2016

PROJECT REPORT:

SPEED DATING EXPERIMENT

PRINCIPLES OF BUSINESS DATA MINING

GROUP 14: VIJETHA SHENOY, SANDEEP RAMESH, MOHAMED NAUMANI, MEGHANA RAI

Contents

DATA BACKGROUND

BACKDROP OF THE EXPERIMENT

Our experiment revolves around the concept of Speed Dating. Speed Dating is a combination of match making and dating, where people are encouraged to socialize with large number of people in short period. In this event, participant will meet another participant for short period (5 -6 mins) for a face to face conversation about any topic that interest them. Once time is up the participants must move on to the next participant. During their conversation, each participant makes a note of whether he would like to meet the person he/she interacted with again before he moves on the next participant. Speed Dating was originally started by Rabbi Yaacov Deyo of Aish HaTorah, as way to help Jewish singles meet and marry.

SPEED DATING EXPERIMENT

The dataset for Speed Dating was compiled by Prof.Ray Fisman and Prof.Sheena Iyengar from Columbia Business School for their paper Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment. Data was gathered from an experimental speed dating event from 2002 - 2004. The participants were asked to rate each participant they interact with based on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. Questionnaire was given to every participant asking them to answer questions about their characteristics, lifestyle and what they look in a potential partner. This was continued throughout the process at different points. This questionnaire is the source of our dataset. Through this experiment, we are trying to find out what attributes attracts two people and finding answers to the following ideas:

- What are the least desirable attributes in a male partner? Does this differ for female partners?
- How important do people think attractiveness is in potential mate selection vs. its real impact?
- How important is it that a person you date be of the same racial/ethnic background to get a match?
- In terms of getting a second date, is it better to be someone's first speed date of the night or their last?

DATASET & ATTRIBUTES

Source for the Speed Dating experiment is **Kaggle.com**. The dataset contains **68 attributes** and **131986 cells**. The class attribute is **Match** (0 /1)

The attributes are:

Attribute Name	Descriptions
age	Unique subject number
age_o	Subject number within wave
amb_o	Participants ambition
amb1_1	Importance of ambitious in Potential date(wave1)
amb2_1	Importance of ambitious in Potential date(wave2)
amb3_1	Importance of ambitious in Potential date(wave3)
art	Participants Interest in art
attr_o	Gender with range of choices
attr1_1	Importance of attractiveness in Potential date(wave1)
attr2_1	Importance of attractiveness in Potential date(wave2)
attr3_1	Importance of attractiveness in Potential date(wave3)
career	Survey made during meet and greet
clubbing	Station number where subject met the partner

concerts	Participants Interest in attending concerts
condtn	Participants limited choice or extensive choice
date	Station number where subjects met
dining	Partner’s IID number
exercise	Participants Interest in exercise
exphappy	Happiness quotient participants are with people they meet in the speed dating event?
from	Race of partner
fun_o	Participants fun level
fun1_1	Importance of fun in Potential date(wave1)
fun2_1	Importance of fun in Potential date(wave2)
fun3_1	Importance of fun in Potential date
gaming	Participants Interest in gaming
gender	Participant’s decision on the night of event
go_out	Rating by the participant on the night of event
goal	Age of participants
hiking	Participants interest in hiking
imprace	Importance of dating a partner of same race for the participant

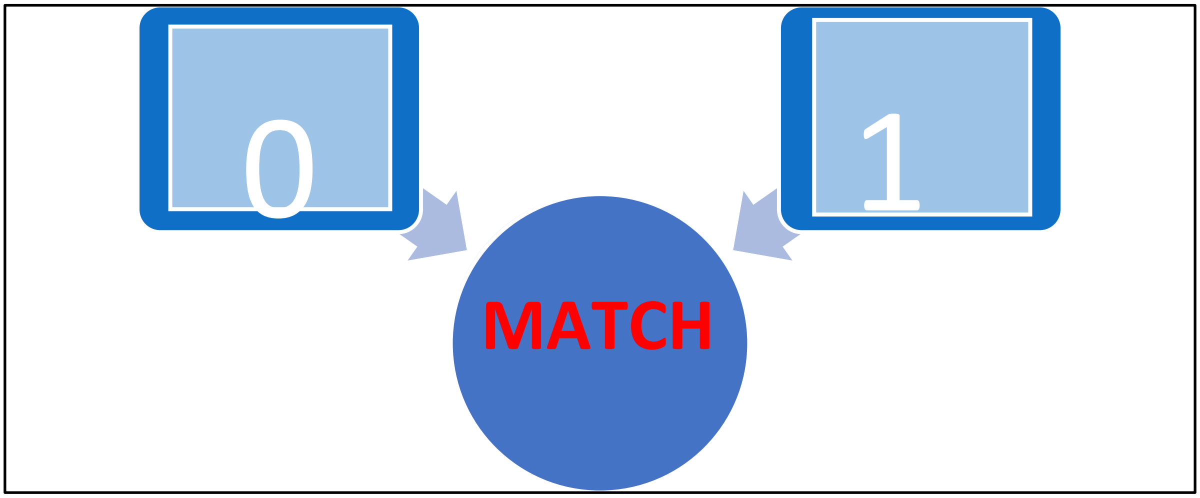
imprelig	Importance of religious background of partner for the participant
income	Median household income of participant
intel_o	Participants intelligence level
intel1_1	Importance of intelligence in Potential date (wave1)
intel2_1	Importance of intelligence in Potential date(wave2)
intel3_1	Importance of intelligence in Potential date(wave3)
length	Frequency of going on dates
Match	Whether the partner is a **Match** or not
movies	Participants interested in movies
museums	Participants interested in visiting museums
music	Participants interested in music
order	Interest in playing sports
pf_o_amb	Interest in ambitious
pf_o_att	Interest in reading
pf_o_fun	Interest in fun
pf_o_int	Interest in intelligence
pf_o_sha	Interest in shared interest

pf_o_sin	Interest in sincerity
position	Overall happiness rating during speed dating event
race	Race of participant
race_o	Race of partner
reading	Level of Sincerity
round	Level of intelligence
samerace	Level of fun
shar_o	Shared interest of partner
shar1_1	Importance of shared interest in Potential date (wave1)
shar2_1	Importance of shared interest in Potential date(wave2)
shopping	Participants Interest in shopping
sinc_o	Sincerity of partner
sinc1_1	Importance of sincerity in Potential date (wave1)
sinc2_1	Importance of sincerity in Potential date(wave2)
sinc3_1	Importance of sincerity in Potential date(wave3)
sports	Length of speed dating
theater	Participants Interest in theatre

tv	Participants Interest in TV
tvsports	Participants Interest in TV sports
wave	Calls made by the participant
yoga	Participants Interest in yoga

CLASS ATTRIBUTE

The class attribute is **Match** which comprises of two nominal values, 0 = **No Match** and 1 = **Match** using different attributes from the attribute list.



DATA CLEANING AND TRANSFORMATION

As the dataset comes from real world, data quality becomes an important factor. Since the data is collected by survey forms filled out by the participants, it is likely to contain one or more data quality issues like noise, missing values, data entry errors, data migration and compilation errors. To overcome this, we will be using few data cleaning techniques which we will discuss in detail.

Data cleaning is the process which involves identifying inaccurate, incomplete, corrupt instances of data from the dataset and then replacing, modifying, or deleting those instances from the dataset.

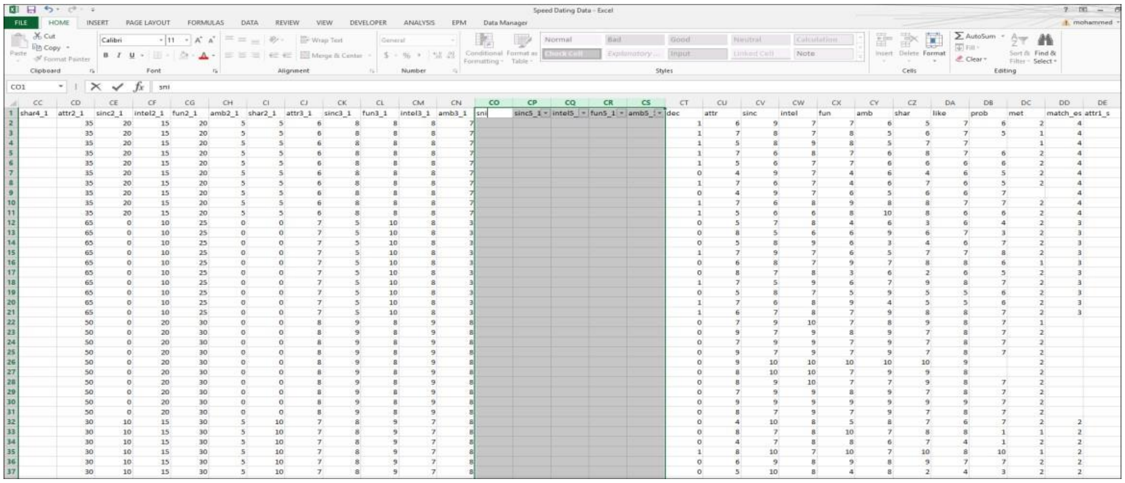
After the process of data cleaning, the dataset will be consistent with other similar instances in the dataset. The inconsistencies detected or removed may have been caused due to user entry errors or during transmission or storage of the dataset.

During the actual data cleaning, the process itself may involve removing errors and validating the dataset against a known list of entities. The validation may be strict (such as rejecting instances that don't match known records) or lenient (such as correcting instances that partially match existing or known records). Some data cleaning solutions involve cross checking the dataset with an already validated dataset. A common technique in data cleaning is data enhancement, where data is made complete by adding enhancements.

DATA CLEANING TOOLS

We have used the below tools for performing cleansing of the data,

- **Microsoft Excel**
- **WEKA**



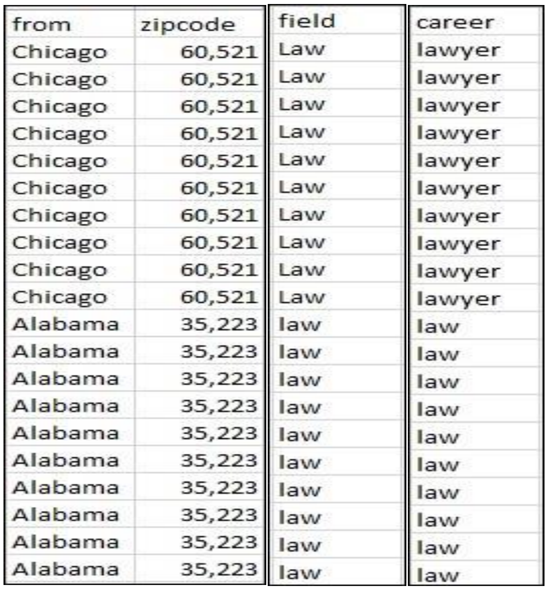
CO1 | *fx* sni

	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA	DB	DC	DD	DE
1	shar4_1	attr2_1	sinc2_1	intel2_1	fun2_1	amb2_1	shar2_1	attr3_1	sinc3_1	fun3_1	intel3_1	amb3_1	sni	sinc5_1	intel5_	fun5_1	amb5_	dec	attr	sinc	intel	fun	amb	shar	like	prob	met	match_es	attr1_s
2		35	20	15	20	5	5	6	8	8	8	7						1	6	9	7	7	6	5	7	6	2	4	
3		35	20	15	20	5	5	6	8	8	8	7						1	7	8	7	8	5	6	7	5	1	4	
4		35	20	15	20	5	5	6	8	8	8	7						1	5	8	9	8	5	7	7		1	4	
5		35	20	15	20	5	5	6	8	8	8	7						1	7	6	8	7	6	8	7	6	2	4	
6		35	20	15	20	5	5	6	8	8	8	7						1	5	6	7	7	6	6	6	6	2	4	
7		35	20	15	20	5	5	6	8	8	8	7						0	4	9	7	4	6	4	6	5	2	4	
8		35	20	15	20	5	5	6	8	8	8	7						1	7	6	7	4	6	7	6	5	2	4	
9		35	20	15	20	5	5	6	8	8	8	7						0	4	9	7	6	5	6	6	7		4	
10		35	20	15	20	5	5	6	8	8	8	7						1	7	6	8	9	8	8	7	7	2	4	
11		35	20	15	20	5	5	6	8	8	8	7						1	5	6	6	8	10	8	6	6	2	4	
12		65	0	10	25	0	0	7	5	10	8	3						0	5	7	8	4	6	3	6	4	2	3	
13		65	0	10	25	0	0	7	5	10	8	3						0	8	5	6	6	9	6	7	3	2	3	
14		65	0	10	25	0	0	7	5	10	8	3						0	5	8	9	6	3	4	6	7	2	3	
15		65	0	10	25	0	0	7	5	10	8	3						1	7	9	7	6	5	7	7	8	2	3	
16		65	0	10	25	0	0	7	5	10	8	3						0	6	8	7	9	7	8	8	6	1	3	
17		65	0	10	25	0	0	7	5	10	8	3						0	8	7	8	3	6	2	6	5	2	3	
18		65	0	10	25	0	0	7	5	10	8	3						1	7	5	9	6	7	9	8	7	2	3	
19		65	0	10	25	0	0	7	5	10	8	3						0	5	8	7	5	9	5	5	6	2	3	
20		65	0	10	25	0	0	7	5	10	8	3						1	7	6	8	9	4	5	5	6	2	3	
21		65	0	10	25	0	0	7	5	10	8	3						1	6	7	8	7	9	8	8	7	2	3	
22		50	0	20	30	0	0	8	9	8	9	8						0	7	9	10	7	8	9	8	7	1		
23		50	0	20	30	0	0	8	9	8	9	8						0	9	7	9	8	9	7	8	7	2		
24		50	0	20	30	0	0	8	9	8	9	8						0	7	9	9	7	9	7	8	7	2		
25		50	0	20	30	0	0	8	9	8	9	8						0	9	7	9	7	9	7	8	7	2		
26		50	0	20	30	0	0	8	9	8	9	8						0	9	10	10	10	10	10	9		2		
27		50	0	20	30	0	0	8	9	8	9	8						0	8	10	10	7	9	9	8		2		
28		50	0	20	30	0	0	8	9	8	9	8						0	8	9	10	7	7	9	8	7	2		
29		50	0	20	30	0	0	8	9	8	9	8						0	7	9	9	8	9	7	8	7	2		
30		50	0	20	30	0	0	8	9	8	9	8						0	9	9	9	9	9	9	9	7	2		
31		50	0	20	30	0	0	8	9	8	9	8						0	8	7	9	7	9	7	8	7	2		
32		30	10	15	30	5	10	7	8	9	7	8						0	4	10	8	5	8	7	6	7	2	2	
33		30	10	15	30	5	10	7	8	9	7	8						0	8	7	8	10	7	8	8	1	1	2	
34		30	10	15	30	5	10	7	8	9	7	8						0	4	7	8	8	6	7	4	1	2	2	
35		30	10	15	30	5	10	7	8	9	7	8						1	8	10	7	10	7	10	8	10	1	2	
36		30	10	15	30	5	10	7	8	9	7	8						0	6	9	8	9	8	9	7	7	2	2	
37		30	10	15	30	5	10	7	8	9	7	8						0	5	10	8	4	8	2	4	3	2	2	

IRRELEVANT ATTRIBUTES

The dataset also contains irrelevant attributes that do not add any value to the information that is present.

In this case, we had attributes "from" and "zipcode" which gave information about the participant's origin city. Similarly, we also had attributes "field" and "career" which gave us information about the participant's occupation details. In such cases, we removed one attribute and maintained the other.

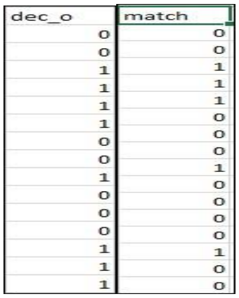


from	zipcode	field	career
Chicago	60,521	Law	lawyer
Chicago	60,521	Law	lawyer
Chicago	60,521	Law	lawyer
Chicago	60,521	Law	lawyer
Chicago	60,521	Law	lawyer
Chicago	60,521	Law	lawyer
Chicago	60,521	Law	lawyer
Chicago	60,521	Law	lawyer
Chicago	60,521	Law	lawyer
Chicago	60,521	Law	lawyer
Alabama	35,223	law	law
Alabama	35,223	law	law
Alabama	35,223	law	law
Alabama	35,223	law	law
Alabama	35,223	law	law
Alabama	35,223	law	law
Alabama	35,223	law	law
Alabama	35,223	law	law
Alabama	35,223	law	law
Alabama	35,223	law	law

FALSE PREDICTORS

False predictors, also called as information leakers, are fields that appear to predict the outcome and record the outcome of the class attribute that is being predicted even before the event occurs.

A simple way to identify false predictors is finding the attribute accuracy using OneR. This method uses each of the attributes in the dataset individually with the class attribute and checks if prediction accuracy is near 100%. If the prediction accuracy is near 100% that means, there are still some false predictors.



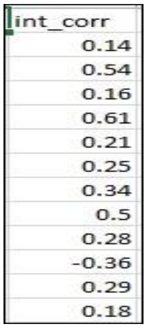
dec_o	match
0	0
0	0
1	1
1	1
1	1
1	0
0	0
0	0
1	1
0	0
0	0
0	0
1	1
1	0
1	0

In this case, using the above technique we could find two false predictors which gave the same values as the class attribute(“**Match**”).

The attributes identified were

- **dec_o**
- **dec**

Another way of identifying false predictors is by using Domain Analysis. Domain Analysis requires domain knowledge related to the attributes present in the dataset. Usually the data miner will collaborate with domain experts to understand the dataset. In this case, the attribute "int_corr**"** field gives a value in the range from -1 to 1 which is based on a combined score of all the shared attributes between two participants, therefore acting as a false predictor.



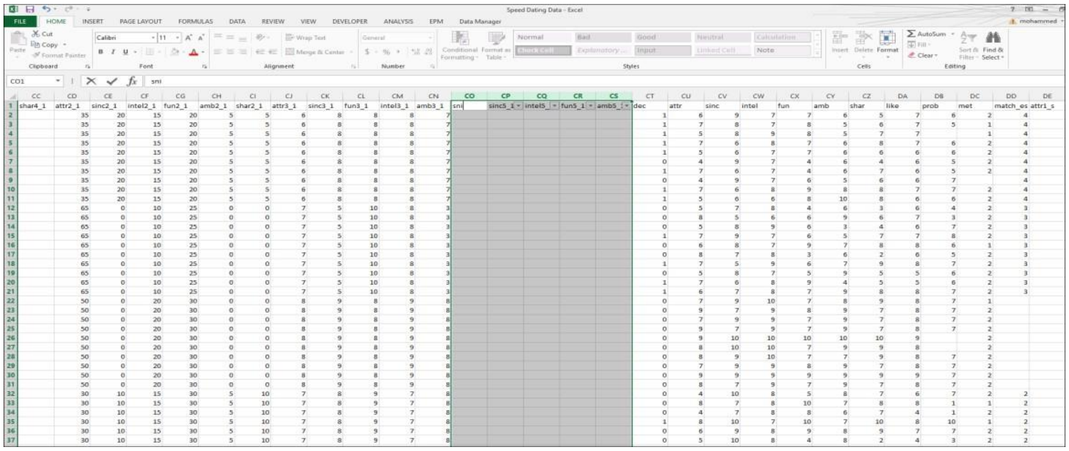
int_corr
0.14
0.54
0.16
0.61
0.21
0.25
0.34
0.5
0.28
-0.36
0.29
0.18

MISSING VALUES

In the original dataset, we found many missing and incomplete values. We tried running the algorithms with these missing values in the dataset but it affected the prediction accuracy of all the algorithms.

We used the below methods to deal with missing values:

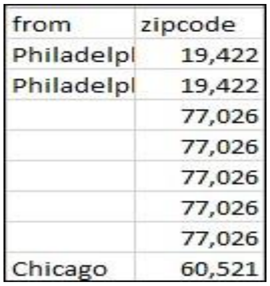
- **Attribute Deletion:** We identified attributes which had majority missing values and tried to predict the missing values for these attributes. Since most of the attributes values in the dataset came from participant's response to survey, we cannot predict such values. Hence, we deleted such attributes.



	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN
1	shar4_1	attr2_1	sinc2_1	intel2_1	fun2_1	amb2_1	shar2_1	attr3_1	sinc3_1	fun3_1	intel3_1	amb3_1
2		35	20	15	20	5	5	6	8	8	8	7
3		35	20	15	20	5	5	6	8	8	8	7
4		35	20	15	20	5	5	6	8	8	8	7
5		35	20	15	20	5	5	6	8	8	8	7
6		35	20	15	20	5	5	6	8	8	8	7
7		35	20	15	20	5	5	6	8	8	8	7
8		35	20	15	20	5	5	6	8	8	8	7
9		35	20	15	20	5	5	6	8	8	8	7
10		35	20	15	20	5	5	6	8	8	8	7
11		35	20	15	20	5	5	6	8	8	8	7
12		65	0	10	25	0	0	7	5	10	8	3
13		65	0	10	25	0	0	7	5	10	8	3
14		65	0	10	25	0	0	7	5	10	8	3
15		65	0	10	25	0	0	7	5	10	8	3
16		65	0	10	25	0	0	7	5	10	8	3
17		65	0	10	25	0	0	7	5	10	8	3
18		65	0	10	25	0	0	7	5	10	8	3
19		65	0	10	25	0	0	7	5	10	8	3
20		65	0	10	25	0	0	7	5	10	8	3
21		65	0	10	25	0	0	7	5	10	8	3
22		50	0	20	30	0	0	8	9	8	9	8
23		50	0	20	30	0	0	8	9	8	9	8
24		50	0	20	30	0	0	8	9	8	9	8
25		50	0	20	30	0	0	8	9	8	9	8
26		50	0	20	30	0	0	8	9	8	9	8
27		50	0	20	30	0	0	8	9	8	9	8
28		50	0	20	30	0	0	8	9	8	9	8
29		50	0	20	30	0	0	8	9	8	9	8
30		50	0	20	30	0	0	8	9	8	9	8
31		50	0	20	30	0	0	8	9	8	9	8
32		30	10	15	30	5	10	7	8	9	7	8
33		30	10	15	30	5	10	7	8	9	7	8
34		30	10	15	30	5	10	7	8	9	7	8
35		30	10	15	30	5	10	7	8	9	7	8
36		30	10	15	30	5	10	7	8	9	7	8
37		30	10	15	30	5	10	7	8	9	7	8

	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA	DB	DC	DD	DE
1	sni	sinc5_1	intel5_	fun5_1	amb5_	dec	attr	sinc	intel	fun	amb	shar	like	prob	met	match_es	attr1_s
2						1	6	9	7	7	6	5	7	6	2	4	
3						1	7	8	7	8	5	6	7	5	1	4	
4						1	5	8	9	8	5	7	7		1	4	
5						1	7	6	8	7	6	8	7	6	2	4	
6						1	5	6	7	7	6	6	6	6	2	4	
7						0	4	9	7	4	6	4	6	5	2	4	
8						1	7	6	7	4	6	7	6	5	2	4	
9						0	4	9	7	6	5	6	6	7		4	
10						1	7	6	8	9	8	8	7	7	2	4	
11						1	5	6	6	8	10	8	6	6	2	4	
12						0	5	7	8	4	6	3	6	4	2	3	
13						0	8	5	6	6	9	6	7	3	2	3	
14						0	5	8	9	6	3	4	6	7	2	3	
15						1	7	9	7	6	5	7	7	8	2	3	
16						0	6	8	7	9	7	8	8	6	1	3	
17						0	8	7	8	3	6	2	6	5	2	3	
18						1	7	5	9	6	7	9	8	7	2	3	
19						0	5	8	7	5	9	5	5	6	2	3	
20						1	7	6	8	9	4	5	5	6	2	3	
21						1	6	7	8	7	9	8	8	7	2	3	
22						0	7	9	10	7	8	9	8	7	1		
23						0	9	7	9	8	9	7	8	7	2		
24						0	7	9	9	7	9	7	8	7	2		
25						0	9	7	9	7	9	7	8	7	2		
26						0	9	10	10	10	10	10	9		2		
27						0	8	10	10	7	9	9	8		2		
28						0	8	9	10	7	7	9	8	7	2		
29						0	7	9	9	8	9	7	8	7	2		
30						0	9	9	9	9	9	9	9	7	2		
31						0	8	7	9	7	9	7	8	7	2		
32						0	4	10	8	5	8	7	6	7	2	2	
33						0	8	7	8	10	7	8	8	1	1	2	
34						0	4	7	8	8	6	7	4	1	2	2	
35						1	8	10	7	10	7	10	8	10	1	2	
36						0	6	9	8	9	8	9	7	7	2	2	
37						0	5	10	8	4	8	2	4	3	2	2	

- **Attribute values based on other attributes:** In the dataset, we had more than one attribute conveying the same information. In such cases, we used one attribute to get the value of the other attribute which had missing values. This was done in case of “from” and “zipcode” attribute. With the help of “zipcode” we could correctly find the missing values for “from” since “zipcode” is unique for each location.

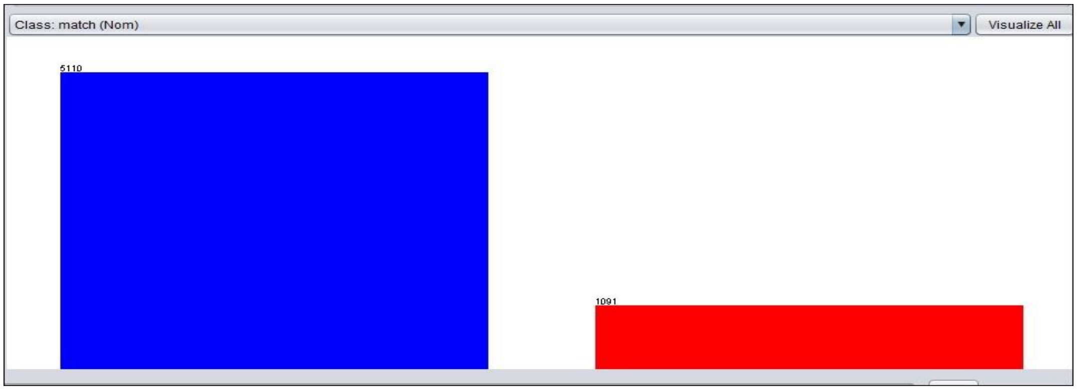


from	zipcode
Philadelpl	19,422
Philadelpl	19,422
	77,026
	77,026
	77,026
	77,026
	77,026
Chicago	60,521

SKEWED DATA

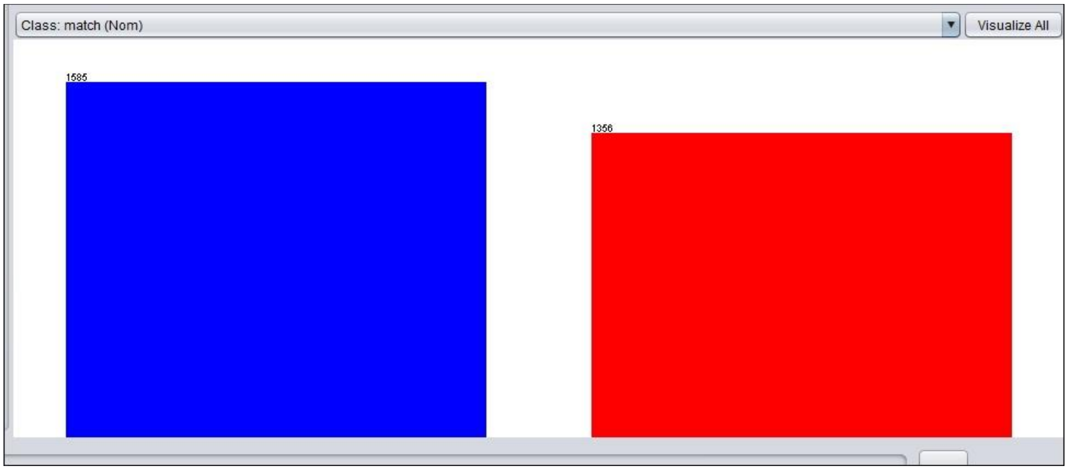
A dataset for modelling is perfectly balanced when the percentage of occurrence of each class is 100/n, where n is the number of classes. If one or more classes differ significantly from the others, this dataset is called skewed or unbalanced. Skewed Data may give high accuracy in training but this will decrease drastically while testing.

In the dataset, class imbalance was found between the two class attributes, “**Match**” (1) and “**No Match**” (0). This difference can be seen from the figure below.



Class: match (Nom)

The challenge we faced here was to balance the class values without losing the important properties of the data. Hence, we used the SMOTE (Synthetic Minority Oversampling Technique) filter in WEKA, which is based on oversampling the values of the minority class. This created a balance between the two class attributes. After data cleaning and application of SMOTE filter in WEKA, the values of the class attribute became balanced as seen below.



Class: match (Nom)

EXPERIMENT DESIGN

CLASSIFIER SELECTION

We have chosen the classifiers based on the comparison of accuracy with other models and the stability with which it is predicting. This gave rise to three models;

- **Naïve Bayes:** It is a classification technique based on the Bayes theorem. It is a probabilistic classifier with the assumption of independence among predictors. Despite its simplicity, it can outperform more sophisticated classification methods.
- **J48:** J48 classifier implements the C4.5 algorithm. The decision tree is formed by using the splitting criterion which is calculated based on the information gain. The values at the top of the tree form the most important attributes.
- **Random Forest:** Random Forest classifier is an ensemble learning method for classification. It operates by constructing a multitude of decision trees thus reducing the variance. Hence this boosts the performance of the model even in the presence of noise.

FOUR CELL EXPERIMENT DESIGN

C1- 0% noise, 10% training set
C2- 0% noise, 80% training set
C3- 10% noise, 10% training set
C4- 10% noise, 80% training set

We are computing 10 runs of each classifier to test the true accuracy of the data. Each run is done with a distinct seed value and an average is calculated to get the true accuracy for all four experimental designs.

Total number of experiment runs = Criteria considered * Number of Classifiers * Total Runs for each classifier = 4 * 3 * 10

= 120 runs

PREDICTION ACCURACY BEFORE EVENT

The results of the classifiers before the event took place has been listed below. This accuracy does not represent the true number but it supports the results of the whole experimental design thus making it more reliable. This accuracy was derived in WEKA by removing all the attributes that were considered after the survey was done.

Classifier	0% Noise,80% Training Set
Naïve Bayes	**77.89%**
J48	**83.84%**
Random Forest	**88.78%**

EXPERIMENT RESULTS

RESULTS FOR EACH CLASSIFIER

The experiment is run for each classifier ten times for all four factors and the results are given below.

- Naïve Bayes Classifier

Seed	C1 Accuracy	C2 Accuracy	C3 Accuracy	C4 Accuracy
1	79.4484	80.102	71.8927	73.4694
2	79.1084	79.0816	70.7216	72.449
3	78.4662	79.5918	70.8349	73.2993
4	80.0151	79.5918	72.5349	74.3197
5	78.7306	80.6122	70.9482	75.8503
6	77.4462	82.6531	72.1949	75.3401
7	78.0506	80.4422	72.2327	72.9592
8	79.184	80.2721	70.4193	72.7891
9	79.6373	79.4218	72.6105	71.9388

10	77.5595	80.9524	71.0616	74.6599
Mean	**78.76463**	**80.2721**	**71.54513**	**73.70748**
StdDev	**0.874164484**	**1.020425926**	**0.828027411**	**1.285719921**

- J48 Classifier

Seed	C1 Accuracy	C2 Accuracy	C3 Accuracy	C4 Accuracy
1	74.5372	82.653	64.9037	74.1497
2	75.4439	82.823	70.1171	71.9388
3	75.6706	84.184	70.646	78.2313
4	76.1617	82.993	67.5859	72.619
5	74.0839	84.354	68.3415	71.9388
6	77.0306	84.864	69.4371	70.2381
7	76.9173	83.844	69.7771	70.5782
8	74.7261	80.952	70.4949	74.3197
9	74.3105	83.844	70.4949	71.2585
10	75.3306	85.034	69.7015	75.1701
Mean	**75.42124**	**83.55**	**69.15**	**73.04**
StdDev	**1.0399997**	**1.226**	**1.786**	**2.45**

- Random Forest Classifier

Seed	C1 Accuracy	C2 Accuracy	C3 Accuracy	C4 Accuracy
1	82.96	89.96	71.93	82.31
2	82.31	88.43	72.95	79.25
3	83.11	90.64	73.47	81.46

4	83.94	90.13	76.08	82.31
5	82.77	88.43	73.47	81.12
6	84.66	89.11	75.63	80.44
7	82.54	90.81	75.06	80.1
8	85.15	90.13	73.78	79.59
9	84.96	89.45	75.59	79.76
10	84.51	89.79	76.53	80.95
Mean	**83.691**	**89.688**	**74.449**	**80.729**
StdDev	**1.073048099**	**0.829736237**	**1.528161204**	**1.087202833**

We consider the mean or accuracy of C2 as the default design of each classifier for selection of our algorithm. C2 design consists of 0% Noise besides 80% Training Set and 20% Test Set. Each design is run 10 times using distinct seed values from 1-10 and the average is calculated. We assume the classifier with the highest mean to have the highest accuracy. From the above results, we interpret Random Forest classifier to have the highest mean and lowest standard deviation. Hence, we choose **Random Forest as our preferred algorithm.**

SUMMARY OF RESULTS

Below table lists the averages of the accuracy derived from the three classifiers in WEKA with different combinations used for testing purpose during the experimental analysis.

Algorithm	C1	C2	C3	C4
Random Forest	**83.69**	**89.69**	**74.45**	**80.73**
J48	**75.42**	**83.55**	**69.15**	**73.04**
Naïve Bayes	**78.76**	**80.27**	**71.55**	**73.71**

From the above accuracy results, we can conclude in terms of percentage split that Random Forest outperforms other algorithms in C2 where 80% of the dataset is used for training and 20% testing with highest accuracy in comparison to C1 where 10% of dataset used for training.

Confusion Matrix Table:

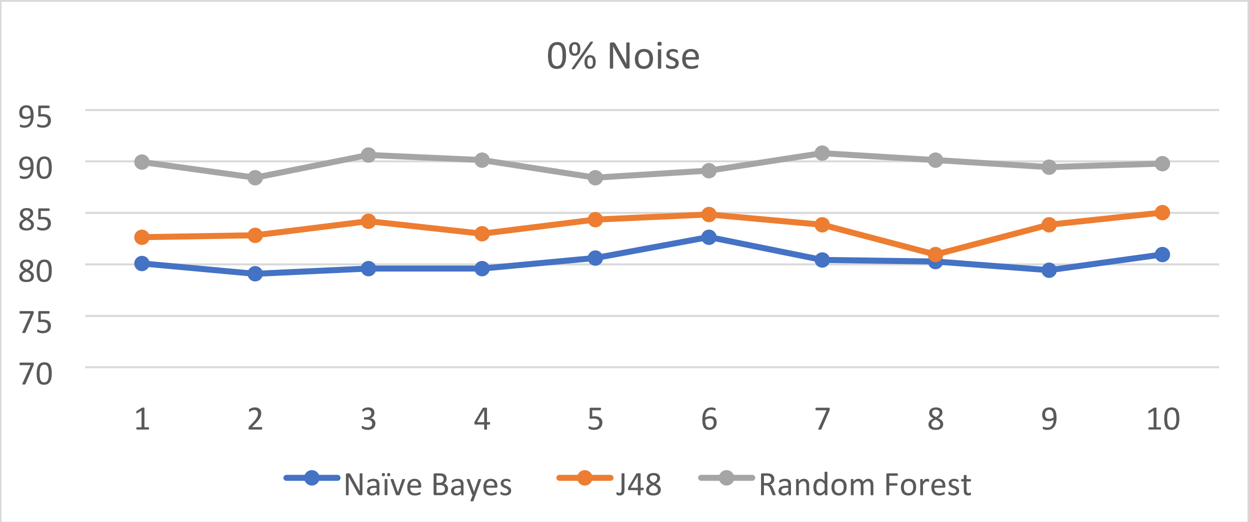
Random Forest		
a	b	
304	15	a=0
43	226	b=1

Naïve Baye		
a	b	
268	51	a=0
66	203	b=1

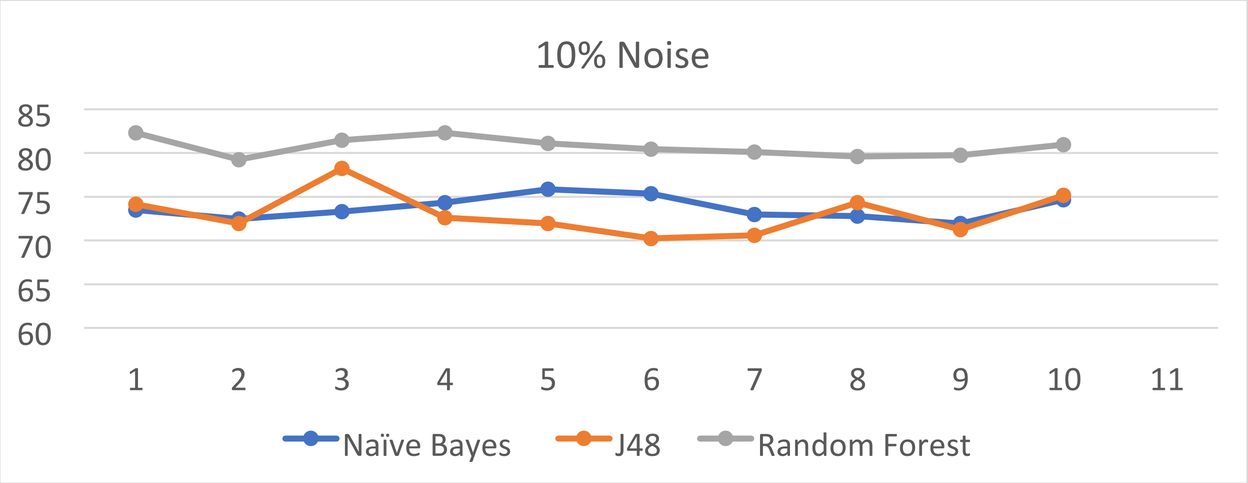
J48		
a	b	
264	55	a=0
47	222	b=1

In the above confusion matrix of the three classifiers, Random Forest classifier has lesser misclassified instances in comparison to Naïve Bayes and J48 classifiers.

Graph Representation (80% Training Set, 0% Noise)



Graph Representation (80% Training Set, 10% Noise)



In presence of 10% noise, the accuracy is impacted and we see a decrease along with slight variance especially in the performance of J48 as seen in the graphs. However, by further analysis of the charts, the algorithms give better accuracy results in combination C4 than C3.

Although Naïve Bayes accuracy results and graphs may show stability, but in terms of higher accuracy and performance, we have chosen Random Forest to be the best classifier among the three classifiers.

Hence based on the accuracy results of the factor experiment design, number of misclassified instances observed in confusion matrix and the graphical representation above, we can conclude Random Forest outperforms other algorithms with and without presence of noise.

ANALYSIS & CONCLUSION

INFERENCES BASED ON THE EXPLORATION IDEAS

Upon analyzing the dataset and the results achieved through data mining, we further explored the dataset from various perspectives to be able to summarize into useful information. This information can be used by organizations for increasing their sales revenue, decrease costs and for decision making.

Some of the questions explored through data mining and their inferences are as follows:

- **What are the least desirable attributes each gender considers in opposite sex?**

This can be determined by analyzing J48 tree visualization in WEKA. First, we filtered the gender attribute on "Male" in our dataset and ran it in WEKA with J48 classifier and upon visualizing the tree, we get two attributes at the bottom- "TV Sports" and "Exercise". This indicates the two least desirable attributes a female considers in a male.

Similarly, we filtered the gender attribute on "Female" in our dataset and ran in WEKA to find out the least desirable attributes in female. "Shared interests" and "Sincerity" are the two least desirable attributes male considers in a female.

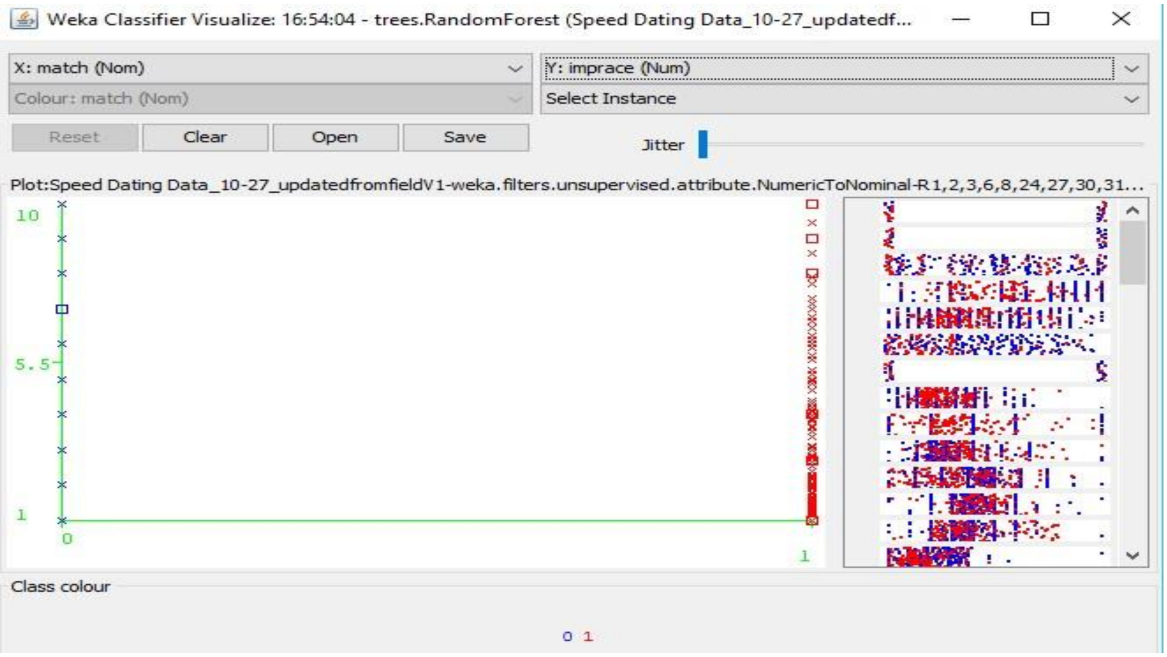
- **How important do people think attractiveness is in potential mate selection vs. it's real impact?**

This is determined by analyzing J48 tree visualization in WEKA. Upon visualization, we observe "shar_o > 5" and "attr_o > 6" attributes make it to the top of the tree. This indicates participants with common or shared interests consider attractiveness as an important quality during potential mate selection. Even in real world scenario, attractiveness is found to play a major role while selecting a partner.

- **How important is it that a person you date be of the same racial/ethnic background to get a match?**

We determined this using Visualize Classifier Error feature of Random Forest algorithm in WEKA. As seen in the below screenshot of the output obtained, we have chosen Y axis as "imprace" which is an attribute where the participant selects on a scale of 1 to 10 on how important it is for them to date someone of the same race/ ethnic background during the experiment. In this survey, 1 implies least importance and 10 implies extreme importance.

We have chosen X axis as **Match** which is a Nominal attribute deciding whether it is 0 (**No Match**) or 1 (**Match**). On further analysis, the graph suggests there are many values falling in the range of 0 to 5.5 marked by dark red points and starts to fade from 5.5 towards 10. Hence, we can deduce that most of the participants do not find race/ethnic background of utmost importance to get a match.



Weka Classifier Visualize: 16:54:04 - trees.RandomForest (Speed Dating Data_10-27_updatedf...

X: match (Nom)

Y: imprace (Num)

Colour: match (Nom)

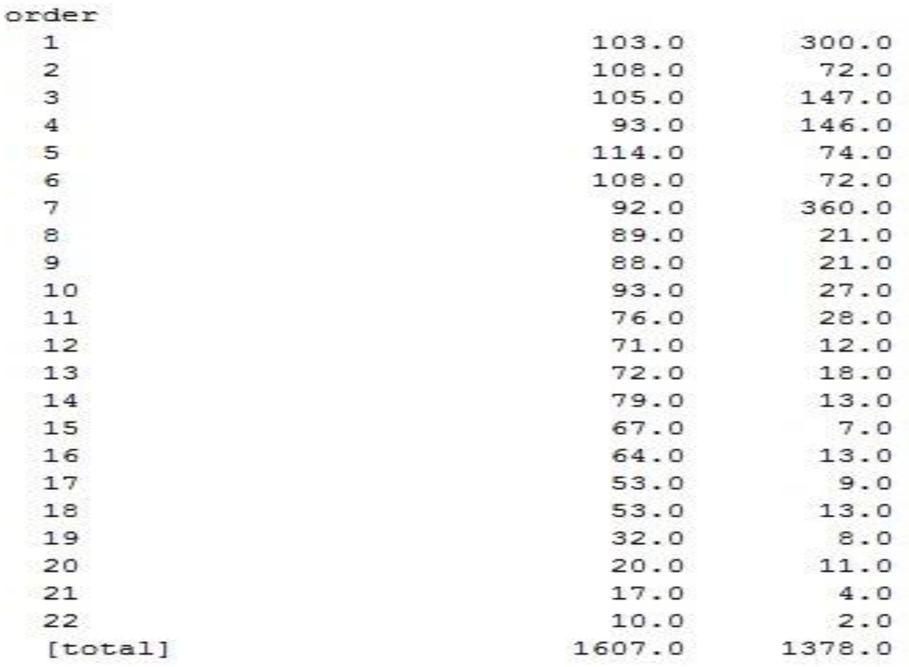
Select Instance

Plot:Speed Dating Data_10-27_updatedfromfieldV1-weka.filters.unsupervised.attribute.NumericToNominal-R1,2,3,6,8,24,27,30,31...

Graphical analysis of Random Forest using Visualize Classifier Error

- **In terms of getting a second date, is it better to be someone's first date of the night or last?**

We analyzed this aspect through Naïve Bayes classifier in WEKA. In this speed dating experiment, a participant meets many people during the date night. As seen in the below screenshot, first column Order denotes the first, second person the participant meets during the night and 22 being the last person he has met. Second column denotes Class 0 with the probabilities of the corresponding people dated being a No Match to the participant based on their interaction and survey results. Similarly, third column denotes Class 1 with probabilities of corresponding people dated being a match to the participant. By observing the Class 1 probabilities, we see that it reduces from the first person dated to last (300 to 2). On analysis of the results, we can deduce that the probabilities of being someone's first date will result into a match in comparison to being someone's last date.



```
order
  1                                  103.0     300.0
  2                                  108.0      72.0
  3                                  105.0     147.0
  4                                   93.0     146.0
  5                                  114.0      74.0
  6                                  108.0      72.0
  7                                   92.0     360.0
  8                                   89.0      21.0
  9                                   88.0      21.0
  10                                  93.0      27.0
  11                                  76.0      28.0
  12                                  71.0      12.0
  13                                  72.0      18.0
  14                                  79.0      13.0
  15                                  67.0       7.0
  16                                  64.0      13.0
  17                                  53.0       9.0
  18                                  53.0      13.0
  19                                  32.0       8.0
  20                                  20.0      11.0
  21                                  17.0       4.0
  22                                  10.0       2.0
  [total]                           1607.0    1378.0
```

CONCLUSION

As per the experimental design and graphical analysis, **Random Forest** classifier emerged as the best among the three classifiers because it displayed higher accuracy and lowest variance in all combinations and had lesser misclassification error in the confusion matrix in comparison to other algorithms.

Further we also explored the dataset and analyzed from various above perspectives. This summarized information can prove beneficial to many organizations. Below listed are some of the real-world applications that could use this information.

- **Sperm Banks** - The important attributes have been determined through this experiment in WEKA. The Sperm Banks can filter the data based on the important attributes and utilize the information to the benefit of their customers. The customers can choose based on their preferences.
- **Online Dating Sites** - This dataset was obtained as part of an experiment conducted in a school. However, it can be utilized by online dating sites for match making online.
- **Product-based companies** - The product based companies can make effective use of this information to send out ads to their customers targeting their needs. For example, people with interests in gaming, athletics, beauty, dining out can be sent out relevant ads and deals targeting the needs and thereby increasing organization profits.

REFERENCES

https://en.wikipedia.org/wiki/Speed_dating

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

http://www.intechopen.com/books/new-advances-in-machine-learning/data-mining-withskewed-data

Data mining with skewed data- By Manoel Fernando Alonso Gadi, Alair Pereira do Lago and Jorn Mehnen