RESEARCH

Introduction of Profile Areas of Data Science :Project 9

Suresh Kumar Choudhary, Frenny Macwan and Aman Jain

Full list of author information is available at the end of the article

Abstract

Goal of the Project: To study the data given, about the students' personality traits, social behaviour and predict how co-related ((Or Different) are the individuals from others in the group along with predicting the individual's personality.

Main Result of the Project: We were able to fetch the information of the students who are highly co-related and highly not-co-related with the others members of the group, along with creating a classification and regression models to correctly classify individual traits. We found best classification with logistic regression model with auc 0.74 with target trait neuroticism and best regression model with mean squared error 0.42 and R^2 (coefficient of de-termination) 0.01 with trait conscientiousness.

Personal Key Leanings: Big five trait, 44 items, Prophet Library, Big Query, ICC (Inter-class Co-relation Coefficient)

Estimate working hours: 24

Project Evaluation: 2

Number of Words: 1628

Keywords: Time Series data trend; Inter-Class Corelation Coefficient; Classification; Regression; Personality Traits

Scientific Background

By usage of Mobile Sensing Methods in four studies, data has been collected from young adults across four different communication channels : Conversations, Phone Calls, Text Messages and Use of Messaging and Social Media, thorough which Behavioural Traits like Duration and Frequency of each channel were fetched along with, the Personality Traits through Surveys were fetched. Variance of individuals were calculated on their behaviour traits.

Goal

To analyze the social life of individuals using the mobile sensing (Majorly the Frequency and Duration of Social and Messaging Apps) and assessment techniques of personality traits .

Data

We used three Sample Data Sets [1] ; S1, S2 and S3 for the tasks.

1 **S1 -** Data of 48 students were taken for a total of 66 days (10 Weeks). Data was acquired by the self tracked psychological experience by the participants and

also behaviours by the smart phone data. The data consist of the participant's age, id, their average call duration and average frequency of the calls of the entire 66 day period, along with the daily count of the duration and frequency of the calls in separate table as well. We create the tables(schema) in the GoogleBigQuery and imported them using gbq commands (gbq.read_gbq). As part of prepossessing, since most of the data which went missing were numeric values, we took the mean of the factor and filled it where the values weren't available.

2 **S2** - Data of 118 1st year Students of a UK College were taken in two different phases of two weeks each, in a gap of three months. Phase 1, where the mobile details were tracked by the app Easy-M faced some technical issues and hence only Phase 2 Data was considered where the behavioural factors like Frequency and Duration of Incoming-Outgoing calls, Incoming-Outgoing SMS were tracked for 28 Students through MyLife Logger App. All tables having _long as the suffix tracked the daily frequency and duration of calls and SMS for each participant, where as the tables having _wide suffix (In this Sample only One), tracked the average behaviours of the participants.

3 **S3** - Data of 137 Students and Employees of Southern German University were taken. The tracking happened for 8 Weeks (60 Days). From Day 2 to Day 31 (Total of 30 Days) data was taken for this study, where Participants were rewarded with 30 Euros in exchange of their Behaviour Tracking on Phone and Personality feed-backs. Two factors (Duration and Frequency) were considered. For each user Behaviour factors (Duration and Frequency) were tracked for 31 days from apps like Messaging, Social, Calls In-Out, SMS In-Out. Along with daily tracking, an average duration and frequency was also given on daily average basis and on average time-of-day basis (Morning, Afternoon, Evening, and Night) for each participant along with their points on each personality traits after their inputs on the questionnaires. The five personality traits given in the data set are Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism.

Result

We have followed the research paper [2] for Task1 and Task 2.

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

- **Goal**: To find out the Interclass Correlation Coefficient (ICC) between the students of a particular group (i.e. Sample S2)
- **Method**: Interclass Correlation Coefficient describes how strongly units in the same group resemble each other. ICC values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability.
- **Result**: To find individual differences in daily social behaviors, we are using the ICC method. We have considered Sample 2 data-set for analysis. We tried to compute between person variance with lower and higher bound values by importing ICC package. For the Sample S2 we have different features like call

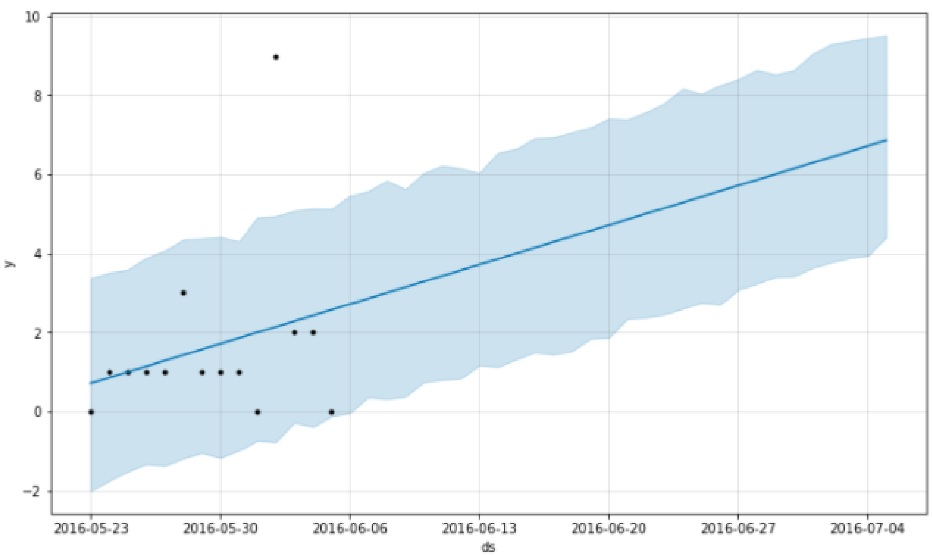
in freq, call out freq, call in duration, call out duration, SMS in freq, SMS out freq, SMS in duration, SMS out duration. For all the features, ICC calculation can be seen in Fig 1

Figure 1 ICC calculation on sample 2 dataset

	CALL IN FREQ	CALL OUT FREQ	CALL IN DUR	CALL OUT DUR	SMS IN FREQ	SMS OUT FREQ	SMS IN DUR	SMS OUT DUR
ICC	0.1115708	0.1944612	0.1060321	0.1464291	0.39286	0.3549673	0.3001231	0.3347373
Lower	0.04356998	0.1025994	0.03976695	0.06790283	0.2598036	0.227266	0.1821257	0.2103536
Higher	0.2397746	0.3518347	0.2317776	0.2885876	0.5768427	0.5391687	0.4811719	0.5182721

When we tried to find variances of an individual, for incoming call frequency (CALL IN FREQ), we found out it varies among individual students. Student ID 173512 shows the highest variance, where Student ID 553102 shows the least variance. That can be seen in Fig 2 and Fig 3

Figure 2 Student id 173512 having high variance



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

- **Goal**: To find out the behavioural traits which are highly linked with the personality traits.
- **Method**: Using the Inter-class Correlation Coefficient, we took Daily Call Out, Daily Call In and Daily SMS In behavioural traits to co relate them with personality traits (Extra-Version).
- **Result**: Out of the three Behaviours we Call Out matches the most with the Extra-version Personality and are positively co-related amongst the three. Daily Call out (Figure : 4), Daily Call In (Figure : 5) and Daily SMS(Figure : 6). It can be seen in the Figure 4: Call out Behaviour is the most co-related with the the Extra-version Personality.

Figure 3 Student id 553102 having low variance

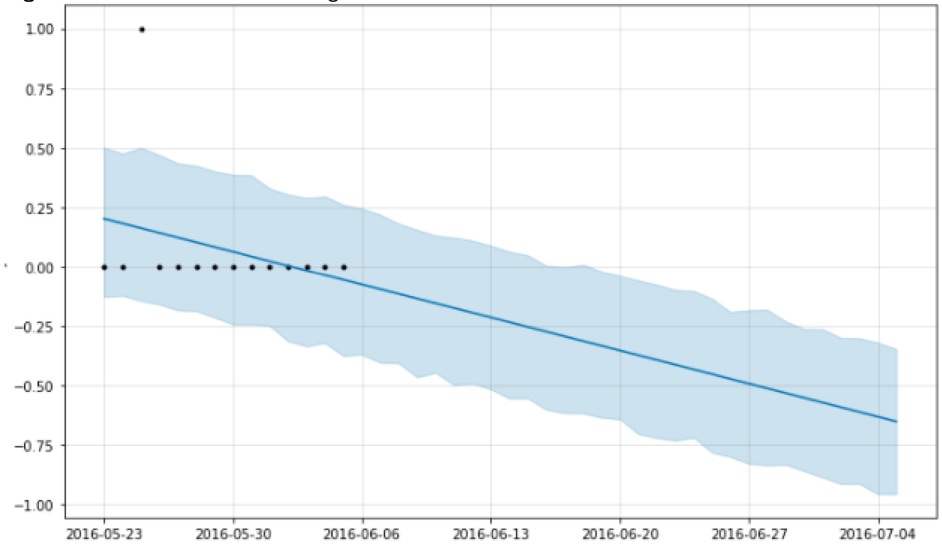


Figure 4 Corelation between Call Out and Personality

```
Daily CALL OUT DUR Extraversion ICC:[1] -0.05082625
Lower ICC :[1] -0.4449291
Upper ICC :[1] 0.2587485

Daily CALL OUT DUR openness ICC:[1] -0.367005
Lower ICC :[1] -0.7814345
Upper ICC :[1] -0.007952622

Daily CALL OUT DUR conscientiousness ICC:[1] 0.148483
Lower ICC :[1] -0.2130452
Upper ICC :[1] 0.4153091

Daily CALL OUT DUR agreeableness ICC:[1] 0.1350457
Lower ICC :[1] -0.2292025
Upper ICC :[1] 0.4050161

Daily CALL OUT DUR  neuroticism ICC:[1] -0.2569622
Lower ICC :[1] -0.6683824
Upper ICC :[1] 0.08756804
```

Task 3 : Can we predict an individual's personality? In this task, our goal was to build a classifier to predict an individual's personality traits (extraversion, agreeableness, conscientiousness, neuroticism, and openness). So we divided this task into two sub tasks which as follows: 1. Classification task: build a classifier to predict an individual traits as categories (high,low) 2. Regression task: build a

Figure 5 Corelation between Call In and Behaviour

```
Daily CALL IN DUR Extraversion ICC:[1] -0.3457488
Lower ICC :[1] -0.7677827
Upper ICC :[1] 0.01497609

Daily CALL IN DUR openness ICC:[1] -0.2941983
Lower ICC :[1] -0.714321
Upper ICC :[1] 0.05939391

Daily CALL IN DUR conscientiousness ICC:[1] 0.06655132
Lower ICC :[1] -0.3135421
Upper ICC :[1] 0.3535099

Daily CALL IN DUR agreeableness ICC:[1] 0.1568647
Lower ICC :[1] -0.2052632
Upper ICC :[1] 0.4228592

Daily CALL IN DUR  neuroticism ICC:[1] -0.1955795
Lower ICC :[1] -0.6095069
Upper ICC :[1] 0.1426439
```

regressor to predict an individual traits as numbers. Sample 3 data set [1] has been used to train and test to address both subtasks. We splitted the dataset into train and test as 85% and 15% respectively to train and test our models. We found one feature 'demogsex' as categorical, so we converted it into numerical feature (m:1 , f:0). We used all features(200) to train our models.

1. Classification Task:: As a part of this task, we used logistic regression scikit learn library model to learn the data. As sample 3 dataset has continous values for personality traits. So we converted those values into high and low categories using the threshold as zero, means values greater than zero are considered high and less than or equal to zero are considered as low. Model parameters as taken as follows: learning rate as 0.01, max iteration as 100 and error threshold as 0.001. As a part of result of this subtask, we found the best model with target trait as neuroticism, which is providing the promising result with area under curve (auc) value as 0.74 and worst model with trait openness with auc 0.39 on test dataset. Other models are slightly better than the random model. The results can be found in table 1 and implementation screenshot can be found in figure 7. Surely a question will raise in mind why we have used the logistic regression model, why not other model. The very simple reason is that it gives the probability of each class that can be used as the score and that can be used to further interpretation of individual trait.

2. Regression Task:: In this task, we tried several models like linear regression, lasso, decision tree regressor, randomforest regressor, support vector regressor, gradientboosting regressor, adaboostregressor using scikit learn library. We tried the reduce the features with autofeature library as well but that didn't work for us.

Figure 6 Corelation between SMS In and Behaviour

```
Daily SMS IN NUM Extraversion ICC:[1] -0.02712702
Lower ICC :[1] -0.07489065
Upper ICC :[1] 0.1544323

Daily SMS IN NUM openness ICC:[1] 0.06718044
Lower ICC :[1] -0.02747302
Upper ICC :[1] 0.3545156

Daily SMS IN NUM conscientiousness ICC:[1] 0.09117272
Lower ICC :[1] -0.01460414
Upper ICC :[1] 0.3956336

Daily SMS IN NUM agreeableness ICC:[1] 0.050627
Lower ICC :[1] -0.0361532
Upper ICC :[1] 0.3241456

Daily SMS IN NUM  neuroticism ICC:[1] -0.03156072
Lower ICC :[1] -0.07700371
Upper ICC :[1] 0.1431406
```

Table 1 Classification Model Result: Logistic regression model to predict the personality traits

Target Personality Trait	AUC
neuroticism	0.74
extraversion	0.55
conscientiousness	0.53
agreeableness	0.52
openness	0.39

Figure 7 Classification model(logistic regresion) result screenshot

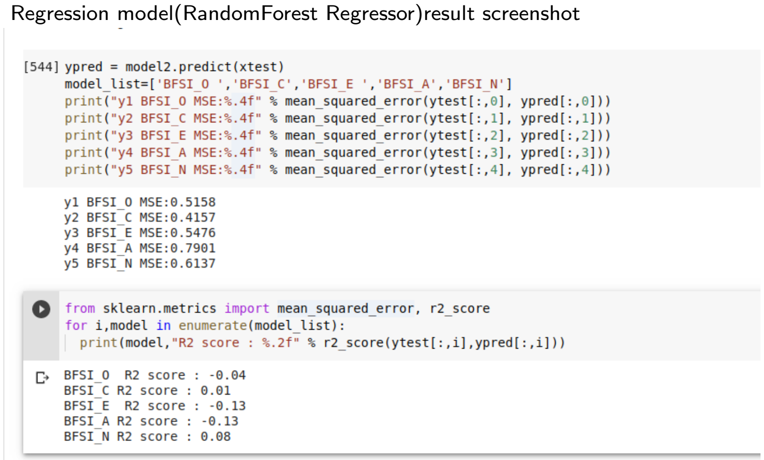
```
[376] ytrain1=np.where(ytrain>0,1,0)
      ytest1=np.where(ytest>0,1,0)

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix
from sklearn import metrics
for i, val in enumerate(model_list):
  model5 = LogisticRegression( C=0.01)
  model5.fit(xtrain, ytrain1[:,i])
  y_pred=model5.predict_proba(xtest)
  #cm = confusion_matrix(ytest1[:,2], y_pred)
  pred=np.where(y_pred[:,1]>0.5,1,0)
  #print(classification_report(ytest1[:,i],pred))
  fpr, tpr, thresholds = metrics.roc_curve(ytest1[:,i],z)
  print(val,'auc:',metrics.auc(fpr, tpr))

BFSI_O  auc: 0.3942307692307692
BFSI_C auc: 0.5333333333333333
BFSI_E  auc: 0.5555555555555556
BFSI_A auc: 0.5181818181818182
BFSI_N auc: 0.7361111111111112
```

However we found good results with randomforest regressor among all these models, so here we are listing only the results of it. the results of other models can be seen in code file. We used the parameter of randomforesregressor as follows: number of estimators=200, max depth=5 , max features='log2'. We used the MultiOutputRegressor class of scikit learn library to train all model together with their respective targets. We evaluated the model with mean squared error and R^2(coefficient of determination) regression score function. Best possible score of R^2 is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0. Further information about the R^2 can be seen at wikipedia [3]. We found the regressor model slightly better than the constant model with target traits neuroticism and conscientiousness. The results of regression models can be found in table 2 and implementation screenshot can be found in Figure 8. Our regression models are not performing well due to less number of training samples and there is a need to finetune the hyperparameters to get better results.

Table 2 Regression Model Result: Randomforest regressor model to predict the personality traits

Target Personality Trait	MSE	$R^2 score$
neuroticism	0.6137	0.08
extraversion	0.5476	-0.13
conscientiousness	0.4157	0.01
agreeableness	0.7901	-0.13
openness	0.5158	-0.04

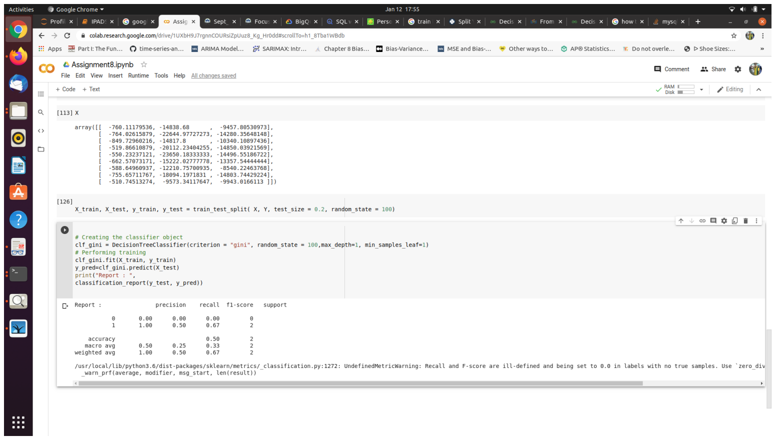
Figure 8 Regression model(RandomForest Regressor)result screenshot



Discussion

As part these kind of analysis, we can understand how individual's daily behavior plays a role in their daily lives. We can obtain basic descriptive details on how much people tend to socialize and when the pattern which is followed. We can map daily life behavior to psychological characteristics (e.g., personality traits, attitudes, values) and life outcomes (e.g., mental health, physical health) through these kind of analysis.

Figure 9 Corelation Between Stress and Conversation



Appendix

Code -

- Aman - Task1
- Suresh - Task3
- Frenny - Task2

Report -

- Aman - Scientific background, Data, Task1
- Suresh - Discussion, Task 3
- Frenny - Abstract, Goal and Task 2.

Author details

References

1. Sample 3 Dataset. https://osf.io/8ztd6/
2. Harari, G.M., Müller, S.R., Stachl, C., Wang, R., Wang, W., Bühner, M., Rentfrow, P.J., Campbell, A.T., Gosling, S.D.: Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. Journal of personality and social psychology (2019)
3. R$^2 ScoreFunction$.https://en.wikipedia.org/wiki/Coefficient$_o f_d etermination$