RESEARCH

# Introduction to the profile areas of data sciences: project 8

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 StudentLife study.

**Main results of the project:** The main results showed which features contributed in stress occurrence among students. Furthermore, Evaluation of student's current state based on other features.

**Personal key learnings:** We learnt how to create a schema/database like structure to handle large and complex datasets. Moreover, we performed data imputation dynamically and found correlation between features using Pearson's correlation coefficient.

**Estimated working hours:** 15

**Project evaluation:** 2

**Number of words:** 1268

## 1 Scientific Background

Many people experience stress, in one form or another, throughout their lives. Stress can be defined as "a state, which is accompanied by physical, psychological or social complaints or dysfunctions and which results from individuals feeling unable to bridge a gap with the requirements or expectations placed on them. Over the years, there has been a significant growth in the use of smartphones for health monitoring. Smartphone applications can be used to monitor physical activity,food intake, sleep quality, the menstrual cycle, and other issues related to health and well-being. Furthermore, monitoring can take place automatically through the sensors embedded within the smartphone, such as accelerometer and microphone, whereas others require users to interact with the application to record data. Subjective self-assessed stress can be measured using smart mobile devices via ecological momentary assessment (EMA). EMA is a collection of methods used to collect "assessments of subjects' current or recent states, sampled repeatedly over time, in their natural environment" Subjective stress can be assessed throughout the day using a time-based EMA where people are prompted to rate or answer questions about their "current stress level"

## 2 Goal

The aim of this project was to analyse the StudentLife study in particular the correlation stress with activity level and other features. This required importing the data into a database-like system as tables. 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. It was also to check other features for correlation with stress. Finally, it was to find out if the student's state can be predicted using other features

## 3 Data

The data used for the project came from the paper of Wang et al. The dataset was collected from all subjects, including automatic sensor data, behavioral interferences and self-reported EMA data with ground truth data including behavioral and mental health outcomes computed from survey instruments, as well as academic performance from spring and cumulative GPA scores provided by the registrar.

## 4 Results

### 4.1 Task 1: Is stress correlated to the activity-level?

Firstly, the data (.json-files containing EMA stress data for each student separately and also csv.files containing the activity data for each student separately) were loaded for students with available data as dictionary and the converted to pandas dataframe respectively. For EMA stress data some errors in the data could be identified (some rows of the column stress_level and location were in the column null_list to find. These incorrection could be solved by moving of respective rows to respective columns). As next the data were also checked for NaN-values. It could be 241 NaNs for stress_level identified and 0 NaNs for activity. After this preprocessing the data were imported to the database system Google Cloud Platform. As next, queries were generated to extract important figures for further statistical evaluations. First, statistics like average and standard deviation values for all data and for each student were extracted (see Figures 1, the output for activity data is for visibility not shown). In the next step, the average daily values as mean values

| index | all_student_id | total_stress_level_average | total_stress_level_std | total_stress_level_min | total_stress_level_max |
|---|---|---|---|---|---|
| 0 | 2408 | 2.269113149470916 | 1.350056056155241 7 | 1 | 5 |

| index | student_id | stress_level_average | stress_level_std | stress_level_min | stress_level_max |
|---|---|---|---|---|---|
| 0 | u00 | 2.263157894736842 | 1.3101493151813 56 | 1 | 5 |
| 1 | u02 | 2.096774193548387 | 1.325514710334 | 1 | 5 |
| 2 | u03 | 2.870967741935484 4 | 1.707667717168086 7 | 1 | 5 |
| 3 | u05 | 3.363636363636363 8 | 1.566698903601280 6 | 1 | 5 |
| 4 | u07 | 3.175438596491227 3 | 1.440768731039389 5 | 1 | 5 |
| 5 | u08 | 1.813186813186813 | 0.855207420091600 9 | 1 | 4 |
| 6 | u09 | 2.428571428571429 | 1.397276262011543 8 | 1 | 4 |
| 7 | u10 | 2.407407407407408 3 | 1.459977192479882 3 | 1 | 5 |
| 8 | u12 | 2.187500000000000 4 | 1.330474084799317 3 | 1 | 5 |
| 9 | u13 | 3.0 | NaN | 3 | 3 |
| 10 | u14 | 2.174999999999999 9 | 0.712075334533701 1 | 1 | 4 |
| 11 | u15 | 2.0 | 1.240347345892084 6 | 1 | 4 |

| index | student_id | stress_level | student_id_count | student_id_ratio |
|---|---|---|---|---|
| 0 | u00 | | 3 | 0.0379746835443038 |
| 1 | u00 | 1 | 32 | 0.4050632911392405 |
| 2 | u00 | 2 | 14 | 0.1772151898734177 2 |
| 3 | u00 | 3 | 11 | 0.1392405063291139 2 |
| 4 | u00 | 4 | 16 | 0.2025316455696202 5 |
| 5 | u00 | 5 | 3 | 0.0379746835443038 |
| 6 | u01 | 1 | 16 | 0.4444444444444444 |
| 7 | u01 | 2 | 6 | 0.1666666666666666 |
| 8 | u01 | 3 | 3 | 0.0833333333333333 3 |
| 9 | u01 | 4 | 9 | 0.25 |
| 10 | u01 | 5 | 2 | 0.0555555555555555 5 |
| 11 | u02 | | 4 | 0.1142857142857142 8 |
| 12 | u02 | 1 | 15 | 0.4285714285714285 5 |
| 13 | u02 | 2 | 7 | 0.2 |
| 14 | u02 | 3 | 1 | 0.0285714285714285 7 |
| 15 | u02 | 4 | 7 | 0.2 |
| 16 | u02 | 5 | 1 | 0.0285714285714285 7 |

**Figure 1** Statistics: average and standard deviation values for EMA stress data

for all students were imputed and extracted (see Figures 2, the output for activity data is for visibility not shown) In the next step, time series for the students u10 and u16 were analysed with the help of Prophet library by extracting components

1 to 25 of 2363 entries | Filter

| index | date | stress_level_average | ChangeIndicator | RowGroup | stress_level_average_FillDown |
|---|---|---|---|---|---|
| 0 | 2013-03-24 08:40:01+00:00 | NaN | 0 | 0 | NaN |
| 1 | 2013-03-24 08:40:58+00:00 | NaN | 0 | 0 | NaN |
| 2 | 2013-03-24 09:46:43+00:00 | 3.0 | 1 | 1 | 3.0 |
| 3 | 2013-03-24 09:47:00+00:00 | 1.0 | 1 | 2 | 1.0 |
| 4 | 2013-03-24 09:47:21+00:00 | NaN | 0 | 2 | 1.0 |
| 5 | 2013-03-24 09:47:56+00:00 | NaN | 0 | 2 | 1.0 |
| 6 | 2013-03-24 09:48:16+00:00 | NaN | 0 | 2 | 1.0 |
| 7 | 2013-03-24 09:48:49+00:00 | NaN | 0 | 2 | 1.0 |
| 8 | 2013-03-24 09:48:56+00:00 | 2.0 | 1 | 3 | 2.0 |
| 9 | 2013-03-24 09:49:20+00:00 | 1.0 | 1 | 4 | 1.0 |
| 10 | 2013-03-24 09:49:24+00:00 | NaN | 0 | 4 | 1.0 |

| index | student_id | stress_level | resp_time | null_list | location | date | ChangeIndicator | RowGroup | stress_level_FillDown |
|---|---|---|---|---|---|---|---|---|---|
| 0 | u00 | | 1364114401 | 43.75908069,-72.32885314 | 43.75908069,-72.32885314 | 2013-03-24 08:40:01+00:00 | 0 | 0 | |
| 1 | u00 | | 1364114458 | 43.75908069,-72.32885314 | 43.75908069,-72.32885314 | 2013-03-24 08:40:58+00:00 | 0 | 0 | |
| 2 | u00 | 1 | 1364169618 | 1 | | 2013-03-25 00:00:18+00:00 | 1 | 1 | 1 |
| 3 | u00 | | 1364177805 | 43.70677151,-72.28746626 | 43.70677151,-72.28746626 | 2013-03-25 02:16:45+00:00 | 0 | 1 | 1 |
| 4 | u00 | 1 | 1364177809 | 1 | | 2013-03-25 02:16:49+00:00 | 1 | 2 | 1 |
| 5 | u00 | 2 | 1364237696 | | 43.70692415,-72.2873929 | 2013-03-25 18:54:56+00:00 | 1 | 3 | 2 |
| 6 | u00 | 2 | 1364268806 | | 43.70555193,-72.28704778 | 2013-03-26 03:33:26+00:00 | 1 | 4 | 2 |
| 7 | u00 | 2 | 1364268814 | | 43.70555193,-72.28704778 | 2013-03-26 03:33:34+00:00 | 1 | 5 | 2 |
| 8 | u00 | 1 | 1364346740 | | 43.70678675,-72.28732051 | 2013-03-27 01:12:20+00:00 | 1 | 6 | 1 |
| 9 | u00 | 1 | 1364437527 | | 43.70508322,-72.28677496 | 2013-03-28 02:25:27+00:00 | 1 | 7 | 1 |
| 10 | u00 | 2 | 1364496192 | | 43.70514505,-72.28786512 | 2013-03-28 18:43:12+00:00 | 1 | 8 | 2 |
| 11 | u00 | 3 | 1364498245 | | 43.70492199,-72.28765189 | 2013-03-28 19:17:25+00:00 | 1 | 9 | 3 |

**Figure 2** Statistics: average daily stress data with imputation by last known values

such as trend, weekly and daily (see Figures **??** and 4). For both students is trend of stress development growing with time. Looking at weekly component, it looks similar for both students, that the stress is high in the beginning and goes down until Friday. By the student u10 the stress grows on weekend again, by student u16 the stress grows on Saturday and goes down on Sunday. Looking at the daily component, it can be noticed that for the student u10 the stress is in the beginning of the day high and goes gradually down. For the student u16 the stress increases for period for 10 till 17 and is low for the rest of time. For both students is trend of

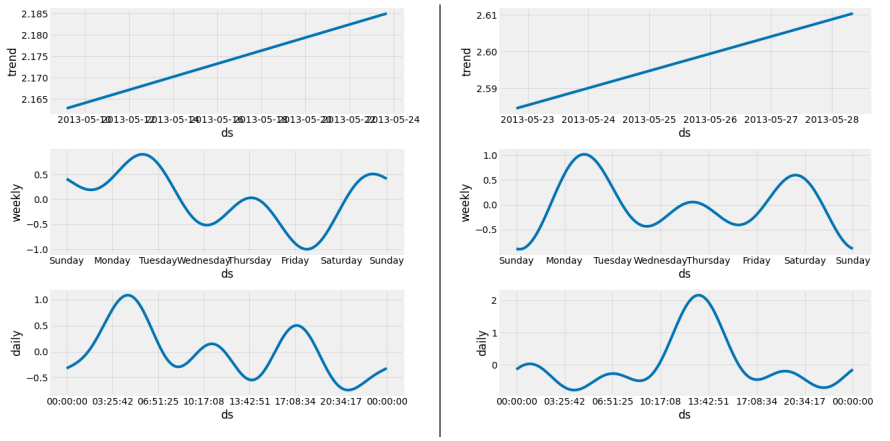


**Figure 3** Time-series for stress data of u10 and u16 : components

activity development growing with time. By weekly component it can be seen the both students have high activity on Tuesday and Wednesday and on Sunday and low activity on Monday and Wednesday. Looking at daily component it is to see that student u10 has low activity from 3am till 10am and high activity from 10 am till 12pm. Student u16 has low activity from 6 am till 10am and high activity from 10am till 6am. To estimate influence of activity on stress the Pearson correlation was estimated for student u10 and u16 for all the time (see Figure 6 and **??**). As can be seen, no significant correlation could be estimated. Performing correlation analysis on components show that activity has influence on the stress (see Figure 6 and 7) As it can be seen stress correlates negatively with activity both for weekly

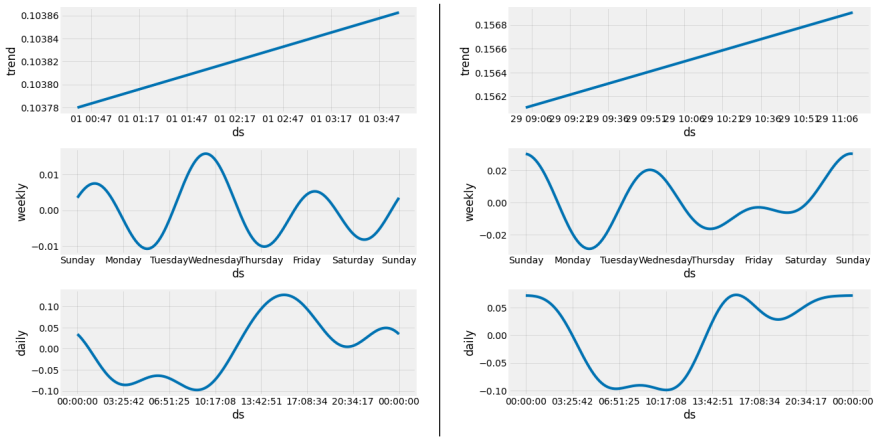


**Figure 4** Time-series for activity data of u10 and u16 : components

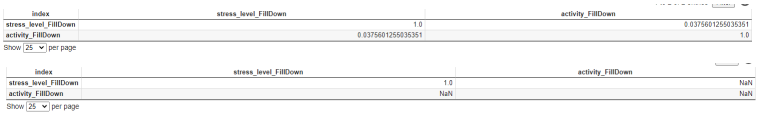


**Figure 5** Time-series for activity data of u10 and u16 : Pearson correlation between stress and activity for u10 and u16

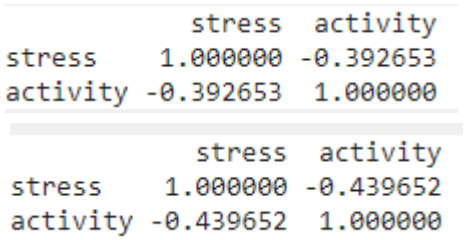


**Figure 6** Time-series for activity data of u10: Pearson correlation between stress and activity for u10 and u16 components: daily and weekly

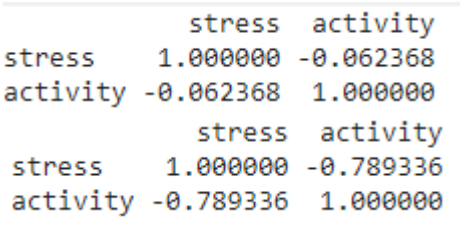


**Figure 7** Time-series for activity data of u16 : Pearson correlation between stress and activity for u10 and u16 components: daily, weekly

and daily components for u10 and correlates negatively only for weekly component for u16 and has no correlation for daily component of u16.

### 4.2 Task 2: What else is stress correlated to?

To identify other features that might be correlated to stress, we first chose student no 10 and 16 as our data. To discover correlations we computed Pearson correlation coefficient also referred to as Pearson's r. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. P-value evaluates how well your data rejects the null hypothesis, which states that there is no relationship between two compared groups. Higher value of p means how likely it is that you would receive these results assuming that there is no correlation or relationship among the subjects. We used scipy stat's pearsonr class to compute r and p values. Below table shows the correlation of stress with sleep, exercise and events. As one would naturally think, stress does relate to sleep. The correlation value r between stress and sleep came out to be 0.040(Shown in table) which suggests a strong positive correlation. Further, we analysed if events or exercise are correlated to stress. Although, computed Pearson's r showed positive correlation of 0.011 and 0.012 between stress and exercise and stress and events respectively, it wasn't as significant when compared to sleep.

Correlations of stress with other features:

| | Other Features | r | p-value |
|---|---|---|---|
| 0 | Stress and Sleep | 0.040342 | 0.672792 |
| 1 | Stress and Exercise | 0.011554 | 0.950811 |
| 2 | Stress and Events | 0.012278 | 0.952529 |

**Figure 8** Correlation of Stress with other features

### 4.3 Task 3: Can we predict a student's state?

The classification using knn Classifier and logistic regression with multi_class set to multinomial was done using GPS, activity, SMS and CALL as features to predict Perceived Stress.As outcome was estimated the accuracy of 0.26, which is too low with knn classifier. Using logistic regression as classifier accuracy 0 was estimated. So it seems that the combination of selected features is not good for predicting StressPerceived values

## 5 Discussion

Such a study can prove to be beneficial in multiple ways on the social system withing a university. Prevalence of smart phones and web provides ease in conducting such a study. One could monitor specific as well as average student's health condition. Based upon on the identified results, the university could draw further analyses on what affected student's health. For instance, if the average students' health is being

impacted by the study or work load then it suggests that revisions can be made in the study program itself. On the other hand, if a student is performing well then his life pattern can be observed and adapted by other students to improve their performance.

# 6 Appendix

| Name | Work Description |
|---|---|
| Michael | |
| Rohan | Abstract, Scientific Background, Results 8.2 and Code, Discussion |
| Natalja | Goal, Data, Results 8.1, 8.3 and Code |

**Table 1** Task responsibilities

**References**