Springboard

Data Science Career Track

Capstone Project 1

What are the factors that contribute to a home team winning an NCAA basketball game?

Milestone Report

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. This report tries to analyse some of the differences between playing on the road and playing at home.

Logistic regression algorithm was used as a supervised machine learning model to predict outcomes (win or losses) in home and away matches for 2019, when given in game statistics for NCAA basketball games from 2003-2018.

Ten ingame statistics (features) were averaged from prior games and used as input into models to predict outcomes of future games. A rolling average technique was done to reduce data leakage in the model.The lagging period used for the past average game statistics to predict upcoming matches was two days. The top five ranked features using a feature selection algorithm were then used to predict the outcome of matches for home and away matches.

A link to the dataset and the code to the account is attached in the link below.

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

In summary,

- Home court does give an advantage in basketball games played in the NCAA.
- The algorithm predicted with up to seventy percent accuracy for home and away games outcomes when selecting all the features for the model.
- When using top five ranked features, as determined by the feature selection algorithm, the accuracy of the model stayed the same for the home matches, but declined to forty eight percent for away games.
- It is advantageous for coaches to field players that have a higher tendency to collect offensive rebounds (taller and more athletic players), than players that are smaller and nimble when playing games at home.

Abstract **2**

Introduction **6**

Objective **6**

Dataset **6**

Data Cleaning & Wrangling **8**

Data Type **8**

Handling Missing Data **9**

Exploratory Data Analysis **9**

Box Plot - points scored by winning team at playing locations 11

Box Plot - points scored by losing team at playing locations 12

Bar Chart - Proportion of games played at different locations in all seasons 13

Line Chart - Points scored per winning team per season 15

Line Chart - Three-Pointers scored per winning team per season 16

Line Chart - Turnovers conceded per winning team per season 17

Bar Chart - Ranking top 15 teams by win percentage for 2003 -2019 18

Bar Chart - Ranking top 15 teams by win percentage for 1985 -2019 19

Applications of Inferential Statistics **20**

ANOVA test & Tukey HSD 21

Heat map with correlation matrix and p value 22

Pair plot matrix 23

Bootstrap Replicates 25

Data manipulation for machine learning 27

Fig: Winning location of matches for home and away games 28

Baseline Modeling: Logistic Regression Model Process **28**

Handling Categorical Data 29

Table :Features used for the analysis 30

Fig : Initial Dataframe 31

Fig : Final Dataframe 31

Splitting of data for Train Test Split 32

Summary of training and test dataset for home games 32

Summary of training and test dataset for away games 32

Fig : Training data set with first two matches averaged out. 33

Extended Modeling: Feature Selection by Recursive Feature Elimination **33**
Implementing the model 36
Findings **36**
Conclusions and Future Work **39**
Recommendations **40**
Works Cited **41**

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 matches[2].

In light of the outbreak of the coronavirus, the usual benefit of home field advantage presented by the home crowd is now neutralized with the presence of guards, engineers, medical personnels and team mates. The benefit of a model that would assist a coach and also influence the outcome of a match would be very important.

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 when playing at an away location.

Dataset

The datasets for this analysis will be obtained from Kaggle[3].These data is associated with an annual competition sponsored by Google.

---

[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 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 games 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.

Regular SeasonCompact Results.csv

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.

Teamspellings.csv

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 our 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 Type

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

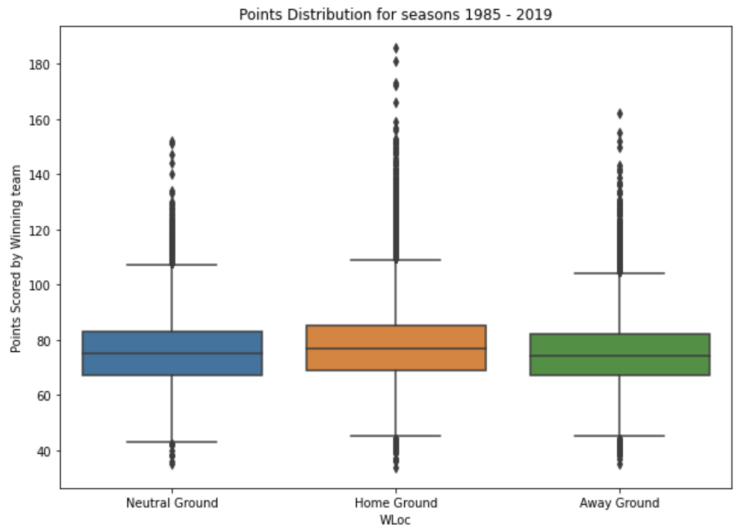
Following the data wrangling of the data, the next step is to interrogate the dataset and ask a series of questions of the dataset. These questions will help identify the contributing factors affecting home games winning matches. The questions being asked of the data are :

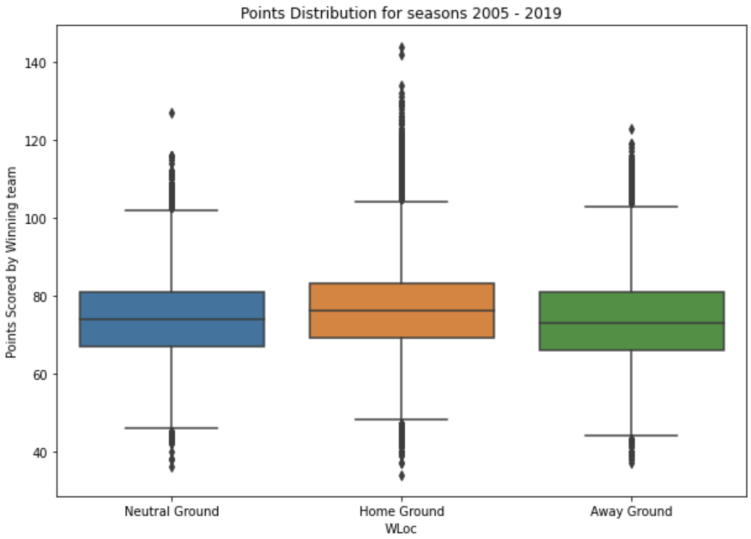
1. Does a winning team score more points when playing at home, than when playing at either a neutral ground or an away ground?
2. Does a losing team score more points when playing at home, than when playing at either a neutral ground or an away ground?
3. Is there a difference in the amount of matches 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?

Box Plot - points scored by winning team at playing locations





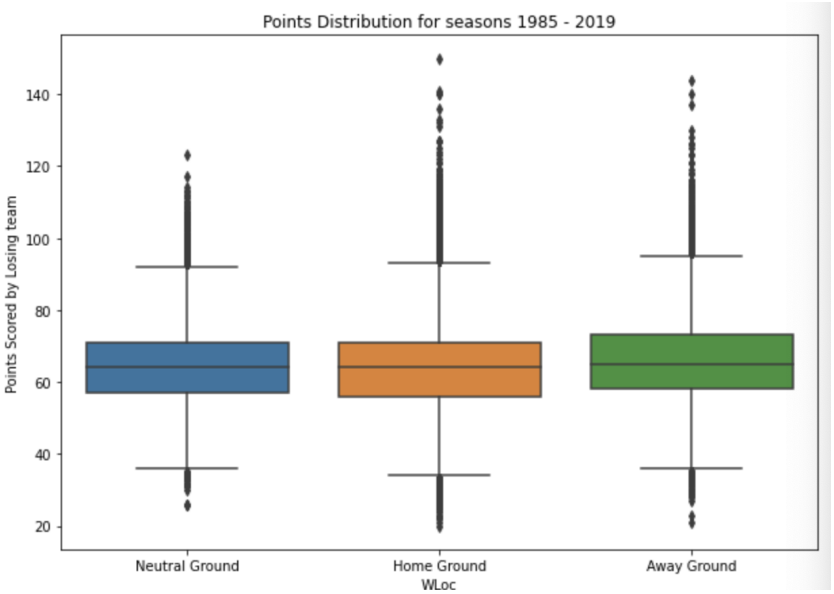
Conclusion

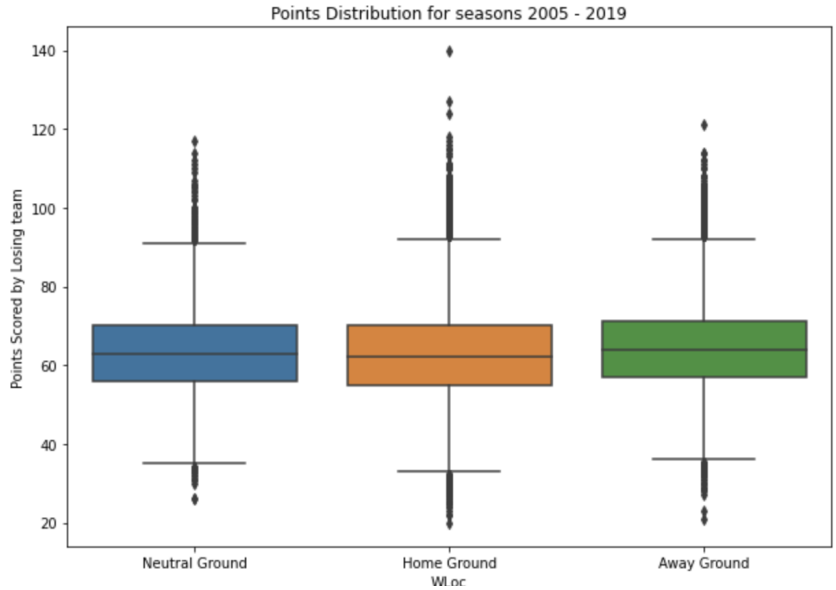
The box plots show that the average number of points for the winning team at their home ground is higher than the points scored when playing at either away or neutral grounds. This result will

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?

Box Plot - points scored by losing team at playing locations





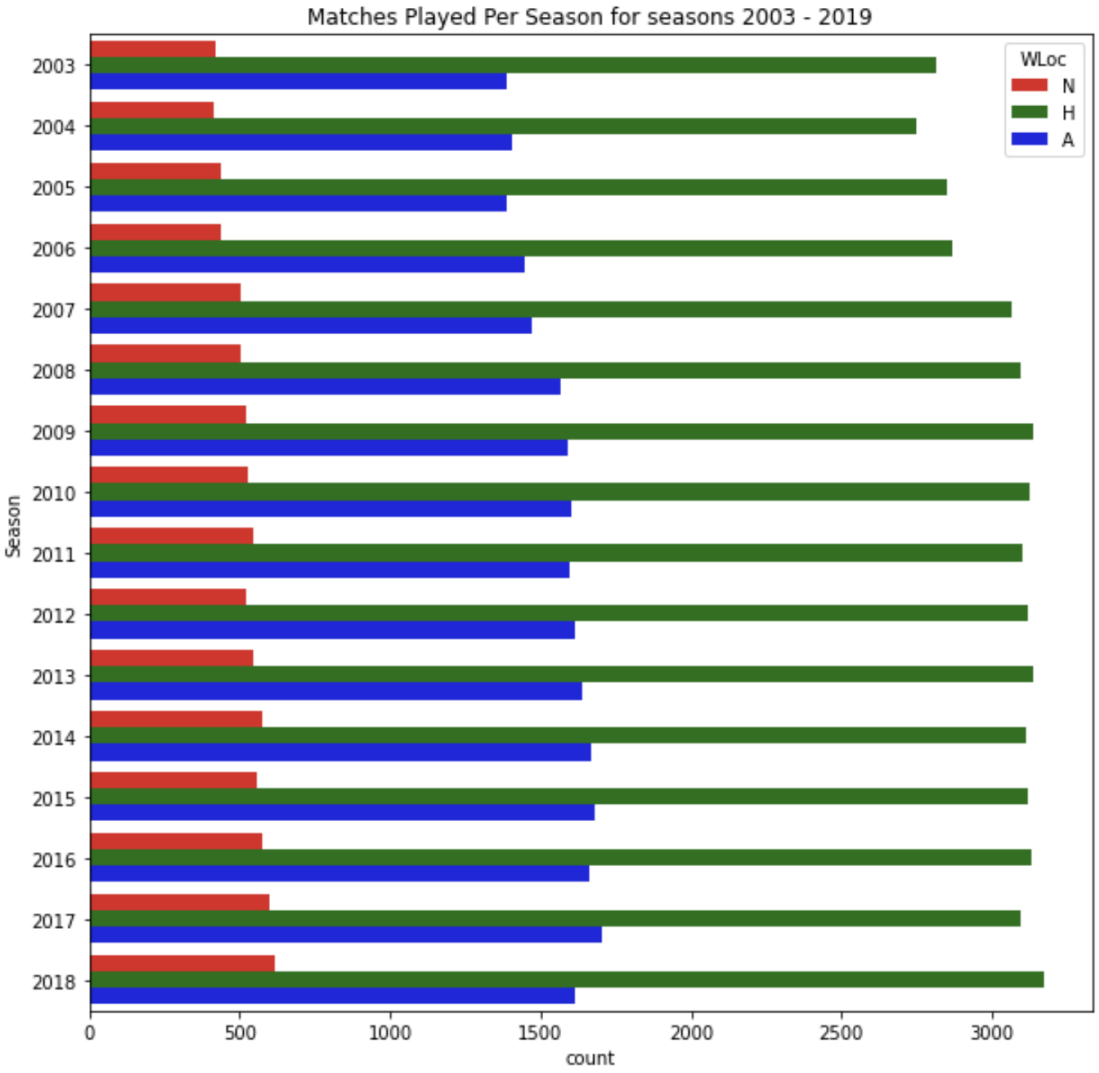
Conclusion

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.

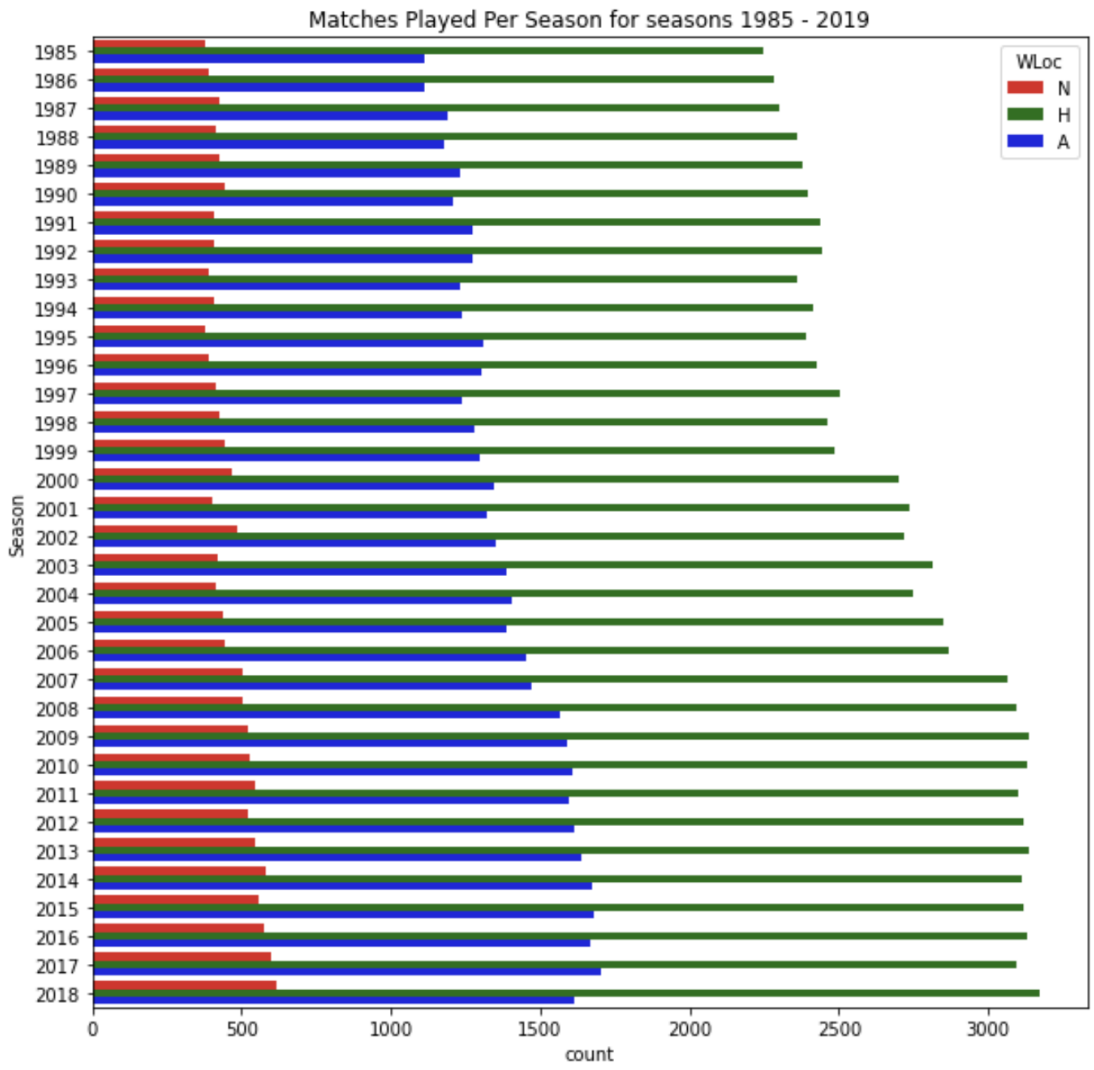
3. What is the difference in the amount of games a team wins either home, away or at a neutral location?

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

[19]



[20]

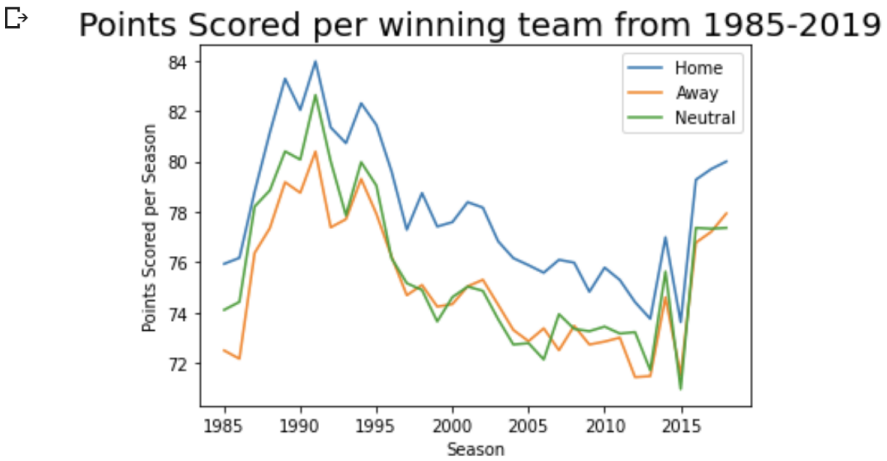


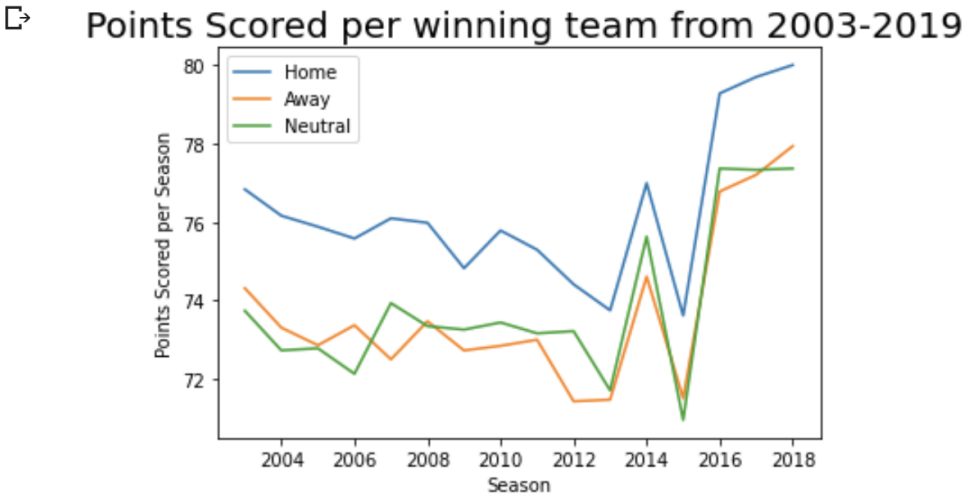
Conclusion

The proportion 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?

Line Chart - Points scored per winning team per season





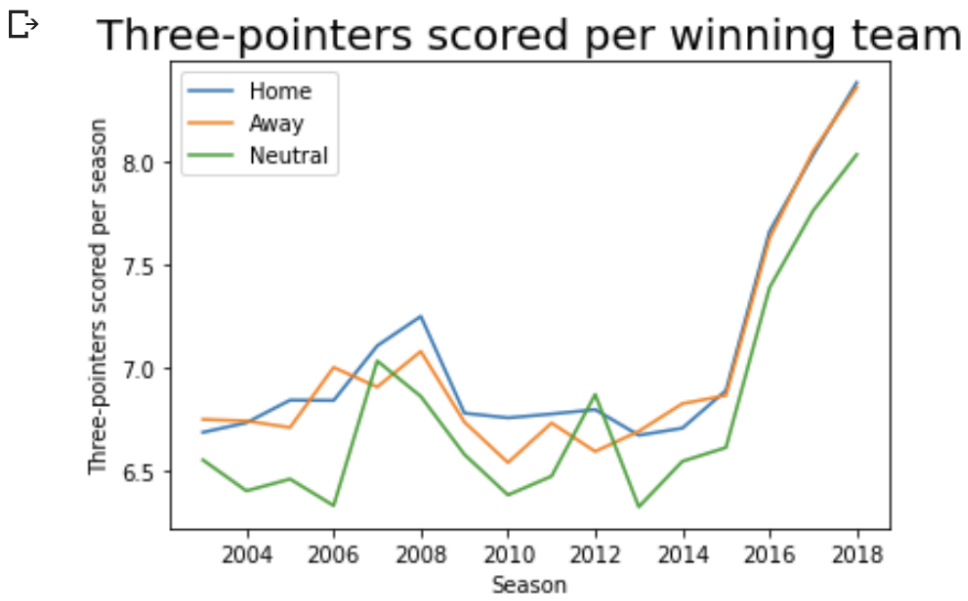
Conclusion

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 average variation in three-pointers scored by the winning team when playing at home, away or a neutral location per season?

Line Chart - Three-Pointers scored per winning team per season



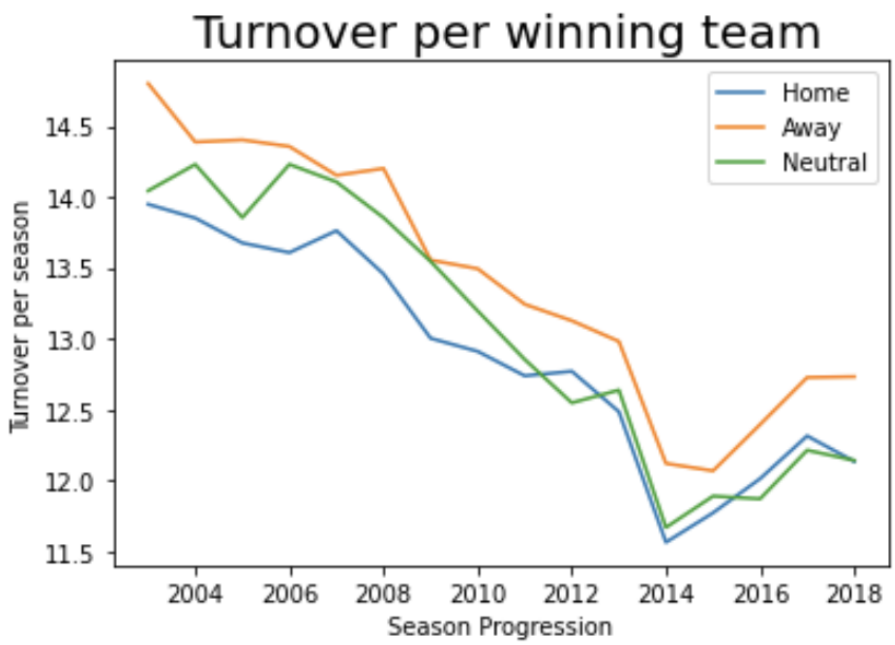
Conclusion

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 average amount of turnovers per season?

Line Chart - Turnovers conceded per winning team per season

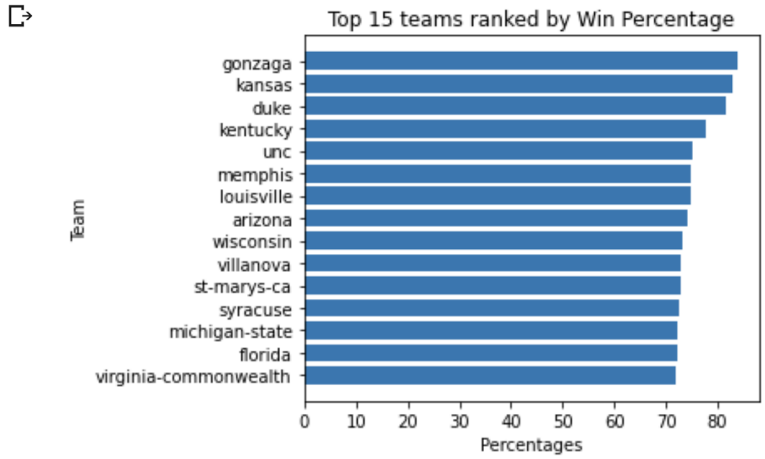


Conclusion

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 is the ranking of the top 15 teams based on the winning percentage per season for seasons 2003 -2019? Winning Percentage is defined as (matches won/total matches played *100)*

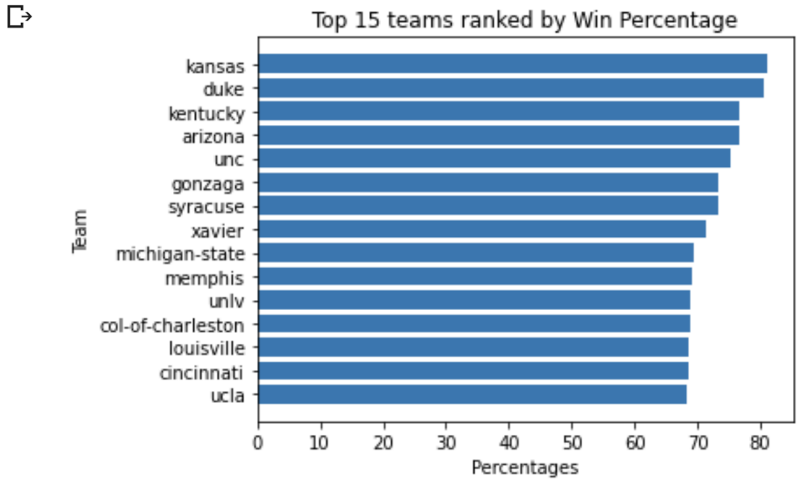
Bar Chart - Ranking top 15 teams by win percentage for 2003 -2019



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

*8. What is the ranking of the top 15 teams based on the winning percentage per season for seasons 1985 -2019? Winning Percentage is defined as (matches won/total matches played *100)*

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 was:

1. Is the difference between a team winning college basketball matches at home locations versus playing at away or neutral locations 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. What is the correlation between game-by-game data for a winning team? A heat map will be used to show the correlations between the game by game data.The correlation matrix from the heat data will show the correlations 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 have the same mean. 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**.

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.

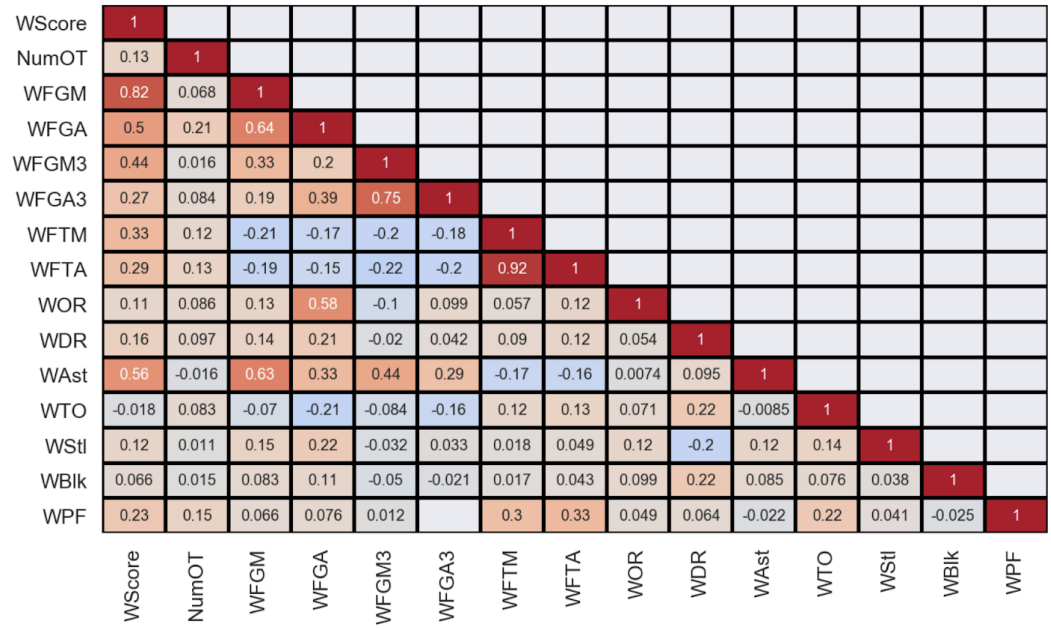
2. What is the correlation between game-by-game data for a winning team?

A heat map analysis was conducted on the data to find correlations between the game by game data. The correlation matrix from the heat map data displayed the correlations that existed between in game statistics for the **winning** team during the season.

This heat map displays the correlation coefficient, and also excludes correlations that have a p-value less than 0.05

Heat map with correlation matrix and p value

In the heat map displayed below orange means positive, and blue means negative. The stronger the color, the larger the correlation magnitude.



The heatmap can be used to investigate possible collinearity among the multiple variables in the dataset. The **.corr()** function was used in the code to show the correlation between the values. This is where we want to set our independent or target variable. Our target variable is "*WScore*". This is the number of points scored by the winning team. We want to find out how all of the other variables affect the points scored by the winning team. In the heatmap, the dark red areas

represent a positive correlation, while light blue represents a negative correlation. It is also normal that the darkest areas are a 1:1 ratio since WScore=WScore, NumOT=NumOT, etc.

While *WScore* is still our independent variable, we can see in the map below that there is little to no correlation between the WTO (-0.018), though a high correlation between WFGM and WAst (0.82, 0.56). these relationships are obvious (assists in a basketball game and field goals made positively correlates with the scores in a match)

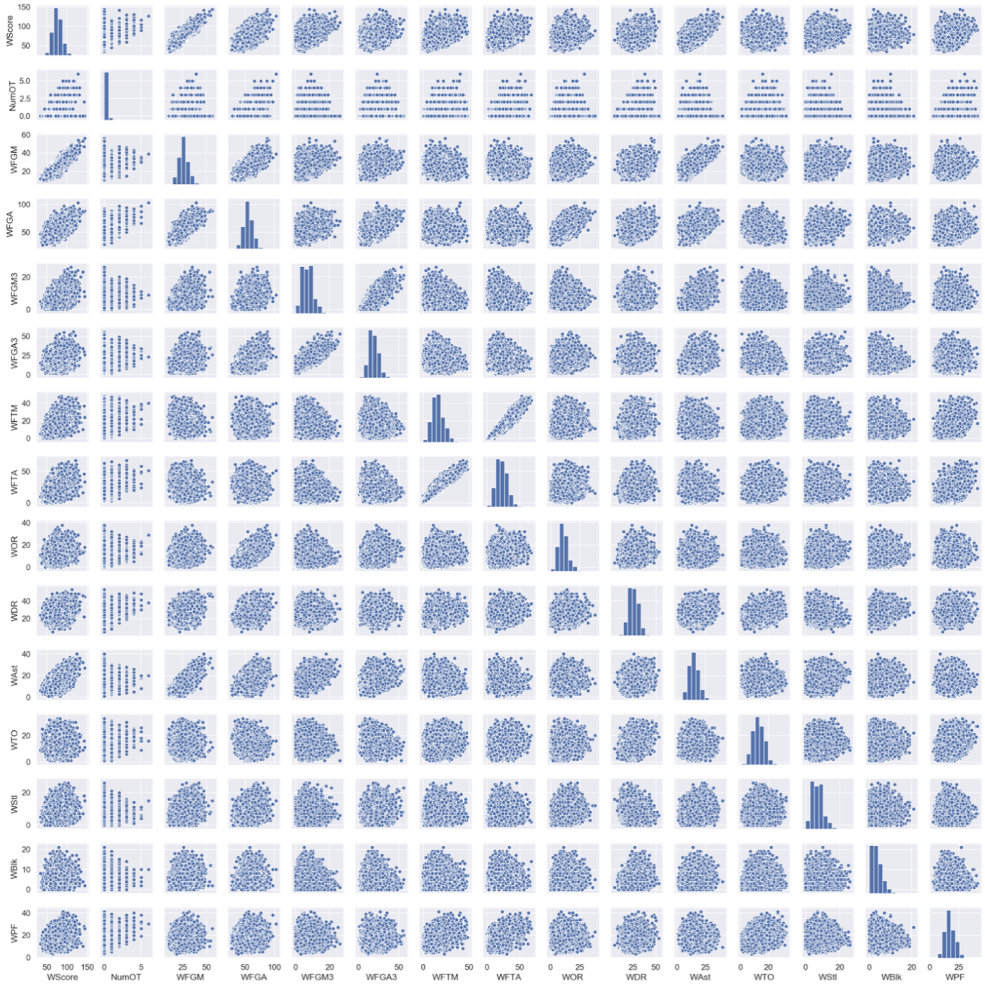
Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. So, when two features have high correlation, we can drop one of the two features. In order to decide to find which features to drop when we decide to create our machine learning model, we compare the relationship between WFGM and WAst, and we get a correlation number of (0.63). Dropping either of these variables when tuning our eventual machine learning model might produce a more accurate prediction model.

In addition to plotting the correlation coefficients in the heat map, only significant p-value correlations (alpha = .05) were plotted. This was achieved with the def corr_sig function. It can be seen from the plot that the relationship between WFGA3 and WPF was not statistically significant.

Pair plot matrix

A pairs 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.

```
In [27]: sns.pairplot(capstone_drop,height=1.5)
Out[27]: <seaborn.axisgrid.PairGrid at 0x1a28349fd0>
```



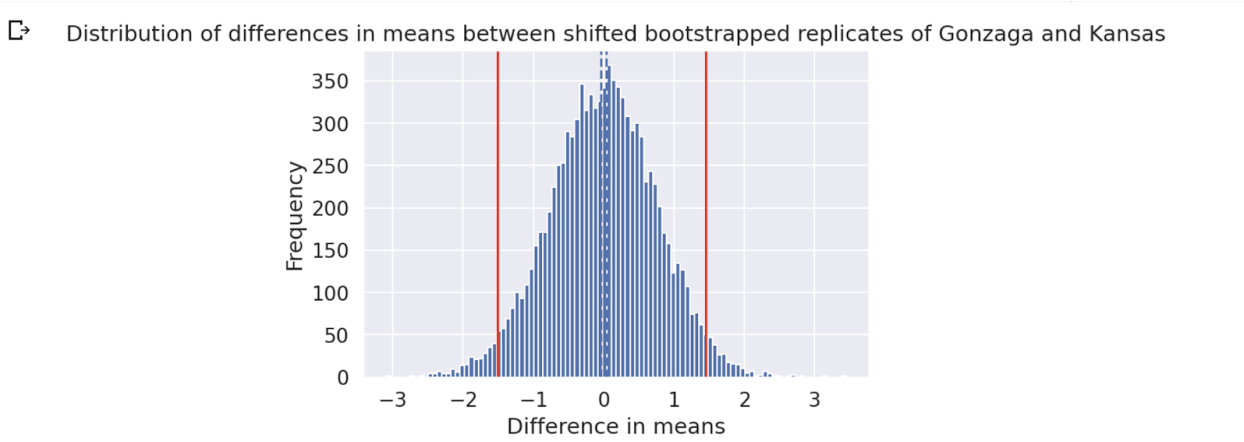
3. What is the 95% confidence interval for the difference between the standard deviations of the winning scores for the two top performing teams? This analysis will be done with a bootstrap sampling analysis, calculating these differences over 10000 replicates. The teams being considered will be Gonzaga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

Bootstrap Replicates



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:

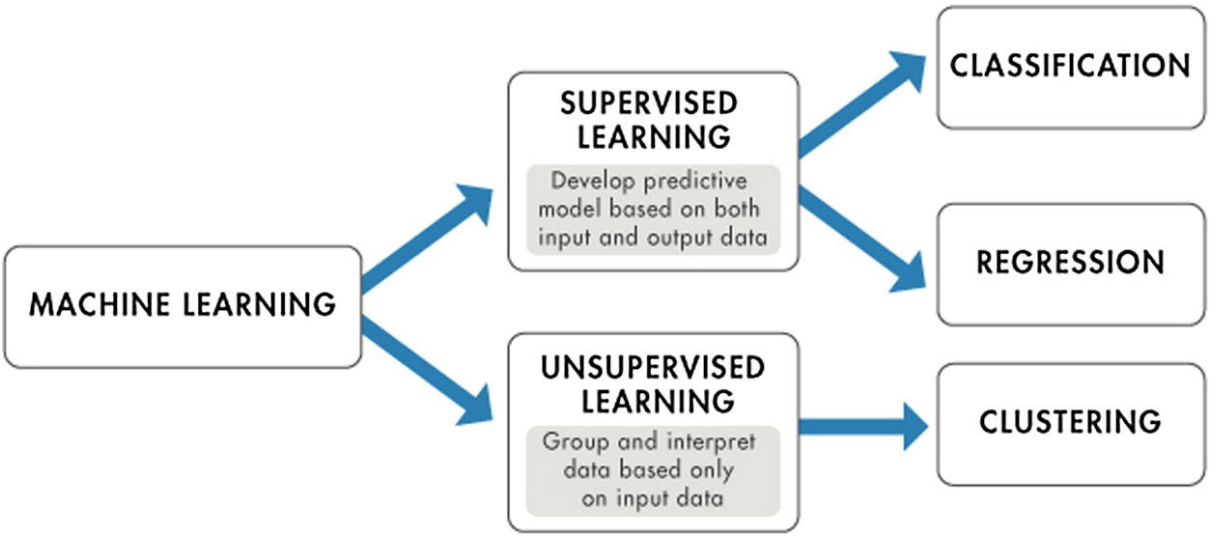
H_0 : there is no statistically significant difference in the means of the winning scores between Kansas and Gonzaga

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 i.e. 0.16>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



The dataset provided is distributed in the following manner:

Home games: 51,825 (59 percent of games played)
Away games: 26,759 (31 percent of games played)
Neutral games: 8,920 (10 percent of games played)

Data manipulation for machine learning

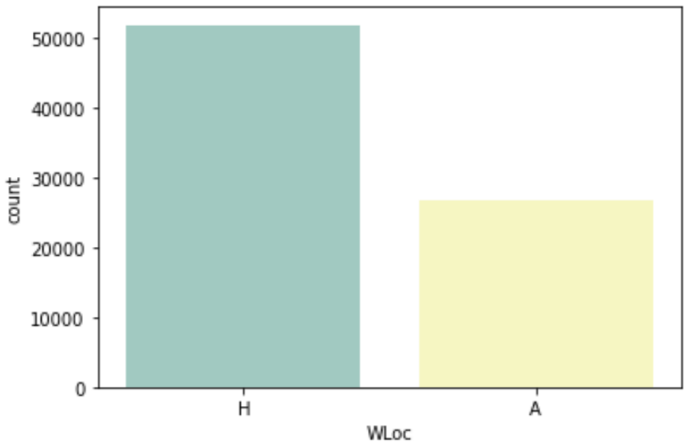
The algorithm used to build the machine learning model is a logistic regression one. The algorithm requires the dependent variable to be part of a binary variable; this is the home and away location of the winning and losing teams. The location of games played in neutral sites will be dropped.

The ratio of matches won at home and away games are displayed with the seaborn count plot in Fig. *Winning location of matches for home and away games*

Fig: Winning location of matches for home and away games

```
sns.countplot (x="WLoc",data =capstonem,palette="Set3")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc357c18588>
```



Percent of matches won at home location: 0.66
Percent of matches won at away location: 0.34

From the plot we can observe that there were more matches won at home locations than away locations.

Baseline Modeling: Logistic Regression Model Process

In order to predict the final scores for the NCAA games, the original data frame was appended , reshaped and arranged so that it can be used for the machine learning model.

A dataset containing all individual game statistics (such as shooting percentages, rebounds, number of turnovers , steals) was used to predict the outcome of final match results, regardless

of whether the team is a home team or an away team. The order of the matches played per individual teams were preserved when manipulating the data .

Since in-game statistics for NCAA games are not usually known until a match has been played, historical in game data (features) are averaged out , and used as features for upcoming matches. Match-related features undergo a separate averaging process before being re-merged with the dependent variable of win and losses (Fig 4) . The dataframe created from this manipulation is then subdivided into two separate dataframes of home and away teams.

The features for the home wins are ranked according to how well they predicted wins and losses of matches. This same step is carried out separately on features for the away team.

Handling Categorical Data

The categorical features from the dataframe were converted to a numerical format.These allowed the machine learning algorithms to handle the features effectively. The conversion to numerical data involved assigning a value of 1 for home games,and a value of 0 for away games. Likewise a value of 1 was assigned for wins and 0 for losses. A sample of this data frame is displayed in

Fig :Training data set with first two matches averaged out.

The model performance was evaluated by classifying match results into home and away. The model was then assessed based on the number of matches that was correctly identified, using a standard classification matrix.The description of the features used for the data modeling is in Table features used for the analysis.

Table :Features used for the analysis

Acronym	**Description**
FGM	Field Goal made
FGA	Field Goal attempts
FTM	Free Throw made
FTA	Free Throw attempts
FGM3	3PT shooting field goal made
FGMA3	3PT shooting field goal attempts
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
Loc	Location of match played
Result	Win or loss
identifier	Unique identifier created from a union of season and DayNum

The initial and the final form of the data frame is shown in the two figures below.

Fig : Initial Dataframe

	Season	DayNum	WTeamID	WScore	LTeamID	LScore	WLoc	NumOT	WFGM	WFGA	WFGM3	WFGA3	WFTM	WFTA	WOR	WDR	WAst	WTO	WStl	WBlk	WPF	LFGM	LFGA	LFGM3
0	2003	11	1458	81	1186	55	H	0	26	57	6	12	23	27	12	24	12	9	9	3	18	20	46	3
1	2003	12	1161	80	1236	62	H	0	23	55	2	8	32	39	13	18	14	17	11	1	25	19	41	4
2	2003	12	1458	84	1296	56	H	0	32	67	5	17	15	19	14	22	11	6	12	0	13	23	52	3
3	2003	13	1166	106	1426	50	H	0	41	69	15	25	9	13	15	29	21	11	10	6	16	17	52	4
4	2003	13	1323	76	1125	48	H	0	25	56	10	23	16	23	8	35	18	13	14	19	13	18	64	8

Fig : Final Dataframe

	Season	DayNum	TeamID	Score	FGM	FGA	FGM3	FGA3	FTM	FTA	OR	DR	Ast	TO	Stl	Blk	PF	Results	identifier	Loc	FGM3Rat	FGMRat	FTMRat
0	2003	011	1458	81	26	57	6	12	23	27	12	24	12	9	9	3	18	W	2003_011	H	0.50	0.46	0.85
1	2003	012	1161	80	23	55	2	8	32	39	13	18	14	17	11	1	25	W	2003_012	H	0.25	0.42	0.82
2	2003	012	1458	84	32	67	5	17	15	19	14	22	11	6	12	0	13	W	2003_012	H	0.29	0.48	0.79
3	2003	013	1166	106	41	69	15	25	9	13	15	29	21	11	10	6	16	W	2003_013	H	0.60	0.59	0.69
4	2003	013	1323	76	25	56	10	23	16	23	8	35	18	13	14	19	13	W	2003_013	H	0.43	0.45	0.70
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
157163	2019	130	1416	55	19	53	6	20	11	19	8	22	7	16	6	6	21	L	2019_130	A	0.30	0.36	0.58
157164	2019	131	1272	58	16	68	4	23	22	26	21	26	4	9	5	5	21	L	2019_131	H	0.17	0.24	0.85
157165	2019	131	1420	49	15	44	8	26	11	14	4	23	9	13	6	4	19	L	2019_131	A	0.31	0.34	0.79
157166	2019	131	1343	77	28	62	6	24	15	20	9	24	9	10	4	1	17	L	2019_131	A	0.25	0.45	0.75
157167	2019	132	1217	85	28	62	10	32	19	24	12	15	7	9	1	2	22	L	2019_132	A	0.31	0.45	0.79

As can be seen in the final dataframe, the final row count is now 157168, up from the original count of 78584. The original data frame which had the result of the matches (win, loss), and winning location, all in one row, is now separated such that the result of matches and location of matches are its own separate columns.

Splitting of data for Train Test Split

The data was originally split into an x-array (which contained the data that we will use to make predictions), and a y-array (which contained the data that was to be predicted).

The x-array are the in game statistics for each NCAA team, while the y-array is the win and losses for each team.

This data frame was later sub divided into two sets: **training data set**, and the **test dataset**. The training data was chosen to be from **seasons 2003 -2018**, while the test data for the model was chosen to be the **2019 season**. The reason for this is that features for the model would be selected based on the information on the training set, not on the whole data set. The test set was kept separate from the training set so as to evaluate the performance of the feature selection. By also separating the training dataset from the test dataset, it would be possible to show how accurately the models could predict a given test data.

The effect of feature selection was measured with the aid of a classification report. This was done by comparing the effect of removing features from the model and seeing the effect on the classifier accuracy.

Summary of training and test dataset for home games

Training dataset shape: (73389, 10) , Testing dataset shape: (4854, 10)

Summary of training and test dataset for away games

Training dataset shape: (73389, 10) , Testing dataset shape: (4854, 10)

Fig : Training data set with first two matches averaged out.

	Season	DayNum	TeamID	Score	OR	DR	Ast	TO	Stl	Blk	PF	identifier	FGM3Rat	FGMRat	FTMRat	Results	Loc
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1	1
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1	1
123	2003.0	11.5	1458.0	82.5	13.0	23.0	11.5	7.5	10.5	1.5	15.5	2003011.5	0.395	0.470	0.820	1	1
360	2003.0	15.5	1458.0	83.5	13.5	26.5	13.0	9.5	7.0	2.5	14.5	2003015.5	0.415	0.465	0.800	1	0
79088	2003.0	22.5	1458.0	76.0	13.0	27.5	12.0	10.0	6.0	5.0	12.5	2003022.5	0.395	0.435	0.790	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
151662	2018.0	101.5	1101.0	72.5	16.0	23.0	11.5	15.0	7.5	3.5	21.0	2018101.5	0.260	0.460	0.695	0	0
151808	2018.0	106.5	1101.0	67.0	16.0	23.5	12.5	16.0	7.0	3.5	19.5	2018106.5	0.325	0.420	0.675	0	1
151920	2018.0	112.0	1101.0	60.5	12.0	22.0	8.0	15.5	7.5	1.5	16.5	2018112.0	0.220	0.385	0.670	0	1
152137	2018.0	115.5	1101.0	69.5	11.0	22.0	9.5	13.0	9.5	3.0	20.0	2018115.5	0.190	0.450	0.565	0	0
73621	2018.0	119.0	1101.0	64.0	11.5	23.0	14.5	16.5	8.5	3.5	19.5	2018119.0	0.300	0.450	0.475	1	0

Extended Modeling: Feature Selection by Recursive Feature Elimination

The feature selection method used for the project was the Recursive Feature Elimination (RFE). Recursive Feature Elimination (RFE) uses an idea to repeatedly construct a model based on the best or worst performing feature. These features are then set aside with the algorithm repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features.The final features would then be ranked based on the algorithm.
A correlation matrix was also created to check the correlation between ingame statistics.

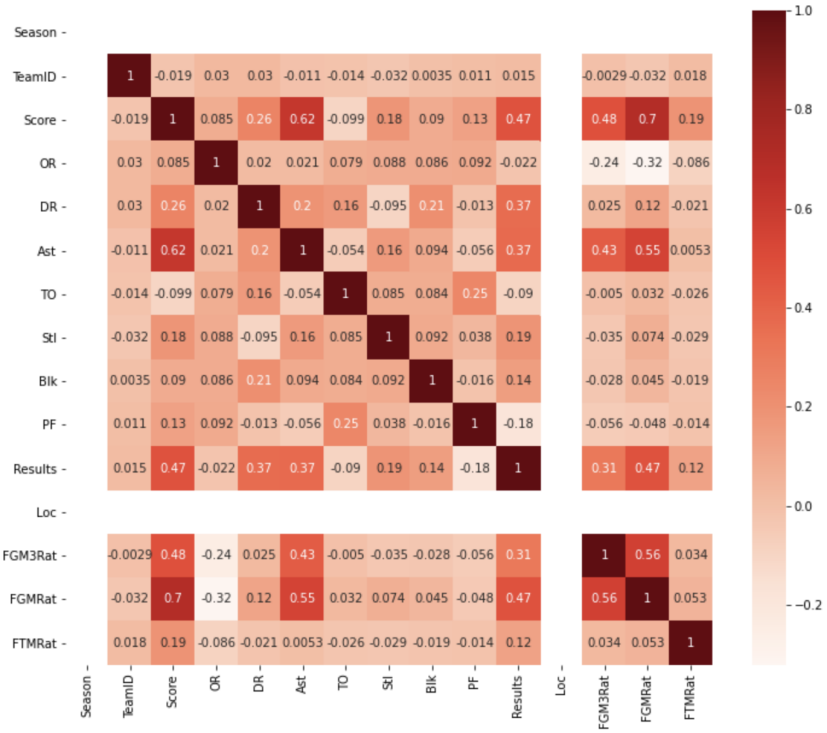
Table: Feature Ranking for Home games

	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

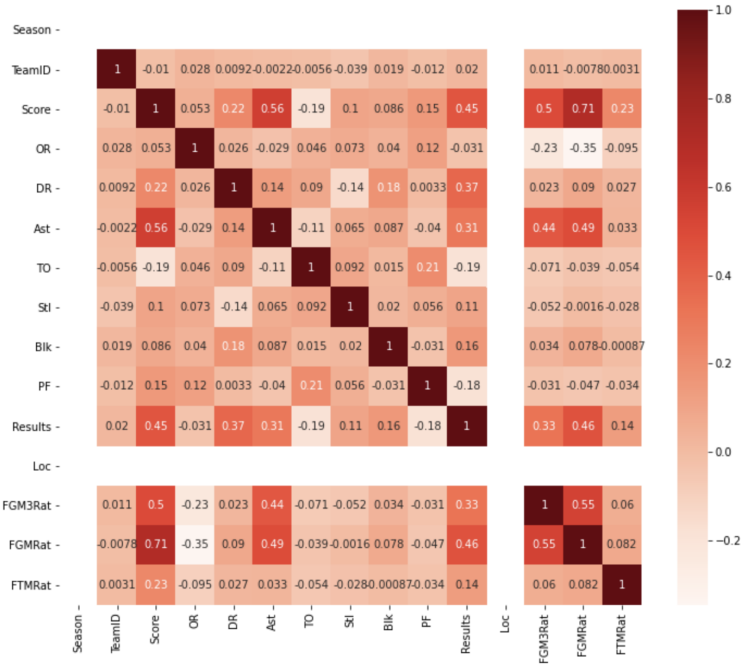
Table: Feature Ranking for Away games

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

Correlation matrix for home games



Correlation matrix for away games



Implementing the model

In creating the model for the logistic regression, a comparison of the accuracy scores were made for the model before and after the RFE model was used for feature selection.The result from the feature selection based on the RFE was fed into the model, with the data being transformed before it could be used in the model. This is because the number of features were changed from 10 to 5 for the RFE model. The top 5 features selected by the RFE model were used for the RFE evaluation.

Findings

Model	Accuracy	Recall	Precision	F1
Logistic Regression All Features Home	0.687	0.69	0.78	0.59
Logistic Regression Top Five Features Home	0.689	0.69	0.74	0.61
Logistic Regression All Features Away	0.692	0.69	0.78	0.61
Logistic Regression Top 5 Features Away	0.477	0.48	0.76	0.42

Based on the performance metrics above, the overall classifier accuracy was similar for when all the features were used to predict the outcome of home games, and when the top five features were used to predict the classifier accuracy. These top five features as denoted by the ranking of 1 in *Table Feature ranking for home games are* : FGM3Rat, FGMRat, OR, FTMRat and BLK.

The accuracy on both outputs is approximately 0.69, which meant that 69 % of the outcome predicted for wins and losses for the 2019 season were correct. This is a good result for our model,especially if it can predict results for win and loss of home games 60 % of the time based on picking five important features.

Also based on the performance metrics above, the overall accuracy is different for both the before and after selection of the top five features used to predict the outcome of home games. These top five features as denoted by the ranking of 1 in Table *Table Feature ranking for home games are* : FGM3Rat, FGMRat, FTMRat, TO and Blk.

Since the accuracy result does not predict as well on the test data , when using the top 5 ranked features, we can say that these features do not do as good a job on predicting the result of away matches. The original 10 features are better predictors of results of away matches.
This is still a good result , especially if the model can predict the outcome of away matches in 2019, seventy percent of the time, when using all the ten features.

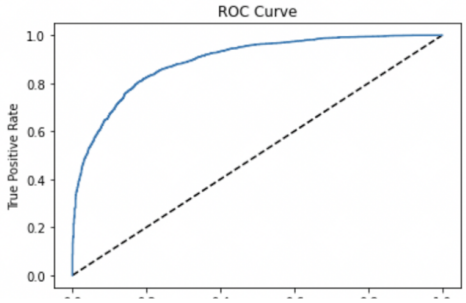
The receiver operating characteristics (ROC) curve is another common tool used with binary classifiers. The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner).

When comparing the AUC curves both the home and away games curves, the results show similar predictions for home and away games.

Fig 6: AUC/ROC curve for Away games

```
AUC: 0.9775666730100043
AUC scores computed using 5-fold cross-validation: [0.47593976 0.53316678 0.59748753 0.49196562 0.49584118]
AUC scores computed using 5-fold cross-validation: [0.47593976 0.53316678 0.59748753 0.49196562 0.49584118]
```

```
Text(0.5, 1.0, 'ROC Curve')
```




```
Text(0.5, 1.0, 'ROC Curve')
```

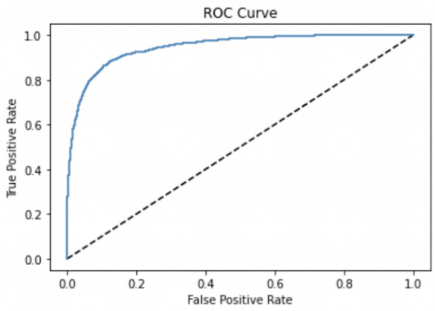


Fig 5: AUC/ROC curve for Home games

```
AUC: 0.9484219828692092
AUC scores computed using 5-fold cross-validation: [0.47300468 0.53468432 0.60063523 0.49177316 0.52491914]
```

Conclusions and Future Work

The model correctly identifies the results of 2019 matches a high percentage of the time when a team is playing at home or away. The accuracy of the classifier for the testing data was at seventy percent when all features were selected.
This logistic regression model does a better job of predicting the test data outcomes for home games than away games when using the test data for the 2019 season, and as would be expected, the most important factors contributing to a team winning a match is the percentage of basket scored in a game i.e. (FGMRat, FTMRat, FGM3Rat). These features as with most of the other top five features track alike for teams winning both at home, and away. In addition to these features, the block (blk) feature is also an important feature in predicting a match outcome.
The two features that were different from the top five features for both the away and the home games are the offensive rebound (OR) and the turnovers (TO). Offensive rebounds are a more favored stat when playing at home , while turnovers (TO) are favored while playing away. Basically, saying that home teams have a tendency to secure the ball better than away teams.
The correlation of the results of matches for home and away matches with turnover is a negative one, while the correlation with offensive rebound is a positive one. This correlation shows that high turnovers can cost a team to lose a match, and high offensive rebounds can help a team win a match.
Another observation from the correlation matrix, is the high positive correlation **(0.62)** observed for the assist feature and the results of matches.

Based on the information provided by the model on assist, rebounds and turnovers, a coach can rotate his team in such a way that players that have a higher tendency of providing assists, limiting turnovers and having a greater tendency of collecting offensive rebounds should be played more minutes when playing at home. Likewise players that have less tendency to turnover the ball, providing assists could be played more when playing at away locations.
In general, players that collect a lot of offensive rebounds and blocks are similar types of players (taller than average players , or players with great athletic vertical leap). A coach should think of giving these types of players more minutes when playing at home, since these features are important in determining the outcome of matches.

Finally, since the five features selected from the overall set of features do not greatly decrease the accuracy of the model 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 a match when playing at home.

Recommendations

Further information that could help with fine tuning this model, could be plotting the average heights of players and minutes played, when playing matches at home versus when playing matches at away locations. The result of this data with the outcome of matches, can help a coach find out whether specific types of players collecting offensive rebounds, deserve more minutes when playing matches at home or away.

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.