Final Project:

Forecasting Outcomes of NCAA Women's Basketball Games

Group Members

Gongjinghao Cheng	A92137907	Applied Mathematics
Zhefan Liu	A13809211	Math-Econ
Shiyi Hua	A13459035	Math-Econ
Wai Siu Lai	A12716401	Data Science
Sainan Chen	A14483750	Statistics

Contents

I. Introduction

Nowadays, it is popular to predict the result of sports competitions. In this project, we are going to predict the results of the 2019 NCAA Division I Women's Basketball Championship. Specifically speaking, we are going to explore different approaches that can predict the probabilities for possible matchups.

NCAA women's basketball tournament is an annual college basketball tournament held each March and can be separated into regional tournaments. There are many basketball games played between Division I women's teams for each season while culminated into the national championship between 64 teams.

Provided with a large amount of historical data, our investigations on this project consists of 5 main parts. At first, we calculate RPI and then examine a logistic regression using difference between RPI. We also explore feature reduction with decision tree and random forest. In addition, we divide the problem into two questions. One is to predict the winning probability of a particular team, given two teams playing against each other in a tournament. The other is to predict the expected number of wins a team will have in the tournament. We answer these two questions by going through the process of regression modeling, predicting and cross-validation. In the further investigation, we use a neural network to classify whether a team's point differential in its won games have an effect on its future probability to win games.

II. Data

In this project, we have a large amount of historical data about this NCAA basketball games and teams that can be mainly divided into four parts. These four parts cover many different aspects, so there are not noticeable data limitations. The first part is data of the basic information, containing 4 files.

Data Section 1 file: WTeams.csv

This file lists 366 teams and their related information:

- Team ID: A 4 digit ID number, ranging from 3000-3999 for NCAA women's team. A college's team ID does not change from one year to the next.
- Team Name: A compact spelling of team's college name that are 16 characters or fewer.

Data Section 1 file: WSeasons.csv

This file lists season-level properties that identify different seasons, such as:

- Season: The year in which the tournament was played.
- DayZero: The date corresponding to daynum=0 during that season
- RegionW, RegionX, RegionY, RegionZ: Whichever region's name comes first alphabetically, this region will be Region W. And whichever Region plays against Region W in the national semifinals will be Region X. For the other two regions, whichever region's name comes first alphabetically, that region will be Region Y, and the other will be Region Z

Data Section 1 file: WNCAATourneySeeds.csv

This file shows the seeds for all teams since 1997-98 season, containing variables:

- Season: The year in which the tournament was played
- Seed: A 3-character identifier of the seed, where the first character is either W, X, Y, or Z and the next two digits tell you the seed within the region
- TeamID: A 4-digit ID number.

Data Section 1 file: WRegularSeasonCompactResults.csv & WNCAATourneyCompactResults.csv

These two files show game-by-game results and tournament results of seasons since 1998 separately with variables such as:

- Season: The year in which the tournament was played

- DayNum: Ranges from 0 to 132 that shows what day the game was played on
- WTeamID: The ID number of the team that won the game
- WScore: The number of points scored by the winning team
- LTeamID: The ID number of the team that lost the game
- LScore: The number of points scored by the losing team
- NumOT: The number of overtime periods in the game
- WLoc: The location of the winning team: whether the winning team was the home team(H), a visiting team(A) or was played on a neutral court(N).

The second part presents game-by-game data at a team level for regular season, conference tournament and NCAA tournament games since 2019-10 season with two files.
Data Section 2 file: WRegularSeasonDetailedResults.csv & WNCAATourneyDetailedResult

- WFGM: field goals made (by the winning team)
- WFGA: field goals attempted (by the winning team)
- WFGM3: three pointers made (by the winning team)
- WFGA3: three pointers attempted (by the winning team)
- WFTM: free throws made (by the winning team)
- WFTA: free throws attempted (by the winning team)
- WOR: offensive rebounds (pulled by the winning team)
- WDR: defensive rebounds (pulled by the winning team)
- WAst: assists (by the winning team)
- WTO: turnovers committed (by the winning team)
- WStl: steals (accomplished by the winning team)
- WBlk: blocks (accomplished by the winning team)
- WPF: personal fouls committed (by the winning team)

The third part of data presents geographical data of regular seasons, conference tournament and NCAA tournament games since 2009-10 season with two files:
Data Section 3 file: WCities.csv
This file is a list of city location where the games played.

- CityID: A four-digit ID number identifying a city

- City: The text name of the city
- State: The state abbreviation of the state that the city is in

Data Section 3 file: WGameCities.csv

This file provides specific information that identifies all games with variables as follow:

- Season: the year in which the tournament was played
- DayNum: ranges from 0 to 132, and tells you what day the game was played on
- WTeamID: the id number of the team that won the game
- LTeamID: the id number of the team that lost the game
- CRType: this can be either Regular or NCAA
- CityID: a four-digit ID number uniquely identifying a city

The fourth part provides additional data, mainly alternative team name and bracket structure in two files.

Data Section 4 file: WTeamSpellings.csv

This file shows the alternative spellings of some team names.

- TeamNameSpelling: Spelling of the team name
- TeamID: a 4 digit id number

Data Section 4 file: WNCAATourneySlots

This file shows how teams are paired against each other, based on their seeds.

- Slot: Unique identifier of one of the tournament games
- StrongSeed: the expected stronger-seeded team that plays in this game
- WeakSeed: the expected weaker-seeded team that plays in this game

III. Background

There are a lot of approaches to predict outcome for not only basketball games but also other sports, such as football. First, based on what Miljković stated, the outcome prediction problem is formalized as a classification problem since the outcomes can be divided into two classes: when the host will win and when the visiting team will win. And he analyzed this problem mainly using Naive Bayes classification method and multivariate linear regression (Miljković et al. 2010)

He described each game with record of 141 attributes related to the outcome and team. Each team has two groups of attributes: the first one is about standard basketball statistics, such as Field goal made per game (FGM), Field goal attempted per game (FGA), 3-pointers made per game (3M), Free throws made per game (FTM), Points per game (P) and so on. The second one is about league standings' information, such as Total number of wins and losses, number of games won and lost at home, number of wins and losses in last 10 games and so on. Miljković used feature selection, normalization and other classification techniques, such as decision trees and k nearest neighbors to improve this system (Miljković et al. 2010).

Besides Naive Bayes classification and multivariate linear regression, other rating methods, such as the Ratings Percentage Index (RPI) and The OLRE Method, a method based on ordinal logistic regression modeling and expectation, can also be used.

Rating Percentage Index is a rating method that has been used in years of selecting and seeding teams and predicting outcomes. This method compare teams based on the winning percentages of them and their opponents. NCAA calculates the RPI for a given college basketball team i in the following way:

$$RPI_i = 0.25 * WP_i + 0.50 * OAWP_i + 0.25 * OOAWP_i$$

Where WP = Winning percentage, OAWP = Opponents' Average Winning Percentage and OOAWP = Opponents' opponents Average Winning Percentage.

However, this method is criticized for its arbitrary calculation method and other important factors being excluded (West et al. 2007).

The main method West showed is The OLRE Method. West regarded this method as a simple and flexible rating method that could come up with a ratings representing predictions of how many wins each team selected for tournament will get. This method investigates on a multivariate set of historical data collected on teams at the end of many regular seasons and examine the patterns of success. The predicted probability of team i winning j games can be written as:

$$\pi_{ij} = \frac{exp(\alpha_j + x_i\prime\beta)}{1 + exp(\alpha_j + x_i\prime\beta)} - \sum_{k=0}^{j-1} \pi_{ik}$$

Where α_j is the intercept for the j-th outcome, x_i is a vector of values for team i on the team-level predictor variables, β is a vector of coefficients associated with the predictor variables, and the last term represents the cumulative sum of predicted probabilities of winning k games (k = 0,... j-1). The predicted probability of getting six wins would be 1 minus the sum of the six predicted probabilities. It will come up with a 64*7 matric of predicted probabilities, considered as a contingency table. Then the expected value of WINS can be calculated for a team i as the follows:

$$E_i[WINS] = \sum_{j=0}^{6} j \times \hat{\pi}_{ij}$$ (West et al. 2007).

In addition to all these rating methods, there is an article by Boulier concluded that rankings are useful predictors by themselves, validated by using the probit function to test the hypothesis. “The higher-ranked men’s and women’s basketball teams beat lower-ranked opponents in 73.5% and 77.7% of the games.” (Boulier et al. 1999). There is also an article by Goller saying that the outcomes of games is decided by the scores of the two teams “that are usually either collapsed into a goal-difference or further aggregated to reflect whether the game ended as a win for the home or away team, or as a draw” (Goller et al. 2018).

IV. Investigation

1. RPI

The rating process index (RPI) is an index used to represent a team's overall performance according to the winning rates for all the teams. It is commonly used in basketball games and football games.

First, in Table 1.1, we calculate the RPI for all the 252 teams who played in Tourney between 1998 and 2018 by finding out their winning rate, their opponents' average winning rates and their opponents' opponents' average winning rate. Higher the RPI, better the team performs, and it also implies the team has higher probability to win in a match.

Team	RPI	Team	RPI	Team	RPI	Team	RPI	Team	RPI
3103	0.27852275	3181	0.57781271	3253	0.28831813	3332	0.50730246	3403	0.5194688
3104	0.49866605	3182	0.46703064	3254	0.32818107	3333	0.57230298	3404	0.28669309
3106	0.26633445	3184	0.25655555	3256	0.53512193	3335	0.2665145	3405	0.21471588
3107	0.32682296	3185	0.27639692	3257	0.54999736	3336	0.52621769	3407	0.28294806
3108	0.27086773	3187	0.26507255	3258	0.28516018	3337	0.24484439	3408	0.34628751
3110	0.22952526	3189	0.22632666	3261	0.5557125	3338	0.50915235	3409	0.41422225
3111	0.30245827	3190	0.23209959	3263	0.33253836	3340	0.26324304	3411	0.28516018
3112	0.46545427	3191	0.27060289	3264	0.2781403	3341	0.28526708	3412	0.59919799
3113	0.50534776	3193	0.27968175	3265	0.41139611	3343	0.30158621	3413	0.27961126
3114	0.38113668	3194	0.27690678	3266	0.39811357	3345	0.56061069	3415	0.26513885
3116	0.51896247	3195	0.37690149	3268	0.56389394	3346	0.46813928	3416	0.25394236
3119	0.27041644	3196	0.42482334	3269	0.21320892	3349	0.4065443	3417	0.50502236
3120	0.44717059	3197	0.23192949	3270	0.25556055	3350	0.26026092	3418	0.2047227
3122	0.2805129	3198	0.43016995	3272	0.17672414	3352	0.28096575	3420	0.32818107
3123	0.51625285	3199	0.5228952	3273	0.25482964	3353	0.54215017	3421	0.27265796
3124	0.58417247	3200	0.23806957	3274	0.35297205	3355	0.35131524	3422	0.24418629
3125	0.25789583	3201	0.2465575	3275	0.2771936	3357	0.23941088	3424	0.26633808
3129	0.26498383	3202	0.30245827	3276	0.40637884	3359	0.26943062	3425	0.44679261
3130	0.52286575	3203	0.46279169	3277	0.49474253	3360	0.26816115	3426	0.25750657
3131	0.32818107	3205	0.1636855	3278	0.51134269	3361	0.50289265	3427	0.27418204
3132	0.38229837	3207	0.50726787	3279	0.50297331	3362	0.27961126	3428	0.47714066
3137	0.27597637	3208	0.51555251	3280	0.55243135	3364	0.40381818	3431	0.39986485
3138	0.49513357	3209	0.28272752	3281	0.42234257	3365	0.25264738	3433	0.26507255
3140	0.42387945	3210	0.42081304	3283	0.43172411	3366	0.28910225	3434	0.25493408
3141	0.50655081	3211	0.49130064	3285	0.25460484	3369	0.27022059	3435	0.50169649
3142	0.25873419	3212	0.26152506	3286	0.25801014	3370	0.26608738	3436	0.34889783
3143	0.48070399	3214	0.26881735	3292	0.35758517	3371	0.24673201	3437	0.47478812
3144	0.28152668	3216	0.38352309	3293	0.28152668	3372	0.31342719	3438	0.44023941

Table 1.1

3145	0.28152668	3217	0.36104058	3294	0.21876266	3373	0.25015319	3439	0.46672791
3146	0.25755762	3218	0.253781	3298	0.25386358	3374	0.39077785	3441	0.2748428
3150	0.26677542	3221	0.25150796	3299	0.28283395	3376	0.58870514	3442	0.26026092
3151	0.28789268	3222	0.35249337	3301	0.47157862	3377	0.27961126	3443	0.34070928
3153	0.35557463	3224	0.24626296	3304	0.41545369	3378	0.40914004	3444	0.27961126
3155	0.46864647	3225	0.29572321	3307	0.3749787	3380	0.27205249	3449	0.50517414
3156	0.28807709	3226	0.23187404	3308	0.26967991	3382	0.54349034	3451	0.2552957
3159	0.30245827	3228	0.48280163	3311	0.2781403	3383	0.32818107	3452	0.43367578
3160	0.49633006	3229	0.27332131	3313	0.28152668	3384	0.25829701	3453	0.37143343
3161	0.44492722	3231	0.42054253	3314	0.54693949	3385	0.47164799	3454	0.26271158
3163	0.66608175	3233	0.27838203	3315	0.26324304	3386	0.46775149	3455	0.25239642
3164	0.29094231	3234	0.41463026	3318	0.26456617	3388	0.44585336	3457	0.28152668
3165	0.32818107	3235	0.47189768	3319	0.28516018	3389	0.22975517	3458	0.220374
3166	0.41653327	3238	0.27690678	3320	0.23077988	3390	0.57149445	3460	0.24531425
3169	0.27074316	3239	0.28910225	3321	0.26320344	3391	0.27562422	3461	0.24987508
3171	0.28745665	3241	0.31309332	3322	0.2898812	3393	0.46905363	3462	0.46528467
3173	0.44240971	3242	0.53530847	3323	0.59746223	3395	0.42275819	3464	0.36678916
3174	0.45025318	3243	0.46814319	3324	0.24031004	3396	0.38312772		
3175	0.24804159	3245	0.23013392	3325	0.25655555	3397	0.61509188		
3176	0.23307629	3246	0.52059069	3326	0.47190912	3398	0.24531425		
3177	0.46692182	3249	0.21508922	3328	0.53069598	3399	0.2423157		
3179	0.36375905	3250	0.22511507	3329	0.46731597	3400	0.48581276		
3180	0.23016239	3251	0.32218932	3330	0.47272194	3401	0.53913523		
3181	0.57781271	3252	0.24804159	3331	0.27001147	3402	0.27690678		

Table 1.1 Continued

2. Logistic Regression Using Difference between RPI

For every team in every match, we calculate the difference between the team's RPI and its opponent's RPI. Then we use the difference between RPI to do logistic regression. We assume two regression models:

The first model: Y = e^(a+bX)/(1+e^(a+bX))=1/(1+e^-(a+bX)), where Y is the dependent variable game result (0 means losing and 1 means winning), X is the independent variable difference between RPI of this team and its opponent (team's RPI - opponent's RPI), and a and b are constants that we want to find.

First, we do a scatter plot to see if there's any possible relationships between game result and difference between RPI.

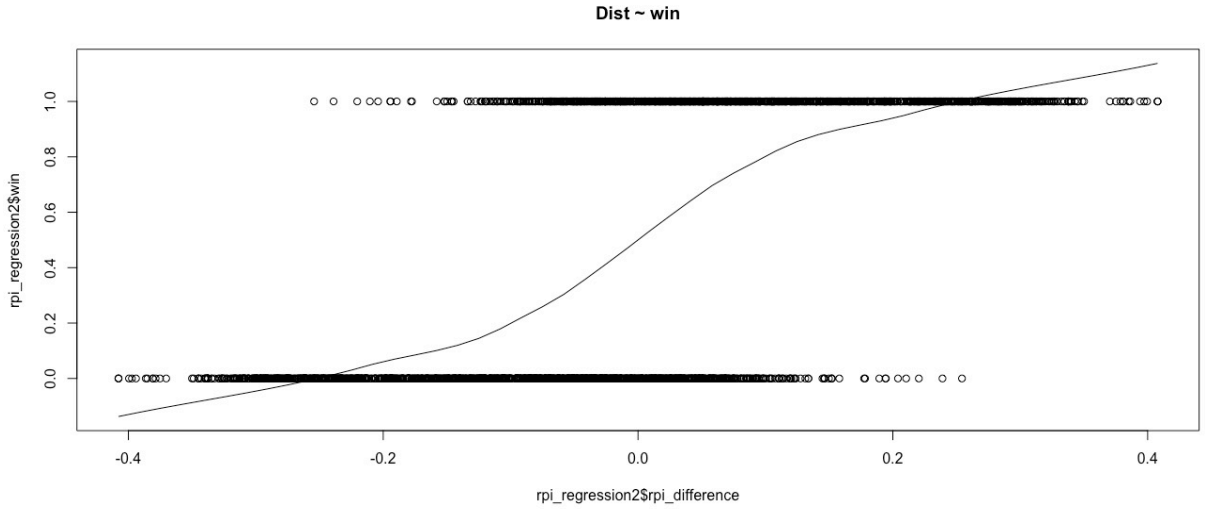


Figure 2.1

Figure 2.1 shows the whole RPI difference distribution mean when the game result equals to 1 are 0.2 larger than the RPI difference distribution mean when the game result equals to 0. It implies that when a team wins in a game, it has relatively higher RPI difference. The correlation between game result and RPI difference is 0.637, so we guess there is a relationship between the winning probability and RPI difference, and we will do a logistic regression in Y = e^(a+bX)/(1+e^(a+bX)).

The result shows a, which is intercept, equals to -6.326*10^-17. It means when the team's PRI and its opponent's RPI are the same, the probability for each team to win is e^a/(1+e^a). b, which is the gradient, equals to 13.47. It means when the difference between the team's RPI and the opponent's RPI increases by 1, the probability for the team to win also increases by some value related to 13.47. In conclusion, our logistic regression model is

Y = 1/(1 + e^(-13.47*(team's RPI - opponent's RPI) - 6.326*10^-17)).

The second model: Y = 1/(1+e^-(aX^2 + bX + c)), where a, b and c are the constants that we want to find.

We then do a scatter plot to see if there's any possible relationships between game result and square of difference between RPI and difference between RPI.

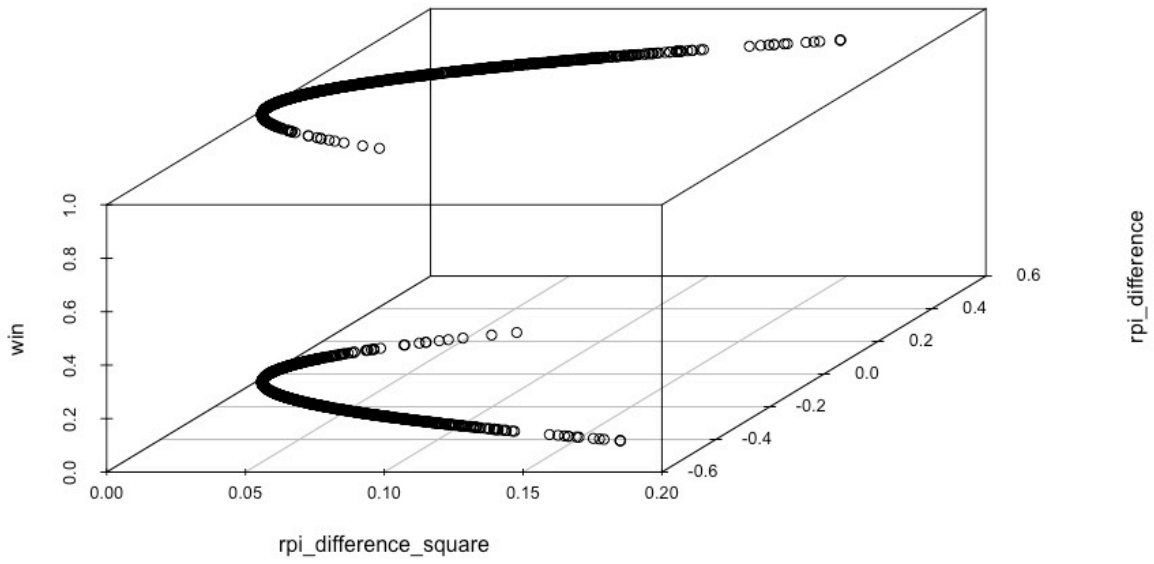


Figure 2.2

Figure 2.2 shows the whole RPI difference distribution mean when the game result equals to 1 are 0.2 larger than the RPI difference distribution mean when the game result equals to 0. It implies that when a team wins in a game, it has relatively higher RPI difference. However, it's not clear to see the relationship between square of RPI difference and game result. The correlation between game result and RPI difference is 0.637, and the correlation between game result and square of RPI difference is 0. We still want to check if there is a relationship between the winning probability, the square of RPI difference and RPI difference, and we will do a logistic regression in Y = 1/(1+e^-(aX^2 + bX + c)).

The result shows c, which is intercept, equals to 3.274*10^-17. It means when the team's PRI and its opponent's RPI are the same, the probability for each team to win is 0.5. b equals to 13.47. It means when the difference between the team's RPI

and the opponent's RPI increases by 1, the probability for the team to win also increases by some value related to 13.47. c equals to -2.709 * 10^(-15). It means every time when square of RPI difference increases by 1, the probability for the team to win also increases by some value related to -2.709 * 10^(-15). In conclusion, our logistic regression model is;

$$Y' = 1/(1+e^{-(-2.709 * 10^{(-15)}*X^2 + 13.47*X + 3.274*10^{-17})})$$

Comparing the first and the second logistic regression model, we can see there's no crucial difference between them, because the coefficient for X^2 is so small that it can be ignored. For calculation convenience, we take the first model as our linear regression model.

To do the prediction, we take 5 teams with highest RPI, which are 3163, 3397, 3412, 3323 and 3376, as an example. We choose one team, and compute its probability to win when it competes with all the other teams.

The result is as following:

Team	Opponent ID	Winning probability	Team	Opponent ID	Winning probability
3163	3397	0.6652623	3412	3323	0.505845
3163	3412	0.7111395	3412	3376	0.5352761
3163	3323	0.7159187	3323	3163	0.2840813
3163	3376	0.7392866	3323	3397	0.4409094
3397	3163	0.3347377	3323	3412	0.494155
3397	3412	0.5533192	3323	3376	0.5294554
3397	3323	0.5590906	3376	3163	0.2607134
3397	3376	0.5879338	3376	3397	0.4120662
3412	3163	0.2888605	3376	3412	0.4647239
3412	3397	0.4466808	3376	3323	0.4705446

3. Feature Reduction

Dimension reduction is a process of reducing the number of random variables and keep only the principle variables. The reason that we want to do this is that most of the time there are many random variables (called features) which may influence the final result that we want to predict. Many of those random variables are correlated therefore redundant. Higher the number of variables, harder to visualize and work on the training set. Large number of random variables might cause overfitting, which makes our prediction imprecise as well.

<u>Correlation between Features</u>

In our data, there are 13 features in total: FGM (field goals made), FGA (field goals attempted), FGM3 (three pointers made), FGA3 (three pointers attempted), FTM (free throws made), FTA (free throws attempted), OR (offensive rebounds), DR (defensive rebounds), Ast (assists), TO (turnovers committed), Stl (steals), Blk (blocks) and PF (personal fouls committed).

We want to find the correlation relationship between every two features. First, we do a scatter plot between features.

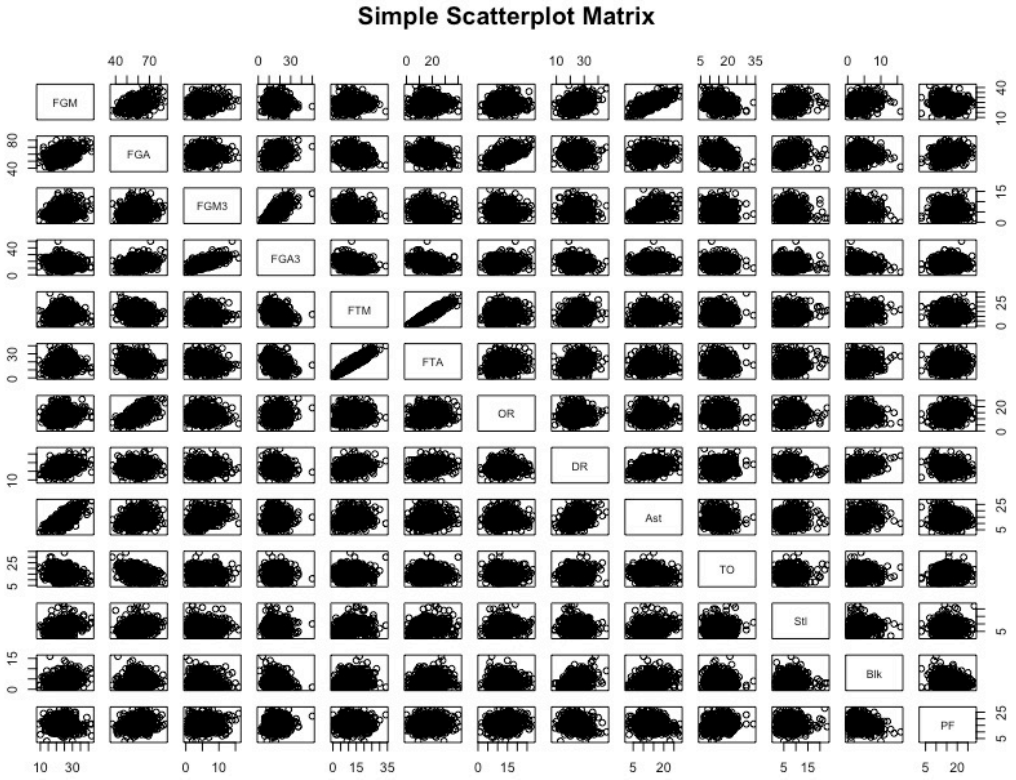


Figure 3.1

From Figure 3.1, we can see most of the plots have points clustered together without obvious correlation relationship. Only the plot of the relationship between FTM and FTA has points in a straight line, implying that FTM and FTA are in a positive relationship.

To get more information about the relationship between features, we make a correlation table.

Note: A positive correlation number means two variables are in a positive relationship, and a negative correlation number means two variables are in a negative relationship.

FGM	FGA	FGM3	FGA3	FTM	FTA	OR	DR	Ast	TO	Stl	Blk	PF
1	0.41	0.215	-0.127	0.103	0.087	0.118	0.332	0.708	-0.209	0.248	0.254	-0.11
0.41	1	0.094	0.276	-0.185	-0.171	0.627	0.025	0.147	-0.376	0.189	0.025	0.157
0.215	0.094	1	0.706	-0.124	-0.138	-0.059	0.016	0.335	-0.024	-0.058	-0.024	0.025
-0.127	0.276	0.706	1	-0.293	-0.294	0.105	-0.124	0.014	-0.076	-0.099	-0.183	0.135
0.103	-0.185	-0.124	-0.293	1	0.934	0.096	0.286	0.114	0.014	0.145	0.157	0.022
0.087	-0.171	-0.138	-0.294	0.934	1	0.164	0.282	0.077	-0.001	0.174	0.153	0.033
0.118	0.627	-0.059	0.105	0.096	0.164	1	0.041	-0.002	-0.053	0.109	0.002	0.118
0.332	0.025	0.016	-0.124	0.286	0.282	0.041	1	0.338	0.029	-0.138	0.375	-0.214
0.708	0.147	0.335	0.014	0.114	0.077	-0.002	0.338	1	-0.077	0.155	0.218	-0.166
-0.209	-0.376	-0.024	-0.076	0.014	-0.001	-0.053	0.029	-0.077	1	0.088	-0.043	0.192
0.248	0.189	-0.058	-0.099	0.145	0.174	0.109	-0.138	0.155	0.088	1	0.046	0.03
0.254	0.025	-0.024	-0.183	0.157	0.153	0.002	0.375	0.218	-0.043	0.046	1	-0.152
-0.11	0.157	0.025	0.135	0.022	0.033	0.118	-0.214	-0.166	0.192	0.03	-0.152	1

Figure 3.2

From Figure 3.2, FGM3 and FGA3, FTM and FTA, Ast and FGM have comparably higher correlation (correlation larger than 0.5).

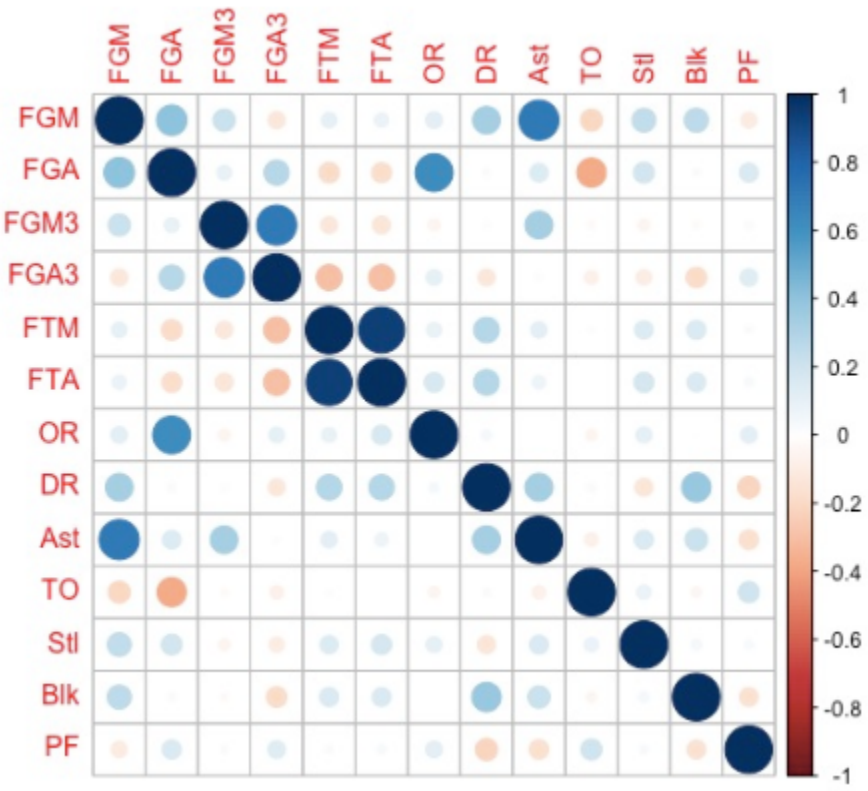


Figure 3.3

In Figure 3.3, the correlation is represented by the size and color density of every circle. A red circle means a negative correlation, and blue circle means a positive correlation. The deeper the color of the circle and larger the circle, the two features are more correlated. we can see clearly that correlations of FGM3 and FGA3, FTM and FTA, Ast and FGM are represented by larger and darker blue dots, and it implys these three pairs of features are highly correlated. Therefore, we can reduce the corresponding two variables into one of them.

Decision Tree

Decision tree is a good way to explore the relationship between variables and output by giving probabilities to achieve the goal for different variables values.

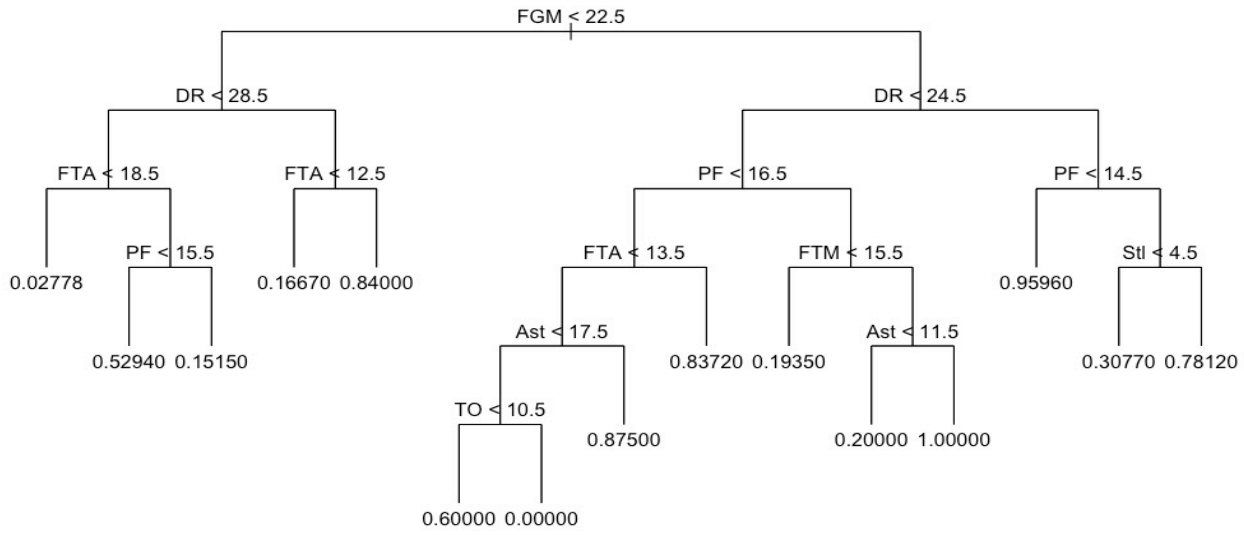


Figure 3.4

Figure 3.4 is a regression tree which shows other important factors that might be related to matchup result. This tree is based on regression over factors including FGM, FGA, FGM3, FGA3, FTM, FTA, OR, DR, Ast, TO, Stl, Blk and PF.

Note that FGA, FGM3, FGA3, OR and Blk do not appear in the tree, which means that they are not influential factors. Factors, such as FGM, on each node that split into two branches and leafs are considered as important.

The values at the end of each branch represents the probability for a team to win a matchup. For instance, teams who has FGM>22.5, DR>24.5, PF<14.5 has a response of 0.960, which means that all of those teams satisfying those categories have a high probability to win a match up. In this manner, we may anticipate matchups' results.

When the team is playing at the home city:

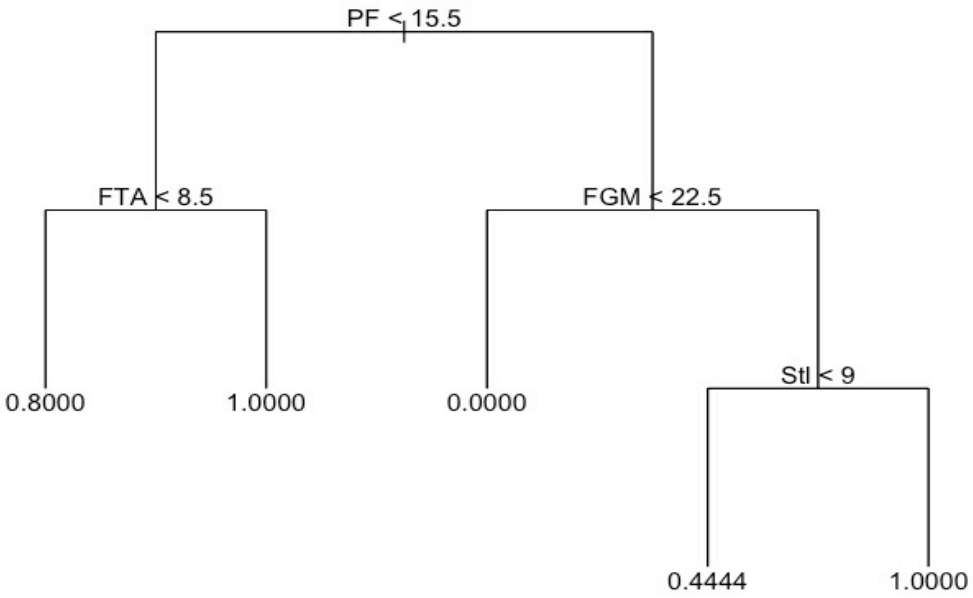


Figure 3.5

In Figure 3.5, when PF<15.5 and FTA>8.5, the team has 100% probability to win. Also, if the team has PF>15.5, FGM>22.5 and Stl<9, it has 100% probability to win. In contrast, if the team get PF>15.5 and FGM<22.5, it has nearly 0% probability for winning.

When the team is playing at a visiting city:

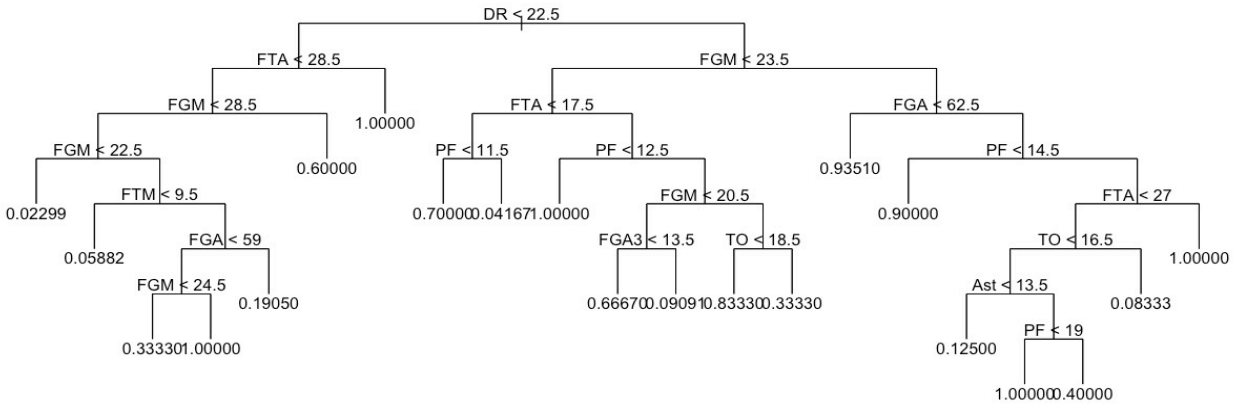


Figure 3.6

In Figure 3.6, when the team is playing at a visiting city, more variables are related to the final result. For example, when DR<22.5, FTA<28.5, FGM>24.5, FTM>9.5 and FGA<59, the probability for the team to win is 100%.

Random Forest

Random forest is a regression method which takes multiple trees at mean time and finally gives the mean of outcomes from all the trees.

When we take 500 as the number of trees and 4 as the number of variables tried at each split, the mean of squared residual/error is 0.1151.

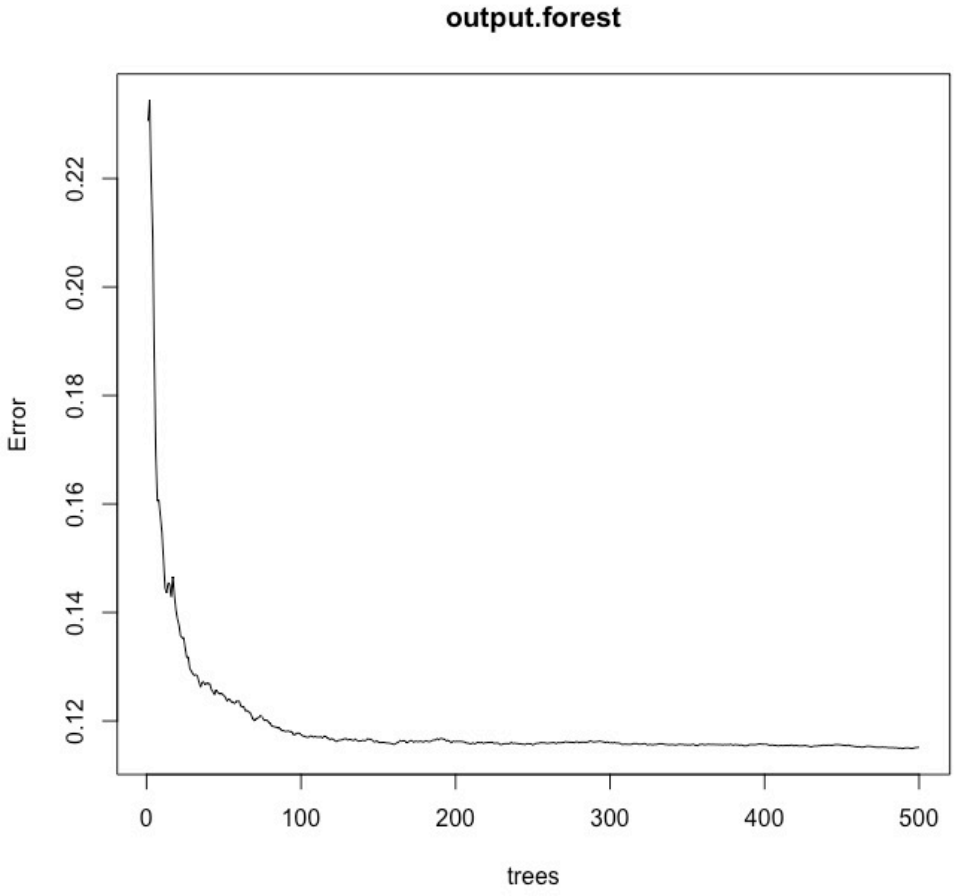


Figure 3.7

Figure 3.7 of Error vs. Number of trees shows the error decreases tremendously to 0.12 when the number of trees increases from 0 to 100, and the rate decreases as the number of trees keeps on increasing. Therefore, we can pick 500 as the number of trees, and the error at this time will be 0.11.

	number.of.predictor	out.of.bag.error	test.error
1	1	0.127	0.112
2	2	0.118	0.102
3	3	0.115	0.098
4	4	0.114	0.098
5	5	0.114	0.097
6	6	0.114	0.097
7	7	0.116	0.098
8	8	0.118	0.098
9	9	0.117	0.097
10	10	0.119	0.097
11	11	0.119	0.098
12	12	0.119	0.097
13	13	0.118	0.098

Table 3.1

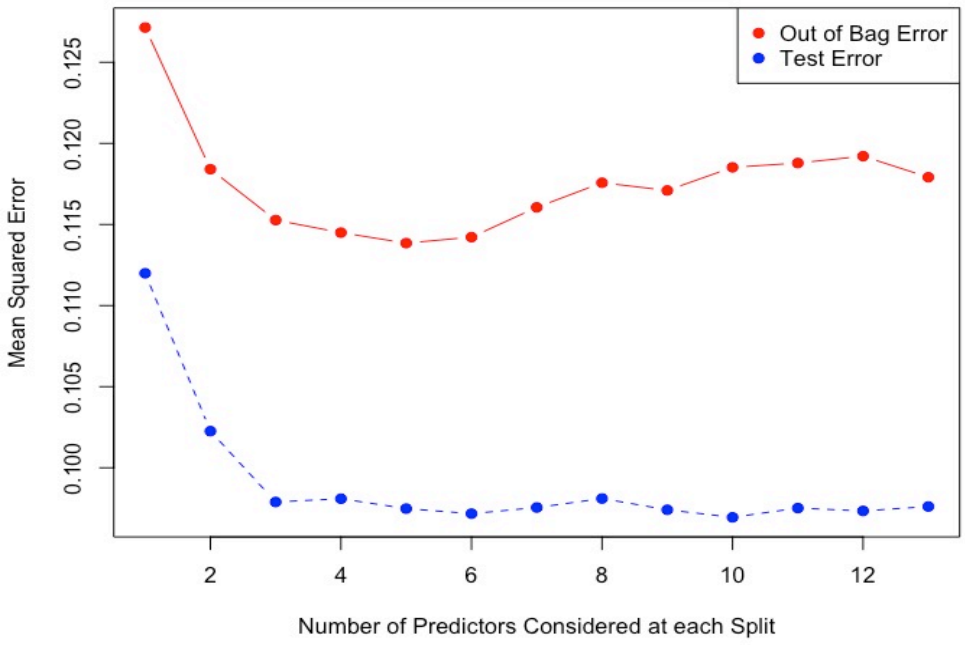


Figure 3.8

Table 3.1 is about out of bag error and test error for different number of predictors.

Figure 3.8 is about out of bag error and test error for different number of predictors.

When number of predictors is higher than 2, the test error doesn't fluctuate greatly when number decreases, and the error keeps under 0.1. The out of bag error keeps decreasing when the number of predictors increases from 1 to 5. However, the out of bag error starts to grow when number of predictors increases from 5. From the graph, we can see the out of bag error is under 0.115 only when the number of predictors is between 3 and 6. Therefore, we can reduce the 13 features to around 6 when the test error and bag error are relatively low.

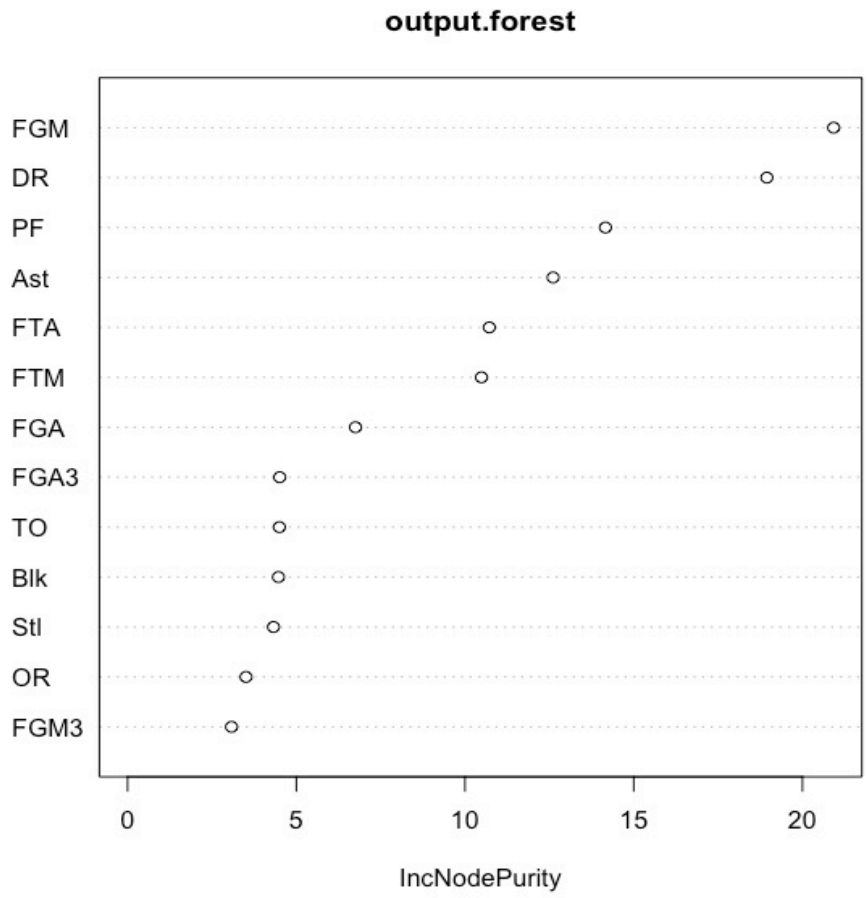


Figure 3.9

Figure 3.9 shows how pure the nodes (features) are at the end of the tree. It also describes the importance of every feature in deciding the game result. Higher the purity of feature, the more important the feature is. From the diagram, FGM, DRm PR Ast, FTA, FTM, FGA and FGA3 are the eight most important features with highest node purity.

4. Given two teams playing against each other in a tournament, denoted as team1 and team2, predict the winning probability of team1.

Modeling

General procedure: In this section, we will try to select influential predictor variables according to the seasonal data and construct a logistic regression with dependent variable (win or loss).

We first create the pool of potential predictor variables as in the detailed data set. After denoting two teams of each match as team1 and team2 respectively, we define the difference between their parameters, including fgm, fga, fgm3, fga3, ftm, fta, OR, DR, ast, blk, by those of team1 minus the counterpart of team2.

To determine the significance of a variable, we first create a tree diagram over all of those variables.

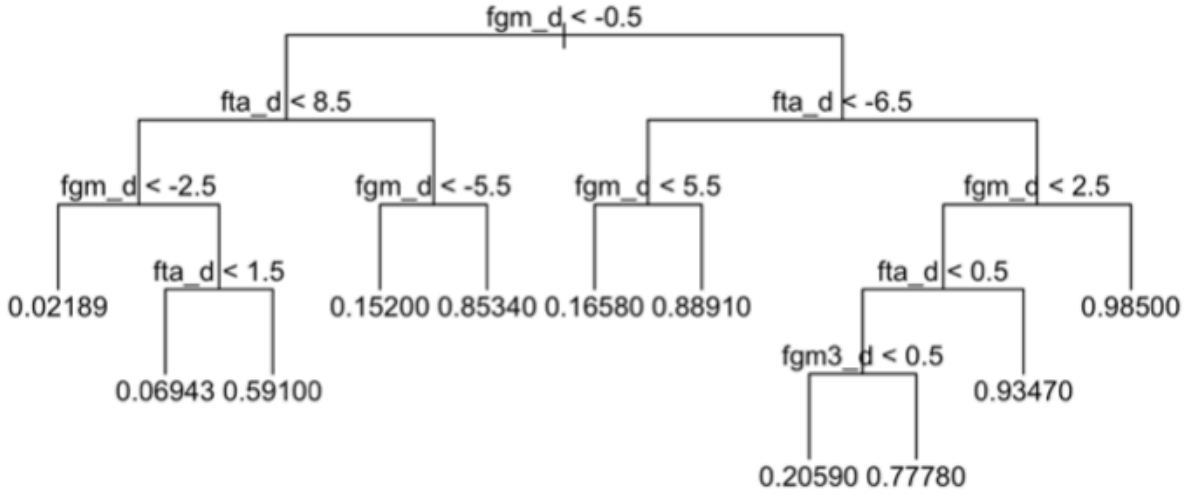


Figure 4.1

Figure 4.1 illustrates that fgm, fta and fgm3 are the potentially proper predictors for our logistic model since they are the only variables shown in our tree's node.

Logistic Model 1

	estimate	Std. Error	z value	p value
intercept	0.02868	0.03419	0.839	0.402
fgm difference	1.59873	0.02849	56.109	< 2e-16
fta difference	0.65014	0.01134	57.342	< 2e-16
fgm3 difference	0.83632	0.01740	48.074	< 2e-16

Residual

min	1Q	median	3Q	max
-3.9805	-0.0048	0.0000	0.0130	3.9661

Table 4.1

Table 4.1 shows the logistic regression model we obtain according to above three variables. That is, denote X = exp (0.0287 + 1.6*fgm_dif + 0.65*fta_diff + 0.836*fgm3_diff). The probability of team1 as winner = X / (1+X). Since we will not know the data of a particular match, we use the average data of each team over the last year to compute those differences. On the other hand, note that the p-value for intercept is 0.4 which is greater than 5%, this indicates the possible error in estimate of the intercept.

To examine the residual, note that the residual follows a chi-squared distribution. The residual of over model is 5788.4 with degree of freedom 45035. The p-value is 1, which is large enough to claim that there is no sign lack of fit.

To check the properness of the model, we decided to compare the predicted winner with real winner in tournament of 2014 and 2018. We first calculate the probability of team1 as winner, and then predict the winner according to the probability (i.e., team1 wins if probability > 0.5, vice versa).

For 2014 tournament, out of 63 matches, we correctly predict results of 45 matches; the rate of correct prediction is 71%.
For 2018 tournament, we successfully predict results of 41 matches; the rate of correct prediction is 65%.

This is a model better than random guess with winning probability of 0.5 for each team.

Due to the large p-value of the intercept, we continue looking for other predictor variables. Indeed, our literature review indicates that fgm, ast, and blk are better predictor variables for the outcomes, therefore we decide to create another logistic model over them.

Logistic Model 2

	estimate	Std. Error	z value	p value
intercept	0.214740	0.014488	14.82	< 2e-16
fgm difference	0.361815	0.004540	79.70	< 2e-16
ast difference	0.125228	0.003423	36.58	< 2e-16
blk difference	0.058706	0.004619	12.71	< 2e-16

Residual

min	1Q	median	3Q	max
-3.05239	-0.42455	0.08841	0.49318	2.91003

Table 4.2

With Table 4.2, note that all of the marginal p-value for those parameters are now approximately 0, which indicates the fit of the model. On the other hand, the residuals are more centralized at origin comparing to residuals of previous model.

We obtain a new model with X = exp (0.2147 + 0.361*fgm_dif + 0.125*ast_diff + 0.0587*blk_diff), and probability of team1 as winner = X / (1+X).

Again, to examine the residual, note that the residual follows a chi-squared distribution. The residual of over model is 29883 with degree of freedom 45035. The p-value is 1, which is large enough to claim that there is no sign lack of fit.

Similarly, we decided to compare the predicted winner with real winner in tournament of 2014 and 2018. We first calculate the probability of team1 as winner, and then predict the winner according to the probability (i.e., team1 wins if probability > 0.5, vice versa).

For 2014 tournament, out of 63 matches, we correctly predict results of 44 matches; the rate of correct prediction is 71%.
For 2018 tournament, we successfully predict results of 53 matches; the rate of correct prediction is 84%.

Obviously, this model is improved compare to the previous model.

Next, we examine the choice of period we average on when we compete those data. For all of the prediction above we obtain average over the whole regular season, now we decide to average over the last 30 days' matches.

	whole regular season	last 30 days
number of correct predictions for 2014	53	42
correct prediction rate for 2014	84%	66.7%
number of correct predictions for 2014	44	40
correct prediction rate for 2014	71%	63.5%

Table 4.3

Accordingly, the data averaging over the whole regular season is more correct than over merely over last 30 days before tournament. We keep our original procedure of estimate over the whole regular season.

Predicting

We are going to predict the results in tournament 2019. We first decide the 64 participants by selecting the teams with highest 64 RPIs. Also, since the regular season of 2019 has not finished, we compute our difference based on data from 2018.

With descending order of RPI, we have participants: 3163 3397 3412 3323 3376 3124 3181 3333 3390 3268 3345 3261 3280 3257 3314 3382 3353 3401 3242 3256 3328 3336 3199 3130 3246 3403 3116 3123 3208 3278 3338 3332 3207 3141 3113 3449 3417 3279 3361 3435 3104 3160 3138 3277 3211 3400 3228 3143 3428 3437 3330 3326 3235 3385 3301 3393 3155 3243 3346 3386 3329 3182 3177 3439.

Due to limit of our content, we only predict the results of matches between first 5 teams, the rest will bear same predicting procedure. The entries are the probability of team1 as winner.

team1\team 2	3163	3397	3412	3323	3376
3163	NA	0.967	0.982	0.801	0.967
3397	0.033	NA	0.704	0.144	0.550
3412	0.028	0.296	NA	0.080	0.389
3323	0.299	0.866	0.82	NA	0.900
3376	0.033	0.450	0.611	0.1	NA

Table 4.4

Cross Validation

We perform same procedure on data that omit all information of 2015, then predict the results of 2015 tournament.

Logistic Model 3

	estimate	Std. Error	z value	p value
intercept	0.206384	0.015381	13.42	< 2e-16
fgm difference	0.360194	0.004803	74.99	< 2e-16
ast difference	0.126826	0.003646	34.78	< 2e-16
blk difference	0.058322	0.004899	11.91	< 2e-16

Residual

min	1Q	median	3Q	max
-3.04872	-0.42363	0.08661	0.49085	2.91186

Table 4.5

According to Table 4.5, the new logistic model has parameters close to the previous model, so do their residuals. This implies the robustness of our regression model. We obtain new regression functions with X = exp (0.2063 + 0.360*fgm_dif + 0.127*ast_diff + 0.0583*blk_diff), and probability of team1 as winner = X / (1+X).

We use this model to predict the results of 2015 tournament. Out of 63 matches, we correctly predict results of 48 matches; the rate of correct prediction is 76%.

5. Given a team in tournament, predict the expected wins they will have in the tournament (possible numbers of wins are 0, 1, 2, 3, 4, 5, 6).

Modeling

General procedure: In this section, we will try to imitate the OLRE (ordinary logistic regression and expectation) method we found in literature review, with some minor modification and omission, since we did not find exact computation process for Jeff Sagarin's rating for teams.

We will perform an ordered logistic regression, with predictor variables of the team's historical winning percentage, averaging point difference and number of its victories against top 20 teams. The regression is based on the data of the 64 participants of tournament in each year.

We first summaries the distribution of our predictor variables.

	winning percentage	averaging point difference	number of victories against top 20 teams
Min	0.2691	11.46	0.0
1st Q	0.5453	13.94	0.0
Median	0.6159	15.32	1.0
Mean	0.6243	15.79	1.5
3rd Q	0.6945	16.88	2.0
Max	0.9421	31.21	12.0

Table 5.1

The Table 5.1 displays a potential outlier in averaging point difference and number of victories against top 20 teams.

coefficient	value	Std. error	t-value	p-value
winning percentage	2.0264	1.07058	1.893	0.06
averaging point difference	0.1935	0.04849	3.991	0
number of victories against top 20 teams	0.5557	0.05422	10.250	0

Table 5.2

According to Table 5.2, we create our ordinal logistic model with Y* being the expected winning number of a given team, X be the vector of the teams' historical winning percentage, averaging point difference and number of its victories against top 20 teams, Beta being the coefficient vector of 2.03, 0.19, 0.5557. We have our predicted Y* = t(X)*Beta.

winning percentage	averaging point difference	number of victories against top 20 teams
7.587039	1.213483	1.743165

Table 5.3

Table 5.3 shows how the unit changes in predictor variables change the prediction. For example, 7.58 for winning percentage means that keeping all other variables constant, when winning percentage increases one unit, it is 7.58 times more likely to be in a higher category. Similar for the other two predictor variables. The coefficients are significant.

To check properness of our model, we apply it to our model to all seeds in tournament from 2010 to 2018. We first compute their probabilities of winning 0,

1, 3, 4, 5, 6 matches respectively, and then compute their expected winning number, round to nearest integer.

The way we compute the probability of the team with each possible winning number is following:

$$\pi_{ij} = \frac{exp(\alpha_j + x_i\prime\beta)}{1 + exp(\alpha_j + x_i\prime\beta)} - \sum_{k=0}^{j-1} \pi_{ik}$$

For each team i, the probability they win j matches are computed as above, with alpha j as the intercept and Xi, Beta defined the same as above.

We obtain that out of 576 teams, we correctly predict the number of winning matches in tournament of 311 of them. Hence, we have a correct prediction rate of 54%.

Indeed, when we compute the average error between the real winning number and predicted winning number, we find that the error is on average 0.6, which means that our prediction is very also to the real number. This is a strong indication of the fit of our model.

Predicting

Again, we select our previous 5 teams for prediction; the prediction procedure for rest will be similar.

Team	expected number of wins
3163	6
3397	3
3412	0
3323	4
3376	3

Table 5.4

Since the average error is about 0.6, it would be fairly safe to extend the confidence interval of expected number of wins by +1 and -1.

Cross Validation

We will omit the data of 2015 and repeat the procedure to create the regression model.

coefficient	value	Std. error	t-value	p-value
winning percentage	2.1426	1.18577	1.807	0.04
averaging point difference	0.1594	0.05123	3.111	0
number of victories against top 20 teams	0.5680	0.05777	9.833	0

Table 5.5

According to Table 5.5, we create our new ordinal logistic model with Y* being the expected winning number of a given team, X be the vector of the teams' historical winning percentage, averaging point difference and number of its victories against top 20 teams, Beta being the coefficient vector of 2.14, 0.16, 0.568. We have our predicted Y* = t(X)*Beta. Note that the coefficients are close to the original model.

With this model, we have correctly predicted the number of wins for 40 teams out of 64 teams. The correct prediction rate is 0.625. The expected error is 0.48, which indicates the fit of the model.

V. Summary

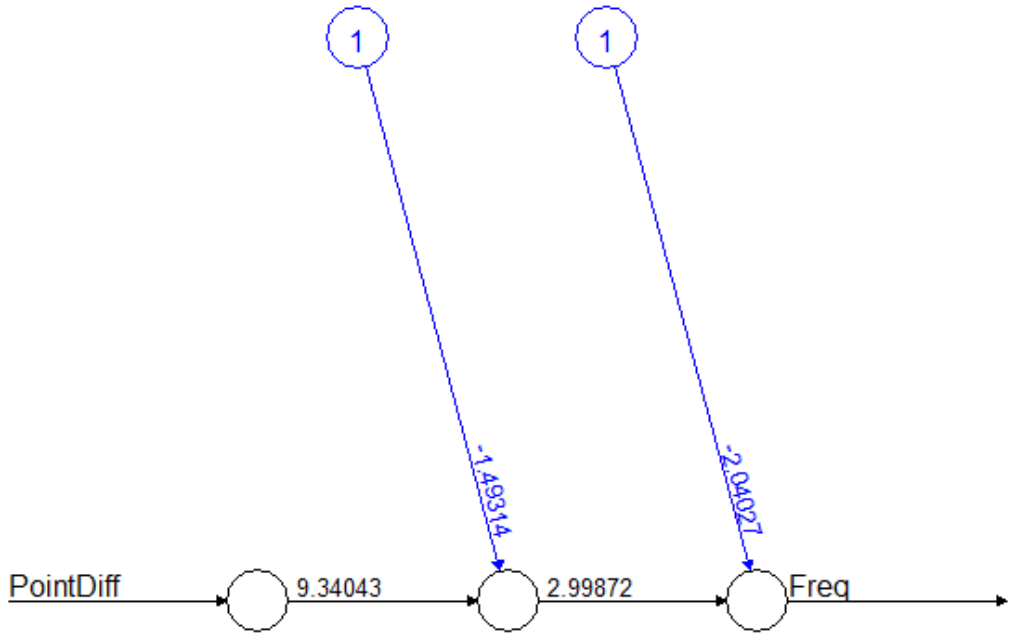
In this project, because some variables are correlated and redundant, we can reduce the features to six most uncorrelated and crucial features. According to the information given, we can see FGM, DRm PR Ast, FTA, FTM, FGA and FGA3 are the eight most important features. However, FTM and FTA are highly correlated, so we only use FTA for regression. Similarly, Ast and FGM are highly correlated, and we only use FGM for regression. Finally, we pick the following six features for regression: FGM, DR, PF, FTA, FGA and TO.

To predict the winner of matches in the coming tournament, we have constructed a logistic regression model by altering predictor variables and measures of parameters. With the data from the newest regular season, we are likely to predict the winner of matches with a probability over 70%.

To predict the expected number of wins with a given team in the tournament, we have created an ordinal logistic regression model with predictor variables specified by OLRE method. We are able to predict the exact number of wins of a given team during a tournament. Indeed, if the predicted number does not match the real number, they are likely to be adjacent number, which implies the accuracy of our model and predictions.

VI. Further Investigation

In this section, the method of a neural network is employed. The goal is to predict a team's winning percentage from a team's point differential across all seasons. The hypothesis is that the higher the winning team's point differential, the higher the probability of a team winning a game.



Error: 0.737491 Steps: 972

Looking at this neural network model, there is one input call PointDiff - which corresponds to the average point difference of each winning team for all seasons. Corresponding to the input variable is a node. This node is known as the input

layer, where the input data is provided. That is, this layer takes input based on existing data. Toward the end there is the output layer, where 'Freq' is the output variable. This variable measures the probability of each particular team winning a game across all seasons. In between the input and output layer, there is a layer called the hidden layer. This layer uses a technique known as back-propagation to optimize the weights of the input variables to enhance the predictive power of the model. The value on the blue line in the hidden layer is similar to a constant that appears in regression equations.

Using the function compute in R, one can interpret that team 3102 was predicted with probability 0.3275129 of winning a game. However, if one looked at the training data, team 3102 had merely the probability 0.13005 of winning a game. This phenomenon can be seen as misclassification. Doing the calculation -1.49314 + (9.34043*0.1342376) = -0.2393031, which is the value for input node two. Performing the sigmoid function common in neural networks, we use the previous value and calculate 1/(1+exp(-0.2393031)) = 0.440458. And this value 0.44 can become the output value for node two in the diagram above. A similar approach is employed to find the input and output of node three. Doing so, the input value for node 3 is -2.04027 + (2.99872 * output2) = -0.7194595. Once again, the sigmoid function is employed,1/(1+exp(-0.7194595)) = 0.327512. Note that this is the same value from the neural net prediction result.

Using the confusion matrix, we seek to calculate the misclassification error of this neural network model. The result is that 126 teams were correctly classified to win a game, and 130 teams were correctly classified to lose a game. The misclassification error for this particular neural work is calculated to be 0.5078125. Similarly from the testing data, 128 teams were correctly predicted to win and 128 teams were predicted to lose. Thus, performing similar calculations on the testing data gives a misclassification error of 0.50. In conclusion, this neural network performs very consistently with training and testing data.

VII. Theory

A. The Ratings Percentage Index (RPI)

1. A particular rating method that continues to receive a great deal of weight in selecting and seeding teams.
2. This rating method is used to compare teams based on their winning percentages and the winning percentages of their opponents.
3. FORMULA: The RPI for a given college basketball team i is calculated as $RPI_i = 0.25 \times WP_i + 0.50 \times OAWP_i + 0.25 \times OOAWP_i$, where WP is Winning Percentage, $OAWP$ is Opponents' Average Winning Percentage, and $OOAWP$ is Opponents' Opponents Average Winning Percentage.

- Note: When calculating the winning percentage, a home win now counts as 0.6 win, while a road win counts as 1.4 wins; inversely, a home loss equals 1.4 losses, while a road loss counts as 0.6 loss.

4. The RPI formula also has many flaws. Due to the heavy weighting of opponents winning percentage, beating a team with a bad RPI may actually hurt your RPI. In addition, losing to a good RPI team can help your RPI.

B. Classification and Regression Tree (CART)

1. An algorithm that can be used for classification when the response is categorical or regression when the response is continuous.
2. Given some variables $x_1, x_2, \cdots, x_n$, we want to predict the value of y if y is continuous random variable or to classify y if y is categorical.
3. The tree is formed by nodes and leaves (i.e., terminal nodes). Each internal node is split in two children on the basis of a splitting rule.

C. Random Forests

1. Refinement of bagged trees.
2. 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 $log_2 p$, where p is the number of features.

3. For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored, which is called the "out-of-bag" error rate.
4. Random forests tries to improve on bagging by "de-correlating" the trees, where each tree has the same expectation.

D. Mean Squared Error (MSE)

1. A measure of the quality of an estimator.
2. MSE of an estimator measures the average of the squares of the errors, which is the average squared difference between the estimated values and what is estimated.
3. Generally, MSE is always positive because of randomness.
4. MSE is the second moment of the error, incorporating the variance of the estimator and its bias. For an unbiased estimator, the MSE is the variance of the estimator.
5. PROOF for the relationship between variance and bias:

$$
\begin{aligned}
&MSE(\hat{\lambda}) = \mathbb{E}[(\hat{\lambda} - \lambda)^2] = \mathbb{E}[(\hat{\lambda} - \mathbb{E}[\hat{\lambda}] + \mathbb{E}[\hat{\lambda}] - \lambda)^2] \\
&= \mathbb{E}[(\hat{\lambda} - \mathbb{E}[\hat{\lambda}])^2 + 2(\hat{\lambda} - \mathbb{E}[\hat{\lambda}])(\mathbb{E}[\hat{\lambda}] - \lambda) + (\mathbb{E}[\hat{\lambda}] - \lambda)^2] \\
&= \mathbb{E}[(\hat{\lambda} - \mathbb{E}[\hat{\lambda}])^2] + \mathbb{E}[2(\hat{\lambda} - \mathbb{E}[\hat{\lambda}])(\mathbb{E}[\hat{\lambda}] - \lambda)] + \mathbb{E}[(\mathbb{E}[\hat{\lambda}] - \lambda)^2] \\
&= \mathbb{E}[(\hat{\lambda} - \mathbb{E}[\hat{\lambda}])^2] + 2(\mathbb{E}[\hat{\lambda}] - \lambda)\mathbb{E}[\hat{\lambda} - \mathbb{E}[\hat{\lambda}]] + (\mathbb{E}[\hat{\lambda}] - \lambda)^2 \\
&= \mathbb{E}[(\hat{\lambda} - \mathbb{E}[\hat{\lambda}])^2] + 2(\mathbb{E}[\hat{\lambda}] - \lambda)(\mathbb{E}[\hat{\lambda}] - \mathbb{E}[\hat{\lambda}]) + (\mathbb{E}[\hat{\lambda}] - \lambda)^2 \\
&= \mathbb{E}[(\hat{\lambda} - \mathbb{E}[\hat{\lambda}])^2] + (\mathbb{E}[\hat{\lambda}] - \lambda)^2 \\
&= Var(\hat{\lambda}) + [Bias(\hat{\lambda}, \lambda)]^2
\end{aligned}
$$

E. Correlation

1. A statistical technique that can show whether and how strongly pairs of variables are related.
2. The main result of a correlation is the correlation coefficient that ranges from -1.0 to +1.0. The closer the correlation coefficient is to -1.0 or +1.0, the more closely the two variables are related.
3. COMPUTATION of the population correlation coefficient denoted as ρ:

Let X and Y be two random variables,X and Y be two random variables,

$$\rho = corr(X, Y) = \frac{cov(X, Y)}{\sigma_x \sigma_y} = \frac{\mathbb{E}[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y}$$

, where μ_x and μ_y are the corresponding expected values, and σ_x and σ_y are corresponding standard deviations.

4. COMPUTATION of the sample correlation coefficient denoted as r:

   Let $(x_1, y_1), \cdots, (x_n, y_n)$ be the pairs of random variables,

   $$r = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{SD(x)} \times \frac{y_i - \bar{y}}{SD(y)}$$

   , where $\bar{x}$ and $\bar{y}$ are the sample means, and $SD(x)$ and $SD(y)$ are the corresponding standard deviations.

F. Logistic Regression

1. A generalized linear model used to model a binary categorical variable using numerical and categorical predictors.
2. Assume a binomial distribution produced the outcome variable. Then model p, the probability of success for a given set of predictors.
3. In regression analysis, logistic regression is estimating the parameters of a logistic model. To finish specifying the logistic model, we commonly use the logit function, which is $$logit(p) = log(\frac{p}{1-p})$$, for $0 \leq p \leq 1$. The logit function takes a value between 0 and 1, and maps it to a value between $-\infty$ and ∞.
4. Inverse logit (logistic) function is given as

   $$g^{-1}(x) = \frac{exp(x)}{1 + exp(x)} = \frac{1}{1 + exp(-x)}$$

   , which takes a value between $-\infty$ and ∞, and maps it to a value between 0 and 1.
5. Note: When interpreting logistic regression, do not treat an odds ratio as a ratio of probabilities.

G. Generalized Linear Mixed-Effects Models

1. Incorporate both fixed-effects parameters and random effects in a linear predictor via maximum likelihood.
2. The linear predictor is related to the conditional mean of the response through the inverse link function defined in the generalized linear model.
3. The expression for the likelihood of a mixed-effects model is an integral over the random effects space.
4. Note: For a generalized linear mixed-effects model, the integral must be approximated.
5. The most reliable approximation for generalized linear mixed-effects models is adaptive Gauss-Hermite quadrature, implemented only for models with a single scalar random effect.

H. Simple Linear Regression vs. Multiple Linear Regression

1. A linear approach to modeling the relationship between a dependent variable and one or more independent variables. Simple linear regression is the case of one independent variable. Multiple linear regression is the case of more than one independent variable.
2. Suppose we have associated variables X and Y. According to simple linear model, expectation of the variabel Y at a given point x is $\mathbb{E}(Y|x) = a + bx + \varepsilon$, where Y is a variable, and x is a fixed point.
3. In general, $\mathbb{E}(Y_i|X_i) = a + bX_i + \varepsilon_i$, where ε_i is true residual. According to Gauss Markov Theorem, if we suppose $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, where $Var(\varepsilon_i) = \sigma^2$ is independent from X_i, then least-square regression is the optimal regression method.
4. In multiple linear regression, we cannot calculate R^2 as the square of the correlation between X and Y since we have multiple x's.
5. With multiple linear regression, we check the constant variance condition using a plot of residuals vs. fitted.

I. Ordered Forest Estimator (Daniel, 2018)

1. Ordered forest estimator generalizes common estimators like ordered probit or ordered maximum likelihood and is able to recover essentially the same output as the standard estimators, such as the probabilities of the alternative conditional on covariates.
2. Consider an ordered outcome variable $Y_i \epsilon (1, \cdots, M)$ with ordered categories m. For a sample of size $N (i = 1, \cdots, N)$, the estimation of the conditional ordered outcome probabilities evaluated at x. That is, $P[Y_i = m | X_i = x]$ is based on an estimation of cumulative probabilities given by binary indicators $Y_{m,i} = 1(Y_i \leq m)$ for m = 1, … , M-1.
3. Then a regression random forest is estimated for all M-1 binary indicators, obtaining the predictions $\hat{Y}_{m,i} = \hat{P}(Y_{m,i} = 1 | X_i = x)$.
4. The prediction for the M-th category is given as $\hat{Y}_{M,i} = 1$ as the cumulative probabilities must sum up to 1.

VIII. Reference

Miljković, D., Gajić, L., Kovačević, A., Konjović, Z. "The use of data mining for basketball matches outcomes prediction." *IEEE 8th International Symposium on Intelligent Systems and Informatics*. 2010, 309-312.

West, Brady T. "A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament" *Journal of Quantitative Analysis in Sports*, 2006, 2(3).

Boulier, B. L., Stekler, B. L. “Are sports seedings good predictors?: an evaluation.” *International Journal of Forecasting*. 1999. 15(1), 83-91.

Goller, D., Knaus, M. C., Lechner M., Okasa, G. “Predicting Match Outcomes in Football by an Ordered Forest Estimator.” 2018, no. 2018-11.