Speed Dating Project

Marlene Berke and Anna Xu

Abstract

A speed-dating study in the United Kingdom generated data from more than 8000 interactions between potential romantic partners. With this data, we tried to predict when a person would want to see their partner again with only demographic and background information as features. Neither decision trees nor logistic regression produced a non-trivial solution. Popularity, on the other hand, was more predictable. A LASSO regression using demographic and background features predicted a participant's popularity on a scale of 1-10 with an out-of-sample MSE of 0.763. We found that race, gender, and career field strongly influenced popularity.

The Dataset

The data[1] captures 21 speed dating events from 2002-2004. Each event and the participants who attended it are called "waves." Each wave had only female-male pairings and no one participated more than once. During the speed dating event, participants were asked to fill out questionnaires about themselves and their partner. Each speed dating participant rated themselves on a scale from 1 to 10 on five key attributes: Attractiveness, Sincerity, Intelligence, Fun, and Ambition. They also rated their date on those same five key attributes, plus Shared Interests. Last, they rated how much they liked their partner (1-10 scale) and gave a yes-or-no decision as to whether they would like to see that partner again.

The dataset also includes demographic data on each participant, like their race, age, hometown, job, income, and hobbies. In the original dataset, each of the 8378 rows represents an interaction between two participants. There were 552 participants total, and 195 features on each interaction (including demographic information and other survey questions).

When a participant or wave did not answer a particular question, an NA was recorded. We omitted all rows that had NAs for any of the features of interest, leaving 5000-7000 interactions, depending on the question.

Practical note: from here on out, we will use second person. Although less formal, it's the clearest way of distinguishing between the individual of interest and his/her partner. "You" will refer to the individual of interest, and "your partner" will refer to your partner.

Exploratory Data Analysis

Question 1: How does your demographic, your intention for speed dating, your rating of your partner's personal attributes, and their liking of you impact how much you like them?

Overall, we found that people liked their partners enough to want to see them again 42% of the time. Females were much choosier than males, only wanting to see their partner again 36% of the time, compared to 48% for men.

We found you really liked people who you rated highly on personal attributes. In retrospect, this seems obvious: if you like someone, of course you think that they are sincere, intelligent, fun, attractive, etc. And if you think that they are sincere, intelligent, fun, attractive, etc., then you probably like them. The correlation matrix shows the high correlation between attribute rating and liking. It also shows a small (0.12) positive correlation between how much your partner liked you and how much you liked your partner.

[1] Kaggle.com: https://www.kaggle.com/annavictoria/speed-dating-experiment

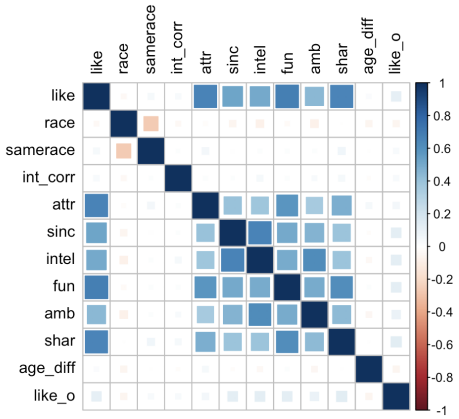

Figure 2 - Correlation Matrix of some of the features.

Another factor that was correlated with your decision was your goal in attending the speed dating event. Though few people stated their goal as "looking for a serious relationship," those who did had the highest probability of saying yes (52%) to seeing their partners again.

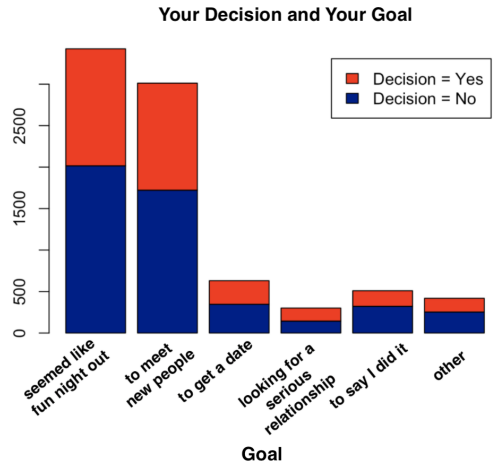

Figure 3 - Breaking down the proportion of "yeses" and "nos" by goal. The y-axis is the number of interactions. Each data point is your goal paired with your decision. Tallying up the positive and negative decisions by goal produces the above graph. Most people went because it seemed like a fun night out and usually said no. Those who went to get a date said yes most often.

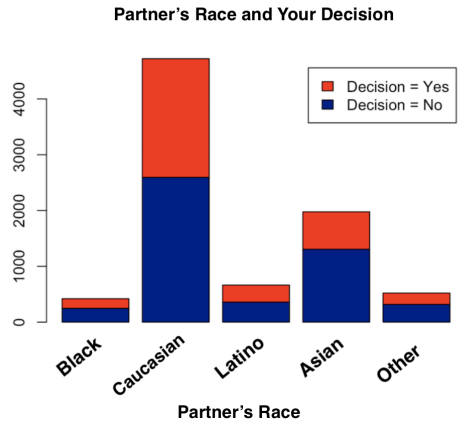

Figure 4 - Breaking down the proportion of "yeses" and "nos" by your partner's race. The y-axis is the number of interactions. Each data point is your decision paired with the race of your partner.

Race also appeared to influence decisions. In particular, Asians indicated that they would like to see their partner again 46% of the time, but their partner only wanted to see them 34% of the time. In comparison, Whites said yes to other people 39% of the time, but their partners said yes to them 45% of the time.

Other factors, such as age difference and interest correlation, didn't seem to be as strongly correlated with how much you liked someone.

Question 2: Who is popular? Who is unpopular? What are the demographics of each group?

To calculate your popularity, we averaged how much all of your partners liked you. This distribution of popularity is shown in Figure 5.

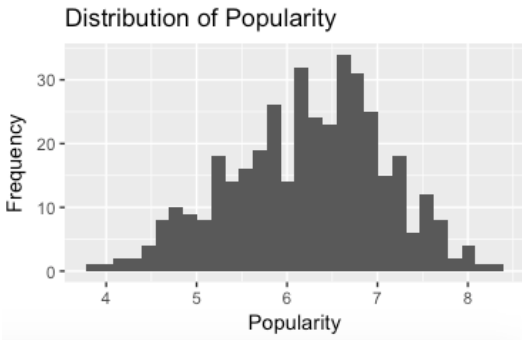

Figure 5 - Popularity looks normally distributed. The mean popularity score was 6.23 and standard deviation was 0.86.

Demographics of all participants	Female	Male
Black/African American	16	10
European/Caucasian-American	140	160
Latino/Hispanic American	24	17
Asian/Pacific Islander/Asian-American	71	64
Other	16	21

Table 1 - Demographic breakdown of the 539 participants in the study who answered all the relevant questions. 55.7% of participants were White and 25.0% were Asian.

We examined the 21 most popular people and 21 least popular people (21 because we wanted 20 but two were tied at the cutoff). The 21 most popular people were those who received an average "liking" score of 7.6 or higher (see Table 2). Of those people, only 2/21 were Asian, despite that 25% of speed dating participants were Asian (see Table 1). There's also a lot more diversity among women in the top 21 than men. Of the 11 popular men, 10 were white. Interestingly, 3 of the 10 most popular women were Latina, while only 9% of women in the study were Latina.

20 Most Popular	Female	Male
Black/African American	1	0
European/Caucasian-American	4	10
Latino/Hispanic American	3	0
Asian/Pacific Islander/Asian-American	1	1
Other	1	0

Table 2 - Demographic breakdown of the 21 most popular participants

The 21 least popular people received an average liking score below 4.7. Of the bottom 21, 15 were male, 8 of whom were Asian. When lumping the genders together, 10 of the 21 least popular speed daters were Asian, while Asians only make 25% of the speed daters. As a whole, race, with a possible interaction with gender, looks predictive of popularity, possibly benefiting Latina women and disadvantaging Asian males. Taken all and all, it appears that women might be choosier and more influenced by their partner's race when making a decision.

20 Least Popular	Female	Male
Black/African American	0	1
European/Caucasian-American	4	5
Latino/Hispanic American	0	0
Asian/Pacific Islander/Asian-American	2	8
Other	0	1

Table 3 - Demographic breakdown of the 21 least popular participants

We were also curious how field of study might affect popularity. Overall, it appears that people in English (6th from left), Business, Political Science, Film, and Fine Arts were most liked, while those in Math, Engineering, and History were least popular (Figure 6).

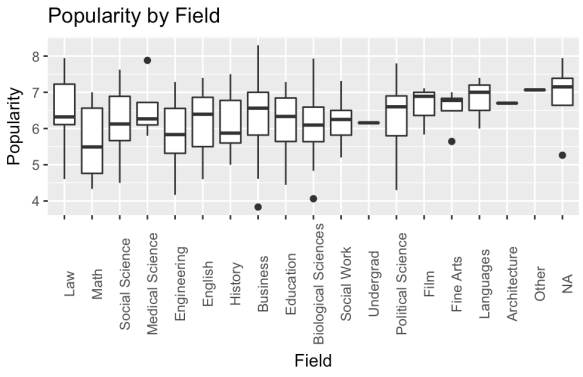

Figure 6 - Boxplots of popularity by career field/field of study.

We wondered how aware the speed-daters were of how other people viewed them, and whether that could predict popularity.

To explore this avenue, we created a self-awareness score for each person by calculating the mean difference between other people's rating of you and your rating of yourself on the 5 key attributes vs how their partners rated them on the 5 key attributes. A negative self-awareness score mean that you overrated yourself, and a positive means you underrated yourself. A score of 0 meant you were perfectly accurate. The results are summarized in Figure 7. So, on average, people rated themselves a point higher on each attribute than other people rated them.

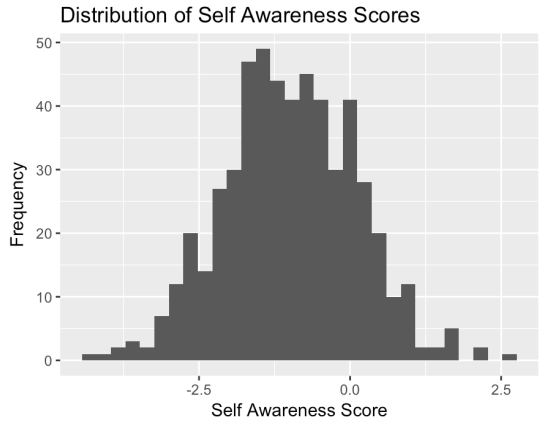

Figure 7 - The Self Awareness Scores look normally distributed. The mean self awareness score was -1.071934 and standard deviation was 1.13.

Participants were also asked how they expected that others would rate them on the 5 key attributes. We calculated an accuracy score comparing how you expected to be evaluated with how you were evaluated, using the same metrics as we used to make the self-awareness score. We call this the expected rating score. For the 313 people who answered both questions, the mean self-awareness score was -1.039492 and the mean expected ratings score was -0.8586617. A paired t-test ($p<10e-05$) showed a significant difference between how you see yourself and how you expected others to see you: on average, you think that people will rate you a bit lower than you would rate yourself. In other words, you partially correct for your overrating of yourself, but not entirely.

We explored whether demographic information like age, gender, race, and career field could predict self-awareness / veridicality of self-image. We found that Latinos and Asians had significantly more accurate self-awareness (according to a linear regression, $p<0.0244$ and 0.0133 respectively), or overestimated themselves less than other groups did. Gender, age, and profession/major had no significant effects.

Although fascinating, neither the self-awareness score or the mean expected ratings score was overly useful in predicting popularity. Because everyone almost universally ranked themselves highly, when someone had a high self-awareness score, it simply meant that their own high opinion of themselves matched up with others' opinions. Self-awareness became a proxy for popularity, and therefore not meaningful in predicting popularity.

Model Questions and Selection

We used about two thirds of the data for training, and one third for testing. We split the data so that 7 of out 21 speed-dating waves were in the test set. Splitting by wave rather than by interaction assured that the data in the training and test sets would be independent since there's no crossover between waves. Had we split by interaction rather than wave, we could have had data on the same person in both the training and test set, which would have made our test set unfairly similar to our training test. Splitting by wave does have one downside: the waves do not all have the same number of interactions in each.

Based on the results of our exploratory data analysis, we decided to modify our questions.

Question 1:

We want to predict how much you would like your partner. Based on the obvious and very strong connection between how you rate your partner on the 5 key attributes and how much you like them, we excluded the 5 key attributes from the model. Especially for applications in online dating, it is useless to know that your liking of a person is related to how attractive and intelligent you think they are. Instead, we predict liking from features we know before two people meet such as race, gender, and age, how often they go out, why they chose to come to the speed dating event, and what expectations they have for it.

Decision Tree

We first grew a decision tree using the features gender, race of you and your partner, age of you and your partner, the correlation of interests, your goal for the event, how often you go out, how often you go on a date, and how happy you expect to be with someone you meet at a speed-dating event. The result was a huge, overfit tree (Figure 8).

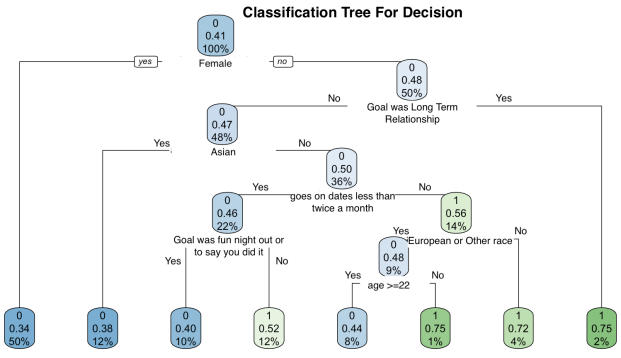

Figure 8 - Overfit Decision Tree. First splits by gender, then goal.

Using 14-fold cross-validation (14 because the training set has 14 waves), we selected the complexity parameter that minimized the error, or misclassification rate. We pruned using this very large complexity parameter. The resulting tree had no splits at all. It indicated that the best model would simply predict that a every decision would be "no." This makes some sense -- in the training set, participants did not want to see their partner again 58.9% of the time, so by always predicting no, the tree could achieve 58.9% accuracy.

There is only one factor that predicted a "yes" decision more 58.9% of the time, and by a narrow margin. This factor was goal. When your goal is "looking for a relationship," you tend to answer "yes" 60% of the time, or answer "no" only 40%. When a factor predicts a yes decision more that 58.9% of the time, then the decision tree would benefit by predicting yes for all interactions with that factor, and no for all the other. We hypothesize that, because we gave the decision tree so many features, it suffered from the curse of dimensionality, and was unable to find the helpful feature that would reduce the Residual Sum of Squares (RSS). When we gave the tree only this feature, it split based on whether or not your stated goal was "looking for a relationship." The split survived the pruning. The resulting decision tree had a classification rate of 59.2%. It is displayed below.

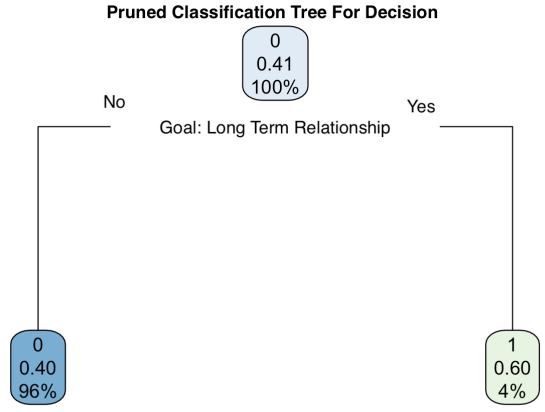

Figure 9 - Decision Tree that produced the lowest error.

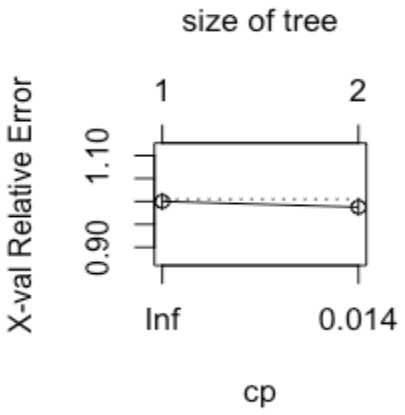

Figure 10 - Improving the size of the tree decreased the error slightly.

The classification rate on the test set was 54.5%. In the test set, participants answered "no" 55.6% of the time. So using goal as a decision factor does not help on the testing set. Our model would have performed better by simply guessing no for everyone.

Because our problem is not variance, but lack of meaningful predictive feature, we decided that bagging and random forests would not improve our model. Using ratings of attractiveness and intelligence would, but as we decided before answering this question, those predictors are uninteresting because the two partners have to meet anyway to assess those attributes.

Logistic Regression

Because of poor results from the decision tree, we decided to see if a logistic regression would be able to better predict whether or not you want to see your partner again. When determining which predictors to use, we also included interaction terms based on what we saw from the exploratory data analysis, such as an interaction between your gender and your partner's race, and an interaction between your own race and your partner's race. When modelling, we also found that in features with multiple levels, such as race and goal, some of the levels were strongly significant and others were not. We simplified these features, so they only contained 2 levels, the significant level and the rest of the levels, which improved our models. For example, we converted race of your partner with 6 levels, to race of your partner with 2 levels: Asian, or not.

From our full model we tried different subsets of features to create different models, and then once again used 14-fold cross validation. When we fit the model with few predictors, we found that the model once again continually predicted "no," which led to a 58.9% accuracy. When we fit the model with many predictors, the model would start to predict some "yeses" and "nos," but the model generalized poorly, and the cross-validation error was lower. Many of our models had cross validation accuracies a few percentage points above 50%. Ultimately, the most accurate model was one where the model always guessed no. This trivial model had a training accuracy of 58.9% and a test accuracy of 55.6%. In other words, 55.6% of people in the test set said "no." The second best model had a test accuracy of 54.44%. This model predicted decisions based on your gender, how often you go on dates, your goal, how happy you expect to be with the people you meet during the event, the age of your partner, if your partner is Asian, and the interest correlation between you and your partner.

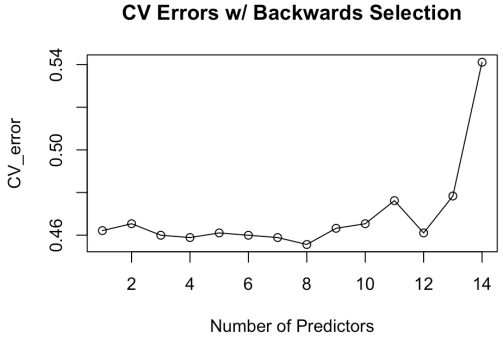

Figure 11 - The 14-fold cross validation error rates

Overall, the model's performance was poor, but considering that we are trying to predict whether or not someone will want to see you again based on demographic, how often you go on dates and not on personality whatsoever, we are reassured that the speed daters are choosing their partners for more meaningful reasons!

Question 2:

Can we predict how popular you are given demographic data, and your goals and expectations for speed dating?

We chose a linear model with the LASSO regularizer to encourage sparsity and remove the less-predictive features. We examined the following features: gender, age, race, pairwise interactions between the three, career field, goals, expectations about happiness with a partner, how often they went out, and how often they went on dates.

We used cross-validation to select the tuning parameter λ. Figure 12 shows the MSE from cross-validation against different values of λ. It achieves a minimum around $\lambda = 0.0305$.

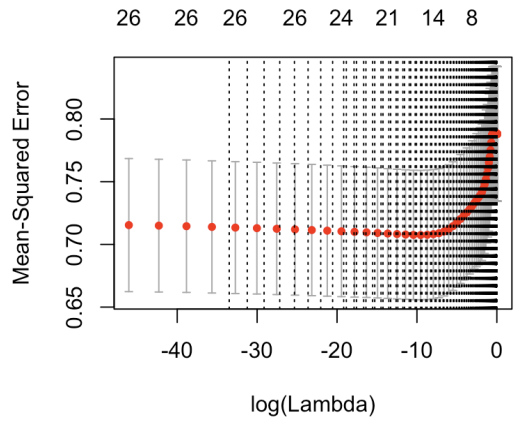

Figure 12 - The Mean Squared Error as we tuned λ

Nine of the 36 parameters were zeroed out. The intercept was 6.95, a bit higher than the mean popularity of 6.06. As we would have expected from our exploratory data analysis, the interaction of Asian and male had a coefficient of -0.15, while on their own, Asian and males both had negative coefficients, although they were very small (-0.09) and (-0.08), respectively. Latinos had a relatively large, positive coefficient (0.20). The 6 largest, non-zero coefficients were in the field of medicine (0.40), mathematics (-0.25), Latino (0.20), field of engineering (-0.16), Asian males (-0.15), and political science / international affairs (0.15). We are confident in the effect of these features, as they are consistent with our exploratory data analysis.

This LASSO regression predicts popularity reasonably well. The MSE over the test set is 0.763. In comparison, had we just used the mean popularity to predict popularity, our MSE would have been 0.893. Using a multitude of predictors and the intercept reduced the out-of-sample error by 0.1. To some extent, demographic information can predict a participant's popularity during speed dating.

Conclusion and Discussion

Could our insights be put to good use? Unfortunately, predicting whether you would want to see your partner again based solely on background information is infeasible. If we had used your ratings of your partner's attributes, then it would have been uninteresting to dating services because you would have to interact with your partner in order to judge those attributes. We were more interested in what could be predicted from objective information that you can gather on a survey.

Perhaps with other features relating to a participant's values and personality, our modeling attempts might have found more

success. Something like the Meyer-Briggs personality test, information on religion, social class identification, and past dating history might have been more telling. It is also regrettable that income data was collected on so few participants.

Popularity was easier to predict. In fact, we discovered some interesting trends. Those in the medical field are well liked (wealth could be a confound). People with careers in math and engineering were less popular, even when accounting for race and gender. Most strikingly, popularity had a strong basis in race. In particular, Latinos were found to be popular, while and interaction between gender and Asian disadvantaged Asian males.

We would be willing to use our results to change dating services so as level the playing field. Our biggest finding was that people (especially women) are influenced by race when looking for dating partners and that the prospects of Asian males suffer the most because of it. However, our findings only relate to a first impression formed in four minutes, not over the course of getting to know someone. Perhaps, after longer periods of interaction, racial bias diminishes. Our results have no bearing on extended interactions like dates or relationships.

Suppose our company wanted to increase matching rates. It could predict a new client's popularity based on the features we outlined, and then show that profile to lots of the opposite gender. This would result in a rich-get-richer and poor-get-poorer scheme, in which people predicted to do well get a boost, while others get further setback. We would not support the use of our research in that capacity.

While machine-made decisions are sometimes seen as objective and fairer than human decisions, this isn't always the case. For example, a classifier trained to decide who gets a loan and who does not might base its decisions on race, overtly or subvertly (as in by zipcode). The EU has taken steps to prevent harmful effects of machine learning, outlined in the 2017 Report on with recommendations to the Commission on Civil Law Rules on Robotics. [2] Our findings represent a microcosm of this issue: if our results were misused by dating companies, it could have discriminative results.

Instead, perhaps our company could use this information to make the dating scene more equitable. It is a well-established phenomenon that people are generally better at recognizing and remembering faces of their own race that of others, constituting a type of racial bias. Recently, it has been shown that training participants to distinguish between faces of another race reduces implicit racial bias.[3] Perhaps, before viewing the profiles of potential dates, our website/app could have its users practice distinguishing between different faces of the opposite race. After such practice, users might be more open to dating other races. If our app especially encourages non-Asian women to do this exercise (perhaps by displaying the task as soon as the app is opened), it might reduce the bias against Asian men. In short, our results could be used to transform the dating scene for the better.

---

[2]Report on with recommendations to the Commission on Civil Law Rules on Robotics, Mady Delvaux (2017): http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//EN

[3]Lebrecht S, Pierce LJ, Tarr MJ, Tanaka JW (2009) Perceptual Other-Race Training Reduces Implicit Racial Bias. PLOS ONE 4(1): e4215.https://doi.org/10.1371/journal.pone.0004215