SPEED DATING DATASET

VIGNESH J MURALIDHARAN

Introduction: In the growing fastness of this world youths are looking for a partner to get marry through internet and dating websites. To reveal these how youth are looking for their partners we have a dataset from one of the online website to analyze and see. Speed dating is a formalized matchmaking process whose purpose is to encourage eligible singles to meet large numbers of new potential partners in a very short period. It was first created in 1998 by a Los Angeles based television. Over 100 companies in the US have created the websites to register and take survey on their experience and one of the companies' website dataset is what I am going to analyze.

Data Reference: Open ML website : http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating/

Relevant paper: Raymond Fisman; Sheena S. Iyengar; Emir Kamenica; Itamar Simonson. Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment. The Quarterly Journal of Economics, Volume 121, 1 May 2006

Data Information: This data was gathered from participants from 2002-2004 during the event the attendees would have 4 min "first date" with every other participant of the opposite sex. At the end of their 4 min, participants were asked if they would like to see their date again. They were also asked to rate their date on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition and shared interest. The dataset also includes questionnaire data gathered from participants at different points in the process. These fields includes demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in the partner and lifestyle information.

Dim (data) => 8378 rows * 123 features

Attributes of the data: Similar Questions were asked with the partner also and their responses also looks similar

1. Gender
2. age: Age of self
3. age_o: Age of partner
4. d_age: Difference in age
5. race: Race of self
6. race_o: Race of partner
7. samerace: same race or not.
8. importance_same_race: partner is of same race?
9. importance_same_religion: partner has same religion?
10. field: Field of study
11. pref_o_attractive: partner rate attractiveness
12. pref_o_sinsere: partner rate sincerity
13. pref_o_intelligence: partner rate intelligence
14. pref_o_funny: partner rate being funny
15. pref_o_ambitious: partner rate ambition
16. pref_o_shared_interests: partner rate having shared interests
17. attractive_o: Rating by partner (about me) on attractiveness
18. sincere_o: Rating by partner (about me) on sincerity
19. intelligence_o: Rating by partner (about me) on intelligence
20. funny_o: Rating by partner (about me) on being funny
21. ambitous_o: Rating by partner (about me) on being ambitious
22. shared_interests_o: by partner (about me) on shared interest
23. attractive_important: you look for in a partner - attractiveness
24. sincere_important: you look for in a partner – sincerity
25. intellicence_important: you look for in a partner – intelligence
26. funny_important: you look for in a partner - being funny
27. ambtition_important: you look for in a partner – ambition
28. shared_interests_important: you look for in a partner - shared interests
29. attractive: Rate yourself – attractiveness
30. sincere: Rate yourself – sincerity
31. intelligence: Rate yourself – intelligence
32. funny: Rate yourself - being funny
33. ambition: Rate yourself – ambition
34. attractive_partner: Rate partner – attractiveness
35. sincere_partner: Rate partner – sincerity
36. intelligence_partner: Rate partner – intelligence
37. funny_partner: Rate partner - being funny
38. ambition_partner: Rate partner – ambition
39. shared_interests_partner: Rate partner - shared interests
40. sports: Your own interests [1-10]
41. tvsports
42. exercise
43. dining
44. museums
45. art

46. hiking
47. gaming
48. clubbing
49. reading
50. tv
51. theater
52. movies
53. concerts
54. music
55. shopping
56. yoga
57. interests_correlate: Correlation between participant's and partner's ratings of interests.
58. expected_happy_with_sd_people: How happy do you expect to be with the people you meet during the event?
59. expected_num_interested_in_me: Out of the 20 people, how many do you expect will be interested in dating you?
60. expected_num_matches: How many matches do you expect to get?
61. like: Did you like your partner?
62. guess_prob_liked: How likely do you think it is that your partner likes you?
63. met: Have you met your partner before?
64. decision: Decision at night of event.
65. decision_o: Decision of partner at night of event.
66. match: Match (yes/no)

Analysis goal / Task:

- The dataset has categorial variables and so it will include lots of data cleaning with preprocessing techniques. Some of the variables will also include count vectorizer using the term frequency. This might make the dataset even a bit longer. Moreover, the dataset has some missing values so imputation techniques needs to be implemented to see the highest number of frequency in similar observations and impute them in the needed columns.
- Classification based on the target variable "match", "decision on both sides" will be made with different classification models.
- Clustering by removing the match variables or the results variable will be implemented and see if it really "match" the target.
- Association will help in this type of data of how the variables relate each other using arulesCBA having the "match" will also be analyzed.
- The comparison between the own and the partner rating will be analyzed if time is available.

Preliminary Results: Since lots of data cleaning is needed, I have started to work on that now.