Default Final Project

Dataset Information

This dataset is a collection of state and national polls conducted from November 2015-November 2016 on the 2016 presidential election. Data on the raw and weighted poll results by state, date, pollster, and pollster ratings are included.

Content

There are 27 variables:

- **cycle**
- **branch**
- **type**
- **matchup**
- **forecastdate**
- **state**:
- **startdate**
- **enddate**
- **pollster**
- **grade**
- **samplesize**
- **populaion**
- **poll_wt**
- **rawpoll_clinton**
- **rawpoll_trump**
- **rawpoll_johnson**
- **rawpoll_mcmullin**
- **adjpoll_clinton**
- **adjpoll_trump**
- **adjpoll_johnson**
- **adjpoll_mcmullin**
- **multiversions**
- **url**
- **poll_id**
- **question_id**
- **createddate**
- **timestamp**

Tasks

1. What are the trends of the polls over time (by month)? Present visualization.

2. Build two different models (M1 and M2) to predict who is going to be the likely winner. Describe your solutions including any preprocessing steps, predictors, and the classification algorithm. Important:

3. Of course the election results are not out yet therefore there is no ground truth as such. How could you possibly use the given dataset to augment it with ground truth and perform 5-fold cross validation. Describe your approach in length.

4. Based on 4, perform 5-fold cross validation and compute precision of M1 and M2. Based on precision, does one of your model better than the other with 5% significance level? Describe your approach in length.

Deliverables:

1. Write up that describes answers to Questions 1-4. Algorithms, preprocessing steps, predictors, visualization, any assumption all should be clearly stated. (5)

2. Source code and actual results coming from the code.

3. Read me that describes how to run your code. (15 for 2 and 3)

presidential_polls.csv