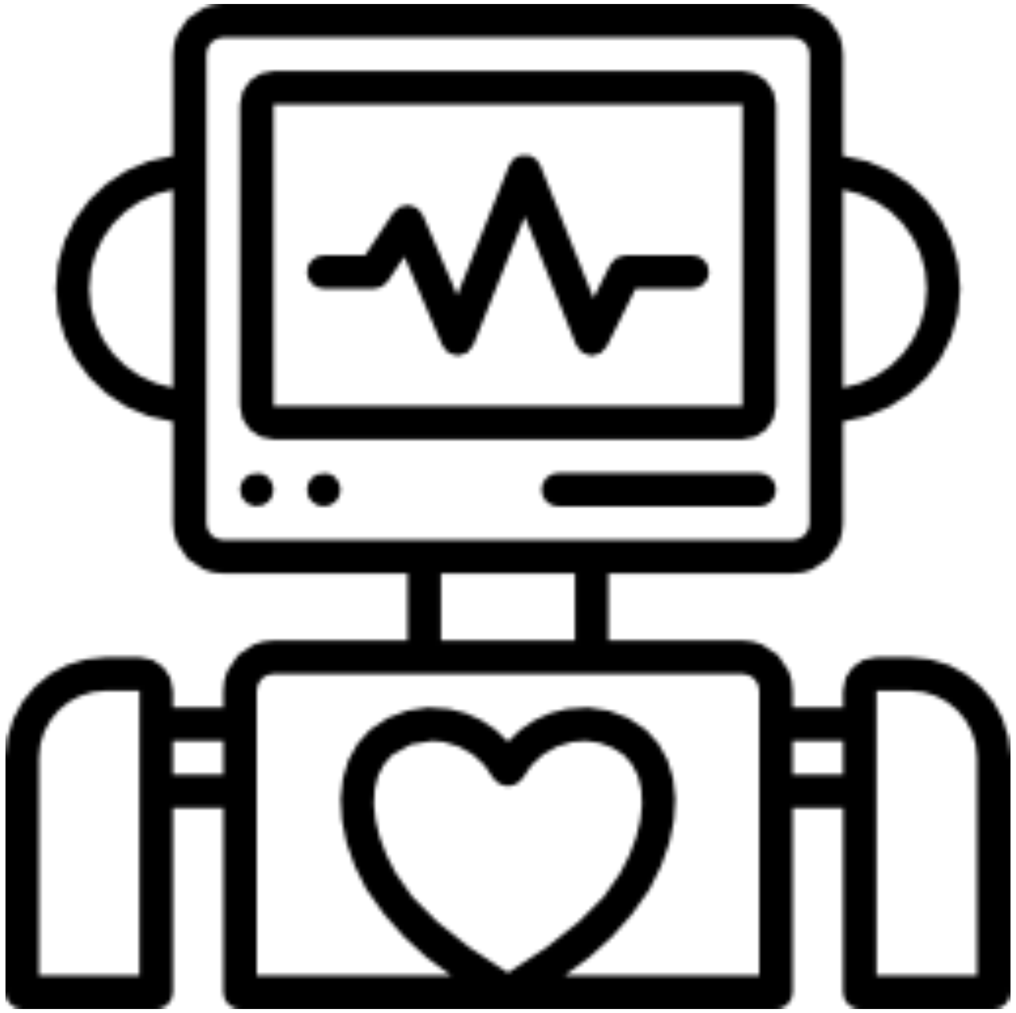


THE MATCHBOT PROJECT

Author: Tais Pancier

INTRODUCTION

- Today’s generations are looking for soul mates and have more opportunities than ever to meet people. The $3 billion dating service industry only in US has exploded in the past few years with the arrival of dozens of mobile apps and dating events
- 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
- Most speed dating events match people at random

OBJECTIVE

- The Matchbot project aims to build an engine that predicts the best matches based on information provided by participants at sign up so the event can pre-match people using machine learning instead of doing it at random. The same can also be used by dating apps to suggest people who could meet.

- The model uses ratings provided by participants at sign up to rank 6 attributes that are important in a partner (Attractive, Sincere, Intelligent, Fun, Ambitious, Shared Interests) and how they evaluate themselves based on the first 5 of these attributes.

DATA SOURCE

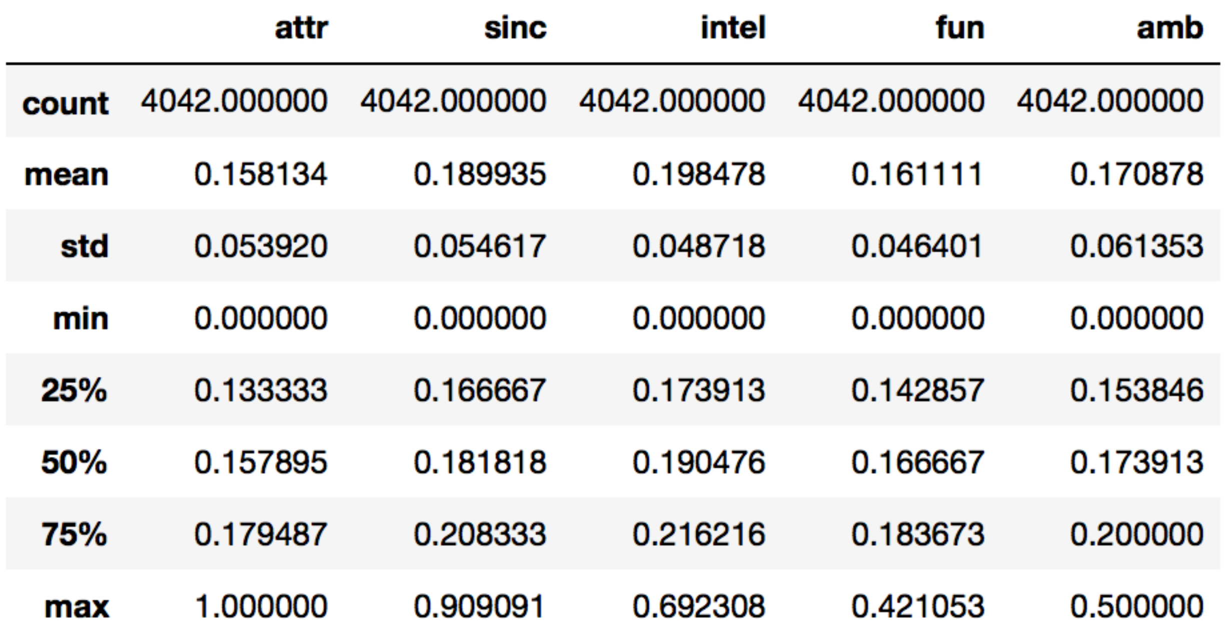
- The dataset is a result of a Speed Dating experiment, in which participants engage in four-minute conversations to determine whether or not they are interested in meeting each other again. Each row represents a speed date between 2 participants.
- The participants were drawn from students in graduate and professional schools at Columbia University.
- At sign up, the participants were asked to rate 6 attributes that were most important for them in a partner (Attractive, Sincere, Intelligent, Fun, Ambitious, Shared Interests) and to evaluate themselves in 5 attributes (Attractive, Sincere, Intelligent, Fun, Ambitious).
- The night of the event, they were asked to evaluate each partner they talked to based on the 6 attributes and to indicate "yes" or "no" to a second date with each partner. If both participants in the date indicated Yes for a second date, there's a match (match = 0 for no match, 1 for match).
- The dataset also contains a column that indicates how much a participant liked people they met in a scale from 1 to 10.

DATA CLEANING

- The raw dataset had 8378 rows and 95 columns. The Dataset after cleaning has 8378 columns, 74 columns, after dropping columns that wouldn’t be used.
- Attributes ratings had different scales (some from 1 to 10, some from 1 to 100) 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.
- Participants were instructed to leave an attribute rating blank if they didn’t consider it important or were not sure. Therefore missing values were replaced by 0 in the attributes ratings.

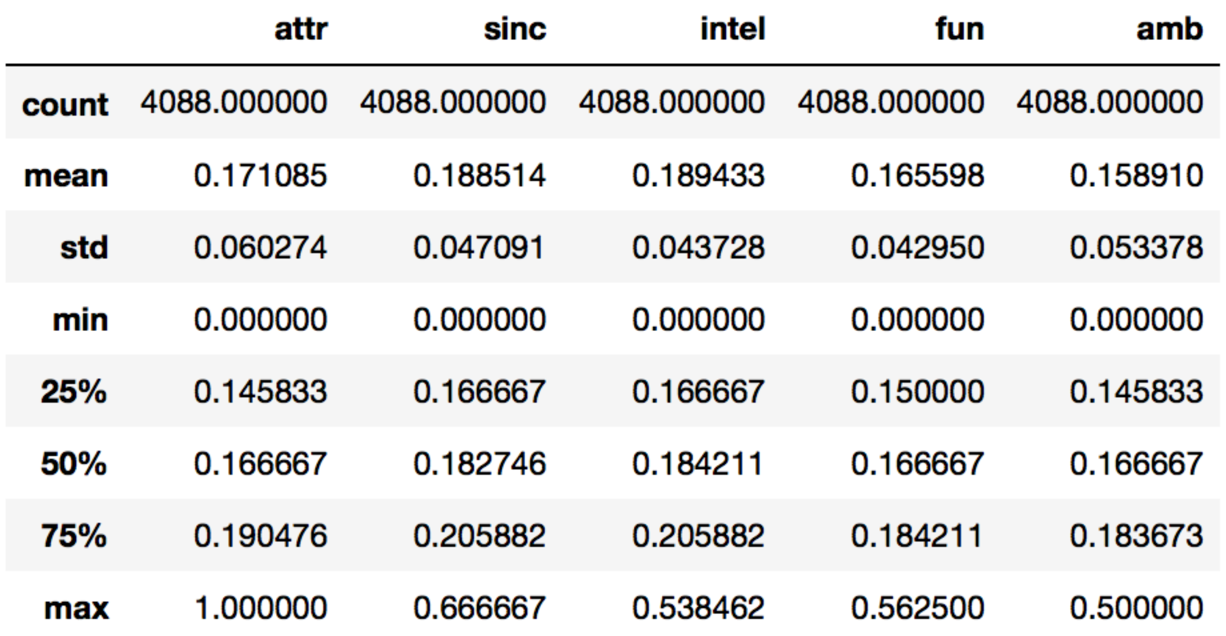
EXPLORATORY DATA ANALYSIS

1. Summary statistics for ratings given by participants to their partners the night of the event:



	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



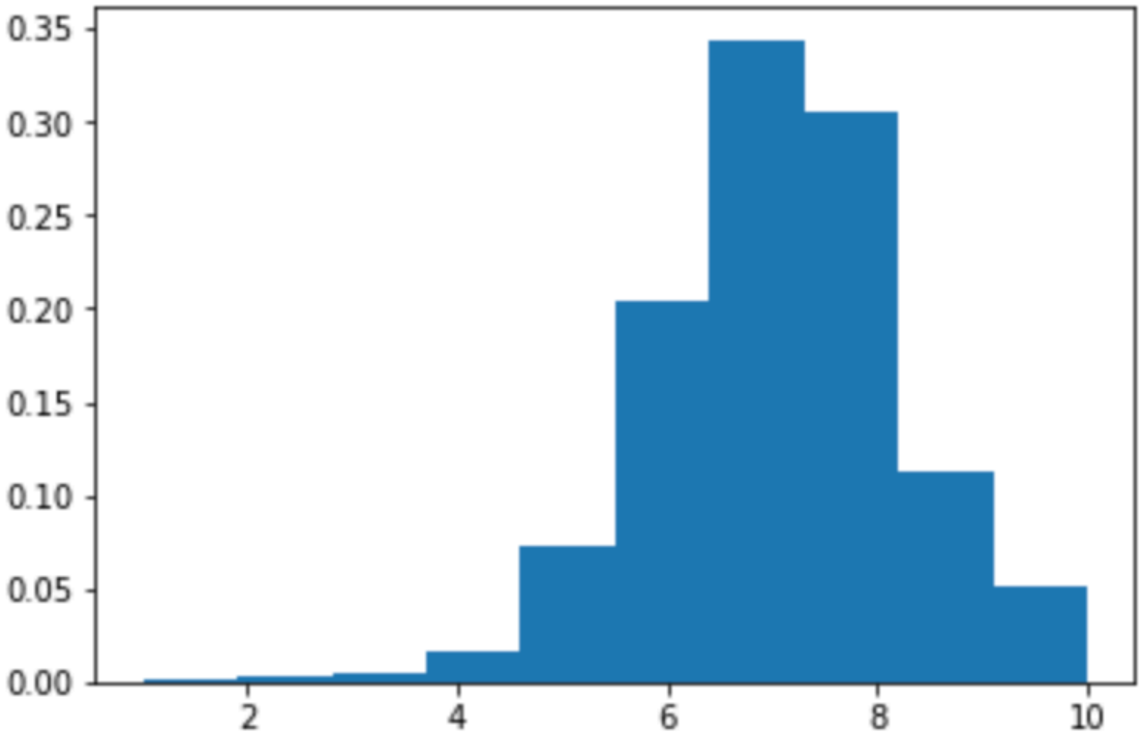
	attr	sinc	intel	fun	amb
count	4088.000000	4088.000000	4088.000000	4088.000000	4088.000000
mean	0.171085	0.188514	0.189433	0.165598	0.158910
std	0.060274	0.047091	0.043728	0.042950	0.053378
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.145833	0.166667	0.166667	0.150000	0.145833
50%	0.166667	0.182746	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

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

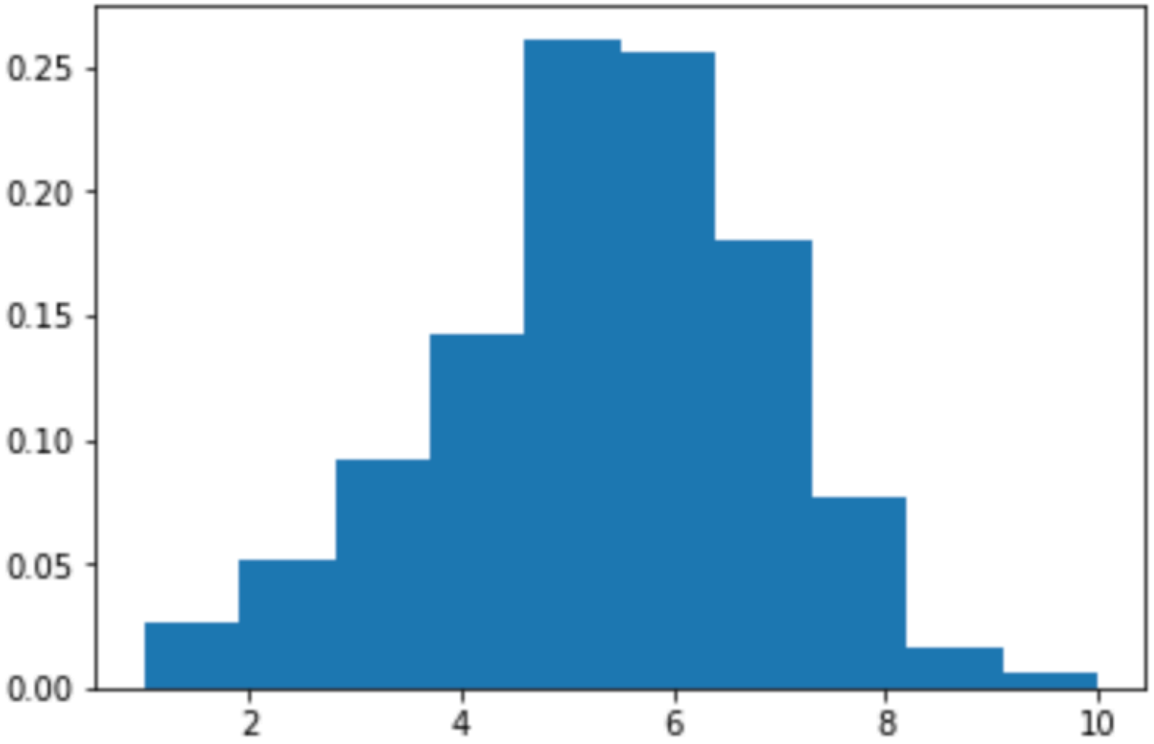
EXPLORATORY DATA ANALYSIS

2. Like scale vs decision ‘yes':



Histogram of Like when decision for a second date is ‘yes’

(mean: 7.22, median:7.0)



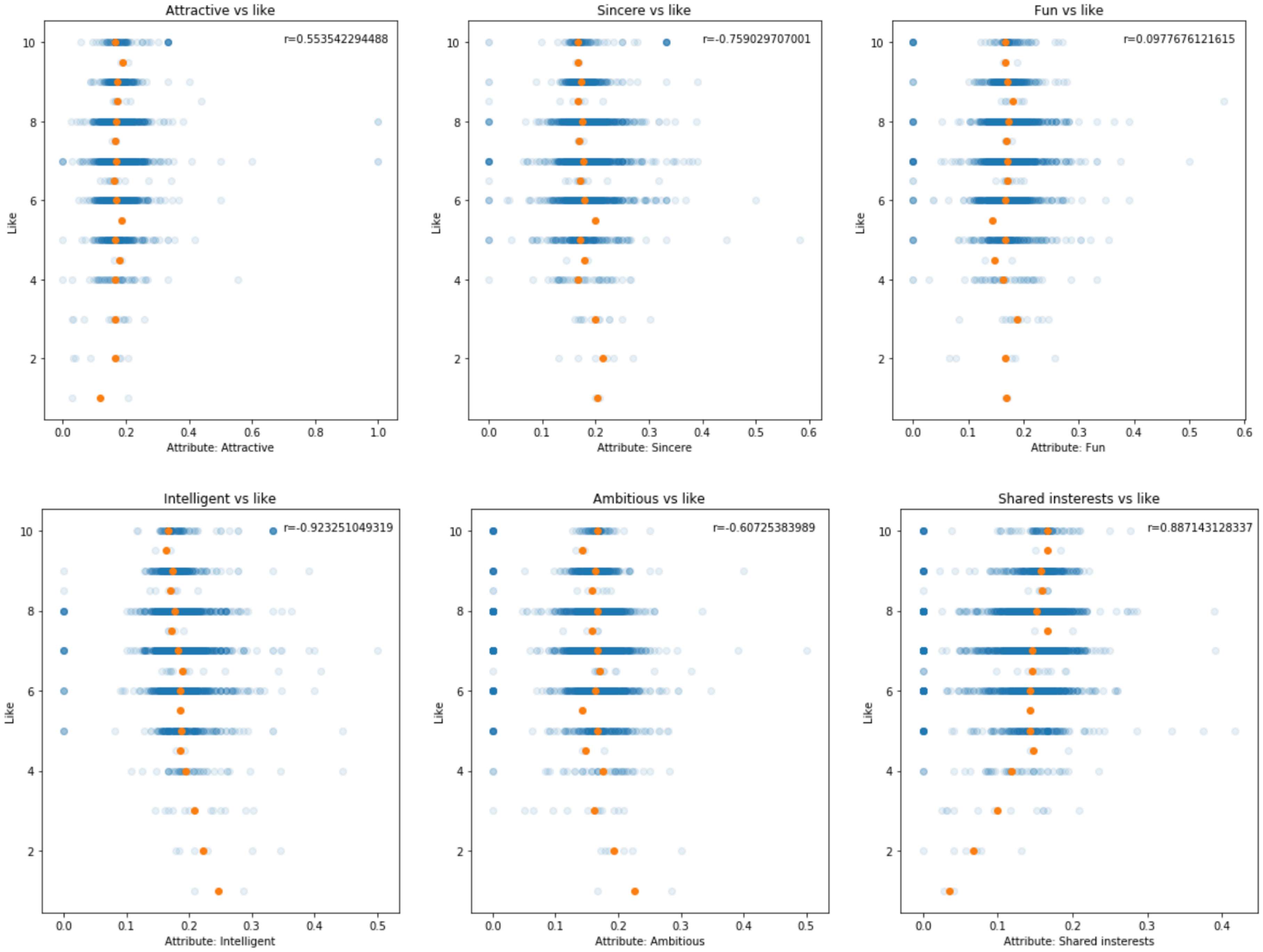
Histogram of Like when decision for a second date is ‘no’

(mean: 5.32, median: 5.0)

- Mean and median 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).

EXPLORATORY DATA ANALYSIS

3. Correlation between median ratings given by each participant to their partners during the event and like scale when decision = yes:



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

EXPLORATORY DATA ANALYSIS

3. Correlation between median of attribute ratings given by each participant to their partners during the event and like scale when decision = yes (cont.):

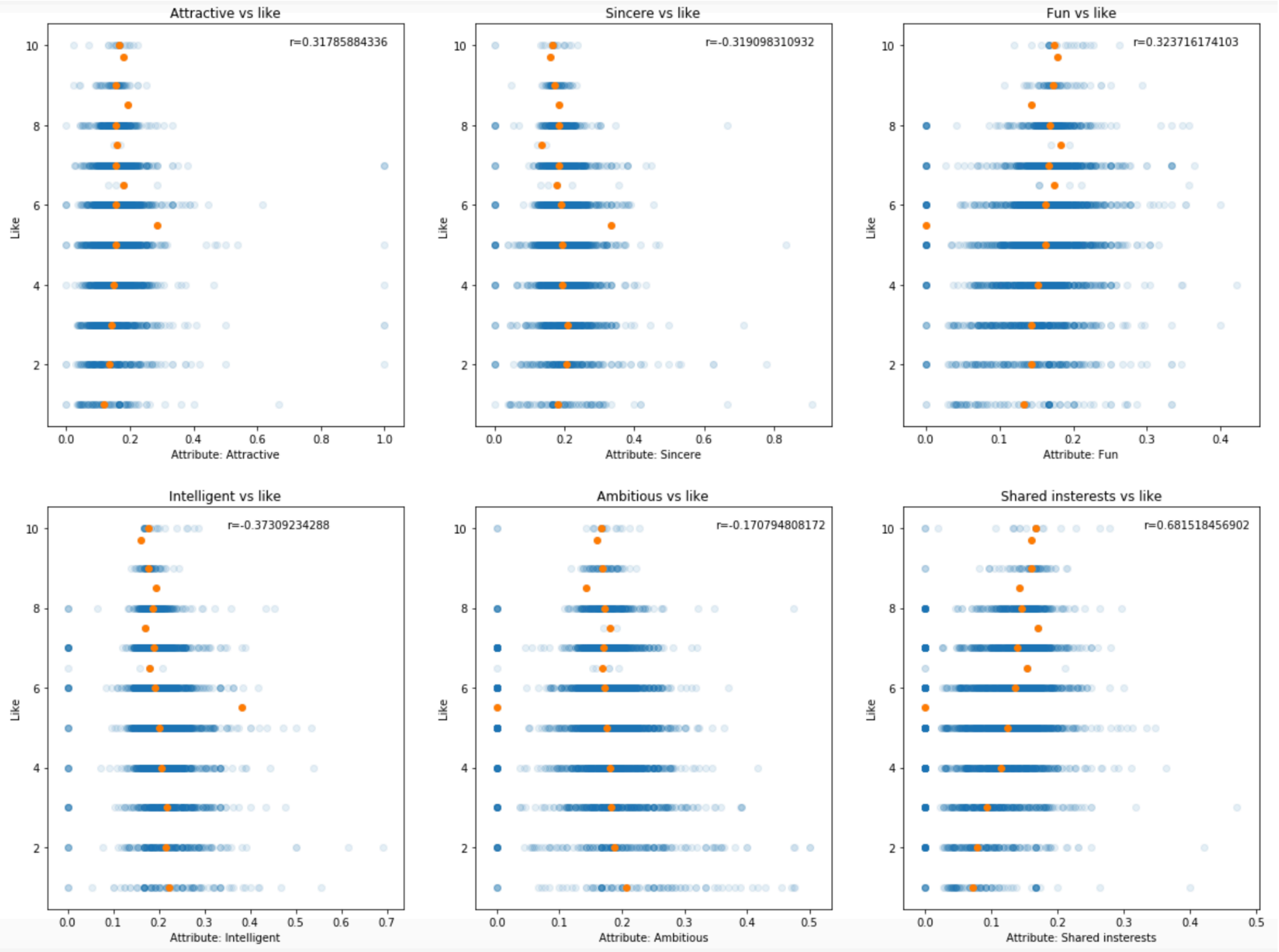
- Attractive median and like: r = 0.553
- Sincere median and like: r = -0.759
- Intelligent median and like: r = -0.923
- Fun median and like: r = 0.097
- Ambitious median and like: r = -0.607
- Shared interest median and like: r = 0.887

➤ 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.

➤ Correlation is negative for ‘sincere’, ‘intelligent' and ‘ambitious'.

EXPLORATORY DATA ANALYSIS

4. Correlation between median of attribute ratings given by each participant to their partners during the event and like scale when decision = no:



Attributes vs Like when decision ‘no' (orange: attribute median for each 'like' datapoint)

EXPLORATORY DATA ANALYSIS

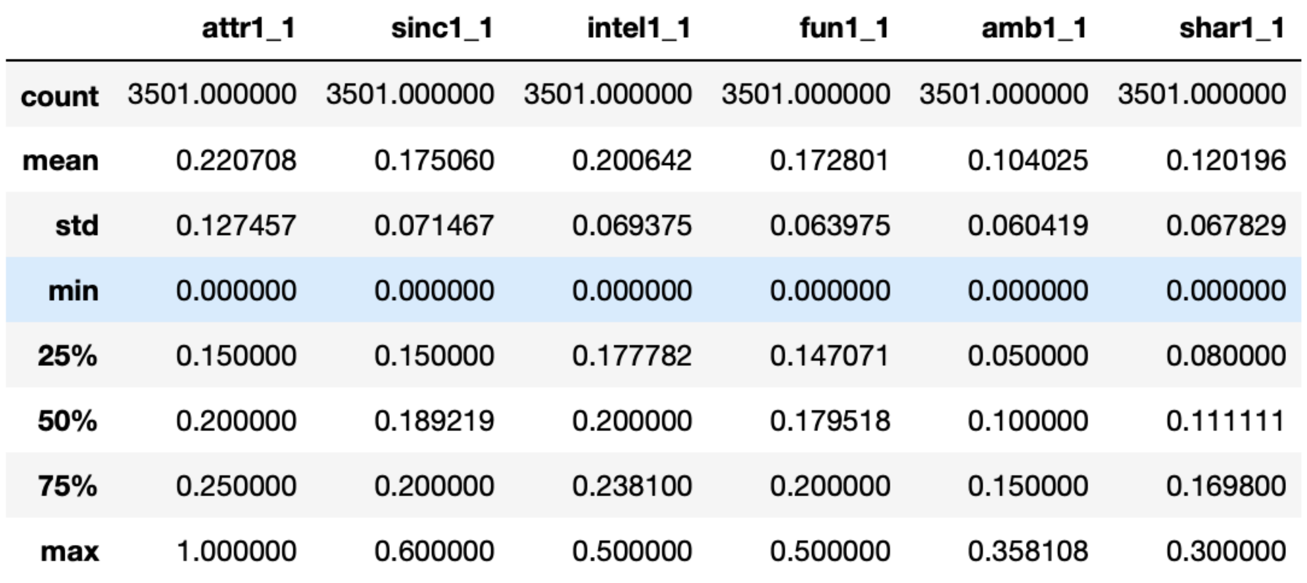
4. Correlation between median of attribute ratings given by each participant to their partners during the event and like scale when decision = no (cont.):

- Attractive median and like: r = 0.318
- Sincere median and like: r = -0.319
- Intelligent median and like: r = -0.373
- Fun median and like: r = 0.323
- Ambitious median and like: r = -0.171
- Shared interest median and like: r = 0.682

➤ Lower correlation between Fun attribute rating and Like for participants who decided for a 2nd date than for participants who said no for a 2nd date. It may indicate that Fun attribute doesn’t strongly determine a decision for a 2nd date.

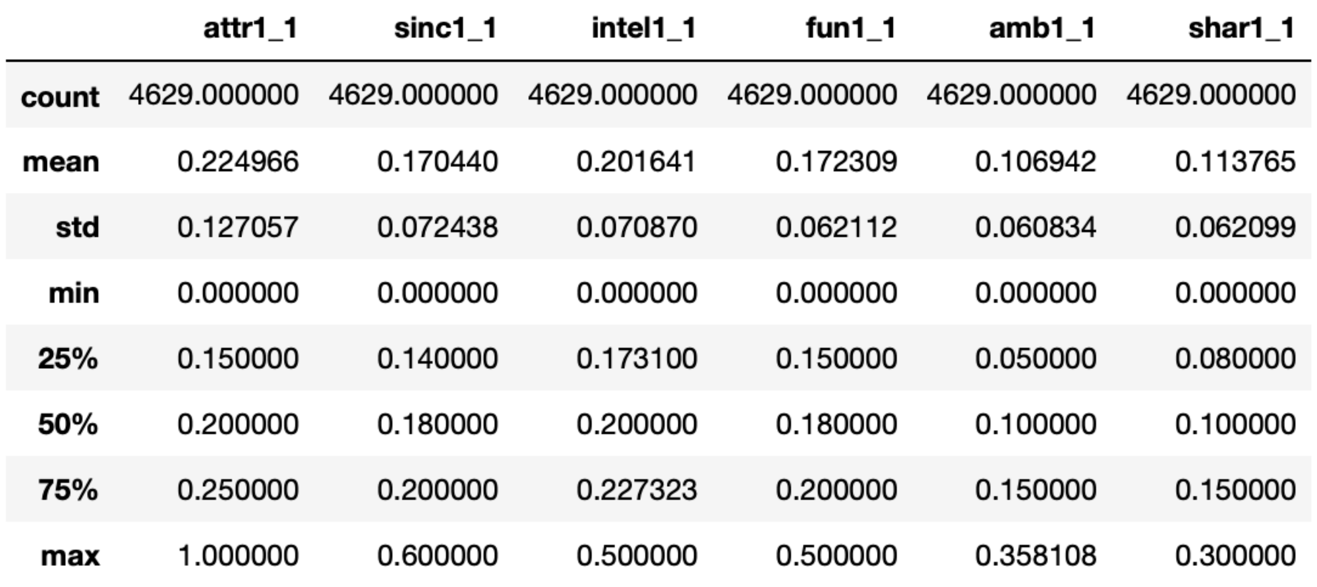
EXPLORATORY DATA ANALYSIS

5. Summary statistics for attributes importance at sign up:



	attr1_1	sinc1_1	intel1_1	fun1_1	amb1_1	shar1_1
count	3501.000000	3501.000000	3501.000000	3501.000000	3501.000000	3501.000000
mean	0.220708	0.175060	0.200642	0.172801	0.104025	0.120196
std	0.127457	0.071467	0.069375	0.063975	0.060419	0.067829
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.150000	0.150000	0.177782	0.147071	0.050000	0.080000
50%	0.200000	0.189219	0.200000	0.179518	0.100000	0.111111
75%	0.250000	0.200000	0.238100	0.200000	0.150000	0.169800
max	1.000000	0.600000	0.500000	0.500000	0.358108	0.300000

Summary statistics of attributes importance at sign up (decision = yes)



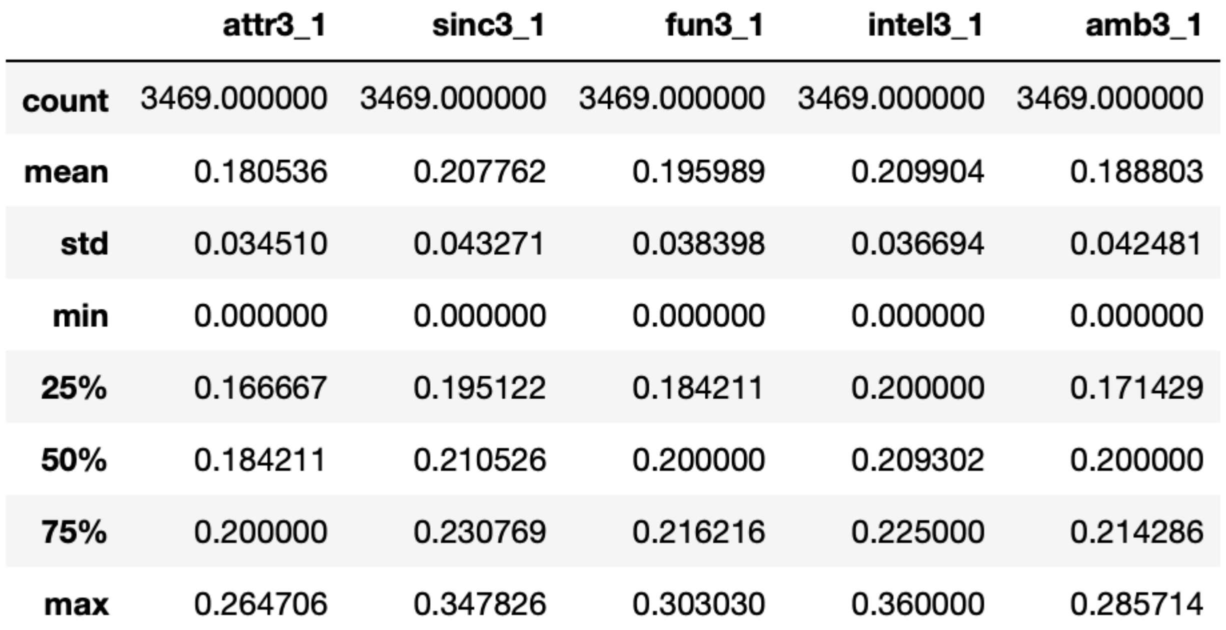
	attr1_1	sinc1_1	intel1_1	fun1_1	amb1_1	shar1_1
count	4629.000000	4629.000000	4629.000000	4629.000000	4629.000000	4629.000000
mean	0.224966	0.170440	0.201641	0.172309	0.106942	0.113765
std	0.127057	0.072438	0.070870	0.062112	0.060834	0.062099
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.150000	0.140000	0.173100	0.150000	0.050000	0.080000
50%	0.200000	0.180000	0.200000	0.180000	0.100000	0.100000
75%	0.250000	0.200000	0.227323	0.200000	0.150000	0.150000
max	1.000000	0.600000	0.500000	0.500000	0.358108	0.300000

Summary statistics of importance attributes at sign up (decision = no)

- Participants who said ‘yes' to a 2nd date have higher mean for ‘Sincerity’, ‘Fun’, ‘Shared Interests’. Participants that said ‘No' to a 2nd date have higher mean for ‘Attractive’, ‘Intelligence’ and ‘Ambitious'.

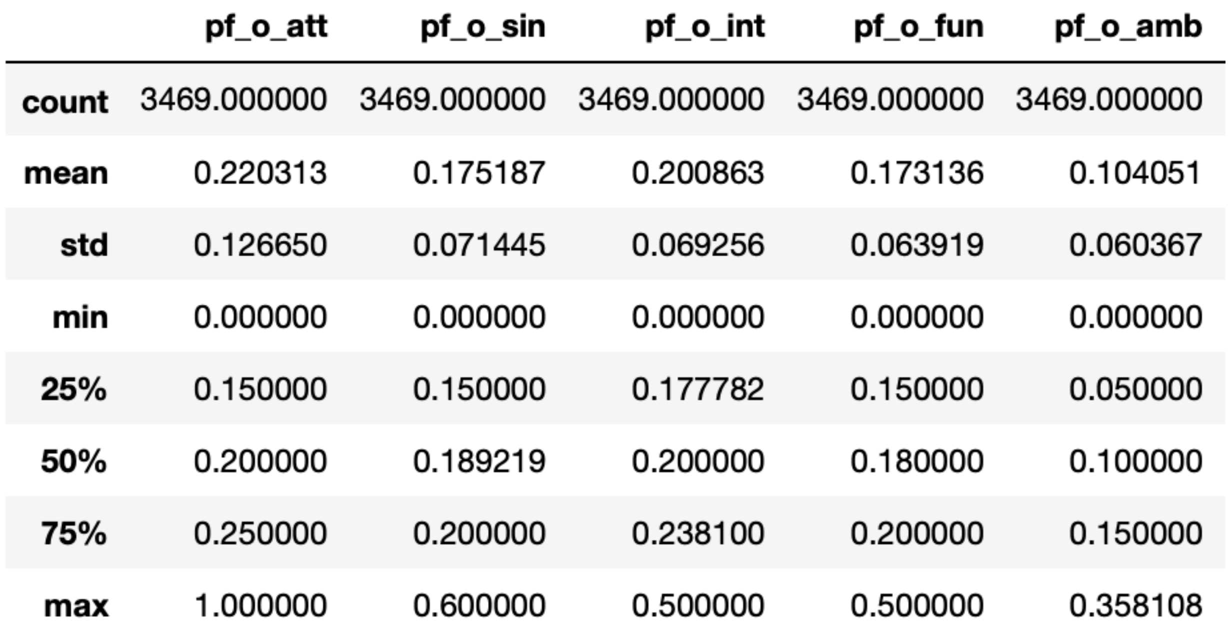
EXPLORATORY DATA ANALYSIS

6. Summary statistics of Self Evaluation vs. how they were rated by their partner when partner's decision for a 2nd date was 'Yes':



	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 partner's decision Yes



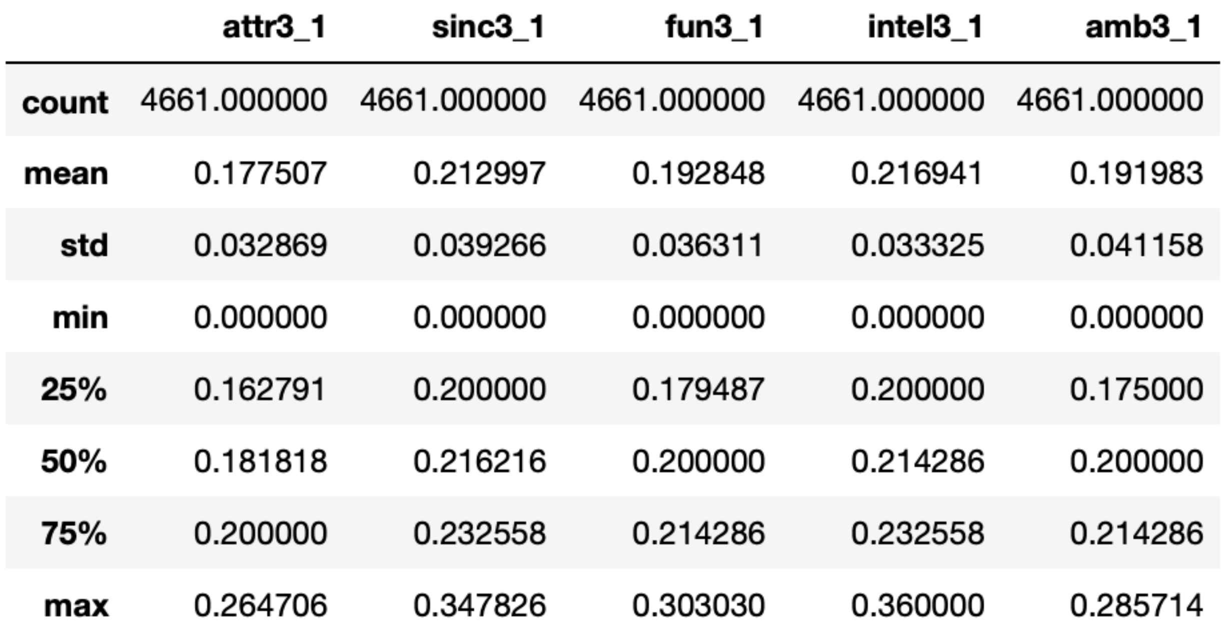
	pf_o_att	pf_o_sin	pf_o_int	pf_o_fun	pf_o_amb
count	3469.000000	3469.000000	3469.000000	3469.000000	3469.000000
mean	0.220313	0.175187	0.200863	0.173136	0.104051
std	0.126650	0.071445	0.069256	0.063919	0.060367
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.150000	0.150000	0.177782	0.150000	0.050000
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 partner's decision Yes

- Participants gave a lower rating for themselves in comparison to the ratings given to them by their partners during the event for Attractiveness. For all other attributes participants gave them higher ratings in comparison to the ratings given to them.

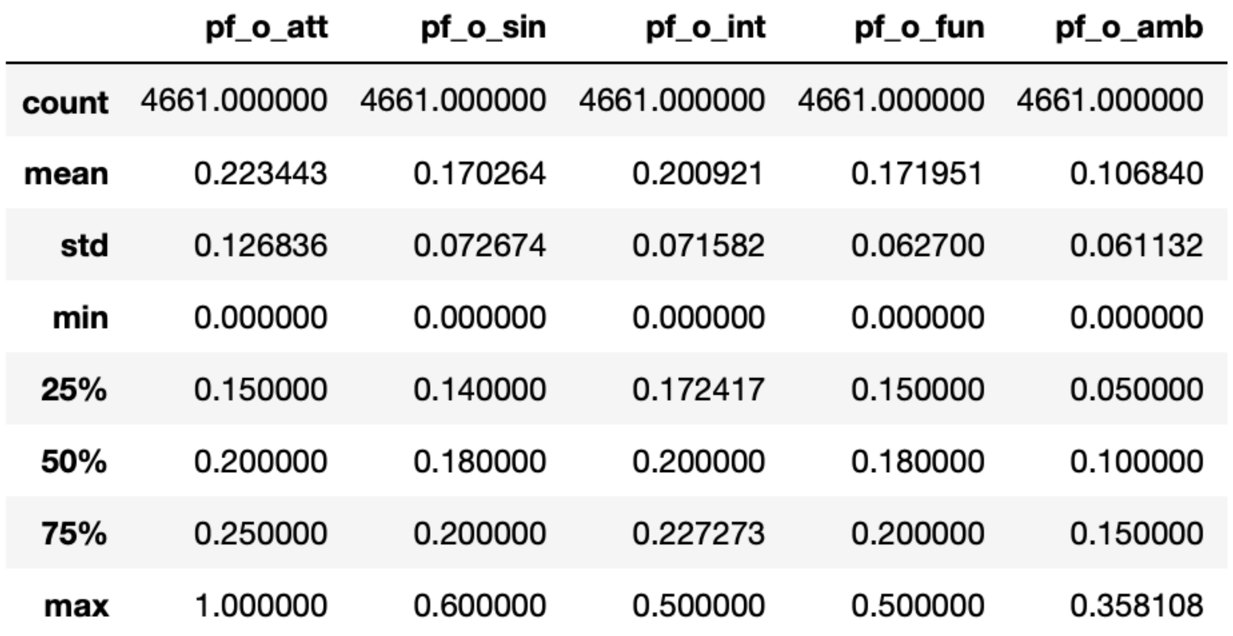
EXPLORATORY DATA ANALYSIS

7. Summary statistics of Self Evaluation vs how they were rated by their partner when partner’s decision for a 2nd date was ‘No’:



	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 partner’s 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 their partners for partner’s 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.

EXPLORATORY DATA ANALYSIS

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

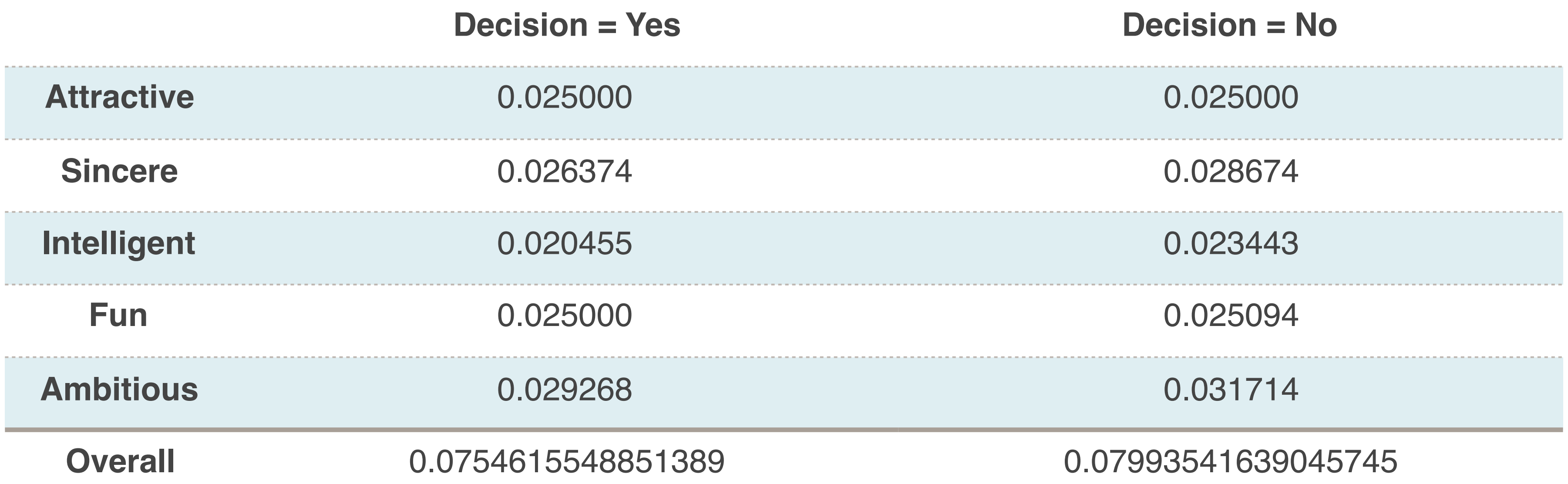
	Decision = Yes	Decision = No
Attractive	0.098995	0.100000
Sincere	0.050000	0.050000
Intelligent	0.050000	0.050000
Fun	0.050000	0.050000
Ambitious	0.050000	0.050000
Shared Interests	0.056812	0.050000
Overall	0.2017422577816702	0.21213203435596428

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

- 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.

EXPLORATORY DATA ANALYSIS

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



	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 between both participants at sign-up

- The distance median for all attributes is lower for participants that had a match, 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.

EXPLORATORY DATA ANALYSIS

Insights from EDA:

- Attributes that most impact like positively: Shared interests, Fun and Attractive
- Attributes that mostly impact like positively 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 had a match
- The distance for self evaluation between 2 participants is smaller for participants that had a match
- 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

The following models using Random Forest were evaluated:

1. Predicting match based on attributes importance at sign up
2. Predicting match based on distance of attributes importance between 2 participants
3. Predicting match based on distance of self evaluation between 2 participants
4. Predicting match attributes importance and self evaluation between 2 participants

Model 4 has highest precision, recall and f1-score.

Model 1

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

Model 2

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

Model 3

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

Model 4

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

MACHINE LEARNING MODEL

Description:

➤ Participants 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.

➤ 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 distances between what participants consider important (in each attribute) and the distance of how they see themselves (in each attribute) drive a match.

➤ Collinearity was investigated to confirm is all these predictors could be used together in the model. Attributes importance distance and self evaluation distance have low correlation for all the attributes, so they can be used as predictor variables.

MACHINE LEARNING MODEL

Model details:

➤ 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 (in scikit-learn)

➤ Hyper parameter optimization: RandomizedSearchCV resulted in the following best parameters:

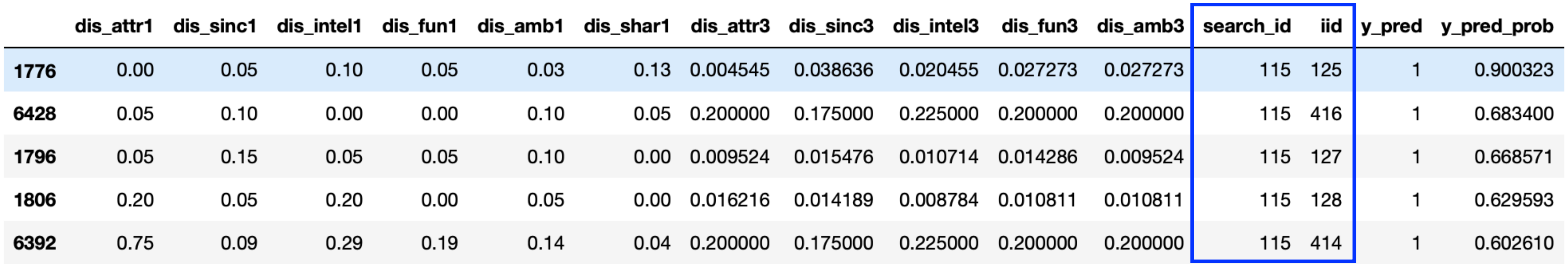
'max_depth': 30, 'max_features': 3, 'n_estimators': 107

➤ Features: 11 distances (6 for attributes importance and 5 for self evaluation)

MACHINE LEARNING MODEL

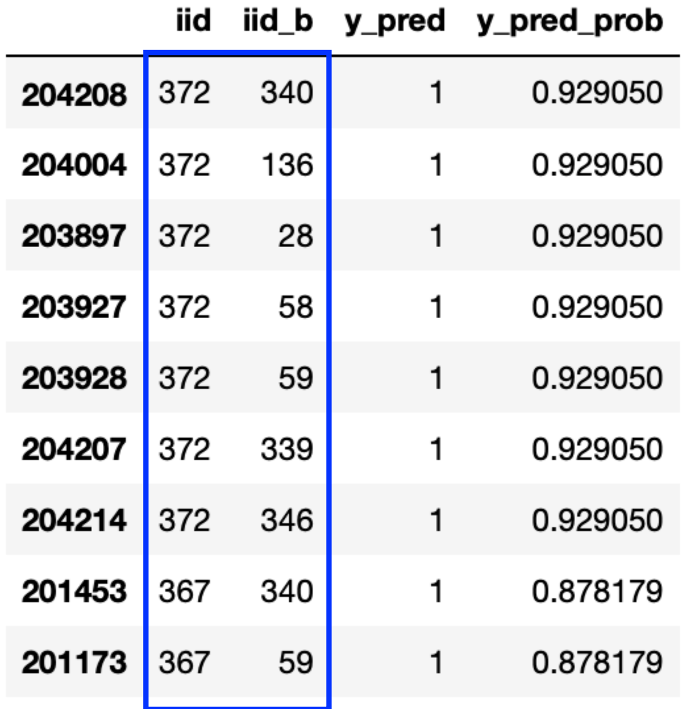
Predictions possibilities:

➤ given one participant (search_id = 115), find other participants (iid) that would give a good match and only return the first matches (5 in the example below).



	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 and iid_b):



	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

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)