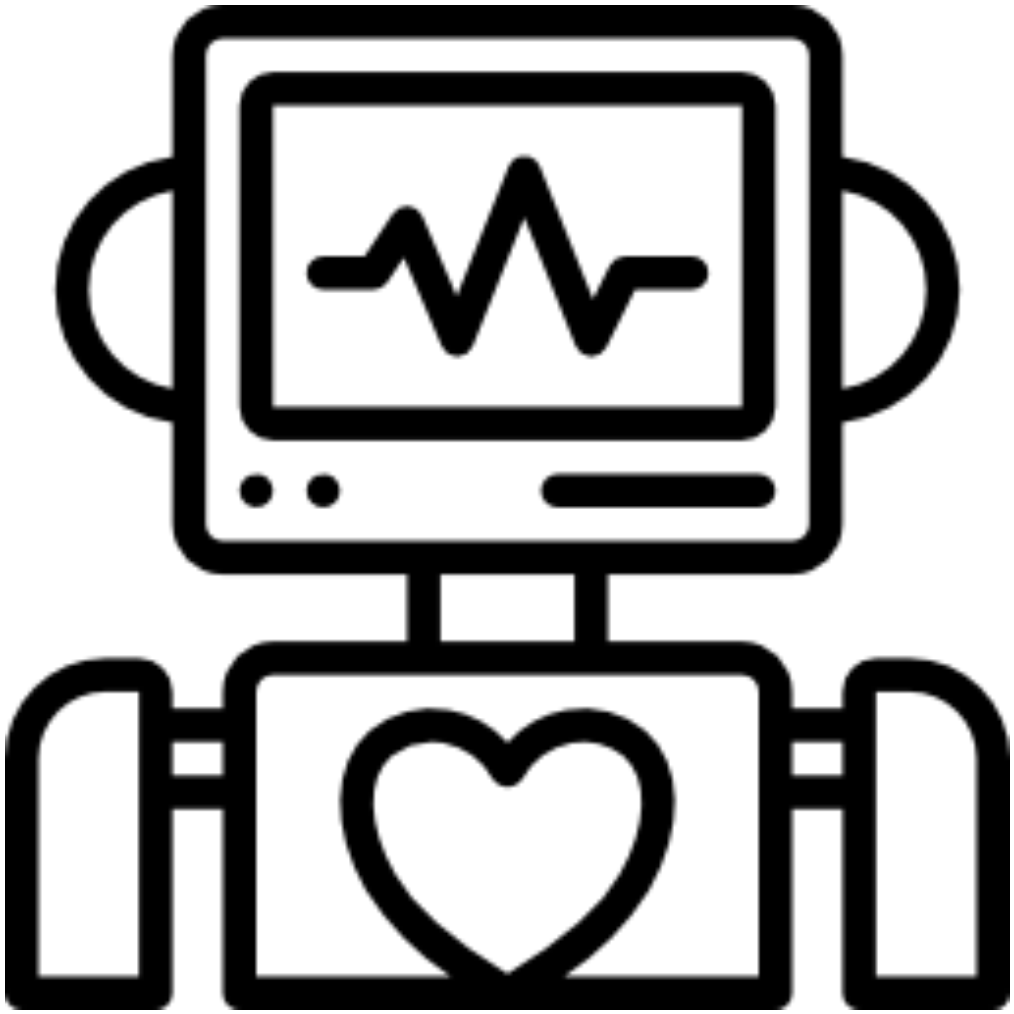
The Matchbot Project

What drives a decision for a 2nd date?

Tais Pancier - May 20, 2018



Introduction

Today's generations are looking for soul mates and have more opportunities than ever to find them. The $2.4 billion online-dating industry has exploded in the past few years with the arrival of dozens of mobile apps and dating events. Facebook has recently joined this market.

Part of this market, Speed dating is a formalized matchmaking process whose purpose is to encourage singles to meet new potential partners in a very short period of time through scheduled events. Men and women are rotated to meet each other over a series of short "dates" usually lasting from three to eight minutes depending on the organization running the event.

Most speed dating events match people at random, and participants will meet different "types" that they might not normally talk to in a club. According to the New York Times, participants in speed dating experience an average of 2 in 10 or 3 in 10 matches. Online dating participants, in contrast, only find a compatible match with 1 in 100 or fewer of the profiles they study.

There is an opportunity of using a model to predict a decision for a 2nd date and use it to pre-match Speed dating participants and possibly increase the average of matches, which is the purpose of this project.

Dataset

The dataset is a result of a Speed Dating experiment, in which participants engage in four-minute conversations to determine whether or not they are interested in meeting each other again. If both people "accept," then each is subsequently provided with the other's contact information.

The participants were drawn from students in graduate and professional schools at Columbia University.

Upon checking in, each participant was given a clipboard with a scorecard, a pen, and a name tag on which only his or her ID number was written. The scorecard was divided into columns in which participants indicated the ID number of each person they met. Participants would then circle "yes" or "no" under the ID number to indicate whether they would like to see the other person again. Beneath the Yes/No decision was a listing of the six attributes on which the participant was to rate his or her partner: Attractive, Sincere, Intelligent, Fun, Ambitious, Shared Interests.

The dataset consists of one csv file obtained in Kaggle:

https://www.kaggle.com/annavictoria/speed-dating-experiment/

The file contains the participants' gender and several sets of the 6 attributes (Attractive, Sincere; Intelligent; Fun; Ambitious; Shared Interests) recorded as a scale. At sign up, the participants were asked to rate the attributes that were most important for them in a partner and to assess how they would evaluate themselves in each of these attributes. The night of the event, they were asked to evaluate each partner they talked to based on these attributes. The dataset also contains a column that indicates how much a participant liked the people they met in a scale from 1 to 10 and a indicator of the decision for a second date. If both participants in the date indicated Yes for a second date, there's a match. Each row represents a speed date between 2 participants.

Data Wrangling

The raw dataset had 8378 rows and 95 columns. The Dataset after cleaning has 8378 columns, 74 columns.

The following steps were performed to clean the dataset:

- identified columns that wouldn't be used in the analysis and could be dropped (26 in total)
- changed 17 columns to categorical to allow proper summary analysis
- Attributes scale was different depending on the date/time of event. These columns were normalized by dividing each rating value by the sum of the related ratings (row-wise) resulting in a scale from 0 to 1.
- Outliers:
    - 1 rating record was originally higher than the max of the scale. As it was 1 record and just 1 point higher than the top of the scale, no special treatment was performed. This value was normalized as well.
- Missing values:
    - 19 columns with missing values were dropped because they were not relevant for the analysis
    - pid column (partner id): filled based on another id field and date/time of event
    - Attribute rating columns: missing data represented attributes that were not important for participants on each dating or they were not sure. NaN replaced with zero to allow correlation analysis
    - NaN in columns that might not be important for the analysis were left as NaN

Exploratory Data Analysis

The following steps were performed as part of the initial analysis:

1) Summary statistics for overall data and by gender:

- for both men and women the attribute ‘intelligence' has the highest mean and median
- Attribute ‘attractive' has higher mean for men than for women
- Attribute ‘ambitious' has higher mean for women than for men

	attr	sinc	intel	fun	amb
count	4042.000000	4042.000000	4042.000000	4042.000000	4042.000000
mean	0.158134	0.189935	0.198478	0.161111	0.170878
std	0.053920	0.054617	0.048718	0.046401	0.061353
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.133333	0.166667	0.173913	0.142857	0.153846
50%	0.157895	0.181818	0.190476	0.166667	0.173913
75%	0.179487	0.208333	0.216216	0.183673	0.200000
max	1.000000	0.909091	0.692308	0.421053	0.500000

Ratings the night of even given by female participants

	attr	sinc	intel	fun	amb
count	4088.000000	4088.000000	4088.000000	4088.000000	4088.000000
mean	0.171085	0.188514	0.189433	0.165598	0.158910
std	0.060274	0.047091	0.043728	0.042950	0.053378
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.145833	0.166667	0.166667	0.150000	0.145833
50%	0.166667	0.182746	0.184211	0.166667	0.166667
75%	0.190476	0.205882	0.205882	0.184211	0.183673
max	1.000000	0.666667	0.538462	0.562500	0.500000

Ratings the night of even given by male participants

2) Correlation between attribute ratings given by each participant to their partners the night of the event (continuous from 0 to 1 after normalization) and how much they like the partners (continuous from 1 to 10 but much less granular than ratings):

- Attractive and like: r = 0.132
- Sincere and like: r = -0.159
- Intelligent and like: r = -0.260
- Fun and like: r = 0.197
- Ambitious and like: r = -0.174
- Shared interests and like: 0.2322 (higher positive correlation)

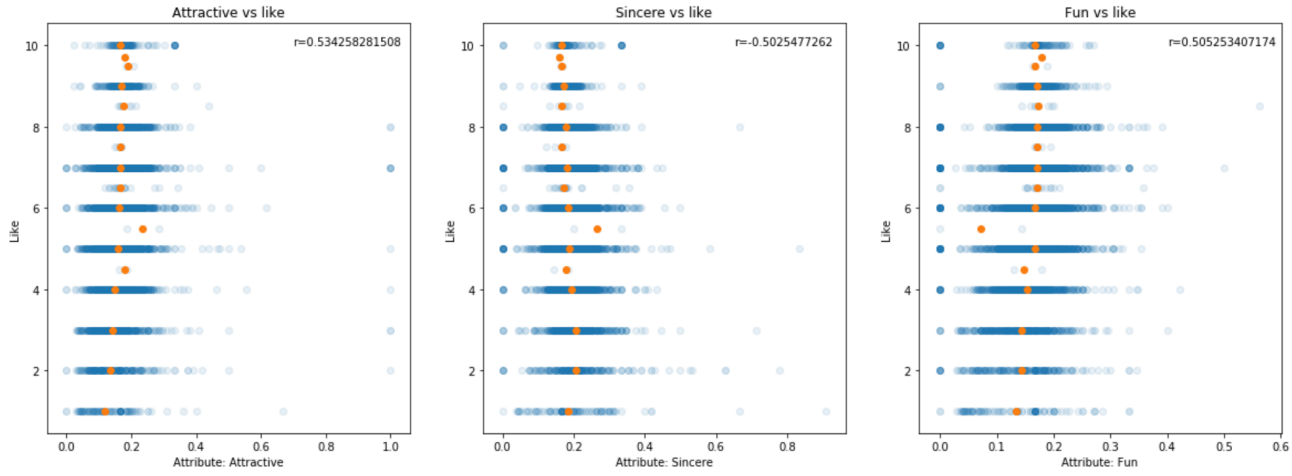
These correlations may indicate that the attributes that most impact whether a participant likes their partner are shared interests, fun and attractive.

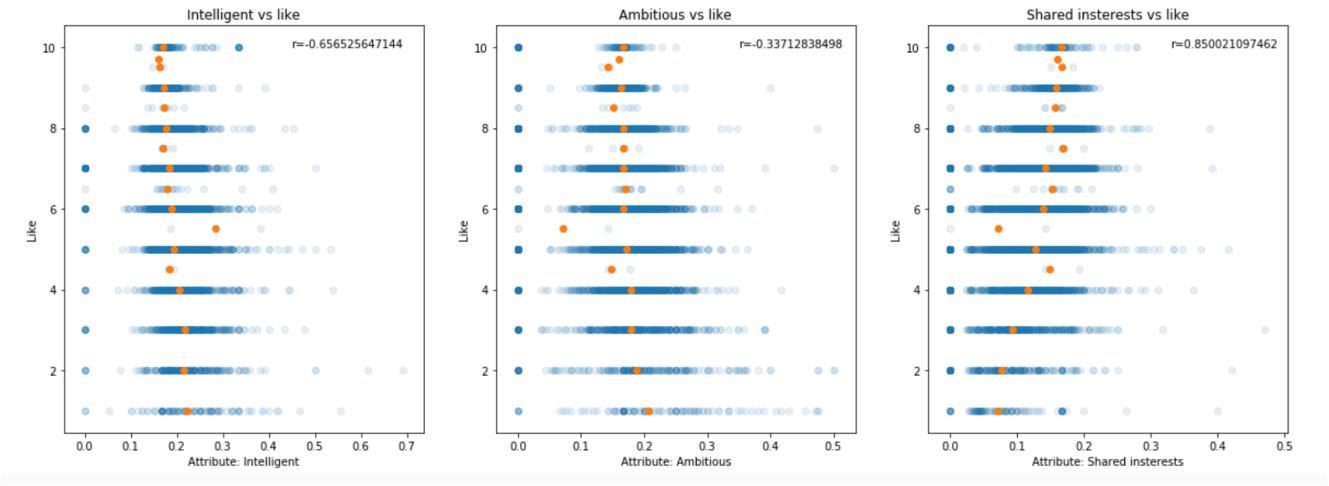
3) Correlation between median of attribute ratings given by each participant to their partners (continuous from 0 to 1 after, after normalization) at each like level (continuous from 1 to 10 but much less granular than ratings):

- Attractive median and like: r = 0.534
- Sincere median and like: r = -0.502
- Intelligent median and like: r = -0.656
- Fun median and like: r = 0.505
- Ambitious median and like: r = -0.337
- Shared interest median and like: r = 0.850

In general, without considering the decision for a second date, the numbers above show stronger correlations for shared interests, fun and attractive.

It shows a negative correlation for sincere, intelligent and ambitious. It may indicate that these attributes are not correctly perceived in a speed dating (5 min), may be interpreted as "trying too hard" to be sincere, intelligent or ambitious.

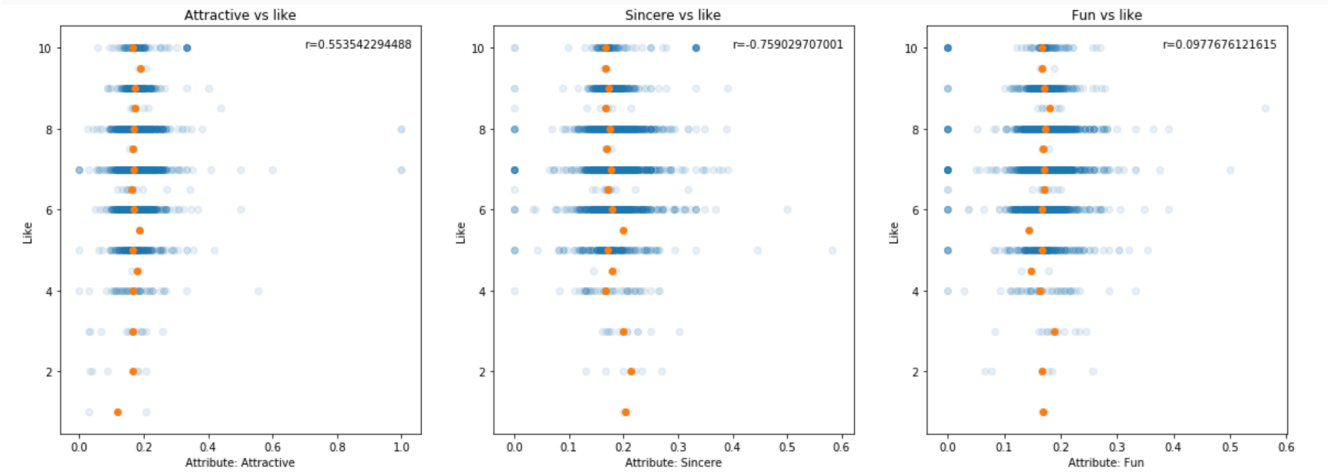




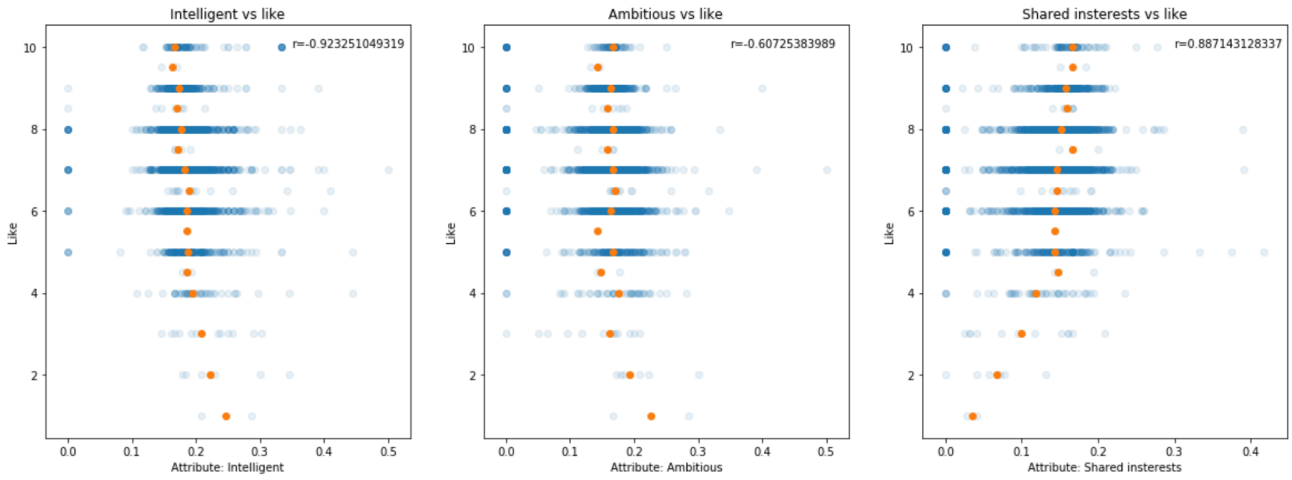
Attributes vs Like (orange: attribute median for each like datapoint)

4) Correlation between median of attribute ratings given by each participant to their partners (continuous from 0 to 1, after normalization) at each like level (continuous from 1 to 10 but much less granular than ratings) for partners that said yes to a 2nd date:

- Attractive median and like: r = 0.553
- Sincere median and like: r = -0.759
- Intelligent median and like: r = -0.923
- Fun median and like: r = 0.097
- Ambitious median and like: r = -0.607
- Shared interest median and like: r = 0.887



For participants that decided on a 2nd date, there is a stronger correlation for shared interests and attractive. Fun correlation is weak for participants who decided on a 2nd date.



Attributes vs Like when decision 'yes' (orange: attribute median for each 'like' datapoint)

5) Correlation between attributes rated based on importance for participants at sign-up and ratings given to partners the night of the event:

	Decision = Yes	Decision = No
Attractive	0.086546249763918573	-0.00051471268108420061
Sincere	0.020917505254446113	-0.039554598538752421
Intelligent	-0.0030664664717438223	0.0055873166593916343
Fun	0.049377289530955416	-0.0057500830536163497
Ambitious	0.097102323614510239	0.07300848738938924
Shared Interests	-0.012589961511008526	-0.0004194855264972844

Decision 'Yes' group: Weak correlation between attributes that were ranked based on importance at sign up and attributes' ranking the night of even for participants that said yes to a second date. It may indicate that the relation between attributes importance and ratings the night of the event is not a strong drive for the decision of a 2nd date.

Decision 'No' group: All but intelligence attribute have negative correlation, which is expected for the group with decision = no. Weak correlation between attributes that were ranked based on importance at sign up and attributes ranking the night of even for participants that said no to a second date.

6) Analysis between self-assessment and ratings received by partner the night of the event:

	attr3_1	sinc3_1	fun3_1	intel3_1	amb3_1
count	3469.000000	3469.000000	3469.000000	3469.000000	3469.000000
mean	0.180536	0.207762	0.195989	0.209904	0.188803
std	0.034510	0.043271	0.038398	0.036694	0.042481
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.166667	0.195122	0.184211	0.200000	0.171429
50%	0.184211	0.210526	0.200000	0.209302	0.200000
75%	0.200000	0.230769	0.216216	0.225000	0.214286
max	0.264706	0.347826	0.303030	0.360000	0.285714

Summary statistics of how participants perceive themselves for Decision Yes

	pf_o_att	pf_o_sin	pf_o_int	pf_o_fun	pf_o_amb
count	3469.000000	3469.000000	3469.000000	3469.000000	3469.000000
mean	0.220313	0.175187	0.200863	0.173136	0.104051
std	0.126650	0.071445	0.069256	0.063919	0.060367
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.150000	0.150000	0.177782	0.150000	0.050000
50%	0.200000	0.189219	0.200000	0.180000	0.100000
75%	0.250000	0.200000	0.238100	0.200000	0.150000
max	1.000000	0.600000	0.500000	0.500000	0.358108

Summary statistics of how participants are perceived by their partners for Decision Yes

	attr3_1	sinc3_1	fun3_1	intel3_1	amb3_1
count	4661.000000	4661.000000	4661.000000	4661.000000	4661.000000
mean	0.177507	0.212997	0.192848	0.216941	0.191983
std	0.032869	0.039266	0.036311	0.033325	0.041158
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.162791	0.200000	0.179487	0.200000	0.175000
50%	0.181818	0.216216	0.200000	0.214286	0.200000
75%	0.200000	0.232558	0.214286	0.232558	0.214286
max	0.264706	0.347826	0.303030	0.360000	0.285714

Summary statistics of how participants perceive themselves for Decision No

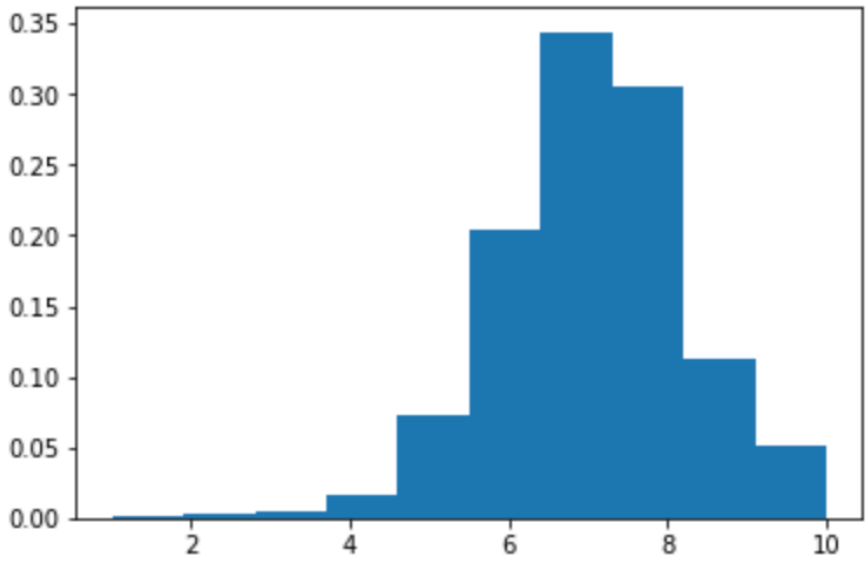
	pf_o_att	pf_o_sin	pf_o_int	pf_o_fun	pf_o_amb
count	4661.000000	4661.000000	4661.000000	4661.000000	4661.000000
mean	0.223443	0.170264	0.200921	0.171951	0.106840
std	0.126836	0.072674	0.071582	0.062700	0.061132
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.150000	0.140000	0.172417	0.150000	0.050000
50%	0.200000	0.180000	0.200000	0.180000	0.100000
75%	0.250000	0.200000	0.227273	0.200000	0.150000
max	1.000000	0.600000	0.500000	0.500000	0.358108

Summary statistics of how participants are perceived by partners for Decision No

Participants gave a lower rating for themselves in comparison to the ratings given to them by their partners in the event for attribute Attractiveness. For all other attributes participants gave them higher ratings in comparison to the ratings given to them. The means of all self-assessed ratings are lower for participants who received a 'Yes' from their partners than for participants who received a 'No', except for Attractiveness.

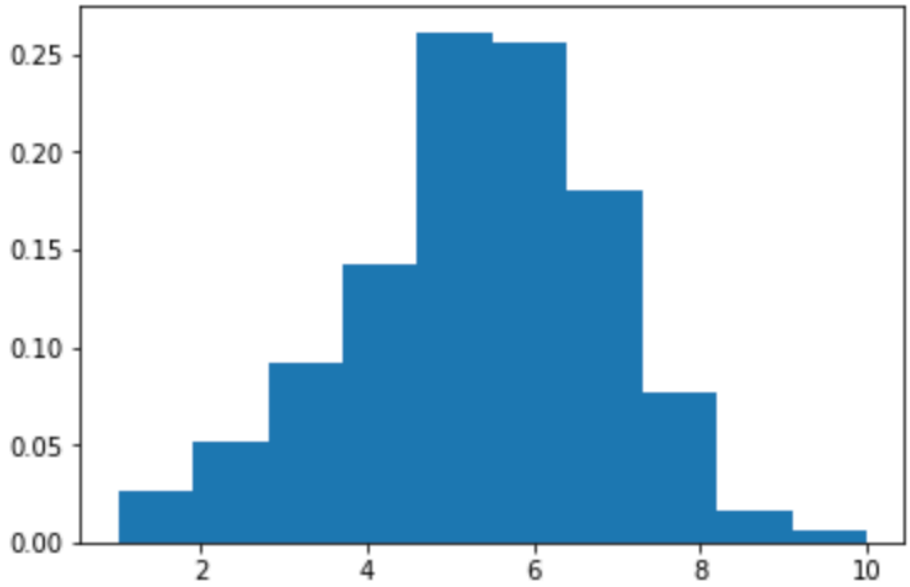
7) Like scale vs decision:

Mean of Like for participants that decided ‘yes’ for a 2nd date was **7.22**.



Histogram of ‘Like' when decision is Yes

Mean of Like for participants that decided ‘No' for a 2nd date was **5.32**.



Histogram of ‘Like' when decision is No

Mean of ‘Like' is higher for participants that decided for a 2nd date compared to the ones that said No for a 2nd date. The histograms above also show that Like needs to be considerably high for someone to decide for a second date.

Some participants said No to a 2nd date even liking the other participant (Like > 8 in histogram).

8) Distance for attributes importance between partners at sign up:

- The distance median for attractiveness is lower for participants that had a match
- The distance median for shared interests is higher for participants that had a match
- The distance is similar for the other attributes.
- The overall distance of all attributes, calculated using Euclidean distance, is lower for participants with a match than for those who didn’t opt for a 2nd date. It may indicate that participants that said yes to a second date are in general more similar in what they value than participants who said no to a 2nd date.

	Decision = Yes	Decision = No
Attractive	0.098995	0.100000
Sincere	0.050000	0.050000
Intelligent	0.050000	0.050000
Fun	0.050000	0.050000
Ambitious	0.050000	0.050000
Shared Interests	0.056812	0.050000
Overall	0.2017422577816702	0.21213203435596428

Distance median of attributes importance ranked by both participants at sign-up

9) Distance for self evaluation between partners at sign up:

- The distance median for all attributes is lower for participants that had a match than for participants that said no to a 2nd date, except for attractiveness which is the same.

- The overall distance of all attributes, calculated using Euclidean distance, is lower for participants with a match than for those who didn't opt for a 2nd date. It may indicate that participants that said yes to a second date are in general more similar in what they value than participants who said no to a 2nd date.

	Decision = Yes	Decision = No
Attractive	0.025000	0.025000
Sincere	0.026374	0.028674
Intelligent	0.020455	0.023443
Fun	0.025000	0.025094
Ambitious	0.029268	0.031714
Overall	0.0754615548851389	0.07993541639045745

Distance median of self evaluation ranked by both participants at sign-up

Insights from EDA

- Attributes that mostly impact like: Shared interests, Fun and Attractive
- Attributes that mostly impact like when the decision for a 2nd date is ‘Yes’: Shared interests and Attractive
- The distance for attributes importance between 2 participants is smaller for participants that said ‘Yes' to a 2nd date.
- The distance for self evaluation between 2 participants is smaller for participants that said ‘Yes' to a 2nd date.
- Features to be explored in a machine learning model:
  - Attributes important at sign up
  - Distance between attributes important at sign up
  - Distance between self evaluation at sign up

Machine Learning Model

In order to make a prediction of match before the event occurs and be able to suggest participants who should meet, the following features were explored:

- Attributes importance: 6 attributes for each participant (attractiveness, sincerity, intelligence, ambition, shared interests)
- Self-evaluation: 5 attributes for each participant (attractiveness, sincerity, intelligence, ambition)

Train and test sets were split using test size of 0.2.

1) Predicting match based on attributes importance at sign up:

Each participant ranked the attributes (attractive, sincere, intelligent, ambitious, fun, shared interests) assigning a scale from 1 to 10 based on what is important for them in a partner. The values were normalized to a scale from 0 to 1. The model considers that how participants rate attributes based on the importance for them influences a match.

Classifier: Random Forest
Features: 'attr1_1', 'sinc1_1', 'intel1_1', 'fun1_1', 'amb1_1', 'shar1_1', 'pf_o_att', 'pf_o_sin', 'pf_o_int', 'pf_o_fun', 'pf_o_amb', 'pf_o_sha'

Scores:

Accuracy: 0.83054892601431984

Confusion Matrix:

	Predicted: 0	Predicted: 1
Actual: 0	1355	45
Actual: 1	253	23

Classification Report:

	Precision	Recall	f1-score	Support
0	0.84	0.97	0.90	1400
1	0.34	0.08	0.13	276
avg/total	0.76	0.82	0.77	1676

Since scores were very low with Random Forest default hyper parameters, hyper parameter optimization was not explored.

2) Predicting match based on distance of attributes importance between 2 participants:

Each participant ranked the attributes (attractive, sincere, intelligent, ambitious, fun, shared interests) assigning a scale from 1 to 10 based on what is important for them in a partner. The values were normalized to a scale from 0 to 1. The distance between each attribute for 2 participants who met is calculated. A low distance may indicate the participants share the same opinion or value about that specific attribute. The model considers that the distance between participants in each attribute drives a match.

Distance: absolute difference between attributes for each participants:

Attractiveness distance = attractiveness for participant 1 - attractiveness for participant 2
Sincerity distance = sincerity for participant 1 - sincerity for participant 2
Intelligence distance = intelligence for participant 1 - intelligence for participant 2
Ambition distance = ambition for participant 1 - ambition for participant 2
Shared interests distance = shared interests for participant 1 - shared interests for participant 2

Classifier: Random Forest
Hyper parameter optimization: RandomizedSearchCV resulted in the following best parameters:

'max_depth': 24, 'max_features': 1, 'n_estimators': 115

Features: 6 distances explained above (dis_att, dis_sinc, dis_sinc, dis_fun, dis_amb, dis_sha)

Scores:

Accuracy: 0.92541766109785206

Confusion Matrix:

	Predicted: 0	Predicted: 1
Actual: 0	1363	37
Actual: 1	88	188

Classification Report:

	Precision	Recall	f1-score	Support
0	0.94	0.97	0.96	1400
1	0.84	0.68	0.75	276
avg/total	0.92	0.93	0.92	1676

3) Predicting match based on distance of self evaluation between 2 participants:

Each participant evaluated themselves on 6 attributes (attractive, sincere, intelligent, ambitious, fun, shared interests) assigning a scale from 1 to 10 based on how they see themselves. The values were normalized to a scale from 0 to 1. The distance between each attribute for 2 participants who met is calculated. A low distance indicates the participants see themselves in a similar way for a specific attribute. The model considers that the distance between participants in each attribute drives a match.

Distance: absolute difference between attributes for each participants:

Attractiveness distance = attractiveness for participant 1 - attractiveness for participant 2
Sincerity distance = sincerity for participant 1 - sincerity for participant 2
Intelligence distance = intelligence for participant 1 - intelligence for participant 2
Ambition distance = ambition for participant 1 - ambition for participant 2

Classifier: Random Forest

Hyper parameter optimization: RandomizedSearchCV resulted in the following best parameters:
'max_depth': 38, 'max_features': 1, 'n_estimators': 140
Features: 5 distances explained above (dis_att, dis_sinc, dis_sinc, dis_fun, dis_amb)

Scores:

Accuracy: 0.94033412887828161

Confusion Matrix:

	Predicted: 0	Predicted: 1
Actual: 0	1385	15
Actual: 1	85	191

Classification Report:

	Precision	Recall	f1-score	Support
0	0.94	0.99	0.97	1400
1	0.93	0.69	0.79	276
avg/total	0.94	0.94	0.94	1676

4) Predicting match attributes importance and self evaluation between 2 participants:

The distances between importance of attributes and self evaluation calculated and used separately in previous models were used in conjunction as features model #4. The model considers that the distances between what participants consider important (in each attribute) and the distance of how they see themselves (in each attribute) drive a match.

Distance: absolute difference between attributes for each participants
Classifier: Random Forest
Hyper parameter optimization: RandomizedSearchCV resulted in the following best parameters:

'max_depth': 30, 'max_features': 3, 'n_estimators': 107
Features: 11 distances (6 for attributes importance and 5 for self evaluation)

Scores:

Accuracy: 0.9564439140811456

Confusion Matrix:

	Predicted: 0	Predicted: 1
Actual: 0	1397	3
Actual: 1	70	206

Classification Report:

	Precision	Recall	f1-score	Support
0	0.95	1.00	0.97	1400
1	0.99	0.75	0.85	276
avg/total	0.96	0.96	0.95	1676

Collinearity test between attributes importance distance and self evaluation distance:

Collinearity was investigated to confirm is all these predictors could be used together in the model.

Attributes	Correlation
Attractiveness	0.19853691445117458
Sincerity	0.30994446956182597

Intelligence	0.26232617147958925
Fun	0.24456190003210818
Ambition	0.11175163477173902

Attributes importance distance and self evaluation distance have low correlation for all the attributes, so they can be used as predictor variables.

Conclusion

Model 4 has the best scores among all the models analyzed and showed that using distances of attributes importance between participants and distances of self evaluation between participants as features give the best prediction algorithm.

This machine learning model allows a speed dating event or dating app to:

- given one participant (search_id = 115), find other participants (iid) that would give a good match, using a cutoff to only show the first 5 matches for example.

	dis_attr1	dis_sinc1	dis_intel1	dis_fun1	dis_amb1	dis_shar1	dis_attr3	dis_sinc3	dis_intel3	dis_fun3	dis_amb3	search_id	iid	y_pred	y_pred_prob
1776	0.00	0.05	0.10	0.05	0.03	0.13	0.004545	0.038636	0.020455	0.027273	0.027273	115	125	1	0.900323
6428	0.05	0.10	0.00	0.00	0.10	0.05	0.200000	0.175000	0.225000	0.200000	0.200000	115	416	1	0.683400
1796	0.05	0.15	0.05	0.05	0.10	0.00	0.009524	0.015476	0.010714	0.014286	0.009524	115	127	1	0.668571
1806	0.20	0.05	0.20	0.00	0.05	0.00	0.016216	0.014189	0.008784	0.010811	0.010811	115	128	1	0.629593
6392	0.75	0.09	0.29	0.19	0.14	0.04	0.200000	0.175000	0.225000	0.200000	0.200000	115	414	1	0.602610

- given all participants who signed up for an event or in the app, match all against each other and find the best matches to suggest the dating couples:

	iid	iid_b	y_pred	y_pred_prob
204208	372	340	1	0.929050
204004	372	136	1	0.929050
203897	372	28	1	0.929050
203927	372	58	1	0.929050
203928	372	59	1	0.929050
204207	372	339	1	0.929050
204214	372	346	1	0.929050
201453	367	340	1	0.878179
201173	367	59	1	0.878179
201142	367	28	1	0.878179

Proposed Next Steps

- Apply this ML model in an event and measure the performance of the predictions vs. real matches

- Enhance model using potential additional features (shared interests, religious, race, ethnicity)