ENSEMBLE MODELING WITH BASKETBALL

JOHNNY THOMAS – M.S. IN STATISTICS AT THE GEORGE WASHINGTON UNIVERSITY


Ursinus
COLLEGE

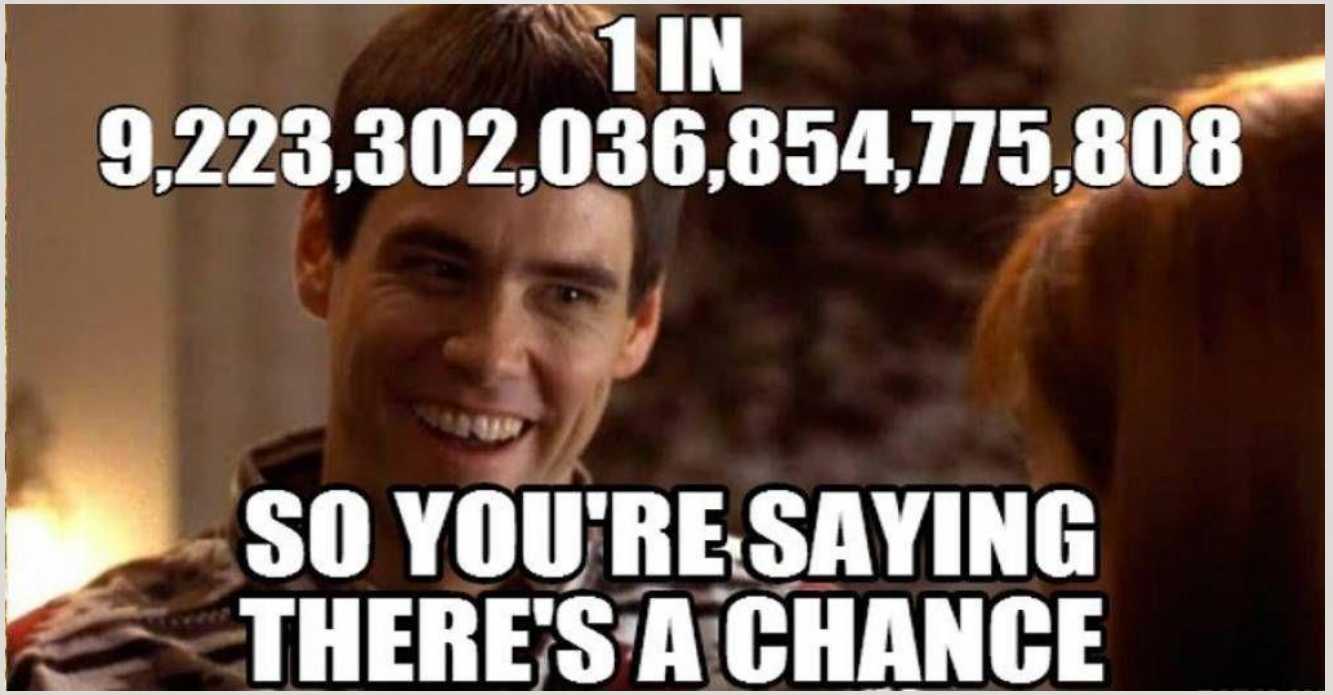

GW


GWU ASA
Student Chapter


76ers

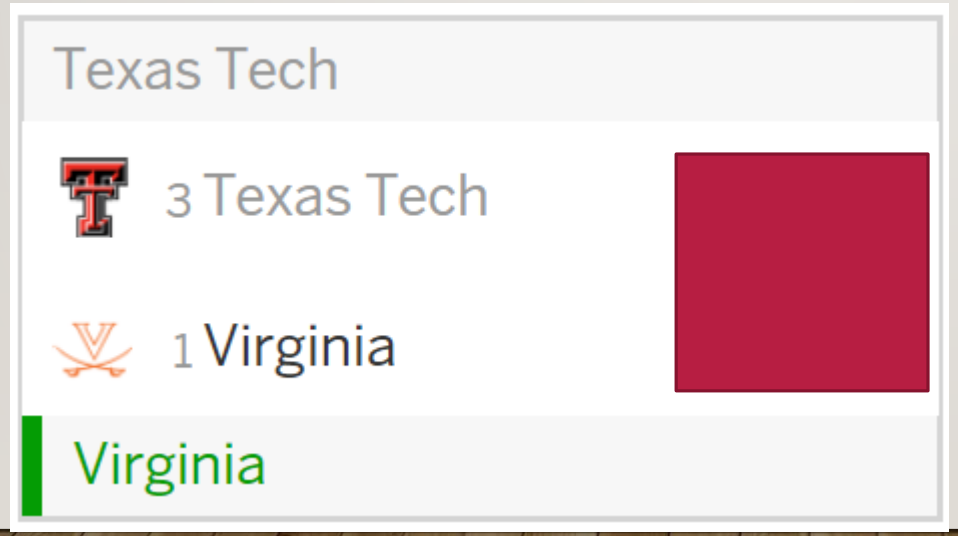

NCAA
MARCH
MADNESS®

UNLIKE THE NBA, NCAA PLAYOFFS ARE HIGHLY UNPREDICTABLE!



ASKING THE QUESTION

- **"Which variables contribute the most to a win in the NCAA Men's Tournament?"**
- *Prediction Problem turned Classification Question*



UVA WINS!

THE DATASET

- Kaggle
  - Google Cloud & NCAA® ML Competition 2019-Men's Challenge
- Regular Season & Playoff Games 2003-2019
  - Six datasets
- 166,178 observations
- 26 Variables

VARIOUS MODELS

Model	AUC Score
Logistic Regression	0.9425262
XGBoost 2 (Parametrized)	0.939129
XGBoost 3 (Parametrized)	0.9348964
LDA	0.9328915
XGBoost	0.9316106
QDA	0.8053575
Decision Trees	0.7201493
SVM	----------

WHAT IS ENSEMBLE MODELING?

SCHWINN

Trek
Bicycle Store

MONGOOSE
BICYCLES

Retrospec

SCHWINN

Trek
Bicycle Store

MONGOOSE
BICYCLES

Retrospec

Interview 1

Interview 2

Interview 3

NO HIRE

HIRE

HIRE

OFFER

NO OFFER

WHAT IS ENSEMBLE MODELING?

- COLLECTIVE DECISION MAKING REDUCES BIAS.

WHAT IS ENSEMBLE MODELING?

1. Create a number of models using various methods
2. Combine predictions of each model on the training set into one new dataframe
3. Try other methods in prediction/classification of a variable
4. Best AUC is your best ensemble model

VARIOUS MODELS

Model	AUC Score
Logistic Regression	0.9401658
XGBoost 2 (Parametrized)	0.939129
XGBoost 3 (Parametrized)	0.9348964
LDA	0.9328915
XGBoost	0.9316106
QDA	0.8053575
Decision Trees	0.7201493
SVM	----------

ENSEMBLING THE MODELS TOGETHER

Actual Results *Predicted Results*

Training Data Obs.	Result		Logistic Regression	XGBoost 2	XGBoost 3	LDA	XGBoost
1	WIN		WIN	WIN	LOSS	WIN	LOSS
2	WIN		WIN	WIN	WIN	WIN	WIN
3	LOSS		LOSS	LOSS	WIN	WIN	LOSS
4	WIN		LOSS	LOSS	WIN	LOSS	LOSS
5	LOSS		LOSS	WIN	LOSS	WIN	WIN
6	LOSS		LOSS	LOSS	LOSS	LOSS	WIN

ENSEMBLING THE MODELS TOGETHER

Actual Results | *Predicted Results*

Training Data Obs.	Result	Ensemble Model	Logistic Regression	XGBoost 2	XGBoost 3	LDA	XGBoost
1	WIN	WIN	WIN	WIN	LOSS	WIN	LOSS
2	WIN	WIN	WIN	WIN	WIN	WIN	WIN
3	LOSS	LOSS	LOSS	LOSS	WIN	WIN	LOSS
4	WIN	WIN	LOSS	LOSS	WIN	LOSS	LOSS
5	LOSS	LOSS	LOSS	WIN	LOSS	WIN	WIN
6	LOSS	LOSS	LOSS	LOSS	LOSS	LOSS	WIN

ENSEMBLING THE MODELS TOGETHER

Model	AUC Score
XGBoost (Parameterized)	0.9425262
Random Forest	0.9333927
Logistic Regression	0.9326131
XGBoost	0.9258744

ENSEMBLING THE MODELS TOGETHER

Original Models	AUC Score
Logistic Regression	0.9401658
XGBoost 2 (Parametrized)	0.939129
XGBoost 3 (Parametrized)	0.9348964
LDA	0.9328915
XGBoost	0.9316106
QDA	0.8053575
Decision Trees	0.7201493
SVM	----------

Ensembled Models	AUC Score
XGBoost (Parameterized)	0.9485262
Random Forest	0.9333927
Logistic Regression	0.9326131
XGBoost	0.9258744

0.0083604 increase!

PROS AND CONS

- **Pros**
  - Improves accuracy of model
    - Less bias, more robust!
  - Will almost always win you coding competitions
  - Capture linear and simple as well non-linear complex relationships in the data.
- **Cons**
  - Time consuming
  - Interpretability can be challenging
  - Can be difficult to choose the right ensemble method

RESULTS

- FTA (Free Throws Attempted)
- DR (Defensive Rebounds)
- Blk (Blocks)
- FGM3 (3-Point Field Goals Made)
- Stl (Steals)
- FGA (Field Goals Attempted)
- Ast (Assists)

John Thomas
Residence Director at The George Washington University

THE END

ACKNOWLEDGEMENTS

- https://www.analyticsvidhya.com/blog/2017/02/introduction-to-ensembling-along-with-implementation-in-r/
- https://www.datasciencecentral.com/profiles/blogs/10-machine-learning-methods-that-every-data-scientist-should-know?utm_source=dlvr.it&utm_medium=linkedin
- Lecture Notes from Dr. Emre Barut, The George Washington University
- Research done by John Thomas, Sam Luxenburg, & Ning Xie