Statistical Slam Dunk: Understanding and Predicting Home Team Victories in NCAA Basketball

Ryan Mersereau
Aiden Bartlett
Nico Field

I. Abstract

This report focuses on analyzing NCAA basketball games to predict home-team wins and explore team performances at home. The goal is to understand why some teams perform better at home and to create a model predicting home-team victories.

Our data comes from the Google Cloud public datasets in BigQuery. This dataset encompasses nearly 30,000 NCAA games from 2014 to 2017 and provides information such as venue details, team information, and team game stats.

After analyzing the data, we found that two-thirds of games are won by the home team, all teams have an advantage at home, crowd attendance does have an impact on the chance the home team wins, and defensive rebounds/field goals made are the most accurate way to predict the winner of a game.

II. Introduction

In the realm of NCAA basketball, there is a recognized advantage of playing at home vs playing away. This prompted our investigation into the disparity between home and away game performances. Upon initial analysis, a substantial discrepancy emerged, compelling us to delve deeper into several pivotal questions: Does the advantage of playing at home extend uniformly across all teams, or do certain teams exhibit a more pronounced advantage? How does the size of the crowd influence the home team's performance? Furthermore, which variables serve as the most influential predictors for determining a team's victory? The analysis hinges on extensive game statistics, team attributes, and venue-related data. Since there are so many possible predictor variables, it was challenging to build an accurate model. Another challenge will be to extract team data based on this dataset since the dataset is a collection of NCAA games; this will require a fair bit of tidying of the data. By exploring predictors of home-team wins and assessing team performances, this analysis aims to provide valuable insights for basketball enthusiasts.

III. Methodology

The methodology for our analysis involved several key steps.

Data Collection and Preprocessing:

The dataset encompassed records of NCAA basketball games, comprising team statistics, game outcomes, venue details, and other relevant attributes. Preprocessing steps involved data cleaning, handling missing values, and encoding categorical variables for analysis readiness.

```
Unset
# Create new variable for point differential and win/loss/tie
library(dplyr)
bball <- bball |>
```

```
mutate(point_differential = h_points_game - a_points_game,
       h_game_result = case_when(
         point_differential > 0 ~ "Win",
         point_differential < 0 ~ "Loss",
       ))
```

Shown above is an example of some of the preprocessing work that was done to prepare for further analysis.

Preliminary Data Analysis:
Our preliminary data analysis involved visualizations such as histograms, scatter plots, and graphs to uncover patterns, trends, and correlations between variables, focusing on home-team performance and game outcomes. These observations gave us the trends we needed to perform the next step.

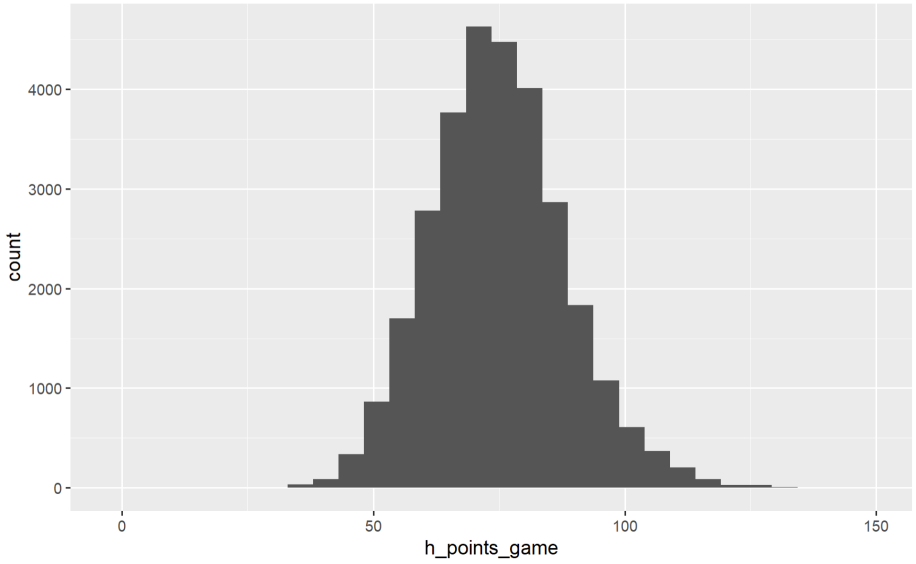
Feature Selection and Engineering:
Identifying pertinent features influencing home-team wins and performance at home was the next crucial step. A subset of the original data frame was taken that only included numerical statistics for the home team in their games. A total of 35 variables were considered to predict our response variable, home game result, which included statistics like turnovers, rebounds, and free throws. After omitting NA values, we used the StepAIC function from the MASS library in R to build a logistic regression model using the best features from the dataset. We used a forward selection of features and our final “best” model was constructed with 14 features all found to be significant in predicting home game results.

Construction of Classification Tree:
Once the logistic regression model was determined, we were able to use that formula in the construction of a classification tree. We also attempted to perform cross-validation and prune the tree to see if that resulted in any improvement, but the tree was unchanged after pruning. Finally, a second tree was constructed with more detail using the rpart library. This tree had some slight changes to the previous tree but was very similar. This tree also allowed us to see detailed statistics of the predicted outcome (win or loss) for each of the terminal nodes of the tree.

IV. Analysis and Results

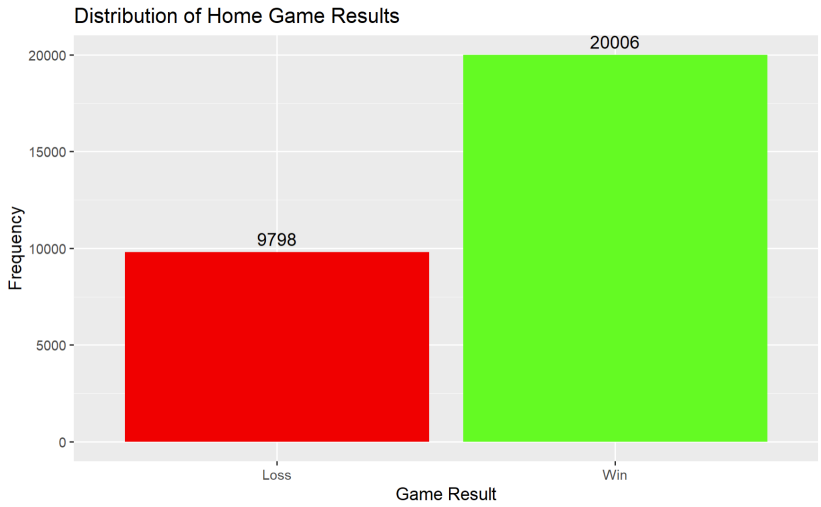
A. Preliminary Data Analysis

The histogram below displays the distribution of points scored by home teams across NCAA basketball games. It provides insights into the frequency distribution of scoring, illustrating the range and density of points achieved by home teams.



From this histogram, we can conclude that home teams generally score between 70-80 points in a single match. We can see that the outliers go beyond 125 points and below 40 points but the vast majority of games are in the 50-100 point range. The histogram is also approximately normally distributed which is not surprising given the nature of points in a basketball game.

The bar chart below depicting home team wins versus losses over the observed games reveals a large difference in the performance trend of teams when playing at home. This visual representation delineates the comparative frequencies of wins and losses, offering an understanding of the home-team success rate.



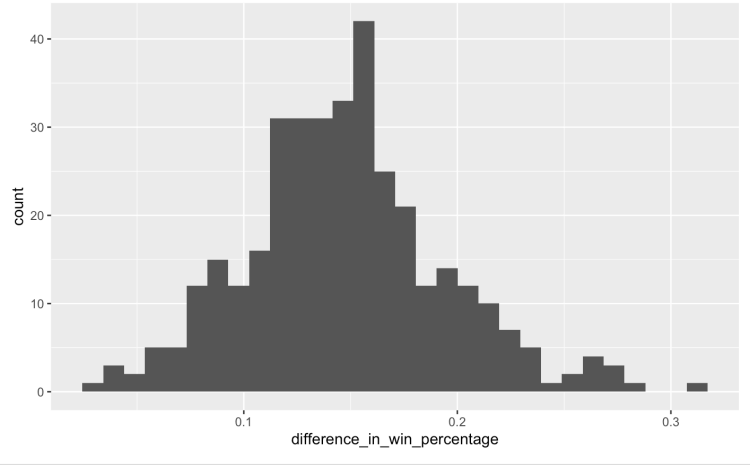
As expected, the home team wins a significant amount more than the away team (20006 / (20006 + 9798) = .671 so the home team wins about two-thirds of games. While this was somewhat expected, the ratio is much higher than what we would have thought before looking at the data.

B. Team Comparison Analysis

From our analysis so far, we have established that the home team has an advantage over the away team. The next question we asked was: Is this true for all teams? Or, do only some teams (like top 10 ranked teams) win all of their home games while others are not affected by this trend?

Given that all we had were games, we had to create a new table that tracked each team's home and away wins and games played. After doing this, we found the difference between the win percentage of each team at home and away.

The following graph shows the distribution of the difference in win percentage. Note that we factored out each team that didn't have at least five away and home games because of the low sample size.



From this graph, we see that all teams had a greater than zero difference in win percentage. This means that every team wins more at home than away. The median of the graph is .16 meaning that most teams win 16% more when they play at home.

However, some teams have a much higher difference in percentage. The highest difference in win percentage can be seen below:

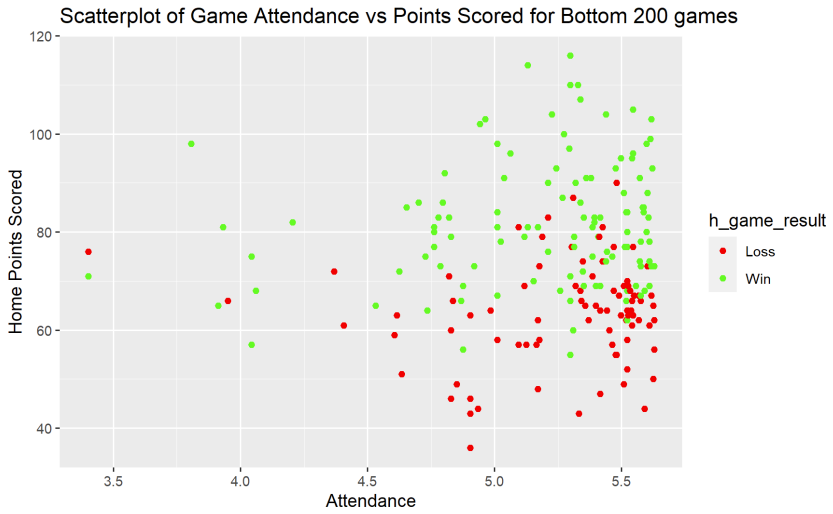
h_name <chr>	h_market <chr>	difference_in_win_percentage <dbl>
Tigers	Texas Southern	0.30906238
Cougars	Brigham Young	0.28730159
Braves	Alcorn State	0.27176221
Bulldogs	South Carolina State	0.27016129
Bulldogs	UNC-Asheville	0.26875000
Huskies	Houston Baptist	0.26469534
Lions	Southeastern Louisiana	0.26399254
Panthers	Prairie View A&M	0.26358669
Tigers	Jackson State	0.25929267
Eagles	Eastern Washington	0.25236686

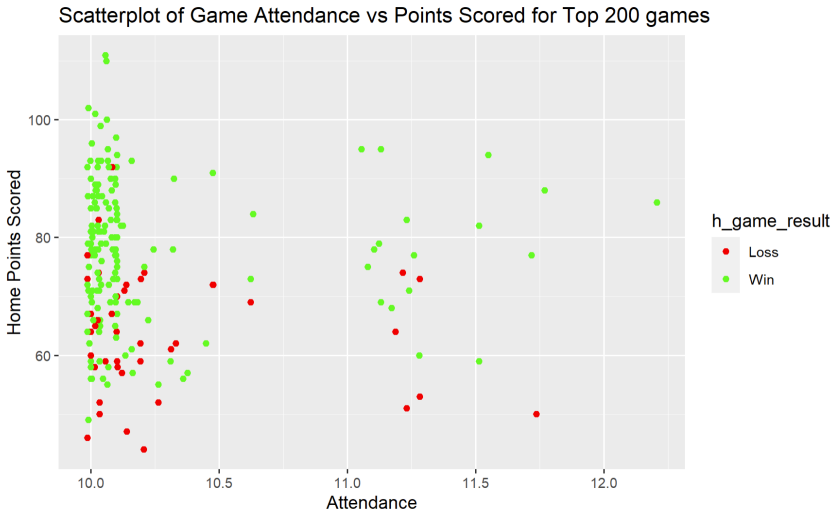
Additionally, of the 355 teams in the dataset, only 37 have a win rate greater than .500 (meaning they win more on the road than they lose). This was a very surprising result to us as this exemplifies the vast difference in performance for all teams based on whether they are the home or away team.

To answer our question from before, we see that not just the top teams with large stadiums have a home-team advantage. Texas Southern, the school with the highest difference in win percentage, was never ranked. Rather, all schools have at least some home-field advantage.

C. Venue Comparison Analysis

As stated before, not just the most popular schools have a home team advantage. Our results encouraged us to look deeper to see how (if at all) attendance affected the home team's win percentage. To analyze attendance, we first created a dataframe for the bottom 200 attended games and one for the top 200 attended games. After plotting logged attendance vs home points scored and win rate for each subset of the data, we got the following graphs:





Just from looking at the game results, it is apparent that the bottom 200 games have significantly more losses than the top 200 games. After performing some calculations on the data, we observed that the home team win rate for the bottom 200 attended games was 58.5% (117/200) while the win rate for the top 200 attended games was 78.5% (157/200).

What this tells us is that there is a correlation between high attendance and the home team winning. This makes intuitive sense because as more fans crowd the stadium to cheer on their team, it follows that the team should perform better. However, we cannot conclude that attendance affects the home team winning. More exploration would have to be done here; we believe that significant games and schools with a large basketball presence would attract a large fanbase, perhaps because the home team wins a lot while attendance is high.

D. Logistic Regression Model Analysis

The results of our logistic regression modeling gave us a final model with 14 predictor variables, as shown below.

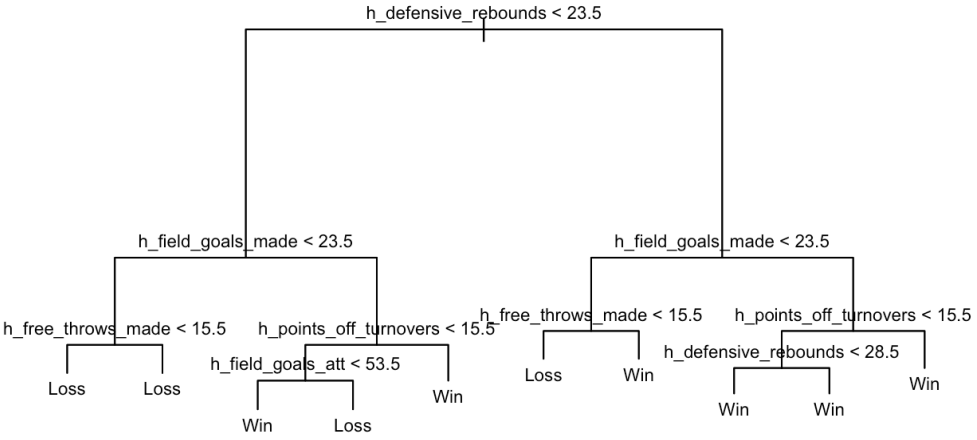
```
final_model <- glm(final_formula, data = bball_vars, family = binomial)
summary(final_model)
```

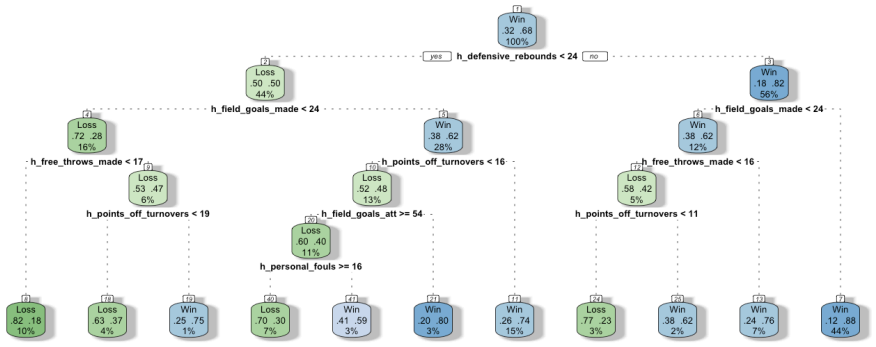
```
Call:
glm(formula = final_formula, family = binomial, data = bball_vars)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.33893    0.65968  -0.514 0.607399
h_field_goals_made        0.51457    0.02277  22.596  < 2e-16 ***
h_defensive_rebounds      0.51670    0.02104  24.563  < 2e-16 ***
h_points_off_turnovers    0.18892    0.01413  13.367  < 2e-16 ***
h_turnovers              -0.44588    0.02255 -19.776  < 2e-16 ***
h_offensive_rebounds      0.42862    0.02549  16.812  < 2e-16 ***
h_team_rebounds           0.44376    0.03483  12.741  < 2e-16 ***
h_steals                  0.35861    0.02903  12.351  < 2e-16 ***
h_three_points_made       0.25990    0.02318  11.214  < 2e-16 ***
h_personal_fouls         -0.14383    0.01669  -8.620  < 2e-16 ***
h_free_throws_made        0.22963    0.02757   8.329  < 2e-16 ***
h_field_goals_att        -0.50674    0.02121 -23.893  < 2e-16 ***
h_free_throws_att        -0.11732    0.02218  -5.289 1.23e-07 ***
h_rank                    0.03901    0.01016   3.840 0.000123 ***
h_coach_tech_fouls       -1.36777    0.35877  -3.812 0.000138 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5648.7  on 4506  degrees of freedom
Residual deviance: 1947.7  on 4492  degrees of freedom
AIC: 1977.7

Number of Fisher Scoring iterations: 7
```

We can see that all variables have a very low p-value, meaning that they are all individually significant in predicting home game results. This model has a null deviance of 5648.7, and a residual deviance of 1947.7, with an AIC value of 1977.7, which is relatively low for a model with this many predictor variables.

E. Classification Model Analysis

Using the equation above, multiple classification models were constructed using the tree and rpart libraries, as seen below:





Classification Tree for Home NCAA Wins

As we can see from the classification trees, defensive rebounds for the home team appear to be the most important predictor of whether or not they win the game. This first split takes place at 23.5 defensive rebounds. Not surprisingly, field goals made by the home team are the next most important predictor, followed by free throws made and points off turnovers. Field goals attempted and personal fouls were also used in the construction of the model.

The first model constructed had 10 terminal nodes, with a residual mean deviance of 0.9466, and a misclassification error rate of 0.2205. Pruning this model resulted in no change, indicating that 10 terminal nodes are the optimal size for this model.

The second model constructed had 11 terminal nodes. The most confident prediction for a home win is the condition where a home team has over 24 defensive rebounds and over 24 field goals made. The model predicted an 88% win chance for the home team with these conditions. The most confident prediction for a home loss is the condition where a home team has less than 24 defensive rebounds, less than 24 field goals made, and less than 17 free throws made. The model predicted an 18% win chance for the home team for these conditions.

A lot of these predictor variables make sense: field goals made would result in more points scored and therefore the team with more points would win. We were surprised though to see that variables like three-pointers made and team rebounds(which we saw earlier have a large impact on the result) did not show up in the final classification tree models.

V. Discussion

Before we summarize our general findings, we want to note that the dataset we used only contained data from 2014-2017. As such, we should be careful generalizing the data outside of that time. While it would make sense to expect similar results, it would be best to use a similar database and similar code to answer questions about different periods or a similar sport/league (maybe we could consider NBA games and see how similar trends are).

Now let us return to our initial questions. Firstly, we found that not just some teams have a home-field advantage but rather all teams have an apparent advantage when playing at their stadium. While it is possible that crowd size affects the result of a game, it is also possible that crowd size and the home team winning are correlated; more data/experimentation would be needed to draw any proper conclusions. We found the best predictor values for a team to be defensive rebounds and field goals made. This implies that while a home-field advantage does provide an edge, the game will still be decided by who has a better performance on the court.

While completing this project, we gained experience in applying statistical techniques learned in class to real data to generate meaningful conclusions. However, we encountered several obstacles that may have hindered our results, and believe that there are still improvements to be made. One big issue was the size of the dataset and having to deal with uncleaned data like several NA values that needed to be removed and other observations that had incorrect values that we needed to ignore. For example, a game with a reported attendance of 200,000, with a venue capacity of 20,000 is inaccurate. We also excluded a 0-0 tie in which the game ended and ruled a tie before play began, which would have influenced our predictions. In constructing the logistic regression model, home team stats of numerical values were all we were able to use to predict the game result. When we tried to include the stats of the away team, the model would not converge, and we encountered several errors.

Although we feel confident that conclusions can be drawn from the logistic regression model, we believe that the model construction process could be improved to generate an even better model. This could be done by splitting the data into training and testing datasets or using a slightly altered model generation process to produce a model with a different number of

variables to reduce the AIC value and/or maximize the Adjusted R-squared value. Additionally, we could examine the residual plots of our model to ensure that the model has no outliers, is approximately normal in distribution, and is sufficient in modeling the data. Other methods of predictive analysis such as Linear Discriminant Analysis (LDA), or Support Vector Machines (SVM) could be examined to see how well they predict the outcome of games.

Some other ideas we had to expand on this project were analyzing scoring or shot percentage trends and how they have changed over time. We could also compare trends across basketball conferences, or see how the performance of teams changes in premier events, such as two highly ranked teams playing each other during the March Madness tournament. I believe this would be a natural evolution of the work we have done so far that could yield some interesting results although not necessarily related to home-field advantage as seen in this paper. Over the course of this project, we learned how to find high-quality data and then clean, modify, and experiment with the data to generate meaningful and interesting results from the data while also experimenting with different data science techniques.

References

BigQuery NCAA Basketball dataset:
https://console.cloud.google.com/marketplace/product/ncaa-bb-public/ncaa-basketball?project=st442-final-project

Link to code:
https://drive.google.com/file/d/1JjX7tmUGMT6SH_-_opldX5iuIuH7j9Fc/view?usp=sharing