ORIE 4741 Project Midterm Report

D'Azevedo, Gloria
gad87@cornell.edu

Yadav, Pihu
py82@cornell.edu

October 28, 2016

1 Introduction

The main goal of this analysis was to develop a model that predicts whether or not two people (a male and a female) will be attracted to one another. In addition, we also want to find strong predictors of compatibility. These predictors can be traits that a person has, activities that they enjoy, or how important is religion to their them and their partner. Some obvious applications for this model will be future speed dating events as well as online surveys that online dating websites or applications will use to determine potential matches for a participant. The ideal model could either work on both genders, and the model should be sparse or relatively small so that the surveys (how we get the results for the predictors) will be shorter and thus people are more likely to complete the whole thing truthfully when they answer them.

The data used in the analysis 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). There are 21 different events, called waves, each of which have between 10 and 45 participants, and the number of men and women are roughly equal in each that occur between October 2002 and April 2004. Each participant is required to fill out a total of 4 surveys in addition to questionnaires about their partners during the events. The first survey occurs when they sign up, the second is halfway through the speed dating event, the third is the first follow-up the day after the speed dating event so that they will get a list of their matches, and the fourth is the second follow-up after the speed dating event which is sent 3-4 weeks after they had been sent their matches.

The types of questions asked on the survey are mostly ordinal or categorical variables with a couple of open-ended text questions. It is noted that some of the open ended questions have a lower completion response rate such as the age field and also the school that you completed your undergraduate degree, if applicable. Since they have a low response rate, for a few of these fields, they cannot be reliably used in the analysis because there could be inherent biases for why the participants did not respond. For example, in the future, surveys can still ask these types of questions on a categorical basis by giving participants a few age range choices that they can choose from so they would not be explicitly stating their age but relative age ranges can yield some information, which is better than no information.

2 Data Cleaning Remarks

During the following report we will take on the convention that a "participant" is the person filling out the survey and the "partner" is the person they interact with during that particular round of the speed dating event (wave) that they are participating in. When scraping data and analyzing the coefficients of the predictors in the model, we choose the participant to be female and the partner is male, so that the data is not double counted. All females have interacted with each male once and vice versa so this is not a problem as there is only one incident of a person not completing her initial surveys.

At the beginning of the analysis, we noted that some cities and jobs had commas in them, so had to remove all those otherwise the comma delimiter during the import would break up all those fields and not import the correct value per column or the correct number of columns. We also removed apostrophe's or rewrote the word so that it would not need to be shortened with an apostrophe (i.e. changing "Int'l" to "International") since that seemed to cause import errors.

Some fields are inter-related so if there are blank values in one of them, then the value cannot be found in the similar field. For example, there are two fields "age_o" and "age". "Age" is the self-reported age of the student and is asked when they signed up. "age_o" is the age of the partner during that round. The values for age_o must have been put in after the survey was asked. For example, is a participant did not report her age, the value for "age_o" for all the partners she had are also null. The problem with age is that there are many possibilities why people did not report their true age. For example, men tend to like young women, so women may report a lower age than the true value. On the other hand, women prefer older, mature men as a measure of stability so men may report a higher age than the true value. For now we did not edit or add any of the "age" or "age_o" values, nor do we use them in the analysis.

While investigating trends in categorical data such as the "field_cd" column (corresponds to a numerical code for the field of study that each person has) or "race", we noted that there were some NA values so we reassigned those NA values to be that column's "Other" category. Even though we would be over-counting the trends in the "Other" category, we hope that the majority of participants had accurately written down their race. If the data field was an ordinal type of variable such as "tv" (measures the interest that a person has in watching TV on a scale from 1 to 10) we reassign NA values to be 0, indicating that they had no interest in that activity and that's why the participants left them blank.

Some of the survey fields requires a participant to rate their preferences about their partner, however, there are different instructions for this variable on the survey for different waves. For example, during the initial survey before the events took place, the participants in waves 1-5 and 10-21 were asked to divide up 100 points among 6 different categories (more points implies higher importance). In contrast, the participants in waves 6-9 were asked to rate each category on a scale of 1-10 for the importance of each attribute in a partner (a rating of 1 indicates that the attribute is not at all important while a rating of 10 is extremely important). However, looking at the data by wave, the participants in waves 6-9 have also divided weights from 100 into the 6 attributes even though the key or instructions have said otherwise. Since each of the attribute weights are all on the same scale for all waves, the values did not need to be normalize the weights to the same scale. The NA values were reassigned to be zeroes so that numerical calculations can be performed. There are many other fields that have weights or ranks as the expected input, so if the fields were not completed, (i.e. the participant left the field blank) then we reassign those blanks or NA values to be 0 indicating that the field is not important to them.

3 Initial Data Analysis and Methods

In the initial sign-up survey, participants are asked how important each of 17 different activities are to them. The activities include yoga, reading, and watching sports, and there may be high correlation between the interest level in some of the activities. For example, a person who likes theater may also really enjoy watching movies. These values are integer values from 0 to 10 and a higher value implies higher interest in the activity. If two people have similar amounts of interest in the activity, then that should reflect in their classifications assuming that the participants do not mis-rank the interest level. We hypothesize that including similar levels of key interests should improve accuracy of classification.

We tried fitting a linear model predicting match with all of the activities for both men and women (total of 34 variables) and taking only the significant variables from those to get a model. In addition, we try using a forward/backward selection algorithm using AIC to find an optimal model using the activities. Unfortunately we did not see that the same activities for both men and women were chosen so we will have to re-evaluate this method in the next steps. One method that we will try is to fit the linear model for matching onto the vector of the absolute differences between the participant and the partner for each activity.

We also consider a norm of a vector of differences in interest that a participant and her partner have in an activity. Using the norm squared versus another error measure such as absolute value will penalize larger differences more, although the maximum amount is 81 in this case for each activity. Using the absolute value (or the 1-norm) will penalize at most 9 for each activity. Then later on, maybe we can use this value as another field in the data instead of using the activities separately.

In one of the initial analyses, we wanted to identify how men value different traits that they want in a partner (attractive, sincere, intelligent, fun, ambitious and having shared interests). We also include the parameters "samerace" (which identifies whether the partner is of the same race) and int_corr (which gives a correlation between the actual interests of the participant and the partner) in the analysis.

4 Next Steps

From the insights that we have gathered so far from our initial analysis, we have a set of predictors that we think are significant to our model. These predictors are useful for determining whether or not two people will initially match or at least if one person will like another person. Since we also have information from follow-up surveys to the participant (a survey from the day after the speed dating event and another survey a few weeks after the first follow-up survey), we hope to use this information as another confirmation of the matches from the event. The response rate for these follow-up surveys are lower than the initial sign-up survey and the survey halfway during the speed dating events since they require more effort than the others, but hopefully there is still enough responses to aggregate relatively accurately. Also, some of the questions and responses for the follow-up surveys are very similar to questions that have already been asked to the participant such as how important are certain traits in the opposite gender or how do they think that they compare across others in terms of attractiveness, sincerity, etc. In theory, their responses should not change over time through the survey but we should run checks just in case.

In addition, so far when casually determining what predictors to use in the model, we have been using the full data set. However, in practice it's best to divide up the data that we have into disjoint training and test data so that we can train different models on the training data and then test the model on new data. We have already developed a function in R to use the boostrap method for resampling but that can also create biases for overfitting the data. Other methods that we have learned are using the k-fold cross validation to compute the error rate or number of misclassifications. Since we only have around 8,000 data points in total, we could theoretically also try a Leave-One-Out-Cross-Validation (LOOCV) method which is a special case of k-fold cross validation where k is equal to the total number of data points.