Google Cloud and NCAA ML Competition 2019-Women's

Zijing Wang, Xuanyu Liang, Kevin Xue, Xiao Wang, Jiacheng Shi, Minsheng Liu, and Junxian Tan.

March 25, 2019

1 *Introduction*

NCAA Division 1 Women Tournament is a basketball tournament that is played between 64 collegiate basketball games. The tournament is usually played in March of each season and the tournament has been popular across the country. Teams that enter the tournament are selected by their overall performance and ranking in different conference games. The first round of the tournament will commence in the late March and the championship game will usually take place in the early April. In the project, we are given data of both regular seasons result and tournament results from the last few years and asked to predict the result of the 2019 NCAA Division 1 Women Basketball Tournament.

For this project, we will first explore the data and discover some potential hypothesis, which we will use statistical methods for validation. After the conclusion of the hypothesis is drawn, we will incorporate them in predicting the 2019 tournament results. In the project, we will first analyze if there is a difference between playing at home and away; or equivalently, testing if there exist home field advantages. For this hypothesis, a numerical summary of the data will be used and a histogram will be drawn before we conduct a hypothesis test. Then we will analyze the difference between a team's performance in regional conferences and the tournament by using graphical and testing methods. Furthermore, we will use different statistical methods, graphical methods, and testings to see if there exist a difference in a team performance due to regions, break-time between consecutive games, and competitors. After the result from these hypotheses is drawn, we will incorporate the results to predict tournament outcomes.

2 *Background*

The following sections will provide information about the background of the study and will include additional research for this project.

NCAA Division 1 Women Basketball Tournament

NCAA Division 1 Women Tournament is a basketball tournament that usually takes place in March of each season. The Women's NCAA Division 1 has gained people's interest and attention over time; it is considered one of the most influential basketball games in the country. NCAA Division 1 Women Tournament is played between 64 collegiate basketball teams. Due to a large number of teams participating, the tournament will be in the format of a single elimination tournament. In *Women's Basketball: Road to the Championship*, NCAA has explained their selection process: 32 teams (Automatic Qualifiers) will receive automatic qualification determined by their conference tournament results and the rest 32 teams (At-Large) will be selected by the NCAA selection committee with some criterions, including their conference records, overall records, strength of the conference, overall record and so on. In the webpage *Selection Criteria*, NCAA has also mentioned how "seeding" works. The NCAA selection committee will create a seed list of 64 teams to offer audience information about teams' competitive ranking on a national scale. The tournament is usually played in March of each season and the tournament has been popular across the country. The first round of the competition will commence in the late March and the championship game will usually take place in the early April.

Game Result Prediction

Researchers have been using past games' statistics to analyze the correlation between some statistics and game output. Multiple studies have shown that the seedings provided by the NCAA are strong indicators of tournament championship since the seeds are assigned based on a team's performance in regional conferences. In *A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament: Updated Results from 2007*, West points out that there may be bias in the seeds assignment as selection committee may have subjective opinions and hence it is beneficial to use statistical methods to develop rating methods to conduct game prediction (2). In the article, West has presented some potential rating methods for predicting, including Rating Percentage Index, a rating that compares team based on its winning percentages and its opponents' and Jeff Sagarin's Computer Ratings (3). Furthermore, West suggests that an ordinal linear regressing modeling and expectation (OLRE) could be an efficient method since it is simple for calculation and flexible in terms of dating usage. In *Heterogeneous Skewness in Binary Choice Models: Predicting Outcomes in the Men's NCAA Basketball Tournament*, Caudill and Godwin have derived a method that is using the skewed logit model that

allows heterogeneity in the skewness parameter and find their model have better fitting when predicting tournament outcomes.

Advanced Statistics in Basketball

In this project, we will use some basketball-specific ratings to analyze the performance of individual teams.

$$\text{Possessions} = 0.96 \times [\text{FGA} + \text{Turnovers} + (0.475 \times \text{FTA})] - \text{Off Rebounds}$$

$$\text{Defensive Rating} = 100 \times (\frac{\text{Opponent's Score}}{\text{Possessions}})$$

$$\text{Offensive Rating} = 100 \times (\frac{\text{Score}}{\text{Possessions}})$$

$$\text{Net Rating} = 100 \times (\text{Offensive Rating} - \text{Defensive Rating})$$

$$\text{Free Throw Rate} = \frac{\text{FTA(Free Throws Attempted)}}{\text{FGA(Field Goal Attempted)}}$$

3 *Data*

The data of this study is composed of the games boxscores from 1998 to 2018, including boxscores results from both regular seasons and tournaments. Detailed box scores are provided for the tournaments between 2008 and 2018 and regular seasons between 1998 and 2018. There are 10 files in the data folder that include information about the season, location, TeamID, seeds, and specific team box scores. Table 1 lists all fields and their descriptions.

4 *Analysis*

Does there exist home-field advantage?

Multiple studies and researches have shown that there are home-advantages in competitive sports.In this section, we are going to examine whether there are home advantage in NCAA Women Basketball games over the last decade including both regular seasons and tourneys.

In figure 1, we observe that it is more likely to win at home than away. In figure 2, this is also true for tourney matches. In fact, we see the difference is even bigger. Since the performance of WNCAA teams varies within the home and away games, home advantage by teams is hereby investigated by analyzing home-win rate and away-win rate. Examining the difference between home and away games is more optimal than just analyzing home-win rate solely because we could have more intuition on home field advantage.

Field	Description
`TeamID`	Identification number (4-digit) of each NCAA Women team
`TeamName`	Name of the team
`Season`	Year of which the tournament is played
`DayZero`	The date corresponding to daynum=0 of the season
`Region`	in W, X, Y, Z; identifier of the region
`Seed`	The team's seed in one of the corresponding region
`DayNum`	The number tells what day the games took place with respect to DayZero
`WScore`	Points Scored by the winning team
`LScore`	Points Scored by the losing team
`NumOT`	The number of overtime in the game
`WLoc`	The location of the winning team (H: home, A: away N:neutral court)
`WFGM`	Winning Team's Field Goals Made
`WFGA`	Winning Team's Field Goals Attempted
`WFGM3`	Winning Team's Three Pointers Made
`WFGA3`	Winning Team's Three Pointers Attempted
`WFTM`	Winning Team's Free Throws Made
`WFTA`	Winning Team's Free Throws Attempted
`WOR`	Winning Team's Offensice Rebounds
`WDR`	Winning Team's Defensive Rebounds
`WAst`	Winning Team's Assists
`WTO`	Winning Team's Turnovers Commiteed
`WStl`	Winning Team's Steals
`WBlk`	Winning Team's Blocks
`WPF`	Winning Team's Personal Fouls Committed

Table 1: All fields and their descriptions.

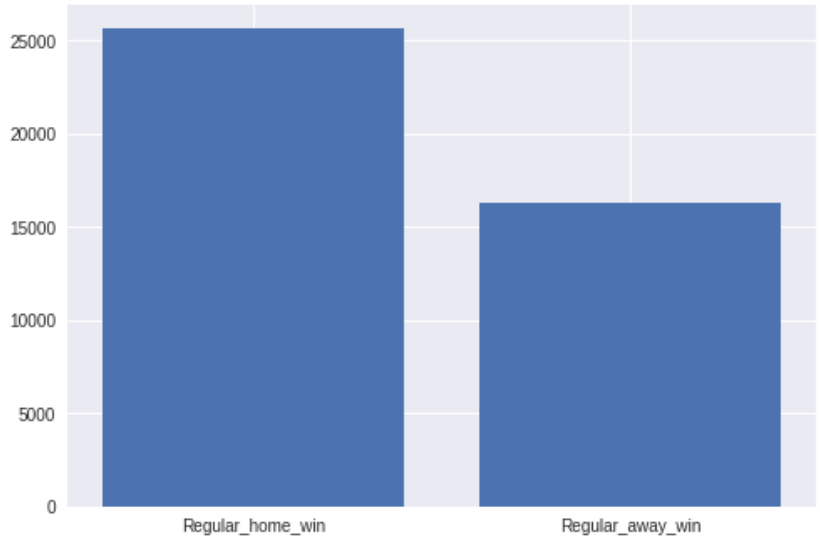

Figure 1: Histogram of number of games win at home or away during regular seasons

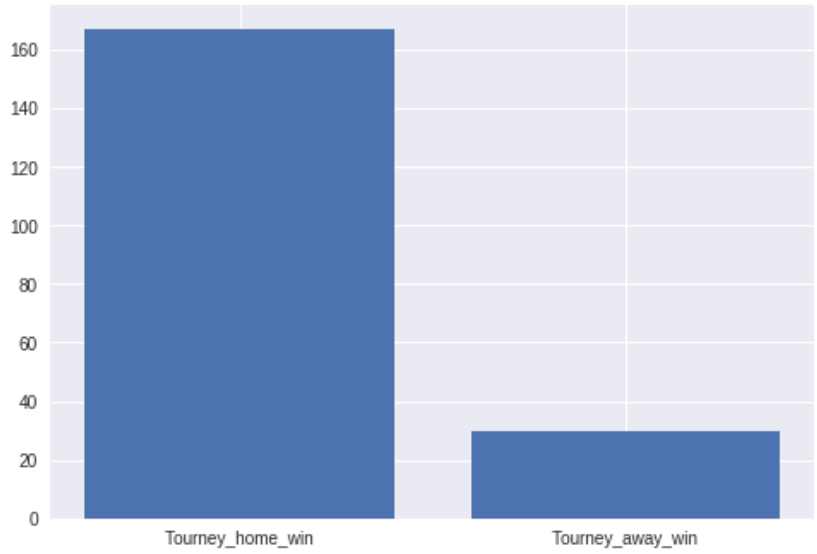

Figure 2: Histogram of number of games win at home or away during tourney

In order to ensure the accuracy of the experiment, we eliminate the influence of competition in a neutral court. Then, we will cover basic results regarding confidence intervals under binomial distribution.

Hypothesis 4.1. Let p be the home-win rate. H_0: $p = 0.5$; There does not exist home-field advantage. H_1: $p > 0.5$ There exist home-field advantage, that is the home team have higher chance of wining.

From Table 1 and Table 2, we observe that the home-winning percentages are higher than 0.50 in both regular seasons and tourneys in 2010-2018 seasons, indicating that home teams have higher chance of winning the game than away teams.

Table 2 records the home/away performance discrepancy over the last decade, done at $p = 0.5$.

Season	Away	Home	Lower	Estimate	Upper	p-value
2010	1770	2912	0.608	0.622	0.636	4.851E-63
2011	1748	2836	0.604	0.619	0.633	1.535E-58
2012	1755	2738	0.595	0.609	0.624	5.622E-49
2013	1814	2743	0.588	0.602	0.616	2.732E-43
2014	1789	2787	0.595	0.609	0.623	1.520E-49
2015	1844	2981	0.604	0.618	0.632	1.147E-60
2016	1858	2855	0.592	0.606	0.620	4.849E-48
2017	1864	2984	0.602	0.616	0.629	1.258E-58
2018	1878	2835	0.587	0.602	0.616	2.257E-44

Table 2: Results of the binomial test done for home/away performance discrepancy in the regular seasons over the last decade.

Table 3 gives the result for tourney.

Season	Away	Home	Lower	Estimate	Upper	p-value
2010	2	21	0.720	0.913	0.989	6.604E-05
2011	6	19	0.549	0.760	0.906	1.463E-02
2012	1	5	0.359	0.833	0.996	2.188E-01
2013	5	22	0.619	0.815	0.937	1.514E-03
2014	2	14	0.617	0.875	0.984	4.181E-03
2016	6	27	0.645	0.818	0.930	3.241E-04
2017	4	30	0.725	0.882	0.967	6.165E-06
2018	4	29	0.718	0.879	0.966	1.093E-05

Table 3: Results of the binomial test done for home/away performance discrepancy in the tourneys over the last decade.

In Table 1 and Table 2, the p-value are all less than 0.5. With significance level α=0.05, we reject the null hypothesis and accept the alternative. By conducting a 95% confidence interval, we are 95% confident that the true home-win rate p is higher than 0.50. Hence, from the one-sided hypothesis test, we conclude that there exist home advantage over the last decade during both regular seasons and tourneys. The results confirmed our assumption, which is competitive sports have home-advantage. Since the home team have psychological advantages and more familiar with the environment at home, they can pay more attention on games rather than struggling on their travels.

Are Tourney Matches More Intense?

Our next question is whether matches in the tourney are more intense than those in the regular reason. Since teams who made into the tourney will be stronger than their not so fortunate peers, it is reasonable to expect that matches between those stronger teams are more competitive. To measure COMPETITIVENESS, we look at the score difference between the two teams in a match: the closer the difference, the more competitive the match is.

If neither team in a match made into the tourney, both are not as strong and hence their score difference could still be small. Hence, for the regular season we only look at matches where one of team made into the tourney. Table 4 shows statistics of score differences in matches from 2010 to 2018.

Stage	Mean	Variance	Skewness
Regular	17.237	166.534	1.290
Tourney	17.139	175.403	1.517

Table 4: Mean, variance, and skewness of score differences in matches from 2010 to 2018.

As stated above, we believe that matches in the tourney should be more competitive. Formally, our null hypothesis is

Hypothesis 4.2. The mean of score differences in the selected matches

from the regular season follow the same distribution as that from the tourney.

Before the test, we use a bar plot to visually compare the mean of score differences, where our data are divided by year, shown in figure 3. Surprisingly, we do not see much discrepancy between the mean from the regular season and that from the tourney. In year 2010 to 2013 as well as 2016, the mean from the regular season is higher than that from the tourney, but the situation is reversed in other years.

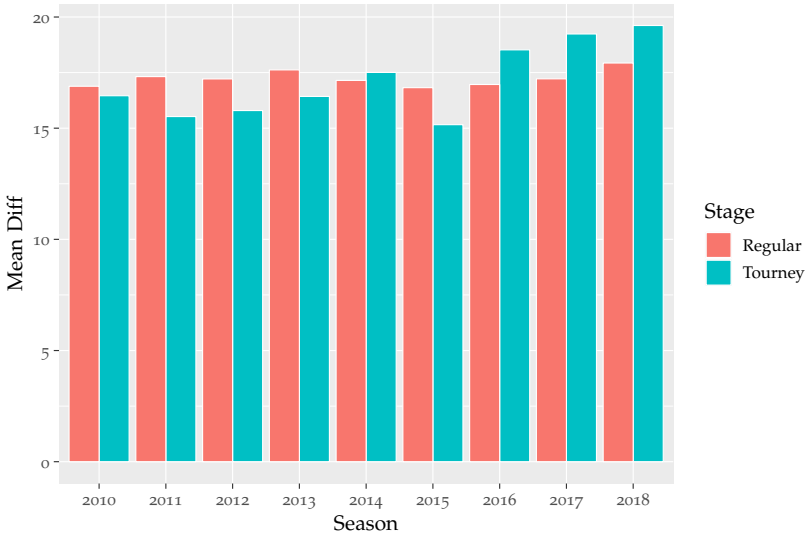

Figure 3: Discrepancy of the mean difference between the regular season and the tourney, separated by year.

In any case, there *is* difference between the regular season and the tourney. We perform the chi-squared test to see if such difference is significant. The test is done separately for each year. Table 5 shows the test results.

Season	Regular	Tourney	χ^2	p-value
2010	16.886	16.460	0.005	0.941
2011	17.321	15.524	0.098	0.754
2012	17.217	15.794	0.061	0.804
2013	17.623	16.429	0.042	0.838
2014	17.149	17.508	0.004	0.951
2015	16.823	15.159	0.087	0.769
2016	16.966	18.524	0.068	0.794
2017	17.220	19.238	0.112	0.738
2018	17.931	19.619	0.076	0.783

Table 5: Results of the chi-squared tests.

As we can see, for all years $p > 0.05$, meaning that we cannot reject hypothesis 4.2, and that we are 95% confident that the mean score

difference between regular season and the tourney is the same. As a conclusion and an answer to our section's question, the competitiveness in regular seasons and tourneys is, contrary to our intuition, not significantly different.

Does a strong team perform better when facing a strong competitor?

In this section, we want to see if there is a difference between match of two powerful teams and match of not too powerful teams. We used the total score of two teams to compare. We will see from the next section that there is a relationship between offensive rating and wins. Offensive rating is a direct consequence of how many points scored. We define strong teams as the teams who entered the tourney that year. Weak teams as the teams that did not enter the tourney that year.

In Figure 4,5 and 6, we can see that in 2016, 2017, and 2018, the matches between strong teams will have a higher score total than the matches between weak teams. The next thing we want to do is to see whether there is a difference between strong teams and weak teams.

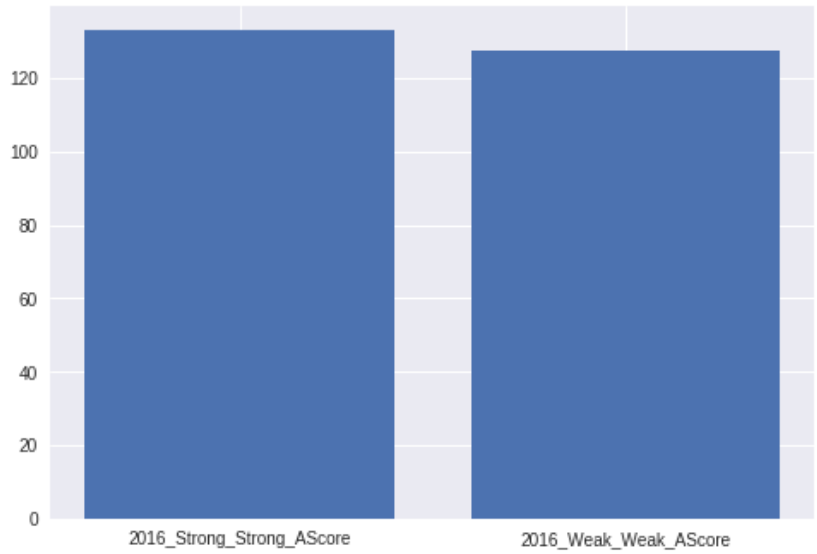

Figure 4: Histogram: mean of total scores in games between two strong teams and two weak teams in 2016

We draw the distribution of strong vs strong teams and weak vs weak teams. Figure 7 is strong vs strong, and Figure 8 is weak vs weak. From these two diagrams, we observe that these two distributions might be normal. So we performed kurtosis and skewness test to test for normality and the results are displayed in Table 6. The result of adjusted kurtosis test (normal will have a result of 0 instead of 3) showed that we are unsure about the distribution of the second one being normal. So we used the K-S test to check since K-S test is a more

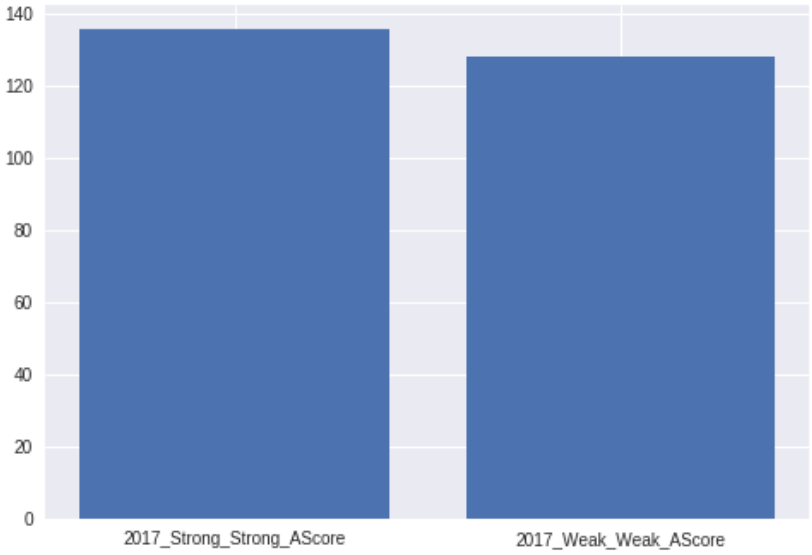

Figure 5: Histogram: mean of total scores in games between two strong teams and two weak teams in 2017

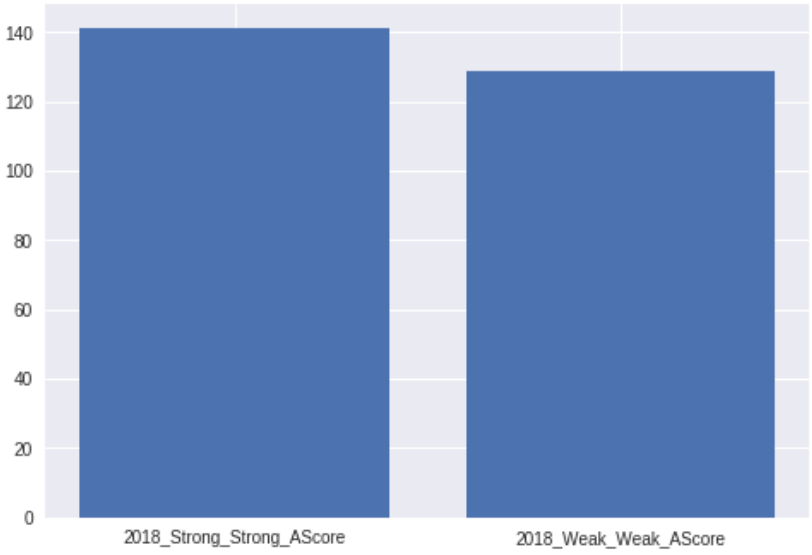

Figure 6: Histogram: mean of total scores in games between two strong teams and two weak teams in 2018

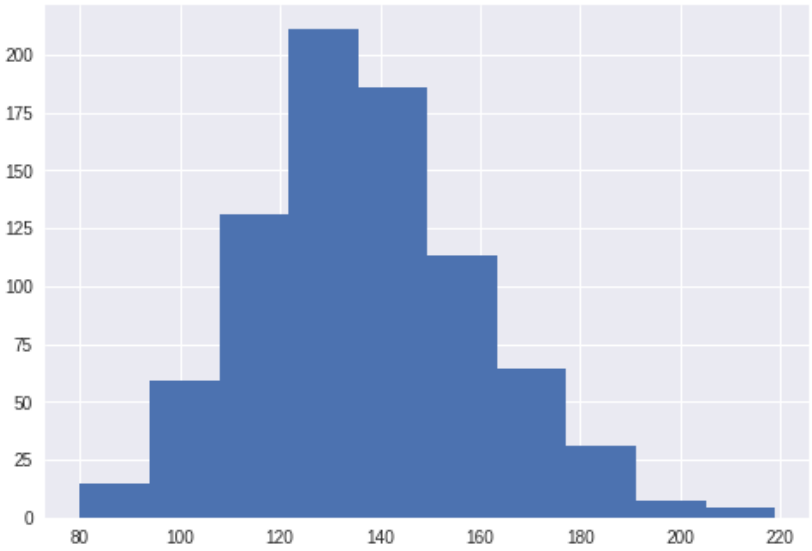

Figure 7: Histogram: distribution of points total per game between two strong teams

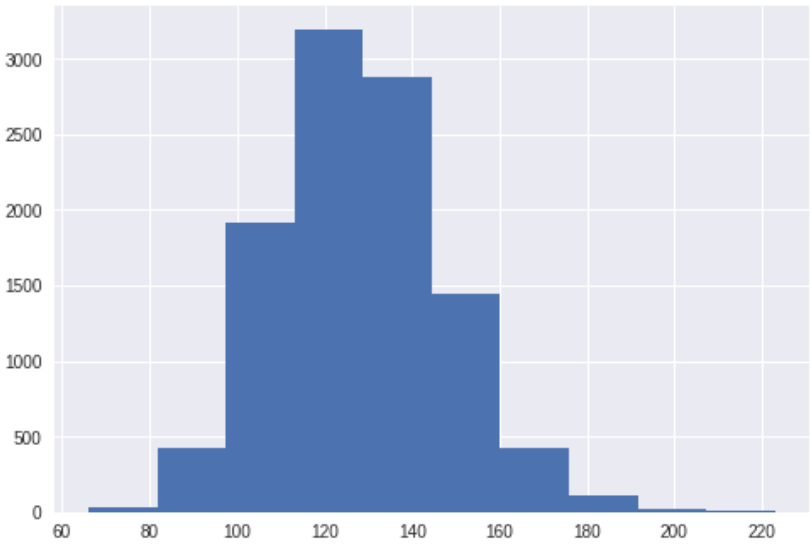

Figure 8: Histogram: distribution of points total per game between two weak teams

powerful test to test for normality in general. We set the null hypothesis of K-S test to be the distribution is normal. The test result gives us a p-value less than 0.05, so we rejected the null hypothesis. The second distribution is not normal.

	Skewness	Kurtosis
Strong vs Strong	0.381942	0.111631
Weak vs Weak	0.396316	0.480225

Table 6: Results of the Skewness test and Kurtosis test

	K-S Statistics	p-value
Strong vs Strong	0.042898	0.094649
Weak vs Weak	0.035909	3.98534e-12

Table 7: Results of Kolmogorov–Smirnov test

var(strong vs strong)	0.045319
var(weak vs weak)	0.037764
t-test stat	944.7691
p-val	0.0

Table 8: Results of Welch-t test(bootstrap)

Next, we tried to perform Bootstrap algorithm to get a normal distribution by using the central limit theorem. We calculated the mean of each bootstrapping result and get 1000 mean from 1000 times of bootstrapping in order to get a normal distribution. We first want to use two sample t test, but realized that the variance is different, so we used Welch t test instead. We get a p-value close to 0, so we being able to reject the null hypothesis. Therefore to conclude that the mean of the total score between strong teams are different than the mean of the total score between weak teams. This might because when a team is facing a strong team, it is more likely to encourage them to perform well because they want to prove themselves.

Offensiveness and Winning

We want to investigate the relationship between offensiveness and the wins. First, we introduce a common statistic in basketball games, "offensive rating".

$$\text{Possessions} = 0.96 \times [\text{FGA} + \text{Turnovers} + (0.475 \times \text{FTA})] - \text{Off Rebounds}$$

$$\text{Offensive Rating} = 100 \times (\frac{\text{Score}}{\text{Possessions}})$$

In general, offensive rating is used to define how "offensive" a team is. Since some teams tend to focus on defense, and others focus on offense, we want to know what competitive teams tend to be.

In figure 9, x-axis is each team's offensive rating and y-axis is its corresponding wins in a specific year. Although there are plots in several years, the fitted line and the trend is very similar so we will take an example of 2013.

In the linear regression, we obtained th following equation: wins = 0.66*Offrtg-45.70 with R-adj = 0.735. This shows a moderately strong relationship between offensive rating and wins.

In figure 9, the residual plot shows residuals are randomly scattered, no pattern is observed, so the linearity is valid. The regression line indicates that on average, every increase in offensive rating is associated with an increase of 0.66 wins. The more offensive the team is, the more the team wins

offensiveness and the wins. First, we introduce a common statistic in basketball game, "offensive rating". Offensive Rating = 100 x (Score / Possessions) Possessions = 0.96 x (FGA + Turnovers + (0.475 x FTA) - Off rebounds In general, offensive rating is used to define how "offensive" a team is. Since some teams tend to focus on defense, and others focus on offense, we want to know what competitive teams tend to be.

In figure 9, x-axis is each team's offensive rating and y-axis is its corresponding wins in a specific year. Although there are plots in several years, the fitted line and the trend is very similar so we will take an example of 2013.

In the linear regression, we obtained the following equation: wins = 0.66*Offrtg-45.70 with R-adj = 0.735. This shows a moderately strong relationship between offensive rating and wins.

In figure 9, the residual plot shows residuals are randomly scattered, no pattern is observed, so the linearity is valid. The regression line indicates that on average, every increase in offensive rating is associated with an increase of 0.66 wins. The more offensive the team is, the more the team wins

What affects the overall chance of winning?

In the last section of analysis, we want to investigate what types of data have a great impact on the chance of winning. The types of data we are looking into include: offensive rating(efficiency), defensive rating(efficiency), net rating, and free throw rate.

The first two diagrams are offensive rating and defensive rating. We can see that the winning team will have a better offensive rating since the winning team and all team have the same distribution, but the curve shifts to the right as a whole for winning teams. On the other head, defensive rating might not have an affect on the outcome of the game. The distribution and curve on defensive rating are identical. The net rating is calculated by offensive rating minus defensive rating.

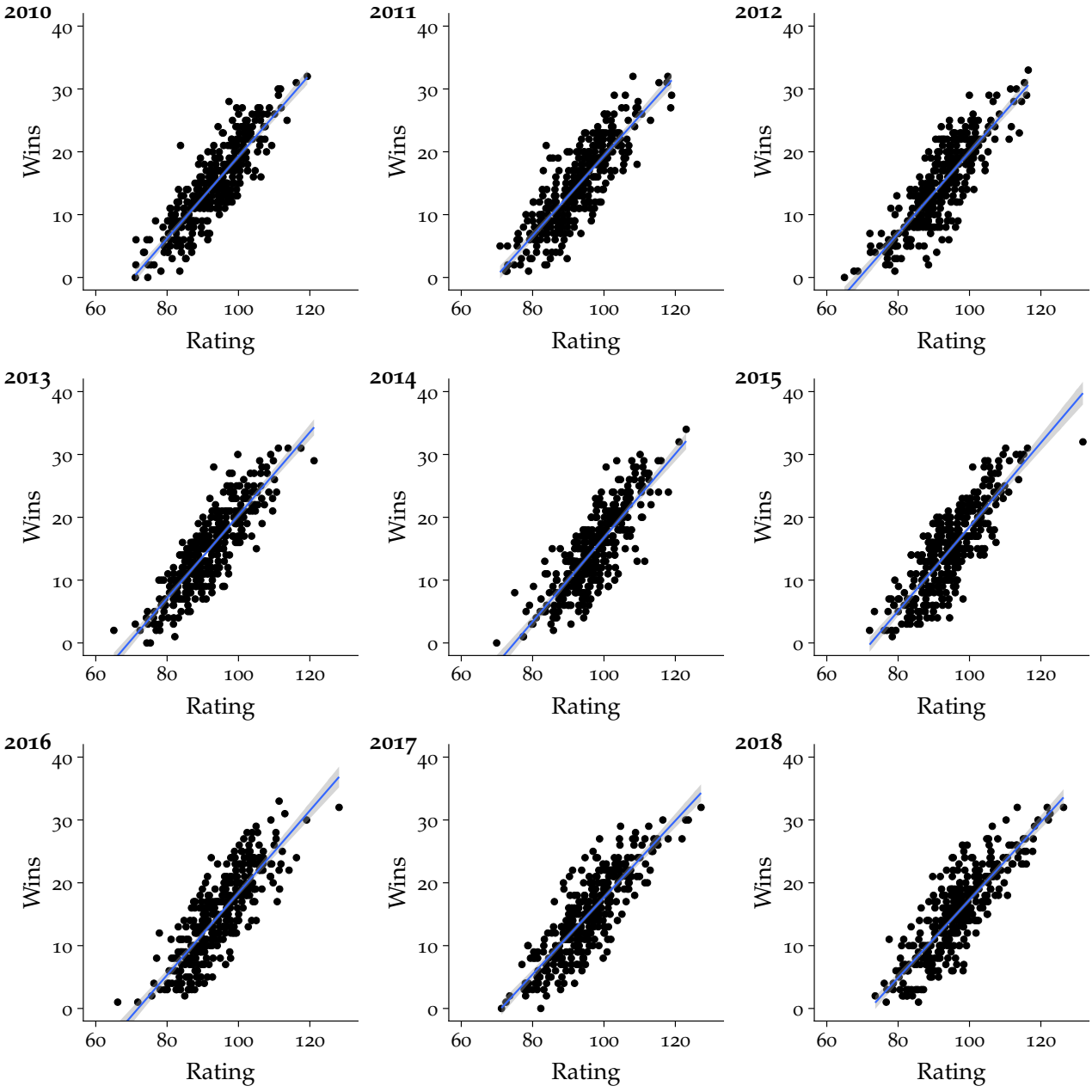

Figure 9: Regression of offensive ratings and winning counts.

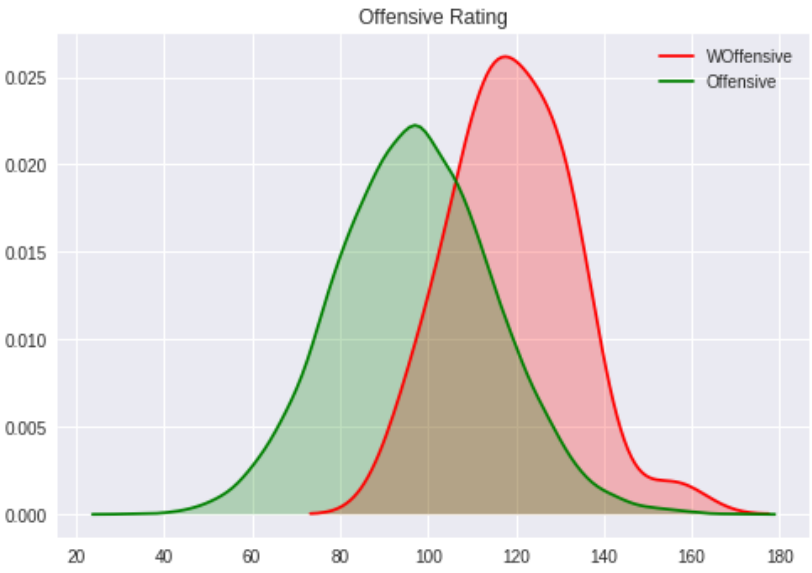

Figure 10: Histogram: Comparison between winning team's offensive rating and the losing's

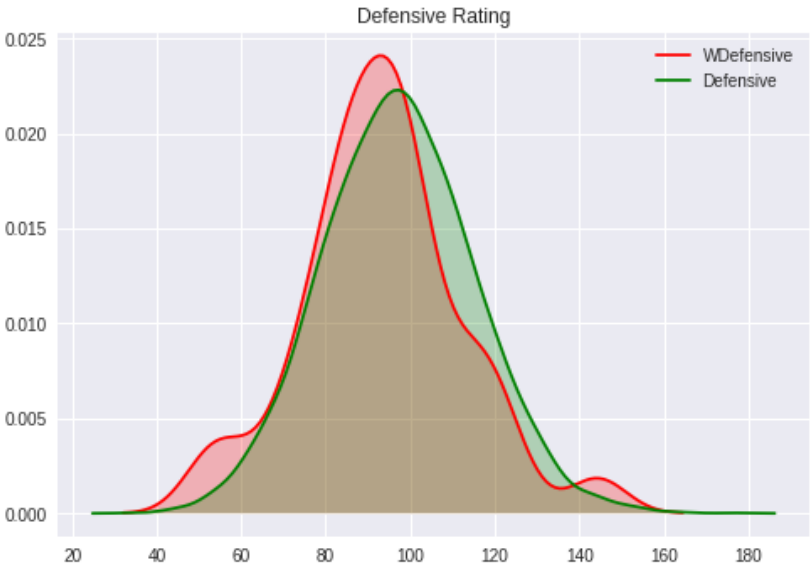

Figure 11: Histogram: Comparison between winning team's defensive rating and the losing's

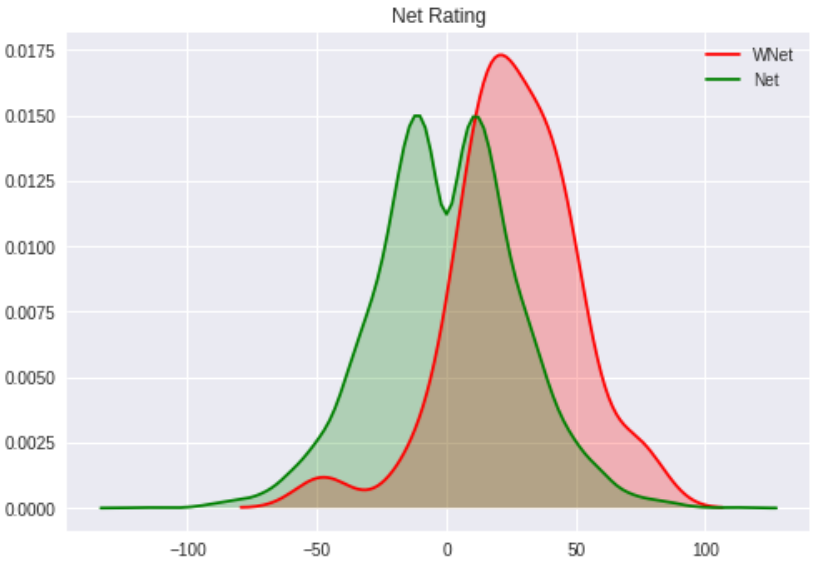

Figure 12: Histogram: Comparison between winning team's net ratings and the losing's

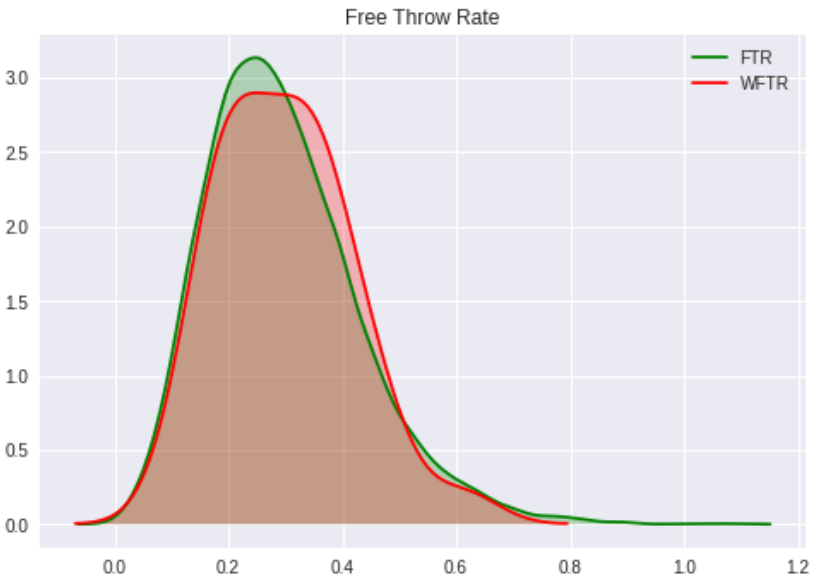

Figure 13: Histogram: Comparison between winning team's FTR and the losing's

It is interesting that the net rating for all teams do not have a normal distribution. This might because it is unlikely for a team to have a zero net rating, since a zero net rating means that this team gets and losses the same amount of points at the same time. But we can see that winning team still will have a higher net rating than all the net rating.

The last thing we look at is the free throw. As we can tell from the graph, there is no significant difference between winning teams and all the teams. But the losing team do have a larger tail to the left, and a higher peak.

5 *Prediction: Who Would Win in the Tourney?*

After our exploratory data analysis, we are now well-quipped for our ultimate goal: predicting who would win in the tourney. In this section, we employ machine learning techniques to solve the problem. We would prepare a common training set and evaluate three models: support vector machine with linear kernels, random forest, and AdaBoost.

We focus on predicting the result for any given match in the tourney, given the two participating teams. The final winner can be predicted by using our estimated predictors repetitively following the tourney rules.

The core idea behind our approach is to build a profile for each team using their performance in the regular season. We assume that there exists a prediction function, which takes two teams' profile as input and output the winner. Our work here is to estimate this function.

In the first hypothesis, we test whether the location of the match has impact on the match results, and the conclusion is positive. As a result, we, in addition to the team profile, also supply location information to our predictor.

Which information should be used to build the team profile? Machine learning algorithms are good at finding important features among all available variables automatically. For that to work, however, we need to supply enough information. Moreover, we need to avoid providing useless information.

Our second and third hypotheses shows that match score could fluctuate depending on whether a team is strong or weak. However, any team that made into the tourney is strong, and the absolute value of score does not determine who wins and who loses. Therefore, we decide to leave out match score from our team profile.

Our fourth and fifth hypotheses checks whether advanced metrics give a good insight in the result of a match. While information like offensive rating is correlated with the outcome a match, some others are

not so helpful. Therefore, instead of doing manual feature engineering and selection, we provide all information required to compute such ratings, like field goal, three pointer, free throws, and so on. The rest of job is delegated to machine learning models.

One might question whether machine learning models are powerful enough to derive all advanced metrics. On the one hand, we need to be cautious regarding the approximation power to avoid overfitting. On the other hand, metrics like offensive rating are nothing but linear combinations of features we use to build the team profile, and it is reasonable to believe that the three machine learning models can learn such linear relationships.

Team performance in any single match could be subject to many variables and have a high variance. It is natural to build a team's profile using records from all matches it plays during the whole reason. For each record, we need to summarize its change in all matches during the season. We choose to compute the mean, median, standard deviation, and skewness for each record, which, hopefully, would provide sufficient information for these models to obtain a good approximation to the changes of those records over the season.

One *pitfall* is the attempt to use results *across* seasons. Since the team members change from year to year, it is reasonable to believe that team profiles also change from year to year. While in two consecutive season a team's profile might not change much, it surely does between, say 2010 and 2018. Hence, for the simplicity of our model, we treat a single team in two different seasons as essentially two teams.

Data Preparation

We first assign each team in each year a unique identifier, which means each team would occur nine times. Then, we compute the mean, median, standard deviation, and skewness for each advanced metric. There are 13 advanced metrics, and we obtain a $13 \times 4 = 52$ dimension vector for each team serving as the team profile.

Then, for each match result in the given data, we get the team profile vector for both teams. We append the row with a two-dimension vector denoting which team is the host. It is similar to the one-hot vector, with a slight difference that both dimensions can be zero if the match happens in a neutral place.

Lastly, we generate the labels, naming the match result. Notice that the data are organized that the first team is the winning one, but this might allow machine learning models to over-fit. Hence, we randomly toss a fair coin and decide for each match, whether the winning or the losing team is the first one.

Results

We reserve 25% of the whole data (about 40k matches) as the validation set. We evaluate three models: linear support vector machine with hard margin, random forest, and AdaBoost. Due to the prohibitively high running cost, we only train these models on a small portion of the training set, but evaluating them on the full validation set.

We choose to train on $500, 1000, \ldots, 7500, 8000$ matches for all three models. Figure 14 summarizes the performance of all three models.

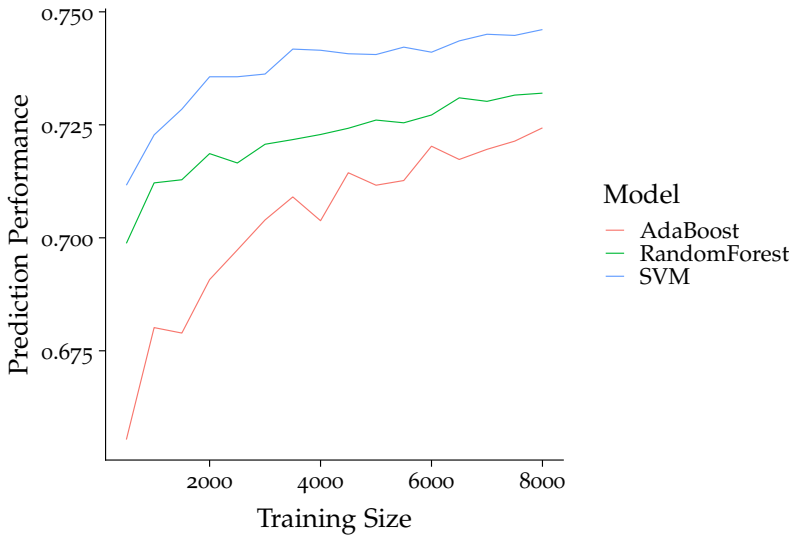

Figure 14: Evaluation results on the validation sets for each model with the given training data size. The y-axis represents the portion of correct labels predicted by the trained model, where 0.5 corresponds to the baseline—randomly guessing the result—and 1.0 the oracle.

The model that performs the best is linear support vector machine, followed by random forest, and the worst-performing one is AdaBoost. The final performance for the three models, where the training set contains 8000 matches, are 74.6%, 73.1%, and 72.6% respectively.

The performance for all three models increases steadily with more training data. Moreover, from the figure we *guess* that we have not yet reached the asymptotic line, that more computational power we can have even better performance.

6 *Theory*

Binomial Distribution: Binomial distribution is a probability distribution for discrete random variables. Binomial distribution gives the probability distribution of x success out of n Bernoulli trials with success probability p. X denotes the number of successes in n Bernoulli

trials.

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

MLE OF BINOMIAL DISTRIBUTION: For a binomial distribution with parameter n and p, the probability mass function is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

In order to calculate the MLE, we first derive the likelihood function:

$$L(p) = \prod_{i=1}^{n} \binom{n}{x} p^{x_i} (1-p)^{n-x_i}$$

Second, we calculate the log-likelihood, and denote constant by C:

$$\ell(p) = \ln(nC) + \sum_{i=1}^{n} x_i \ln(p) + (n - \sum_{i=1}^{n} x_i) \ln(1-p)$$

Then, take the first and second derivative of the log-likelihood function:

$$\frac{d\ell}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p}$$
$$\frac{d^2\ell}{dp^2} = \frac{-x}{p^2} - \frac{n - \sum_{i=1}^{n} x_i}{(1-p)^2} < 0$$

At last, set the first derivative equals to zero and derive the MLE as the following:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

OFFENSIVE EFFICIENCY: For calculating the offensive efficiency, we count the number of points a team scores every 100 possessions.

DEFENSIVE EFFICIENCY: For calculating the defensive efficiency, we count the number of points a team allows every 100 possessions.

MEAN: the mean of the sample is taking the sum of all the data in sample and divide it by sample size, the mean of X is usually denoted by $\bar{X}$,

MEDIAN: it is a value that indicates the middle of the data sample, located in the middle of the data sample.

STANDARD DEVIATION: Standard deviation is the summation of the distance between data points and the median of the data.

VARIANCE: The variance is the square of the standard deviation

QUANTILE: The First Quantile represents the cutoff point between the first 25 percent of the data and the remaining 75 percent of the data.The Third Quantile represents the cutoff point between the first 75 percent of the data and the remaining 25 percent of the data.

LINEAR REGRESSION: Linear regression is a linear approach to model the relationship between two variables.

SKEWNESS AND KURTOSIS: Skewness and kurtosis measure the normality of the probability distribution of a real-valued random variable. It is defined as follows

$$\text{skew}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S}\right)^3$$
$$\text{kurt}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S}\right)^4$$

where $\bar{X}$ is the expectation of X and S is the standard deviation:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 .$$

For a normally distributed random variable X, $\text{skew}(X) = 0$ and $\text{kurt}(X) = 3$.

HISTOGRAM: Histogram is a representation of data by its frequency. It is usually used as a method to understand the distribution of a continuous variable. By dividing data into bins, histogram shows the density of data.

BOX PLOT: Box plot is a graphical method to visualize a set of data in quantiles. The maximum, minimum, median, first quantile , third quantile, and outliers will be shown on the boxplot. Two or more distributions could be graphed in the same plot for comparison.

QUANTILE-QUANTILE PLOT: Quantile-quantile plot (Q-Q Plot) is a plot that is usually used to compare if the data is normal.

WILCOXON RANK SUM TEST: Wilcoxon Rank Sum Test is a order test by ranking each set of data in ascending data. Then we compare the order of the data to test if there exists difference of distribution between two observation.

CHI-SQUARE TEST: χ^2 test is used to determine if there exists a difference between the expected frequencies and the observed frequencies. Assuming the null hypothesis is true, it evaluates the probability of the observations that are made in the data. The chi-square test is calculated as the below equation:

$$X^2 = \sum_{i=1}^{k}\frac{(x_i - m_i)^2}{m_i}$$

MULTIPLE REGRESSION: Multiple regression predicts the outcome of a response variable using multiple explanatory variables. It explores a linear relationship between the independent variables and dependent variable. For i=n observations, the formula for multiple regression is the following:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$$

where y_i is the dependent variable, x_i are the explanatory variables, β_1 is the interception point of y axis, β_p are the slope coefficients for each explanatory variable, and ϵ is the model's error term.

WELCH T-TEST: Welch T-test is a two sample test that is used to discuss the difference between the two correspond means, with the assumption that both samples are normally distributed.

Let $\bar{X_1}$, s_1^2 ,and N_1 be the mean, sample variance, and the sample size of the first sample.

$$\text{Welch's t} = \frac{\bar{X}1 - \bar{X}2}{\sqrt{\frac{S_1^2}{N1} + \frac{S_2^2}{N2}}}$$

KOLMOGOROV–SMIRNOV TEST: We use the Kolmogorov–Smirnov test to determine whether there exist some differences of the given two distributions. As a result, the KS test gives a p-value which indicates the probability of the given two distribution come from the same distribution. The test statistic of KS test for a given cdf is

$$D_n = \sup_x |F_n(x) - F_n|$$

CENTRAL LIMIT THEOREM: When independent random variables are added into any given probability distribution, the central limit theorem states the sum of that particular probability distribution would tends be a normal distribution. This theorem states the fact that a sum of many independent and identically distributed random variables can be approximated by a normal distribution.

BOOTSTRAP: Under the condition that the test has a random sample with replacement, bootstrap allows the data to be more precise using confidence intervals, prediction error, etc. It estimate the properties of an estimator by drawing information from the approximated distribution. A common use of bootstrapping is inference the population from sample data so that the sample data can be remodeled.

RANDOM FOREST: Random Forest is a flexible and popular machine learning algorithm for classification, regression and other tasks. at each tree split, a random sample of m features is drawn, and only those m features are considered for splitting. Typically $m = \sqrt{p}$ or $\log_2(m)$, where p is the number of features. It corrects for decision trees' over-fitting to their training set. Each tree has the same expectation.

SUPPORT-VECTOR MACHINE(SVM): One of the most widely used clustering algorithms in industrial applications, SVM is one supervised learning algorithm with associated learning algorithms defined by a separating hyper-plane. Given labeled training data, SVM algorithm builds an optimal hyper-plane that assigns new examples to categories. The kernels of SVM algorithms transform the problem

based on some linear algebra, which is called kernel trick. It has linear, polynomial, and exponential kernels.

ADA BOOSTING: adaptive boosting is a popular machine learning meta-algorithm for classification, regression and other tasks, which can incorporate multiple weak classifiers into a single strong classifier. Basically, Ada boosting can decide how much weight should be given to each classifier. Also, based on the results of the original classifier, we can select the training set for new classifier. The following introduction is given by Trevor Hastie from Stanford University.

1. Initialize the observation weights $w_i = \frac{1}{N}, i = 1, 2, ..., N$.
2. from $m = 1$ to M repeat the following steps
   a. Fit a classifier $C_m(x)$ to the training data using weight w_i
   b. Compute weighted error of newest tree
   $$\text{err}_m = \frac{\sum_{i=1}^{N} I(y_i \neq C_m(x_i))}{\sum_{i=1}^{N} w_i}$$
   c. Compute
   $$\alpha_m = \log[\frac{1 - err_m}{err_m}]$$
   d. Update weights

   for i = 1,....,N:$w_i \leftarrow w_i \times exp[\alpha_m \times I(y_i \neq C_m(x_i))]$and renormalize to w_i to sum to 1.
3. Output C(x)= sign $[\sum_{m=1}^{M} \alpha_m \times \text{C}_m(x)$

7 *Limitation*

Even though we have access to a large database, we still encounter various limitations during the project.

1. When investigating whether home field advantage exists in both regular seasons and tournaments, we have simply eliminated the games that are played in a neutral court, which could result in a loss of data.
2. The objective of this project is to provide a valid prediction for the tournament result; however, the data given does not include any data about the referees. It is clear that referees could have some impacts on the games results, but we are unable to clarify such impact.
3. To investigate hypothesis 4, we have employed two metrics: defensive rating and offensive rating. Since some of the teams in the

regular season are unqualified to enter the tournament, errors may exist in both offensive rating and defensive rating for the teams that enter tournament.

4. In hypothesis three, we have used bootstrap to randomly generate more than 11,000 data points out of 800 data points we have since we need to have same sample size for two samples. There could be biases in the bootstrap process.

8 Conclusion

In the project, we attempt to make a prediction on the 2019 NCAA Division 1 Women Basketball Tournament methods. In the beginning, we have conducted researches on the tournament format as well as some prevalent methods of predicting tournament results.

Before using advanced machines learning methods, investigation on factors that affect winning rates are conducted. In the beginning, we want to see if there exist home field advantage: home teams have a higher chance of winning than the away team in a game. To accomplish this objective, we have first eliminated games that are played in neutral courts and then conducted a one-sided hypothesis test and confidence intervals. Results from both regular seasons and tournament have shown that home teams have a higher chance of winning than away teams in any basketball game, the home-win rate is significantly higher in the tournament.

For the second hypothesis, we have investigated if the tournament games are more intense than matches in regular seasons. To measure competitiveness, we have measured the score difference between two teams in a match: a smaller difference in points implies a more competitive match. After using numerical representation and conducting hypothesis test, we have shown that there does not exist a statistically significant points difference between regular seasons and tournaments.

Then we want to analyze if a strong team perform better when facing a strong competitor. By drawing histograms and performing KS test and Welch's t-test, we conclude that a team will perform better when facing a strong competitor. In hypothesis 4, we want to see if there is a correlation between offensiveness and winning rate of a team. We have used an additional metric named offensive rating to run linear regression against the overall winning rate. Results have clearly shown that there is a positive correlation between a team's offensive ratings and its winning probability.

For the last hypothesis, we have investigated what factors will affect the overall chance of winning. By drawing distribution plots, we

conclude that offensive rating and net rating could affect the chance of winning while free throw rates and defensive ratings do not seem to affect the winning chance.

Built upon our knowledge, we use advanced machine learning to provide a prediction for the 2019 tournament result.

9 *Works Cited*

1. Caudill, Steven B., and Norman H. Godwin. "Heterogeneous skewness in binary choice models: Predicting outcomes in the men's NCAA basketball tournament." *Journal of Applied Statistics* 29.7 (2002): 991-1001

2. Hastie, Trevor, *Trees, Bagging, Random Forests and Boosting*, Lecture

3. "Road to the Championship." *Women's Basketball Road to the Championship*, NCAA, https://www.ncaa.com/womens-final-four/road-to-the-championship.

4. "Selection Criteria" *Women's Baksterball: Selection 101*, NCAA, http://www.ncaa.org/about/resources/media-center/womens-basketball-selections-101-selections

5. West, Brady T. "A simple and flexible rating method for predicting success in the NCAA basketball tournament: Updated results from 2007." *Journal of Quantitative Analysis in Sports* 4.2 (2008).