The Matchmakers: Data Analysis Report

D'Azevedo, Gloria
gad87@cornell.edu

Yadav, Pihu
py82@cornell.edu

December 5, 2016

1 Executive Summary

The goal of this analysis is to create a model that identifies whether or not a male and a female are a match in the speed dating context. The expectation of the project is to develop a model, that will identify important traits people look for in a partner so that effective onboarding surveys for speed dating events or online dating will have high accuracy rates in predicting matches. The data has several types of formats including ordinal variables, open ended text variables, and binary variables. The 3 types of models that are tested are a nonregularized linear model with best subset selection, logistic models, a regularized model using lasso with cross-validation, and a K-Nearest-Neighbors model. Out of these models, the model obtained using cross-validation and a lasso regularizer has the lowest misclassification rate without overfitting the data nor having too many variables. In the future we hope to incorporate other variables indicative of a "match" such as one person calling the other or the two people going on a date. These fields, that were asked in surveys taken after the speed dating exercise, have low response rates and a high bias so they may not be completely reliable but could still give good insights about the data if there were more responses.

2 Introduction

The main goal of this analysis is to develop a model that predicts whether or not two people (a male and a female) will be attracted to one another. We want to do this by identifying what traits individuals look for and value most in their partners. Some obvious applications for this model will be organizing future speed dating events as well as online surveys that online dating websites or mobile applications use to determine potential matches for a participant. The ideal model should work on both genders, and the model should be sparse or relatively small so that the surveys will be shorter and have high accuracy and response rates.

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. (https://www.kaggle.com/annavictoria/speed-dating-experiment). There are 21 different speed dating events that occur between October 2002 and April 2004, each of which have between 10 and 45 participants, and the number of men and women are roughly equal in each. Each participant is asked 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. However, it is important to note that the response rates for each of the 4 surveys are not 100% and the missing data poses a problem.

The types of questions asked on the survey are mostly ordinal or categorical variables with a couple of open-ended text questions. Some of the open ended questions have a low completion response rate such as the age field and also the school that you completed your undergraduate degree. These fields therefore cannot be reliably used in the analysis because there could be inherent biases for why the participants did not respond. In the future, surveys can still ask these types of questions as a categorical or multiple-choice variable by giving participants a few age ranges that they can choose from so they would not be explicitly stating their age (a potential reason for nonresponse).

3 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 wave. All females have interacted with each male once and vice versa so this is not a problem. There is only one incident of a person not completing any of the surveys (including the initial survey).

For many questions in all the surveys, participants are asked in the surveys to divide up 100 points to 6 attributes by importance. These main 6 attributes are attractivness, sincerity, intelligence, fun personality, ambition, and shared interests. However, there are several sections where the shared interests are either not recorded in our data or the survey forgot that attribute when asking the participant. It may be possible to impute that value for some participants (assuming that the question was on the survey but the data was not recorded) by subtracting the sum of the points for the other attributes from 100 to get the points for the shared interests attribute as long as there is no non-response in the other questions.

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" (integer corresponding to a type of race), we noted that there were some NA values so we reassigned those NA values to be that predictor'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 or what field they classify themselves in. If the data field was an ordinal variable such as "tv" (measures the interest that a person has in watching TV on a scale from 1 to 10) we can reassign NA values to be 0, indicating that they had no interest in that activity and that's why the participants left them blank. Obviously, imputing too much data affects the results significantly, so we only do that when trying to implement the KNN algorithm in Section 5.4.

4 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 in-

terest 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.

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 in the implementation we would have to have different onboarding surveys for men and women or we would include the nonsignificant variables on both surveys for the sake of consistency. We also considered using a vector of absolute differences ($||\cdot||_1$) or the sum of squared differences ($||\cdot||_2^2$) between the interest level in an activity between two people.

In another analysis, we want to identify how men value different traits that they want in a partner. Initially we computed the summary statistics for several high response traits in all instances where men said 'yes' to their partner. However, comparing summary statistics is not an adequate indicator of the significance of certain predictors, so the information is used to develop a more detailed model.

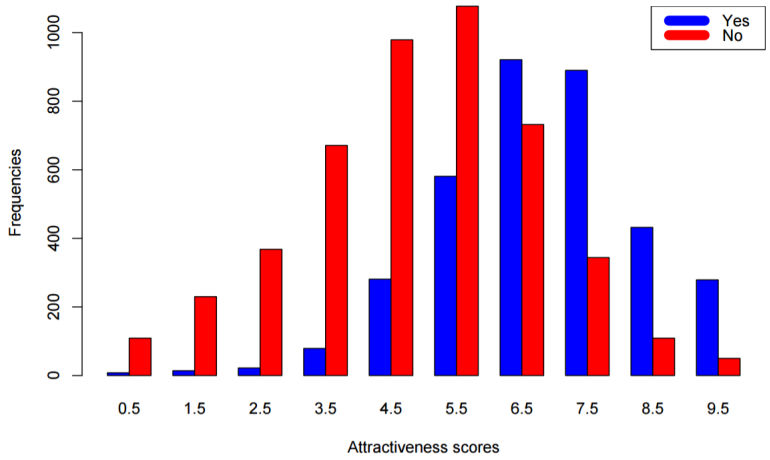
We also plotted the scores given to partners for different characteristics, along with whether the participant

Table 1: Summary table for key statistics

Statistic	N	Mean	St. Dev.	Min	Max
Same Race	3,518	0.409	0.492	0	1
Attractiveness	3,507	7.285	1.533	1	10
Sincerity	3,480	7.596	1.514	0	10
Intelligence	3,474	7.757	1.330	3	10
Fun	3,449	7.334	1.512	0	10
Ambition	3,271	7.159	1.592	0	10
Shared Interests	3,167	6.463	1.841	0	10

said 'yes' to them. This plays a strong role in understanding whether high scores for a particular attribute correspond to more 'yes' values than 'no' values, which means that it is am important trait that people value in their partners. Attractiveness scores are later found to be a strong attribute for predicting the decision of a participant and we see from the graph below that partners who were given low attractiveness scores generally get more 'no' values whereas partners with high attractiveness scores get more 'yes' values.

Figure 1: Histogram of Attractiveness Scores



5 Model Development

5.1 Logistic Regression

We want to identify how men value the main 6 attributes in a partner that were identified in Sesction 3. Characteristics such as whether the partner is of the same race is also included in the analysis. Since the decision variable (of saying 'yes' or 'no' to a partner) is boolean we have chosen to use logistic regression for the analysis. In this analysis all the values with missing data have been removed from the data when computing the coefficients for the logistic regression, as it would be misleading to impute the values of scores given to partners or assign the average value when the answer has been left blank by the participant. We do not expect this to create problems in our analysis since more than 85% of the data is still available. In logistic regression, instead of directly fitting the binary output to a linear model, we fit the data to the model denoted in Equation 1 where p is the probability that the outcome is 1 and j denotes the number of variables in the model. An additional parameter is required as a threshold of classification, t. The model outputs a probability so any output above the threshold will be classified as a success or 1 while all other values are classified as a failure or 0.

$$ln(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{d} \beta_j x_j \tag{1}$$

The model with all of the above predictors are fit first, then using only the significant variables from that one, the model is refit as the final logistic model. The resulting model uses attractiveness, sincerity, being fun, ambition, and having the shared interests as significant variables and has a misclassification rate of about 22.5% with a threshold of $t = 0.5$. In other words, the model correctly classifies about 77% of men's decisions correctly.

A similar analysis is done for the female data. The amount of missing data is about 17%, skewed slightly high by one female who didn't complete any surveys. The same 7 attributes are fit to the data at first then the resulting significant variables are refit. The final model has 6 variables (and an intercept) which are all 6 of the main attributes as mentioned in Section 3 and has a misclassification rate of about 22%.

5.2 Best Subset Selection Using Linear Models

The best subset selection model uses an exhaustive search to study all possible combinations of predictors for a given model size. For this analysis we model the decision that a participant gives their partner by using the main 6 variables (detailed in Section 3) as well as other variables such as how much they liked the partner, the probability that they think their partner will say yes to them, how happy they expect to be with the speed dating exercise, how often they go out (to see if extroverted behavior correlates with saying yes to more people), their goal behind attending the speed dating exercise, whether they have met the partner before and how often they go on dates. We only include variables that could directly impact outcome and have disregarded those that have a large number of missing values.

The best subset in our model is selected using the Bayesian Information criterion (BIC) which is a function that uses both the maximum likelihood value as well as the number of parameters in the model to determine the best model overall. The specific calculation of the BIC is found below.

$$BIC = -log(L) + d * log(n) \tag{2}$$

The BIC value takes into account how accurately the model fits the data while at the same time penalizing models that have a large number of parameters (if the likelihood does not increase enough from adding another parameter). Hence, when determining the best model overall model, it yields an accurate as well as a simple model.

The best model of size 1 used the score for how much they liked the partner. This model is reasonable since how much you like a person should have a very strong correlation with whether you say yes to them. For a model with two predictors, the best possible combination was found to be the score for how much they liked the partner and the score they gave for attractiveness of the partner. Similarly, best possible subset combinations were predicted for models of size 1 to 13. The best subset among all possible combinations of predictors and number of predictors was found to be the one containing 9 variables, which were the scores given to the partner for attractiveness, sincerity, being fun, ambitious, having shared interests, how much they liked the partner, the probability they expect their partner to say yes to them, how often they go out, and how happy they expect to be with the speed dating exercise. The corresponding misclassification rate for this model is $\approx 24\%$.

While most of the predictors are easy to justify, it is interesting to see that participants were more likely to say yes to a partner when they thought it was highly probable that their partner would say yes to them. Also, participants who go out more often were also likely to say yes to more partners and those who expected to be happy with the speed dating exercise also said yes to more people. This indicates the possibility of a feedback loop, so participants who expect the speed dating exercise to go well for them are able to find more people that they are interested in.

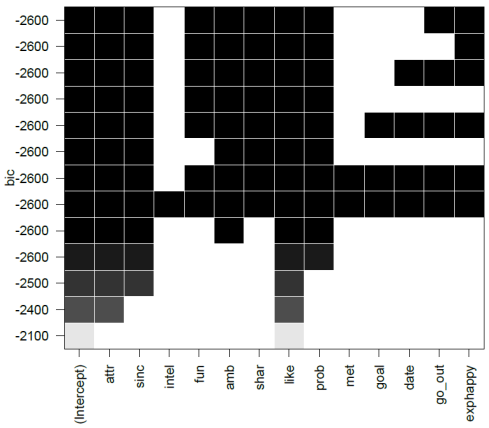


Figure 2: Results from Best Subset Selection for model size up to 13

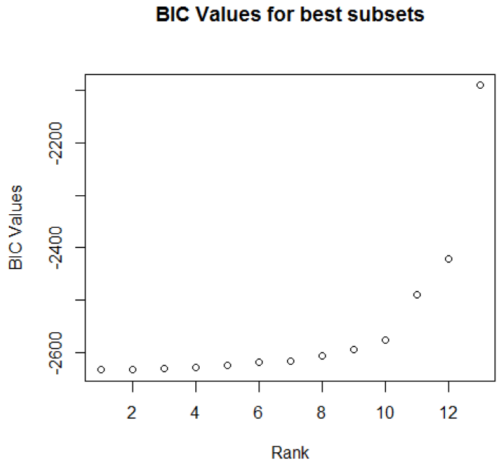


Figure 3: BIC values for the best subsets

In Figure 2, the BIC values for the top 10 models look to be about the same. However, explicitly calculating them for each of the model sizes in Figure 3, we can see that the top 5 models have BIC within a margin of 10 so they are all very good fits and the top model of size 9 has essentially the same BIC value as the second best model of size 8 so if we wanted to further decrease variance but not sacrifice fit, we could also use the second best model of size 8.

5.3 Cross-Validation with Lasso Regularizer

In the logistic regression models described in Section 5.1, we used the entire dataset to create our model. While this method creates a comprehensive model, it may not predict new data accurately. One way to estimate test error is to perform cross-validation, a technique that picks the best model on the basis of how well it performs when tested on the data it has not trained on. In addition, a lasso regularization is implemented to force a more sparse model, in which only the truly relevant parameters have non-zero coefficients, thus creating a simple and accurate solution. For this model our input parameters and output remain the same as the ones used in the Best Subset analysis (see Section 5.2) and we use 20 folds in the our cross validation procedure.

Using cross validation the best value of λ (the coefficient of the lasso regularizer) is calculated by training the model on a subset of data with a number of different values of λ. Then, on the remaining testing data, the test error is calculated and the λ of the model that yields the minimizes the cross validation test error is used. Hence, with all the above details we are able to implement our model, and get a good classification model. Rather than choosing the model with the lowest misclassification rate, we choose the simplest model that has a misclassification rate within one standard deviation of the minimum misclassification rate. This model has nine predictors, which were the scores given to the partner for attractiveness, sincerity, being fun, ambitious, having shared interests, how much they liked the partner, the probability with which they expect their partner to say yes to them,how often they go out, and how happy they expect to be with the speed dating exercise. It is very similiar to the one predicted by the best subset model, implying that it is a strong prediction model for our dataset. The misclassification rate with this model is $\approx 23\%$.

5.4 K-Nearest-Neighbors (KNN)

In order to implement the K-Nearest-Neighbors (KNN) algorithm to predict a match between two people, the data must not have any missing values nor can it use variables with text. Thus, some cleaning scripts were applied and some fields that had NA values were inputed to an "Other" category or if the missing field had some form of weights or rankings, the values were imputed to be 0 (just for the application of this algorithm). In addition, since there is relatively little data, a cross validation method with 10 folds is used to train and test the model on disjoint sections of data. This method of reassigning or imputing may cause some bias in the data so we do not draw significant conclusions from this estimate. In addition, it may be the case that the algorithm does not yield the same result for a match depending on the view (male or female as the participant) which has a skewed interpretation so we also try to fit whether or not the participant "likes" their partner. Another thing to note is that since their partner has to have a different gender than the participant, gender should probably not be included for finding the outcome of similar people.

Initial trials of KNN showed that a few number of nearest neighbors to implement KNN is too flexible and we have a high misclassification rate with the best case scenario having a misclassification rate of around 85%. More tests can be done to lower this rate by increasing the number of nearest neighbors used, also using less fields in the algorithm, and using only complete data so there's no bias from imputing variables.

6 Results and Conclusions

A summary table of the models and their results are as follows. The models with lowest misclassification rates

Table 2: Summary table of models and corresponding misclassification rates

Model Name	Model Size	Misclassification Rate
Logistic Regression for Men	5	23%
Logistic Regression for Women	6	22%
Best Subset for Linear Regression	9	24%
Cross-Validation with Lasso Regularizer	9	23%
K-Nearest Neighbors	33	85%

are the logistic regressions for each male (Logistic Male) and female (Logistic Female)and also the linear models that included cross-validation and lasso regularization (Cross-Validated Lasso). Each of these types of models has advantages and disadvantages. The logistic regressions are smaller models yielding a similar misclassification rates as Cross-Validated Lasso but the data was the same for both training and making predictions so there's a high probability that the model is overfit to the data. In addition, the number of variables that were made available to the respective lasso regressions was less than the number for the Cross-Validated Lasso method so there is a possibility that the logistics models underfit the data. The best model that we recommend from the aforementioned list of models is the Cross-Validated Lasso since it is tested on new data and minimizes the BIC to optimize fit and model size.

7 Next Steps

In the future we can utilize the 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) as another measure of confirmation of the matches from the event, despite their low response rates and potential bias. However, when there are responses, they may give a lot of insight into what drives initial dates and calls.

We have already developed a function to use the boostrap method for resampling but that can also create biases and overfit the data. Another method that we will use for the linear or logisitic models is the k-fold cross validation method to compute the error rate or number of misclassifications. This method begins by randomly divides the data into k equally sized, disjoint partitions. Then each fold is taken in turn to be the test data set evaluated on the model that was calculated on the other $k-1$ folds as the training set. In addition since there are only around 8,000 data points in total, the 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.

Another way to utilize our current data is to incorporate more interaction effects of different variables. For example, if a participant assigns high importance (above some threshold to be determined) to a partner having shared interests as them, then we would compare the activites that the participant and their partner like. This can be implemented using a variation of decision trees so that if the threshold of shared interests priority is not met, the model may not need to incorporate more data to make a prediction for a positive decision or match.