

Predicting the Outcomes of March Madness Games

By Aviva Mazurek

Objectives and Methodology

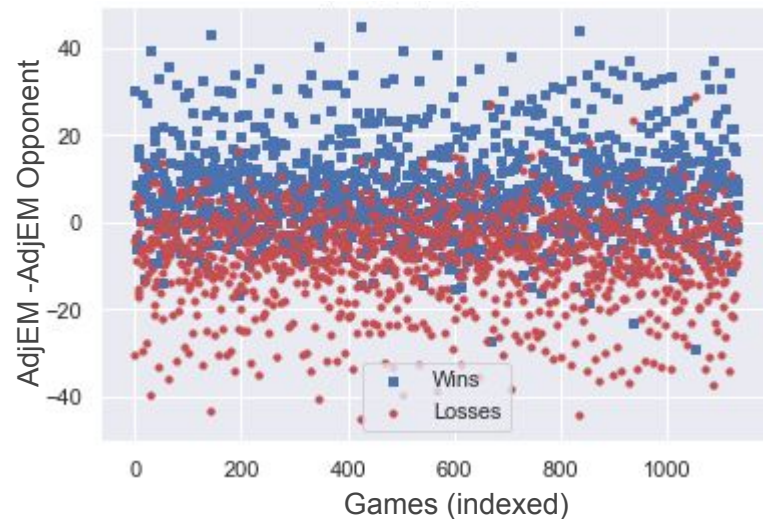
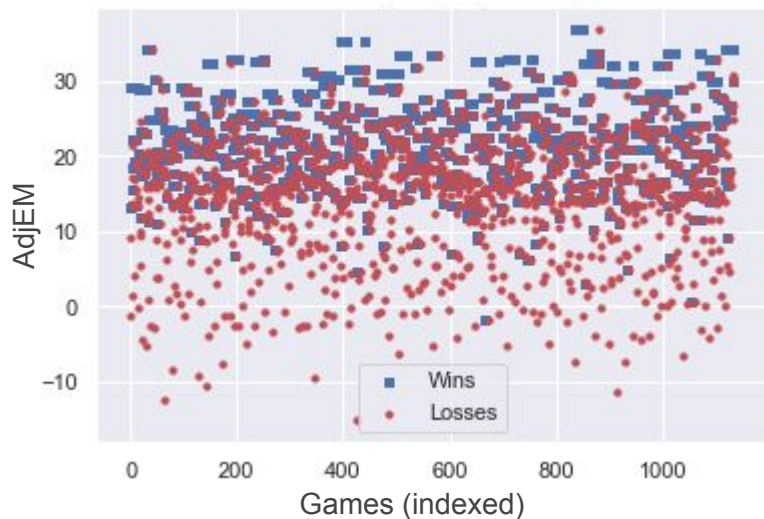
Objective

- Predict individual outcomes of march madness games
 - Final SVM model predicts game outcomes with 86% accuracy

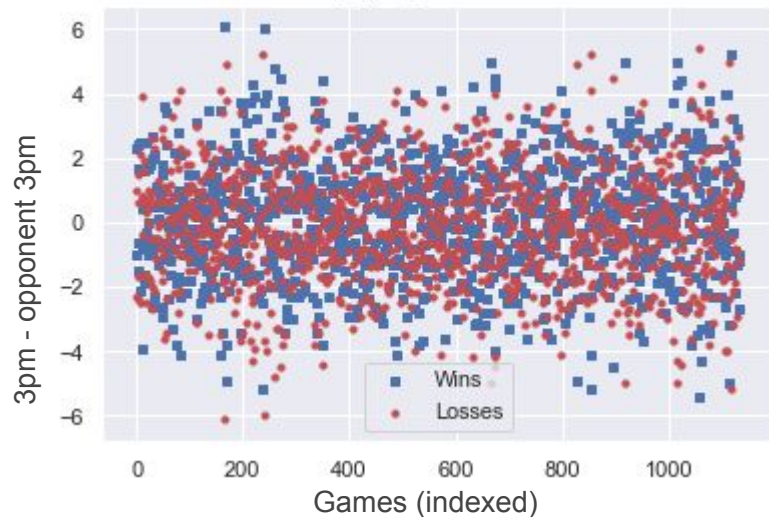
Methodology

- Data sources
 - Sports Reference - historical brackets, seeding, advanced stats from regular season
 - ESPN - Stats from regular season
 - Kenpom - Ranking System, wins/losses per team in regular season
- Created multiple dictionaries to merge all data from 2002-2019

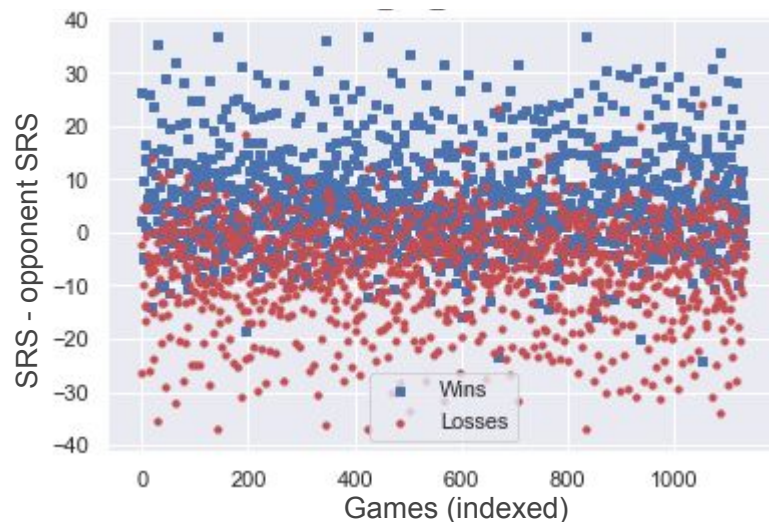
Feature engineering creates larger disparity between wins and losses - allows model to learn better



Which features are more significant?



vs.



→ Most significant features: Kenpom ratings, wins from regular season, losses from regular season, SOS adv, SRS adv, seeding

Modeling Data - Trained and Tested on Data From 2002-2019

Model	Accuracy (%)
Decision Tree	75.62
KNN	69.68
XGBoost	85.14
Logistic Regression	84.43
Support Vector Machine (SVM)	86.20

**The accuracy is an average of 10 random train test split accuracies (except for the decision tree)

Experiment: Used SVM model to predict individual game outcomes for 2019 brackets

- Trained and tested data on all data from 2002-2018 (excluded 2019 data)
- Used the model to predict 2019 march madness results

Results 

Eastern Region

1 Duke 85	✓	1 Duke 77		1 Duke 75		1 Duke 67
16 North Dakota State 62	✓		at Columbia, SC			
8 VCU 58	✓					
9 UCF 73	✓	9 UCF 76	✓			
5 Mississippi State 76	✗			✓	at Washington, DC	
12 Liberty 80	✗	12 Liberty 58				
4 Virginia Tech 66	✓		at San Jose, CA	4 Virginia Tech 73		
13 Saint Louis 52	✓	4 Virginia Tech 67	✓			
6 Maryland 79	✓					
11 Belmont 77	✓	6 Maryland 67				2 Michigan State
3 LSU 79	✓		at Jacksonville, FL	3 LSU 63		
14 Yale 74	✓	3 LSU 69	✓			
7 Louisville 76	✓					
10 Minnesota 86	✓	10 Minnesota 50		✓	at Washington, DC	2 Michigan State 68
2 Michigan State 76	✓					
15 Bradley 65	✓	2 Michigan State 70	✓			

Western Region

1 Gonzaga 87		1 Gonzaga 83		1 Gonzaga 72		1 Gonzaga 69
16 Fairleigh Dickinson 49	at Salt Lake City, UT		at Salt Lake City, UT		at Anaheim, CA	
8 Syracuse 69		9 Baylor 71				
9 Baylor 78	at Salt Lake City, UT					
5 Marquette 64		12 Murray State 62				
12 Murray State 83	at Hartford, CT		at Hartford, CT	4 Florida State 58		
4 Florida State 76		4 Florida State 90				
13 Vermont 69	at Hartford, CT					
6 Buffalo 91		6 Buffalo 58				
11 Arizona State 74	at Tulsa, OK		at Tulsa, OK	3 Texas Tech 63		
3 Texas Tech 72		3 Texas Tech 78				
14 Northern Kentucky 57	at Tulsa, OK					
7 Nevada 61		10 Florida 49				
10 Florida 70	at Des Moines, IA		at Des Moines, IA	2 Michigan 44		
2 Michigan 74		2 Michigan 64				
15 Montana 55	at Des Moines, IA					3 Texas Tech 75

Midwest Region

1 UNC 88		1 UNC 81		1 UNC 80	
16 Iona 73	✓ at Columbus, OH		at Columbus, OH		
8 Utah State 61	✓ at Columbus, OH	9 Washington 59	✓		
9 Washington 78					
5 Auburn 78	✓ at Salt Lake City, UT	5 Auburn 89		✓ at Kansas City, MO	5 Auburn 77
12 New Mexico State 77	✓				
4 Kansas 87	✓ at Salt Lake City, UT	4 Kansas 75	✓	5 Auburn 97	
13 Northeastern 53	✓				
6 Iowa State 59					
11 Ohio State 62	✗ at Tulsa, OK	11 Ohio State 59			
3 Houston 84	✓ at Tulsa, OK	3 Houston 74	✓	3 Houston 58	
14 Georgia State 55	✓				
7 Wofford 84	✓ at Jacksonville, FL	7 Wofford 56		✓ at Kansas City, MO	2 Kentucky 71
10 Seton Hall 68	✓				
2 Kentucky 79	✓ at Jacksonville, FL	2 Kentucky 62	✓		
15 Abilene Christian 44	✓				

Southern Region

[illegible]

Final 4

1 Virginia		63			
5 Auburn	✓	62	1 Virginia	85	
3 Texas Tech		61			
2 Michigan State	✗	51	3 Texas Tech	77	1 Virginia

Conclusions

- Can predict individual march madness games with 86% accuracy
 - Feature engineering and inclusion of Kenpom stats increased accuracy tremendously

Future Work

- T-Pot pipeline determined logistic regression coupled with SGDC Classifier produces 90% accuracy
 - Model difficult to interpret - requires investigation
- Incorporate point outcomes of games
- Conduct same experiment as for 2019, but for every year and determine where the model fails the most
- Incorporate more features/engineering