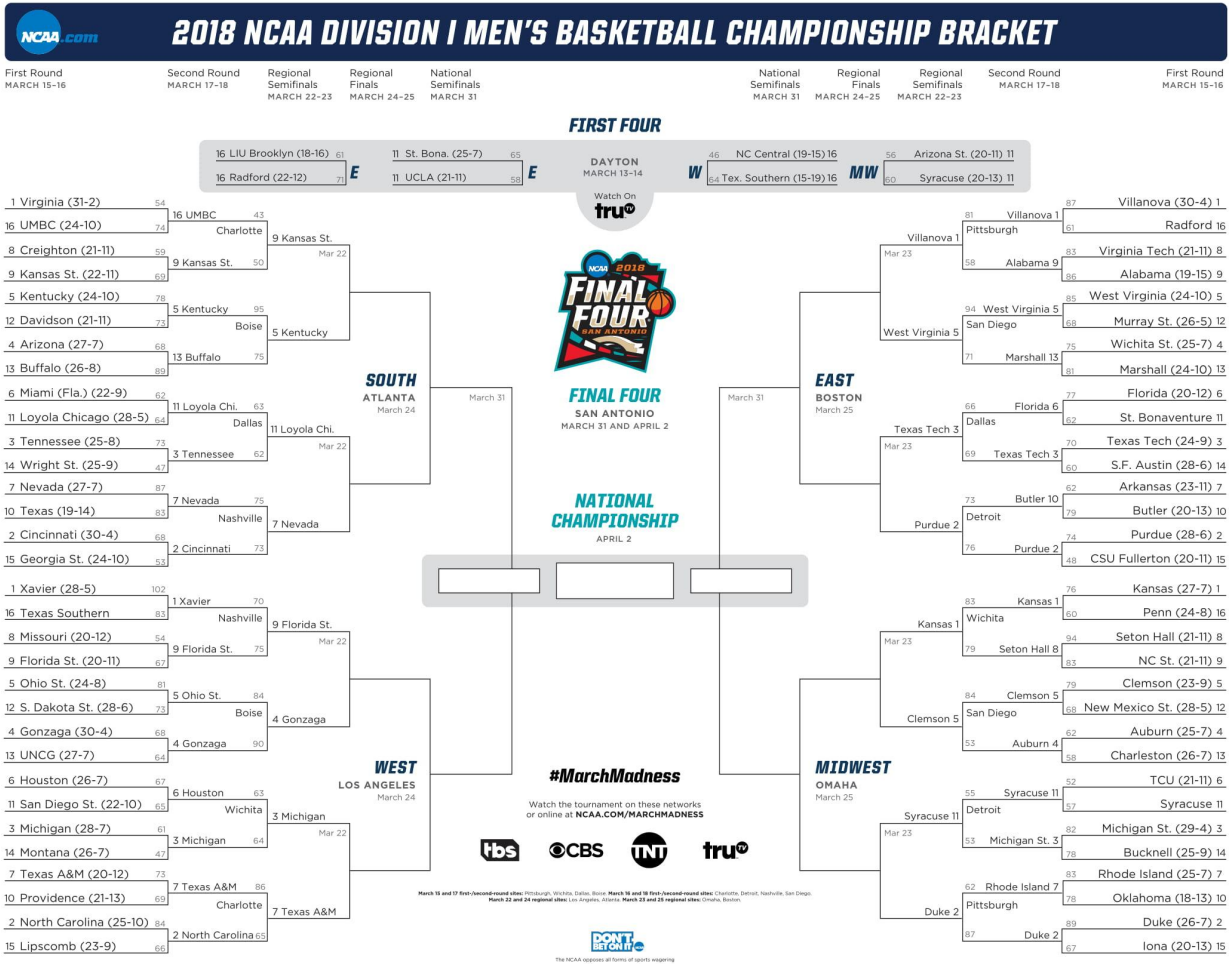
March Data Crunch Madness Final Group Report

Team No Idea

Minghan Wang
Xin Li
Xinjie Guo
Yinglun Xu

Introduction

- 68 teams, single elimination tournament
- Contributed over 80% total revenue from NCAA last year
- Over 10 million TV viewers and 60 million live stream audience on average this year; most watched in 24 years across television
- Duration: 3 weeks starting from middle March
- Odds of a perfect bracket: 1 in 9.2 quintillion
- What we do: using statistical methods and historical NCAA match data to predict the results of all games this year and respective winning probabilities of all teams.



Our Predicted Elite 8

Methodology

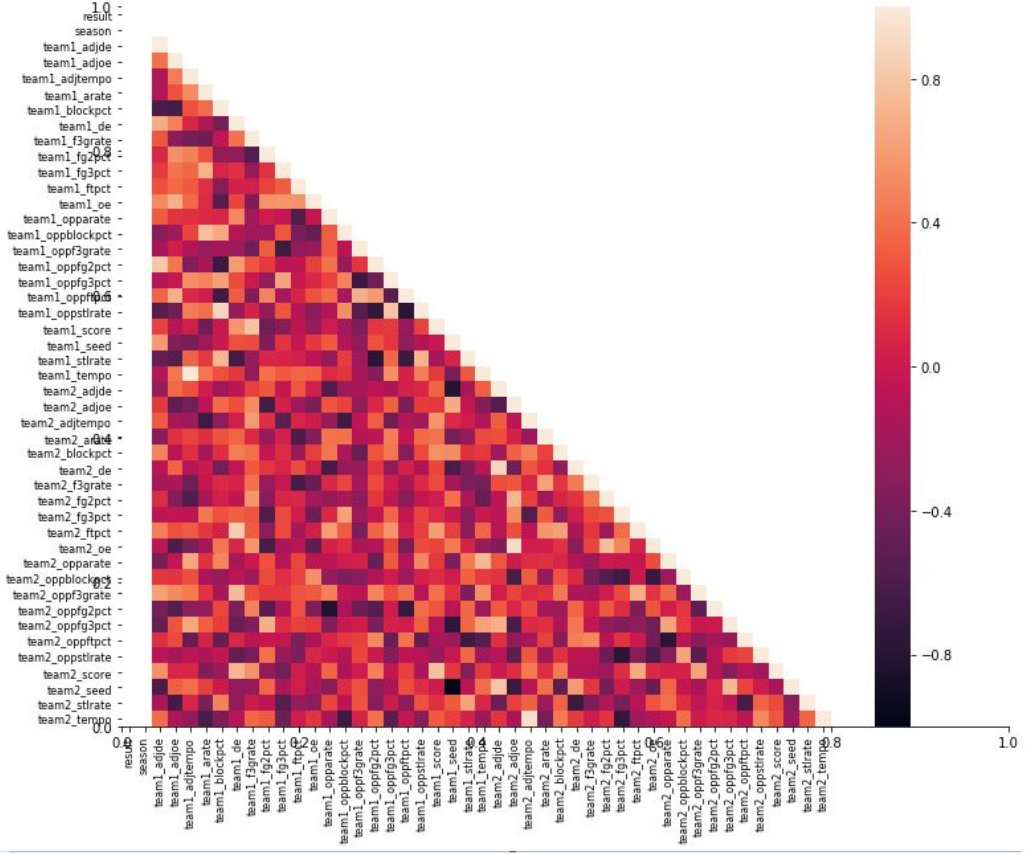
1. Data Cleaning. Removing null value.
2. Feature Selection
3. Calculate new features
4. Data normalization, scale dataset into 0,1 range.
5. Model Building
    a. Use two classification model, svm and logistic regression
    b. Split history data into training and testing
    c. Train the two models
    d. Evaluate the performance
6. Select the better performance model and pass in the new data.

Feature Selection

Sklearn.Feature_Selection.SelectFromModel()

- First filtered out variables that are not important. Variables reduced from 100 to 28 left.
- Manually selected most important variables from 28 leftover variables according to basketball knowledge.
- 5 final key variables.
  - Seed Difference
  - RPI Difference
  - Adj Defense Efficiency Difference
  - 2 Point FG Difference
  - Adj Offense Efficiency Difference

Correlation between different features



Data Standardization

sklearn.preprocessing. MinMaxScaler()

Model performances was bad on dataset that was not standardized as the individual features may not look like standard normally distributed data.

To achieve the robustness to very small standard deviations of features and to improve the performance of the models, features were scaled to have 0 mean and standard deviation of 1.

Statistical Results

```
Accuracy for SVM :  0.643171806167
```

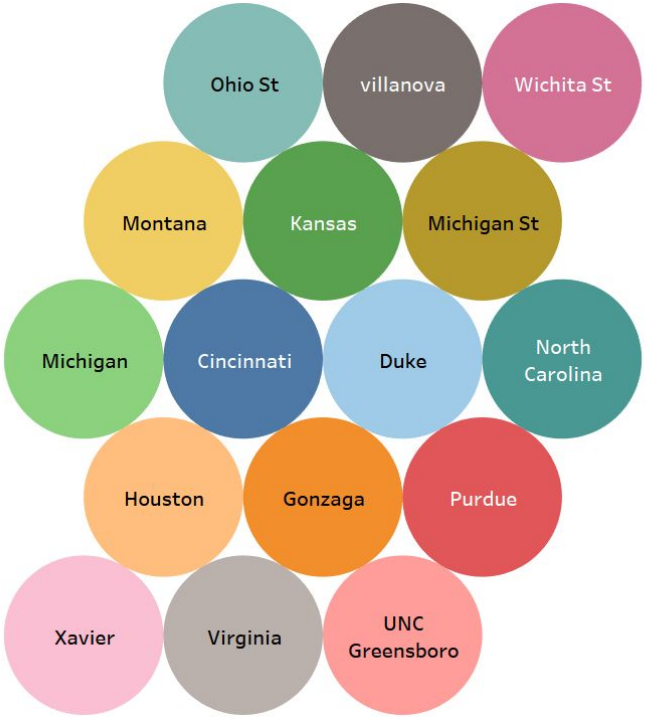
Above is the accuracy without scaling the data.

```
Accuracy for SVM :  0.704845814978
```

Above is the accuracy with scaling the data into 0 and 1 range.

```
Accuracy for Logsitic Regression:  0.647577092511
```

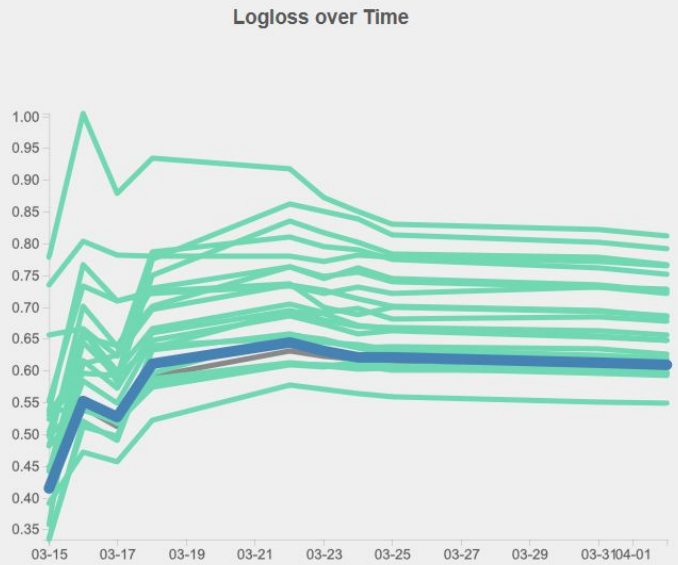
Prediction of TOP 16 teams



Current Prediction Results

Final Logloss: 0.61

Team	Current Logloss
Fantastic4	0.55
QuadraKill	0.59
Class Median	0.6
Team_Hoop	0.6
Eagles	0.6
No_Idea	**0.61**
Gigamassette	0.61
Athletic_Nerds	0.62
KillerWhales	0.63
DataDunkers	0.65
team_Analyst_K	0.65
Go_Mad	0.66
CDNY	0.68
Berserker	0.68
Orange	0.69
B_Girl_Team	0.72
Unofficial_Intelligence	0.72
ruby	0.73
MembaNerverOut	0.75
Bracketology	0.77
Buzzer_Beater	0.77
Team_A	0.79
Resistance	0.81



Model Improvements

- Include more real-time data such as injuries, which actually contributed to many surprising upset wins this year (16th seed UMBC eliminated 1st seed Virginia in first round for the first time in NCAA history. The 6th man of Virginia missed the game due to injury).
- Explore more statistical methods to select features. Regression analysis methods such as LASSO might help improve the accuracy of our model.