



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

Adeyemi Adejuwon

Capstone Project

Springboard Data Science

July 2020

Introduction

- Home Field Advantage is a phenomenon in sports that is not yet understood
 - The most important factors believed to be responsible for this advantage are ¹:
 - Crowd
 - Travel factor
 - Size of home team arena
- Examples of sports and their statistics for home field advantage are known ²:
 - NFL :57.6%
 - NHL: 58.0%
 - Premiership Soccer:63.1%
 - WNBA:61.7%

[1] [Home-Field Advantage \(SOCIAL PSYCHOLOGY\) - iResearchNet](#).

[2] https://www.harrywalker.com/media/2001/scorecasting_si_1-17-111.pdf



Source: Sports Illustrated, Jan. 17, 2011 "What's Really Behind Home Field Advantage?"
by Tobias J. Moskowitz and L. Jon Wertheim

Objective

- Game Factors that affect a team winning or losing a NCAA Division 1 Men's Basketball game
 - Field goal percentage (Field goal made/Field goal attempted)
 - Three point percentage (Three point made/Three point attempted)
 - Free throw percentage (Three point made/Three point attempted)
 - Turnovers per team
 - Assists
 - Steals
 - Blocks
 - Personal Fouls
 - Offensive Rebounds
 - Defensive Rebounds

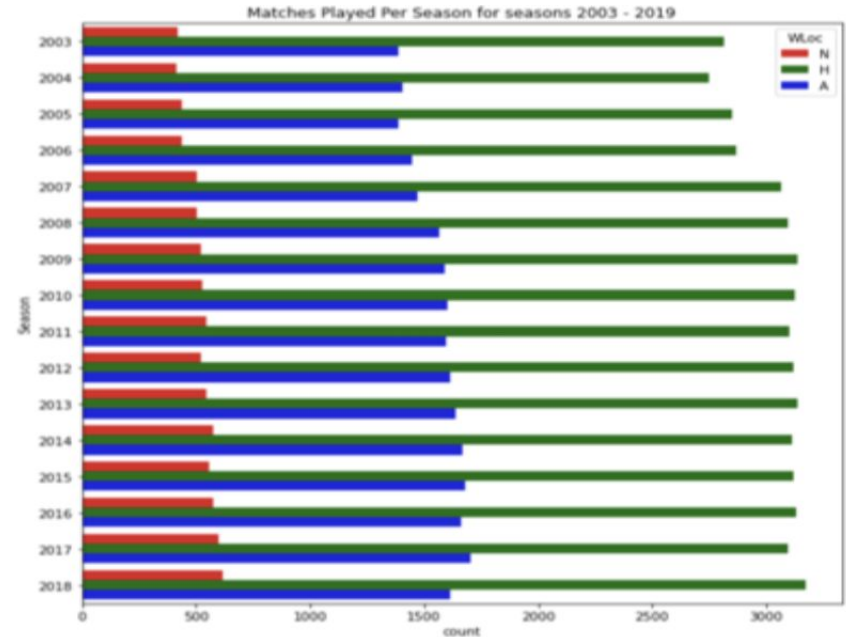
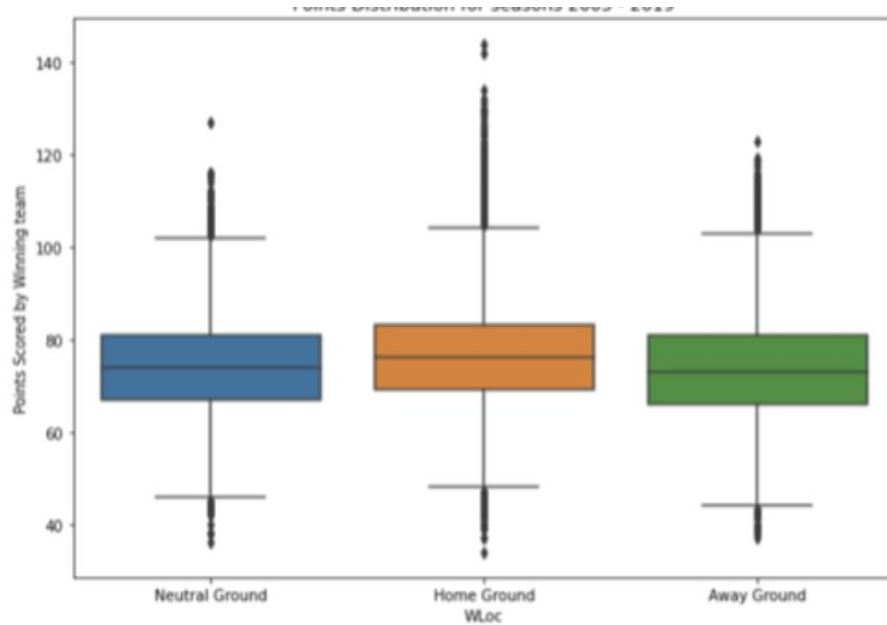


Dataset



- NCAA basketball season for 2003 -2019 season (no March Madness)
- 351 basketball games - 157,164 data points

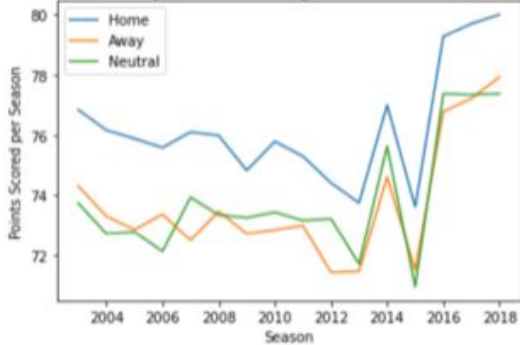
Statistics of Matches in Different Locations



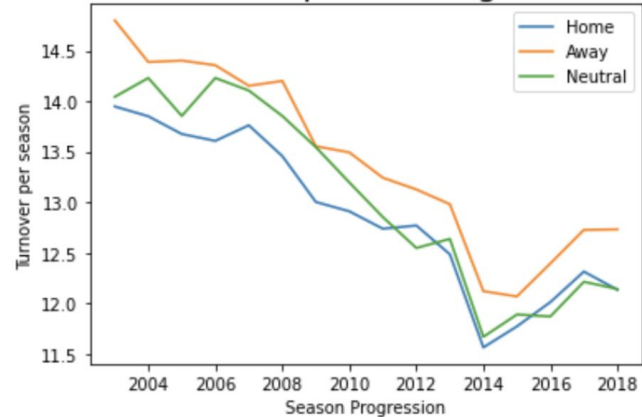
The box plots show that the average number of points for the winning team at their home ground is higher than the points scored when playing at either away or at neutral grounds. Likewise for the histograms, the number of games won at home, away and neutral grounds were all similar across all seasons

Statistics Comparing Matches in Different Locations

Points Scored per winning team from 2003-2019



Turnover per winning team



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

Machine Learning Approach

- Logistic Regression:

Using one model to predict outcome of matches. This is a binary classification problem described with two classes:

- Dependent variables
 - Home team winning
 - Away team winning
- Independent variables
 - Moving averages of the historical performance metrics



*Games played in neutral location will not be used in this project

Machine Learning Approach

Data Partitioning



Training Data: Games played in 2003 -2017 season

- 48422 home win data points
- 24967 away win data points

Test Data: Games played in 2018 season

- 3141 home win data points
- 1713 away win data points

Model Performance

Train Classification Report

```
[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
```

Test Classification Report

```
[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
```

Interpretation

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.



Feature Selection

Original Model is extended using a Feature selection algorithm called: Recursive Feature Elimination

- The five important features assigned:

1. Offensive Rebound
2. Block
3. Field Goal Percentage
4. Free Throw Percentage
5. Three Point Percentage



Classification report using Feature Selection

[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	4854
macro avg	0.81	0.57	0.53	4854
weighted avg	0.77	0.69	0.61	4854

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.

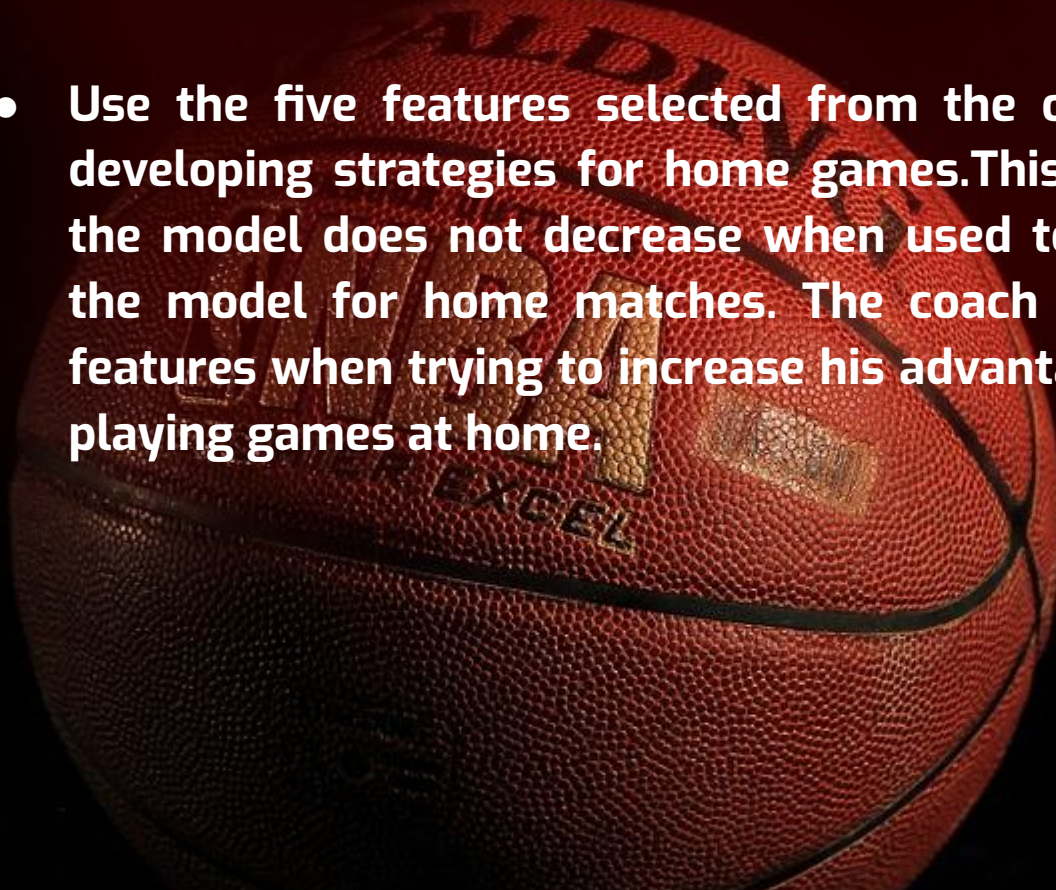


FINAL CONCLUSIONS

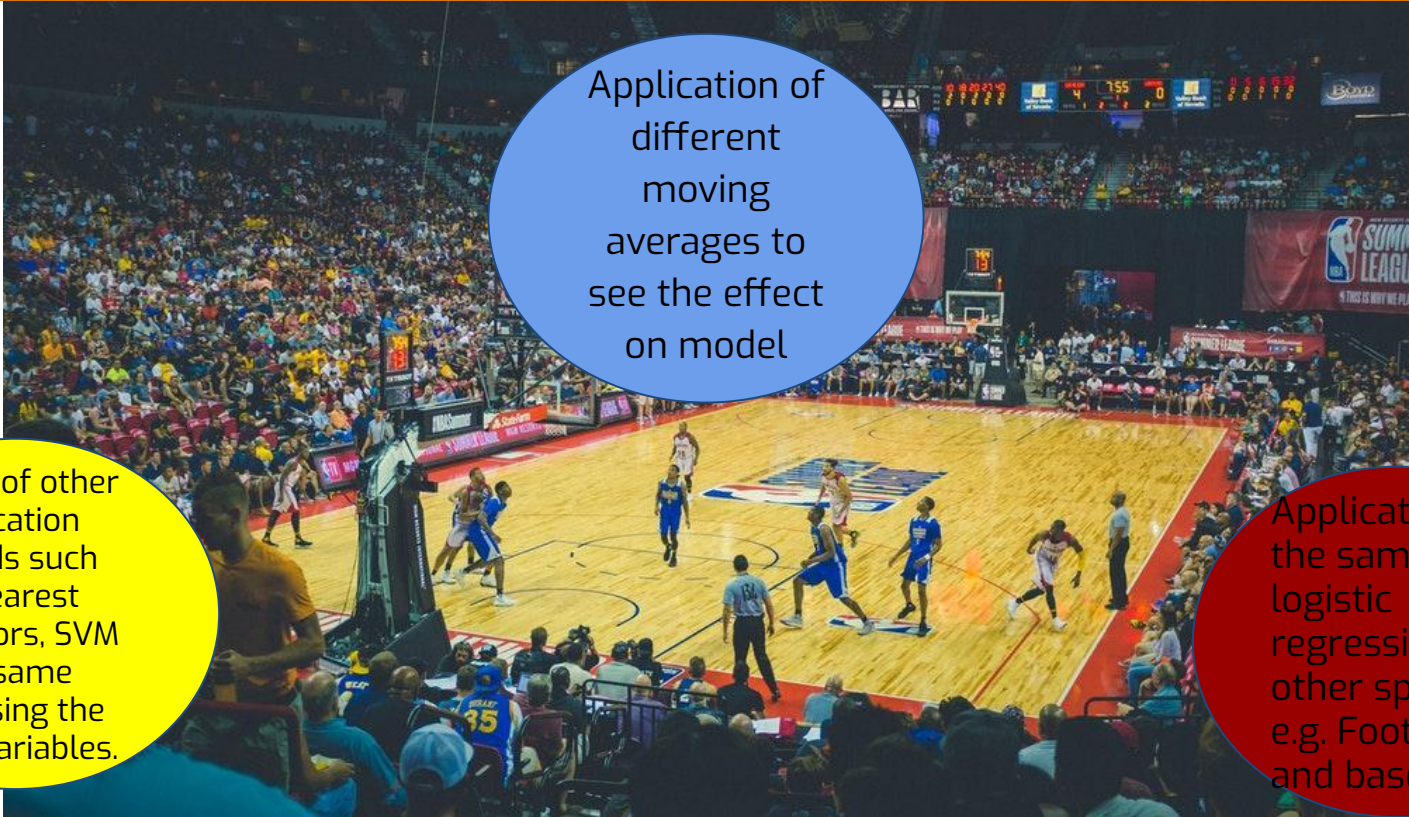
1. The model is good when predicting when the home team is going to win, and not good at predicting when an away team is going to win
2. From the perspective of improving the global accuracy of the Logistic Regression model using RFE, the most important features are: field goal percentage, free throw percentage, field throw percentage, blocks and offensive rebounds

Recommendations to Client

- 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



Application of
different
moving
averages to
see the effect
on model

Review of other
classification
methods such
as: k-Nearest
Neighbors, SVM
on the same
data, using the
same variables.

Application of
the same
logistic
regression to
other sports
e.g. Football
and baseball

Thank you!



BIBLIOGRAPHY

1. "The Home Court Advantage: Evidence from Men's College Basketball." *The Sport Journal*, 13 Feb. 2017
2. Bunker, Rory P., and Fadi Thabtah. "A Machine Learning Framework for Sport Result Prediction."
3. "Home-Field Advantage (SOCIAL PSYCHOLOGY) - IResearchNet." *Psychology*, 21 Jan. 2016
4. Ncaa. "NCAA Basketball." *Kaggle*, 20 Mar. 2019, www.kaggle.com/ncaa/ncaa-basketball.
5. D. Buursma, Predicting sports events from past results "Towards effective betting on football matches".



Presentation design

- FONT:
Exo
Russo One
- IMAGES:
www.pixabay.com
www.unsplash.com
- ICONS:
www.iconfinder.com
- DESIGN BY
www.slidespower.com