March Data Crunch Madness

Team Jumpman:

Yen-Shao Chou
Xueqing Jin
Hengda Shen
Zhihan Yang

Overview

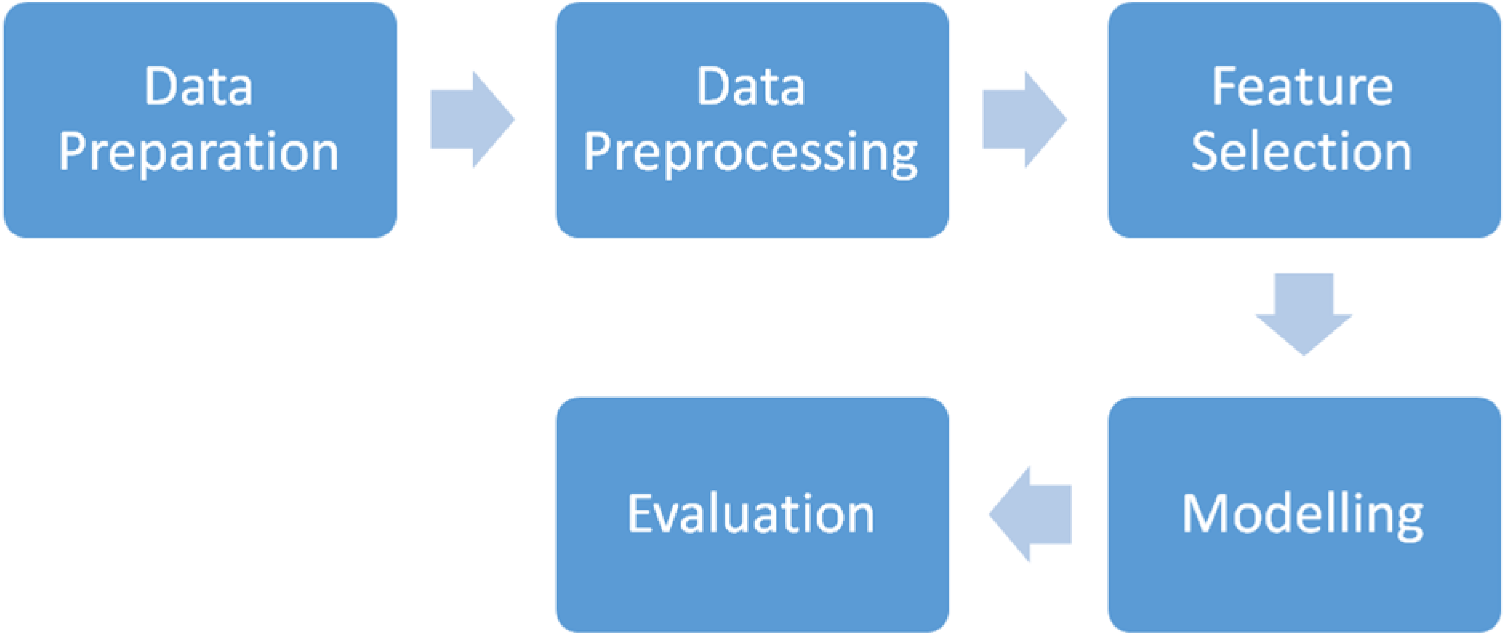
- Introduction
- Data description
- Methodology
- Data preprocessing
- Model
- Evaluation
- Interesting results

Introduction

- The NCAA Men’s Basketball Tournament, a.k.a. March Madness, is held every spring in U.S. with 68 college basketball teams to determine the national championship.

- In March Data Crunch Madness, teams predict the probability that a team wins in all possible matchups. Since March Madness was cancelled in 2020, there will be no Logloss as evaluation. Historical data was given and teams were encouraged to modify the dataset from sources.

Methodology



Data Preparation

Data Preprocessing

Feature Selection

Modelling

Evaluation

Data preparation

- Dataset:

  NCAA_Tourney_2002_2019, Team_RPI, Team_SOS, Team_MasseyOrdinals

- Data sources

  Team_RPI(Rating percentage index): teamrankings.com

  Team_SOS(Strength of schedule): kenpom.com

  Team_MasseyOrdinals: A ranking system from masseyratings.com

Data preprocessing

- We merged datasets we crawled online, namely Team RPI, Team_SOS and Team_MasseyOrdinals to the given dataset

- We shuffled the positions of team1 and team2 and their stats accordingly to produce balanced results

- We generated new features by subtracting the team2 game stats from team1

- We handled the missing values

Features selection

- Variance Inflation factor (VIF)

  Examine the multi-collinearities of features

- Tree-based feature selection

  Analyze the factor importances of features

- L1 - Regularization

  Remove insignificant features

Variance Inflation Factor

- We removed features with VIF greater than 10 one by one and re-run VIF testing to ensure least multicollinearity and max explainability

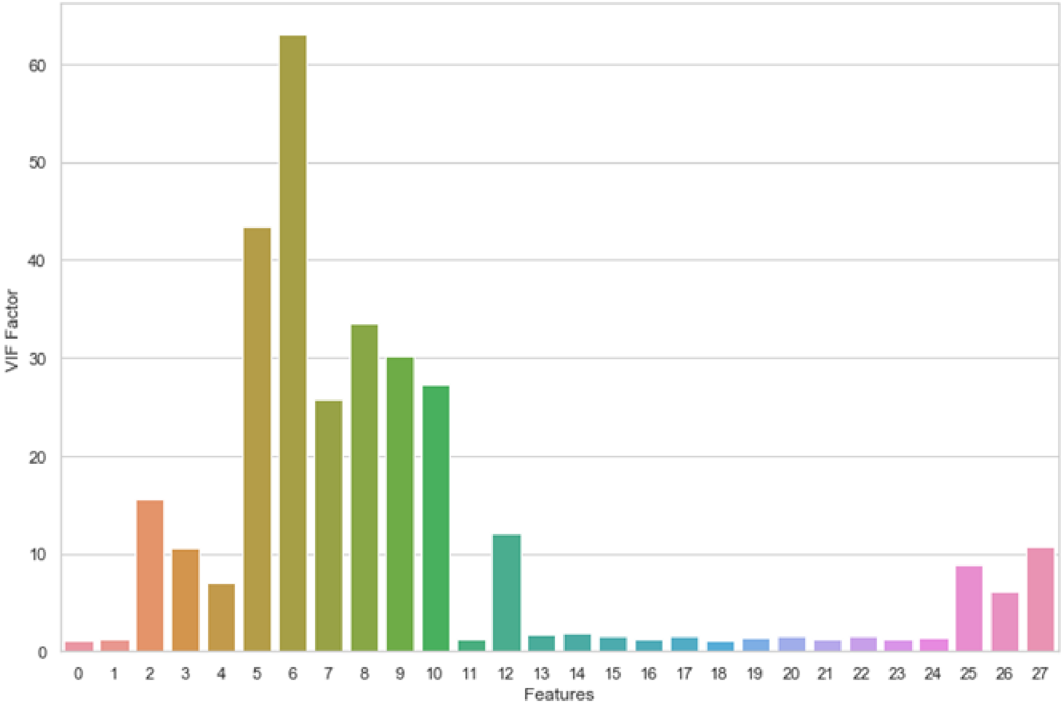


Figure 1. VIF

Tree - Based Feature Selection

- Method: ExtraTreeClassifier from Scikit-Learn

- Top 5 features:

  1. Seed
  2. rpi
  3. Massey Ordinals
  4. Adjusted Offensive efficiency
  5. Adjusted Defensive efficiency

- We keep the top 15 features

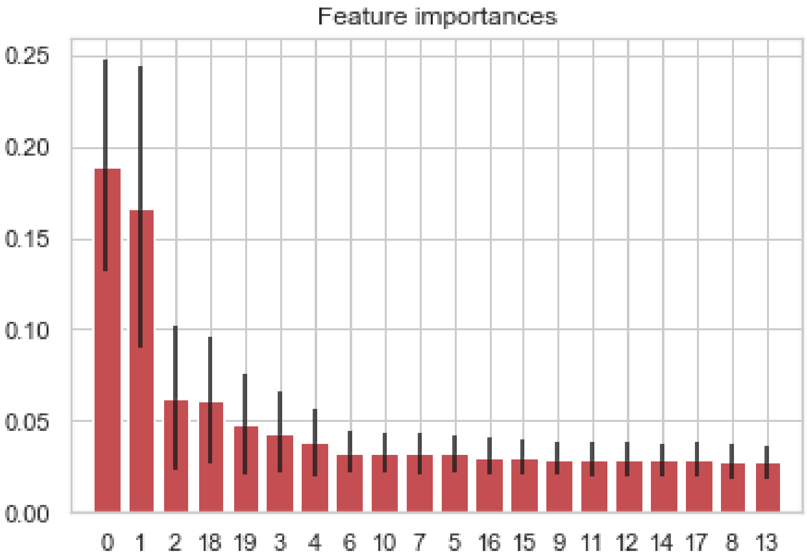


Figure 2: Feature Importance

L1 Regularization

- Also known as Lasso Regression

- Lasso Regression adds “*absolute value of magnitude*” of coefficient as penalty term to the loss function so the coefficient of less important features will be zero.

- We get rid of the insignificant features

Features-Top 5

Features	Description
Seed	Team’s seed in the tournament. There are 4 brackets of teams seeded 1 through 16.
RPI	The rating percentage index is a quantity used to rank sports teams based upon a team's wins and losses and its strength of schedule.
Massey Ordinals	The ratings are totally interdependent, so that a team's rating is affected by games in which it didn't even play. The final ratings represent a state of equilibrium in which each team's rating is exactly balanced by its good and bad performances.
Adjusted Offensive efficiency	An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average D-I defense.
Adjusted Defensive efficiency	An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average D-I offense

Modelling

- We utilized an Ensemble Model with 10-folds Cross Validation
    - Lightgbm (weight = 0.3)
    - XGboost (weight = 0.7)

- To determine the optimized weight, we iterated and evaluated the ensemble model

Evaluation

	Log Loss	Overall F1 - score
Lightgbm	0.442	0.80
XGboost	0.442	0.80
Logistic Regression	0.517	0.75
Ensemble	0.442	0.81

- We splitted three seasons (17, 18, 19) as our testing set.
- We included Logistic Regression as a comparison
- Overall precision of the ensemble model is around 80%.

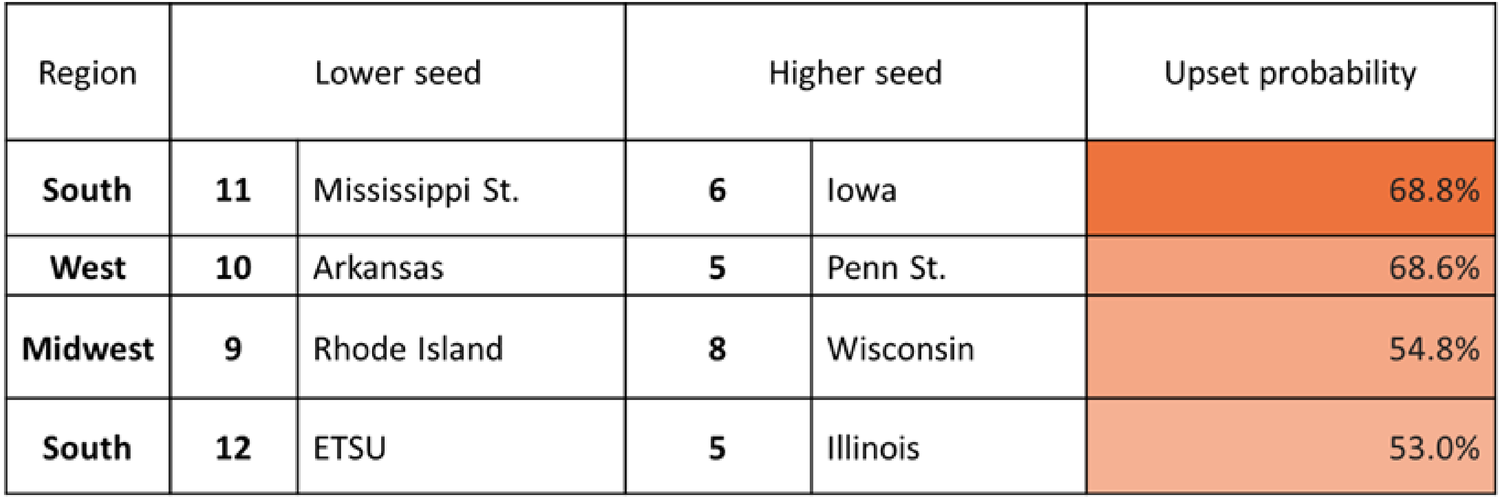
Region Overview

- The table gives the average winning probability of seed-one teams versus the other teams in the same region respectively

- Since the Mean Win_prob of region Y is much higher, and other teams of Y are also competitive according to the history. We present that seed-one of region Y, namely Kansas has a high chance of winning Final Four, or even Championship

Region	Mean Win_prob
W (South)	0.878
X (East)	0.858
Y (Midwest)	0.921
Z (West)	0.868

Upsets

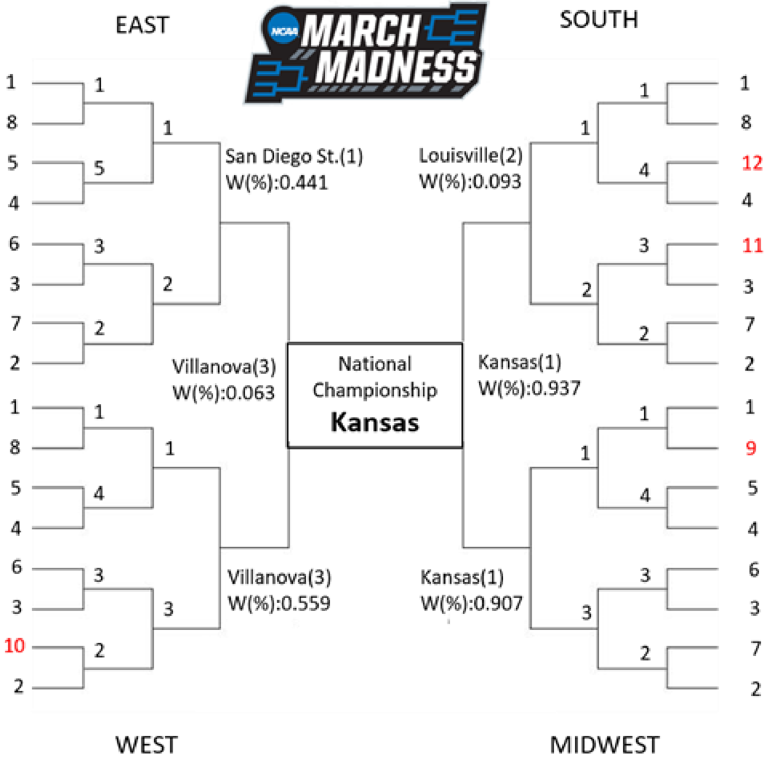


Region	Lower seed		Higher seed		Upset probability
South	**11**	Mississippi St.	**6**	Iowa	68.8%
West	**10**	Arkansas	**5**	Penn St.	68.6%
Midwest	**9**	Rhode Island	**8**	Wisconsin	54.8%
South	**12**	ETSU	**5**	Illinois	53.0%

- Upsets predicted by our model for the first round
- Our model suggested that there would be 4 upsets in 2020. Mississippi State(11th) has the highest probability to win as a underdog against Iowa(6th).
- East Tennessee State was predicted to win as a underdog with the largest seed difference.

Championship prediction

- The NCAA 2020 bracket of the prediction starts from 2nd round.
- We assumed region matchups(E v.s. W, S v.s. M) would be the same as 2019.
- Highlighted numbers refers to our predicted upsets in 1st round.
- Based on the model, the final four would be San Diego St., Villanova, Louisville, Kansas.
- **The predicted NCAA 2020 championship is Kansas with a high winning percentage against Villanova, it coincides to we proposed early.**


EAST
MARCH MADNESS
SOUTH
1
8
5
4
6
3
7
2
1
5
3
2
1
2
San Diego St.(1)
W(%):0.441
Louisville(2)
W(%):0.093
1
4
3
2
1
2
1
8
12
4
11
3
7
2
Villanova(3)
W(%):0.063
National Championship
Kansas
Kansas(1)
W(%):0.937
1
8
5
4
6
3
10
2
1
4
3
2
1
3
Villanova(3)
W(%):0.559
Kansas(1)
W(%):0.907
1
4
3
2
1
3
1
9
5
4
6
3
7
2
WEST
MIDWEST