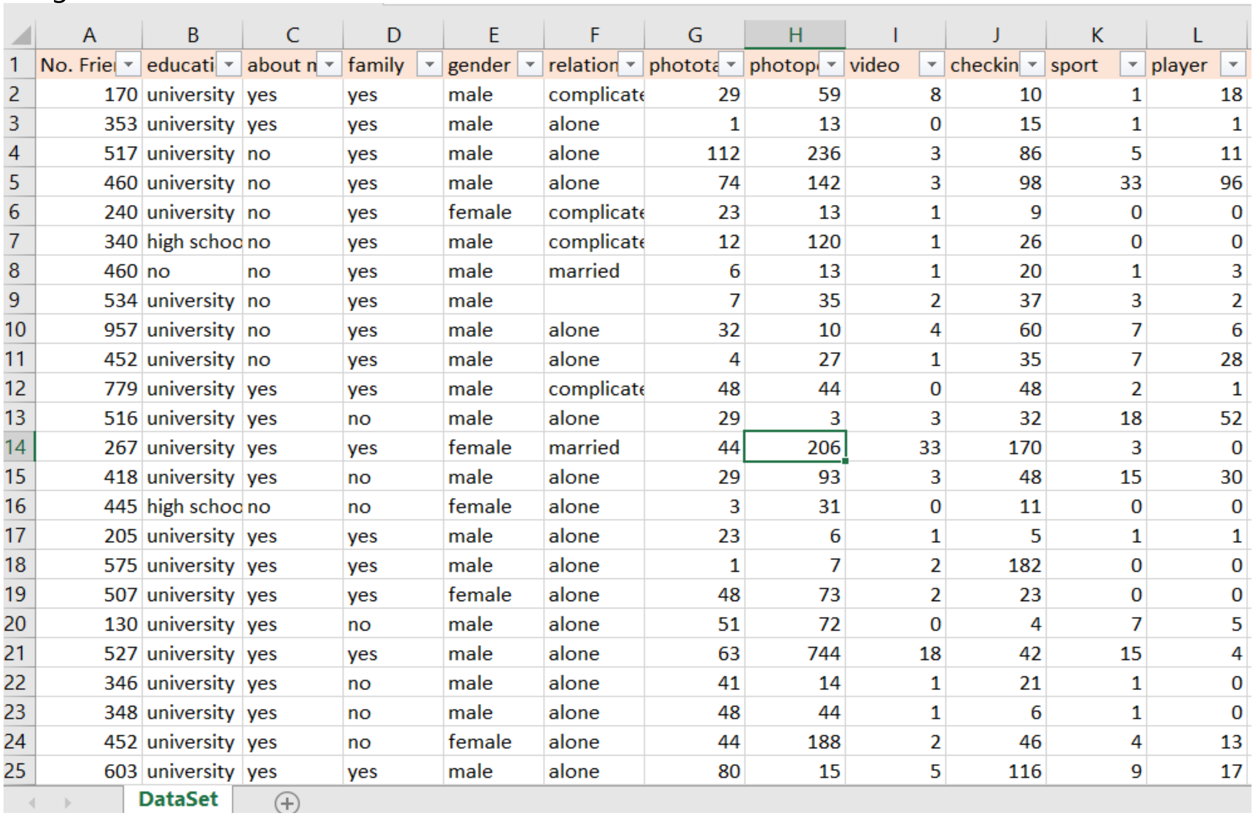
Project Proposal II, MIS -637-A
Suguna Bhargavi Bontha
CWID: 100440658

Project Statement:
Classify and study fake/spammer accounts and genuine accounts in social media.

Overview for choosing this topic:
Today, security challenges are a major concern for users in many industries like telecom, banking and social networks etc. In this project, the idea is to classify and study the fake accounts and genuine accounts based on similarity between the user's accounts details for Facebook (a large database for user accounts). Malicious users seek to violate the privacy of other users and abuse their names and credentials by creating fake accounts, which has become a concern for users. Hence, trying to detect and study the pattern of malicious users and fake accounts is important in order to eliminate them. Thus, the study can also help in prediction of spammers and thus prevent crime/fraud.

Dataset:
Fake and real accounts Facebook.

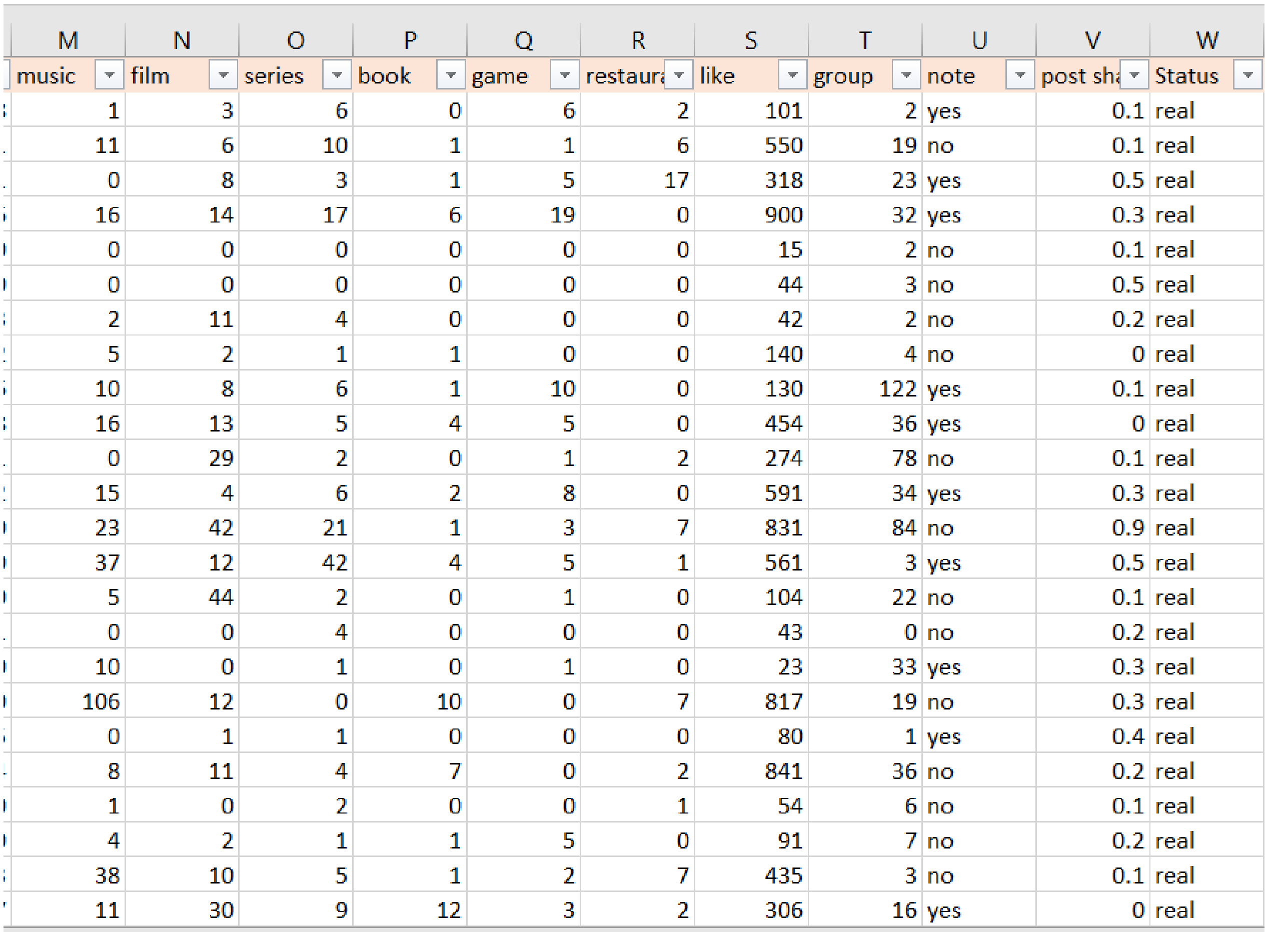
Sample data set is shown below. Image 1 shows the first 12 columns and image 2 has the last 11 columns.
Image 1



	A	B	C	D	E	F	G	H	I	J	K	L
1	No. Frie	educati	about n	family	gender	relation	photota	photop	video	checkin	sport	player
2	170	university	yes	yes	male	complicate	29	59	8	10	1	18
3	353	university	yes	yes	male	alone	1	13	0	15	1	1
4	517	university	no	yes	male	alone	112	236	3	86	5	11
5	460	university	no	yes	male	alone	74	142	3	98	33	96
6	240	university	no	yes	female	complicate	23	13	1	9	0	0
7	340	high schoo	no	yes	male	complicate	12	120	1	26	0	0
8	460	no	no	yes	male	married	6	13	1	20	1	3
9	534	university	no	yes	male		7	35	2	37	3	2
10	957	university	no	yes	male	alone	32	10	4	60	7	6
11	452	university	no	yes	male	alone	4	27	1	35	7	28
12	779	university	yes	yes	male	complicate	48	44	0	48	2	1
13	516	university	yes	no	male	alone	29	3	3	32	18	52
14	267	university	yes	yes	female	married	44	206	33	170	3	0
15	418	university	yes	no	male	alone	29	93	3	48	15	30
16	445	high schoo	no	no	female	alone	3	31	0	11	0	0
17	205	university	yes	yes	male	alone	23	6	1	5	1	1
18	575	university	yes	yes	male	alone	1	7	2	182	0	0
19	507	university	yes	yes	female	alone	48	73	2	23	0	0
20	130	university	yes	no	male	alone	51	72	0	4	7	5
21	527	university	yes	yes	male	alone	63	744	18	42	15	4
22	346	university	yes	no	male	alone	41	14	1	21	1	0
23	348	university	yes	no	male	alone	48	44	1	6	1	0
24	452	university	yes	no	female	alone	44	188	2	46	4	13
25	603	university	yes	yes	male	alone	80	15	5	116	9	17

Project Proposal II, MIS -637-A
Suguna Bhargavi Bontha
CWID: 100440658

(continued)
Image 2



M	N	O	P	Q	R	S	T	U	V	W
music	film	series	book	game	restaur	like	group	note	post sh	Status
1	3	6	0	6	2	101	2	yes	0.1	real
11	6	10	1	1	6	550	19	no	0.1	real
0	8	3	1	5	17	318	23	yes	0.5	real
16	14	17	6	19	0	900	32	yes	0.3	real
0	0	0	0	0	0	15	2	no	0.1	real
0	0	0	0	0	0	44	3	no	0.5	real
2	11	4	0	0	0	42	2	no	0.2	real
5	2	1	1	0	0	140	4	no	0	real
10	8	6	1	10	0	130	122	yes	0.1	real
16	13	5	4	5	0	454	36	yes	0	real
0	29	2	0	1	2	274	78	no	0.1	real
15	4	6	2	8	0	591	34	yes	0.3	real
23	42	21	1	3	7	831	84	no	0.9	real
37	12	42	4	5	1	561	3	yes	0.5	real
5	44	2	0	1	0	104	22	no	0.1	real
0	0	4	0	0	0	43	0	no	0.2	real
10	0	1	0	1	0	23	33	yes	0.3	real
106	12	0	10	0	7	817	19	no	0.3	real
0	1	1	0	0	0	80	1	yes	0.4	real
8	11	4	7	0	2	841	36	no	0.2	real
1	0	2	0	0	1	54	6	no	0.1	real
4	2	1	1	5	0	91	7	no	0.2	real
38	10	5	1	2	7	435	3	no	0.1	real
11	30	9	12	3	2	306	16	yes	0	real

The entire dataset contains of 23 attributes: input variables are the first 22 attributes shown above, they tell us about the Facebook account features, and Status is the output/target variable. The target variable “status” is a Boolean and has two values – real and fake. It will tell us if the account is real or not.

This dataset has no unnecessary attributes and no missing values. The number of records in this dataset are 889.

The data source is Kaggle.

The link to the data set is https://www.kaggle.com/khahu132/fake-and-real-accouts-fakebook.

Project Proposal II, MIS -637-A
Suguna Bhargavi Bontha
CWID: 100440658

Goals:

1) To classify fake/spammer accounts and genuine accounts.
2) Identify factors that majorly contribute to deciding the nature of the account and thus, can predict fake accounts and genuine accounts.

Process:

- The first step would be to choose the training, validation and testing sets from the datasets.
- The next step is to learn the decision trees using the training sets.
- Using that decision tree, we predict the if the account is fake or not for the testing set and report the accuracy, precision and recall.

Below are the methods used to proceed with the project.

Algorithm to be Used:

Standard CART (Classification and Regression Trees) algorithm is chosen since the target variable is a binary variable (which means it takes only two values "yes" or "no"). The algorithm is used to select the best split for the decision tree.

Software Package:

Salford Predictive Modeler is used to implement CART. It is a platform for data mining and predictive analytics like Random Forests, CART etc.