Deloitte.

NCAA®

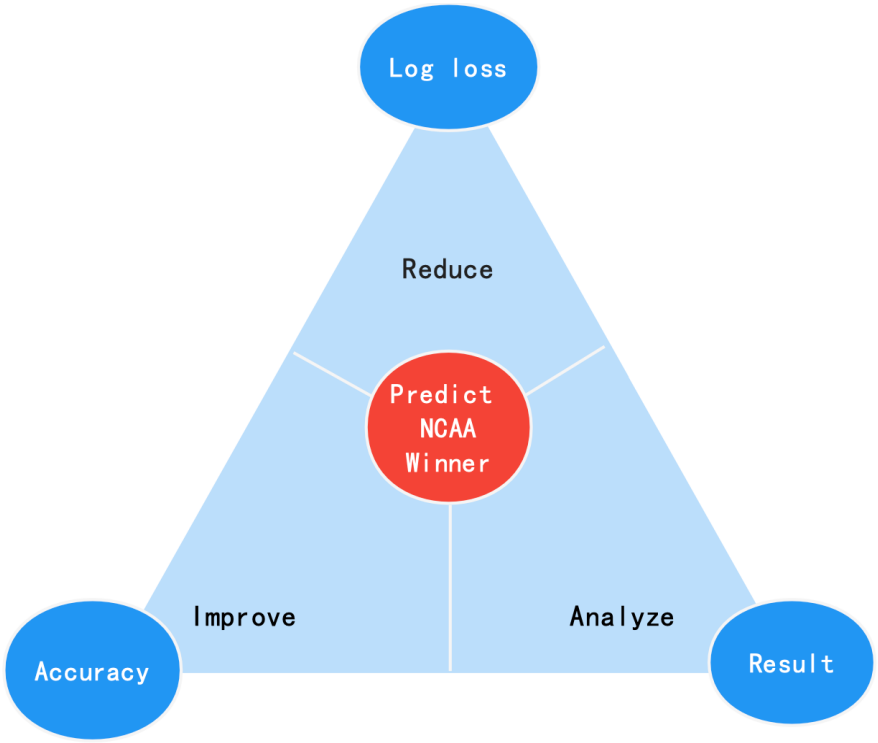
MARCH MADNESS

2022 March Madness: NCAA Men's Tournament Bids Report

Presented by Goal Diggers

-Nancy Wang
-Stephanie Zhao
-Tony Wang
-Vandana Agarwal

Objectives



We predict NCAA winner with following objectives:

1. Reduce log loss
2. Improve accuracy
3. Analyze result

Hypotheses

The team's seed number has a great impact in their probability of winning because a higher seeding means a weaker opponent and a home court advantage

Offensive Rebounds Off, Defense Rebounds Def, Tempo Factors play an equally important role in each game outcomes

We ignore the impact of athlete turnover due to graduation on the level of the team Predictors will remain sustainable in the foreseeable future.

Challenges

Our dataset has 104 features, and determining the impact of these features on the outcome of each match may be a complex process

All the observations share the same result (team1 wins) which will not satisfy the rule of model training part

A single model may lead to some errors in the results due to its limitations so that we need to build an ensemble model to improve the accuracy of our prediction

Methodology

Data Preprocessing

- *Swap/Transfer Data*
- *Add Dependent Variable['result']*

Data Exploration

- *Relevance Detection*

Feature selection

- *Institutional knowledge*
- *Random Forest Algorithm*

Model Building

- *Single Model (Logistic Regression Random Forest Gradient Boosting Deep Neural Network)*
- *Ensemble Model*

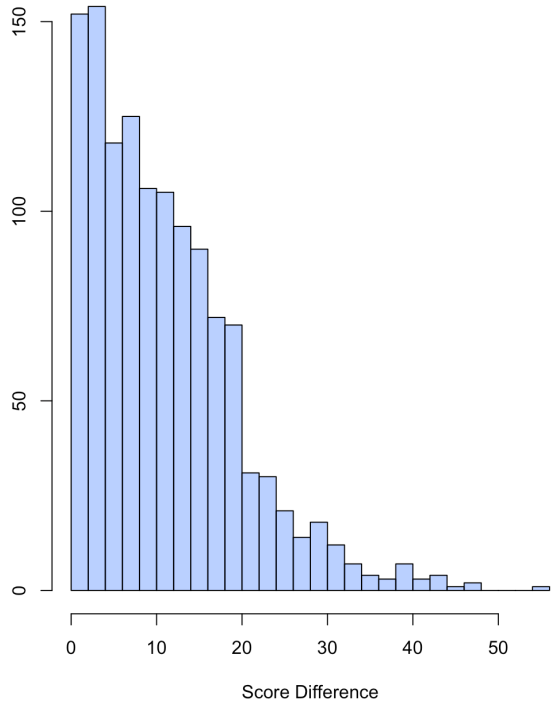
Model Evaluation

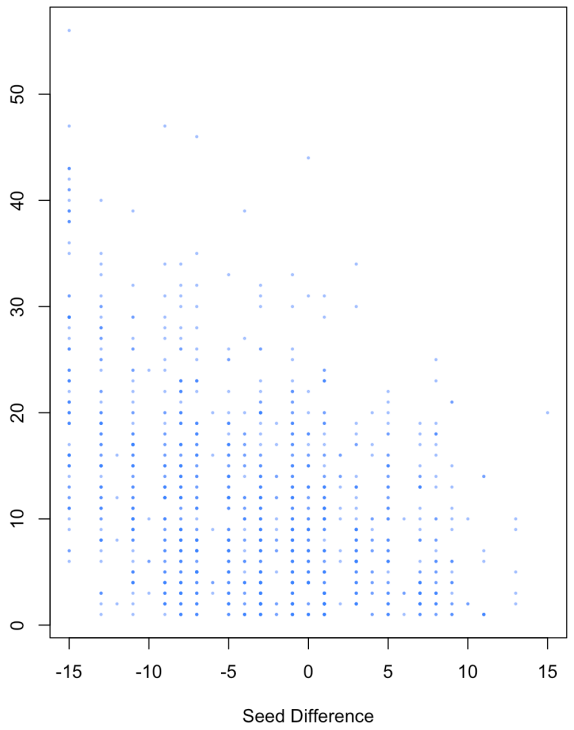
- *Accurancy*
- *Log Loss*

Conclusion & Improvement

Data Exploratory

Score Difference : Distribution and correlation with Seed Difference





Findings:

Score difference shows exponential distribution

The score difference tends to be large when the seed difference is large negative

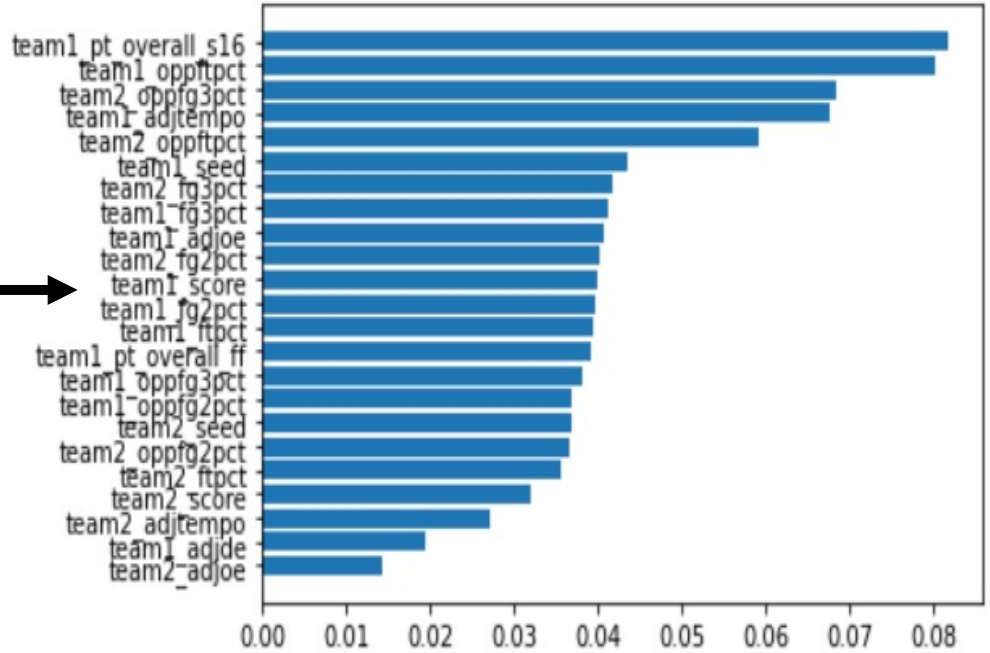
**** Definition:***

Score difference = team1 score – team2 score

Seed difference = team1 seed – team2 seed

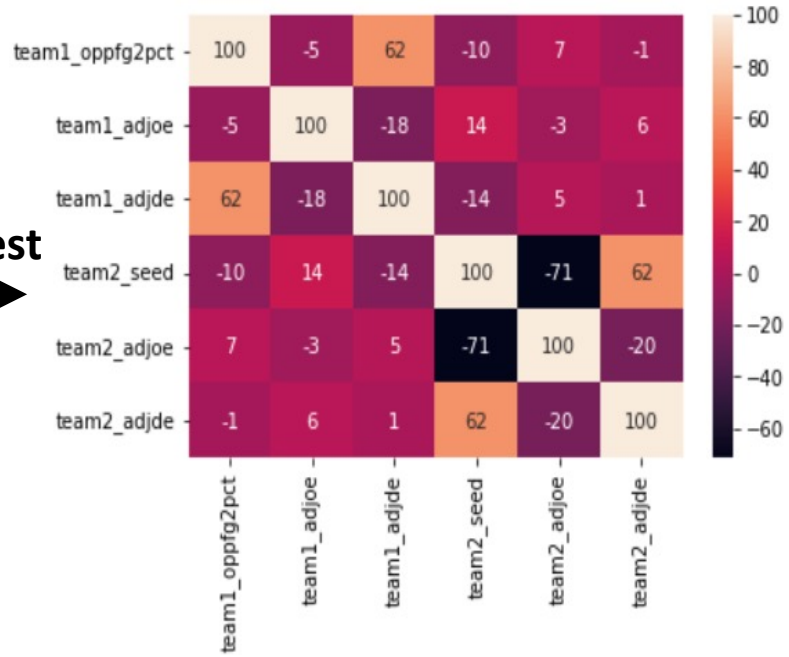
Feature Selection

Institutional Knowledge



Random Forest

cut_off = 0.05



104 Features → **24 Features** →

6 Features

'team1_oppfg2pct', 'team1_adjoe', 'team1_adjde', 'team2_seed', 'team2_adjoe', 'team2_adjde'

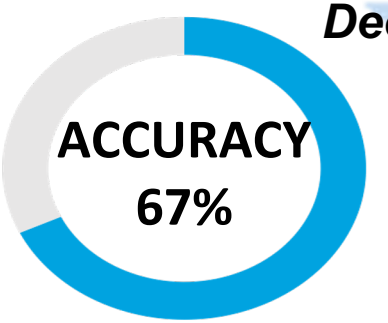
Deloitte.

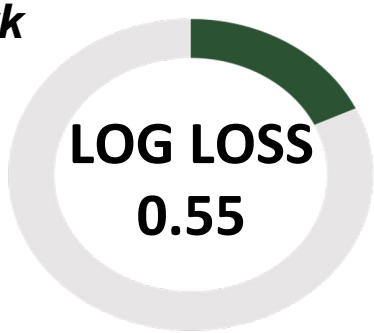
Single Model Comparison

Model	*Accuracy*	*Log Loss*	*Efficiency(high/modest/low)*
Logistic Regression	0.7112	0.5641	HIGH
Deep Neural Network	0.6789	0.5527	HIGH
Random Forest	0.7005	0.6268	MODERATE
Gradient Boosting	0.6658	0.6618	MODERATE
K-Nearest Neighbor	0.7032	0.7316	LOW

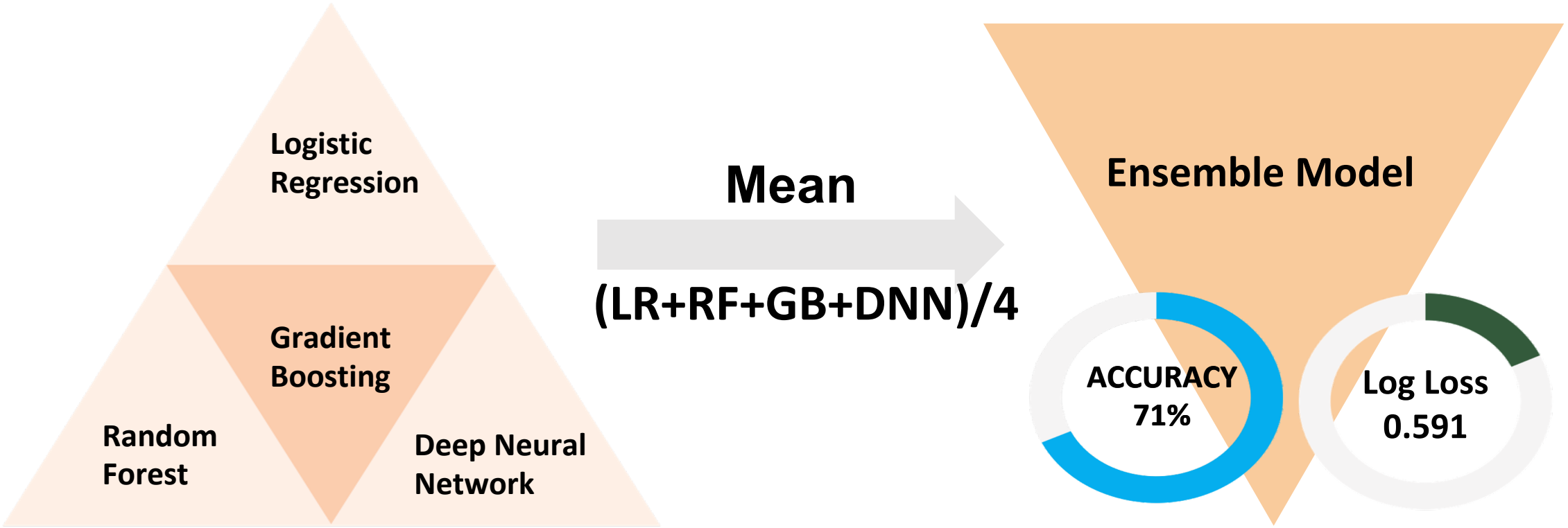
Highest Model Performance

Deep Neural Network





Model Building - Ensemble



Deloitte.

1ST ROUND	2ND ROUND	REGIONAL ROUND	SEMI-FINAL	CHAMPIONSHIP	SEMI-FINAL	REGIONAL ROUND	2ND ROUND	1ST ROUND
1 Gonzaga	Gonzaga	Gonzaga	Gonzaga	CHAMPION Gonzaga	Arizona	Arizona	Arizona	1 Arizona
16 Georgia St.								16 WRIGHT/BRY
8 Boise St.	Boise St						Seton Hall	8 Seton Hall
9 Memphis		Gonzaga						9 TCU
5 UConn	Connecticut					Arizona	Houston	5 Houston
12 New Mexico St.		Connecticut						12 UAB
4 Arkansas	Arkansas					Houston	Illinois	4 Illinois
13 Vermont								13 Chattanooga
6 Alabama	Alabama	WEST	Gonzaga		Arizona	SOUTH	Colorado St	6 Colorado St.
11 RUT/ND		Alabama				Colorado St		11 Michigan
3 Texas Tech	Texas Tech						Tennessee	3 Tennessee
14 Montana St.		Alabama				Loyola-Chicago		14 Longwood
7 Michigan St.	Michigan St						Loyola-Chicago	Ohio St.
10 Davidson		Duke				Loyola-Chicago		10 Loyola Chicago
2 Duke	Duke						Villanova	2 Villanova
15 CS Fullerton								15 Delaware
1 Baylor	Baylor		Gonzaga		Arizona		Kansas	1 Kansas
16 Norfolk St.		Baylor				Kansas		16 TEXSO/TXAMCC
8 North Carolina	North Carolina						San Diego St	8 San Diego St.
9 Marquette		Baylor				Kansas		9 Creighton
5 Saint Mary's	St Mary's CA						Iowa	5 Iowa
12 WYO/IND		St Mary's CA				Iowa		12 Richmond
4 UCLA	UCLA						Providence	4 Providence
13 Akron			Baylor		Kansas			13 South Dakota St.
6 Texas		EAST				MIDWEST	LSU	6 LSU
11 Virginia Tech	Texas							11 Iowa St.
3 Purdue		Texas				Wisconsin	Wisconsin	3 Wisconsin
14 Yale	Purdue							14 Colgate
7 Murray St.		Texas				Auburn	USC	7 USC
10 San Francisco	San Francisco							10 Miami (FL)
2 Kentucky		San Francisco				Auburn	Auburn	2 Auburn
15 Saint Peter's	Kentucky							15 Jacksonville St.

1 Gonzaga Bulldogs

1 Baylor Bears

1 Arizona Wildcats

1 Kansas Jayhawks

- Date by March 23

- Correct Prediction
- Incorrect Prediction
- Predicted Top 8 Teams

Deloitte.

Final 4 Conclusion

Champion Probability Prediction

89.71%

1 Gonzaga Bulldogs

89.89%

1 Arizona Wildcats

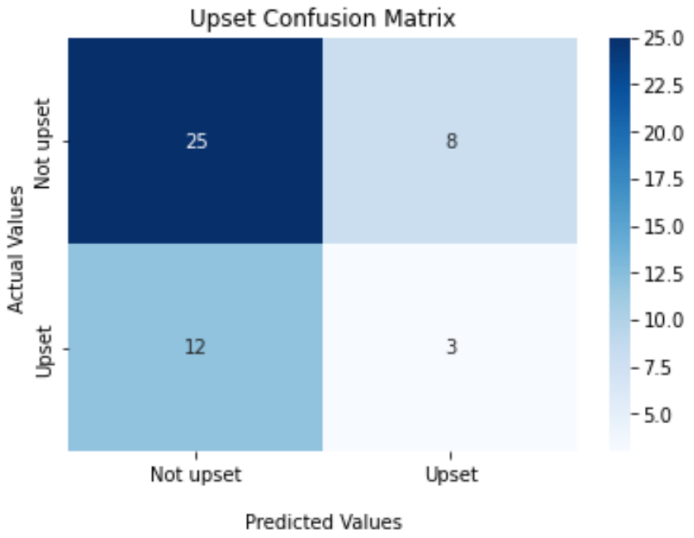
54.10%

1 Baylor Bears

36.06%

1 Kansas Jayhawks

Upset Prediction



Recall	Precision
0.20	0.27

PREDICTED FIRST ROUND UPSETS

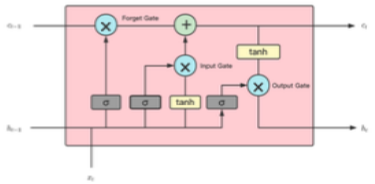
San Francisco
51.63%
10 vs 7

Loyola
53.57%
10 vs 7

Improvement

IMPROVEMENT FLOWCHART

Data	PreProcessing	Dimension reduction	Model
Numerical value Categorical value Text data (news/ social media)	Fill missing value: Hot Deck Imputation [1] Standardize Normalize Remove unicode characters Remove stopwords Stemming/ Lemmatization	Bayesian neural networks [3] Principal Component Analysis [2] Random Forest feature selection Lasso regression	Binary prediction model: 1. Logistic 2. SVM 3. Boosting tree 4. Random forest Classifier 5. Deep neural network 6. K nearest neighborhood LSTM [4]



What we have done

What can be improved

Reference

[1] Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, *78*(1), 40-64.

[2] Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.

[3] Carreira-Perpinán, M. A. (1997). A review of dimension reduction techniques. Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09, 9, 1-69.

[4] Yao, L., & Guan, Y. (2018, December). An improved LSTM structure for natural language processing. In 2018 IEEE International Conference of Safety Produce Informatization (IICSPI) (pp. 565-569). IEEE.

[5] https://en.wikipedia.org/wiki/NCAA_Division_I_Men%27s_Basketball_Tournament