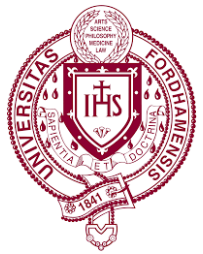
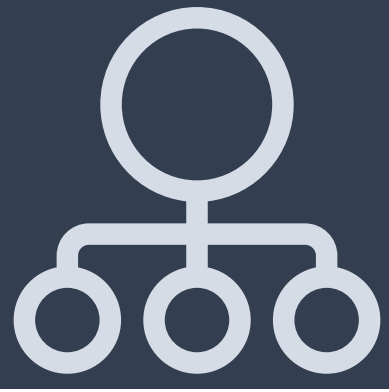
Building Machine Learning Model to Predict the 2019 NCAA March Madness Bracket



Katie Cao

Table of Contents

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

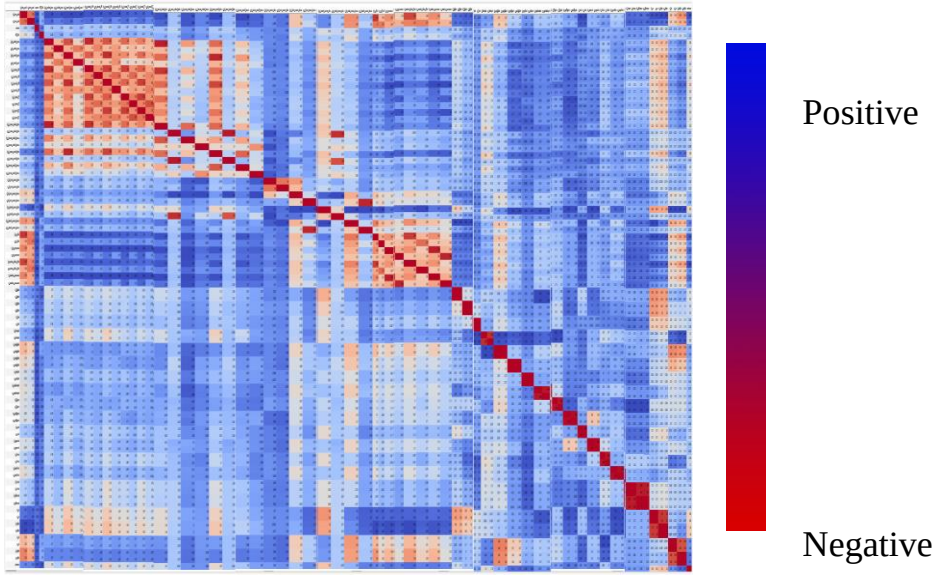


Figure 1. Correlation Matrix of All Variables

~~> 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 12 variables with the highest predictor importance as input variables.

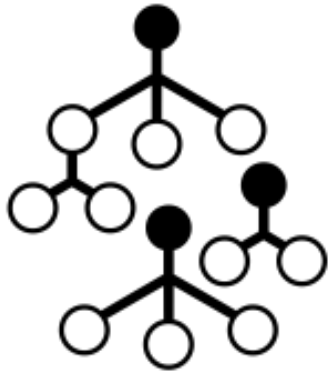
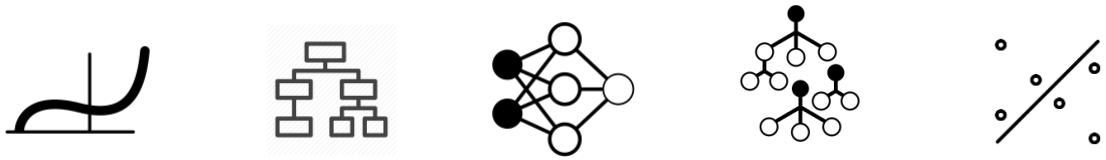


Figure 2. Table Showing 12 Selected Variables With Explanation

	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_pt_overall_s16**	Career overall number of NCAA Sweet Sixteen appearances 16 [quotient]
4	**q_ap_final**	The final AP Poll ranking of each team (top 25 only) 16 [quotient]
5	**d_ap_preseason**	The preseason AP Poll ranking of each team (top 25 only) [difference]
6	**q_ap_preseason**	The preseason AP Poll ranking of each team (top 25 only) 16 [quotient]
7	**d_coaches_before_final**	The most recent Coaches Poll rankings before the final [difference]
8	**q_coaches_before_final**	The most recent Coaches Poll rankings before the final 16 [quotient]
9	**d_oppfg3pct**	Opponent's shooting percentage on 3 point field goals [difference]
10	**d_oe**	Points scored per 100 offensive possessions [difference]
11	**d_adjoe**	An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average D-I defense [difference]
12	**d_adjde**	An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average D-I offense [difference]

Step 4: BUILD & EVALUATE DIFFERENT MODELS



	Logistic Regression	Decision Tree	Artificial Neural Network	Random Forest	SVM	XGBoost
Accuracy	70.89%	71.56%	68.44%	68%	71.33%	70.89%
Precision (For outcome = 1)	70.54%	71.07%	67.19%	66.93%	70.11%	70.37%
Recall (For outcome = 1)	73.91%	74.78%	74.78%	73.91%	76.52%	74.35%

Figure 3. Table Comparing The Performance Of Five Models

Step 4: BUILD & EVALUATE DIFFERENT MODELS – CON'T

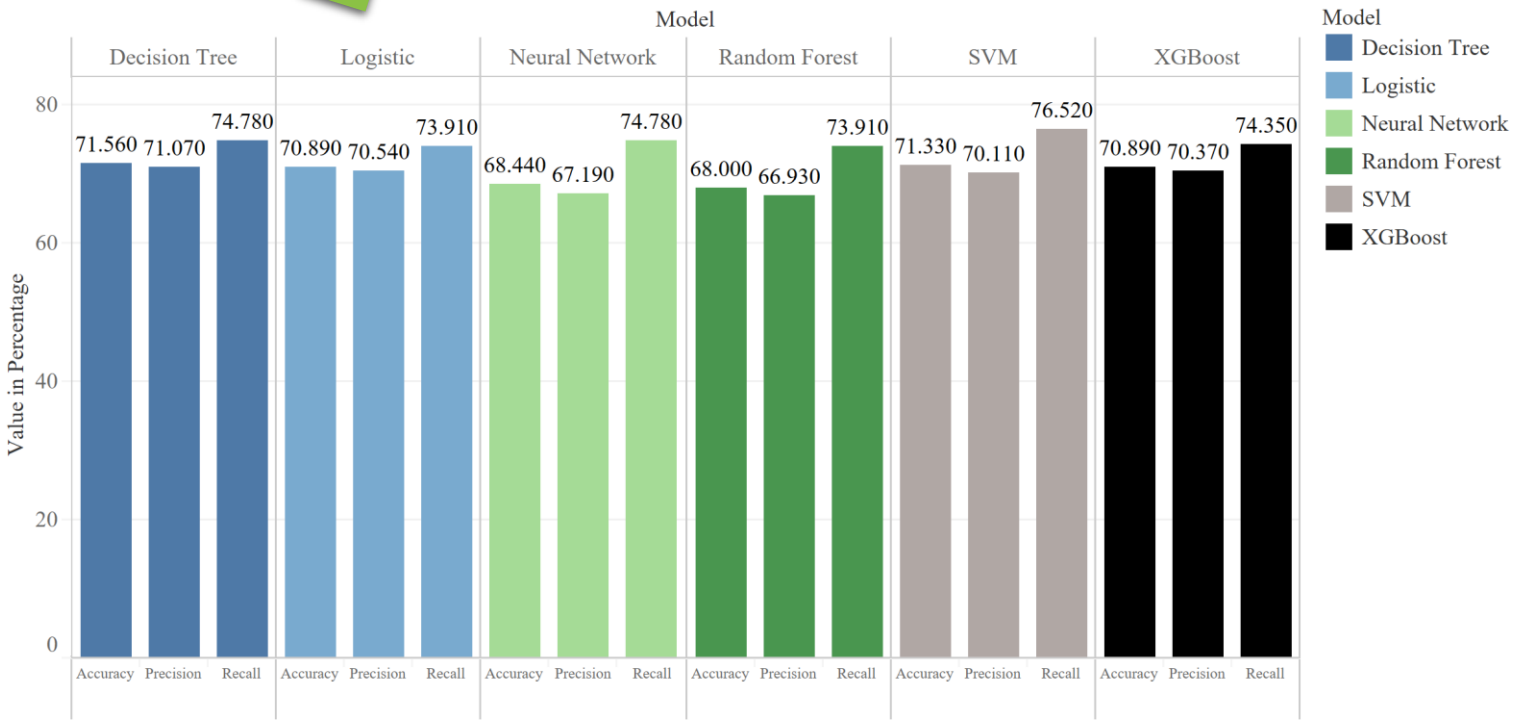
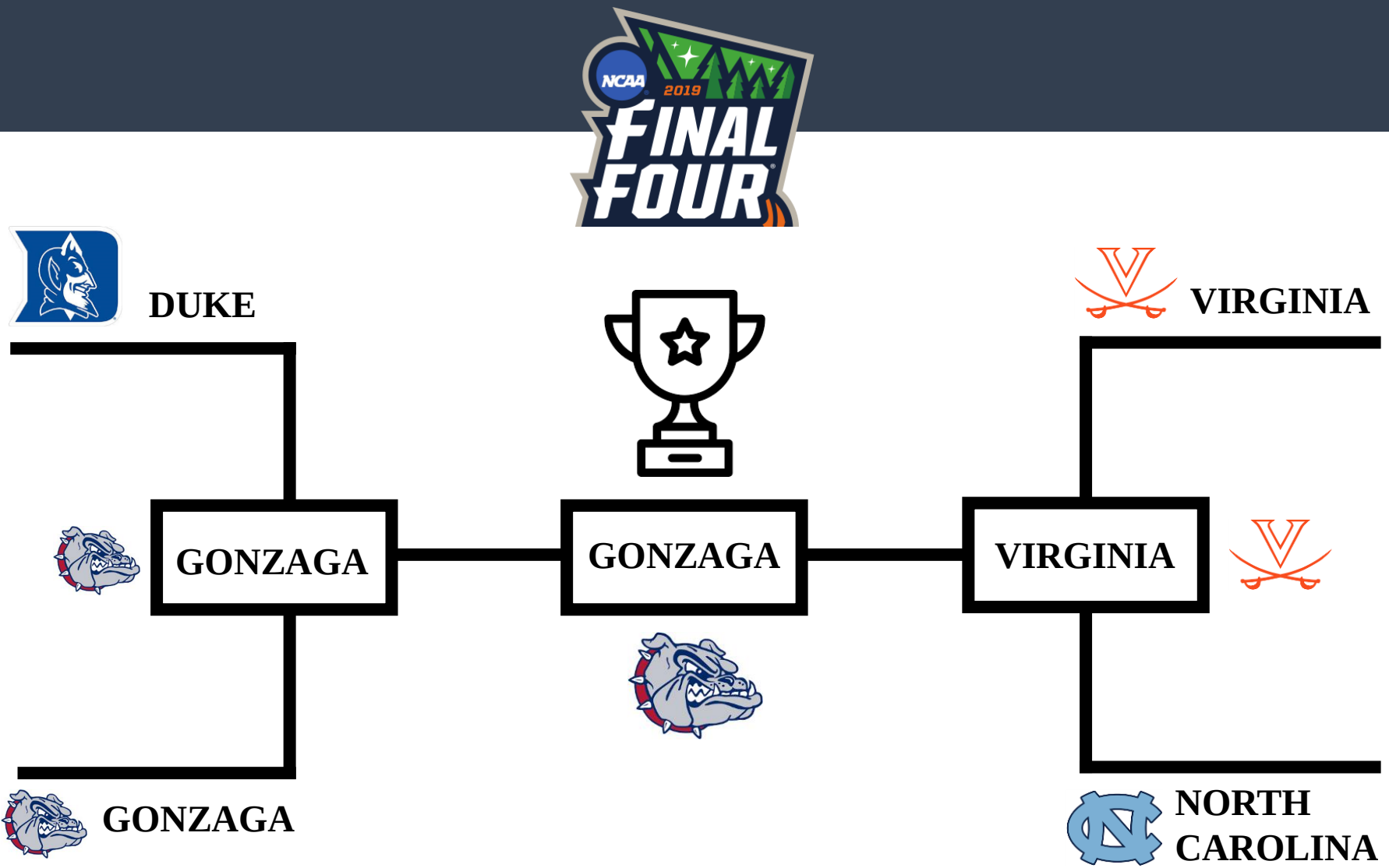


Figure 4. Bar Chart Comparing The Performance Of Five Models

Logistic regression model is the most optimal model, given its performance on accuracy, precision and recall and its ability to generate a probability of prediction.

Step 5: FINAL FOUR AND CHAMPIONSHIP PREDICTION



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 **0.49**

Model Accuracy **70.89%**

PIPELINE

TOOLS USED

Python

Data Preprocessing
Feature Engineering

SPSS Modeler

Model Construction
Model Evaluation

Excel

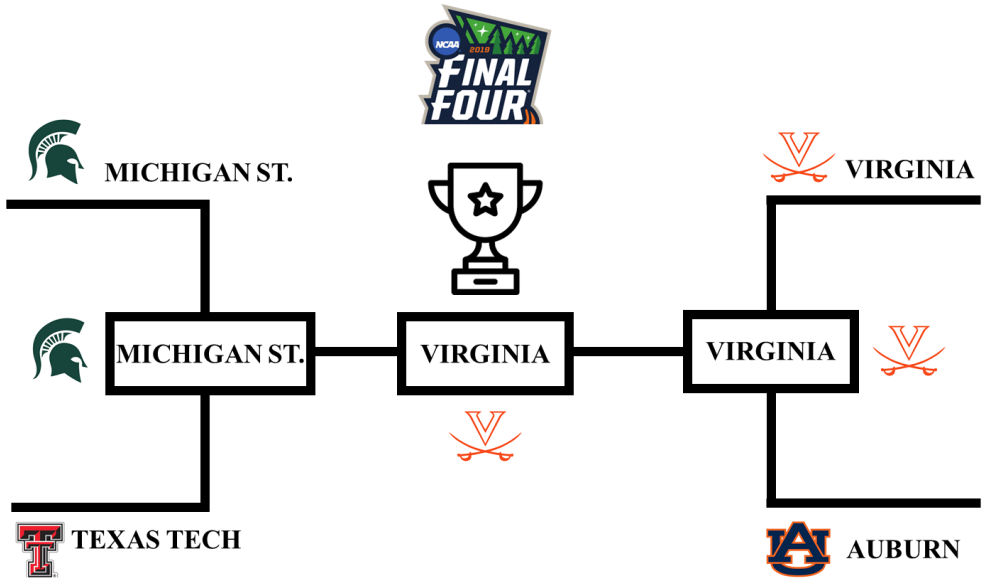
Log Loss Calculation
Data Preview

Tableau

Data Visualization

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

Thank you.