Kavindra Sahabir
CSC 590

CSC Final Project

In our knowledge of the animal kingdom, every animal has their own way of mating. Peacocks have mating rituals that involve showing off their bodies, birds have elaborate songs that they sing, and humans have the recent invention of speed dating. A phenomenon that is a direct result of our desires to be as efficient as possible, speed dating involves a short 4 to 5 minute conversation between two participants who have the steep objective of sussing out whether or not they are interested in each other in the same amount of time it takes to get coffee from Starbucks. My interest is in seeing what kind of attributes lead to speed dating participants deciding that they've built up enough interest to go on another date. Therefore, I chose to perform data analysis and classification of a dataset containing data gathered from participants in speed dating events from 2002 to 2004. The dataset contained more than 100 attributes, some of which measured the participants' ratings of their dates based on six attributes: attractiveness, sincerity, intelligence, fun, ambition, and shared interests. Other attributes include demographic data like race and field of work.

My analysis revolved around one question: what attributes could be used to predict whether or not a pair would be likely to have a second date?

The first step I took towards answering this question was to find the correlation between all the different attributes of the dataset. I found the Pearson correlation coefficient for each of the relationships between the attributes, and as an output, got a list of correlation coefficients as well as the attributes that have this correlation. Based on this list, I edited the DataFrame to include only these selected attributes. The attributes included are: the ages of the participants per date, the difference in ages of the participants per date, and the rankings of attractiveness, sincerity, intelligence, fun (humor), ambition, and shared interests that each participant in each date gave to each other (where attributes ending in _o represent rankings given to participant 2 from participant 1 and _partner are rankings given to participant 1 from participant 2). Also included are whether or not the participants per date were of the same race, and a measure of how much the interests of both participants were related (this attribute was included with the dataset and not calculated independently). These specific attributes are:
gender, age, age_o, d_age, samerace, match, shared_interests_o, shared_interests_partner, funny_o, funny_partner, attractive_o, attractive_partner, intelligence_partner, intelligence_o, sincere_partner, sinsere_o, ambition, ambitous_o, interests_correlate.

With the collection of all this data, there was bound to be missing values especially in the context of something as personal as dating. However, in order to analyze the dataset in the context of the initial question, I needed to have only fully complete data from as many dates as possible. I considered imputing the values, but decided that would be inappropriate as the data

in question is categorical rather than continuous. Imputation could have been performed by replacing missing values with the most common values per attribute. I decided against this in recognition of the uniqueness of the values that would come from each participant in each date. Therefore, I went ahead and dropped any rows with incomplete data. This left me with a dataset of 1048 rows, which I considered to be a big enough dataset to go ahead and analyze.

Fig. 1

	Index	Index	0
1	museums	art	0.863136
2	funny_partner	like	0.703506
3	like	shared_interests_partner	0.682596
4	shared_interests_o	funny_o	0.673013
5	like	attractive_partner	0.668993
6	shared_interests_partner	funny_partner	0.666318
7	intelligence_o	sinsere_o	0.652441
8	movies	music	0.637356
9	intelligence_o	ambitous_o	0.632021
10	sincere_partner	intelligence_partner	0.630665
11	music	concerts	0.626656
12	museums	theater	0.622002
13	funny_o	attractive_o	0.619802
14	art	theater	0.616101
15	tvsports	sports	0.609148
16	ambition_partner	intelligence_partner	0.598927
17	importance_same_religion	importance_same_race	0.578448
18	attractive_partner	funny_partner	0.57767
19	decision	match	0.542849
20	decision	like	0.538018
21	decision_o	match	0.533405
22	shared_interests_o	attractive_o	0.527187
23	ambition	intelligence	0.523992
24	decision_o	attractive_o	0.522838

A sample of the list of correlation coefficients generated. The first two columns show the attributes associated with each correlation coefficient

After finding this correlation data and editing the dataframe to "zoom in" on the relationships highlighted here, I wanted to analyze any differences in the correlations based on the gender of the initial participants. To this end, I created visual representations of the correlations to see what differences might exist.

Fig. 2
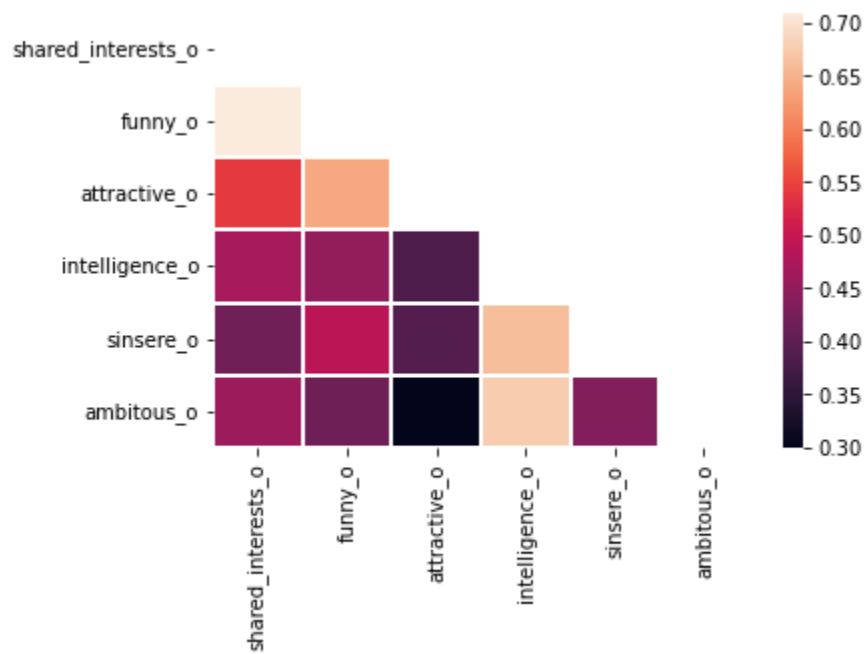
The visualization of the correlations between selected attributes for participants whose sex was specified to be males.

Fig. 3
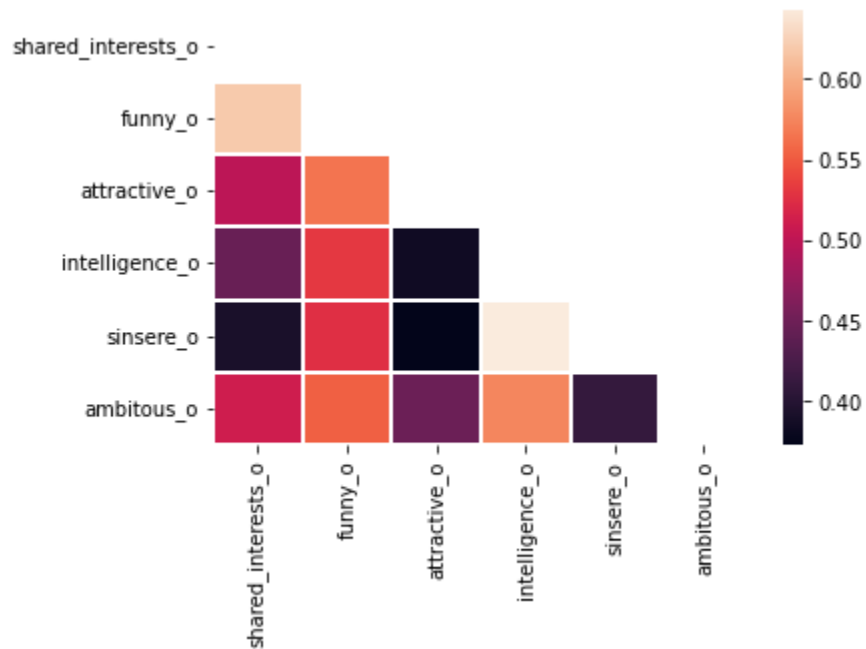
A visualization of the correlations between selected attributes for participants whose sex was specified to be females.

As seen from the heatmaps, there were some attributes that seemed to be just as significant for males as they are for females. An example is the correlation between a participant's ranking in terms of funniness and their ranking in terms of how much that participant shares interests with their partner. There doesn't seem to be attributes that have significantly different correlations for the two sexes measured.

After these correlations were discovered, a Naive Bayes classifier was used on this dataset with the hopes of confirming that the attributes included in the edited dataframe are indeed good for predicting whether or not the participants would indicate a willingness for a second date. As hoped, the classifier resulted in fairly high accuracies around 85% with a maximum accuracy of 89.5%.

Fig. 4

df_norm - DataFrame

Index	gender	age	age_o	d_age	samerace	match	...red_interest	..._interests_p	funny_o	...unny_partne	attractive_o	...ractive_partn	...lligence_part	...ntelligence_c	...ncere_partne	sinsere_...
0	0	0.176471	0.529412	0.428571	0	0	0.6	0.5	0.8	0.7	0.6	0.6	0.7	0.8	0.9	0.8
1	0	0.176471	0.235294	0.0714286	0	0	0.5	0.6	0.7	0.8	0.7	0.7	0.7	1	0.8	0.8
3	0	0.176471	0.294118	0.142857	0	1	0.8	0.8	0.8	0.7	0.7	0.7	0.8	0.9	0.6	0.8
4	0	0.176471	0.352941	0.214286	0	1	0.7	0.6	0.6	0.7	0.8	0.5	0.7	0.9	0.6	0.7
5	0	0.176471	0.411765	0.285714	0	0	0.7	0.4	0.8	0.4	0.7	0.4	0.7	0.8	0.9	0.7
6	0	0.176471	0.705882	0.642857	0	0	0.7	0.7	0.5	0.4	0.3	0.7	0.7	0.7	0.6	0.6
8	0	0.176471	0.588235	0.5	0	1	0.9	0.8	0.8	0.9	0.7	0.7	0.8	0.8	0.6	0.7
9	0	0.176471	0.352941	0.214286	0	0	0.6	0.8	0.6	0.8	0.6	0.5	0.6	0.6	0.6	0.6
10	0	0.352941	0.529412	0.214286	1	0	0.4	0.3	0.9	0.4	0.8	0.5	0.8	0.6	0.7	0.7
11	0	0.352941	0.235294	0.142857	1	0	0.5	0.6	0.6	0.6	0.7	0.8	0.6	1	0.5	0.6
12	0	0.352941	0.235294	0.142857	0	0	1	0.4	1	0.6	1	0.5	0.9	1	0.8	1
13	0	0.352941	0.294118	0.0714286	1	1	0.9	0.7	0.9	0.6	0.9	0.7	0.7	0.9	0.9	0.9
14	0	0.352941	0.352941	0	0	0	0.8	0.8	1	0.9	1	0.6	0.7	1	0.8	1
15	0	0.352941	0.411765	0.0714286	1	0	0.7	0.2	0.5	0.3	0.7	0.8	0.8	0.7	0.7	0.8
16	0	0.352941	0.705882	0.428571	1	0	0.3	0.9	0.3	0.6	0.5	0.7	0.9	0.4	0.5	0.3
17	0	0.352941	0.529412	0.214286	1	0	0.5	0.5	0.7	0.5	0.7	0.5	0.7	0.7	0.8	0.7

A sample of the edited dataframe, which was normalized per attribute and includes the attributes discussed above

We then used the .argmax function to select the highest probability and generate a list of predictions according to the class with the highest probability per row in the testing dataframe.

The results of this analysis shows some nicely insightful data that gives a good indication of what speed dating participants prioritize when considering their willingness to go on a second date. For one, the strongest correlation (that makes sense in the context of the attributes considered) is between the funninness of participant 2 and whether or not the participant 1 would say they like participant 2. Another interesting correlation is between how much participant 2 believes they share in terms of interests, and whether or not they like their partner. Another interesting insight is that one's intelligence and their perceived sincerity are also well correlated. Also included is the cliche that the attractiveness of a person and whether or not they are liked, which seems to be backed up by the correlation shown between those two attributes. It would seem that, even more than attractiveness, speed dating participants prize having shared interests and having partners who have a nice sense of humor. This signals a shift away from dated ideas in what humans look for in partners, although to be fair this data is collected from participants in a first world country. The results of this dataset's analysis may not hold up in different countries with different societies/different standards of qualities to look for in partners.