March Madness Report

GUCCIGANG

Tram Ngoc Le, Yanqi Shi,
Bao Ngoc Dinh, Kefan Zha

Introduction

- One of the most famous and popular sporting events annually in the United States
- Formed by 68 Division I level college basketball team
- To generate a national champion ship, the competition is divided by several rounds into:
    - First Four
    - First and Second Round
    - Sweet 16/Elite Eight
    - Final
- The prediction on results of March Madness is a popular culture in the United States.

Problem Statement

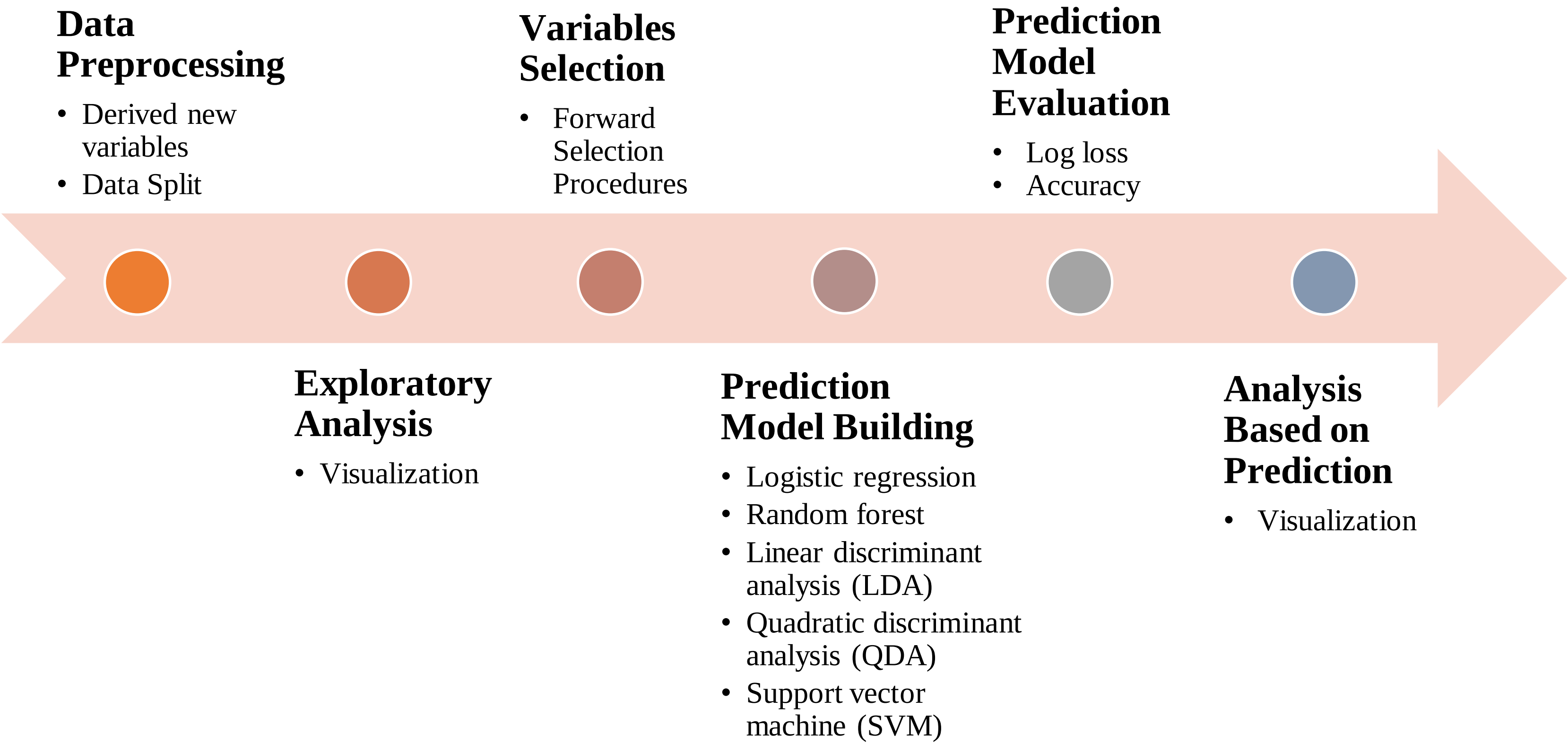
What are we proposed to do?

- Descriptive analytics with visualization on data from 2002 to 2019.
- Prediction model building on data from 2002 to 2019.

Why is this important?

- A reference for people who are interested in predicting the game.
- Used by NCAA teams to make improvement for their performance.
- Used by NCAA to adjust their marketing strategies such as ticket price sales.

Methodology Diagram



Data Preprocessing

- Derived new variables
- Data Split

Exploratory Analysis

- Visualization

Variables Selection

- Forward Selection Procedures

Prediction Model Building

- Logistic regression
- Random forest
- Linear discriminant analysis (LDA)
- Quadratic discriminant analysis (QDA)
- Support vector machine (SVM)

Prediction Model Evaluation

- Log loss
- Accuracy

Analysis Based on Prediction

- Visualization

Data Preprocessing

Derived Novel Variables

- Team’s winning rate – Calculated as ‘wins/(wins+losses)’.
- Team’s win-loss ratio – Calculated as ‘wins/losses’.
- Coach’s winning rate – Calculated as ‘wins/(wins+losses)’.
- Coach’s win-loss ratio – Calculated as ‘wins/losses’.
- Seed Differences – Calculated as ‘strong seed - weak seed’.
- Whether team 1 wins – Derived by scores after randomly switch team 1 and team 2.

Variables Removal

- Remove two teams’s variables in association with wins and losses.

Split Data (Used in prediction model building)

- Training Data: 2002 – 2018; Test Data: 2019
- Data needed prediction: 2020

Hypotheses

- **Hypothesis 1:**

A team’s defense and offense are both important. Thus, a team with excellent offense and defense will have a bigger likelihood of winning.

- **Hypothesis 2:**

The team’s seed number is a big factor in their probability of winning. In other words, the stronger the seed, the bigger the likelihood of winning.

Exploratory Analysis I

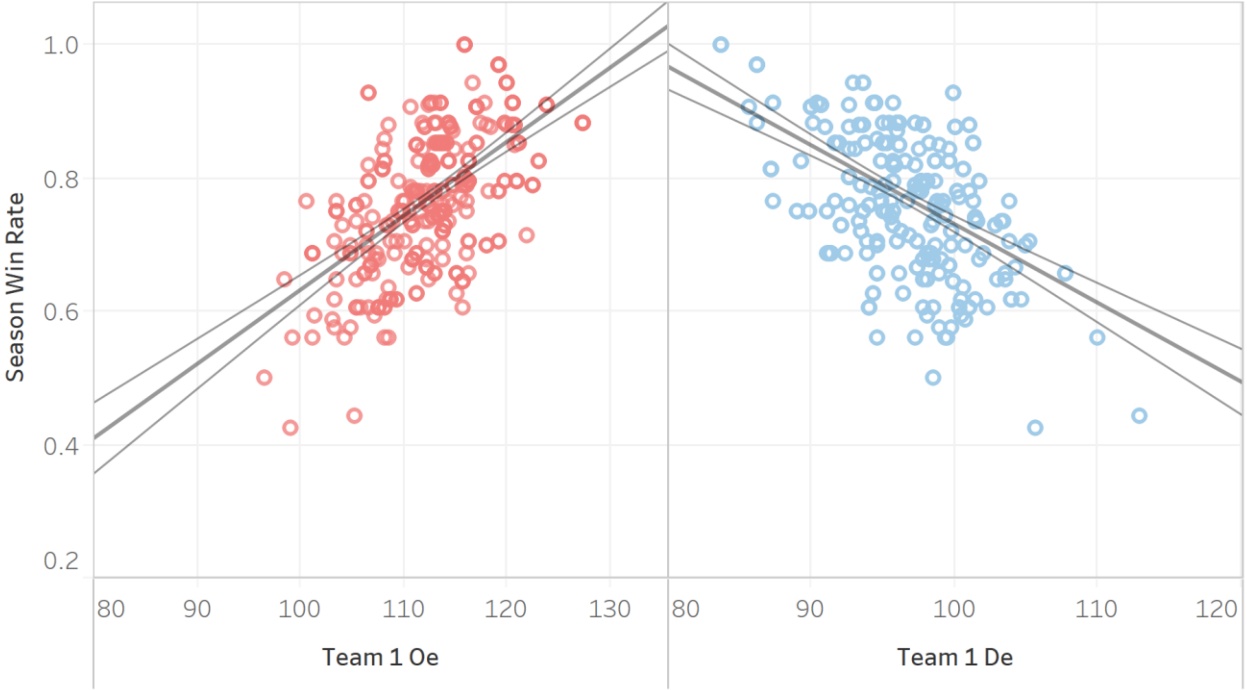


Figure 1. Scatter Plot – Trend Line: Season Win Rate vs. Offense Points and Defense Points (2015-2019)

- Figure 1 show the relationship between the season win rate for a team vs. its offensive points and defensive points.
- There is a positive correlation between offense points and season win rate. The higher offensive possession points a team scored, the higher win rate the team will get. There is a negative correlation between defense points and season win rate. The higher defensive possession points a team allows its opponent to score, the lower win rate the team will get.
- Both offense power and defense power affect the win rate.

Exploratory Analysis II

- Figure 2 demonstrate a comparison of the number of NCAA Sweet Sixteen appearances at current school vs career overall number of NCAA Sweet Sixteen appearances in 2019.
- The smaller the seed number, the bigger the amount of appearances, whether it's overall or at current school. For seed number bigger than 12, the appearance at sweet sixteen is very rare.
- Seed can be considered as a big factor in predicting the probability of winning.

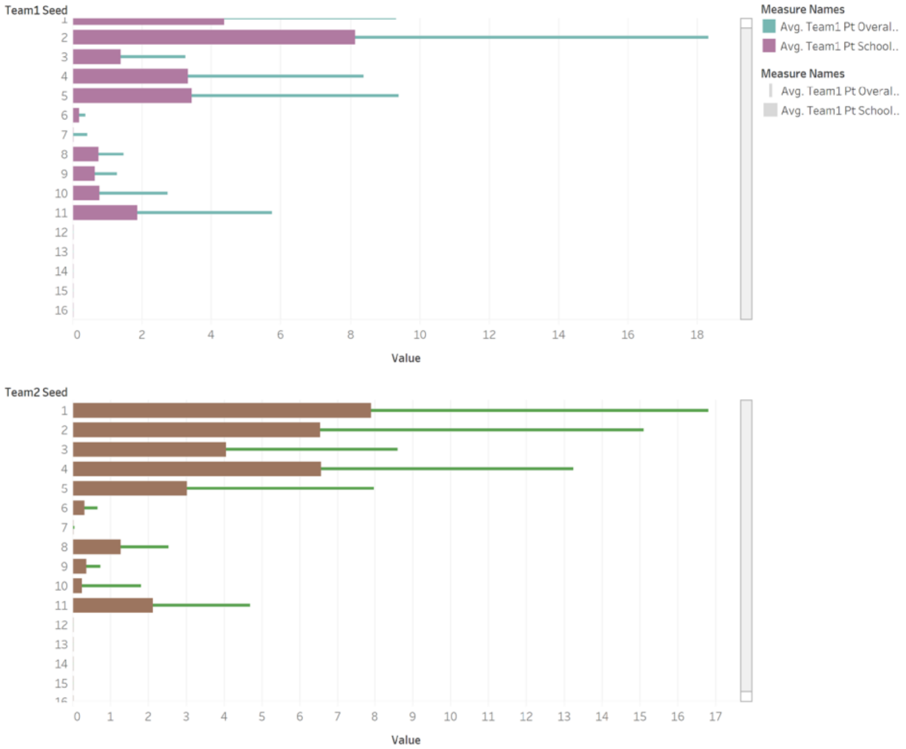


Figure 2. Bar Chart – Comparison of number of NCAA Sweet Sixteen appearances at current school and career overall number of NCAA Sweet Sixteen appearances (2019)

Variable Selection

- Forward Selection Procedures – Choose the most significant variables.
- Pick the explanatory variables combination with lowest BIC (Figure 3), which consists of 9 variables:

  Team1_seed, Team2_seed, Team1_oe, Team2_oe, Team1_de, Team2_de, Team2_pt_school_s16, Team1_pt_overall_s16, Team1_arate
- Since we randomly switch team 1 and team 2 in data preprocessing, we add relevant variables above for both team 1 and team 2.
- These selected variables correspond to our exploratory analysis.
- Detailed data dictionary is shown on next slide.

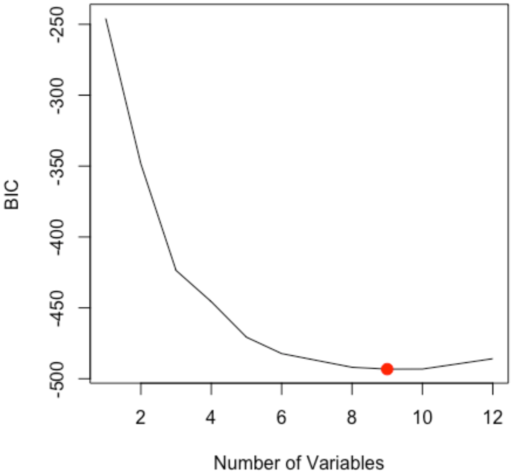


Figure 3: BIC value in subset selection

Variable Name	Scale	Description	Variable Type
Team1_seed	Ordinal	Team 1’s seed in the tournament, ranking team power (1 is the strongest)	Explanatory Variable
Team2_seed	Ordinal	Team 2’s seed in the tournament, ranking team power (1 is the strongest)	Explanatory Variable
Team1_arate	Ratio	Team 1’s percentage of field goals that were preceded by an assist	Explanatory Variable
Team2_arate	Ratio	Team 2’s percentage of field goals that were preceded by an assist	Explanatory Variable
Team1_oe	Continuous	Team 1’s offensive power - Points scored per 100 offensive possessions	Explanatory Variable
Team2_oe	Continuous	Team 2’s offensive power - Points scored per 100 offensive possessions	Explanatory Variable
Team1_de	Continuous	Team 1’s defensive power - Points allowed per 100 defensive possessions	Explanatory Variable
Team2_de	Continuous	Team 2’s defensive power - Points allowed per 100 defensive possessions	Explanatory Variable
Team1_pt_school_s16	Continuous	Team 1’s number of NCAA Sweet Sixteen appearances at current school	Explanatory Variable
Team2_pt_school_s16	Continuous	Team 2’s number of NCAA Sweet Sixteen appearances at current school	Explanatory Variable
Team1_pt_overall_s16	Continuous	Team 1’s career overall number of NCAA Sweet Sixteen appearances	Explanatory Variable
Team2_pt_overall_s16	Continuous	Team 2’s career overall number of NCAA Sweet Sixteen appearances	Explanatory Variable
Team1_pt_team_season_winrate	Ratio	The possibility of a team 1 to win in this season	Explanatory Variable
Team2_pt_team_season_winrate	Ratio	The possibility of a team 2 to win in this season	Explanatory Variable
Team1_pt_coach_season_winrate	Ratio	The possibility of winning if this coach of team 1 is in charge this season	Explanatory Variable
Team2_pt_coach_season_winrate	Ratio	The possibility of winning if this coach of team 2 in charge this season	Explanatory Variable
Team1_WoLI	Binary	Team 1’s result of win or loss (1=WIN, 0=LOSS)	Response Variable
Game_id	Nominal/Numeric	The identification number of a game	Unique Identifier
Season	Date	The year of season	Control Variable
Team1_id	Nominal/Numeric	Team 1’s identification number	Control Variable
Team2_id	Nominal/Numeric	Team 2’s identification number	Control Variable

Prediction Model Building

- Logistic regression
- Random forest
- Linear discriminant analysis (LDA)
- Quadratic discriminant analysis (QDA)
- Support vector machine (SVM)

	Logistic regression	Random forest	LDA	QDA	SVM
Log loss	0.5402133	0.5227263	0.5378359	0.9806795	0.6019181
Accuracy	71.64%	71.64%	73.13%	67.16%	67.16%

Model Summary – Random Forest

- 5 variables are randomly sampled as candidates at each split
- 600 trees to grow
- Random forest would give different vote results each time since it is ‘random’
- Repeat training procedure 500 times to pick the best ‘forest’
- Final model’s performance on test data:

Log loss: 0.4997394

Accuracy: 71.64%

Analysis Based on Prediction - Likelihood

Likelihood	Probability
Most Likely to Win	p >= 0.95
Likely to win	0.68 <= p < 0.95
50-50 case	0.32 < p < 0.68
Likely to Lose	0.05 < p <=0.32
Most Likely to Lose	p <= 0.05

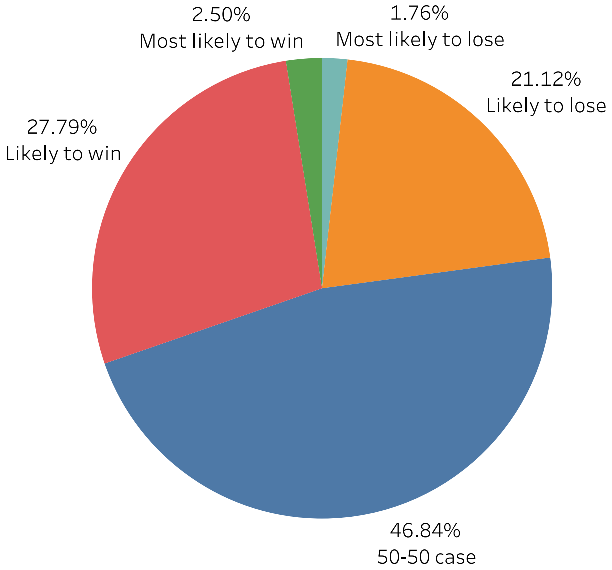


Figure 4. Pie Chart: Team 1 Winning Probability Distribution

- Figure 4 shows the percentage of likelihood for team 1 winning the game.
- Nearly 50% of the games will be a 50-50 case. Only 4.2% of the games will be ‘easy games’.
- People can choose 50-50 case games to watch since these games may be in a stalemate.

Analysis Based on Prediction - Upsets

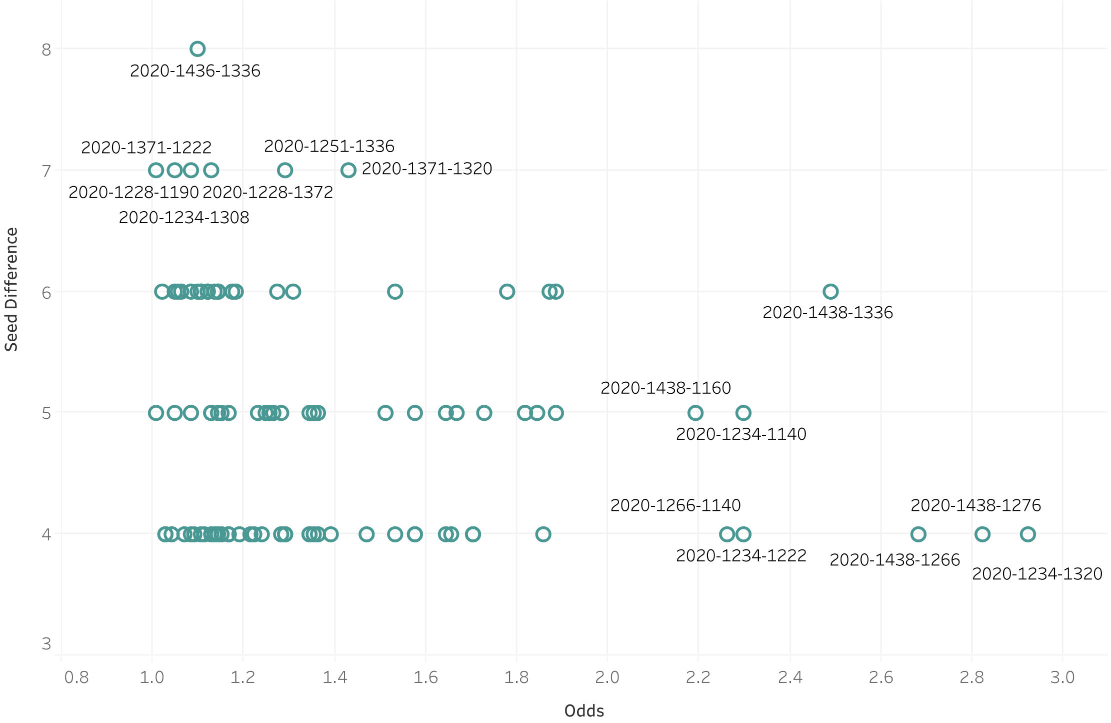


Figure 5. Scatter plot: Seed Difference vs. Winning Odds in Upsets

- Figure 5 shows the predicted upsets with seed difference larger than 4. Upsets with large seed difference or with high odds are labeled.
- 8 games are predicted to be upsets with high odds, which means the weaker team has relatively high probability to beat the stronger team; 7 games are predicted to be upsets with large seed difference, which means a much weaker team may beat a much stronger team.
- Both audience and NCAA should give attention to these games since they may have results against all expectations.

Conclusion & Improvement

Conclusion

- A reasonable model of March Madness prediction has been established.
- The likelihood of winning for games in March Madness 2020 has been obtained.
- Some potential upsets are predicted for people who are interested in.

Limitation & Improvement

- Limited sample size and large time span of the training data could weaken our model.
- More implicit variables and more complex algorithms may improve the accuracy and reduce the log loss.

Reference

- Adit, D. (2017, March 12). Applying Machine Learning To March Madness. Retrieved from https://adeshpande3.github.io/Applying-Machine-Learning-to-March-Madness
- Conor, D. (2018, March 15). Machine Learning Madness: Predicting Every NCAA Tournament Matchup. Retrieved from https://towardsdatascience.com/machine-learning-madness-predicting-every-ncaa-tournament-matchup-7d9ce7d5fc6d
- Kaggle. (2017, April 4). March Machine Learning Mania 2017. Retrieved from https://www.kaggle.com/c/march-machine-learning-mania-2017
- Lotan, W. (2019, April 21). How We Predicted March Madness Using Machine Learning. Retrieved from https://medium.com/@lotanweininger/march-madness-machine-learning-2dbacc948874