Final draft for CMPUT 566 Mini Project

Speed Dating Matching

Name: Yuhan Ye

Student No: 1463504

Data: November 26, 2019

Introduction

Online dating sites have become popular platforms for people to seek love. Many online dating sites now provide recommendations on compatible partners based on algorithms that take users' profile and interests as parameters. The purpose of this project is to find out a good model to predict whether two individuals can successfully match.

This project includes three different machine learning algorithms: Neural Network, Random Forest and Gradient Boosted Tree. They will be tested on the real-world data.

Finally, evaluate three algorithms to find out which one works the best.

Description of the Data set

The data is a collection of surveys taken from many different waves of people. Each respondent gave his basic information and ratings of other people they met. There are total of 8738 samples and 169 attributes. Here are a few attributes,

go out:

How often do you go out (not necessarily on dates)?

Several times a week=1

Twice a week=2

Once a week=3

Twice a month=4

Once a month=5

Several times a year=6

Almost never=7

race:

Black/African American=1

European/Caucasian-American=2

Latino/Hispanic American=3

Asian/Pacific Islander/Asian-American=4

Native American=5

Other=6

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?

The value of each attribute mainly has two forms.

1. Coded representations of the attribute, like go_out attribute above
2. Rate on a scale of 1-10 of the attribute, like imprace attribute above

There is also a column “match”. The value 0 means fail and 1 means success.

See Appendix for all information of the survey and each attribute.

The data set is not sorted at all — there are lots of “missing data” since most respondents didn’t fill out the all the information for the survey.

Some attributes are irrelevant,e.g.

round: number of people that met in wave

position: station number where met partner

Since the data set is not perfect, we need to clean the data. See Design of Experiments for details.

Learning Approaches

Three learning algorithms are chosen for this project: Neural Network(NN), Random Forest(RF) and Gradient Boost Tree(GBT).

- Neural Network:

Neural Network has very good performance for most of classification problems. In this project Adam is used since Adam is robust and commonly used nowadays.

Adam MLP Learner with parameter settings below:

Maximum number of iterations: 1000

Number of hidden layers:2/4/8

Number of hidden nodes:4/8/10

Learning rate: 0.01/0.05/0.1

- Random Forest:

Random Forest is a method based on tree search. It is a resemble of Decision Tree, using random feature selections from bootstrap training samples. The predictor is built by averaging the outputs of all Decision Trees. RF provides a reliable feature importance estimate, which is essential for this project and this is why I chose this approach.

For RF Learner, the parameter settings are below:

Split criterion: Information gain ratio

Number of models of the forest: 50/100/200

No limit for tree depth and child node size for the tree

- Gradient Boosted Tree:

Gradient Boosted Tree is also a method based on tree search, and is a resemble of Decision Tree. The difference

between GBT and RF is the way the trees are built. GBT build trees one at a time, where each new tree helps to correct errors made by previously trained tree. It has been shown that GBT performs better than RF if parameters tuned carefully and I considered that GBT shall works very well for this project.

For GBT Learner, the parameter settings are below:

Tree depth: 4/8/no limit

Number of models: 50/100/200

Learning rate: 0.01/0.05/0.1

Design of Experiments

All the experiment reported in this study were conducted using KNIME, the Konstanz Information Miner, a free and open-source data analytics, reporting and and integration platform.

Workflow of the project

1. Read data set file
2. Data manipulation

(1) Let 'match' be target column and others as attributes.

(2) String to number, coding representation of some attributes, whose values are string.

(3) Use column filter to manually select relevant attributes.

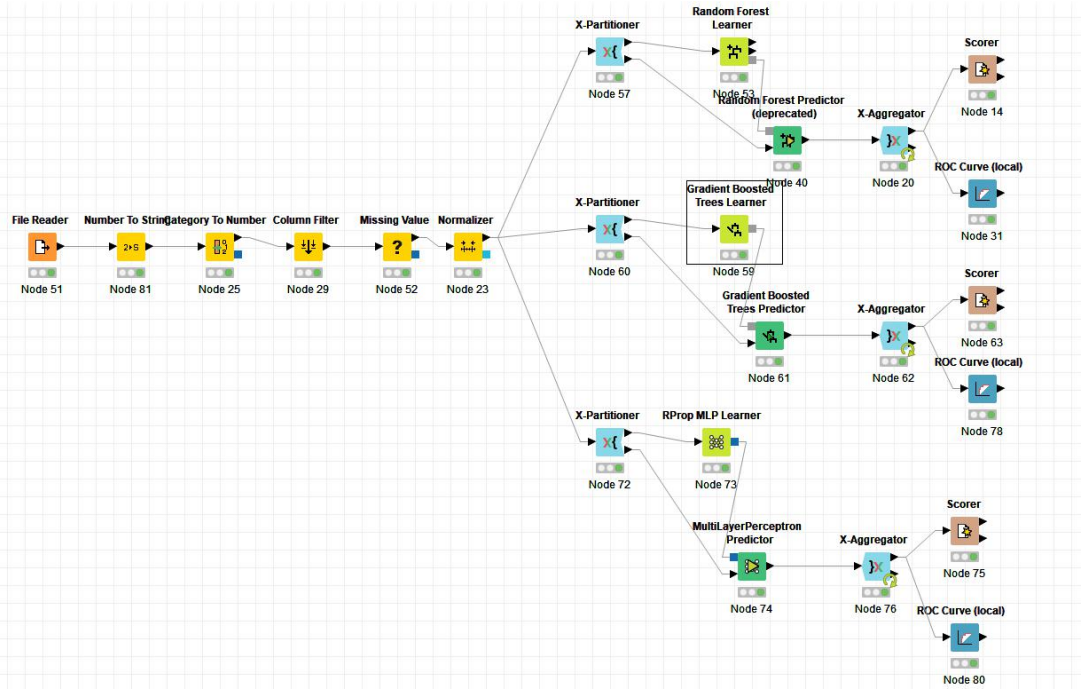
(4) Use mean value to fill in missing data.

(5) Gaussian Normalization. All attributes were standardized to zero mean and one standard deviation.

3.Cross-validation for each algorithm.Manually choose different meta-parameters to improve performance and get best meta-parameters that give minimum error.

4. Scorer and ROC analysis for each algorithm.

5. In order to compare all three algorithms, Anova test(with significance level 0.05 is used for significance **test**



Result

The best meta-parameters tested for each algorithm in this project are below:

Random Forest:

Number of models: 200

Gradient Boosted Tree:

Tree depth: not limited

Number of models: 200

Learning rate: 0.1

Neural Network:

Number of nodes: 10

Number of layers: 4

Learning rate: 0.01

The error rate, scorer table including correct classified, wrong classified, accuracy, error rate and Cohen's kappa for each algorithm are shown below:

Random Forest:

Row ID	D Error in %	I Size of Test Set	I Error Count
fold 0	17.243	1676	289
fold 1	17.542	1676	294
fold 2	17.134	1675	287
fold 3	16.289	1676	273
fold 4	15.164	1675	254

error rate of Random Forest

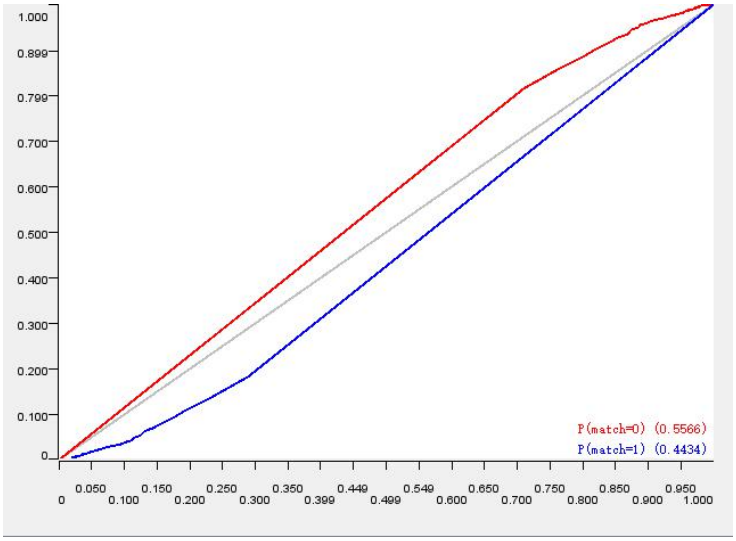
match \ Prediction (RF)	0	1
0	6943	55
1	1342	38

Correct classified: 6,981 Wrong classified: 1,397

Accuracy: 83.325 % Error: 16.675 %

Cohen's kappa (K) 0.031

Scorer table of Random Forest



ROC curve of Random Forest

Gradient Boosted Tree:

Row ID	Error in %	Size of Test Set	Error Count
fold 0	17.243	1676	289
fold 1	17.184	1676	288
fold 2	17.015	1675	285
fold 3	16.11	1676	270
fold 4	15.164	1675	254

error rate of Gradient Boosted Tree

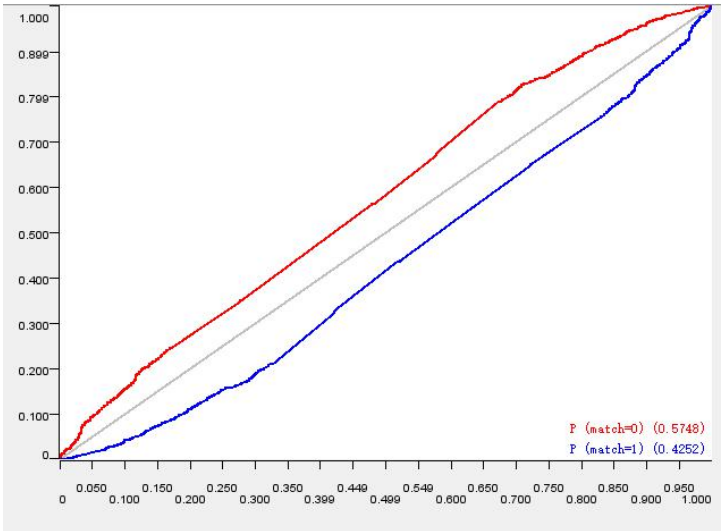
match \ Prediction (gbt)	0	1
0	6960	38
1	1348	32

Correct classified: 6,992 Wrong classified: 1,386

Accuracy: 83.457 % Error: 16.543 %

Cohen's kappa (K) 0.029

Scorer table of Gradient Boosted Tree



ROC curve of Gradient Boosted Tree

Neural Network:

Row ID	Error in %	Size of Test Set	Error Count
fold 0	17.064	1676	286
fold 1	17.184	1676	288
fold 2	16.358	1675	274
fold 3	17.363	1676	291
fold 4	17.134	1675	287

error rate of Neural Network

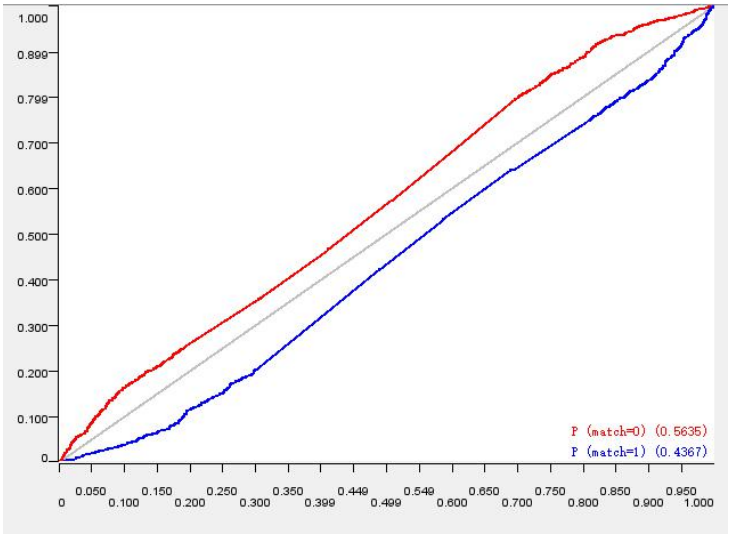
match \ Prediction (nn)	0	1
0	6907	91
1	1335	45

Correct classified: 6,952 Wrong classified: 1,426

Accuracy: 82.979 % Error: 17.021 %

Cohen's kappa (K) 0.031

Scorer table of Neural Network



ROC curve of Neural Network

From the results above, we can find that Gradient Boosted Tree is the best performing model with a 83.457% accuracy on our data set while Neural Network performs the worst with a 82.979% accuracy. Since the data set is imbalance with too many match 0 and small amount of match 1, the errors tend to be a false negative. While for this type of data would most likely to be used by a dating service, it would be fine for the dating service to not match you with someone you would have good with.

Anova Test:

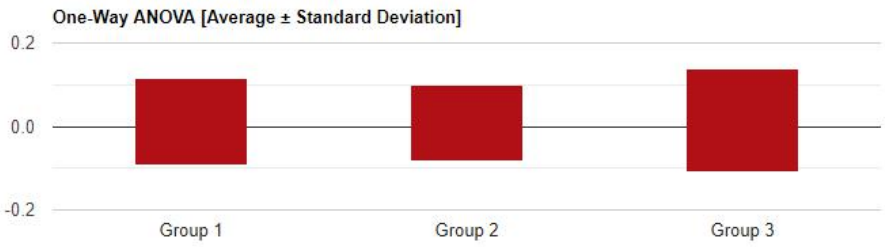
Analysis of Variance Results

F-statistic value = 9.34554

P-value = 0.00009

Data Summary				
Groups	N	Mean	Std. Dev.	Std. Error
Group 1	8378	0.0111	0.1048	0.0011
Group 2	8378	0.0084	0.091	0.001
Group 3	8378	0.0155	0.1236	0.0014

ANOVA Summary					
Source	Degrees of Freedom DF	Sum of Squares SS	Mean Square MS	F-Stat	P-Value
Between Groups	2	0.2152	0.1076	9.3455	0.0001
Within Groups	25131	289.35	0.0115		
Total:	25133	289.5652			



From the results of our Anova test, we can find that the p-values are far less than the set threshold 0.05. Small p-value suggests that with strong statistical evidence, the means of all input distributions are different. The performance results of three learning approaches are significantly different.

Conclusion

Matching users with mutual interest in each other is an important task for online dating sites. In this project, three different machine learning algorithms are used: Random Forest, Gradient Boosted Tree and Neural Network to try to predict whether two individuals could successfully match. The data is a collection of surveys taken from many different waves of people. On this data set, Gradient Boosted Tree method outperforms the other two algorithms(Random Forest and Neural Network) with a 83.457% accuracy. The methods and results can provide valuable guidelines to the design and performance of matching systems for dating services.

Appendix

Speed Dating Data Key

iid: unique subject number, group(wave id gender)

id: subject number within wave

gender: Female=0

Male=1

idg: subject number within gender, group(id gender)

condtn:

1=limited choice

2=extensive choice

wave:

Wave #	Date	Preference Scale	Variations	# Males	# Females
1	October 16th '02	100 pt alloc.		10	10
2	October 23rd '02	100 pt alloc.		16	19
3	November 12th '02	100 pt alloc.		10	9
4	November 12th '02	100 pt alloc.		18	18
5	November 20th, '02	100 pt alloc.	undergrads	10	10
6	March 26th '03	1-10 scale		5	5
7	March 26th '03	1-10 scale		16	16
8	April 2nd '03	1-10 scale		10	10

9	April 2nd '03	1-10 scale		20	20
10	September 24th '03	100 pt alloc.		9	9
11	September 24th '03	100 pt alloc.		21	21
12	October 7th '03	100 pt alloc.	Budget: only allowed to yes yes to 50% of the people that met	14	15
13	October 8th '03	100 pt alloc.	Different M.C.	9	10
14	October 8th '03	100 pt alloc.	Different M.C.	18	20
15	February 24th '04	100 pt alloc.		19	18
16	February 25th '04	100 pt alloc.		8	6
17	February 25th '04	100 pt alloc.		14	10
18	April 6th '04	100 pt alloc.	brought a magazine	6	6
19	April 6th '04	100 pt alloc.	brought a book	15	16
20	April 7th '04	100 pt alloc.	brought a book	8	6
21	April 7th '04	100 pt alloc.	brought a magazine	22	22

round: number of people that met in wave

position: station number where met partner

positin1: station number where started

order: the number of date that night when met partner

partner: partner's id number the night of event

pid: partner's iid number

match 1=yes, 0=no

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

age_o: age of partner

race_o: race of partner

pf_o_att: partner's stated preference at Time 1 (attr1_1) for all 6 attributes

dec_o: decision of partner the night of event

attr_o: rating by partner the night of the event, for all 6 attributes

signup/Time1:

[Survey filled out by students that are interested in participating in order to register for the event.]

age:

field: field of study

field_cd: field coded

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

undergrd: school attended for undergraduate degree

mn_sat: Median SAT score for the undergraduate institution where attended. Taken from Barron's 25th Edition college profile book. Proxy for

intelligence.

tuition: Tuition listed for each response to undergrad in Barron's 25th Edition college profile book.

race:

Black/African American=1

European/Caucasian-American=2

Latino/Hispanic American=3

Asian/Pacific Islander/Asian-American=4

Native American=5

Other=6

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?

imprelig:

How important is it to you (on a scale of 1-10) that a person you date be of the same religious background?

from:

Where are you from originally (before coming to Columbia)?

zipcode:

What was the zip code of the area where you grew up?

income:

Median household income based on zipcode using the Census Bureau website: http://venus.census.gov/cdrom/lookup/CMD=LIST/DB=C90STF3B/LEV=ZIP

When there is no income it means that they are either from abroad or did not enter their zip code.

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 a week=1

Twice a week=2

Once a week=3

Twice a month=4

Once a month=5

Several times a year=6

Almost never=7

go out:

How often do you go out (not necessarily on dates)?

Several times a week=1

Twice a week=2

Once a week=3

Twice a month=4

Once a month=5

Several times a year=6

Almost never=7

career:

What is your intended career?

career_c: career coded

1= Lawyer

2= Academic/Research

3= Psychologist

4= Doctor/Medicine

5=Engineer

6= Creative Arts/Entertainment

7= Banking/Consulting/Finance/Marketing/Business/CEO/Entrepreneur/Admin

8= Real Estate

9= International/Humanitarian Affairs

10= Undecided

11=Social Work

12=Speech Pathology

13=Politics

14=Pro sports/Athletics

15=Other

16=Journalism

17=Architecture

12. How interested are you in the following activities, on a scale of 1-10?

sports: Playing sports/ athletics

tvsports: Watching sports

excersice: Body building/exercising

dining: Dining out

museums: Museums/galleries

art: Art

hiking: Hiking/camping

gaming: Gaming

clubbing: Dancing/clubbing

reading: Reading

tv: Watching TV

theater: Theater

movies: Movies

concerts: Going to concerts

music: Music

shopping: Shopping

yoga: Yoga/meditation

exphappy:

Overall, on a scale of 1-10, how happy do you expect to be with the people you meet during the speed-dating event?

expnum:

Out of the 20 people you will meet, how many do you expect will be interested in dating you?

We want to know what you look for in the opposite sex._

Waves 6-9: Please rate the importance of the following attributes in a potential date on a scale of 1-10 (1=not at all important, 10=extremely important):

Waves 1-5, 10-21: You have 100 points to distribute among the following attributes -- give more points to those attributes that are more important in a potential date, and fewer points to those attributes that are less important in a potential date. Total points must equal 100.

Attractive	+
Sincere	+
Intelligent	+
Fun	+
Ambitious	+
Shared Interests	+

attr1_1

Attractive

sinc1_1

Sincere

intel1_1

Intelligent

fun1_1

Fun

amb1_1

Ambitious

shar1_1

Has shared interests/hobbies

Now we want to know what you think MOST of your fellow men/women look for in the opposite sex.

Waves 6-9: Please rate the importance of the following attributes on a scale of 1-10 (1=not at all important, 10=extremely important):

Waves 10-21 : You have 100 points to distribute among the following attributes -- give more points to those attributes that you think your fellow men/women find more important in a potential date and fewer points to those attributes that they find less important in a potential date. Total points must equal 100.

attr4_1

Attractive

sinc4_1

Sincere

intel4_1

Intelligent

fun4_1

Fun

amb4_1

Ambitious

shar4_1

Shared Interests/Hobbies

What do you think the opposite sex looks for in a date?

Waves 6-9: Please rate the importance of the following attributes on a scale of 1-10 (1=not at all important, 10=extremely important):

Waves 1-5 and 10-21: Please distribute 100 points among the following attributes -- give more points to those attributes that you think are more important to members of the opposite sex when they are deciding whether to date someone. Total points must equal 100.

attr2_1

Attractive

sinc2_1

Sincere

int2_1

Intelligent

fun2_1

Fun

amb2_1

Ambitious

shar2_1

Has shared interests/hobbies

How do you think you measure up?

Please rate your opinion of your own attributes, on a scale of 1-10 (be honest!):

attr3_1

Attractive

sinc3_1

Sincere

int3_1

Intelligent

fun3_1

Fun

amb3_1

Ambitious

And finally, how do you think others perceive you?

Please rate yourself how you think others would rate you on each of the following attributes, on a scale of 1-10 (1=awful, 10=great)

attr5_1

Attractive

sinc5_1

Sincere

int5_1

Intelligent

fun5_1

Fun

amb5_1

Ambitious

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)?: _____

Half way through meeting all potential dates during the night of the event on their scorecard:

Hold up! Now that you are half way through your Speed Dates, we have a few questions for you…

We want to know what you look for in the opposite sex.

Please rate the importance of the following attributes in a potential date on a scale of 1-10: (1=not at all important, 10=extremely important).

attr1_s

Attractive_____

sinc1_s

Sincere _____

intel1_s

Intelligent _____

fun1_s

Fun _____

amb1_s

Ambitious _____

shar1_s

Shared Interests/Hobbies _____

Please rate your opinion of your own attributes, on a scale of 1-10 (1=awful, 10=great) --Be honest!

attr3_s

Attractive_____

sinc3_s

Sincere _____

intel3_s

Intelligent _____

fun3_s

Fun _____

amb3_s

Ambitious _____

followup/Time2:

[Survey is filled out the day after participating in the event. Subjects must have submitted this in order to be sent their matches.]

satis_2:

Overall, how satisfied were you with the people you met? (1=not at all satisfied, 10=extremely satisfied)

length:

Four minutes is:

Too little=1

Too much=2

Just Right=3

numdat_2:

The number of Speed "Dates" you had was:

Too few=1

Too many=2

Just right=3

Now, think back to your yes/no decisions during the Speed Dating event. Try to distribute the 100 points among these six attributes in the way that best reflects the actual importance of these attributes in your decisions. Give more points to those attributes that were more important in your decisions, and fewer points to those attributes that were less important in your decisions. Total points must equal 100.

attr7_2

Attractive

sinc7_2

Sincere

intel7_2

Intelligent

fun7_2

Fun

amb7_2

Ambitious

shar7_2

Has shared interests/hobbies

We want to know what you look for in the opposite sex.

Waves 1-5 and 10-21: You have 100 points to distribute among the following attributes -- give more points to those attributes that are more important in a potential date, and fewer points to those attributes that are less important in a potential date. *Total points must equal 100.*

Waves 6-9: Please rate the importance of the following attributes in a potential date on a scale of 1-10 (1=not at all important, 10=extremely important):

attr1_2

Attractive

sinc1_2

Sincere

intel1_2

Intelligent

fun1_2

Fun

amb1_2

Ambitious

shar1_2

Has shared interests/hobbies

What do you think MOST of your fellow men/women look for in the opposite sex?

You have 100 points to distribute among the following attributes -- give more points to those attributes that you think your fellow men/women find more important in a potential date, and fewer points to those attributes that they find less important in a potential date.

Total points must equal 100.

attr4_2

Attractive

sinc4_2

Sincere

intel4_2

Intelligent

fun4_2

Fun

amb4_2

Ambitious

shar4_2

Shared Interests/Hobbies

What do you think the opposite sex looks for in a date?

Please distribute 100 points among the following attributes -- give more points to

those attributes that you think are more important to members of the opposite sex when they are deciding whether to date someone. *Total points must equal 100.*

attr2_2

Attractive

sinc2_2

Sincere

intel2_2

Intelligent

fun2_2

Fun

amb2_2

Ambitious

shar2_2

Has shared interests/hobbies

How do you think you measure up?

Please rate your opinion of your own attributes, on a scale of 1-10 (1= awful and 10=great). Be honest!

attr3_2

Attractive

sinc3_2

Sincere

int3_2

Intelligent

fun3_2

Fun

amb3_2

Ambitious

And finally, how do you think others perceive you?

Please rate yourself how you think others would rate you on each of the following attributes, on a scale of 1-10 (1=awful, 10=great)

attr5_2

Attractive

sinc5_2

Sincere

int5_2

Intelligent

fun5_2

Fun

amb5_2

Ambitious

followup2/ Time3:

[Subjects filled out 3-4 weeks after they had been sent their matches]

SINCE HURRYDATING…

1. Of the matches that you received:

you_call:

(a) How many have you contacted to set up a date?

them_cal:

(b) How many have contacted you?

date_3:

Have you been on a date with any of your matches?

Yes=1

No=2

If you have been on at least one date, please answer the following:

numdat_3:

(a) How many of your matches have you been on a date with so far?

num_in_3

If yes, how many?

What do you look for in the opposite sex?

Please distribute 100 points among the following attributes -- give more to attributes that were more important in your decisions when Hurrydating, and less to attributes that were less important. Total points must equal 100.

We want to know what you look for in the opposite sex.

Please rate the importance of the following attributes in a potential date on a scale of 1-10 (1=not at all important, 10=extremely important):

attr1_3

Attractive

sinc1_3

Sincere

intel1_3

Intelligent

fun1_3

Fun

amb1_3

Ambitious

shar1_3

Has shared interests/hobbies

Now, think back to your yes/no decisions during the night of the Speed Dating event. Try to distribute the 100 points among these six attributes in the way that best reflects the actual importance of these attributes in your decisions. Give more points to those

attributes that were more important in your decisions, and fewer points to those attributes that less less important in your decisions. Total points must equal 100.

attr7_3

Attractive

sinc7_3

Sincere

intel7_3

Intelligent

fun7_3

Fun

amb7_3

Ambitious

shar7_3

Has shared interests/hobbies

Now we want to know what you think MOST of your fellow men/women look for in the opposite sex.

Please rate the importance of the following attributes on a scale of 1-10 (1=not at all important, 10=extremely important):

attr4_3

Attractive

sinc4_3

Sincere

intel4_3

Intelligent

fun4_3

Fun

amb4_3

Ambitious

shar4_3

Has shared interests/hobbies

What do you think the opposite sex looks for in a date?

Please rate the importance of the following attributes on a scale of 1-10 (1=not at all important, 10=extremely important):

attr2_3

Attractive

sinc2_3

Sincere

intel2_3

Intelligent

fun2_3

Fun

amb2_3

Ambitious

share2_3

Has shared interests/hobbies

Please rate your opinion of your own attributes, on a scale of 1-10 (1= awful and 10=great). Be honest!

attr3_3

Attractive

sinc3_3

Sincere

intel3_3

Intelligent

fun3_3

Fun

amb3_3

Ambitious

And finally, how do you think others perceive you?

Please rate yourself how you think others would rate you on each of the following attributes, on a scale of 1-10 (1=awful, 10=great)

attr5_3

Attractive

sinc5_3

Sincere

int5_3

Intelligent

fun5_3

Fun

amb5_3

Ambitious