 B03705006 侯舜元 B03705037 陳主牧

Speed Dating Experiment

• What is the data analytic problem?

The experiment we would lilke to reproduce is the following:

https://github.com/ColinLeverger/speed-dating-experiment-r/blob/master/BIBA-Colin-LEVERGER.pdf

This study is based on the Speed Dating Experiment dataset containing matching records in 21 waves. Speed dating is a new type of matchmaking which encourage people the match in a short period of time. The dataset records lots of feature about the date, such as gender, hobbit of participants, race of participants, and, most importantly, whether they matched in the end.

This study tried to find what variables are influential in a successful match and some interesting problems like “are people more willing to meet someone working in the same field?”. We will extend its model and use other techniques taught in class, like KNN, logistic regression to see if we can find more insight.

• Why the problem is interesting?

Today, many NTU students spend most of their time dating, devoting themselves in activities just to find their “soulmate”. Dcard, a combination of forum and matchmaking platform founded by NTU IM graduate, is great example showing how big dating and mating market is. Dcard somehow practices the spirit of speed dating and that is probably the season why it is so popular now. With Dcard as well as many mating softwares, like Tinder, Justdating, Paktor and more, finding people to date is no longer a hard problem. The problem is how can a person date successfully? What kind of personality or interest will be helpful when matching?

• What datasets are available?

This is dataset is available on kaggle:
https://www.kaggle.com/xanonimoxp/notebook-9d16aaca3f2c517b5603/data

There are 8378 observations and 195 features.

• Provide a short description of the datasets.

This dataset was compiled by Columbia Business School professors Ray Fisman and Sheena Iyengar and was gathered from participants in experimental speed dating events from 2002-2004. During the events, the attendees would have a four minute "first date" with every other participant of the opposite sex. After the four minute, the participants will be asked if they would like to see their date again, and also rate their date on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. Also, the dataset gathers different features about the attendees with questionnaire, such as demographics, dating habits, interests and more.

• Provide sample data of the datasets.

iid	id	gender	idg	condtn	wave	round	position	positin1	order	···	imprace	imprelig	from	zipcode	income	goal	date	go_out	career	ca
1	1	0	1	1	1	10	7	NA	4	···	2	4	Chicago	60,521	69,487.00	2	7	1	lawyer	
1	1	0	1	1	1	10	7	NA	3	···	2	4	Chicago	60,521	69,487.00	2	7	1	lawyer	
1	1	0	1	1	1	10	7	NA	10	···	2	4	Chicago	60,521	69,487.00	2	7	1	lawyer	
1	1	0	1	1	1	10	7	NA	5	···	2	4	Chicago	60,521	69,487.00	2	7	1	lawyer	
1	1	0	1	1	1	10	7	NA	7	···	2	4	Chicago	60,521	69,487.00	2	7	1	lawyer	
1	1	0	1	1	1	10	7	NA	6	···	2	4	Chicago	60,521	69,487.00	2	7	1	lawyer	

career_c	sports	tvsports	exercise	dining	museums	art	hiking	gaming	clubbing	···	tv	theater	movies	concerts	music	shopping	yoga	exphappy
NA	9	2	8	9	1	1	5	1	5	···	9	1	10	10	9	8	1	3
NA	9	2	8	9	1	1	5	1	5	···	9	1	10	10	9	8	1	3
NA	9	2	8	9	1	1	5	1	5	···	9	1	10	10	9	8	1	3
NA	9	2	8	9	1	1	5	1	5	···	9	1	10	10	9	8	1	3
NA	9	2	8	9	1	1	5	1	5	···	9	1	10	10	9	8	1	3
NA	9	2	8	9	1	1	5	1	5	···	9	1	10	10	9	8	1	3

There are 190+ different features for each data point, the detailed description for each column can be found on Kaggle.

• Provide a summary of what other people have done

The report have done studies about what activities make a man or woman more willing to please his/her date. The article used methods like random forest to find the most important features for different genders. We could apply different methods and take more features(not only activities) into consideration if we can find more insight.
Results from the report is listed below.