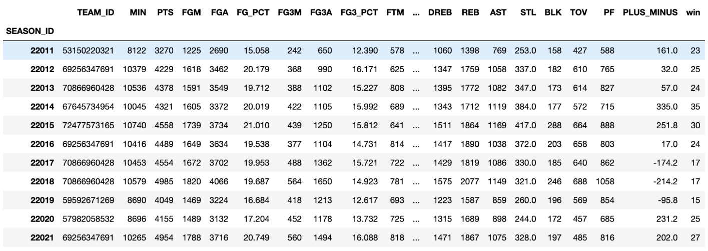
The metric we'll be trying to predict will look like this:

$$\frac{\text{number of wins at home}}{\text{number of home games}} - \frac{\text{number of wins away}}{\text{number of away games}}$$

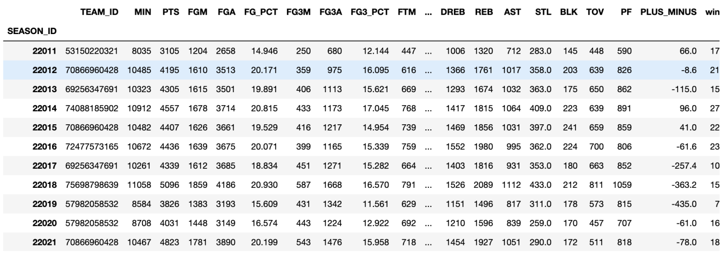
The goal here is to predict this metric using two different models—a GLM and a random forest. The GLM we'll be using is a multi-linear regression, predicting the proportion difference between wins at home and wins away (a number between zero and one).

The Data Set

Throughout this analysis, the same dataset will be used. Through an NBA DATA API a set of two data frames per NBA team were manufactured. The first holds season stats pertaining to **home** games that a particular team played in. The second holds season stats pertaining to **away** games that a particular team played in. Below is an example of how each table looks like for a specific team—we'll be using the Atlanta Hawks for this example.



SEASON_ID	TEAM_ID	MIN	PTS	FGM	FGA	FG_PCT	FG3M	FG3A	FG3_PCT	FTM	...	DREB	REB	AST	STL	BLK	TOV	PF	PLUS_MINUS	win
22011	53150220321	8122	3270	1225	2690	15.058	242	650	12.390	578	...	1060	1398	769	253.0	158	427	588	161.0	23
22012	69256347691	10379	4229	1618	3462	20.179	368	990	16.171	625	...	1347	1759	1058	337.0	182	610	765	32.0	25
22013	70866960428	10536	4378	1591	3549	19.712	388	1102	15.227	808	...	1395	1772	1082	347.0	173	614	827	57.0	24
22014	67645734954	10045	4321	1605	3372	20.019	422	1105	15.992	689	...	1343	1712	1119	384.0	177	572	715	335.0	35
22015	72477573165	10740	4558	1739	3734	21.010	439	1250	15.812	641	...	1511	1864	1169	417.0	288	664	888	251.8	30
22016	69256347691	10416	4489	1649	3634	19.538	377	1104	14.731	814	...	1417	1890	1038	372.0	203	658	803	17.0	24
22017	70866960428	10453	4554	1672	3702	19.953	488	1362	15.721	722	...	1429	1819	1086	330.0	185	640	862	-174.2	17
22018	70866960428	10579	4985	1820	4066	19.687	564	1650	14.923	781	...	1575	2077	1149	321.0	246	688	1058	-214.2	17
22019	59592671269	8690	4049	1469	3224	16.684	418	1213	12.617	693	...	1223	1587	859	260.0	196	569	854	-95.8	15
22020	57982058532	8696	4155	1489	3132	17.204	452	1178	13.732	725	...	1315	1689	898	244.0	172	457	685	231.2	25
22021	69256347691	10265	4954	1788	3716	20.749	560	1494	16.088	818	...	1471	1867	1075	328.0	197	485	816	202.0	27



SEASON_ID	TEAM_ID	MIN	PTS	FGM	FGA	FG_PCT	FG3M	FG3A	FG3_PCT	FTM	...	DREB	REB	AST	STL	BLK	TOV	PF	PLUS_MINUS	win
22011	53150220321	8035	3105	1204	2658	14.946	250	680	12.144	447	...	1006	1320	712	283.0	145	448	590	66.0	17
22012	70866960428	10485	4195	1610	3513	20.171	359	975	16.095	616	...	1366	1761	1017	358.0	203	639	826	-8.6	21
22013	69256347691	10323	4305	1615	3501	19.891	406	1113	15.621	669	...	1293	1674	1032	363.0	175	650	862	-115.0	15
22014	74088185902	10912	4557	1678	3714	20.815	433	1173	17.045	768	...	1417	1815	1064	409.0	223	639	891	96.0	27
22015	70866960428	10482	4407	1626	3661	19.529	416	1217	14.954	739	...	1469	1856	1031	397.0	241	659	859	41.0	22
22016	72477573165	10672	4436	1639	3675	20.071	399	1165	15.339	759	...	1552	1980	995	362.0	224	700	806	-61.6	23
22017	69256347691	10261	4339	1612	3685	18.834	451	1271	15.282	664	...	1403	1816	931	353.0	180	663	852	-257.4	10
22018	75698798639	11058	5096	1859	4186	20.930	587	1668	16.570	791	...	1526	2089	1112	433.0	212	811	1059	-363.2	15
22019	57982058532	8584	3826	1383	3193	15.609	431	1342	11.561	629	...	1151	1496	817	311.0	178	573	815	-435.0	7
22020	57982058532	8708	4031	1448	3149	16.574	443	1224	12.922	692	...	1210	1596	839	259.0	170	457	707	-61.0	16
22021	70866960428	10467	4823	1781	3890	20.199	543	1476	15.958	718	...	1454	1927	1051	290.0	172	511	818	-78.0	18

Home Games Away Games

Note because of the ever changing NBA, we decided to use data over the last 11 seasons. This number was selected based on our domain knowledge pertaining to the NBA – the past 11 seasons accurately represents one “era” of basketball that will allow us to compare apples to apples. Additionally, this number affords us enough data points in order to comfortably test our hypothesis against a large distribution. Since we’re going to be bootstrapping from this data-set, the size of the population (our data set) will be important.

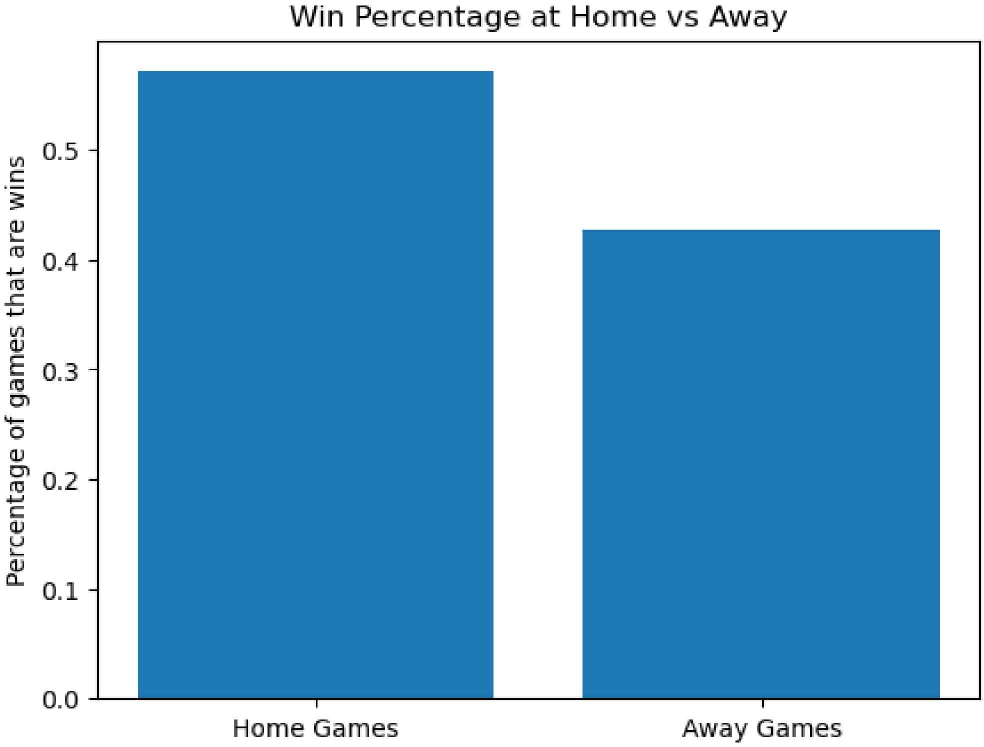
Exploratory Data Analysis (EDA)

The objective of this section is to do two things: the first is to establish whether or not the win proportion at home is different than the win proportion away. Establishing this will allow us to gain motivation in the hypothesis testing we will undergo. The second is to explore which metrics may contribute to a team’s high win rate for both home and away proportions.

Before diving into the figures, below is a quick snapshot as to how the win rate at home and win rate away are defined.

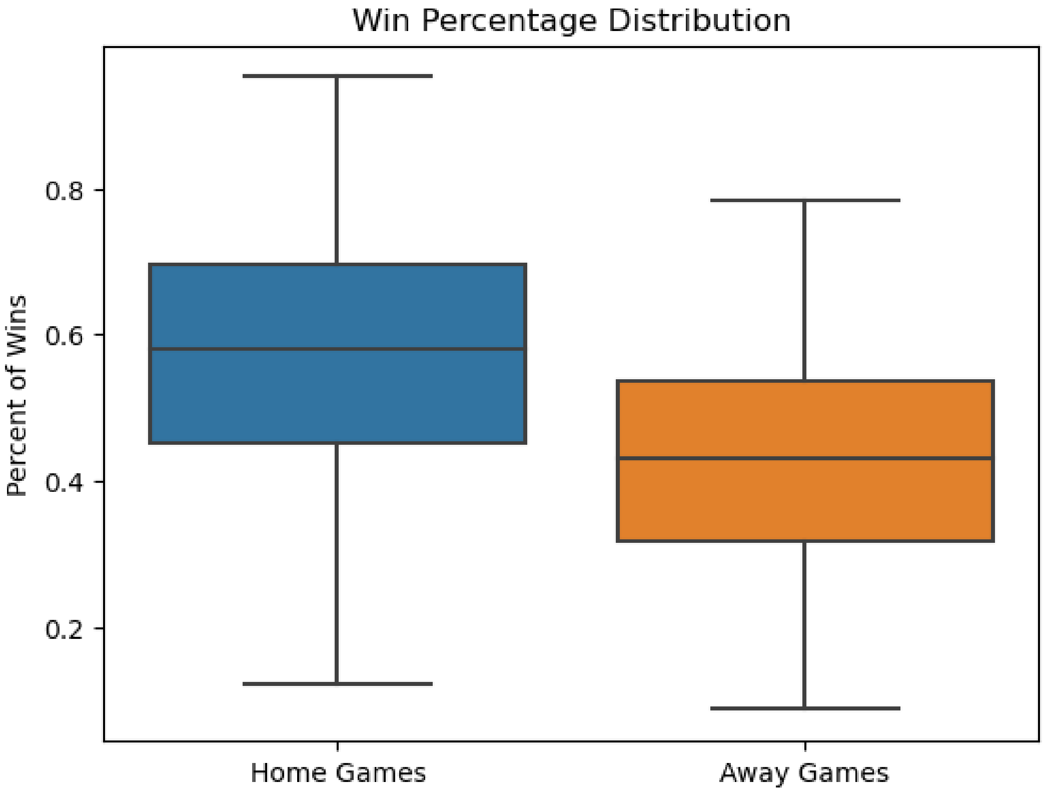
$$\textit{Win Rate Away} = \left(\frac{\textit{The Number of Wins Away}}{\textit{The Number of Games Away}}\right) \qquad \textit{Win Rate At Home} = \left(\frac{\textit{The Number of Wins at Home}}{\textit{The Number of Games at Home}}\right)$$

Taking the average of these metrics across all teams and all seasons, we found that the average Win Rate At Home reached 60% while the average Win Rate Away was 40%.



The findings here validates that there is some difference between win rates at home and win rates away. Therefore, it provides us with motivation to test whether or not the difference in proportions here are due to chance or not. This is going to be the basis of the first part of our investigation.

Since the means across all seasons and teams could be influenced by skewed distributions. We also looked at the distribution of average season win rates for both home and away games. The difference here is that we are not averaging the win rates across teams and seasons, rather we are looking at the distribution of data pooled from 11 NBA seasons across 30 teams. In total the figure alone has 330 data points for the Home games box plot, and 330 data points for the away games box plot (30 teams, 11 seasons per team).

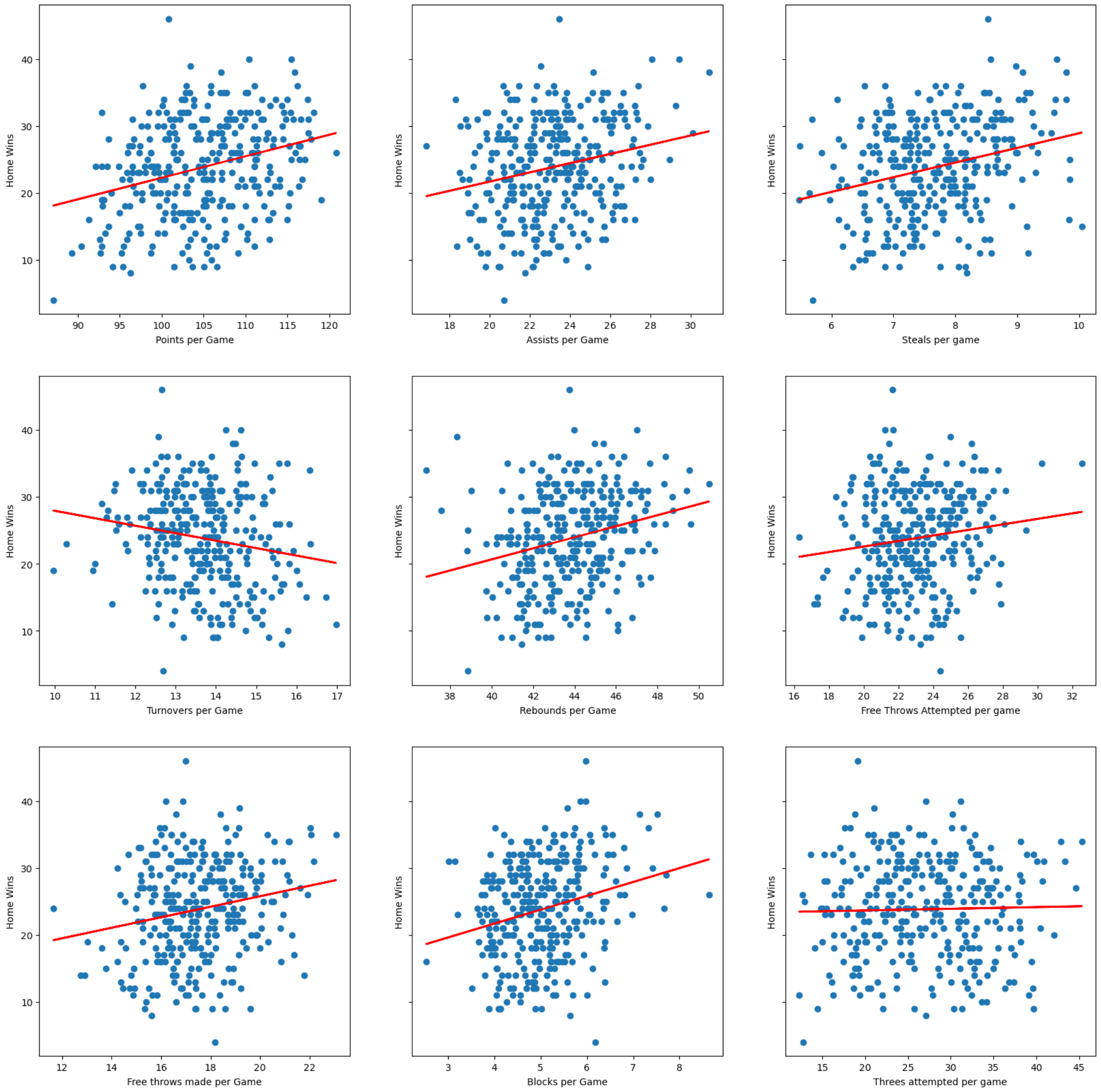


Aligning with the findings from the bar plot above, the distribution of home game wins is centered at a higher value than away game wins. Again this points towards a higher win percentage for home games than away games.

Next we looked at the features that might impact and influence the win rates of teams, for both home and away games. Doing this would highlight which features are most important in predicting wins. This will be crucial for the second half of our research project, in which we are predicting win rate ratios.

Below are a series of scatter plots, with one scatter plot per potentially significant feature. Each scatter plot has 330 total data points (30 teams, with 11 seasons per team). Therefore, the home and away wins per point will represent the number of wins that point (a team during a certain season) obtained.

Home Game Stats Vs Home Wins

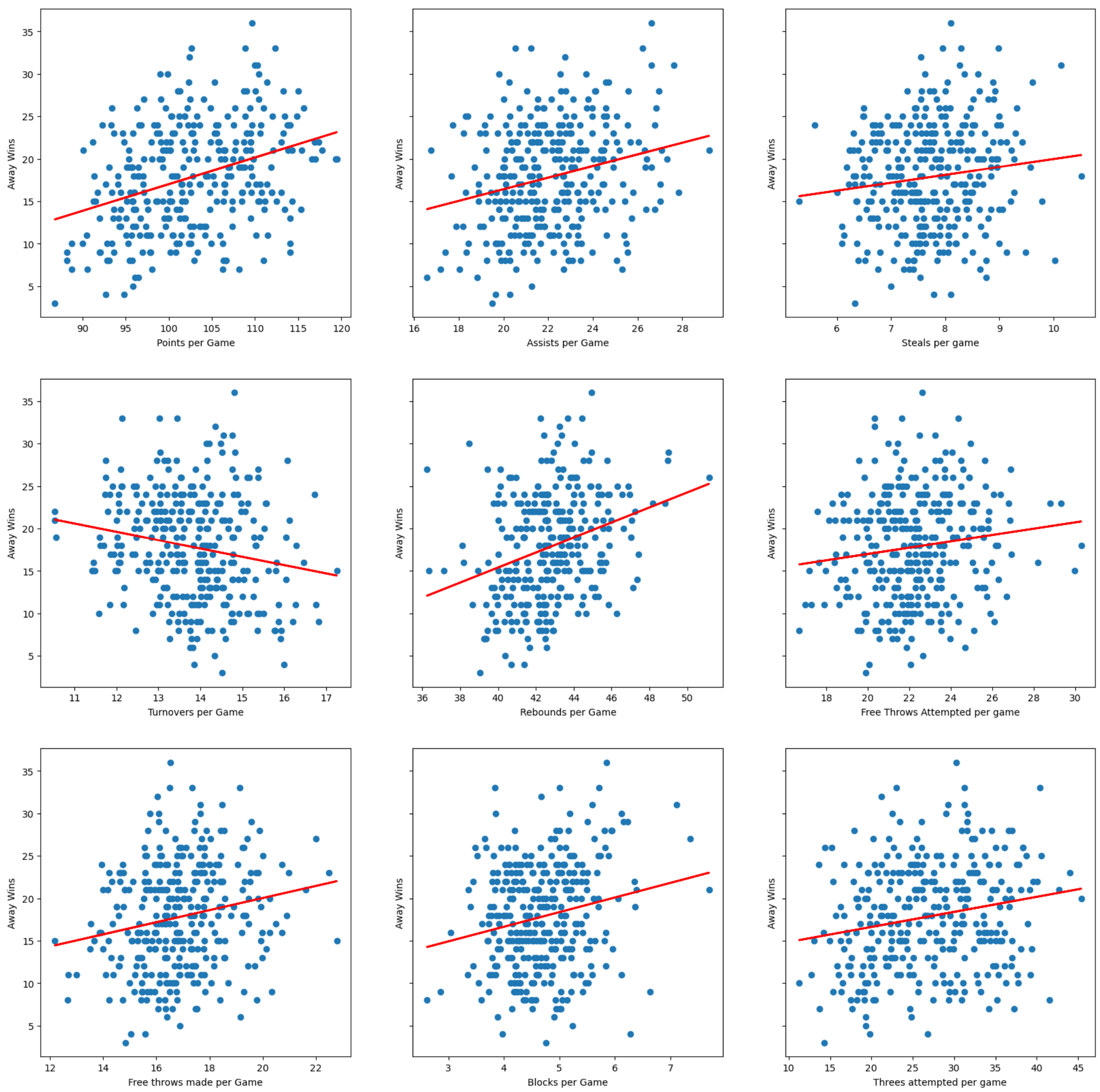


The following statistics were deemed as important in predicting home game wins:

1. Points per game (positive association)
2. Assists per game (positive association)
3. Rebounds per game (positive association)
4. Turnovers per game (negative association)
5. Blocks per game (positive association)
6. Free throws made per game (positive association)
7. Free Throws attempted per game (positive association)
8. Steals per game (positive association)

Notably, the number of threes attempted per game was not associated with the number of wins at home.

Away Game Stats Vs Away Wins



The following statistics were deemed as important in predicting away game wins:

1. Points per game (positive association)
2. Assists per game (positive association)
3. Rebounds per game (positive association)
4. Turnovers per game (negative association)
5. Blocks per game (positive association)
6. Free throws made per game (positive association)
7. Free Throws attempted per game (positive association)
8. Steals per game (positive association)
9. **Threes attempted per game**

A key difference between home and away games resides in the threes attempted. This difference should be accounted for when predicting the ratio of home to away wins—as this variable will mostly impact the number of away games won.

Two things to note here:

1) We are not making causation claims. We realize that we would need to control for teams, and other variables in order to create causal type claims. These scatterplots are merely to highlight certain features that we could use to differentiate between home and away win rates, in order to help us predict ratios later on.
2) With assumption 1 in mind, we could not use these scatter plots as evidence that the home win rate is higher or lower than away win rate. A higher association between threes attempted per game for away teams than home teams, for example, could be due to many confounding variables; however this differentiation will be helpful in our prediction later on.

Hypothesis testing

Context

Again our overarching goal is to see whether or not there is a statistical difference between the proportion of wins at home and away. To do this we will conduct a multiple hypothesis test using our bootstrapped distribution (explained below), and one test statistic per team meaning we'll be running 30 tests per defined statistic.

Implementing the Bootstrap

Why bootstrap? We have access to the past 10 years of data for away and home win rates, but those 10 years alone are not a sufficient sample to test our hypotheses on. Thus we can utilize the bootstrap, sampling with replacement from the last 10 years, where one row represents a given team's performance for a season. Utilizing the bootstrapping method allows us to have a large enough sample for which we can draw conclusions from. This in itself simulates the null distribution, thereby allowing us to test a variety of test statistics.

Note here that we only use the past 10 years as the NBA has changed a lot of the years, and thus the last 10 years is an attempt at looking at the most recent iteration of the game.

Process breakdown:

- Parameters -
  - N = the size of the sample
  - T = the test statistic you want to use.

1) Define a test statistic for a given test (say difference in average points per game at home versus away).
2) Create a table with this test statistic—so you'll have 330 rows (the subtraction of points per game at home versus away for each team and season).
3) Take two samples of size N with replacement from your population in (2).
4) Calculate the test statistic between these two samples.
5) Repeat this process many times.

Result:

- The result is a distribution that resembles the null distribution for a given test statistic.

Focusing on correctional methods

Since we're repeating each hypothesis test (for each test statistic) thirty times—one for each team—we need to ensure that we are adjusting our p-value threshold in order to prevent false discoveries or false positives.

The Bonferroni Correction:
This controls the family wise error rate, which is essentially the probability of getting a false positive. Note that this is the probability across all tests.

The B-H correction:
A procedure that ranks the P-values and compares them to a linearly increasing function. P-values under the increasing function are discoveries. This controls the false discovery rate, which is the probability that one of our discoveries is actually not a discovery.

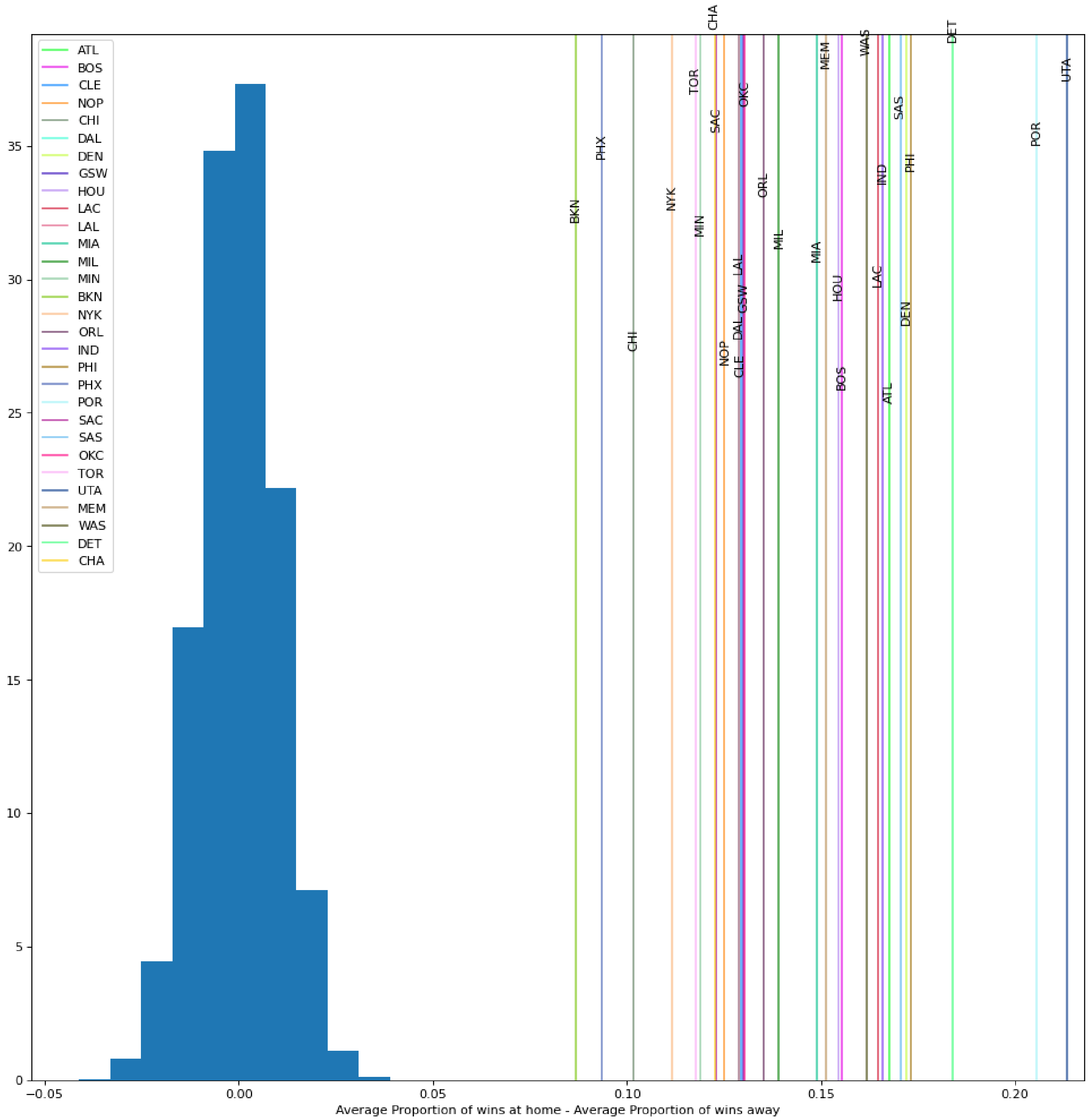
The Bonferroni correction is much more strict, setting the cutoff lower.

We'll be applying these corrections to each one of our tests below.

Proportion of wins at home versus away

The distribution and p-values below show that across every single NBA team, the proportion of wins won at home is different than the proportion of wins won away at a statistically significant level. Because the p-values for each observed statistic is 0, the Bonferroni and B-H procedure do not change the number of discoveries made.

As you can see below, the proportion of wins at home is statistically significant across all teams. This is apparent as the null distribution is in blue (our bootstrapped distribution for the proportion of wins at home being equal to the proportion of wins away) and our p-values displayed as lines to the right of the distribution.



Exploring other features—*the why*

While we can now say that the difference between home and away proportion of wins is not due to chance (with teams at home winning more), we need to get a better understanding as to *why* this is the case.

As a result, running this type of analysis, in the form of hypothesis testing as seen above, will allow us to see which features do differ significantly between home and away games giving us context as to why home teams may perform better in relation to wins. For example, if it is found that the number of points per game is larger at home than away for all NBA teams, then this is a potential factor as to why home teams win more (more points could lead to more wins)!

The end goal here is to shortlist a set of features that are most important to predicting the proportion of wins at home versus the proportion of wins away. That is, identifying the features most statistically significant across all teams will help in our prediction section below.

Initial list

These are the features that we conducted hypothesis testing for.

Offensive stats:

1) Points per game
2) Rebounds per game
3) Assists per game

Defensive stats:

1) Blocks per game
2) Steals per game

Other stats:

1) Plus Minus average per game
2) Free throws made per game

Because of the sheer size of each image, I will not include each test statistic's histogram. Instead is a table below that summarizes the results. If you're interested in seeing the distribution for each test statistic, there will be an appendix at the end where you can access our code.

Results summarized

The Test Stat (AVG home - AVG away)	**The # of discoveries (After each correction)**
Win proportion (baseline)	Bonferroni: 30 B-H: 30
Points per game	**Bonferroni: 26** **B-H: 28**
Rebounds per game	**Bonferroni: 26** **B-H: 29**
Assists per game	Bonferroni: 23 B-H: 24
Blocks per game	**Bonferroni: 27** **B-H: 28**
Steals per game	Bonferroni: 9

	B-H: 15
Free throws made per game	Bonferroni: 18 B-H: 21
Team Plus Minus per game	**Bonferroni: 30** **B-H: 30**

The statistics with the highest number of discoveries in order:

1. **Team Plus Minus per game**
2. **Rebounds per game**
3. **Blocks per game**
4. **Points per game**

These are the statistics with the highest number of discoveries. That is, these statistics vary not due to chance greatly, across all (and in some cases almost all) NBA teams. Thus we should shortlist these variables as the ones most important for the prediction algorithm below!

A Note on interpretation
23 discoveries under Assists per game, for example, means that 23 of the 30 NBA teams differ in their number of assists per game between home and away games.
Above we are just taking the maximum discoveries after both corrections.

A deeper dive into Plus Minus

Traditionally, plus minus is assigned to a player, and represents the number of additional points a team will score when a player plays.

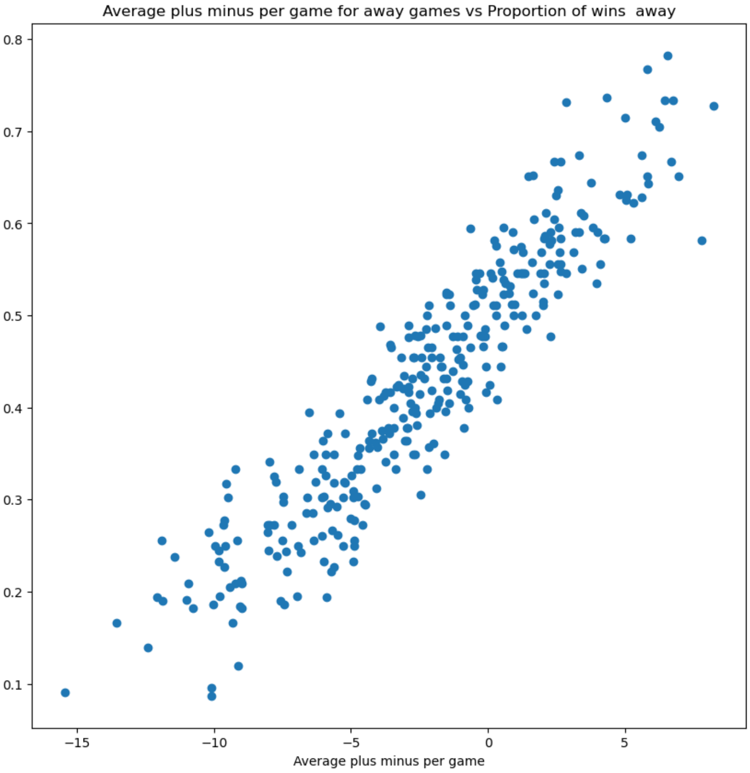
Generalizing this to a team level per game, we can take the total plus minus for a team by adding each player's plus minus on that team. Finally we can get the average team's plus minus per game by dividing by the number of games played in a season.

This metric is then the average team plus minus per game.

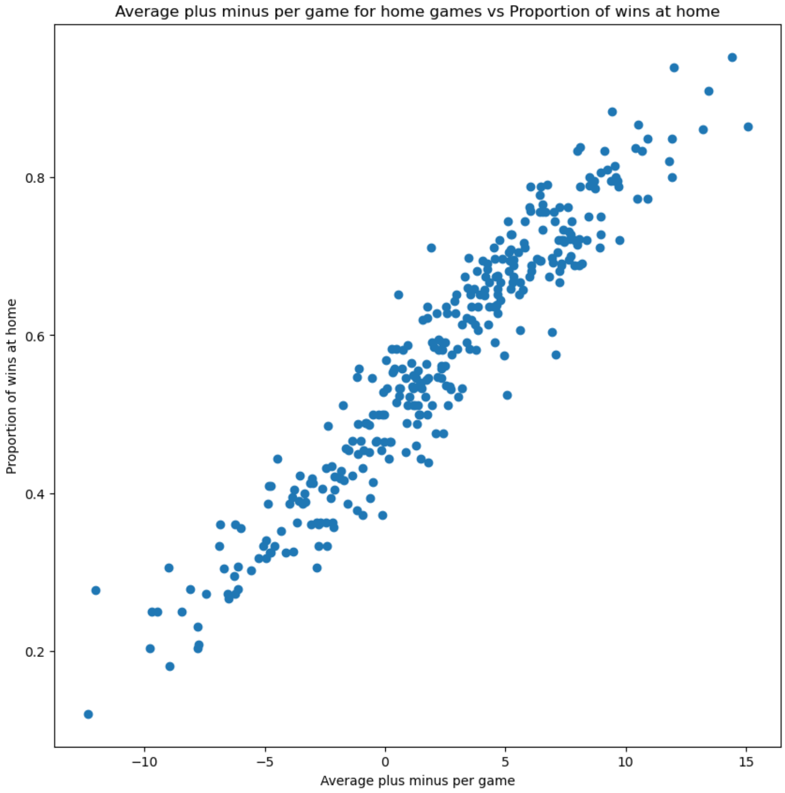
Interpretation
A plus minus score for a team per game of 5 would mean that on average a player on said team contributes 5 more points when they play.

Thus comparing this metric at home games versus away (through hypothesis testing above) allowed us to assess average player performance per team at home versus away.

This is directly related to the number of games won at home versus away. If a team's players perform better at home, then they'll win more at home. This is proven by the scatterplots below.



Average plus minus per game for away games vs Proportion of wins away



Average plus minus per game for home games vs Proportion of wins at home

Conclusion for Hypothesis testing section

There is an obvious statistically significant difference between the proportion of games won at home and proportion of games won away across all NBA teams. The difference indicated that all NBA teams win more games at home than away.

In order to see where this result is potentially derived from, we shortlisted 7 potential factors that could differ between home and away games resulting in more wins at home.

1) Points per game
2) Team plus minus per game
3) Assists per game
4) Rebounds per game
5) Blocks per game
6) Steals per game
7) Free throws made per game

The statistics with the highest proportion of NBA teams seeing statistically significant results between home and away games were:

1) Points per game
2) Team plus minus per game
3) Rebounds per game
4) Blocks per game

These statistics should be used in the prediction section below, as they most clearly differ across all teams between home and away! In other words there is some association between having a higher team plus minus per game and winning more games at home.

Predicting

Context

Having come up with the features that are most differentiated between home and away, we’re now able to create a model to predict the proportion of home games won minus the proportion of away games won. The logic is that if a given feature is statistically significant in its difference between home and away (meaning the statistic differs between home games and away games not due to chance) then we can use this feature to predict the difference between home and away outcomes—as the feature would presumably impact game outcomes.

In order to evaluate our models, we will be testing them on the 2021-2022 season and determining whether or not our model was representative of the true results.

Note our training data consists of NBA data of the last 20 years as we needed slightly more data to build the model.

Technical overview

We’ll be using a random forest and a linear regression to predict the following metric:

$$\frac{\text{number of wins at home}}{\text{number of home games}} - \frac{\text{number of wins away}}{\text{number of away games}}$$

We’ll be assessing the performance of each and comparing them with one another.

Multi Linear regression results

Again just to recap, the hypothesis testing section highlighted the following variables to be used in our multi linear regression model.

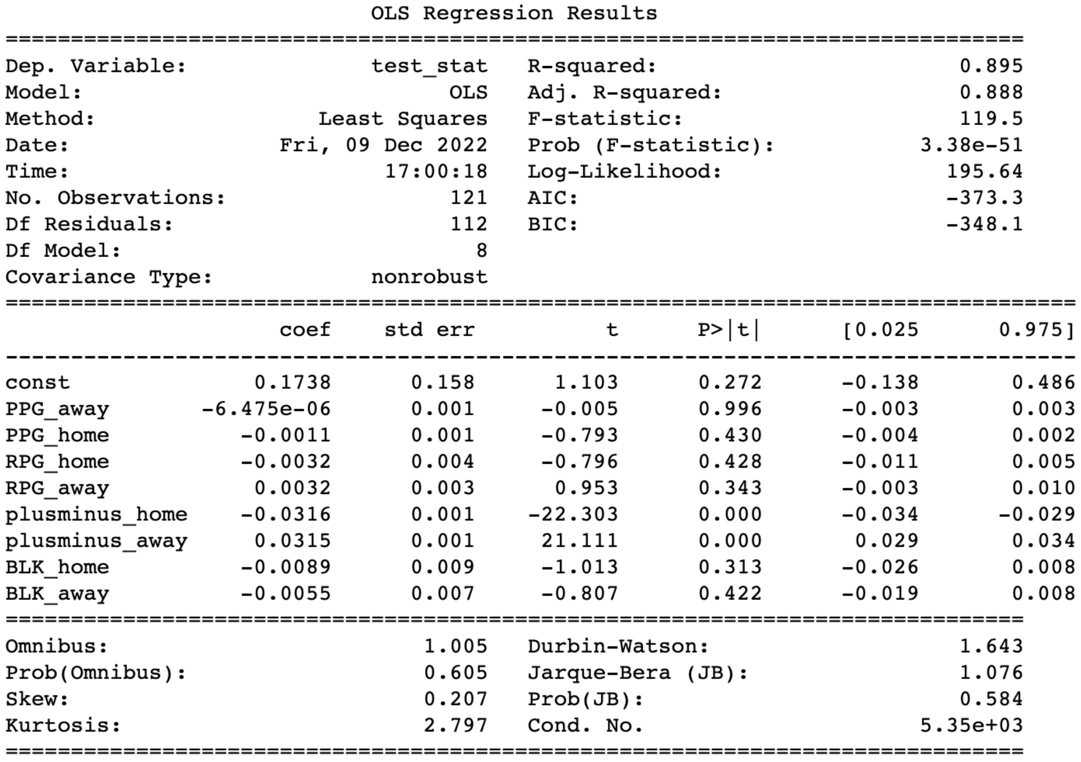
1) Team plus minus per game
2) Points per game
3) Rebounds per game
4) Blocks per game

To estimate how well our model did, we used the mean squared error which essentially measures how much we vary from the mean.

Assessing the performance of our Multi Linear Regression

The mean of the test statistic we want to predict (the proportion of wins at home - away) for the 2021-2022 season was about .09. Our root mean squared error was **triple** that, meaning on average our predictions vary pretty greatly from the mean. To put into perspective, if a given test statistic is 0, then our model on average would predict a value that varies by **.28** in either direction.

If we look at the coefficients of the linear regression, we'll see the following.



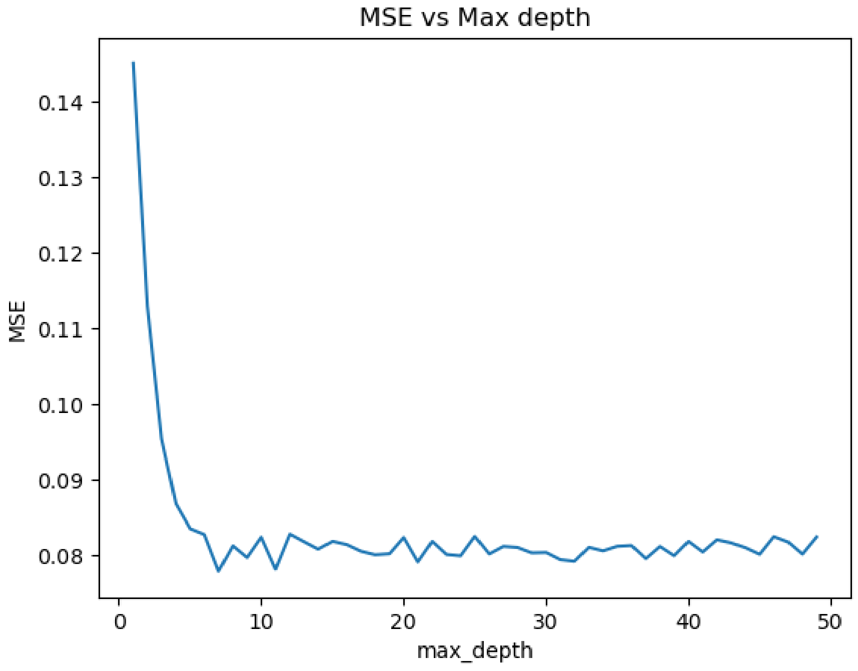
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              test_stat   R-squared:                       0.895
Model:                            OLS   Adj. R-squared:                  0.888
Method:                 Least Squares   F-statistic:                     119.5
Date:                Fri, 09 Dec 2022   Prob (F-statistic):           3.38e-51
Time:                        17:00:18   Log-Likelihood:                 195.64
No. Observations:                 121   AIC:                            -373.3
Df Residuals:                     112   BIC:                            -348.1
Df Model:                           8
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              0.1738      0.158      1.103      0.272      -0.138       0.486
PPG_away       -6.475e-06      0.001     -0.005      0.996      -0.003       0.003
PPG_home          -0.0011      0.001     -0.793      0.430      -0.004       0.002
RPG_home          -0.0032      0.004     -0.796      0.428      -0.011       0.005
RPG_away           0.0032      0.003      0.953      0.343      -0.003       0.010
plusminus_home    -0.0316      0.001    -22.303      0.000      -0.034      -0.029
plusminus_away     0.0315      0.001     21.111      0.000       0.029       0.034
BLK_home          -0.0089      0.009     -1.013      0.313      -0.026       0.008
BLK_away          -0.0055      0.007     -0.807      0.422      -0.019       0.008
==============================================================================
Omnibus:                        1.005   Durbin-Watson:                   1.643
Prob(Omnibus):                  0.605   Jarque-Bera (JB):                1.076
Skew:                           0.207   Prob(JB):                        0.584
Kurtosis:                       2.797   Cond. No.                     5.35e+03
==============================================================================
```

Thus, the model itself holds all variables except the plus_minus home and away as non significant. Which means these are the only two variables that the model shows to hold a relationship with the proportion of wins per game not due to chance when controlling for the other variables.

Below we'll use a non parametric model in order to predict the test statistic in hopes of achieving lower MSE.

Random forest results

After optimizing our by performing a grid search on the max_depth parameter, we found that the max_depth of 8 yields the lowest mean squared error. Below I've attached the results as to how our figure looks like.



When predicting the 2021-2022 season home vs away proportion, we found that the random forest regressor achieved a mean squared error of .029. Interestingly enough if we decrease our max depth to 3, we can achieve a result of .0144(which is much stronger). Despite this, if the goal of our model is to predict future seasons, we want to have it be generalizable. Therefore, a max depth of 6 may be optimized for the 2021-2022 season, it may not be for others. On the other hand a max_depth of 8 was derived from a cross validation of our current training data, and is therefore more **robust**.

Assessing Performance of the Random Forest

The mean of the test statistic we want to predict (the proportion of wins at home - away) for the 2021-2022 season was about .09. Our root mean squared error was **double** that, meaning on average our predictions vary pretty greatly from the mean. To put into perspective, if a given test statistic is 0, then our model on average would predict a value that varies by **.17** in either direction.

Naturally, this is not the most accurate model—and rightfully so. Despite using the last 20 years of NBA data to train the model, each year teams often change and therefore it becomes difficult to correctly predict their numerical proportion. **However, the result of the random forest was much stronger than the linear regression above!** This could be due to the fact that our random forest does not assume anything about the distribution of data while the linear regression assumes that our data is normal.

Attempting to improve the Random Forest model

In order to attempt to improve our model's MSE, I tried to include a variable that was not as statistically significant across all teams (between home and away). This variable was assists per game. Unsurprisingly, adding this variable to the model actually decreased the test mean squared error. This makes intuitive sense because assists per game is not statistically significant across all teams, and therefore is not a complete differentiator between home and away performance (like the other variables used). Thus, it adds more noise to the feature table, leading to a lower performing model.

Future improvements

A root mean squared error of .17 of a model that predicts a dependent variable with a mean off .09 is far from perfect. Having said this, there are a few things that could be done in the future to improve the model itself.

1) Including more data.
    a) Though 20 years of NBA data is sufficient, our API was structured in a way that prevented us from being able to actually use all twenty seasons.
    b) Instead we were able to only access about 30% of these seasons resulting in a much smaller than anticipated training set.
    c) Including more data overall should improve the model itself creating more patterns for the random forest to learn to split on.
2) Including and testing advanced metrics.
    a) There are some advanced NBA statistics that were not tested between home and away games for the sake of simplicity.

b) For example, efficiency ratings were left out.
c) If these metrics were found to vary significantly between home and away games, then they could be included in our model to improve it overall.

Conclusion

To conclude, the proportion of wins at home minus the proportion of wins away was statistically significant. This led us to conduct a hypothesis test for each team metric that may contribute to this higher win proportion at home.

Using the variables that were statistically significant between home and away allowed us to create two distinct models. The first model was a parametric multi linear regression which resulted in a mean root mean squared error of .27. This model found the strongest relationship between the plus minus statistics, listing the other variables as non statistically significant.

Once we utilized the random forest regression model, we were able to achieve a root mean squared error of .17 which is greatly improved from the performance of the multi linear regression.

Overall, we were successful in creating a model that somewhat accurately predicts the proportion of home wins minus the proportion of wins away. In order to achieve an even higher accuracy we would need an even larger data set. We could also potentially experiment with other nonparametric regression methods such as Gradient Boosting!

Appendix

We're using the NBA API which scraped the data from NBA.com. The API can be accessed here:
https://github.com/swar/nba_api

Some things to note:

1) We needed to write many functions to combine the data in an appropriate way for our purposes. For example, a function was created to separate home versus away games for each NBA team. This is what allowed us to create the proportion comparison between both types of games.
2) The full data set (over 10 years for the hypothesis testing portion) contained about 25k rows which needed to be aggregated by team and season, which again was handled by one of our many functions.
3) The notebook has some commentary as to what types of work we did.