FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

FBAS MARCH MADNESS DATA CRUNCH

Deloitte.

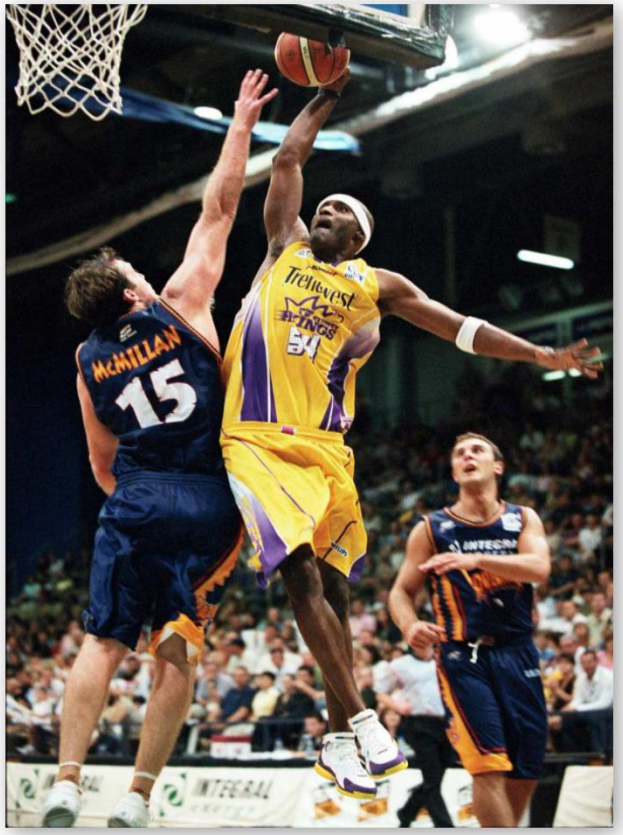
MARCH MADNESS DATA CRUNCH

MARCHKETEERS

FOURdham MARCHKETEERS

Sumi Choudhury Wen Zhang
Chuanyue You Minxia Ji

Introduction



The March Madness Data Crunch Challenge

BKGD

Background

The NCAA Men's Basketball Division I Tournament is a single-elimination tournament played each spring in the U.S., currently featuring 68 National Collegiate Association (NCAA) basketball teams, in order to determine the national championship.

OBJ

Objective

Using data mining and machine learning techniques, FOURdham Marchketeers will choose the best model to predict the winning brackets for this year’s competitions.

Analysis Process

Data Preprocessing

Data Cleaning

Add Distance & AdjEM

Manhattan Distance

Euclidean Distance

Data Transform

Delta

Adjusted Seed

Feature Scaling

Standardization

Feature Selection

Reactive Data Visualization

Domain Knowledge

Features Analysis

Correlation Matrix

Distribution

Recursive Elimination

Modeling

Model Building

Polynomial SVM

Logistic Regression

Artificial Neural Network

Decision Tree

Naive Bayes

Model Evaluation

ROC Curve

Performance Validation (Log loss)

Model Improvement

Learning Curve

Feature Reselection

Summary

Interesting Analysis

Conclusion

Suggestions

Data Preprocessing

Add Distance & AdjEM

Manhattan Distance:

|host_lat - team_lat| + |host_long-team_long|

Euclidean Distance:

sqrt (|host_lat - team_lat|^2 + |host_long-team_long|^2)

Euclidean Distance surpasses Manhattan Distance in our models

AdjEM:

Adjusted Efficiency Margin

AdjEM = AdjOE -AdjDE

Delta

Interpretation

Delta means the difference between TEAM1 and TEAM2 on a specific attribute

e.g.

Delta of seed = team1_seed - team2_seed

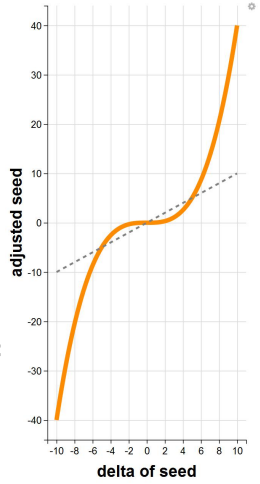
Adjusted Seed

Idea

Enlarge delta of seed when it is large; decrease when it is small

Formula

adjusted seed = 0.04 * (delta seed)^3



Feature Scaling

Standardization

Makes the values of each feature have zero-mean and unit variance

Formula

$X_i' = (X_i - mean(X)) / sd(X)$

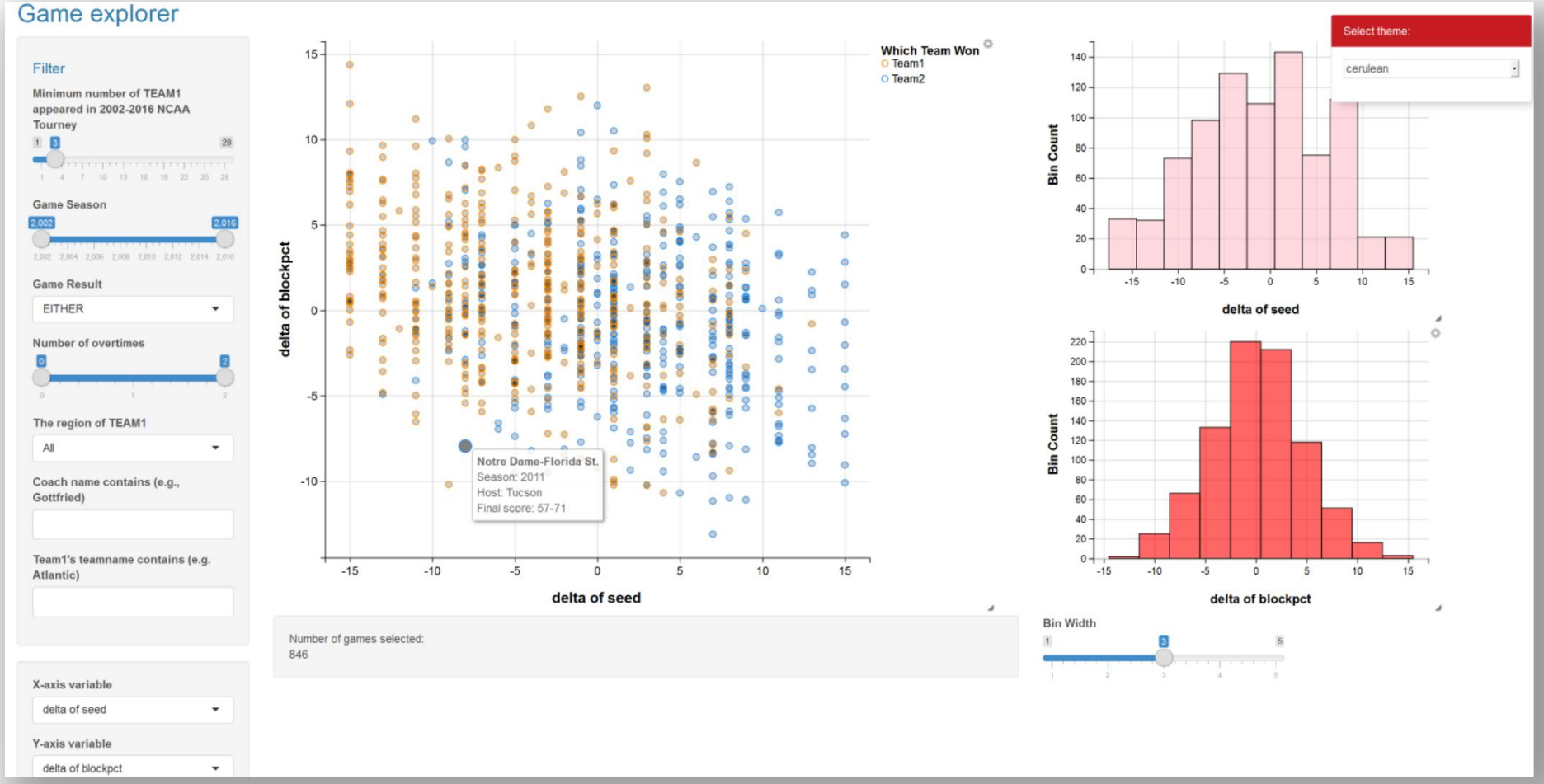
Reactive Data Visualization

powered by R Shiny

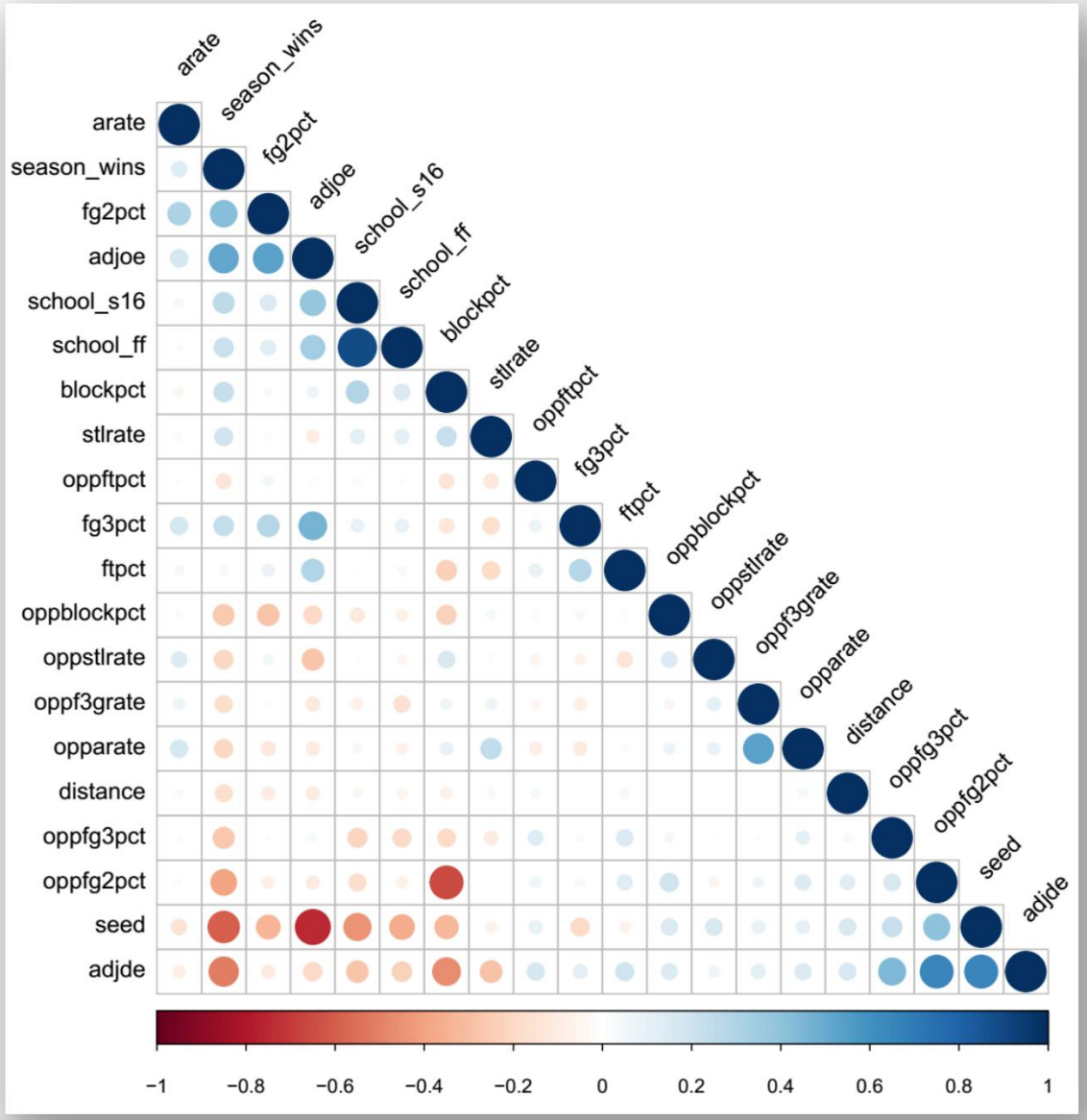
intro

Game Explorer

FOURdham Marchketeers designed a unique web app for game data visualization. It can filter game data by frequency of which team1 appeared in dataset, game season, result, team name, coach name, etc. Choose whatever x/y variable you want and try it out!



Feature Selection



CM

Correlation Matrix

Range = -1 to 1
-1/1 = Perfect Negative/Positive Correlation
0 = No Correlation
Blue = Postive Correlation
Red = Negative Correlation
Color Intensity & Size = Proportional to Correlation

!

Notice

Few features have linear correlation
Be careful when choosing these features

Feature Selection

FDH

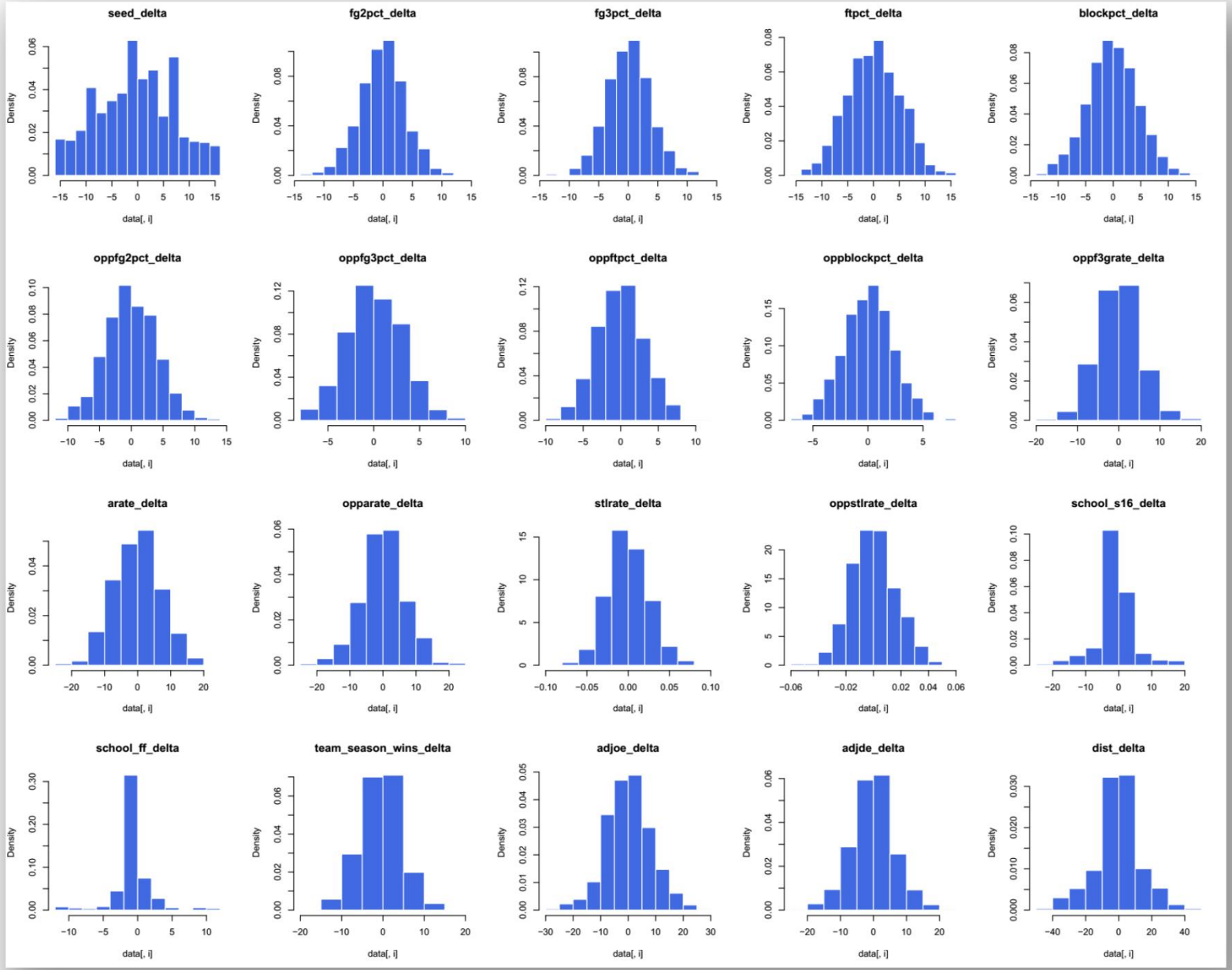
Feature Distribution Histogram

Most features are normally distributed

99→6

Collapse 99 Features to 6

Seed, adjoe, adjde, oppf3grate, blockpct & dist



Optimization Objective

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$$

N = number of games played
Pi = predicted probability that team 1 beats team 2
Yi = 1/0 depending on whether team 1 wins / team 1 loses

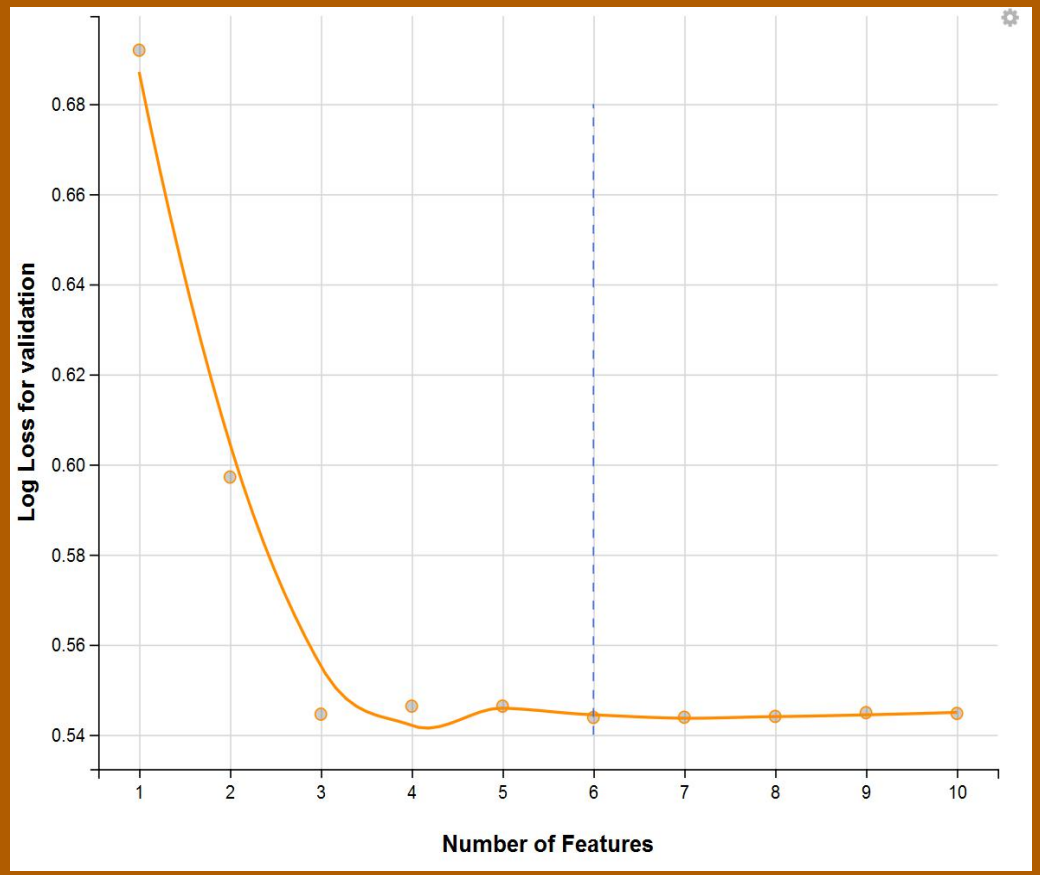
Luck

What's Luck Got To Do With It?
Some models can exhibit higher accuracy compared to its log loss in testing due to major game 'upsets'

Minimizing the Log loss

Log loss heavily penalizes predictions that are both confident and wrong
Create an optimal model by minimizing log loss!

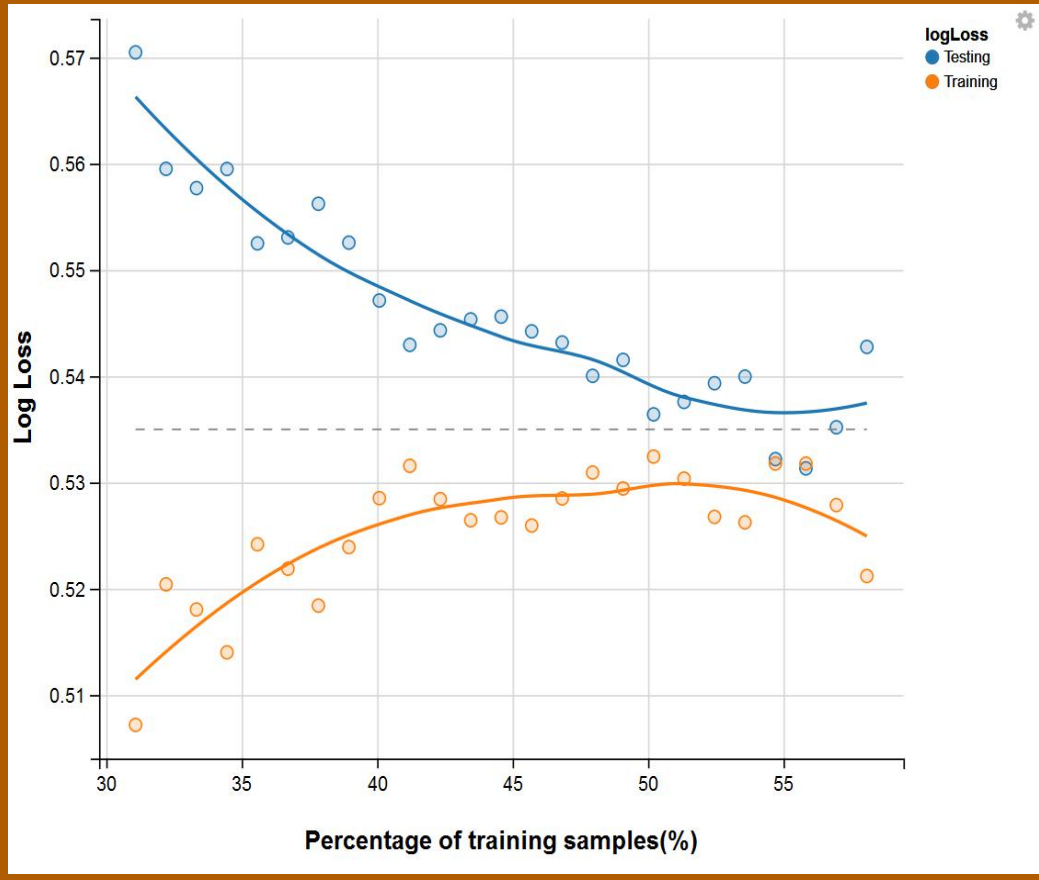
Model Improvement



LC

Learning Curve

Partition ratio is equal to 53:47 (Training:Testing), testing set reaches desired log loss



FS

Feature Selection

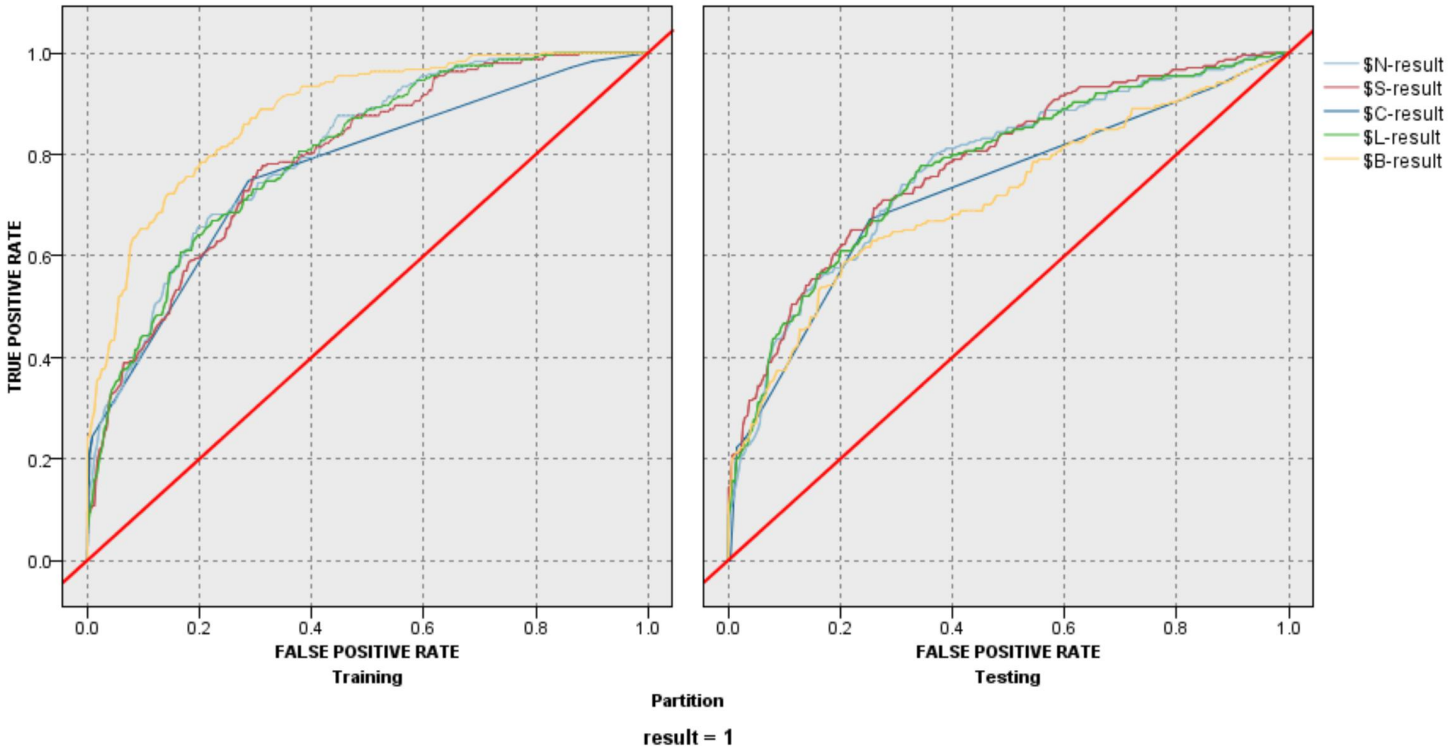
Number of features selected is equal to 6, touching the minimum log loss for validation

Model Evaluation

ROC

ROC Curve

Polynomial SVM is slightly better than other models in Testing.



MODEL	LOG LOSS	ACCURACY
Polynomial SVM	0.537	72%
Logistic Regression	0.539	71.2%
Artificial Neural Network	0.556	71%
Decision Tree	Not Applicable	71.2%
Naive Bayes	0.690	68.3%

PV

Performance Validation (Log loss)

Once again, Polynomial SVM has lowest log loss together with highest accuracy.

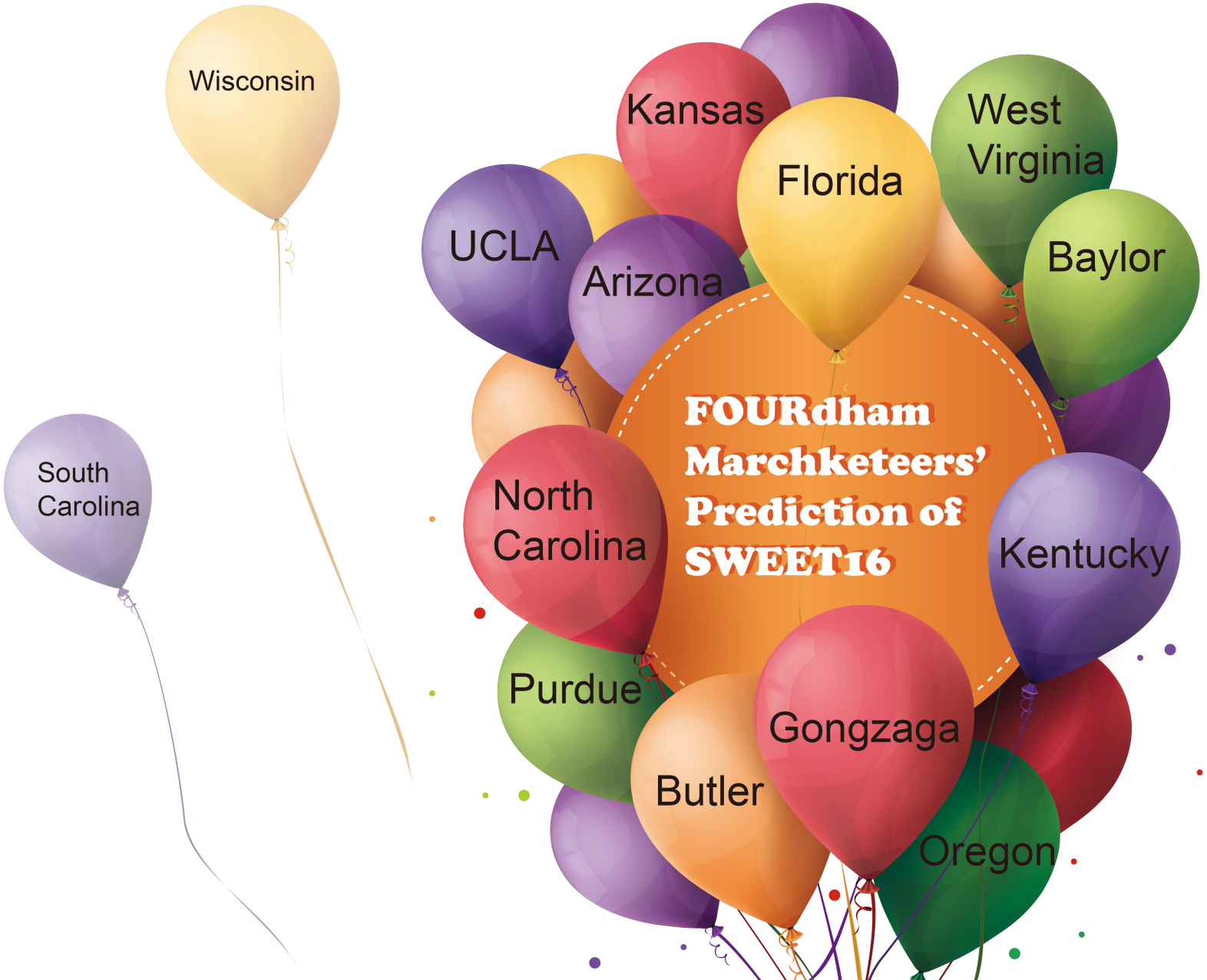
ENS

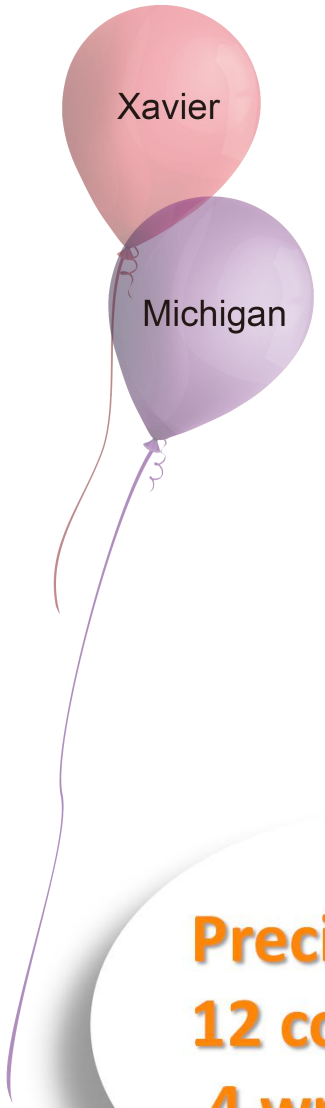
Ensemble

Confidence-Weighted Voting among TOP 3 models (SVM, LR & ANN)

Prediction for 32

Prediction for SWEET 16

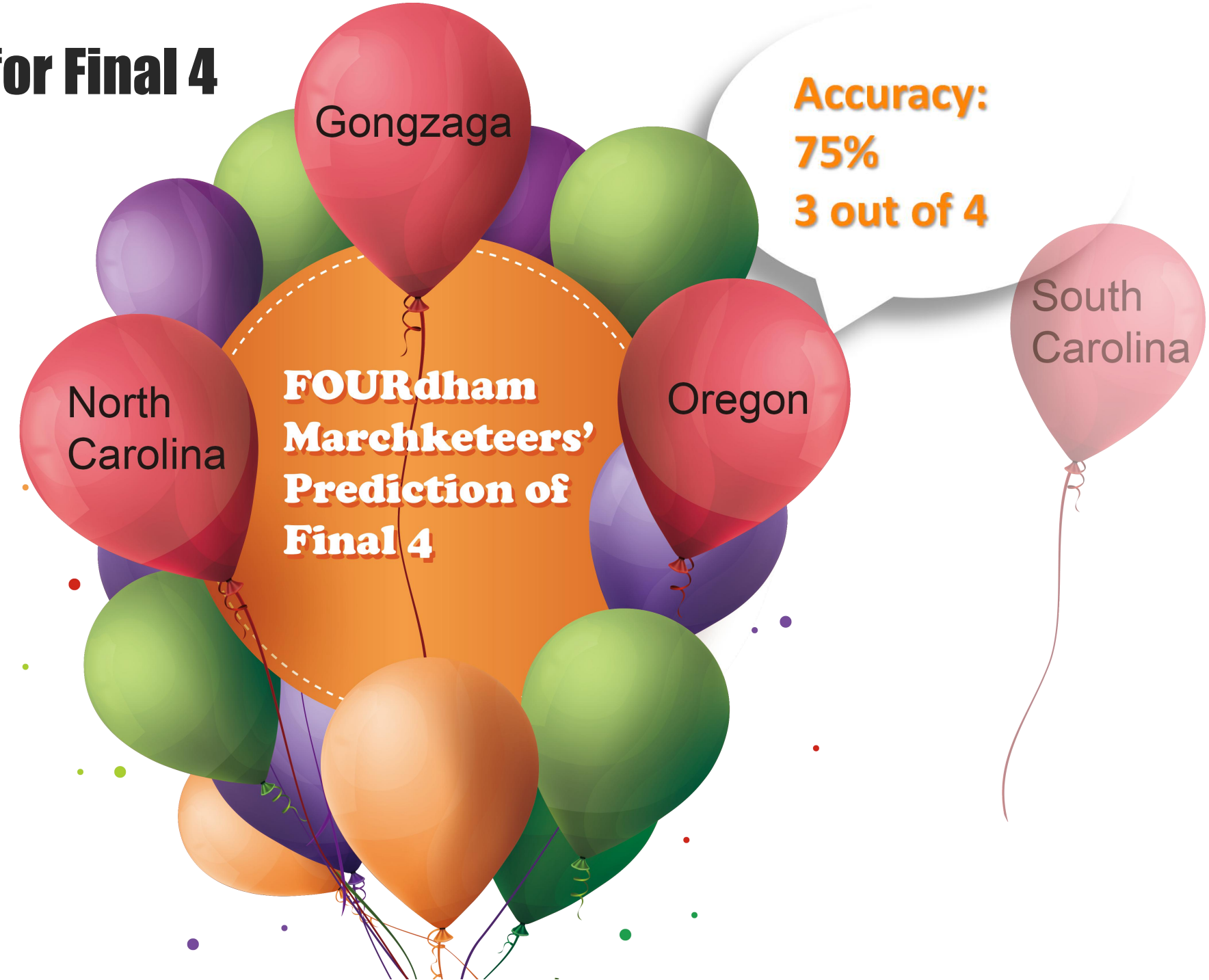




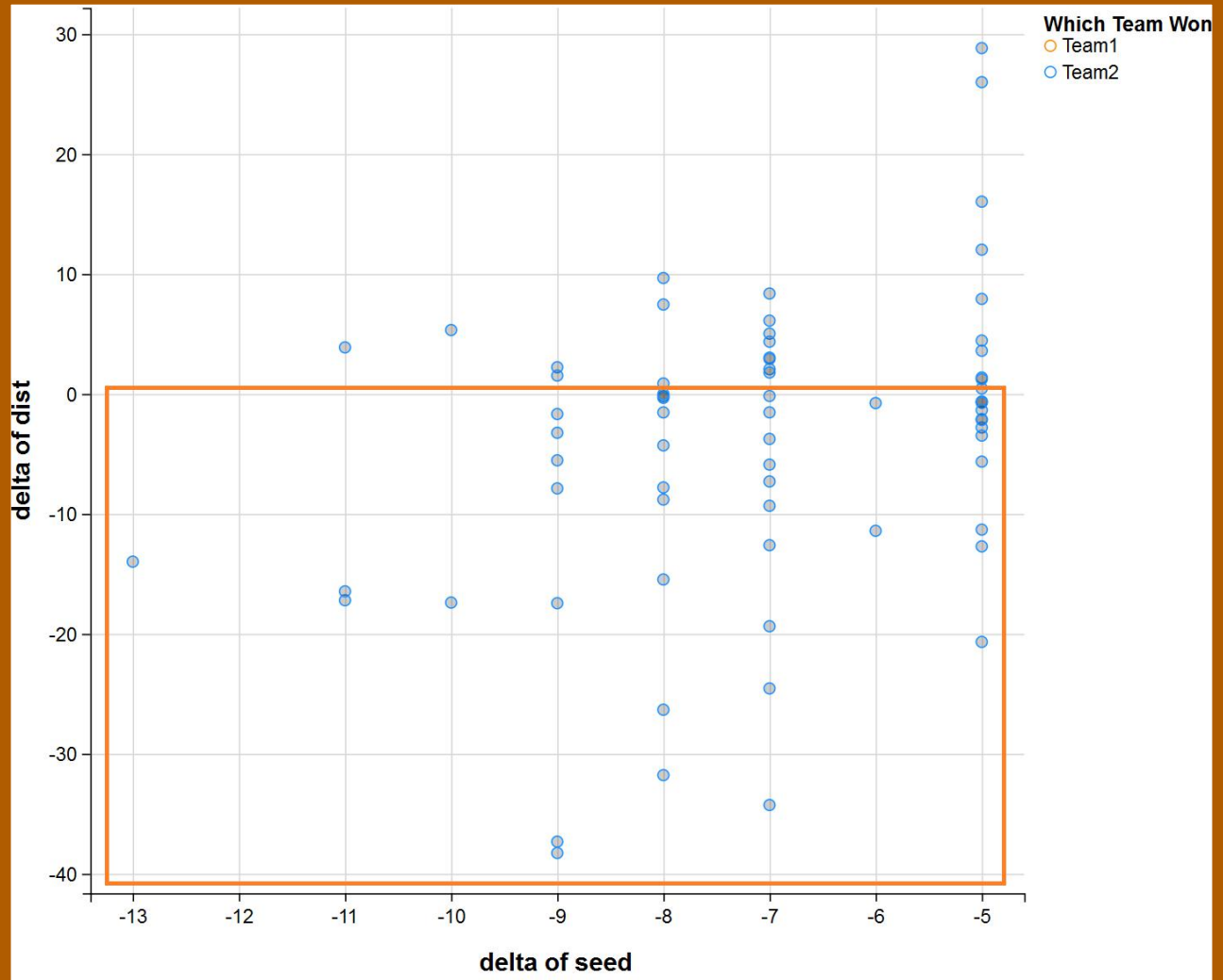
Precision:75%
12 correct
4 wrong

Prediction for Elite 8

Prediction for Final 4



Interesting Analysis



How we did the analysis

Two Conditions in selecting games:
1) delta of seed ≤ -5, which means Team1 had much better seed than Team2
2) only blue points, which means Team1 lost the game (even though Team1 was expected to win due to its better seed)

Our Intuition

Team1's failure may be related to some unfavorable outside reasons (e.g. far away from the host city)

In Fact

we find just the opposite! In the majority of these games (marked by orange rectangle), Team1 was closer to the host city than Team2

Notice!

Be careful about predicting the results of these games in which one of the teams has both better seed and distance advantage

Conclusion

- **SEED is an important indicator, but sometimes might mislead**

  Adding these following features have greatly helped our prediction model:

OA — Opponent Attributes

Percentage of Opponent's 3-Point Field Goal Attempts

EM — Efficiency Metrics

Adjusted Offensive Efficiency
Adjusted Defensive Efficiency
Block Shots Percentage

GA — Game Attributes

Distance of Host Location from Team's Home Campus

- **FOURdham Marchketeers' Prediction Accuracy**

  Up Until Final FOUR: 76.67%

Suggestions

The addition of the following data may have been helpful:

1. Team's level of confidence and preparedness for the Tournament
2. NCAA rule changes, such as an additional 9-inches extension of the three point line in 2007
3. Non-traditional data such as Experience: A team with more Seniors and Juniors receives a higher Experience ranking than a team that consists mostly of Freshmen and Sophomores
4. Campus Support: Measured by the average attendance at home games throughout the course of a season