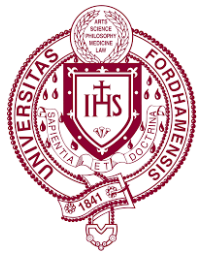
Building Machine Learning Model to Predict the 2019 NCAA March Madness Bracket



Katie Cao

Table of Contents

Instead of analyzing the stats of each team….

Transform the features into DIFFERENCE and QUOTIENT in each of 63 games and predict the probability

Current Log Loss **0.53**

Model Accuracy **79.10%**

PIPELINE

FEATURE ENGINEERING → DATA PREPROCESSING → FEATURE SELECTION → MODEL COMPARISON → PREDICTION

TOOLS USED

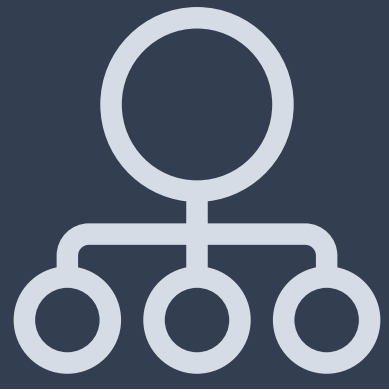
Python

Data Preprocessing
Feature Engineering
Model Construction
Data Visualization

Excel

Log Loss Calculation
Data Preview
Prediction Submission

Step 1: FEATURE ENGINEERING



VARIABLE DIFFERENCES ARE CONSIDERED

Contrast the variable differences between 2 teams in each game, instead of analyzing each individual team

TRANSFORMATION OF FEATURES

Transform all the features of 2 teams in each game into Difference and Quotient.

$$\text{Difference} = \text{Team 1 Feature n} - \text{Team 2 Feature n}$$

$$\text{Quotient} = \frac{\text{Team 1 Feature n}}{\text{Team 2 Feature n}}$$

SIMILAR FEATURES ELIMINATED

Ignore the similar features of each team such as Team1_blockpct, Team2_blockpct, etc. because those are irrelevant in this method.

Step 2: DATA PREPROCESSING

1. Shuffle and Switch

Original Dataset	How We Processed the Data
Sorted by year	Shuffle the year
Team 1 is always the winning team	Switch Team 1 and Team 2 data

2. Replace NULL values in 'ap_ranking' with 45

Each team has four ranking variables where only the top 25 teams will have a value. We assigned teams with NULL value a rank of 45.

3. Replace all 0 with 0.1

Each team has 10 variables that include 0 value. We replaced 0 with 0.1 to avoid those being considered as null in further statistical analysis.

Step 3: FEATURE SELECTION

REMOVING CORRELATED VARIABLES

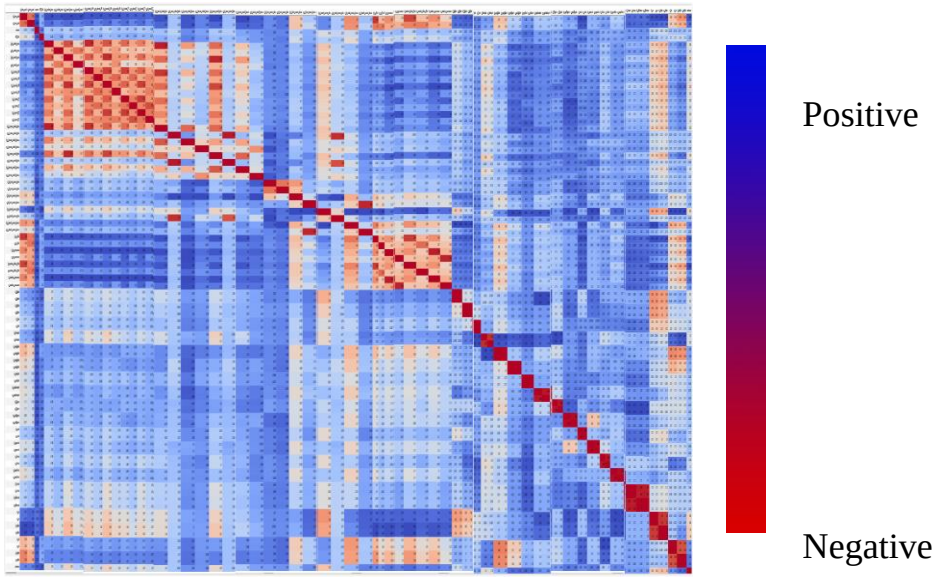


Figure 1. Correlation Matrix of All Variables

~~> 0.9~~

- Calculated a correlation matrix
- Removed variables that have a correlation higher than 0.9.

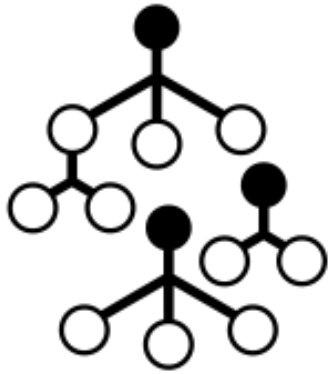
- As a result, 30 variables were dropped

This was done in order to avoid multicollinearity problem.

Step 3: FEATURE SELECTION – CON'T

USING RANDOM FOREST TO PERFORM FEATURE SELECTION

Use Random Forest to select top 11 variables with the highest predictor importance as input variables.



	Selected Variables	Explanation
1	**d_team_seed**	Team seed in the tournament [difference]
2	**q_team_seed**	Team seed in the tournament [quotient]
3	**q_ap_final**	The final AP Poll ranking of each team (top 25 only) 16 [quotient]
4	**d_ap_preseason**	The preseason AP Poll ranking of each team (top 25 only) [difference]
5	**q_ap_preseason**	The preseason AP Poll ranking of each team (top 25 only) 16 [quotient]
6	**d_coaches_before_final**	The most recent Coaches Poll rankings before the final [difference]
7	**q_coaches_before_final**	The most recent Coaches Poll rankings before the final 16 [quotient]
8	**d_oppfg3pct**	Opponent's shooting percentage on 3 point field goals [difference]
9	**d_oe**	Points scored per 100 offensive possessions [difference]
10	**d_adjoe**	An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average D-I defense [difference]
11	**d_adjde**	An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average D-I offense [difference]

Figure 2. Table Showing 11 Selected Variables With Explanation

Step 4: BUILD & EVALUATE DIFFERENT MODELS

	Logistic Regression	Random Forest Classifier	Support Vector Classifier	Gradient Boosting
Accuracy	74.62%	79.10%	70.15%	74.63%
Log Loss	0.5701	0.5252	0.5618	0.5469

Figure 3. Table Comparing The Performance Of Four Models

Random Forest Classifier is the most optimal model:

- Accuracy: 79.10%
- Log Loss: 0.5252
- 5-fold Cross Validation Score: 0.7420
- F-1 Score: 0.76 (Team 2 Wins) 0.82 (Team 1 Wins)

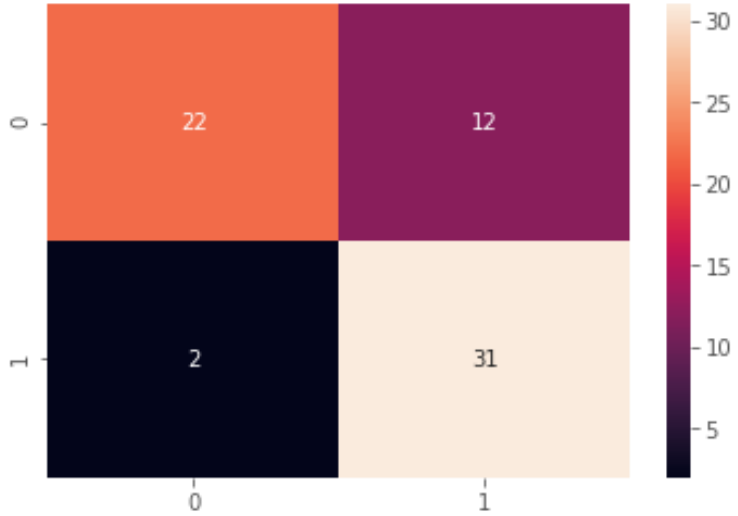
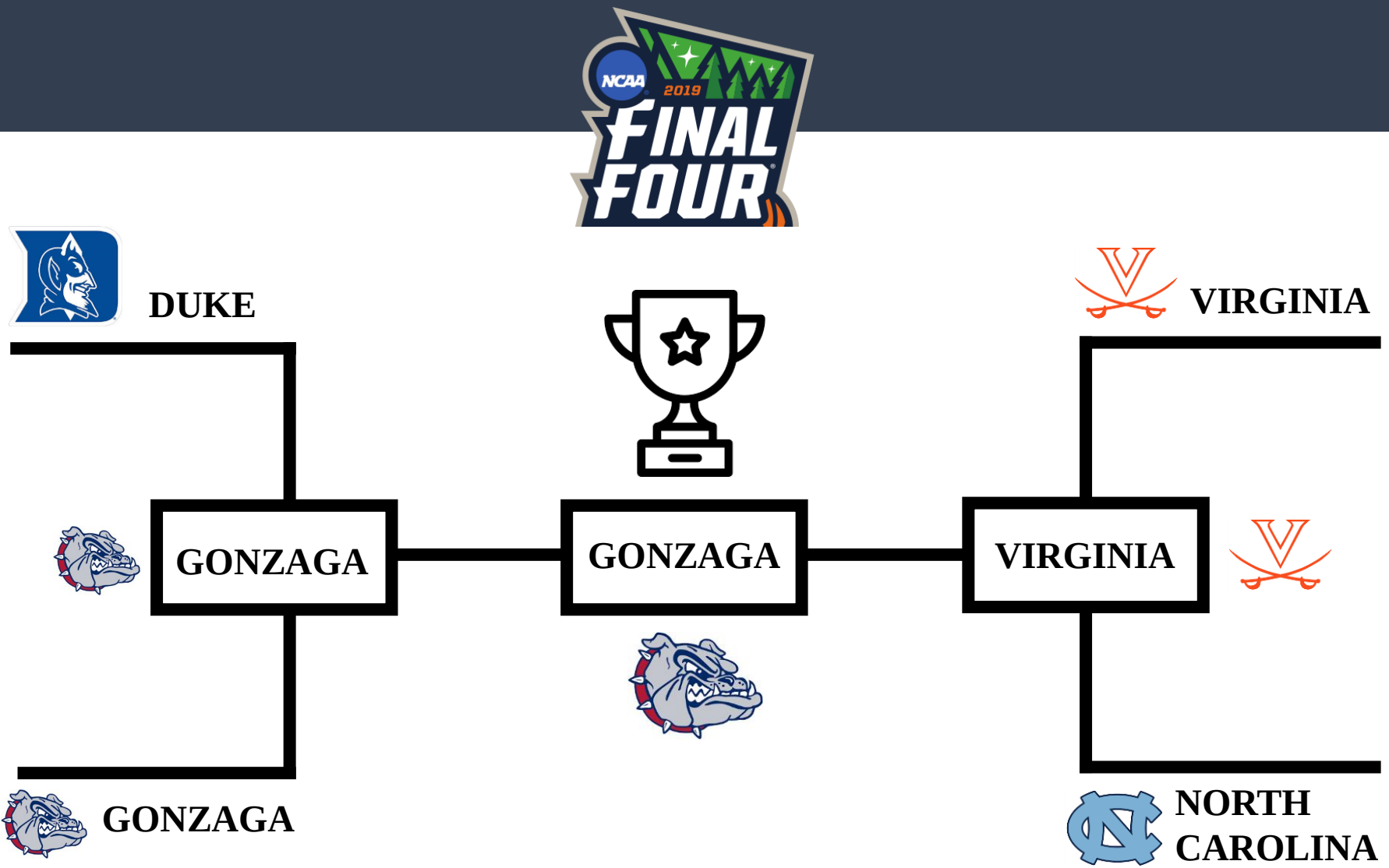


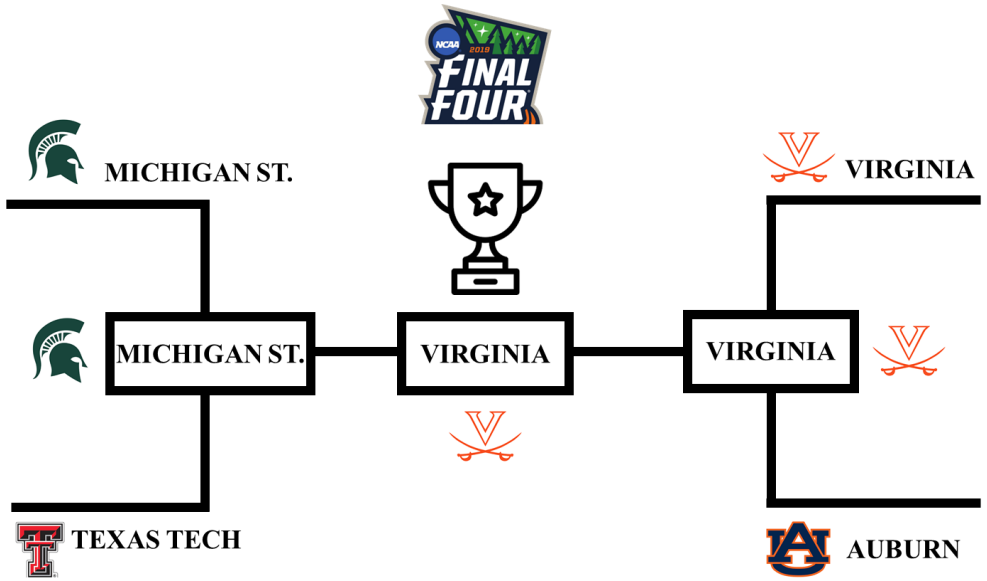
Figure 4. Confusion Matrix for Random Forest Classifier

Step 5: FINAL FOUR AND CHAMPIONSHIP PREDICTION



CHAMPION PREDICTION

BASED ON ACTUAL CURRENT FINAL FOUR AS OF MARCH 2019



Only .02 percent of people correctly predicted the 2019 Final Four in the NCAA Bracket Challenge Game

WHO THE WORLD PICKED TO BE THE FINAL FOUR

CORRECT FINAL FOUR PICKS	PERCENT
0	43.75%
1	45.07%
2	10.32%
3	0.84%
4	0.02%

YEAR	PERCENT OF BRACKETS WITH PERFECT FINAL FOUR
2011	0.0
2012	0.31
2013	0.0
2014	0.006
2015	1.61
2016	0.09
2017	0.003
2018	0.003
2019	0.02

Figure 5. Data Based on Brackets Entered Into NCAA's Bracket Challenge Game

Limitations & Further Suggestions

LIMITATIONS

- **Upsets happen** every season and that could easily mess with your log loss
- Model is based heavily on features team seed, AP Ranking and Coach Ranking. There should be **a random noise feature added** to account for upsets, especially in the last few matches
- **Bias in machine learning model**: Models are black boxes. By assessing several ways to evaluate a model only then can we commit to reduce biases.

FURTHER SUGGESTIONS FOR IMPROVEMENT

- Spend more time **tuning the hyperparameters** using Grid Search and other methods.
- Collect and quantify **alternative data points**: sports betting data, expert opinion, fan polls
- Use PCA and other **dimension reduction methods** to predict considering sports analytics data always a vast amount of variables but maybe fewer rows

Thank you.