RESEARCH

Introduction to the profile areas of data sciences: project 9

Rohan Chitte] Natalja Amiridze] Michael Onyebuchi Ohaya

Abstract

Goal of the project: The goal of this project was to perform data handling as well as data analysis on study of social behaviour and personality traits

Main results of the project: Furthermore, Evaluation of student's current state of social behaviour based on sensing data using smarthphone based mobile sensing methods

Personal key learnings: We learnt how to handle large and complex datasets. Moreover, we performed data imputation and learnt and estimated distinct statistical estimates, such as intraclass correlation coefficient, ICC, correlation between features using Pearson's correlation coefficient

Estimated working hours: 14

Project evaluation: 2

Number of words: 1495

1 Scientific Background

When we observe the people around us, one of the first things we notice is how different people are from each other. Some people are very talkative, while others are very quiet. Some are active, while others are couch potatoes. Some worry a lot, while others almost never seem anxious. Every time we use one of these words like "talkative", "quiet", "active" or "anxious" to describe the people around us, we are talking about a person's personality - the characteristic differences between people. One challenge in documenting trends in social behaviour is the large number of channels through which socialisation can occur, both in person and through digital media. To study individual differences in everyday social behaviour, smartphone-based methods are increasingly being used.

2 Goal

The aim of the project was to analyse parts of the Harari study[1], especially the correlation of the BIG FIVE TRAITs (Neuroticism: a tendency to easily experience unpleasant emotions such as anxiety, anger or depression; Extroversion: energy, vivacity and a tendency to seek stimulation and the company of others; Agreeableness: a tendency to be compassionate and cooperative towards others rather than suspicious and antagonistic; Conscientiousness: a tendency to show self-discipline, act dutifully and strive for achievement; openness to experience social behaviour patterns of the study participants) with social behaviour of young adults. For this, the data had to be imported into a table (DataFrame). Then it was necessary to create and compare summary statistics for the data, pre-process the data with imputation of missing values, perform time series and correlation analyses. Finally, it was necessary to find out whether the student's feature could be predicted by social behaviour features with a trained classifier of one's own choice.

3 Data

The data came from the Harary study to investigate individual differences in everyday social behaviours. To do this, four studies S1,S2,S3,S4 (N=927 in total) used smartphone-based mobile sensing methods (MSM) to collect real-world data on young adults' social behaviour across four communication channels: conversations, phone calls, text messages and use of messaging and social media apps. In each study, participants were first informed about the purpose of the data collection and their approval regarding mobile sensing apps usage and its use of recorded data was taken into consideration prior to their participation. In S1,Participants were given Android phones to use throughout the study duration of 10-week and the no. of students were 48. S2 had 26 students and the data collection wave was divided into 2 phases, each consisting of 2 weeks. 137 students were monitors for 8 weeks in sample no. 3 and 716 students for 2 weeks in Sample no. 4. Along with the social behaviour data captured through mobile sensor with the help of mobile app, students in sample 3 and 4 underwent a self-reported personality disposition. This was captured using the Big Five Inventory that consists of 44 questions that ultimately determines individual's personality. Big Five Inventory consists of 5 trait ratings (i.e., 1) extraversion , 2) agreeableness, 3) conscientiousness, 4) neuroticism, and 5) openness).

4 Results

4.1 Task 1: Are there individual differences in the daily social behaviour?

In order to investigate the the differences between individuals, we calculated intraclass correlation coefficients for different daily social activities (see Figure 1). The table shows for different samples between-person variance (BPV: represents ICC1 estimate), which is the percentage of variation in the observed daily social behaviors that can be explained by individual factors and mean consistency (S4IMR: represents ICC(3,k) estimate), which is the average individual stability of the daily social behaviour assessments across days. These estimates were calculated using package rpy2 in python, with which it is possible to use packages from R and to calculate ICC estimates in similar way. The calculated estimates were identical with that in the paper by Harari et.al [1]. Investigating the sample 4 the between-person variance is more high for conversation frequency and duration (0.55 and 0.52), incoming texting frequency (0.52), outcoming texting frequency (0.57). Also high BPVs are seen for app message frequency, app social frequency and app social duration in sample 3. All other parameters of the samples have small BPVs. All mean consistency values are high for all four samples. The high BPV could be due to the different daily social behaviour of young adults. The timeseries of two participants, one with low variance (Figure 2) and one with high variance (Figure 3).

	convofreq	convodur	callinfreq	callindur	calloutfreq	calloutdur	textinfreq	textinlen	textoutfreq	textoutlen	appmessgfreq	appmessgdur	appsocialfreq	appsocialdur
S1BPV	[0.3, 0.24, 0.39]	[0.35, 0.28, 0.44]	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
S1IMR	[0.97, 0.95, 0.98]	[0.97, 0.96, 0.98]	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
S2BPV	NaN	NaN	[0.11, 0.06, 0.21]	[0.11, 0.05, 0.2]	[0.2, 0.12, 0.32]	[0.15, 0.08, 0.26]	[0.39, 0.28, 0.54]	[0.3, 0.2, 0.44]	[0.35, 0.25, 0.5]	[0.33, 0.23, 0.48]	NaN	NaN	NaN	NaN
S2IMR	NaN	NaN	[0.68, 0.5, 0.82]	[0.67, 0.49, 0.81]	[0.81, 0.7, 0.89]	[0.75, 0.62, 0.86]	[0.92, 0.87, 0.95]	[0.88, 0.81, 0.93]	[0.9, 0.85, 0.95]	[0.9, 0.84, 0.94]	NaN	NaN	NaN	NaN
S3BPV	NaN	NaN	[0.16, 0.13, 0.21]	[0.2, 0.16, 0.25]	[0.33, 0.28, 0.38]	[0.37, 0.33, 0.43]	[0.17, 0.14, 0.21]	[0.11, 0.09, 0.14]	[0.19, 0.16, 0.23]	[0.14, 0.11, 0.18]	[0.7, 0.66, 0.75]	[0.51, 0.46, 0.56]	[0.67, 0.61, 0.72]	[0.62, 0.57, 0.67]
S3IMR	NaN	NaN	[0.85, 0.82, 0.89]	[0.88, 0.85, 0.91]	[0.94, 0.92, 0.95]	[0.95, 0.94, 0.96]	[0.87, 0.84, 0.89]	[0.8, 0.75, 0.84]	[0.88, 0.85, 0.9]	[0.83, 0.8, 0.87]	[0.99, 0.98, 0.99]	[0.97, 0.96, 0.97]	[0.98, 0.98, 0.99]	[0.98, 0.98, 0.98]
S4BPV	[0.55, 0.52, 0.57]	[0.52, 0.5, 0.55]	[0.3, 0.26, 0.34]	[0.42, 0.38, 0.47]	[0.32, 0.28, 0.37]	[0.26, 0.22, 0.31]	[0.52, 0.47, 0.57]	[0.43, 0.38, 0.48]	[0.57, 0.52, 0.62]	[0.45, 0.41, 0.51]	NaN	NaN	NaN	NaN
S4IMR	[0.97, 0.97, 0.97]	[0.97, 0.96, 0.97]	[0.92, 0.9, 0.93]	[0.95, 0.94, 0.96]	[0.93, 0.91, 0.94]	[0.9, 0.88, 0.92]	[0.97, 0.96, 0.97]	[0.96, 0.95, 0.96]	[0.97, 0.97, 0.98]	[0.96, 0.95, 0.97]	NaN	NaN	NaN	NaN

Figure 1 Variability and Reliability of Daily Social Behaviors

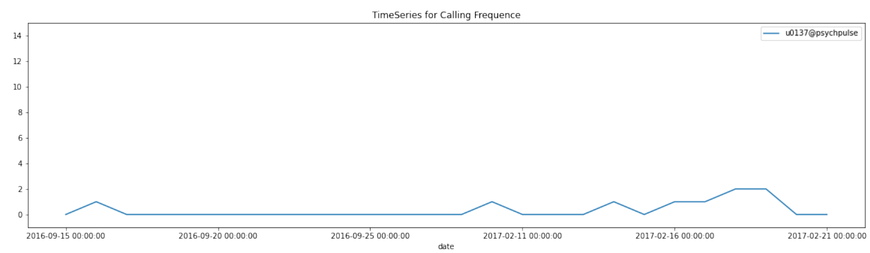

Figure 2 Timeseries of one participant with low variance from the sample 4

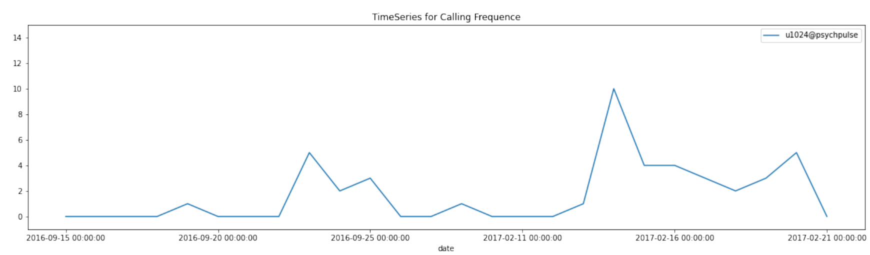

Figure 3 Timeseries of one participant with high variance from the sample 4

4.2 Task 2: Which behavioral dispositions are related to personality traits?

To evaluate the extent to which behavioral sociability tendencies map on to standard self-reported measures of personality traits[Particularly Extraversion], we selected Sample no. 3 as our dataset which had 10 daily social behavior. The Spearman's Rank Correlation Coefficient is used to discover the strength of a link between two sets of data. With calculation of Spearman correlation it was possible to estimate how social behavior activities relate with personality traits (see Figure 5). The table [Figure 4] displays the correlational estimates, associated 95% confidence intervals and exact p values for the correlational analyses conducted in S3 where N=137. Most BIG FIVE TRAITS showed low values of correlation with social behavior dispositions. From the table, it can be observed that extraverts tend to engage in a longer calling behaviour [duration] during the night time ($r = 0.10$ to $r = 0.33$). A strong correlation is found during night, evening and weekend for calling frequency (r = 0.10 to r = 0.28 , r = 0.10 to r = 0.23 r = 0.12 to r = 0.21). Texting behaviour at weekday, evening and morning evidently shows higher correlation (r = 0.12 to 0.22, r = 0.15 to 0.21 r = 0.16 r = 0.21). As for the application messaging duration, maximum correlation was evident across morning and weekend time with value r = 0.28. Figure 5 shows the correlation between all the features and all five big five traits.

Correlations Between Time of the Day/Week Social Behavior Tendencies and Extraversion

CALL OUT DURATION

	Variable	r	ci	p-value
0	Morning	0.067001	(-0.1018591675536811, 0.23210818774972677)	0.436619
1	Afternoon	0.115695	(-0.05304943295583833, 0.2780163047484722)	0.178206
2	Evening	0.134760	(-0.03371785017953301, 0.2957899577752689)	0.116406
3	Night	0.330748	(0.17260760758212476, 0.4722663857180523)	0.000079
4	Weekday	0.092950	(-0.0759499255672151, 0.25666388570837767)	0.280000
5	Weekend	0.143516	(-0.024796522911973838, 0.3039160935415045)	0.094309

CALL IN DURATION

	Variable	r	ci	p-value
0	Morning	0.074287	(-0.09460727811732919, 0.23902439469617903)	0.388285
1	Afternoon	0.061641	(-0.10718255105843766, 0.22700963396192522)	0.474263
2	Evening	0.047848	(-0.12083737173275416, 0.21384721659867822)	0.578737
3	Night	0.107541	(-0.06127967677260702, 0.2703799037277814)	0.210995
4	Weekday	0.041250	(-0.12734684590027479, 0.20752941016818247)	0.632227
5	Weekend	0.080939	(-0.08797071210987562, 0.24532423406519038)	0.347100

CALL OUT FREQ

	Variable	r	ci	p-value
0	Morning	0.135354	(-0.03311322983150563, 0.29634219667753)	0.114789
1	Afternoon	0.190428	(0.02346180299179199, 0.3470591828217841)	0.025819
2	Evening	0.238631	(0.07387240662476555, 0.3907097921373484)	0.004982
3	Night	0.284499	(0.12263476184928929, 0.43161957657572037)	0.000753
4	Weekday	0.182117	(0.014854865071530972, 0.3394636151615405)	0.033177
5	Weekend	0.215116	(0.049175026521131496, 0.3695008685309908)	0.011589

CALL IN FREQ

	Variable	r	ci	p-value
0	Morning	0.107604	(-0.06121577106062771, 0.27043935960246895)	0.210725
1	Afternoon	0.114453	(-0.054304655725312234, 0.2768543491064921)	0.182943
2	Evening	0.100131	(-0.06873931285769022, 0.26342226559253556)	0.244348
3	Night	0.107776	(-0.06104239736316285, 0.27060064832475994)	0.209992
4	Weekday	0.125804	(-0.04281474387167832, 0.2874545047476547)	0.142966
5	Weekend	0.122840	(-0.045819265953581295, 0.2846904369958763)	0.152701

SMS IN FREQ

	Variable	r	ci	p-value
0	Morning	0.167129	(-0.000603155016352276, 0.32571488607706656)	0.050935
1	Afternoon	0.158854	(-0.009104145791840648, 0.318094625424333)	0.063726
2	Evening	0.218851	(0.05308409528880508, 0.37288016894311765)	0.010189
3	Night	0.088319	(-0.08059065178285614, 0.2522969426818806)	0.304761
4	Weekday	0.186112	(0.01898943662819745, 0.34311763929797506)	0.029444
5	Weekend	0.200332	(0.0337500617853631, 0.3560831442972565)	0.018916

SMS OUT FREQ

	Variable	r	ci	p-value
0	Morning	0.213417	(0.047398069290110555, 0.3679619234285066)	0.012279
1	Afternoon	0.144181	(-0.024117667112394978, 0.30453248828422785)	0.092777
2	Evening	0.152647	(-0.0154641804470533, 0.312365729078761)	0.074950
3	Night	0.070785	(-0.09809546195053798, 0.2357018953139512)	0.411104
4	Weekday	0.223231	(0.05767497669410366, 0.37683807822885407)	0.008739
5	Weekend	0.121026	(-0.047656774400154485, 0.28299727319945106)	0.158901

```
SMS OUT LEN
     Variable         r                                            ci   p-value
0     Morning  0.200829     (0.0342677737103988, 0.3565356527569358)  0.018616
1   Afternoon  0.135284   (-0.03318443525376472, 0.2962771715981983)  0.114979
2     Evening  0.116723  (-0.05201111774893315, 0.27897673813261553)  0.174358
3       Night  0.052850   (-0.1158923963552185, 0.21862794966840404)  0.539639
4     Weekday  0.175188  (0.007698929655130184, 0.33311615280593904)  0.040599
5     Weekend  0.112975  (-0.05579748697027167 6, 0.275471178128538 6)  0.188699
```

```
APP SOCIAL FREQ
     Variable         r                                             ci   p-value
0     Morning  0.091608   (-0.07729457808812298, 0.25539995786590736)  0.287027
1   Afternoon  0.121864  (-0.046808298434835624, 0.28377935370476076)  0.156015
2     Evening  0.121356     (-0.04732245893885078, 0.2833054801054091)  0.157759
3       Night  0.102350     (-0.06650716766672735, 0.2655078275193804)  0.233996
4     Weekday  0.103643     (-0.06520642618113974, 0.2667217190569903)  0.228111
5     Weekend  0.140987     (-0.02737607962245623, 0.3015713676998645)  0.100320
```

```
APP MESSAGING DURATION LONG
     Variable         r                                            ci   p-value
0     Morning  0.220585     (0.05490016650207302, 0.3744472400754819)  0.009592
1   Afternoon  0.213817     (0.04781657767925533, 0.3683245334494291)  0.012113
2     Evening  0.184391    (0.017207447256730762, 0.3415439666196005)  0.031006
3       Night  0.144474  (-0.02381825409691095 5, 0.30480426506452535)  0.092108
4     Weekday  0.203222     (0.03675978596485389, 0.3587116939292214)  0.017230
5     Weekend  0.151259   (-0.01688465549286011, 0.3110829489969786)  0.077667
```

```
APP MESSAGING DURATION
     Variable         r                                           ci   p-value
0     Morning  0.280084    (0.11790693958731717, 0.4277079844305011)  0.000917
1   Afternoon  0.173848  (0.006316383463346876 6, 0.3318863965835155)  0.042185
2     Evening  0.235597    (0.07067457728292904, 0.3879824780823676)  0.005581
3       Night  0.174782  (0.007280223974987388, 0.33274383657766504)  0.041074
4     Weekday  0.282035    (0.11999511897488341, 0.4294370839925019)  0.000841
5     Weekend  0.180391   (0.013070875378463895, 0.3378839411257648)  0.034909
```

Figure 4 Correlations between social Behaviours by time of the day/week and self-reported BIG FIVE Trait[Extraversion] in the sample 3

	extra.r	extra.ci	extra.p	agr.r	agr.ci	agr.p	con.r	con.ci	con.p	neur.r	neur.ci	neur.p	ope.r	ope.ci	ope.p
morning callinfreq	0.11	[-0.06, 0.27]	0.211	0.06	[-0.11, 0.23]	0.467	-0.07	[-0.24, 0.1]	0.392	-0.08	[-0.24, 0.09]	0.372	-0.00	[-0.17, 0.17]	0.995
afternoon callinfreq	0.11	[-0.05, 0.28]	0.183	0.03	[-0.14, 0.2]	0.719	-0.06	[-0.22, 0.11]	0.503	-0.08	[-0.25, 0.08]	0.327	0.04	[-0.13, 0.2]	0.684
evening callinfreq	0.10	[-0.07, 0.26]	0.244	-0.03	[-0.2, 0.14]	0.714	-0.09	[-0.25, 0.08]	0.306	-0.02	[-0.19, 0.15]	0.820	-0.03	[-0.19, 0.14]	0.753
night callinfreq	0.11	[-0.06, 0.27]	0.210	0.06	[-0.11, 0.23]	0.486	-0.19	[-0.34, -0.02]	0.029	-0.01	[-0.18, 0.16]	0.923	0.05	[-0.12, 0.22]	0.542
weekday callinfreq	0.13	[-0.04, 0.29]	0.143	-0.01	[-0.18, 0.16]	0.924	-0.10	[-0.27, 0.07]	0.234	-0.05	[-0.22, 0.12]	0.546	0.01	[-0.15, 0.18]	0.872
weekend callinfreq	0.12	[-0.05, 0.28]	0.153	0.04	[-0.13, 0.21]	0.633	-0.02	[-0.19, 0.15]	0.806	-0.04	[-0.2, 0.13]	0.668	-0.01	[-0.18, 0.16]	0.891
morning callindur	0.07	[-0.09, 0.24]	0.388	0.05	[-0.12, 0.21]	0.576	-0.05	[-0.21, 0.12]	0.574	-0.05	[-0.21, 0.12]	0.598	-0.02	[-0.19, 0.15]	0.827
afternoon callindur	0.06	[-0.11, 0.23]	0.474	0.05	[-0.12, 0.22]	0.548	-0.07	[-0.24, 0.1]	0.407	-0.05	[-0.22, 0.12]	0.538	0.03	[-0.13, 0.2]	0.694
evening callindur	0.05	[-0.12, 0.21]	0.579	-0.04	[-0.21, 0.13]	0.624	-0.10	[-0.26, 0.07]	0.251	0.04	[-0.12, 0.21]	0.604	-0.04	[-0.21, 0.12]	0.604
night callindur	0.11	[-0.06, 0.27]	0.211	0.07	[-0.1, 0.23]	0.446	-0.19	[-0.35, -0.03]	0.024	-0.01	[-0.17, 0.16]	0.939	0.04	[-0.13, 0.21]	0.619
weekday callindur	0.04	[-0.13, 0.21]	0.632	-0.00	[-0.17, 0.16]	0.970	-0.09	[-0.25, 0.08]	0.321	0.02	[-0.15, 0.19]	0.830	-0.00	[-0.17, 0.17]	0.976
weekend callindur	0.08	[-0.09, 0.25]	0.347	0.01	[-0.16, 0.18]	0.929	0.01	[-0.16, 0.17]	0.953	0.02	[-0.15, 0.19]	0.809	-0.02	[-0.19, 0.15]	0.802
morning calloutfreq	0.14	[-0.03, 0.3]	0.115	-0.04	[-0.21, 0.13]	0.638	-0.07	[-0.23, 0.1]	0.433	-0.05	[-0.21, 0.12]	0.601	-0.06	[-0.22, 0.11]	0.498
afternoon calloutfreq	0.19	[0.02, 0.35]	0.026	-0.01	[-0.18, 0.16]	0.913	-0.10	[-0.26, 0.07]	0.254	-0.11	[-0.27, 0.06]	0.199	0.02	[-0.15, 0.18]	0.837
evening calloutfreq	0.24	[0.07, 0.39]	0.005	-0.00	[-0.17, 0.17]	0.993	-0.14	[-0.3, 0.03]	0.113	-0.08	[-0.25, 0.09]	0.344	0.02	[-0.15, 0.18]	0.844
night calloutfreq	0.28	[0.12, 0.43]	0.001	0.11	[-0.06, 0.27]	0.216	-0.07	[-0.23, 0.1]	0.436	-0.11	[-0.27, 0.06]	0.222	0.14	[-0.03, 0.3]	0.107
weekday calloutfreq	0.18	[0.01, 0.34]	0.033	-0.04	[-0.21, 0.13]	0.653	-0.09	[-0.26, 0.07]	0.270	-0.09	[-0.25, 0.08]	0.321	-0.00	[-0.17, 0.17]	0.975
weekend calloutfreq	0.22	[0.05, 0.37]	0.012	0.05	[-0.12, 0.22]	0.565	-0.07	[-0.24, 0.09]	0.389	-0.08	[-0.25, 0.09]	0.347	-0.00	[-0.17, 0.16]	0.960

Figure 5 Correlations between social Behaviours by time of the day/week and self-reported BIG FIVE Traits in the sample 3

4.3 Task 3: Can we predict an individual's personality?

Logistic regression classification was performed using assessed daily social behaviour data (conversations, phone calls, message texting and social app use) as traits to predict extraversion as a personality trait. We extracted most important features and found that some attributes have the highest feature importance like app_social_num_avg_morning and app_social_num_avg_afternoon, call_out_num_avg_weekdays, call_in_num_avg_weekdays, app_social_num_avg_weekends and sms_out_num_avg_weekends for logistic regression (LR) classifier see Figure 6.

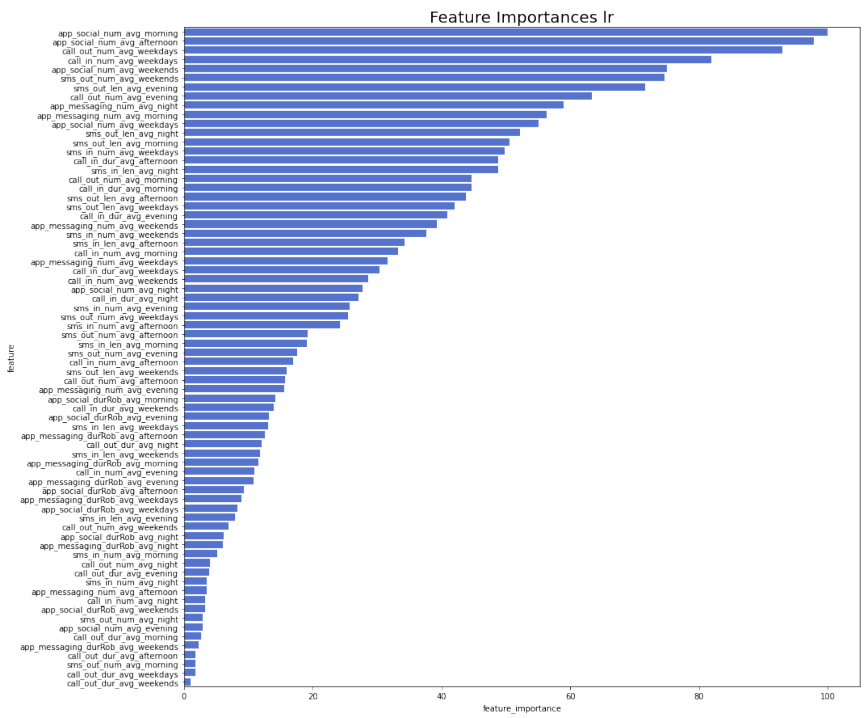

Figure 6 feature importance of LR

Here are the results from analysis with the test dataset (data were split to train and test set to 70:30 ratio) (see Figure 7). The confusion matrix shows that 26 data were correctly identified as low and high level Extraversion and misclassified 16 entries. ROC curve shows that the model is evaluated at 69% reliablility.

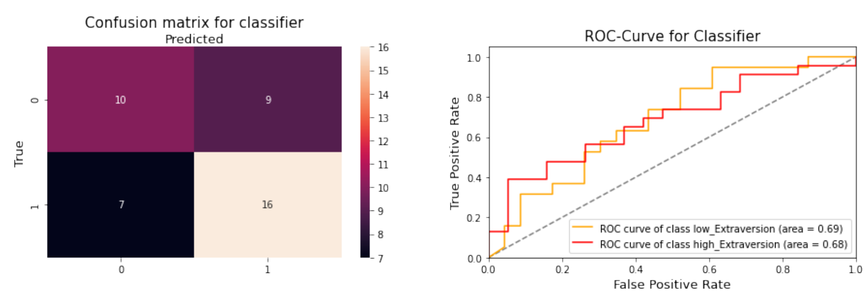

Figure 7 Classification evaluation with training set: Confusion matrix and ROC curve of LR with test set

The accuracy of LR for test set is 62% (see Figure 8). 0 means low level of Extraversion and 1 means high level of Extraversion. The low and high Extraversion groups were separated by a mean value. For the low level of Extraversion a low score was received for precision and slightly lower recall (59% and 53% respectively), so that slightly less correct labels are returned but most of these predicted labels are still correct compared to the training labels. This means that the classifier provides here not so accurate result. But for the high level of Extraversion the result shows higher values for predictions and recall (64% and 70% respectively). So given daily behavioural characteristics, it is possible to predict groups with high and low levels of Extraversion only with low accuracy.

```
Cross validation accuracy:
         0.6190476190476191

Cross validation classification report

              precision    recall  f1-score   support

           0       0.59      0.53      0.56        19
           1       0.64      0.70      0.67        23

    accuracy                           0.62        42
   macro avg       0.61      0.61      0.61        42
weighted avg       0.62      0.62      0.62        42
```

Figure 8 Crossvalidation of LR with test set

5 Discussion

Such a study can prove useful in several ways. The widespread use of smartphones and the internet makes it easier to conduct such a study. One could monitor both the specific and the average social behaviour of the students. Based on the results obtained, further analysis can be drawn on what social behaviours influence the personality of the students. Such analysis can be used to measure a person's most important personality characteristics, and help the person understand which roles would fit the best for him/her. A person who is outgoing might have excellent social, engaging, presentation etc skills which can be efficiently used in a large organization for roles involving large scale interaction(typically sales or manager). Where as a shy and introvert person will unlikely feel at home in a high-pressure positions like sales department or business development etc, where he or she has to make lots of telephone calls and interact, for instance. Recruiters can especially use it to find people who have the personality, as well as the skills, to fit the roles that they are hiring for as when the personality doesn't fit the role, it is not only the organizations lose but employees' as well.

6 Appendix

Name	Work Description
Michael	
Rohan	Data, Results, Discussions, 9.3 and 9.4 Code
Natalja	Scientific Background, Goal, Data, Results, Discussion, 9.1, 9.2, 9.3 and 9.4 Code

Table 1 Task responsibilities

References

1. Harari, G.M., Mueller, S.R., Stachl, C., Wang, R., Wang, W., Buehner, M., Rentfrow, P.: Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. Journal of personality and social psychology **1**, 204–228 (2020)