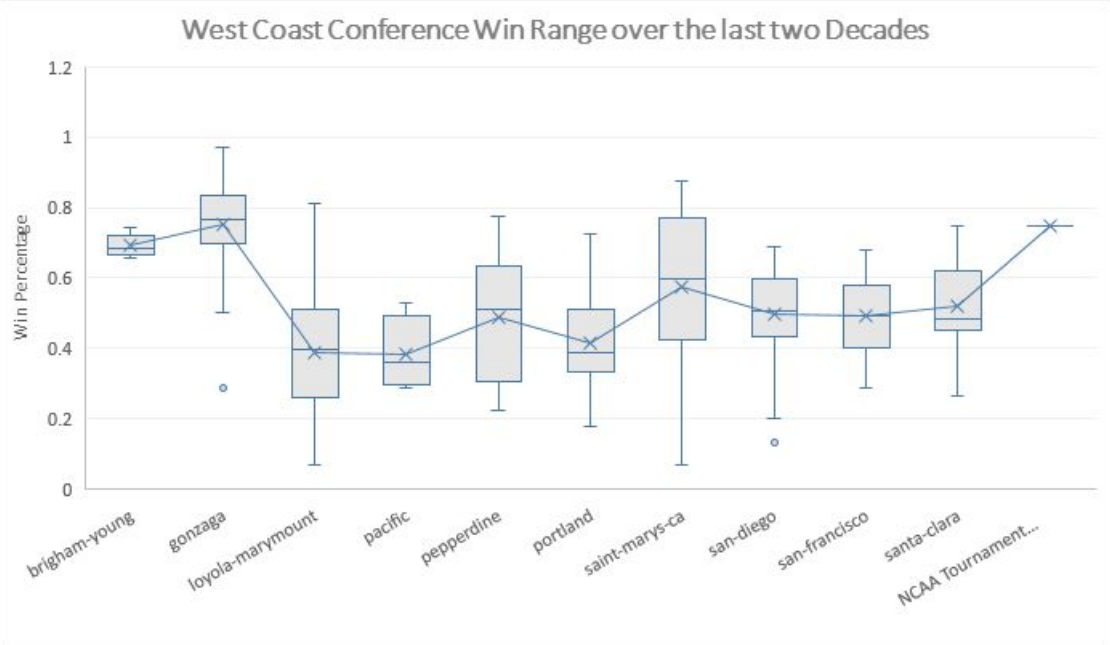
How To End a 20-Year March Madness Drought

Data Analysis of NCAA Men's Basketball and Santa Clara University Hoops

FNCE 3490: Machine Learning

Carson Badger & Kyle Vandenberg

The Problem: SCU Men's Basketball team has not made the annual NCAA Tournament since 1996



Our Approach:

1. **Data Exploration:** How to navigate the data?
2. **Data Analysis:** Can machine learning help determine winning efforts?
3. **Data Engineering:** Are advanced statistics better predictors of success?
4. **Business Applications:** How can our analysis improve SCU basketball?

Data Exploration

First, we needed to obtain the right data

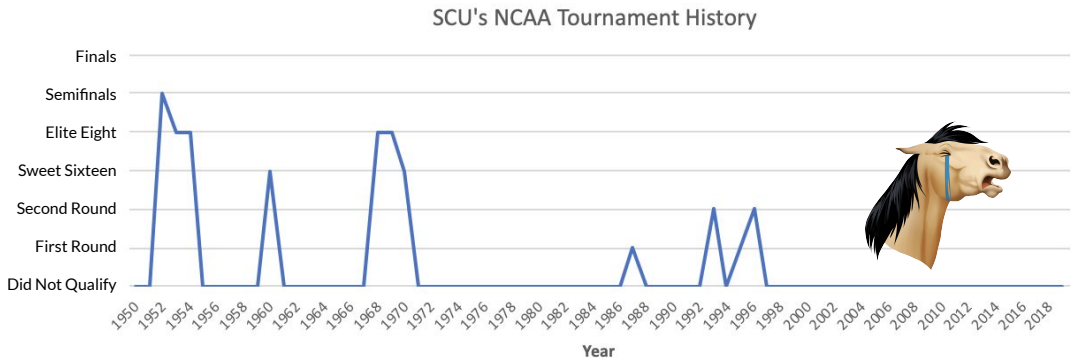
- We tried looking at game-level data, before pivoting to season-long data
- Employed web scraping on a sports stats site

School Stats NCAA = Tournament Appearance Share & more ▾ Glossary

		Overall						Conf.		Home		Away		Points	
Rk	School	G	W	L	W-L%	SRS	SOS	W	L	W	L	W	L	Tm.	Opp.
1	Abilene Christian NCAA	34	27	7	.794	-1.91	-7.34	14	4	13	2	10	4	2502	2161
2	Air Force	32	14	18	.438	-4.28	0.24	8	10	9	6	3	9	2179	2294
3	Akron	33	17	16	.515	4.86	1.09	8	10	14	3	1	10	2271	2107
4	Alabama A&M	32	5	27	.156	-19.23	-8.38	4	14	4	7	0	18	1938	2285
5	Alabama-Birmingham	35	20	15	.571	0.36	-1.52	10	8	11	5	6	6	2470	2370
6	Alabama State	31	12	19	.387	-15.60	-7.84	9	9	8	3	3	13	2086	2235
7	Alabama	34	18	16	.529	9.45	9.01	8	10	10	6	4	8	2448	2433
8	Albany (NY)	32	12	20	.375	-9.38	-6.70	7	9	6	8	6	10	2150	2216

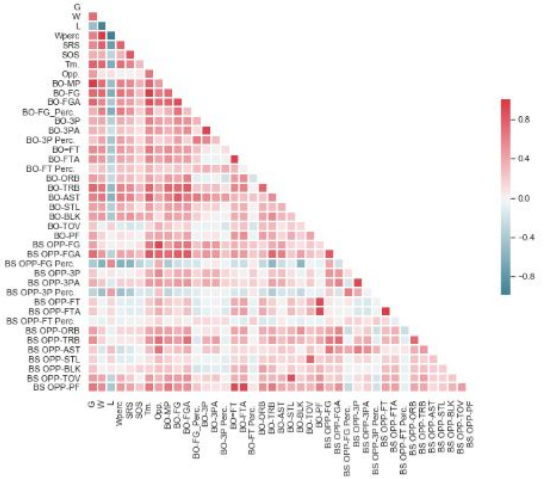
Second, we needed to clean up the data

- Got rid of “noisy” columns (e.g. # of games played)
- Cut down to 6 most recent seasons of data
- Added a binary Y/N feature for whether the team made the NCAA Tournament that season

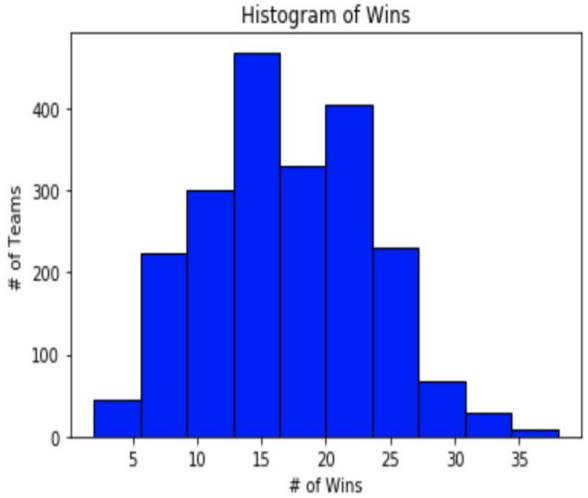


Third, we did some basic data exploration

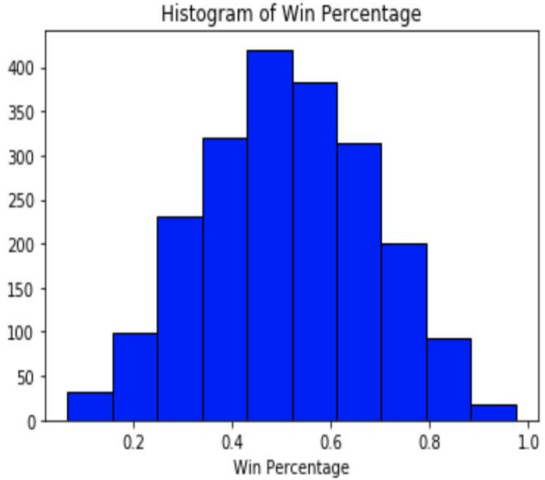
- Our targets: Wins, Winning Percentage, and SRS (“Simple Rating System”: includes average point differential and strength of schedule)
- Generated a correlation heat map, but didn’t find any counter-intuitive correlations with Wins or Win Percentage)



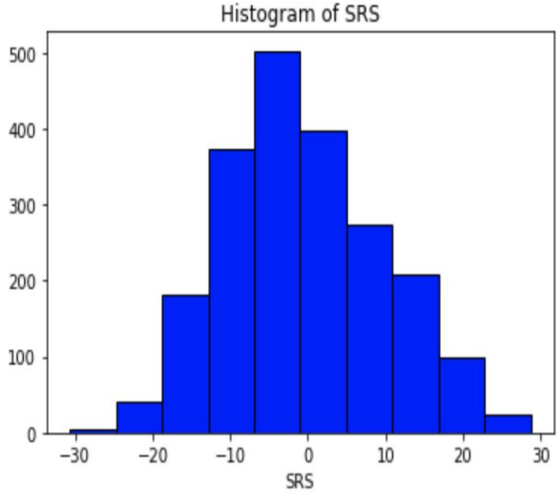
- Found approximately normal distributions for our target features



SCU's Wins: 16



SCU's Win %: 0.516

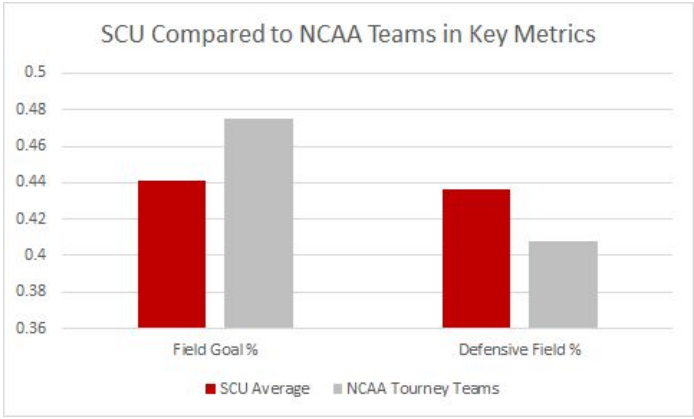


SCU's SRS: -1.67

Data Analysis

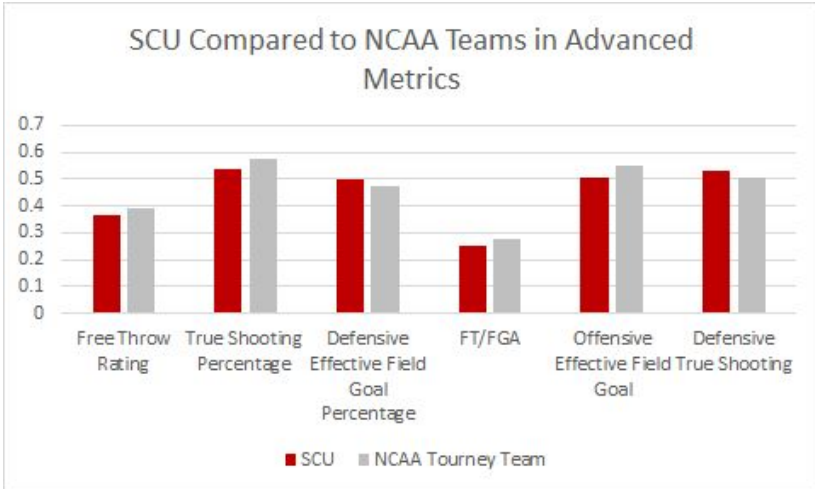
First, we employed linear regression on each of our three targets from a basic statistics data set

- Basic basketball statistics include field goals made, rebounds, steals, blocks, etc.
- Features with both high coefficients and statistical significance were own and opponent's field goal and 3-point shooting percentage
    - We'd be laughed out of the locker room if these were our data-driven insights for coaching staff!



Second, we employed linear regression on our three targets using an advanced statistics dataset

- "Advanced statistics" generated by feature engineering from basic statistics, e.g. "3-Point Attempt Rate" as 3-Point Attempts / sum of all field goal attempts
- Features with both high coefficients and statistical significance:
  - "Free Throw Attempt Rate": free throws attempted per field goals attempted
  - "True Shooting Percentage": weights shooting percentage by each shot's point value



This leads to a seemingly contradictory lesson for SCU basketball coaching staff:

- Should SCU try to draw fouls to win more free throws to increase "Free Throw Attempt Rate," or try to increase volume and accuracy of 3-point shooting to boost "True Shooting Percentage"?
- Answer: it depends! The data seem to point to two distinct strategies; there's more than one way to successfully play a complex game.

Predicting NCAA Tournament Qualification

Finally, we wanted to see how well basic vs. advanced statistics predicted qualification to the NCAA Tournament

- We tried Logistic Regression, SVM, Random Forests, and AdaBoost
- The best-performing prediction algorithm in both cases was AdaBoost:
  - On the basic statistics dataset: 85.8% accuracy and 58.6% f1-score for the test data
  - Slightly improved results on the advanced statistics dataset: 87.5% accuracy and 62.9% f1-score for the test data

Business Insights for SCU Hoops

What can the SCU basketball program immediately take away from this analysis?

- Use the advanced metrics to more robustly track performance
- Set benchmarks such that if SCU hits them, they'll be predicted to be in the running for qualifying for the March Madness tournament
- Choose a distinct strategy for boosting "Free Throw Attempt Rate" and "True Shooting Percentage"
- Recruiting, coaching, and practicing time / resources are limited; everything is an opportunity cost, so SCU should sacrifice statistics that are less strong predictors for ones that are more efficient at moving the needle

How would we take the analysis to the next level in the future?

- Differentiate between teams that qualify for March Madness in different ways (winning their conference vs. receiving an “at-large” bid)
- Rather than using a binary Y/N label for NCAA Tournament qualification, distinguish between rounds of the tournament, and weight deeper tournament runs more heavily
- For each upcoming game, look at the previous few seasons of data for that specific opponent and find most important features for teams who beat them

It's too late for us, but hopefully future SCU students will be able to enjoy basketball glory!