

Springboard

Data Science Career Track

Capstone Project 1

**What factors contribute to a home team winning an NCAA
basketball game?**

By

Adeyemi Adejuwon

July, 2020

Abstract

From 1985-2019, teams in the regular season of NCAA Division 1 Men's basketball games won 92,732 of their home games but only won 47,547 of their away games, and 15,805 of matches played in neutral locations. This home court advantage is further confirmed with a Tukey HSD test, where the statistically significant difference between the pairs home and away games, the pairs home and neutral games, versus the pair away and neutral games was calculated.

Many articles have explained that the reason for a team's home-court success is the presence of fans and the arena, as opposed to the fact that they are simply the better team in a particular matchup. Assuming fans are the reasons for this home court advantage, the odds are already stacked against a visiting team, when playing these basketball matches. My first capstone project tries to analyse some of the differences between playing at home vs playing away.

A logistic regression algorithm was used to build a model, predicting whether a match finishes 0 (home win) or 1 (away win) given input derived from game statistics aggregated over a window of two past games.

All deliverables required by this project can be found in my GitHub repository, a link to which follows:

<https://github.com/dreamtx01/Springboard/tree/master/Folders/Capstone%20Project%201>

The results of my capstone 1 are:

- Home court does give an advantage in basketball games played in the NCAA.
- The models predicted up to sixty-nine percent accuracy for home games when selecting all the features for the model.
- When using top five ranked in-game statistics (features), as determined by a feature selection algorithm, the performance metrics for the model stayed the same.

Abstract	2
Introduction	5
Objective	5
Datasets	5
Data Cleaning & Wrangling	7
Data Types	7
Handling Missing Data	7
Exploratory Data Analysis	8
Fig 1: Box Plot - points scored by winning team at playing locations	9
Fig 2: Box Plot - points scored by losing team at playing locations	10
Fig 3: Bar Chart - Proportion of games played at different locations in all seasons	11
Fig 4: Line Chart - Points scored per winning team per season	13
Fig 5: Line Chart - Three-Pointers scored per winning team per season	14
Fig 6: Line Chart - Turnovers conceded per winning team per season	15
Fig 7: Bar Chart - Ranking top 15 teams by win percentage for 2003 -2019	16
Fig 8: Bar Chart - Ranking top 15 teams by win percentage for 1985 -2019	16
Applications of Inferential Statistics	17
Fig 9: ANOVA test & Tukey HSD	18
Fig 10: Pairplot Matrix	19
Fig 11: Bootstrap Replicates	20
Machine Learning Models	22
Data manipulation for machine learning	22
Baseline Modeling: Logistic Regression Model Process	23
Fig 12: Features used for the analysis	24
Fig 13: Data frame after two day moving average calculation	25
Fig 14: Original data frame before two day moving average calculation	25
Fig 15: Dataset with Home games	26
Fig 16: Confusion Matrix and Classification Report for Training Set	26
Fig 17: Confusion Matrix and Classification Report for Test Set	27
Extended Modeling: Feature Selection by Recursive Feature Elimination (RFE)	28
Fig 18: Confusion Matrix and Classification Report for Test Set, after building model with the features selected by RFE	29

Conclusions	31
Recommendations For the Client	32
Future Work	33
Works Cited	34

Introduction

The phenomenon of home field advantage is nothing new in the world of sports ¹. Home teams generally have the advantage of having the support of larger, more enthusiastic crowds, and it has been suggested in some studies, that it is the home court atmosphere that enhances the home teams opportunities for winning matches².

Objective

The goal of my project is to study game factors that affect a home team winning an NCAA Division 1 Men's basketball game, and likewise affect an away team losing a match. The information derived from these analysis can help a basketball coach tune his game play tactics and thereby increase his odds of winning a game.

Datasets

The datasets for this analysis will be obtained from Kaggle³. The datasets from NCAA March Madness are provided by a Kaggle competition sponsored by Google.

The regular season detailed results file identifies the game-by-game match play data and results, for regular season matches for the years 2003 - 2019. The dataset for the regular detailed season games consists of 87,366 data points for 350 college basketball teams playing in the NCAA since 2002. This dataset includes 34 variables such as number of assists, three-point percentages per game, win and loss records per season, location of matches played.

¹ "Home-Field Advantage (SOCIAL PSYCHOLOGY" [Home-Field Advantage \(SOCIAL PSYCHOLOGY\) - iResearchNet](#). Accessed 14 May. 2020.

² "The Home Court Advantage: Evidence from Men's College" 9 Mar. 2017, <http://thesportjournal.org/article/the-home-court-advantage-evidence-from-mens-college-basketball/>. Accessed 14 May. 2020.

³ <https://www.kaggle.com/ncaa/ncaa-basketball>

The regular season compact results identify just the team losses and wins from 1985-2016. The dataset for regular season compact games consists of 156,089 entries and 8 variables. These variables do not include in-game data.

The team spellings file is used to correlate TeamID numbers with their associated names.

- WTeamID - this identifies the id number of the team that won the game.
- WScore - this identifies the number of points scored by the winning team.
- LTeamID - this identifies the id number of the team that lost the game.
- LScore - this identifies the number of points scored by the losing team.
- WLoc - this identifies the "location" of the winning team. The home team is given the value "H", while the visiting team is given the value "A", and the value "N" is given to a match played on a neutral location.
- NumOT - this indicates the number of overtime periods in the game, an integer 0 or higher.
- WFGA - field goals attempted (by the winning team)
- WFGM3 - three pointers made (by the winning team)
- WFGA3 - three pointers attempted (by the winning team)
- WFTM - free throws made (by the winning team)
- WFTA - free throws attempted (by the winning team)
- WOR - offensive rebounds (pulled by the winning team)
- WDR - defensive rebounds (pulled by the winning team)
- WAst - assists (by the winning team)
- WTO - turnovers committed (by the winning team)
- WStl - steals (accomplished by the winning team)
- WBlk - blocks (accomplished by the winning team)
- WPF - personal fouls committed (by the winning team)

Data Cleaning & Wrangling

Before beginning analysis of the data, it is essential to explore there are no missing values in the dataset. It is also essential for the data in our table to be of the correct data type. This upfront work will give more confidence in interrogating the data and would allow better conclusions to be made as regards the dataset. The libraries used for the data cleaning and wrangling of the data sets are:

- numpy for scientific computing of the numerical arrays
- pandas for data analysis and manipulation,
- matplotlib for visualization

Data Types

The next step in the data wrangling stage was to determine the type of data type in the dataset. As can be observed in the table below, there are 87,366 entries, with no missing values in any of the 34 columns. Additionally, all but one column takes integer values, whereas the lone column (WLoc) takes a string entry.

In fact from the data set description we know that the WLoc column will take only three values each one representing the location of games played. To confirm the entry of the WLoc column we call the unique () function on that column.

- “H” stands for “Home game
- “A” stands for away (visiting to opponent’s site)
- “N” is the location of games played at a neutral location

Handling Missing Data

The next step in the data wrangling stage is to check for any gaps in the dataset. This is confirmed with the function “is null” and “value_counts”. The isnull function finds the null

value in the data set, while the `value_counts` function displays the amount of the categorical variables in `WLoc`.

Based on these results, we can observe that there are **no missing** values in the dataset.

It should be noted that if there were missing values in the column we would either drop them or fill them in. This is because some of the techniques in the data exploratory will not allow for missing data.

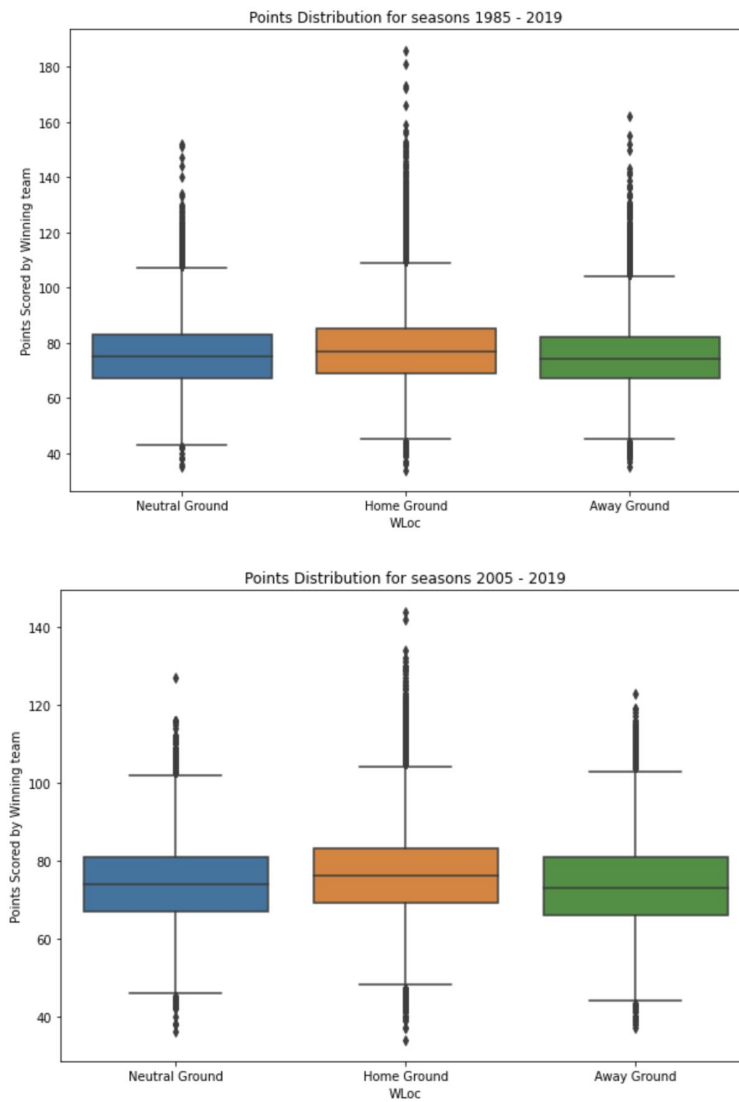
Exploratory Data Analysis

Following the data wrangling of the data, the next step is to interrogate the dataset and ask a series of questions of the dataset. These questions will help identify the contributing factors affecting home games winning matches. The questions being asked of the data are :

1. Does a winning team score more points when playing at home, than when playing at either a neutral ground or an away ground?
2. Does a losing team score more points when playing at home, than when playing at either a neutral ground or an away ground?
3. Is there a difference in the amount of matches a team wins at home, away or a neutral location?
4. What is the average variation in points scored for the winning team per season?
5. What is the average variation in three-points scored by the winning team when playing at home, away or a neutral location per season?
6. What is the average turnover by the winning team when playing at home, away or a neutral location per season?
7. What is the ranking of the top 15 teams based on the winning percentage per season for seasons 1985-2019?
8. What is the ranking of the top 15 teams based on the winning percentage per season for seasons 2003-2019?

1. Does a winning team score more points when playing at home, than when playing at either a neutral ground or an away ground?

Fig 1: Box Plot - points scored by winning team at playing locations



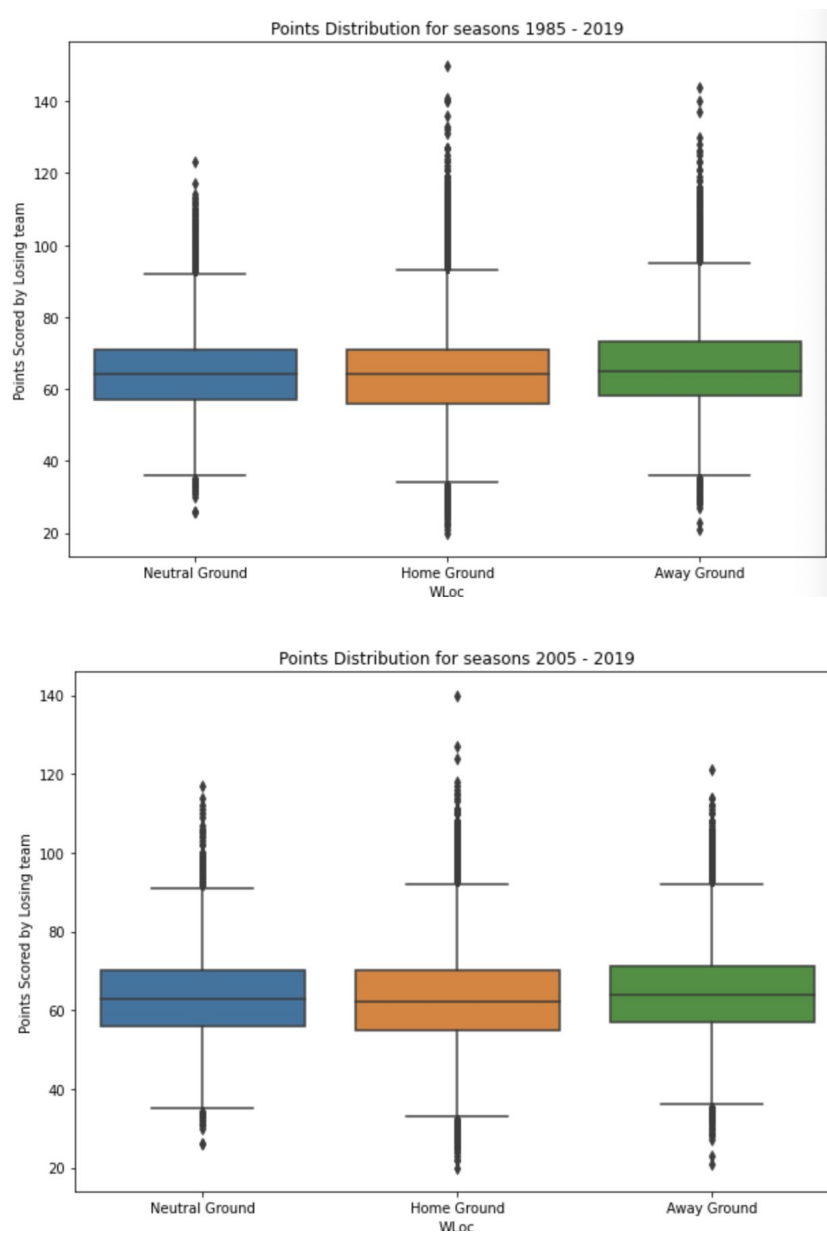
Conclusion

The box plots show that the average number of points for the winning team at their home ground is higher than the points scored when playing at either away or neutral grounds. This result will

be further investigated in the applications inferential statistics section of the report. This box plot confirms the phenomenon of home advantage being present for home teams.

2. Does a losing team score more points when playing at home than when playing at either a neutral ground or an away ground?

Fig 2: Box Plot - points scored by losing team at playing locations



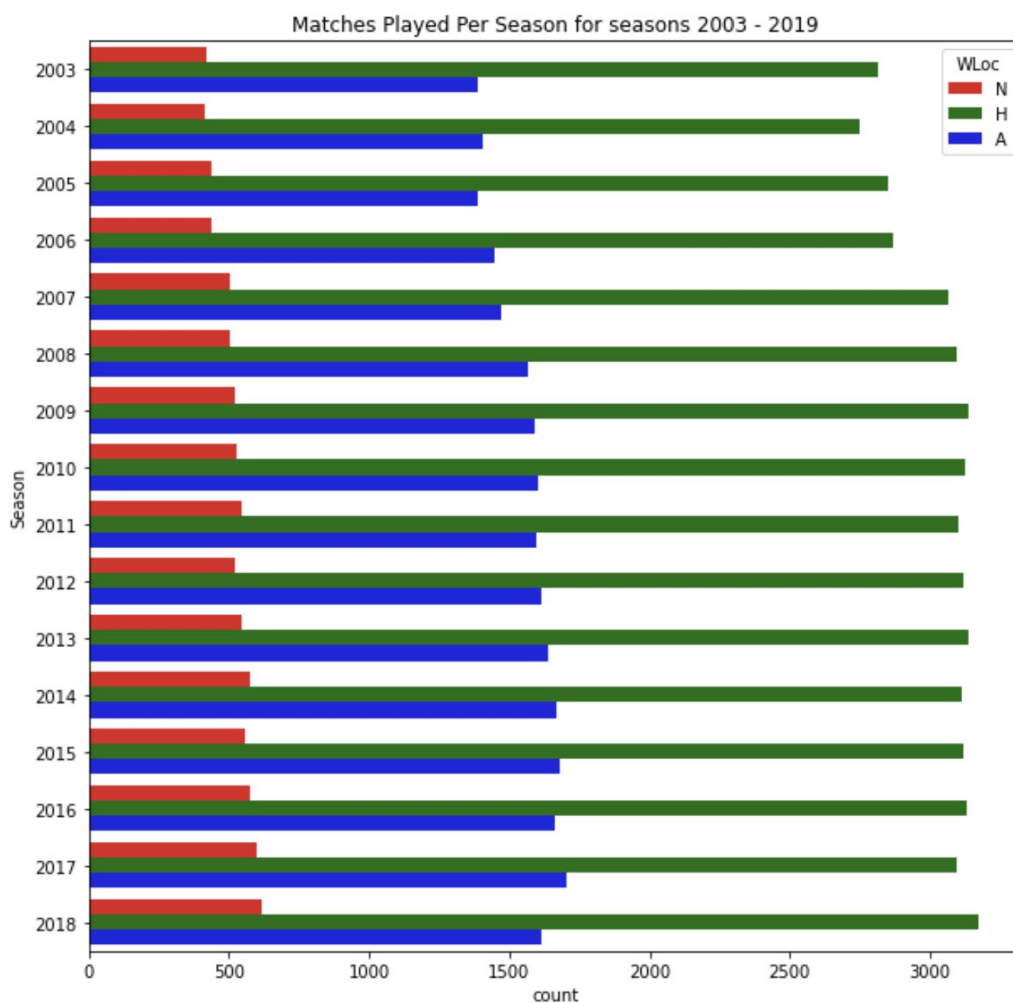
Conclusion

The plots seem to show that the losing teams score more points when playing at an away ground versus when playing at their neutral or home ground. However, there seems to be no difference to the average number of points scored by the losing team when playing in either of the neutral or home locations.

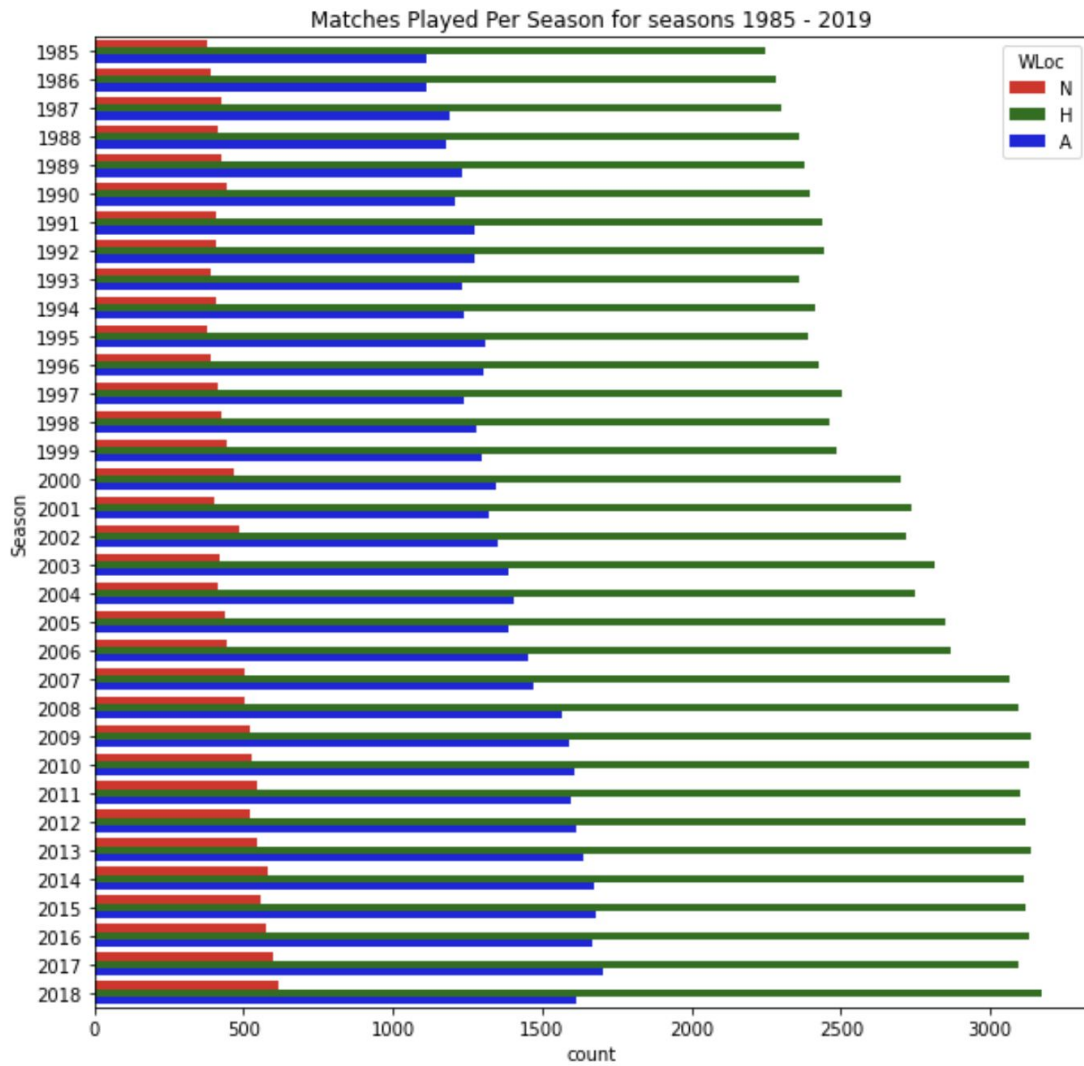
3. What is the difference in the amount of games a team wins either home, away or at a neutral location?

Fig 3: Bar Chart - Proportion of games played at different locations in all seasons

[19]



[20]



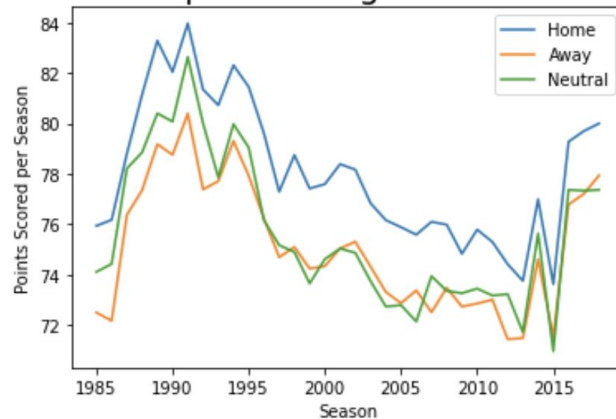
Conclusion

The number of all games won at home, away and in a neutral ground were all similar across all seasons.

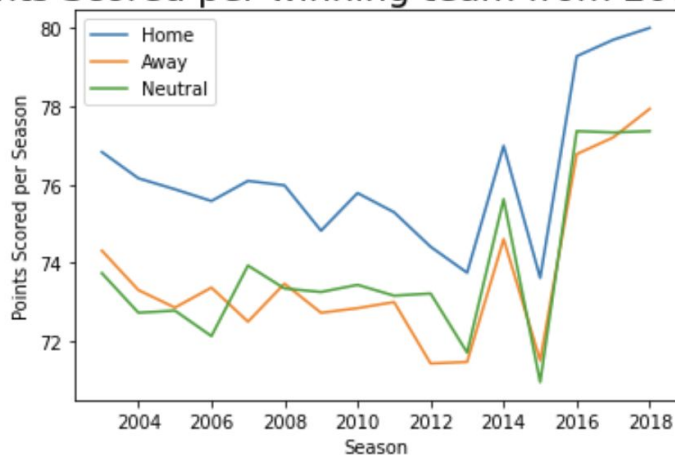
4. What is the average variation in points scored for the winning team per season?

Fig 4: Line Chart - Points scored per winning team per season

Points Scored per winning team from 1985-2019



Points Scored per winning team from 2003-2019



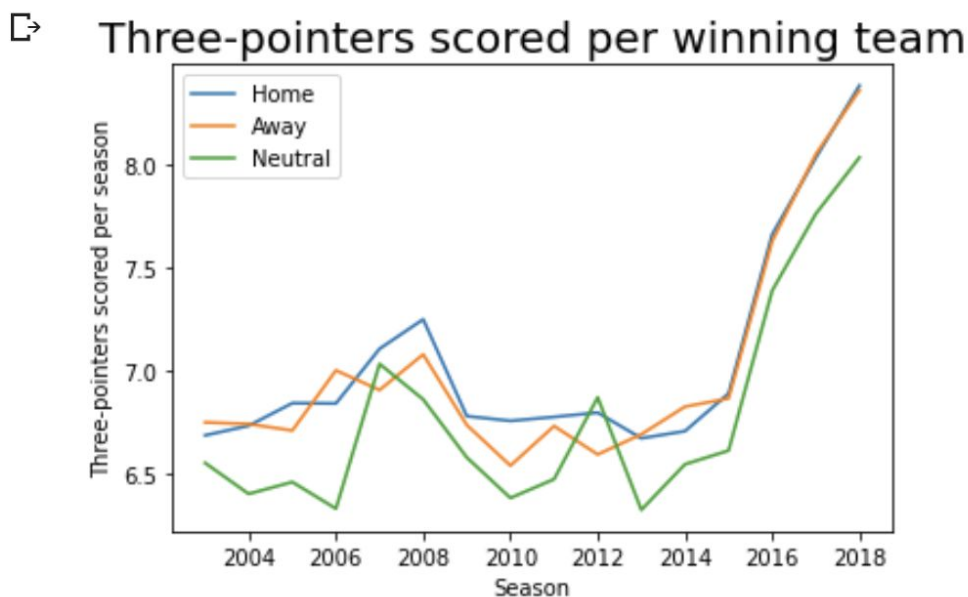
Conclusion

For all seasons the number of points scored by the winning team while playing at home is consistently greater than points scored when playing at an away or a neutral location. After an initial upward trend in points scored from 1985 -1990, the total number of points scored per winning team started decreasing, and later started following a general upward trend in the points

scored per season, from year 2016 onwards. Some suggestions for the uptick in scoring are that maybe teams are becoming more efficient in scoring, or it could also be that the rules of basketball have changed to favor the scoring team. Another observation from this analysis is that in 2015 there was a dip in scoring in all locations. This observation would require further investigation as to why there was a specific drop in scoring this particular year.

5. *What is the variation in three-pointers scored by the winning team when playing at home, away or a neutral location per season?*

Fig 5: Line Chart - Three-Pointers scored per winning team per season

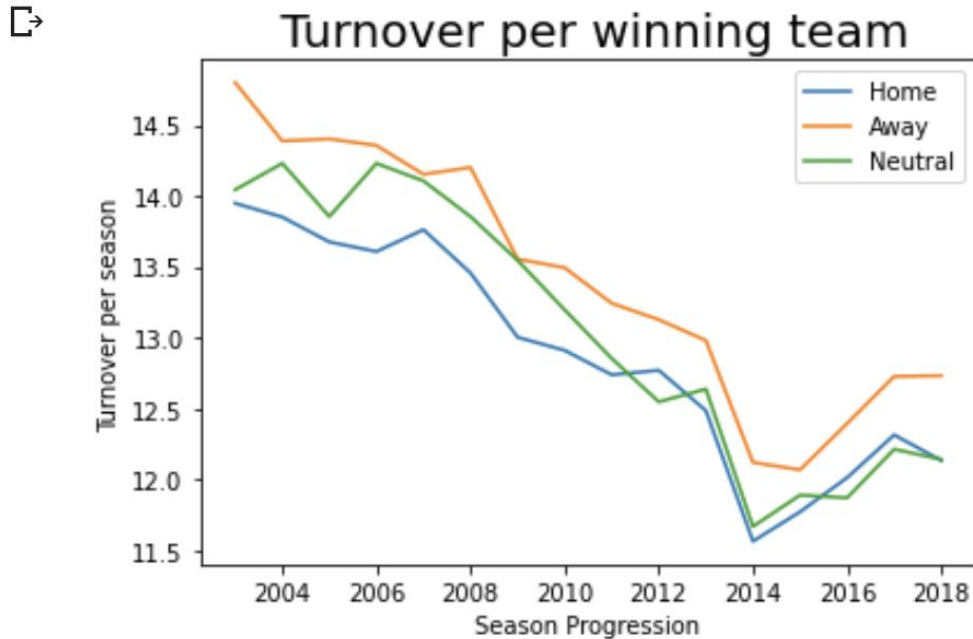


Conclusion

There seems to be a general upward trend in three pointers scored as the years go by. It can also be observed for most seasons the number of three-pointers scored by the winning team while playing at home is greater than three-pointers scored when playing at an away, or a neutral location. Some of the reasons for this general increase in three-pointers could be the game of basketball is evolving to more teams scoring more three pointers.

6. What is the amount of turnovers per season?

Fig 6: Line Chart - Turnovers conceded per winning team per season



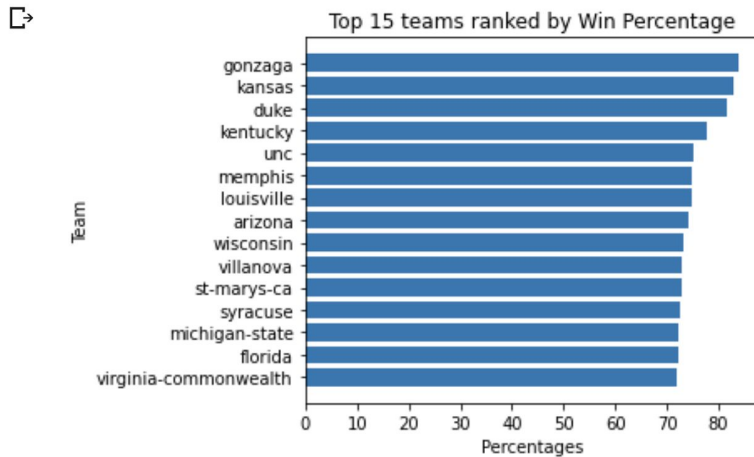
Conclusion

For all seasons the number of turnovers conceded while playing at an away location for the winning team is greater than turnovers conceded when playing at an home or a neutral location.

There also seems to be a decrease in number of turnovers through the years.

7. What are the winning percentages for the top 15 teams in seasons 2003 -2019?

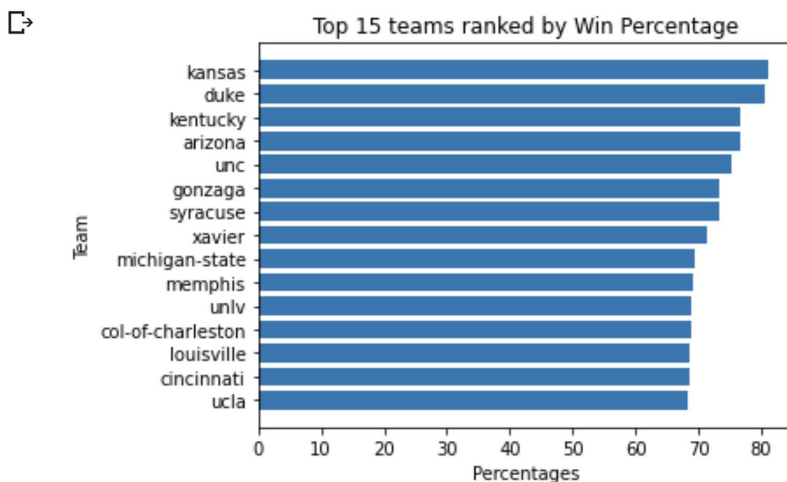
Fig 7: Bar Chart - Ranking top 15 teams by win percentage for 2003 -2019



The team with the highest winning percentage is Gonzaga followed by Kansas

8. What are the winning percentages for the top 15 teams in seasons 1985-2019?

Fig 8: Bar Chart - Ranking top 15 teams by win percentage for 1985 -2019



The team with the highest winning percentage is Duke followed by Kansas

Applications of Inferential Statistics

Upon exploration of the data and making initial observations about the data, the next step will be to find out whether there are any correlations within the in-game statistics, and also confirm whether some of our initial observations are statistically significant. The questions that was asked of the data were:

1. Is the difference between the variance of games won by home and away teams statistically significant ? This question will be answered with the aid of an ANOVA (Analysis of Variance) test, followed by a pairwise Tukey HSD (Honest Significant Difference) correlation.
2. Is there a relationship between in-game statistics for a winning team? The correlation matrix from the heat data will show the relationships that exist between in game statistics during the season.
3. What is the 95% confidence interval for the difference between the standard deviations of the winning scores for the two top performing teams? This analysis will be done with a bootstrap sampling analysis, calculating these differences over 10000 replicates. The teams being considered will be Gonzaga and Kansas.

1. Is the difference between a team winning college basketball matches at home locations statistically significant from when playing at away or a neutral location? This question will be answered with the aid of an ANOVA (Analysis of Variance) test, followed by a pairwise Tukey HSD correlation.

When comparing the three numerical datasets, ANOVA (Analysis of Variance) was used to test the null hypothesis that all of the datasets, thereby investigating if the difference observed among the mean of variables is statistically significant. If we reject the null hypothesis with ANOVA,

we're saying that at least one of the sets has a different mean; however, it does not tell us which datasets are different.

We can use the SciPy function `f_oneway` to perform ANOVA on multiple datasets of winning scores of teams playing in home, away and neutral locations. The `f_oneway` function takes in each dataset as a different input and returns the **t-statistic and the p-value**.

Fig 9: ANOVA test & Tukey HSD

```

[>] Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====
group1 group2 meandiff p-adj  lower  upper  reject
-----
      a      b -2.5265  0.001 -2.7284 -2.3247   True
      a      c -2.253  0.001 -2.561 -1.9449   True
      b      c  0.2736  0.1247 -0.055  0.6022  False
=====

```

H_0 = all three locations have the same mean score in the basketball game

H_a = at least one of the locations have different means in the basketball game.

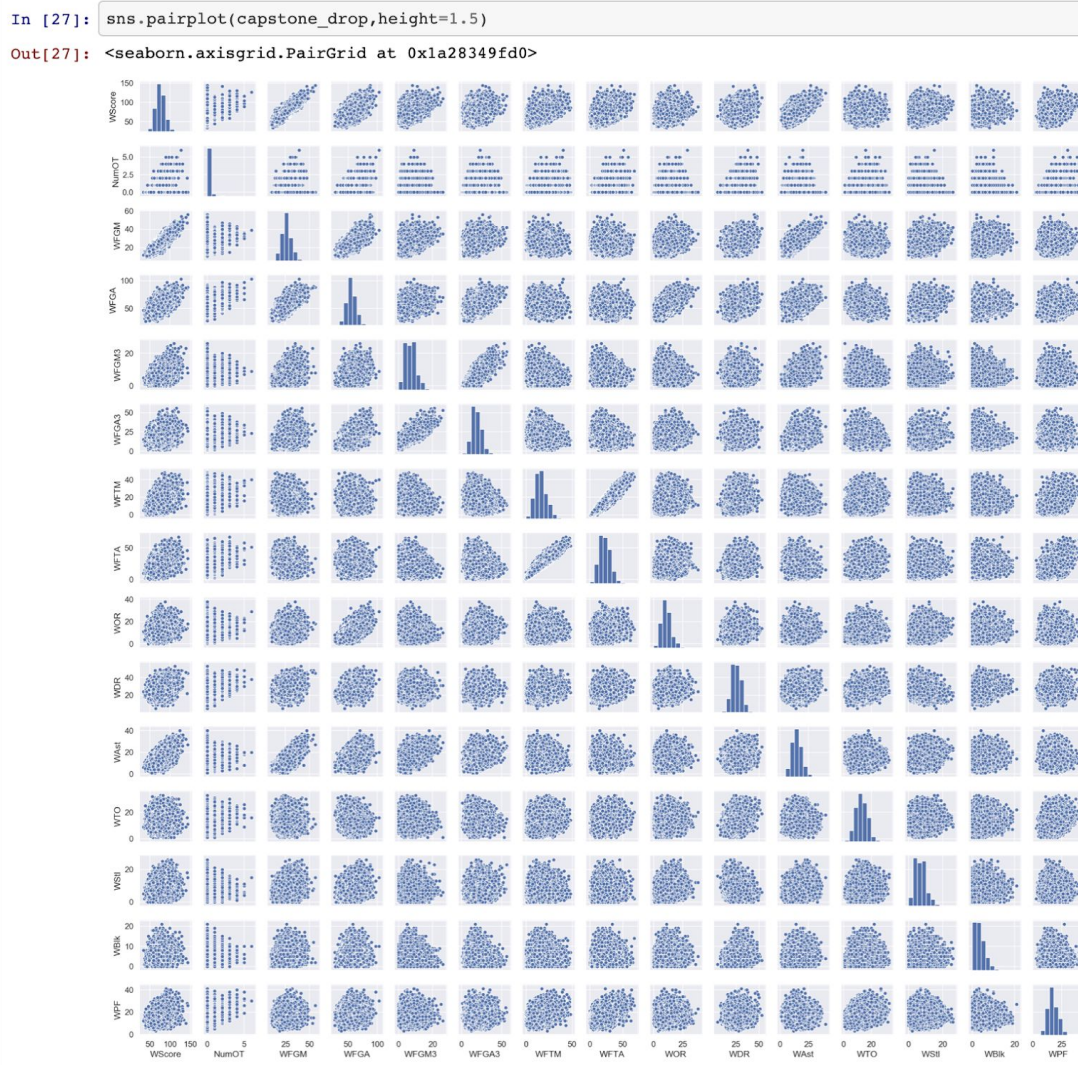
We will reject this null hypothesis for the pairs a & b and a & c (since we are getting a p-value less than 0.05). We are reasonably confident that a pair of datasets is statistically significantly different. After using only ANOVA, we can't make any conclusions on which two populations between the home, away and neutral locations have a significant difference.

There is a significant difference between the pairs home and away games, and the pairs home and neutral games, but there is not a significant difference between the pair away and neutral games.

2. What are the correlations among the game-by-game statistics for a winning team?"

A pair plot allows us to see both the distribution of single variables and relationships between pairs of variables. A visual scan through of the pairs plot shows variables that seem highly correlated. The relationship between (WScore and WFGM), and (WFTA and WFTM) shows what seems to be a strong linear correlation between these variables. This positive correlation was also confirmed by the heatmap.

Fig 10: Pairplot Matrix



3. What is the 95% confidence interval for the difference between the standard deviations of the winning scores for the two top performing teams? This analysis will be done with a bootstrap sampling analysis, calculating these differences over 10000 replicates. The teams being considered will be Gonzaga and Kansas.

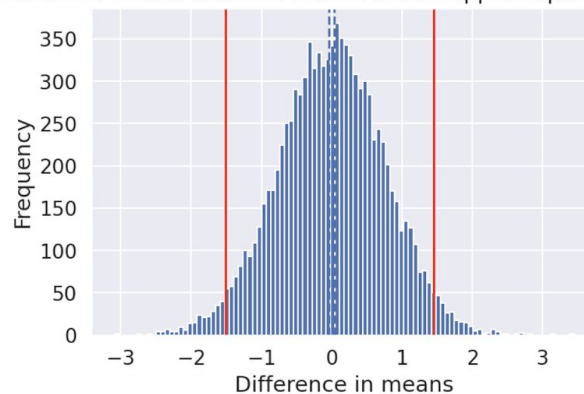
A bootstrap method will be used to compare the means of winning scores of two teams over multiple seasons. The two chosen are Gonzaga and Kansas. These two teams were chosen because they have the highest winning percentages i.e. winning Percentage is defined as (matches won/total matches played *100) . We will proceed by defining both the null and the alternative hypothesis for our calculations

H0: there is no difference in standard deviations in the winning scores between Kansas and Gonzaga

Ha: there is a difference in standard deviations in the winning scores between Kansas and Gonzaga

Fig 11: Bootstrap Replicates

↳ Distribution of differences in means between shifted bootstrapped replicates of Gonzaga and Kansas



The solid red vertical lines correspond to the 95% lower and upper confidence intervals of expected random differences in means of bootstrap replicates of Gonzaga and Kansas samples.

The dashed blue vertical lines correspond to the observed difference in means between Gonzaga and Kansas samples.

Our Null and Alternative Hypotheses were as follows:

H_0 : there is no statistically significant difference in the means of the winning scores between Kansas and Gonzaga

H_a : there is a statistically significant difference in the means of the winning scores between Kansas and Gonzaga

The calculated p value of 0.16 is greater than the significance value of 0.05 . Therefore we fail to reject the null hypothesis, and we cannot say that there is a statistically significant difference in the means of the winning scores between Kansas and Gonzaga.

Our bootstrap replicates with a 95% confidence interval indicate that the difference in means between the two groups have a 95% chance of lying within [-1.4236463156146701 , 1.4045422356365809]. Our calculated difference in means is 0.71. Since the value is within the 95% confidence range, we therefore fail to reject the null hypothesis, and say there is no statistically significant difference in means between the Kansas and Gonzaga winning scores.

Machine Learning Models

Data manipulation for machine learning

Data is collected and organised in two types, (i) game statistics, (ii) season results. Game statistics includes statistics of each team in a game, such as the offense, defense, field goal percentages, rebounds, assists, turnovers etc. Hence, each row in the data represents a single team in a match and its statistics. Season results data files include game-level results, i.e. each row corresponds to a game. The data file stores the day number, home and away teams and their scores. A description of the data file with combined game statistics and season results is displayed in Fig. 13.

Fig. 13. Game Statistics and Season Results

	Season	DayNum	TeamID	Score	OR	DR	Ast	TO	Stl	Blk	PF	Results	identifier	Loc	FGM3Rat	FGMRat	FTMRat
73730	2019	01	alabama	82	16	29	17	20	3	6	24	W	2019_01	H	0.38	0.49	0.68
73731	2019	01	arizona-state	102	20	38	11	17	3	2	26	W	2019_01	H	0.38	0.42	0.59
73732	2019	01	army-west-point	73	6	22	14	6	5	1	22	W	2019_01	H	0.30	0.43	0.69
73733	2019	01	auburn	101	19	23	24	13	9	8	21	W	2019_01	H	0.47	0.51	0.50
73734	2019	01	ball-state	86	16	26	14	9	7	5	18	W	2019_01	H	0.37	0.49	0.75
...
157148	2019	126	eastern-michigan	43	14	18	8	13	6	4	16	L	2019_126	H	0.24	0.29	0.50
157151	2019	127	jackson-state	49	9	21	10	11	11	6	18	L	2019_127	H	0.27	0.32	0.62
157152	2019	127	st-francis-pa	76	15	14	15	10	4	2	16	L	2019_127	H	0.52	0.48	0.24
157162	2019	129	unlv	55	11	32	12	10	2	4	23	L	2019_129	H	0.18	0.32	0.69
157164	2019	131	memphis	58	21	26	4	9	5	5	21	L	2019_131	H	0.17	0.24	0.85

4854 rows x 17 columns

Baseline Modeling: Logistic Regression Model Process

The outcome of home and away matches is a binary classification problem, i.e an observation must be classified as 0 (away win) or 1 (home win) . Based on the symmetry of this problem, a single model will be used to identify features that are important for a home team winning a game. This means that the same model can be used to identify features that are important for an away team to win a game, since if p is an estimation of the probability of a home team winning a game, then $1-p$ is an estimation of an away team winning a game.

The algorithm used to build the machine learning model is Logistic Regression. The algorithm requires the dependent variable (a.k.a the target) to have values denoting the classes being modeled. In this case, there are two classes as indicated above--0 denoting the away team winning and 1 denoting that the home team winning.

Since the models in this project are built under the assumption that each game has a home and away team, the games played in neutral sites will not be used in this project, and the dataset to be used in the model spans the seasons 2003- 2019.

Models built in this project use the ten in-game statistics shown in Fig 12. To model a given game, an aggregation of these statistics for prior games are computed, for both the home and away teams. The lagging period used for the prior game statistics to model a game is two games. This lagged aggregation method is used to avoid “leaking” information into the model that belongs to “the future”.

Fig 12: Features used for the analysis

Acronym	Description
Stl	Steals per team
Ast	Assists per team
TO	Turnovers per team
Blk	Blocks per match per team
PF	Personal fouls per team
OR	Offensive Rebounds per team
DR	Defensive Rebounds per match per team
FTMRAT	Field throw percentage
FGMRat	Field goal percentage
FGM3Rat	Three point percentage

To handle games at the beginning of the season, the design decision adopted in this project is to ignore the first two games.

An example of a prior game statistics being used as input to predict future games is illustrated in Fig. 13 and Fig. 14 below. In Fig. 13, the offensive rebound (OR) for “Duke” in row 562 (with value 17.5), is calculated by taking an average of the offensive rebound for rows 92 and 163 (with values of 19.0 and 16.0, respectively) in Fig. 14. The same method is applied for the rest of the features in Figures 13 and 14.

Fig 13: Data frame after two day moving average calculation

	Season	Score	OR	DR	Ast	TO	Stl	Blk	PF	FGM3Rat	FGMRat	FTMRat
92	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
163	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
562	2003.0	98.0	17.5	29.0	18.0	17.5	11.0	8.0	16.5	0.335	0.440	0.715

Fig 14: Original data frame before two day moving average calculation

	Season	DayNum	TeamID	Score	OR	DR	Ast	TO	Stl	Blk	PF	identifier	FGM3Rat	FGMRat	FTMRat
92	2003	019	duke	101	19	29	17	14	11	6	13	2003_019	0.35	0.43	0.67
163	2003	021	duke	95	16	29	19	21	11	10	20	2003_021	0.32	0.45	0.76
562	2003	033	duke	81	16	24	12	13	14	5	23	2003_033	0.17	0.48	0.62

Splitting of dataset for Training and Testing the Models

Dataset splitting allows checking for overfit issues, by splitting data into a training set, which is used to fit our model, and a test set, which is used to confirm that the model (fitted on training data) gives a similar rate of correct predictions on a new but similar data which was not part of fitting the model.

In this project, the games played by a team during 2003 - 2018 was used for training/fitting the model, while the data for games played during the 2019 season was held for testing.

The dataframe in Fig. 15, is split into two subsets, the 2003- 2018 seasons for feature selection and training, and the final season 2019 for testing.

It should be noted that due to the imbalance in the number of home and away games, the classification problem could have been resolved using a stratified train/test split. This splitting technique would have split the dataset into train and test sets, preserving the same proportions of examples in each class as observed in the original dataset.

The classification report and confusion matrix for the training and test dataset is displayed in Fig. 16 and 17.

Fig 15: Dataset with Home games

	Season	DayNum	TeamID	Score	OR	DR	Ast	TO	Stl	Blk	PF	Results	identifier	Loc	FGM3Rat	FGMRat	FTMRat
73730	2019	01	alabama	82	16	29	17	20	3	6	24	W	2019_01	H	0.38	0.49	0.68
73731	2019	01	arizona-state	102	20	38	11	17	3	2	26	W	2019_01	H	0.38	0.42	0.59
73732	2019	01	army-west-point	73	6	22	14	6	5	1	22	W	2019_01	H	0.30	0.43	0.69
73733	2019	01	auburn	101	19	23	24	13	9	8	21	W	2019_01	H	0.47	0.51	0.50
73734	2019	01	ball-state	86	16	26	14	9	7	5	18	W	2019_01	H	0.37	0.49	0.75
...
157148	2019	126	eastern-michigan	43	14	18	8	13	6	4	16	L	2019_126	H	0.24	0.29	0.50
157151	2019	127	jackson-state	49	9	21	10	11	11	6	18	L	2019_127	H	0.27	0.32	0.62
157152	2019	127	st-francis-pa	76	15	14	15	10	4	2	16	L	2019_127	H	0.52	0.48	0.24
157162	2019	129	unlv	55	11	32	12	10	2	4	23	L	2019_129	H	0.18	0.32	0.69
157164	2019	131	memphis	58	21	26	4	9	5	5	21	L	2019_131	H	0.17	0.24	0.85

4854 rows x 17 columns

Fig 16: Confusion Matrix and Classification Report for Training Set

(73389, 10)

```
[[ 2535 22432]
 [ 2181 46241]]
```

[Train Classification Report]

	precision	recall	f1-score	support
0	0.54	0.10	0.17	24967
1	0.67	0.95	0.79	48422
accuracy			0.66	73389
macro avg	0.61	0.53	0.48	73389
weighted avg	0.63	0.66	0.58	73389

[Train] Accuracy score (y_predict_training, ytrain_home): 0.6646227636294267

Fig 17: Confusion Matrix and Classification Report for Test Set

(4854, 10)

```
[[ 225 1488]
 [  10 3131]]
```

```
[Test Classification Report]
              precision    recall  f1-score   support

      0         0.96         0.13         0.23        1713
      1         0.68         1.00         0.81        3141

 accuracy              0.69        4854
 macro avg           0.82         0.56         0.52        4854
weighted avg           0.78         0.69         0.60        4854
```

```
[Test] Accuracy score (y_predict_test, ytest_home): 0.6913885455294603
```

Interpretation

The closeness of the training and the test accuracy score shows that the model will “generalize well”, and that the model will predict well , when new data is presented to it.

Accuracy alone is not a reliable metric to assess the performance of the logistic regression model, a classification report was used to assess the performance of the model.

The classification report for Fig. 16 has 73389 predictions for the home games, and out of the predictions , the confusion matrix predicted 22432+46241 times for home wins, while predicting 2535+2181 wins for away games. While in reality there were 2535+22432 away wins, and 2181+46241 wins in home matches.

However, in the classification report for the test set in Fig. 17 the total the model made 4854 predictions for the home games, out of these predictions, the confusion matrix predicted wins **1488+3131** times for home games, while predicting **225+10** wins for away games. While, in reality there were **225+1488** away wins, and **10 + 3131** wins in home matches.

The precision value for the minority class (0) tells us how often our model is correct when predicting away wins, while the recall value for the majority class (1) tells us out of all the home wins, how many did our model correctly identify.

For the minority class of the classification report for both the training and test data, a high precision value and a low recall value was obtained. This is basically saying that the model is correctly predicting a high percentage of the time. However, of all the away wins in the dataset, the model did not catch many of them, hence the low recall value.

Extended Modeling: Feature Selection by Recursive Feature Elimination (RFE)

The original model is extended by using a feature selection algorithm called Recursive Feature Elimination (RFE). The feature selection was done to reduce the number of features needed to predict the outcome of matches, while also maintaining the accuracy of the model. The feature selection process would also tell us which features contribute most to the outcome of a home team winning or losing a match. The RFE selection process works by using the model accuracy to identify which combination of attributes contribute most to predicting the outcome of matches i.e. home win or away win. The RFE algorithm works by fitting the machine learning algorithm used in the model (**logistic regression**), ranking features used in the model by importance, and discarding the least important features, and re-fitting the model ⁴. The description of the features used for the data modeling is in Fig. 12. From these features, the RFE algorithm was used to select the five most important features contributing to the outcome of matches for home games below. The result of the feature selection method can be seen in the classification report of Fig. 18. These features were assigned a value of 1 and displayed in Fig. 19

⁴ "Recursive Feature Elimination (RFE) for Feature Selection in" 25 May. 2020, <https://machinelearningmastery.com/rfe-feature-selection-in-python/>. Accessed 14 Jul. 2020.

Fig 18: Confusion Matrix and Classification Report for Test Set, after building model with the features selected by RFE

```
[Confusion Matrix for Test Data using RFE]
[[ 247 1466]
 [  15 3126]]
```

```
[Test Classification Report]
              precision    recall  f1-score   support

     0           0.94         0.14         0.25         1713
     1           0.68         1.00         0.81         3141

 accuracy              0.69         0.69         0.69         4854
 macro avg           0.81         0.57         0.53         4854
weighted avg           0.77         0.69         0.61         4854
```

```
[Test] Accuracy score (y_predict_test, ytest_home): 0.6948908117016893
```


Comparing Full Model vs. Reduced Model

Comparing the performance metrics from Fig. 17 and 18, the accuracy score was slightly improved (0.694 vs. 0.691) when the top five features (reduced model) were used to predict the classifier accuracy.

The full model does a better job predicting away wins versus the reduced model. This is because the precision value for the minority class for the full model is 0.96 versus 0.94 for the reduced model. This shows us that the original 10 features are better predictors of away wins.

These top five features as denoted by the ranking by the ranking of 1 in the RFE algorithm in Fig. 19 are : FGM3Rat, FGMRat, OR, FTMRat and BLK.

Fig 19: Feature Rankings for Home Games



	Feature	Ranking
0	OR	1
5	Blk	1
7	FGM3Rat	1
8	FGMRat	1
9	FTMRat	1
3	TO	2
4	Stl	3
1	DR	4
2	Ast	5
6	PF	6

Conclusions

Logistic Regression was used to build models to predict the outcome of matches in NCAA basketball games for 351 different teams. More specifically, the models estimate the probability of a home team winning a match and an away team winning a match. Data was partitioned into training and test for model validation. The training set included data from 2003-2018, and the test data was for the 2019 season. A performance matrix was generated both for the training and test data.

The global accuracy prediction for the test and training data was sixty nine percent for model prediction. However, for the performance matrix for both the training and test data, a low recall value was obtained for the minority class (away wins) versus the recall value obtained for the majority class (home wins).

This implies that the model is good when predicting when the home team is going to win, and not good at predicting when an away team is going to win

From the perspective of improving the global accuracy of the Logistic Regression models using RFE, the most important features are: field goal percentage made, field throw percentage made, free throw percentage made, blocks and offensive rebounds. This **does not** mean that these features are the most important ones with respect to a home team winning a game.

In order to determine which features contribute the most to the home teams winning a game, through a Logistic Regression (LGR) model, would be to analyze the coefficients of the LGR models.

Recommendations For the Client

Based on the information provided by the model on rebounds and turnovers, coaches can place more emphasis on the top five features selected by the RFE algorithm when playing home games. Players that collect a lot of offensive rebounds and blocks are given more minutes when playing home games, since these features are important in determining the outcome of games at home.

A coach can therefore,

- Rotate his team in such a way that players that possess the top five features be played more minutes when playing at home.
- Use the five features selected from the overall set of features when developing strategies for home games. This is because the accuracy of the model does not decrease when used to predict the performance of the model for home matches. The coach can focus on these top five features when trying to increase his advantage in winning matches when playing games at home.

Future Work

There were few key limitations that future research might be able to determine for this model and some of these are:

- Determination of the most important features for the dataset as identified by the logistic regression model.
- Trying different rolling average windows to build the features used to model each game. When trying different rolling average windows, an optimum rolling average could be determined that could help improve the accuracy of the models. Buursma ⁵ followed this process for the sports of football, and found, through experimentation, that using an average across the 20 matches resulted in the best classification accuracy
- Trying different models and seeing the effect of these models on the performance metrics. The project was limited to exploring the logistic regression algorithm. It would be interesting to see how other classification methods such as Vector Machine, Random Forest , etc would perform on the same data using the same variables.
- Including players' information and features, when creating models so as to improve accuracy of prediction.
- Inclusion of games played in March Madness. Since our training and test data only pertained to regular season games only, and not the end of the year tournament(i.e. March Madness), performance metrics could be different from those collected during the regular season.

⁵ D. Buursma, Predicting sports events from past results "Towards effective betting on football matches", in: Conference Paper, presented at 14th Twente Student Conference on IT, Twente, Holland, 21 January 2011, 2001

Works Cited

- Bunker, Rory P., and Fadi Thabtah. "A Machine Learning Framework for Sport Result Prediction." *Applied Computing and Informatics*, No Longer Published by Elsevier, 19 Sept. 2017, www.sciencedirect.com/science/article/pii/S2210832717301485.
- "The Home Court Advantage: Evidence from Men's College Basketball." *The Sport Journal*, 13 Feb. 2017, thesportjournal.org/article/the-home-court-advantage-evidence-from-mens-college-basketball/.
- "Home-Field Advantage (SOCIAL PSYCHOLOGY) - IResearchNet." *Psychology*, 21 Jan. 2016, psychology.iresearchnet.com/social-psychology/control/home-field-advantage/.
- Li, Susan. "Building A Logistic Regression in Python, Step by Step." *Medium*, Towards Data Science, 27 Feb. 2019, towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8.
- Markham, Kevin. "In-Depth Introduction to Machine Learning in 15 Hours of Expert Videos." *R*, 24 Sept. 2014, www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/.
- Ncaa. "NCAA Basketball." *Kaggle*, 20 Mar. 2019, www.kaggle.com/ncaa/ncaa-basketball.
- D. Buursma, Predicting sports events from past results "Towards effective betting on football matches", in: Conference Paper, presented at 14th Twente Student Conference on IT, Twente, Holland, 21 January 2011, 2001.