Data Science and Business Intelligence

Speed Dating

Group Project Report

Section

BU. 330. 780. 52

Group Members:

Hang Zhao, Qinya Chen, Siqi Zhao,

Shuqiao Li, Xingyu Gu, Yanshu Liu

May 9, 2019

I. Background & Problem Statement

Compared to the 4.4% annual growth rate of the dating service industry, casual dating only takes a 2% annual growth rate. In such a fast-growing lucrative industry ($3,744 million in 2018)[1], casual speed dating faces fierce competition from online dating and matchmaking services. In addition, speed dating shares the least average revenue per paying user. To provide better speed dating services and attract more paying users, the speed dating company requires our team to provide a detailed data analysis report.

In this project, we will simulate several models which can be used to predict what kind of participants are more likely to find a match and predict the probability of "match" between randomly selected two participants. From the research and empirical experience, we found out 24 independent variables and categorized them into demographics information and dating survey-based information.

Problems to solve:

1. Which kind of participants are more likely to find a match?

2. How to predict the probability of "match" between randomly selected two participants?

II. Data Description

1. Variable selection

Our data was gathered from participants in experimental speed dating events from 2002-2004 and compiled by Columbia Business School. The response variable is **match (binary variable),** and independent variable are divided into two categories as follows:

Demographics data: gender, race, age, from, field_cd,

Dating survey-based information: iid, pid, int_corr, samerace, imprace, goal, date, go out, match_es, attr_m, sinc_m, intel_m, fun_m, amb_m, shar_m, like_m, max1, max2. For the detailed description, please refer to the Appendix, Figure 1.

2. Data processing

In order to make it easier for building the model, we formatted all places and replaced partner-rated attributes score ("attr_o","sinc_o") with the mean ("attr_m","sinc_m"). Besides, we calculated the most valued attribute and the second important attribute of female and male, and renamed them with max1 and max2.

III. Descriptive Analysis and Data Visualization

To better understand the data and select the most appropriate model, we did some descriptive analysis and data visualization to find out what factors or attributes may affect the match success rate.

1. Interest correlation vs match

We made two boxplots of female and male (Appendix, Figure 2. & 3.) to find the relationship between interest correlation and success match. We can tell that interest correlation has a significant effect on match or not. Higher correlation tends to have a higher match probability.

2. Samerace vs. match & gender

We also made a plot (Appendix, Figure 4.) to see whether the same race will affect the match success rate. It shows that same race or not may don't affect the match success rate.

3. Age vs match & gender

Through the plot (Appendix, Figure 5.) we made, there is no obvious difference of theage distribution between match or not match. Therefore, we made a regression model to verify our conclusion. The results show age has no effect on match or not.

4. The difference between what are woman and man looking for

We can see that there is a great difference between what male and female participants are looking for (Appendix, Figure 6.):

- For male participants, the attractiveness of a female is given a lot more weight, and the ambitiousness or if they have any shared interest are ranked not as high.

- For females, the points are more evenly distributed across all the attributes, with intelligence ranked slightly higher compared to others.

To sum up, men are looking for attractive women, and less concerned about ambition and shared interests. On the other hand, women are looking for a well-rounded male and value intelligence in a man.

5. Male/Female's stated interest compared to actual influence on decision

We want to know whether people really know what they want, so we compared what people want as stated before the event and what influences their decisions. We can tell there are dramatic differences between the dreams and reality (Appendix, Figure 7.)

Males still attach much importance to attractiveness, but they underestimate the influence of shared interest and fun scores for the female. Besides, male don't care as much as they expected

about sincerity and intelligence of a female as they thought, as these personalities do not contribute so much to men's decision making.

While females' stated interests are far away from actual influence of these attributes. They underestimate the power of attractiveness, shared interest, and fun, while overestimate sincerity, intelligence and ambitiousness.

IV. Demography Analysis

Based on the previous descriptive analysis, we will conduct the further analysis into the relationship between successful match and demography attributes, to see the effect of some personal information.

We separate the demographic analysis into three steps and interpret the results to draw some conclusions:

1. Clean the data and select variables:

In our dataset, a participant will meet several different people and receive different comments records but with the one and only iid, so we delete the repetitive records based on the iid to keep only one record of personal information.

In the following analysis, we will use "match" as target binary variable and the demographic indicators in this case include race, the importance of same rate, age, field of career,

participant goal, date frequency and go out frequency, which include some dating survey-based information.

2. Convert the data types and conduct logic regressions:

To get a whole idea of the demographic analysis, we set "match" as factor type, and conduct a general logic regression on all the variables firstly. From the regression results, we find out that only the date frequency and go out frequency are statistically significant variables.

Then we convert the date frequency and go out frequency variables as factor type and run logic regression to each set of variables again, to see which kind of participants will be easier to find a match. The result shows that Date 5, 6 and 7 and Go out 1 are significant. Take the meaning of variables into consideration, the result tells us that the participants with lower date frequency (once a month, several times a year and almost never) are hard to find a match while those with higher go out frequency (several times a week) are easier to find a match in this event.

3. Plot the Classification tree and analyze the result:

The CP table of the classification tree indicates a big problem of overfitting if we use cp as the control parameter, so we turn to the max depth. To plot and interpret the tree clearly, we set the max depth as 5 and the results are like the logic regression (Appendix, Figure 8).

The first node is the date frequency, date = 3, 5, 6, 7 would provide the most information gain at the first step. Then the successful match groups are the people with higher date frequency, consider race is not a big problem (< 6/10) and younger than 26 and with some other restraints. For example, if a participant works in one of the following fields, he or she will have 80% chance to find a match: math, social science, medical science, history, education, undergraduate /undecided, political science, languages and architecture. If he or she doesn't work in any of those fields, but not a Black/African American nor a European/Caucasian-American, he or she still have

more than 50% chance to find a match.

V. Random Match Model

After analyzing which attributes can increase the possibility to get matched, we also want to take a step further, building up a model to predict whether two daters would be a match given their demographic information.

1. Modify the dataset:

To build up the model, we first must modify the dataset. Instead of using each person as input objects, we recognize each speed date as input objects. To create the object, we use two participants' id in each date to find their demographic data and also the match result as dependent variable. Also, we create a new variable to see whether these two people have the same most importance evaluated attributes. With the new dataset, we begin to build up prediction model by classification tree.

2. Build a classification tree of prediction model:

First of all, we conduct a training/test split at 80/20 for further validation. Then, we build the classification tree model using the training set, picking cp=0 to print out the full tree. From the full tree, variables used in tree construction include the attributes of two people and their occupation. That glimpse gives us some idea about how the tree would look like.

But the cp table (Appendix, Figure 9) shows that the best *nsplit* is 0 when the *xerror* is the minimum, where the classification tree is a node. By using `plotcp()`, we check how the cross-validation error rate changes as the complexity of the model increases (Appendix, Figure 10). From the plot, we choose the tree with 12 splits which is also within the range of standard deviation of the minimum error rate.

After pruning the tree, we create the final classification tree for random match model (Appendix, Figure 11). In this model, it shows that basically the match success probability is determined by two daters' personal attributes, such as *funny, likely, sincerity,* and *shareable,* instead of one's age or race. That's an interesting result, illustrates that the love is more about personalities rather than personal biological background.

3. Choose a best threshold:

Also, we want to choose a best threshold to complete our prediction model. So, we repeat threshold from 0.10 to 0.90 and predict the match in test data. After comparing the test result with the actual result, we find that when threshold is 0.5, the prediction has highest accurate rate, at 81.94% satisfyingly. Also, we plot the ROC where AUC reaches 0.593 (Appendix, Figure 12).

VI. Conclusion and Limitation

From the logic regression of demography analysis, the result tells us that the participants with lower date frequency are hard to find a match while those with higher go out frequency are easier to find a match in the speed dating. Besides, the classification tree provides a similar result, the successful match groups are the people with higher date frequency, consider race is not a big problem and younger than 26 and with other constraints. For example, if a participant works in any of the following fields: math, social science, medical science, history, education, undergraduate/undecided, political science, languages and architecture. If he or she doesn't work

in any of those fields, but not a Black/African American nor a European/Caucasian-American, he or she still have more than 50% chance to find a match.

From the random match model, we find that instead of considering people's biological background, such as age and race, the classification tree gives more weight on personalities such as *funny, sincerity* and *shareable.* Also, with the training/test split, our model is evaluated and has an accurate rate at 81.94%.

Admittedly, there are some limitations in our analysis and model.

First of all, as an important demographic information, citizenship is excluded given that that data is unstructured in the raw dataset. The inputs in this variable are hugely different, varying from country name to state abbreviation, so it's extremely difficult to structure this piece of data into our analysis and we determine to abandon it.

Also, the heterogeneity of sample is limited. As the experiment took place in university, most participants are students with similar age and education level. So, we recommend further study to emphasize on the selection of participants for heterogeneity assurance.

Besides, when conducting the classification tree, the preliminary result is not quite satisfying, as the best model is when the tree only contains one node. Also, the final tree model we choose is reasonable, there still place to modified. With better data processing and raw data collection, we believe the tree can be improved.

VII. Reference

1. eServices Report 2019 - Dating Services (2018). Retrieved from https://www-statista-com.proxy1.library.jhu.edu/study/40456/dating-services-report/
2. Speed Dating Experience (2016). Retrieved from https://www.kaggle.com/annavictoria/speed-dating-experiment

VIII. Appendix

Figure 1. Variable Selection and Description

Demographics Data	
gender	Female=0, Male=1
race:	Black/African American=1 European/Caucasian-American=2 Latino/Hispanic American=3 Asian/Pacific Islander/Asian-American=4 Native American=5 Other=6
age	
from	Where are you from originally (before coming to Columbia)?
field_cd:	1= Law 2= Math 3= Social Science, Psychologist 4= Medical Science, Pharmaceuticals, and Bio Tech 5= Engineering 6= English/Creative Writing/ Journalism 7= History/Religion/Philosophy 8= Business/Econ/Finance 9= Education, Academia 10= Biological Sciences/Chemistry/Physics 11= Social Work 12= Undergrad/undecided 13=Political Science/International Affairs 14=Film 15=Fine Arts/Arts Administration 16=Languages 17=Architecture 18=Other
Dating survey-based information	
iid	unique subject number, group(wave id gender)
pid	partner's iid number
Int_corr	correlation between participant's and partner's ratings of interests in Time 1
samerace	participant and the partner were the same race. 1= yes, 0=no
Imprace	How important is it to you (on a scale of 1-10) that a person you date be of the same racial/ethnic background?
goal	What is your primary goal in participating in this event? Seemed like a fun night out = 1 To meet new people=2 To get a date=3 Looking for a serious relationship=4 To say I did it=5 Other=6
date	In general, how frequently do you go on dates? Several times aweek=1 Twice a week=2 Once a week=3 Twice a month=4 Once a month=5 Several times ayear=6 Almost never=7
go out	How often do you go out (not necessarily on dates)? Several times aweek=1 Twice a week=2 Once a week=3 Twice a month=4 Once a month=5 Several times ayear=6 Almost never=7
match_es	How many matches do you estimate you will get (a match occurs when you and your partner both check "Yes" next to decision)?
attr_m	The mean of this person's grade for Attractiveness after first date
sinc_m	The mean of this person's grade for Sincere after first date
intel_m	The mean of this person's grade for Intelligent after first date
fun_m	The mean of this person's grade for Fun after first date
amb_m	The mean of this person's grade for Ambitious after first date
shar_m	The mean of this person's grade for Shared Interests/Hobbies after first date
like_m	The mean of how much this person was liked by people she/he dated
max1	Among 6 attributes, the attribute this person care about the most
max2	Among 6 attributes, the attribute this person thinks the second most important

Figure 2. Interest Correlation vs. Match (Female)

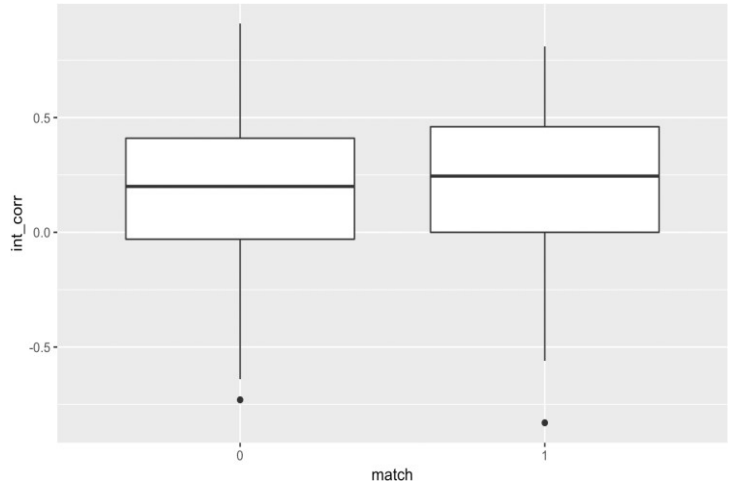

Figure 3. Interest Correlation vs. Match (Female)

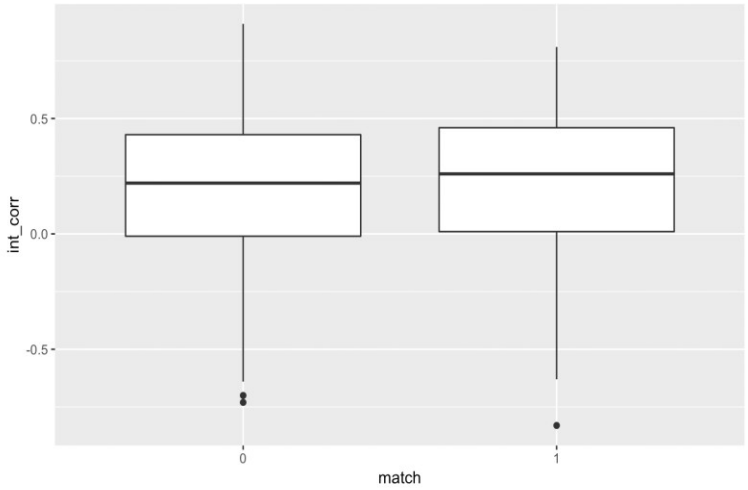

Figure 4. Same Race vs Match & Gender

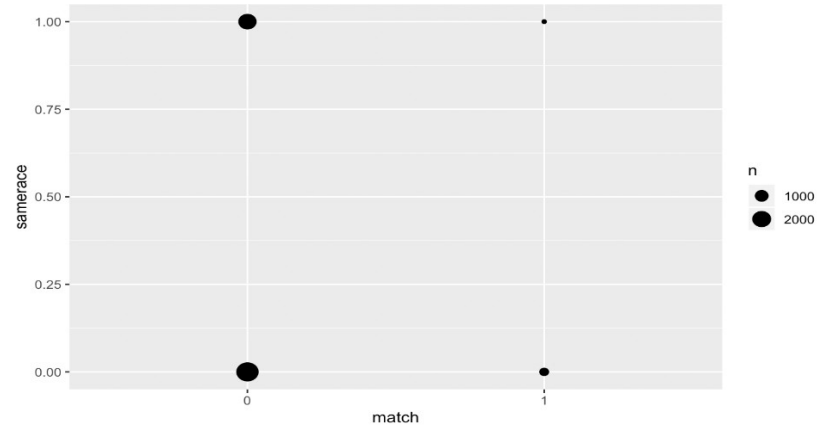

Figure 5. Age vs Match & Gender

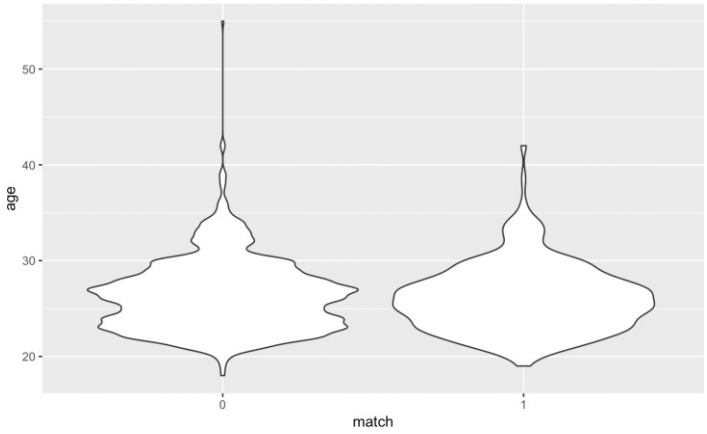

Figure 6. The Difference Between What Are Women and Men Looking For

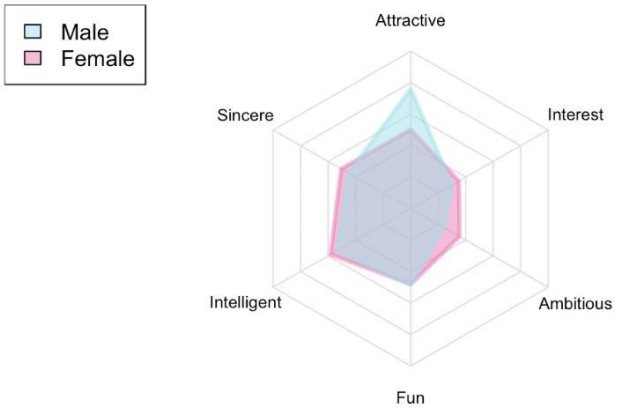

Figure 7. Male/Female's Stated Interest Compared to Actual Influence on Decision

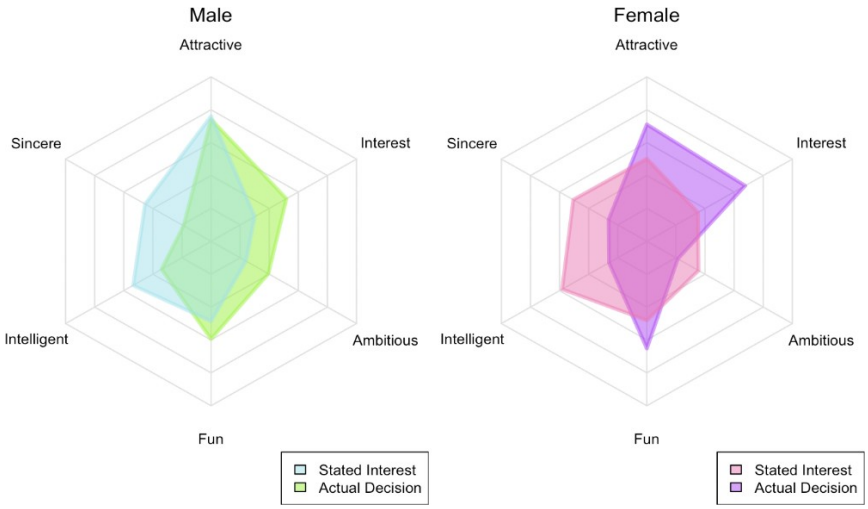

Figure 8. The Classification Tree of Demography Analysis (maxdepth = 5)

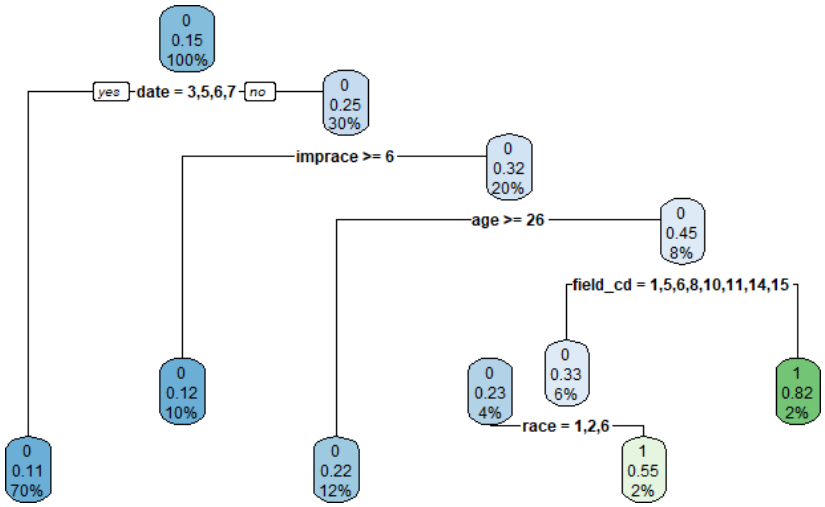

Figure 9. The CP Table for the Classification Tree of the Random Match Tree Model

	CP	nsplit	rel error	xerror	xstd
1	0.00829493	0	1.00000	1.0000	0.052144
2	0.00645161	12	0.84194	1.0774	0.053733
3	0.00430108	16	0.81613	1.0839	0.053860
4	0.00322581	19	0.80323	1.1065	0.054302
5	0.00161290	47	0.68387	1.2419	0.056785
6	0.00053763	53	0.67419	1.3355	0.058344
7	0.00000000	59	0.67097	1.3548	0.058652

Figure 10. The CP Plot for the Classification Tree of the Random Match Model

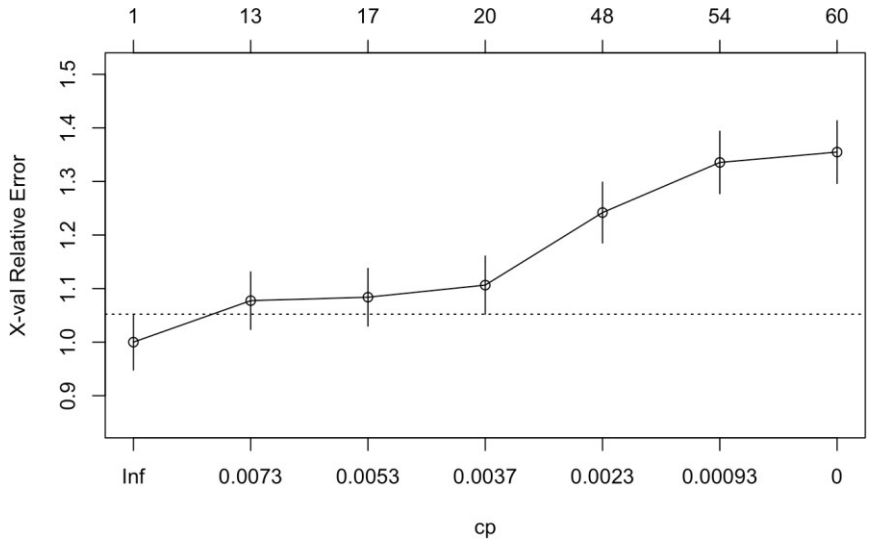

Figure 11. Classification Tree for Random Match Model

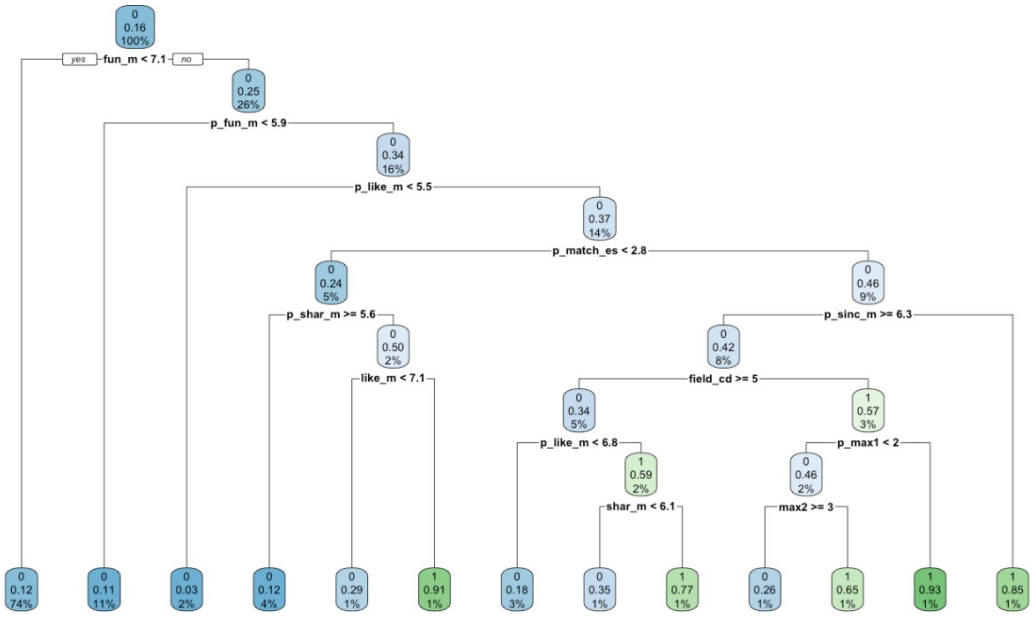

Figure 12. ROC Plot for Random Match Model