March Madness

Philip, Bryan, Bart, Baiyu

Reason for Selection

- **Cultural and financial interest**
  - **Big upside to success, with lots of extant resources to use**
- **Difficult application of ML**
  - **Push our understanding of ML to the furthest extent possible**

Datasource

- Our data came from the NCAA website
  - High certainty of accuracy with respect to data
  - Data is all in one place, and goes back as far as we could hope
    - College basketball is subject to fairly frequent rule changes, limiting how far back our data can reach without too much pollution from rule changes
- Initially, our features of interest were in separate csvs
  - We assembled them into a single dataset by hand
  - We also made further modifications, focusing on round wins instead of overall wins

Hypotheses

- We have maintained a couple standards for success so far
  - 1: H0= ML model is more predictive than a coin flip (accuracy score > .5), Ha= accuracy score < .5
    - Once we started focusing on a per-round model, this evolved to our model-generated bracket beating random brackets (brackets filled out by coin flips)
  - 2: H0= ML model creates brackets with more points than experts, Ha= ML model brackets are scored lower than expert brackets
    - Note that points for number 2 refer to the accuracy of the bracket, with points awarded for correct predictions, increasing as the tournament goes on

Data Exploration

- Initially, our features of interest were all contained in separate csvs
  - We combined them into a single dataset for all seasons, to avoid the time panel element to our modeling
- We started with simple linear and logistic regressions, while some of these had good accuracy scores, we noticed a problem
  - Our target parameter, “tournament wins”, had losses coded as zeros
  - This meant that all teams who did not make it to March Madness were skewing the results of the model
  - Our model could achieve a deceptively high accuracy score because of how loaded our data was with zeros

Analysis

- Moving forward, we decided to try more sophisticated models and reshaping our data
  - We reshaped our data to break up tournament wins from one column into ‘round wins’
    - This lead to a greater emphasis on teams who made it to the tournament, reducing the noise from zeros in our data
    - This also had the effect of minimizing the dataset overall, becoming quite small towards the later rounds

Model

- We started with simple linear regressions, but these proved insufficient
  - There were too many features for this kind of regression, especially with the ‘conference’ column encoded into multiple columns
- Following that, we developed a logistic regression and random forest model in tandem
  - While both models had similar predictive power, the random forest proved better over multiple rounds

Prediction Dashboard

Dashboard Concept

Our HTML page will include Overview of the project, as well as Bar, Line graphs to demonstrate the outcomes of March Madness Predictions using Machine Learning.