PREDICTING NCAA MEN'S BASKETBALL TOURNAMENT RESULTS

Frederick McCollum

1

OVERVIEW

Background

Objectives

Methodology

Results

Uses of this Model

2

BACKGROUND

3

OCCURS EVERY YEAR BEGINNING IN MARCH (MARCH MADNESS)

THERE HAS NEVER BEEN A PERFECTLY PREDICTED BRACKET

63 TOTAL GAMES IN THE TOURNAMENT

9,223,372,036,854,775,808 TOTAL BRACKET COMBINATIONS

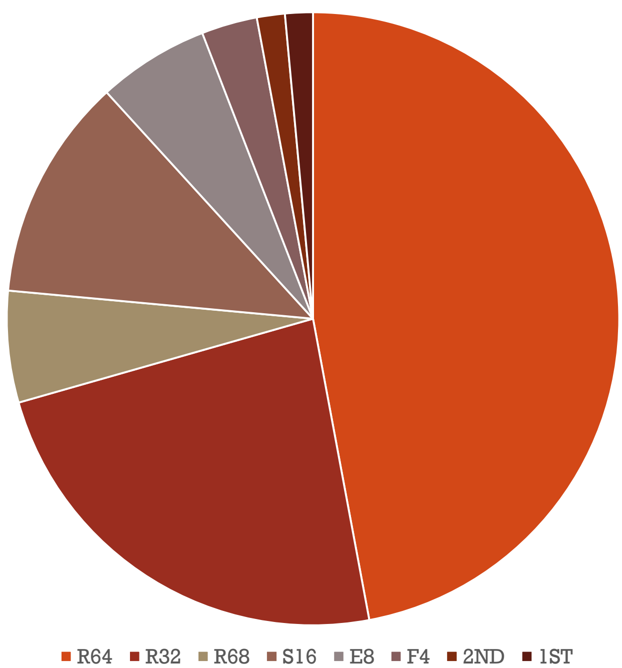
NCAA MEN'S BASKETBALL TOURNAMENT

4

DATA OVERVIEW

- Data obtained from Kaggle (Andrew Sundberg)
- NCAA men's basketball team statistics
- 2015 – 2019 (5 years)
- Contains team's final result

5



DATA DISTRIBUTION

6

OBJECTIVES

7

- Predict NCAA Men's Basketball Tournament results
- Identify the best type of predictive model
- Determine variables with the most significant impact on predictions

OBJECTIVES

8

METHODOLOGY

9

1. Load data
2. Select features
3. Preprocess data
4 Train model(s)
5. Test model(s)
6. Compare results

METHODOLOGY OVERVIEW

10

FEATURE SELECTION

TEAM	CONF	G	W	WINPERCENT	ADJDE	BARTHAG
EFG_O	EFG_D	TOR	TORD	ORB	DRB	FTR
FTRD	TP_O	2P_D	3P_O	3P_D	ADJ_T	WAB
	ADJOE	SEED	YEAR	POSTSEASON		

11

DATA WAS PREPROCESSED TO INCREASE PREDICTIVE EFFICACY OF THE DATA

- Identify and remove near-zero variance predictors (none removed)
- Remove highly correlated predictors (none removed)
- Center and scale numeric predictors
- Apply Yeo-Johnson transformation to numeric predictors
- Convert factor variables to dummy variables

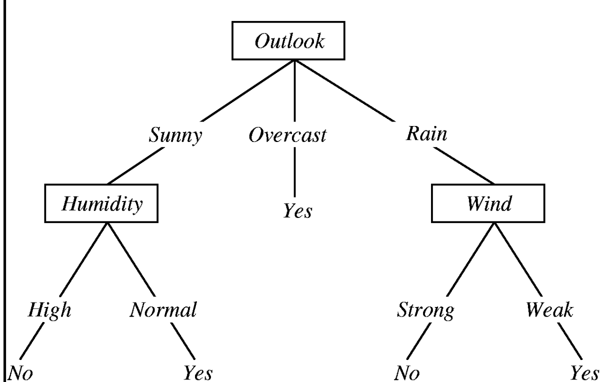
12

THREE DIFFERENT MODELS WERE CREATED

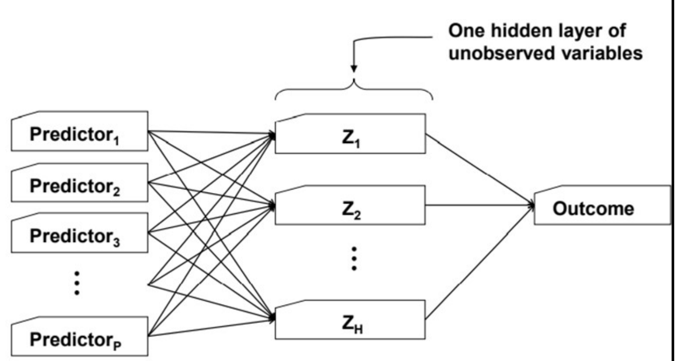
RANDOM FOREST	NEURAL NETWORK	STOCHASTIC GRADIENT BOOSTING
• Decision tree based • Ensemble model (aggregate) • Random sample each time	• Constructed to resemble human brain • Highly complex • Large "network" of decision-making functions	• Decision tree based • Ensemble model (stepwise) • Each tree improves on the last

13

THREE DIFFERENT MODELS WERE CREATED



towardsdatascience.com



14

THREE DIFFERENT MODELS WERE CREATED

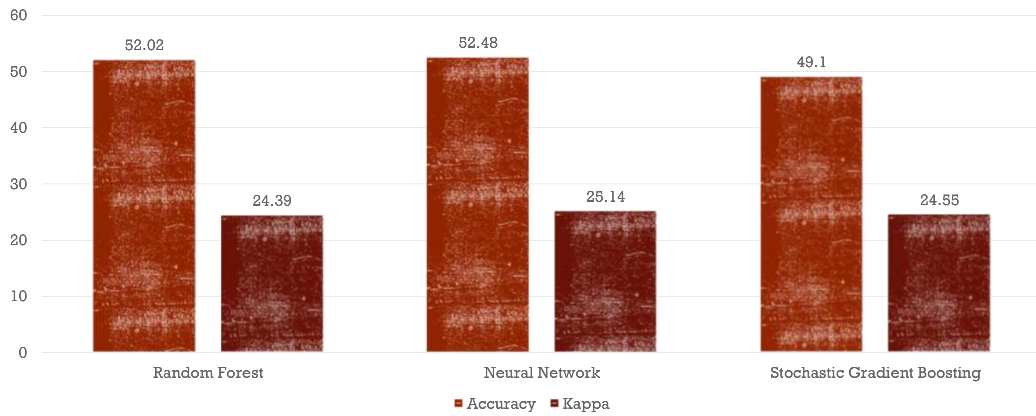
RANDOM FOREST	NEURAL NETWORK	STOCHASTIC GRADIENT BOOSTING
Pros • Less prone to overfitting • Less variance • Handles noisy data well Cons • May not perform as well for regression models • May not learn complexities as well as GBM	Pros • Powerful learner • Flexible Cons • Complexity • Require more data to be effective • Processing power	Pros • Variance similar to random forest • Perform well for unbalanced data Cons • More prone to overfitting than RF • More difficult to tune than RF

15

RESULTS

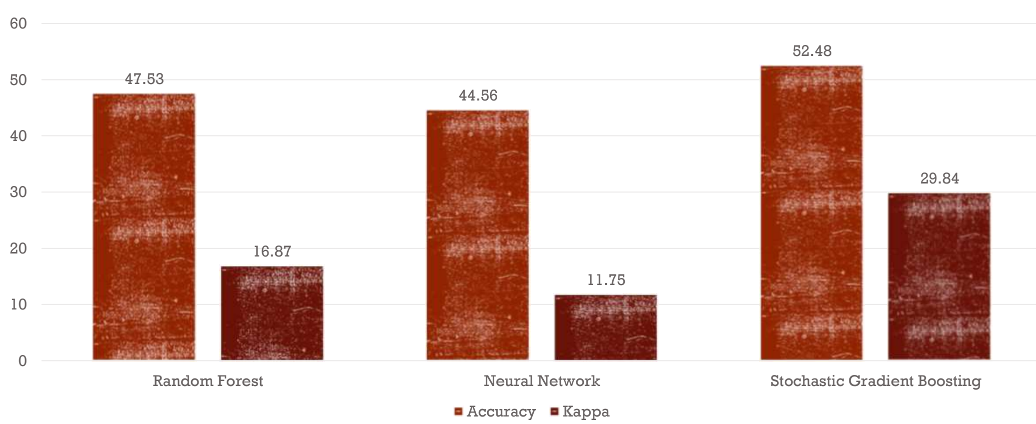
16

MODEL RESULTS – TRAINING DATA



17

MODEL RESULTS – TEST DATA



18

MOST SIGNIFICANT FEATURES

1. BARTHAG: 100.00
2. FTR: 32.00
3. ADJDE: 31.67
4. ADJOE: 28.41
5. WINPERCENT: 24.67

19

USES OF THIS MODEL

20

POTENTIAL MODEL USES

BRACKET CHALLENGE COMPETITIONS

TV NETWORK VIEWERSHIP / REVENUE FORECASTING

FUN

21

CONCLUSION

22

23