Forecasting Outcomes of 2018 Women's March Madness

W207 Final Project

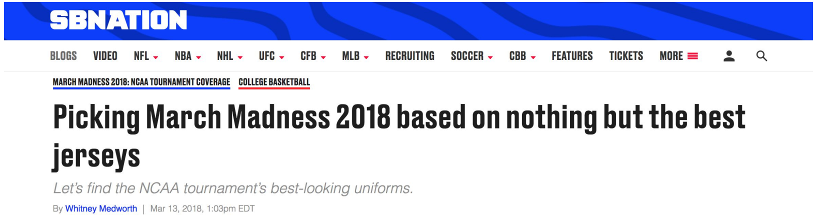
Julia Buffinton, Charlene Chen, Arvindh Ganesan, Prashant Kumar Sahay

Content

- Goals and Impact
- Data Understanding and Preparation
- Feature Engineering and Selection
- Modeling
- Evaluation and Deployment

What are we trying to achieve? Why do we care?

- Goal: predict game outcomes (win/loss) for 2018 “March Madness” tournament play
  - Use historical data and machine learning approach to make informed predictions
- No one has ever made a perfect bracket*
  - Record for streak for correct picks in men’s is 39 games (2017)
  - No verified brackets that have been perfect in the Sweet 16
- We will predict pairwise matchups
  - All possible games in 2018 tourney

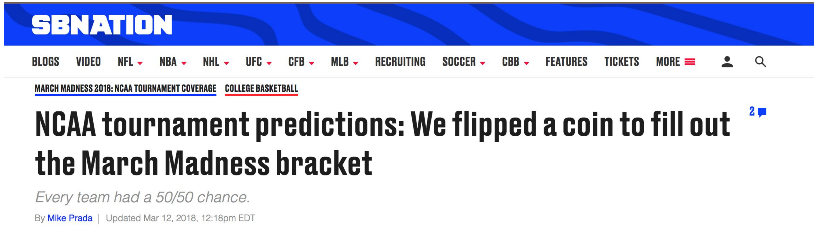

SBNATION
BLOGS VIDEO NFL NBA NHL UFC CFB MLB RECRUITING SOCCER CBB FEATURES TICKETS MORE

MARCH MADNESS 2018: NCAA TOURNAMENT COVERAGE COLLEGE BASKETBALL

Picking March Madness 2018 based on nothing but the best jerseys

Let’s find the NCAA tournament’s best-looking uniforms.

By Whitney Medworth | Mar 13, 2018, 1:03pm EDT


SBNATION
BLOGS VIDEO NFL NBA NHL UFC CFB MLB RECRUITING SOCCER CBB FEATURES TICKETS MORE
2

MARCH MADNESS 2018: NCAA TOURNAMENT COVERAGE COLLEGE BASKETBALL

NCAA tournament predictions: We flipped a coin to fill out the March Madness bracket

Every team had a 50/50 chance.

By Mike Prada | Updated Mar 12, 2018, 12:18pm EDT

*https://www.ncaa.com/news/basketball-men/bracket-beat/2017-03-14/march-madness-longest-perfect-bracket-streak-we-know

Data Understanding and Preparation

- Data used from Kaggle competition includes:
  - City locations, 1998-2017 tournament results, 1998-2018 regular season results, 1998-2018 tournament seeds
- Additional data used:
  - Brought in our own league information from NCAA
- Focus on detailed results from regular season/tourney - each unit is a matchup
  - Contain game-level statistics from game played; merge with team-level info (seeds, league competitiveness)
- For prediction:
  - Most of the fields will be generated using feature engineering based on input above
  - Can't use game-level stats because games haven't been played, so use historical data and other metrics to represent relative strength of team to predict outcome

Feature Engineering and Selection

- Win Probability
  - Engineered feature from regular season
  - Reflects a team's annual performance relative to its league
    - Field goals attempted / made
    - Free throws attempted / made
    - Blocks
    - Rebounds
    - Assists
    - Steals
    - 2 point goals
    - 3 point goals
    - Point Opportunities Developed
    - Opportunity Conversion Rate

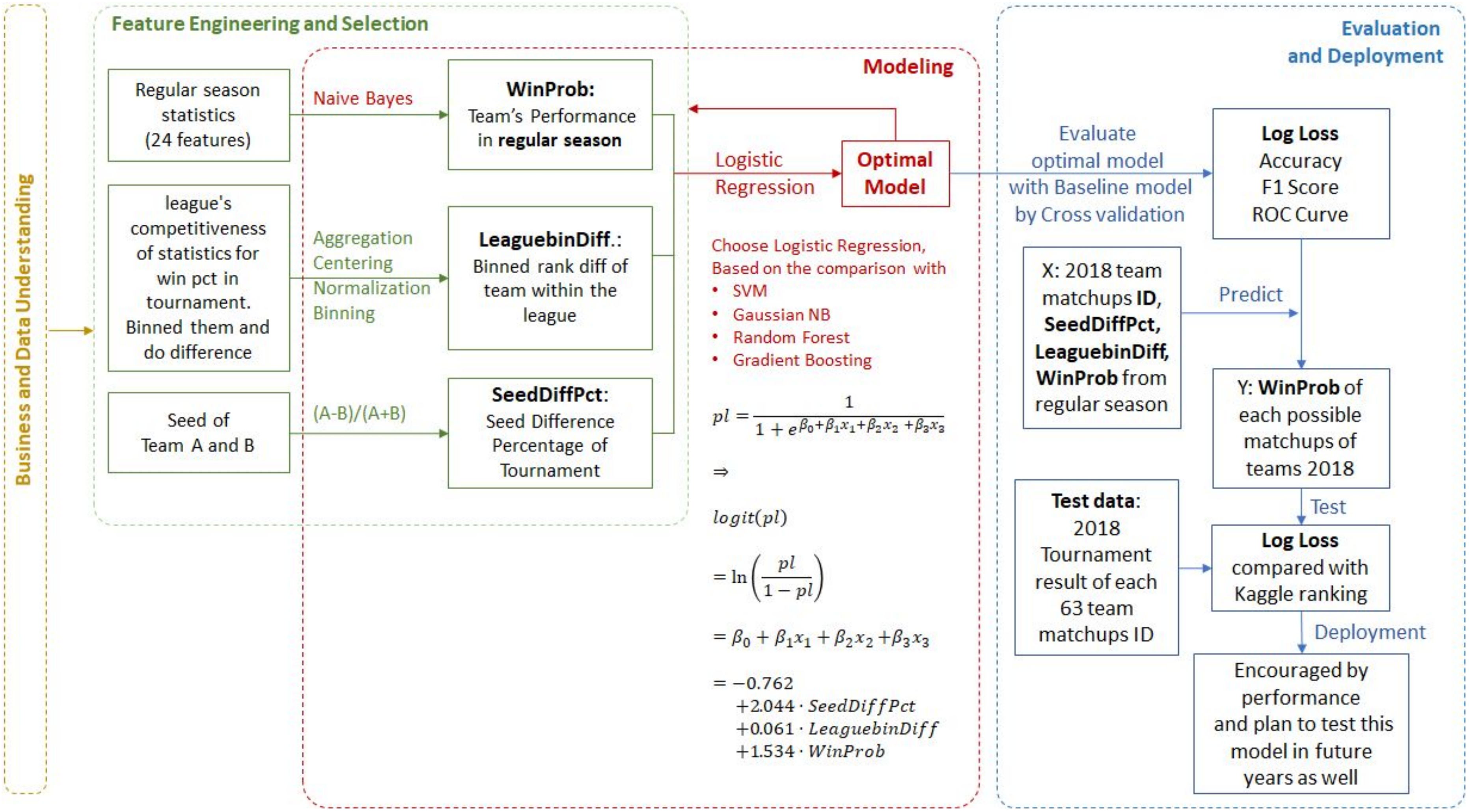
- League Bin Difference
  - Difference in strength of leagues that each team belongs to
  - Reflects the competitiveness of the opponents in games that generated win prob

- Seed Difference Percentage
  - Difference in seed between teams
  - Relative to the values of each seed
    - 1 vs. 2 = (2-1) / (1+2) = 1/3
    - 15 vs. 16 = (16-15) / (15+16) = 1/31

Modeling - Two-Stage

- Generate ‘win probabilities’
  - Use regular season statistics for each team in each season to predict game outcomes
- Gaussian Naive Bayes
  - Use the predicted probabilities as ‘win probability’ features

- Predict Tournament Outcomes using Logistic Regression
  - Win probability
  - League bin difference
  - Seed difference percentage


Business and Data Understanding
Feature Engineering and Selection
Regular season statistics (24 features)
league's competitiveness of statistics for win pct in tournament. Binned them and do difference
Seed of Team A and B
Naive Bayes
Aggregation
Centering
Normalization
Binning
(A-B)/(A+B)
WinProb: Team's Performance in regular season
LeaguebinDiff.: Binned rank diff of team within the league
SeedDiffPct: Seed Difference Percentage of Tournament
Modeling
Logistic Regression
Optimal Model
Choose Logistic Regression, Based on the comparison with
• SVM
• Gaussian NB
• Random Forest
• Gradient Boosting
$pl = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}$
$\Rightarrow$
$logit(pl)$
$= \ln\left(\frac{pl}{1 - pl}\right)$
$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
$= -0.762$
$+2.044 \cdot SeedDiffPct$
$+0.061 \cdot LeaguebinDiff$
$+1.534 \cdot WinProb$
Evaluation and Deployment
Evaluate optimal model with Baseline model by Cross validation
Log Loss
Accuracy
F1 Score
ROC Curve
X: 2018 team matchups ID, SeedDiffPct, LeaguebinDiff, WinProb from regular season
Predict
Y: WinProb of each possible matchups of teams 2018
Test
Test data: 2018 Tournament result of each 63 team matchups ID
Log Loss compared with Kaggle ranking
Deployment
Encouraged by performance and plan to test this model in future years as well

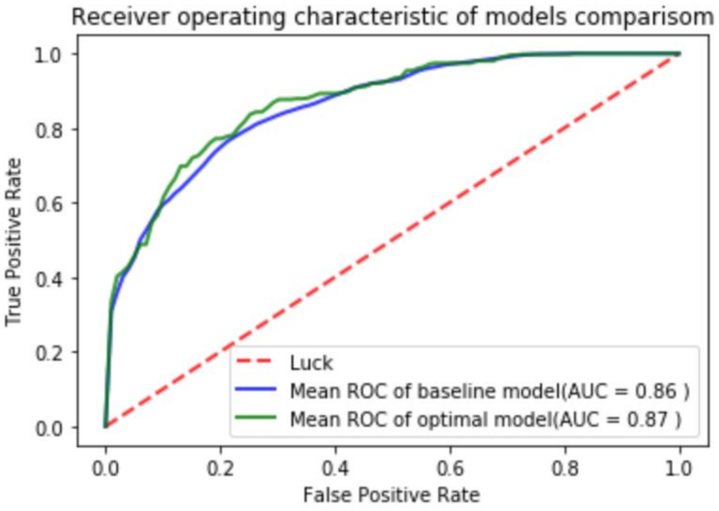
Evaluation

Evaluation with Cross validation:

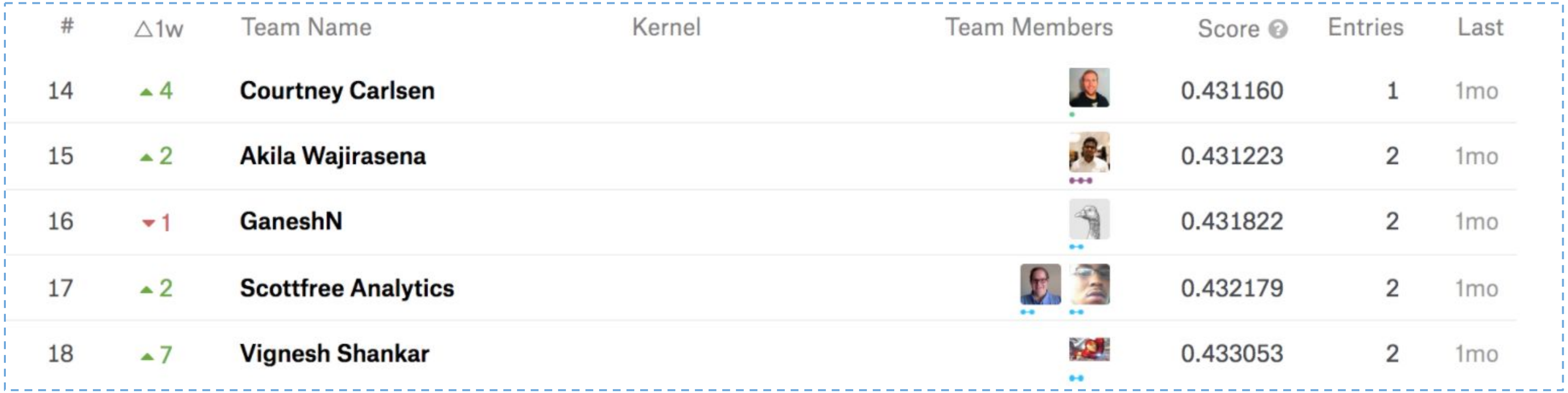
- Optimal model v.s. Baseline model

Metrics	Optimal model	Baseline model
Log Loss	0.442	0.457
Accuracy	0.782	0.781
F1 measure	0.782	0.781

- ROC Curve



Deployment



#	△1w	Team Name	Kernel	Team Members	Score	Entries	Last
14	▲4	**Courtney Carlsen**			0.431160	1	1mo
15	▲2	**Akila Wajirasena**			0.431223	2	1mo
16	▼1	**GaneshN**			0.431822	2	1mo
17	▲2	**Scottfree Analytics**			0.432179	2	1mo
18	▲7	**Vignesh Shankar**			0.433053	2	1mo

- **Prediction:**
  - Win likelihood of every potential matchup in the 2018 NCAA Division I Women's Basketball Tournament
  - Competitively against other teams in the Kaggle competition with 0.43159 Log Loss score on test data (#16)
- **Extension and Support:**
  - Encouraged by our performance and plan to test this model in future years as well

https://www.kaggle.com/c/womens-machine-learning-competition-2018/leaderboard

Thank you!

Julia, Prashant, Arvindh, Charlene