ORIE 4741 Project Proposal

D'Azevedo, Gloria
gad87@cornell.edu

Yadav, Pihu
py82@cornell.edu

September 23, 2016

The problem that we will solve is determining whether or not two people who interact in a speed dating event are compatible. We also hope to determine which predictors are most important for determining compatibility, so that a user could have multiple opportunities to match up, and whether or not predictor preferences are different for people based on gender, race and other factors. In addition, we may also compare users opinions of their own sex and the opposite sexs preferences with the actual preferences by comparing their stated preferences with those of the people they actually selected.

The data set that we will use is the Kaggle Speed Dating Experiment data set which was obtained from speed dating events conducted by Columbia Business School professors, this presumably has mostly Columbia students as they are more likely to see or hear advertisements about the event (https://www.kaggle.com/annavictoria/speed-dating-experiment).

We hypothesize that certain factors will automatically influence compatibility such as expected length of relationship with their potential significant other. It is also noted that there may be inherent biases towards speed dating which could manifest in a person not choosing someone based on the fact that they are in the data set or if the set of people that the Columbia Business School professors picked are a true random sample of the population. If the sample is not a true random sample, then predicting compatibility between people who are not similar to this data set may be wildly inaccurate. In addition, if there was early success in the speed dating waves (i.e. if two people decide that theyre compatible during the first round), then they may not return for consequent waves and thus we may have less information about people with their characteristics and their consequent preferences.

Some of the characteristics included are hometown, undergraduate institution, expected career, religion, and ethnicity. There are several questions that have an ordinal response such as How important is it to you (on a scale of 1-10) that a person you date be of the same religious background? Many fields have categorical responses such as ethnicity and field of study, where there is a finite list of options. Age is given as an integer; but it may be beneficial to bucket the ages accordingly to account for inaccurate reporting or to bundle similarly mature individuals. There are also questions asking users to assign weights based on their preferences to characteristics such as attractiveness, sincerity, ambition etc. using discrete values that add up to 100. When looking at each question, the weights should be appropriately rounded, possibly to the nearest 5 or 10 points since it would be silly to distinguish between a user who assigned 23 points and 25 points to the weight for thinking that the opposite date looks for attractiveness. Rounding will also help normalize the responses (if any) that do not add up to 100 for the question.