Data Log

Haotian Li (hl3gy), Jonathan Bae (jjb2mu)

Dating Cleaning and Exploratory Analysis:

1. Import data
2. Import packages
3. Create male data frame
4. Change 8 categorical variables into factor
5. Remove all rows without any rating to males
6. For rows with partial missing ratings on attributes, impute based on average ratings from other customers
7. Confirm there are no missing values in attributes
8. Plot age and remove rows with no age
9. Check missing values of field\_cd; find that all missing values are actually Operations Research and make the corresponding change
10. Plot field\_cd
11. Plot Race and check missing values
12. Check distribution and missing values of goal
13. Analyze the goal distribution of different age groups
14. Remove roles without the variable date and plot the histogram
15. Check whether there is missing value in go\_out and plot the histogram
16. Check missing values of career\_c; find that all missing values are actually tech professional and make the corresponding change
17. Check missing values of exhappy and plot the histogram
18. Change 0, 7 and missing values in met to 2
19. Check number of NAs for each attribute (to look at whether any activity has NAs)
20. Add a dummy variable magazine to represent whether it is a speed dating with book/magazine
21. Get rid of observations in wave 5 and wave 12 because they have different features

Sentiment Analysis:

1. Load licenses into correct empty folders
2. Set date from 2000-01-01 to 2016-10-31
3. Run analysis of tweets for general population for “speed dating”
   1. Create word cloud for this general search
   2. Found some interesting words that may correspond with findings
   3. Run more analysis with geo-location of Manhattan and Charlottesville
   4. Not enough results found to create word cloud.
4. Think of similar words pertaining to speed dating; thought of romance and love
   1. Run analysis of tweet for romance for Charlottesville, Manhattan, and everywhere
   2. Create word clouds
      1. Romance and Manhattan did not come up with any results thus do not use
   3. Run analysis of tweet for love for Charlottesville, Manhattan, and everywhere
5. Look at sentiment scores; not telling enough to present this information
   1. Romance was generally negative
   2. Love was very positive
   3. Speed Dating was generally neutral
6. Look at emotions
   1. Romance and speed dating were mainly in the joy category (after unknown)
   2. Love was classified as joy

Classification Model:

1. Subset to get all variables reflecting the characteristics of males/events
2. Create training and testing samples
3. Construct a tree for all variables; show the tree and its performance
4. Construct a ctree for all variables; show the ctree and its performance
5. Feature selection using random forest
6. Feature selection using Recursive Feature Elimination
7. Subset using a new list of variables based on the result of feature selection and intuition
8. Construct a tree for the new set of variables; show the tree and its performance
9. Construct a ctree for the new set of variables; show the ctree and its performance