Forecast the NCAA Division I Women's Basketball Championships in 2019

Gang Yang, Kening Zhang, Qingying Luo, Taiyou Chen, Yifei Ke

March 2019

1 Introduction

Sports forecasting's history is as long as the history of sports gamble. What would a man in 10^{th} century think when he was betting on his brother's muscle in a arm-wrestling game? The chance of win. And, he would not bet base on his relationship with the player, but the information(data) he got about his brother (such as the diameter of the biceps). It is important for sports fans, team managers, sponsors, the media and the growing number of punters who bet on online platforms (Vlastakis et al., 2007).

This project is a machine learning competition held by Google Cloud. The goal is to attempt to forecast the outcomes of March Madness during 2019's NCAA Division I Women's Basketball Championships with the bracket that we pick using a combination of NCAA's historical data.

We have two focuses about this project. First we investigate whether the home court condition increases our possibility of winning the basketball game, and how the home court contribute to the difference in score. We used graphs, tables, and several tests to finish that. Secondly, we aimed to use our previous data to predict the final outcome of this NCAA basketball game. The models we used include support vector machine, random forest, and logistic regression.

Our forecasting begins with determination of the influence on success or failure of the so-called "home court advantage". We used the method of KS test, two sample T-test to test our hypothesis, and concluded that a home court team has better chance to win the particular play. With the method of support vector machine, we took one more step to quantify the influence that the home court advantage can make to the score regardless of which side are the winner. Therefore, we got some power to predict how good a team could do on a typical NCAA game based on which side is playing "at home".

Furthermore, inspired by Eric Scot Jones' study *Predicting Outcomes of NBA Basketball Games* and Martin Spann and Siera Bernd's *Sports Forecasting: A Comparison of the Forecast Accuracy of Prediction Markets, Betting Odds and Tipsters*, we designed a model to evaluate the fractions of each technical data of a team that might affect the result of the play. What this model do is to study a team's historical records against another team, then calculate out an expected

result of the two teams' future play. By applying the model to all 64 teams (pairing by the official order published by the NCAA) of 2019's tournament and adding alternating quantity of home court advantage, we finally got a prediction of the result of the first round and eventually the national championship.

2 Background

The National Collegiate Athletic Association, as known as NCAA, is a member-led organization dedicated to the well-being and lifelong success of college athletes. Its members including 1117 colleges and universities, 100 athletics conferences and 40 affiliated sports organizations. Nearly half a million college athletes make up the 19750 teams that send more than 52500 participants to compete each year in the NCAA's 90 championships in 24 sports across 3 division.

Google Cloud Platform is the official public cloud provider of the NCAA, and the holder of the ML Competition. It has been supporting NCAA to serving the needs of schools, their teams and students by helping the NCAA use data and machine learning.

NCAA Division I Women's Basketball Tournament is one of the NCAA championship programs, this tournament is an annual college basketball tournament for women, held each March (March Madness). It follows a single-elimination tournament of 64 teams, the regions are: Albany Reginal; Chicago Regional; Greensboro Regional; Portland Regional;

The official website is:

http://www.ncaa.org/championships/division-i-womens-basketball

Here is a brief of the format of the tournament, The game is usually palyed in March and April by a total of 64 qualified teams. NCAA Selection Committee holds the ultimate decision of the qualification. The tournament is split into four regional tournaments, and each regional has teams seeded from 1 to 16. In the first row of game, the top-seeded team in each region gets to play the 16th team, the second seeded team plays the 15th, etc. Eventually, each region would have one representative to play the National Semifinals and eventually the National Championship.

To understand the format easier, please see Figure 1.

As described in the format part, the "seed" is a very important evaluation system, which indicates the strangeness of a team. It is used to construct each year's bracketology. In order to "seed", a team, the selection process evaluate several factors, including team rankings, win-loss records, and Ratings Percentage Index (RPI) data. Historically, each regions' 1 seed team is usually end up to the best 4 of that years tournament.

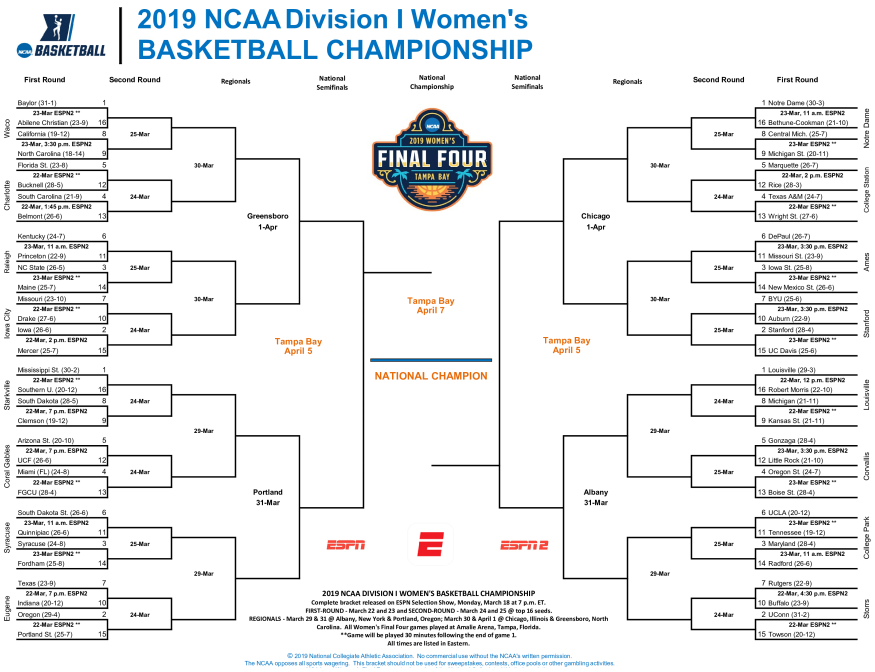

Figure 1: 2019 NCAA Women's Tournament Bracket Schedule

3 Data

Provided by Google Cloud, we were able to access the data of past tournaments. Which had been migrated by NCAA including more than eighty years of historical and play-by-play data, from ninety championships and twenty-four sports.

Some of the data files only provide basic information, such as the team name and its ways of spelling of the particular team ID. Our process and calculation would not use these files, but for forming the conclusion of the project.

We would be mainly use the data sets of the technical information of NCAA tournaments through season 1998 til 2018:

WNCAATourneyCompactResults.csv and WNCAATourneyDetailedResults.csv

Additionally, the regular seasons' information (1998 - 2018) would be used in order to tune the process.

WRegularSeasonCompactResult.cav and WRegularSeasonDetailedResults.csv

The description of each variable can be found under:

https://www.kaggle.com/c/womens-machine-learning-competition-2019/data.

4 Investigation

4.1 Does whether 'home' or not influence winning?

We saw there are 'H', 'A' and 'N' in the winning location column in our data set and we began to wonder is the location where they play the games relative to our winning result or not. So we started to dissect this question in two aspects. The first one is that players have more chance to win when they play at the home location. The second is that the players will win more scores when they play in home location.

4.1.1 First hypothesis

Our first assumption is that **the players will become more likely to win if they play at their home places**. From the data in regular games, we have information from 1998 to 2018, and we focus on only 'H' and 'A'. We can count how many 'H' and how many 'A' separately in the 'WLoc' column from each year, so we will have two vectors, each of length 21. From now on, we use H to denote the data that represents the number of H counted in years from 1998 to 2018, and A follows the similar logic.

After we get the two lists of data, we check their means and draw them in a **boxplot**.

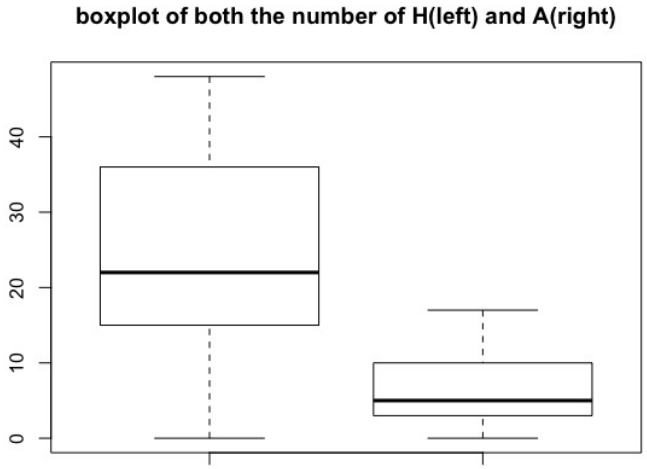

Figure 2: Boxplot of H and A from 1998 to 2018

From Figure 2, we can see the H group is nearly bell shape, while the A

group is a little right skewed. Since the upper tail is longer in the right box plot, which corresponds to A case.

We also calculate the mean and median for each group.

mean of H group	mean of A group
24.80952	6.095238
median of H group	median of A group
22	3

Table 1: Statistics for H and A group

From Figure 2 and the table above, we can see that the mean for 'H' group is greater than the mean for 'A' group, but in order to confirm ourselves that the mean is actually different, we need to perform some rigorous test.

We set our hull hypothesis as that players have equal chance to win whether at home or not, which is $H_0 : \mu_H = \mu_A$, where μ_H comes from the 'H' vector and μ_A comes from the 'A' vector. Since we want to test our first assumption that the home location will result in more winning outcomes, we will set our alternative hypothesis as $H_1 : \mu_H > \mu_A$. Now we want to compare the means of the two and use two sample T-test to see whether there is any significant difference between the two means.

In order to use two sample T-test, we have to **check the normality** of the data. So we plot the histograms of the two.

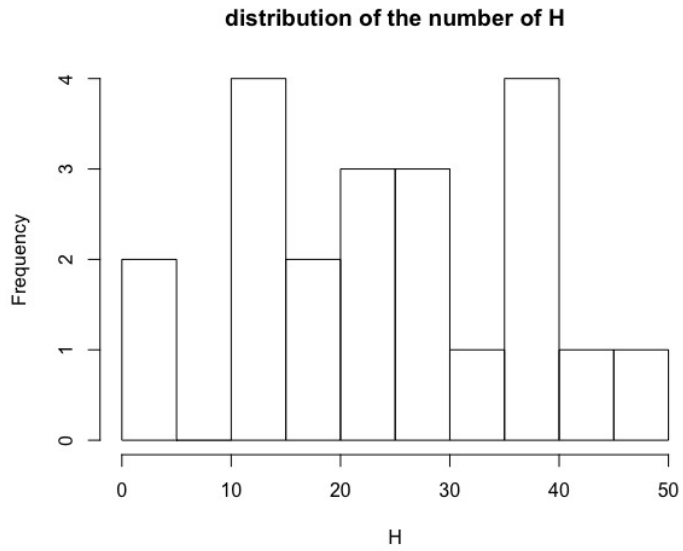

Figure 3: Histogram of the number of H in winning location from 1998 to 2018

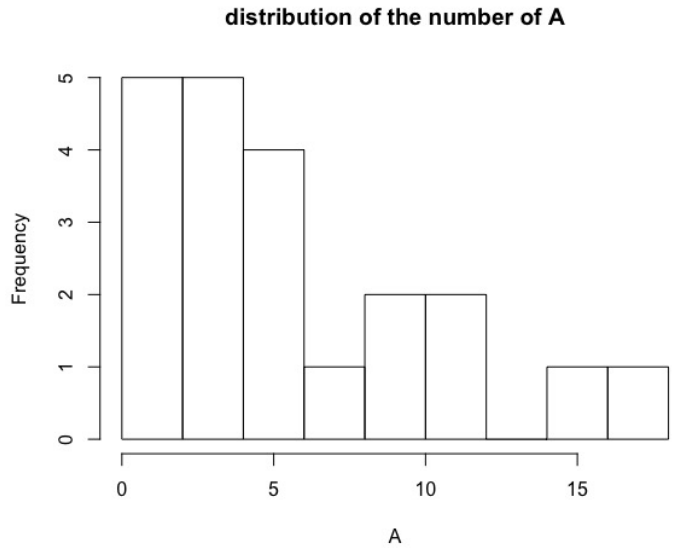

Figure 4: Histogram of the number of A in winning location from 1998 to 2018

In Figure 3 and Figure 4, we can see their distributions but it is hard for

us to tell whether they are normal distribution or not. So for the graphical visualization, we also use **qqnorm** and **qqline** to visualize that.

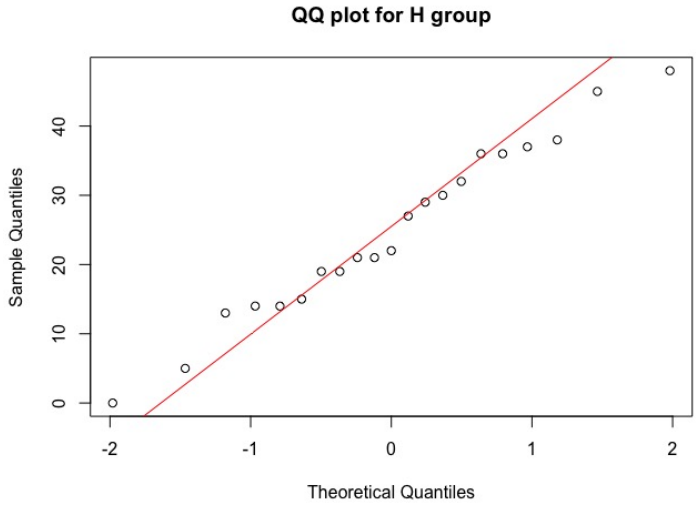

Figure 5: QQ plot for the distribution of H group

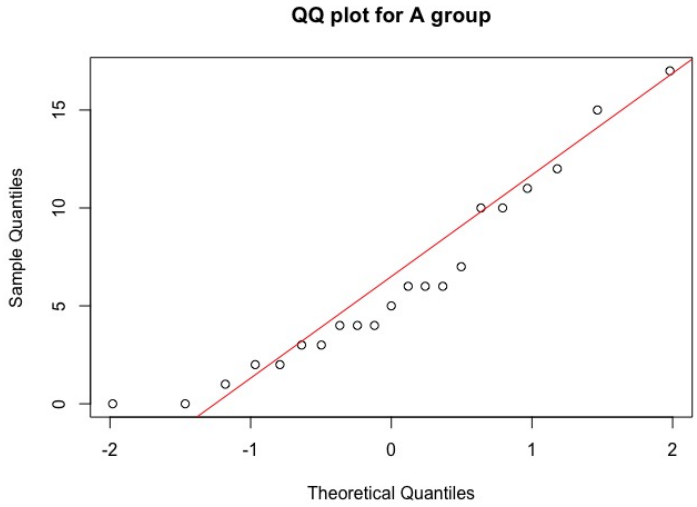

Figure 6: QQ plot for the distribution of A group

From Figure 5 and Figure 6, we can see that they are approximately normal.

Then we use rnorm to generate 21 data from normal distribution with the mean and standard deviation from H group and A group separately, and then use two-tailed KS test to compare whether the two distribution differ or not, with the **null hypothesis that the two distributions are the same**.

KS test for H group	mean	standard deviation	p-value
KS test for H group	24.80952	12.65156	0.8407
KS test for A group	mean	standard deviation	p-value
KS test for A group	6.095238	4.773937	0.3581

Table 2: Table for KS test

From the table above, because the **p-values are all greater than 0.05**, we cannot reject the null hypothesis in both cases.

But to confirm ourselves, we decide to use **kurtosis** and **skewness** to evaluate their normality further. We use **Monte Carlo** to simulate 1000 times, each consisting of 21 numbers from standard normal distribution. We then get the kurtosis and skewness of each, and will have two distributions so that we can calculate the 95% interval and see whether the number we get fall in the extreme categories. The lower bound is 0.025 quantile of the Monte Carlo simulation, and the upper bound is 0.975 quantile.

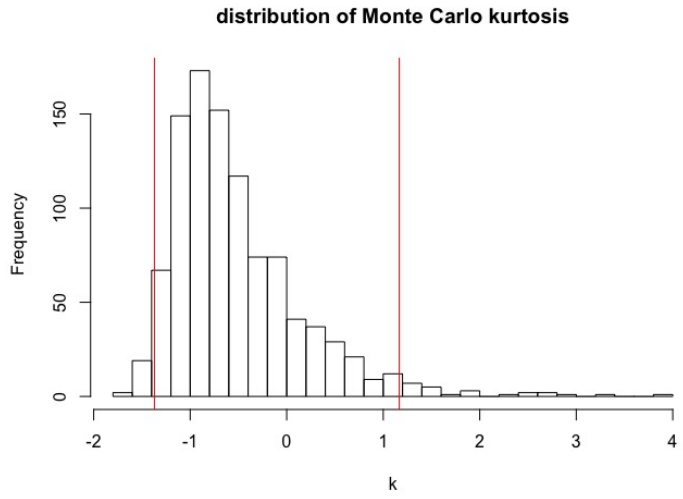

Figure 7: Histogram of simulated kurtosis

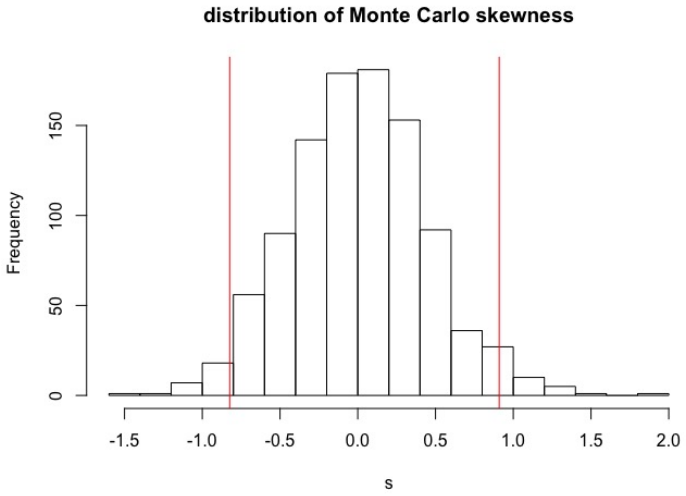

Figure 8: Histogram of simulated skewness

Figure 7 and Figure 8 are visualization of the distribution of our simulated kurtosis and skewness from Monte Carlo. We use red line to show the 0.025 quantile and 0.975 quantile of the two distributions and record the number in the following table.

kurtosis interval	kurtosis of H	kurtosis of A
(-1.368867,1.166228)	-0.8913558	-0.5312404
skewness interval	skewness of H	skewness of A
(-0.8239712,0.9094824)	-0.01125381	0.7159253

Table 3: Table for kurtosis and skewness

From the above table, we can see the kurtosis and skewness of both H group and A group fall in the 95% confidence interval. So we conclude that they are from normal distribution and we move on to the **two sample T-test**.

We want to test if the players will be more likely to win at their home places, so we want to know if the mean of the counts of H will be greater than the mean of the counts of A. We set our null hypothesis as that the mean of the two distributions is the same, whereas the alternative hypothesis states that the mean of H will be greater than the mean of A, which shows that the players will tend to win at their home places. Since we do not know the true variance, the test statistic will have a t-distribution, and the **var.equal will be FALSE**.

From the table above, we see that the **p-value is less than 0.05**, which means we should **reject the null hypothesis** and conclude that the two means differ significantly.

Two sample T-test	Null hypothesis	Alternative hypothesis	p-value
Two sample T-test	$H_0 : \mu_H = \mu_A$	$H_1 : \mu_H > \mu_A$	5.516e-07

Table 4: Table for two sample T-test

4.1.2 Conclusion on first hypothesis

Since the means from distributions of the number of H and the number of A differ significantly, and the mean of H is greater than the mean of A, we can conclude that the female basketball game players will have a greater chance to win when they play at their home places. We suggest that all colleges who want to win should strive for more opportunities to get the games held at their own universities. Since players are often trained at their local universities, they will be more familiar with their own colleges and environments, and result in better performances, like more likely to win the game.

4.1.3 Second hypothesis

Now we already know the home places will lead more chances to win. We want to further investigate of all the winning cases on how does the 'H' and 'A' influence the degrees they win. More specifically, how does 'H' and 'A' influence the difference between winning scores and losing scores.

Our assumption is that **in all of the winning games, players who play at their home places will make higher differences in winning scores and losing scores**. In order to test that, we separate the the winning games into H group and AN group, with the first one meaning that the players win at their home places, and the latter one meaning that the players win not at their home places ('A' and 'N'). Then we get the difference between winning scores and losing scores, and make them into two vectors, namely H and AN.

First we gather some basic information of the two vectors. We can see from

Statistics	H group	AN group
mean	23.77246	14.37
median	21	12
standard deviation	15.96482	10.80161
length	167	400

Table 5: Statistics for H and AN group

the table the mean of H will higher than the mean of AN. And the number of H, 167 is less than 400, sum of the number of A and the number of N.

Again, we should use graphical approach to visualize the data. We first check the histogram of two distributions.

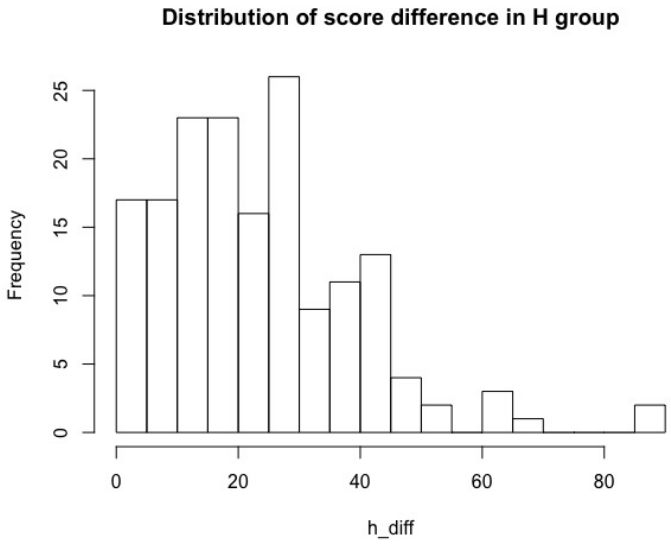

Figure 9: Histogram for score difference in H group

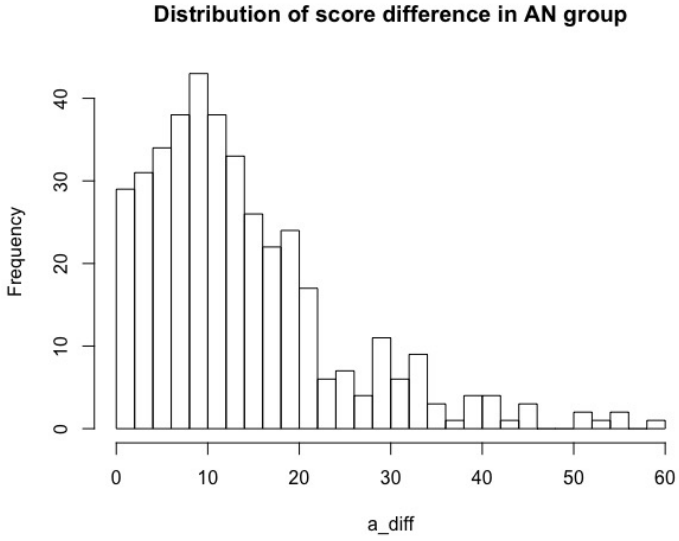

Figure 10: Histogram for score difference in A group

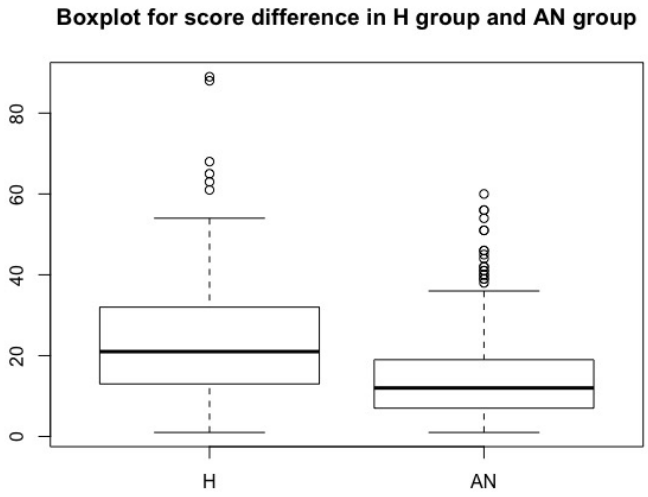

Figure 11: Boxplot for score difference in H and AN group

From the Figure 11, we can see that the mean of H is higher than the mean of AN. There are many outliers at the upper quantiles of the two distributions, and they all seem to be right skewed. So we doubt the normality of these two distributions.

In order to check the normality of these two distributions, we again use qqnorm and qqline.

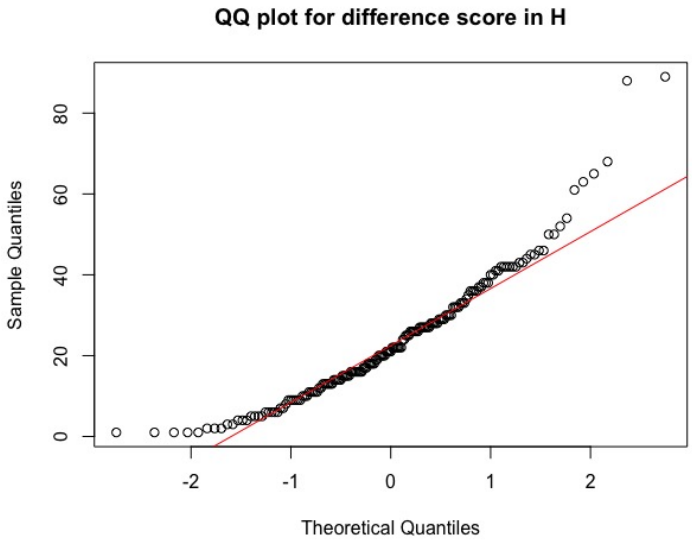

Figure 12: QQ plot for score difference in H group

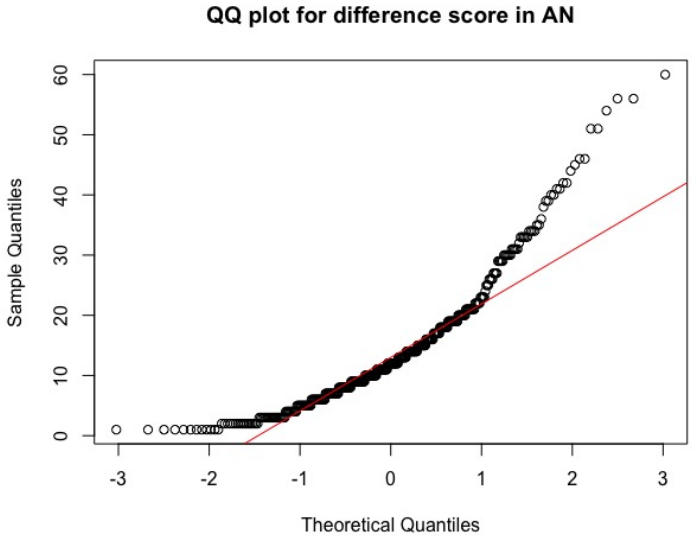

Figure 13: QQ plot for score difference in AN group

From Figure 12 and Figure 13, we question their normality. So we use two-tailed KS test with the null hypothesis being that the two distributions are the

same, and we use Monte Carlo simulation on kurtosis and skewness again to check its normality. The following table is the summary of the details.

KS test for H	null hypothesis	p-value	Decision
KS test for H	two distributions the same	0.03492	reject null
Kurtosis	95% interval	Kurtosis of H	In the range
Kurtosis	(-0.6167333, 0.8344197)	2.165023	FALSE
Skewness	95% interval	Skewness of H	In the range
Skewness	(-0.3559927, 0.3495534)	1.153879	FALSE

Table 6: Statistics for H group

KS test for AN	null hypothesis	p-value	Decision
KS test for AN	two distributions the same	0.0002468	reject null
Kurtosis	95% interval	Kurtosis of AN	In the range
Kurtosis	(-0.4216925, 0.4732243)	2.301868	FALSE
Skewness	95% interval	Skewness of AN	In the range
Skewness	(-0.2520398, 0.2544279)	1.42589	FALSE

Table 7: Statistics for AN group

From the above two tables, we see that the two distributions are not normal. We cannot use the two sample T-test without normality assumption. So we move on to the non parametric test **Mann–Whitney U test**, which is essentially the same as the t-test for independent samples. We set the null hypothesis as the distributions of both populations are equal, and **alternative hypothesis as one distribution is stochastically greater than the other**.

Test	alternative	p-value	Decision
Mann–Whitney	greater	2.746e-13	reject null

Table 8: Mann–Whitney U Test Result

From the table above, we see that we should **reject the null hypothesis**. And there is significant difference between the mean of two groups, with the mean from H group higher.

4.1.4 Conclusion on second hypothesis

We can conclude that **when players play at their home places, they will win more scores and result in a larger score difference**. So we encourage

all colleges take every advantage of home court.

4.2 How to predict the results of 2019/2018 Tournament based on the data in hand

The purpose of this Kaggle competition is to predict the results of the 2019 Tournament of women basketball game. Given two team in a game, we need to output a probability that the first game would win. For all 64 games, we should output all combination of game which is $\binom{64}{2} = 2016$. Since we have the statistics of all past games, we could get the information of each team's performance in "WRegularSeasonDetailedResults" file for known features as training data and apply a machine learning method Supported Vector Machine to learn the linear separable decision boundary between wining team and losing team.
The results of 2019 game is not out yet so we can not test the accuracy of our model. Therefore we use the year starting from 2010 to 2017 as our training data and the year 2018 as the testing data for predication since we have the 63 testing results.

4.2.1 Data Pre-process

At the first glance, we decide to use past Tournament data as our training data since we assume that players are performing more than 100 percent in Tournament to get the championship rather than the regular season.
We treat every game every year as a brand new game. We concatenate the final game statistics, 13 for each team from the Field Game Made(FGM) to Personal Fouls Committed(PFC) as the descriptive vector for that team.
Therefore, for every game, the **feature X** is the 26 vectors composed of the first team statistics and the second team. The **label Y** 0 or 1 is the label indicating if the first team wins the second team. 0 for lose and 1 for win.
Note: Since the data provided is structured as wining team the first and losing team the second, which leads to a label of 1 for every game, we **change the order of two teams for odd index as the case for label 0** and **randomly shuffle** the data for the balance of training set.
For testing data, we need a vector describing the general performance of two teams as a relative comparison between them. Because the information in Tournament Game is limited, We utilize the 13 descriptive information in Regular season in 2017 as an evaluation for the performance of a team. For each team, we average the performance of it in 2017 regular season as a vector of length 13. Therefore the testing data for a game is a combination of two team vector with length 26. The length of testing data is 63, which are the 63 pairs of game in Tournament.

4.2.2 prediction-SVM

Since we only have 13 descriptive information of 2 teams, some of which is highly related to the performance of the game. For example, more Offensive Rebounds

indicates more opportunities to shot in a new around, therefore is more possible to get more points. We could therefore conjecture a positive association between OR and probability to win. The feature would therefore be linear separable.
Linear Support Vector Machine would therefore be a proper model for the classification task. More detailed description of SVM would be introduced in Theory part.
We could see that the most important coefficient is the first and the 14^{th} feature, which is the Field Goal Made by the first and the second team respectively. The coefficients indicates that the more FGM the first team made, and the less FGM the second team made, the more possible the first team would win the game.

```
clf.coef_
```

```
array([[ 0.77881826,  0.01938597,  0.37927752,  0.00609028,  0.36381826,
         0.0313458 , -0.02961228,  0.09639579,  0.0292662 , -0.04341945,
         0.01798265,  0.01111754, -0.02637719, -0.75828745, -0.04390671,
        -0.41841601,  0.00185671, -0.35790642, -0.03823474,  0.04678375,
        -0.07060259, -0.02257428,  0.04246143, -0.01606106,  0.00430183,
         0.01926071]])
```

Figure 14: feature weight in SVM

4.2.3 prediction-Random Forest

Random Forest is another method we applied. The input training data and the testing data is the same as the SVM model. Similar to SVM model, for random forest, we check if the coefficient of the model is intuitive. We see that the 1^{st} and the 14^{th} weights are more important than other variables as the average FGM is truly an important indicator of the result of a game.

```
: print(tree_clf.feature_importances_)
```

```
[0.14380534 0.01272644 0.01693188 0.01027305 0.06027263 0.04360455
 0.00915474 0.05192739 0.08636355 0.01841547 0.01588243 0.00799176
 0.0272193  0.15275456 0.01256205 0.01628492 0.01008662 0.05999621
 0.04855271 0.00895298 0.04633821 0.06714368 0.01906085 0.01592123
 0.00870082 0.02907665]
```

Figure 15: feature weight in random forest

4.2.4 prediction-Logistic Regression

Logistic regression is another widely-used model when the response is categorical. If there are two possible outcomes, we use the binomial distribution. Glmnet is a package that fits a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elastic net penalty at a grid of values for the regularization parameter lambda. In this

case, we take into account all of the variables except for the winning scores and losing scores. We input both the features of two teams and let the model to select the weight on each. We use **glmnet** and **elastic-net penalty** inside which alleviates regularizes and selects variables as well.

After we fit the glmnet with the **family='binomial'**, we also use the **10-fold cross validation**. -0

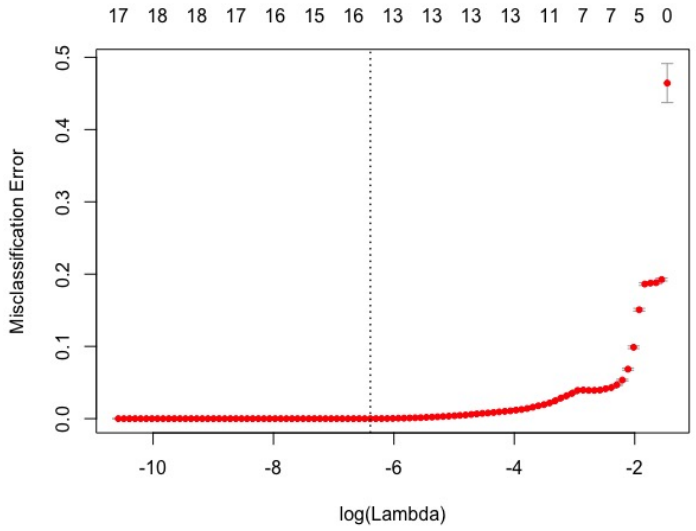

Figure 16: Error vs. lambda

In Figure 16, we plot the **misclassification error** and show the optimal value of , which is 0.00167673. We will then use the set that has the lowest error in our cross validation to do the prediction.

We get the coefficients under the minimum , then put the testing data and evaluate the result. We classify all the numbers ¿ 0.5 as 1, and 0 otherwise, and compare with the ground truth data from 2018 NCAA competition. We finally get **an accuracy of 0.6666667**.

4.2.5 prediction-Naive Model

Based on the above predication models, we decide to design a new formula calculating the score of teams in each game based on Field Goals Made, Three Points Made and Free Through Made. We find the average score S for a team should be $S = 2 * (FGM - FGM3) + 3 * FGM3 + FTM$
which is $S = 2 * FGM + FGM3 + FTM$
(since FGM is the total count of 2 points and 3 points goal made while FTM is 1 point goal)

We directly check if the calculated score of the first team is greater than the second team for the prediction of result. The final accuracy is actually pretty close to the prediction using Logistic Regression.

Model	Training Accuracy	Testing Accuracy
SVM	99.9	65.1
Random Forest	100.0	68.5
Logistic Regression	96.5	66.7
Naive Model	-	66.7

Table 9: Training and Testing Accuracy of Different Models

4.3 The relationship between the difference of rebound grabbed by two teams and the games' result

4.3.1 Hypothesis

H_0: there is no relationship between the difference of rebound grabbed by two teams and the games' result, such that all the coefficients are 0.
H_1: there is a relationship between the difference of rebound grabbed by two teams and the games' result, such that at least one coefficient is 0.

4.3.2 Data Pre-process

Before applying the regression, we first extract the columns of offensive and defensive rebound for both teams. Then, we add those rebounds to get the total rebound for both teams. After that, we get the rebound difference between two teams. (Winning team's rebound-Lost team's rebound) Also, we get the score difference between the two teams. (Winning team's score-Lost team's score) Then, we draw the scatterplot based on the data mentioned above.

4.3.3 Regression

For null hypothesis, we assume there is no relationship between the difference of rebound grabbed by two teams and the games' result. So, we decide to apply the linear regression line at first. However, based on the residual of TotalRebound (Figure 17) shown below, it has a pattern. So, we cannot apply the linear regression, and the linear regression line shown in Figure 15 does not fit the data indeed. Then, we apply the quadratic regression, which fits the data better.
Furthermore, from the scatterplot, we find an outlier on the left of the plot. Maybe in this match, the winning team performs better on other aspects, like field scoring or hit rate.

Based on the Table 11, we find that the p-value of for β_1=0 and β_2=0 is very small. So, we reject the null hypothesis.

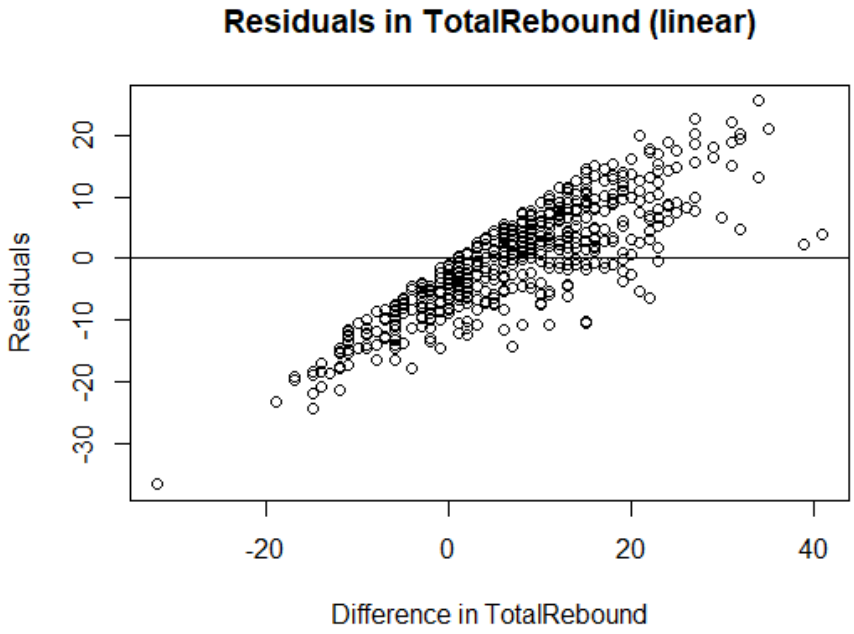

Figure 17: residuals in TotalRebound

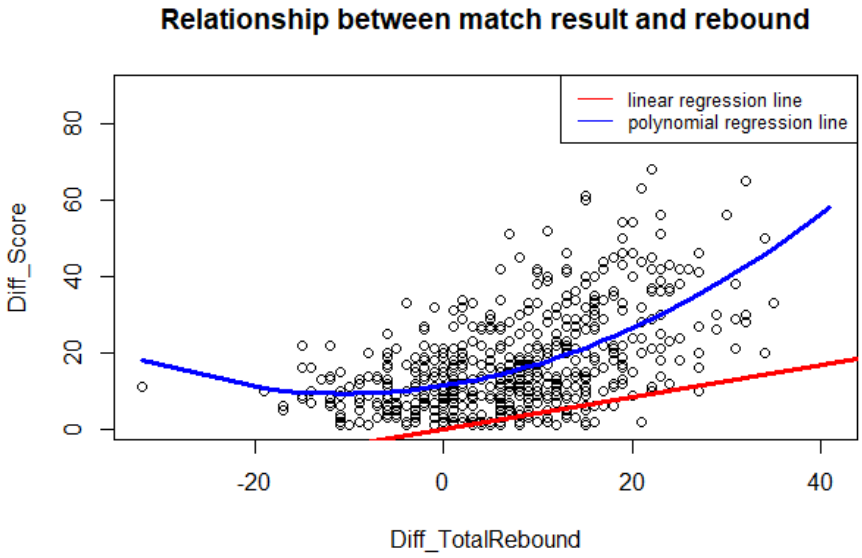

Figure 18: Relationship between match result and rebound

type of regression line	β_0	β_1	β_2
linear	0.09289	0.41638	none
polynomial	17.14	166.66	69.57

Table 10: The value of the coefficient and corresponding P-value for each regression line

type of regression line	p-value (β_1)	p-value (β_2)
linear	<2e-16	none
polynomial	< 2e-16	3.28e-10

Table 11: The P-value of coefficient for each regression line

5 Theory

Two sample proportion test: a paired difference test is a type of location test that is used when comparing two sets of measurements to assess whether their population means differ.

standardized residual formula: $\frac{SampleCount - ExpectedCount}{\sqrt{Expectedcount}} = \frac{N_j - \mu_j}{\sqrt{\sqrt{\mu_j}}}$

Support-vector machine (SVM): A set of supervised learning methods used for classification, regression and outliers detection. A SVM model represents the examples as points in space, so that the examples of the separate classes are divided separately. The objective function includes the minimization of margin between two classes, which leads to a robust decision boundary.

Classification Trees: Each internal node represents a value query on one of the variables, and each internal node represents a value query on one of the variables. The tree is grown using training data, by recursive splitting. New observations are classified by passing their X down to a terminal node of the tree, and then using majority vote.

Random Forest: A supervised learning algorithm for tasks (such as classification and regression) that need to construct a decision trees at training time. This algorithm outputs the class that is the mode of the classes or mean prediction of the individual trees.

KS test: A non-parametric test of the equality of continuous, one-dimensional probability distributions that is used to examine if a sample follows some distribution in some population.

Two sample t-test: A type of location test that is used when comparing two sets of measurements to access whether their population means differ.

Monte-Carlo simulation (Monte Carlo Method): A computerized mathematical technique that is used to predict numerical results through repeated random sampling. Due to the intervention of random variables, the probability of different outcomes can be difficult to predict.

Mann Whitney U Test (Wilcoxon Rank Sum Test): A non-parametric alternative test used to test the equality of means in two independent samples. A non-parametric test is appropriate when the dependent variable is not normally distributed.

Linear regression: A linear model of the relationship between a dependent variable and independent variables.

Polynomial regression: a model for a single predictor X and expected value Y that is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_n X^n + \epsilon$$

Elastic net: A Linear regression with combined L1 and L2 priors as regularizer.

Logistic regression: A statistical model using a logistic function to model a binary dependent variable. Logistic Regression is used when the dependent variable is categorical.

6 Conclusion

For hypothesis test, we find that **the home court condition** will make players more likely to win, and it will result in larger score difference.
For the prediction part, we find out that even though the training accuracy is almost at 100, there is a gap between the training and testing accuracy, which is the **over-fitting** problem.
Two possible issues would hide behind the gap: first is that the relationship between the feature and the final performance **may not be linear** . As our third hypothesis test illustrates a **quadratic relationship** between the Difference of rebound and the final score difference.
Second, the performance of the team would not only depend on past performance. **The emergency situation** such as injury and psychological issue of the players would not be predicable by past statistics. To better predict the performance of the teams on tournament game, we need a more comprehensive investigation with more aspects of the Team included.

7 Works Cited

Heit Evan, Price Paul C., and Bower Gordon H., *A Model for Predicting Outcomes of Basketball Games*. faculty.ucmerced.edu/sites/default /files/eheit/ files/basketball.pdf.

Jones Eric Scot, *PREDICTING OUTCOMES OF NBA BASKETBALL GAMES*, North Dakota State University of Agriculture and Applied Science, 2016

Spann, Martin, and Bernd Skiera. "Sports Forecasting: a Comparison of the Forecast Accuracy of Prediction Markets, Betting Odds and Tipsters." *Journal of Forecasting*, vol. 28, no. 1, 2009, pp. 55–72., doi:10.1002/for.1091.