Predicting the Madness

Tom Davich
General Assembly
Data Science 20

What is March Madness?

- NCAA Division I Championship Basketball Tournament
- Single Elimination
- 68 Teams (currently)
- 5 Rounds with 1 round of 4 “play-in” games
- The probability of picking a perfect bracket….

1 in 128 **Billion**

(Billion, with a **B**)

The Challenge: Kaggle March Madness Mania

- Goal: Predict the **probabilities** for each possible outcome
- Evaluation: Log Loss of predicted probability vs outcome
  - High penalty for confidently predicting the wrong outcome
- Kaggle competition closed for submissions to evaluate probabilities...

- Next best option: **predicting Wins / Losses of tournament games**
  - **Train: 2012-2014**
  - **Test: 2015**

kaggle™

Data Files: Complete Historical Data for 30 years

- Season
  - Year, Day, and Tournament Region information (East, West, etc)
- **Regular** (& Tourney) Season Compact Results:
  - Game / Teams / Score / Location
- **Regular** (& Tourney) Season Detailed Results:
  - Shots attempted / made
  - Blocks, Steals, Fouls, Assists
- **Tourney Seeds**
  - Seeding for each team
- Tourney Slots
  - Compares paired teams as Stronger or Weaker than expected

No missing values. **Bold = used in model**

The plan: (Data) Science!

1. Feature Selection
2. Data Transformation
3. Models!
    a. Decision Trees
    b. Random Forest
    c. Boosting
4. Profit

Feature Selection: **To Infinity... and Beyond!**

1. Season Averages per team:
    a. Points Score, Points Allowed, Shots, Blocks, Rebounds, Assists, Fouls, etc
2. Wins / Losses
    a. Last 6 games of Regular Season
    b. Against Tournament teams
    c. Margin < 2 (close games)
    d. Margin > 7 (blow outs)
3. Away game winning percentage
4. Tournament Seed

Total of **26 features per team**.

Data Transformation: building the Training Set

1. Create 26 feature Data Frame for Team A
2. And then for Team B
3. Game results
4. Join together!

1

TEAMID	A_TWPCT	A_WST6	A_SEED
511	0.8	5	3
515	0.6	2	12
519	0.63	5	14
527	0.643	2	11
539	0.63	4	12
581	0.815	5	9

2

TEAMID	B_TWPCT	B_WST6	B_SEED
511	0.8	5	3
515	0.6	2	12
519	0.63	5	14
527	0.643	2	11
539	0.63	4	12
581	0.815	5	9

3

Matchup	Win
A_515_729	1
A_555_559	0
A_576_666	1
A_577_581	0
A_551_604	0
A_539_629	0

4

Matchup	Win	A_ID	B_ID	A_TWPCT	A_WST6	A_SEED	B_TWPCT	B_WST6	B_SEED
A_515_729	1	515	729	0.6	2	12	0.778	3	5
A_555_559	0	555	559	0.517	5	16	0.938	6	1
A_576_666	1	576	666	0.893	6	12	0.759	4	5
A_577_581	0	577	581	0.6	4	8	0.815	5	9
A_551_604	0	551	604	0.643	3	9	0.679	5	8
A_539_629	0	539	629	0.63	4	12	0.733	5	5

Process: serious data wrangling in Pandas

- Iterating with nested **for loops**
  - for each game, for each team, for each season
- .loc, .groupby, .agg, .index, .isin
- 4 functions to create DF for each feature “type”
- **Unexpected**: Kentucky’s perfect Regular Season!
  - 2014-2015: Can’t divide by 0...

Credit: Statsguys

stack overflow

Models and Results

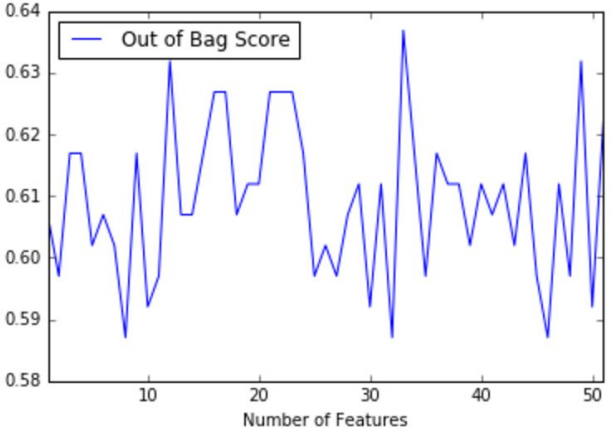
	Model	Decision Tree Classifier	Random Forest Classifier
Initial	OOB		0.601
	CV	0.575	0.621
	Test	0.597	0.776
Tuned	OOB	0.595	0.602
	CV		0.62
	Test	0.686	**0.791**
	Parameters	Depth: 2; Leaf: 3	Features: 7

Most important features:

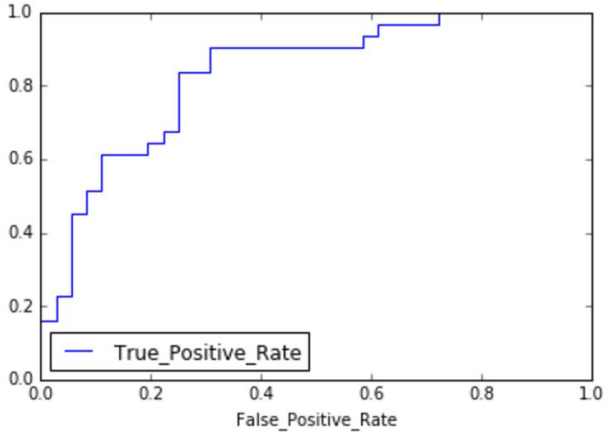
1. Winning % against Tournament Opponents: 0.08
2. Seed: 0.08
3. Avg. Margin of Victory: 0.04

My thoughts:

- Surprised by the variability in OOB tuning
- 2015 (test) scored much higher than CV (2012-2014)
- Would have tried boosting with more time



RF: OOB Tuning (2012-2014)



AUC: Test - 2015 Results

Learnings & Future Plans

Learnings:

1. Data wrangling is time consuming
2. Predicting outcomes is much harder than probabilities
3. 2015 seemed more easily predictable...

Future Plans:

1. Submit to 2017 Competition
    a. Score probabilities instead of outcomes
2. Benchmark teams vs an average opponent
3. Get Vegas Odds for first round games
4. Investigate Bradley-Terry model

Wisconsin ending Kentucky's perfect season (2015)

Fin!

Tom Davich
tdavich@gmail.com
LinkedIn
Github