March Madness!

NCAA Men's Basketball Games Analysis and Prediction

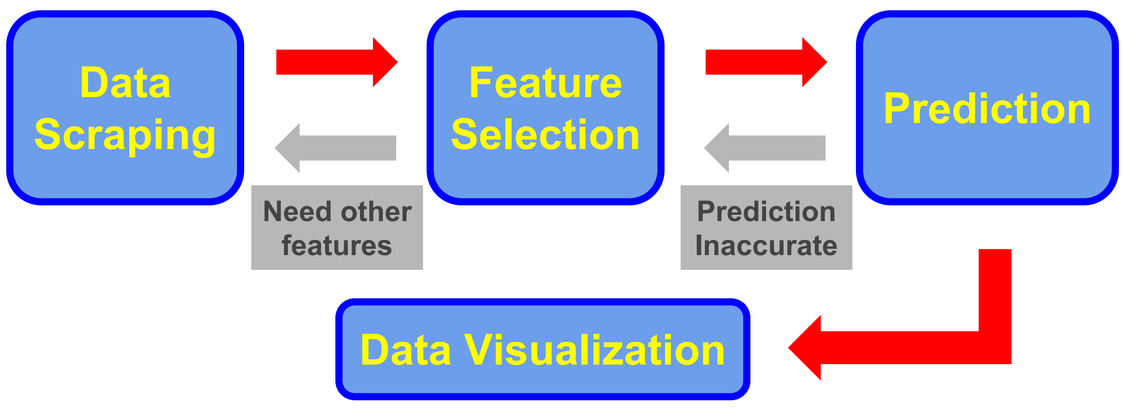
Georgia Tech College of Computing
Computational Science and Engineering

Team 41: Tzu-Wei Huang, Somayeh Hosseini Porgham, Yi-Chi Shao, Pao-Yang Tsai

Motivation

- Every year hundreds of teams and thousands of players do everything they can to win games and fulfill their dream: -The National Champion!
- Being enthusiastic basketball fans, we can't wait to participate in this "out--of--stadium" basketball competition!

Implementation Process



Data Scraping

- Scrape statistical data from *College Basketball @ Sports-Reference.com*
- Use *Scrapy* which is an open source and collaborative framework for extracting the data from websites
- Get "All games outcomes", "All teams conference total", "All players conference average" tables
- Totally 32,756 games, 2,084 teams and 26,497 players in 6 years (2011~2016)

Feature Selection

- Long-term factors
  - The school's average status in the conference
  - The average of 2/3-Point Field Goals, Free Throws, Assists, Steals, etc.
- Short-term factors
  - The school's average status of previous 5 games
- Bag-of-players
  - The composition of a team is also an important factor to be considered
  - Use bag-of-words method to represent a team as a bag of players
  - Use K-means to cluster players in the dataset into K groups
  - Create histograms of K types of player for both team in a game which could be used as features that reflect the composition of a team

Conference Team and Opponent Stats

	G	MP	FG	FGA	FG%	2P	2PA	2P%	3P	3PA	3P%	FT
Team	18	3600	462	1062	.435	366	777	.471	96	285	.337	245
Rank			5th	3rd	11th	2nd	2nd	12th	14th	14th	9th	10th
Opponent	18	3600	461	1033	.446	371	742	.500	90	291	.309	283
Rank			8th	8th	9th	14th	13th	10th	1st	1st	3rd	14th

Conference Per Game

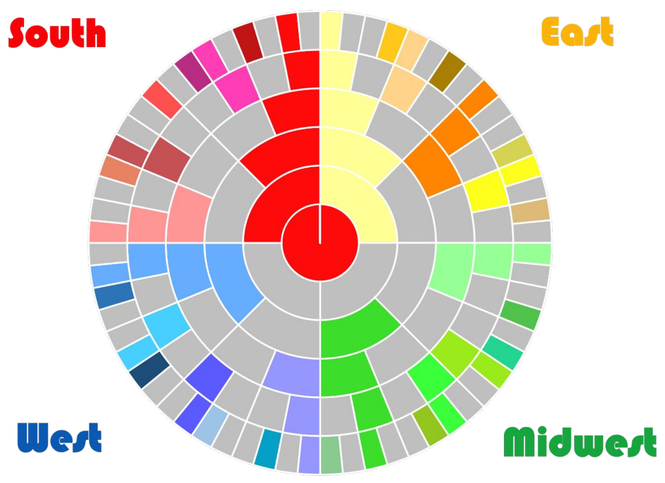
Rk	Player	G	MP	FG	FGA	FG%	2P	2PA	2P%
1	Marcus Georges-Hunt	18	35.1	5.7	11.7	.486	4.7	8.6	.545
2	Adam Smith	18	32.6	5.2	13.6	.384	2.2	5.8	.375

Prediction

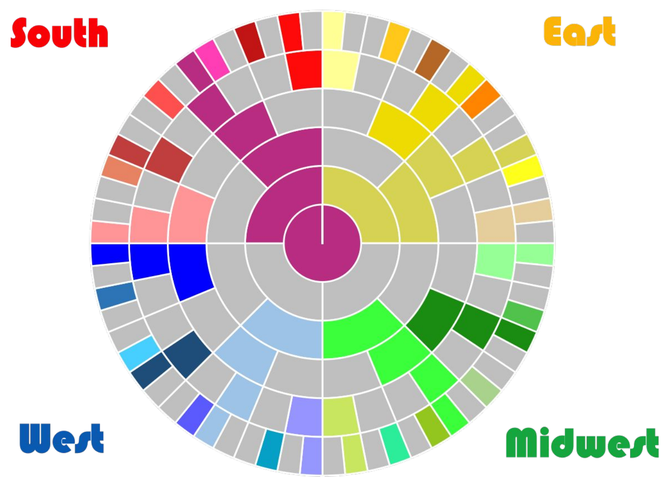
- Use Logistic Regression model from Python Scikit-Learn
- Use the data from 2011~2015 seasons to train the model
- Predict the outcome of 2016 NCAA tournament to test the accuracy of the model
- To predict the outcome of a match between Team A and Team B, we calculate Team A's win probability of both A v.s. B and B v.s. A (switch the order of features) and use the average to determine the outcome
- Bag-of-players increases the accuracy by 2%
- The overall accuracy is 70%

Result & Visualization

Heatmap: Win Rate of games



Bracket: Actual Outcome



Bracket: Our Prediction

- Heat Map
  - The Heat map on the left shows the win rate of each battle combination among 64 teams participating in 2016 NCAA tournament
  - The map should be read based on the teams on each row vs. teams on each column
- The Doughnut-Shaped Bracket
  - The doughnut-shaped brackets on the right demonstrate the actual outcomes of 2016 NCAA tournament and our prediction.
  - The inaccuracy might come from the imbalanced strength distribution among different conferences