RESEARCH

Introduction to the profile areas of data sciences: project 8

Eva Aßmann* and Paul Vogler

*Correspondence:
eva.assmann@fu-berlin.de
Institute of Mathematics and Informatics, Berlin, Takustr. 9, Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Goal of the project: Analyse the relationship of student's stress response data to the tracked activity and other sensoring data. Evaluate if it is possible to predict a student's psychological state based on the collected data.

Main results of the project: The correlation between the conversation frequency, duration and the stress level did not show any significant connection between the three features. Also the classifier for student's stress prediction showed poor accuracy.

Personal key learnings: Complex raw data sets are messy and it takes time to explore, process and analyse them. With complex data sets come various possibilities to filter, process, combine and mine. Exploring the StudentLife study fast and effective was highly correlated to our stress level over the last week.

Estimated working hours:10

Project evaluation:1

Number of words: 1934

Suggestions for improvements of the project:

1 Background and goal of the project

The StudentLife study collected activity, sociability, sleep and stress data for a class of 48 Dartmouth students over a 10 week term to assess the impact of workload on the student's mental health, academic performance and behavioral trends [1]. With mobile phones, passive continuous sensing data and active ecological momentary assessment (EMA) responses were tracked and yielded a broad amount of raw data. The goal of this project was to analyse the StudenLife data set comprising multiple tables for each instance regarding the relationship between stress responses and activity data, as well as the correlation between stress and other sensoring data. Also, it was investigated, if a student's stress level or psychological state could be predicted based on the collected data.

2 Brief description of the data set

All information collected within the scope of the study was categorised by the type of information and was structured such that within each information category the data for each study object (i.e. student) was stored in a single file which contained the anonymized student identifier (ID) in it's name. Only educational data and pre- and post-study-survey responses were stored such that each data file contained the information in a table with the student IDs as row index. Within this project, only EMA response and sensoring data was analysed. EMA data contained the EMA question definitions and possible responses in one file and stored the EMA responses as one file for each question and student ID. EMA data was stored in .json format. EMA stress responses stored the student's answers to the stress EMA. There were six possibilities of how to answer the question which indicate different levels of stress from one to five or no answer at all. In addition to the student's answer, the response time point and location were recorded. Location was recorded in GPS latitude and longitude or was unknown, the response time stamp was measured in Unix time within the eastern time zone (Tab. 1). On average, 3-13 EMA questions were fired each day, but the EMA schedules were set up on a week-by-week basis. Around assignments, multiples stress EMAs were scheduled per day and on some days, the same EMAs were administered multiple times.

question	'Right now, I am ...'	location
question_id	level	location
options	[1]A little stressed, [2]Definitely stressed, [3]Stressed out, [4]Feeling good, [5]Feeling great, NaN	GPS latitude & longitude

Table 1: Collected information from EMA stress responses.

Amongst other behaviours, sensoring data comprised the student's phyiscal activity which was detected using a classifier trained on accelerometer stream data from the study phones. Activity labels were generated every 2 seconds 24/7. To avoid draining the battery, activity inferences were made continuously for 1 minute and were then paused for 3 minutes before restarting again. Physical activity data for each student was stored in a single file containing the time stamps and inferred labels of the sensored activity (Tab. 2). The time stamps were measured in Unix time within eastern time zone.

Inference ID	Description
0	Stattionary
1	Walking
2	Running
3	Unknown

Table 2: Collected information from physical activity sensoring.

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

EMA Stress response data and activity sensing data for student IDs u10 and u16 were uploaded into Google Cloud Platform as two data sets containing one table for each ID. The JSON files with the EMA Stress response data were reformatted into new line delimited JSON using the command-line JSON processor jq in order to be used with BigQuery. The stress tables of ID 10 and 16 both contained 112 entries with response time point, location at the time of EMA response and reported stress level. For some time points ranging from march 24th to may 23rd), the answers to location or stress level were not assigned the correct column but were collectively stored in a column called 'null'. There were missing values in level and location for some time points which could have partly been filled with values from the null column. After we got lost in our tryings with Google BigQuery and SQL, we switched to python only, to at least get some insights into the data within the first task. Since we loaded the data into a pandas data frame, what surprisingly made the null column disappear, the idea of using also the null column for imputation came too late. Both u10 and u16 stress response data frames had missing location and level data for six and five time points, respectively. Missing values only concerned the first collection day and where imputed with the data from the next known day. Time series analysis using the Prophet library showed that stress response for u10 varied a lot over the complete value range in april, decreased very fast in the beginning of may and then slowly increased from beeing a little stressed to feeling great until the end of may (Fig. 1). Stress response values of u16 varied a lot over the range of april and may, to then beeing steadily increasing from beeing definitely stressed in the end of may to feeling good in june. This course of stress response over the months was also observed in the respective yearly trend components, as well as in the time series analysis for the joined stress response data set of u10 and u16 (Fig. 2). The very frequent variation between feeling a little stressed out, beeing definitely stress and almost feeling good during april in addition to the reduced stress in may suggests an examination phase. The observation that u16's stress phase duration was almost one month longer than for u10 could suggest, that either u10 is more resilient and recovers faster from stress or that u16 had to take more or make-up exams.

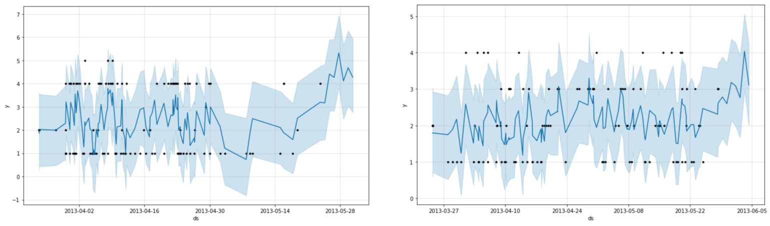

Figure 1: Time series data for IDs 10 (left) and 16 (right).

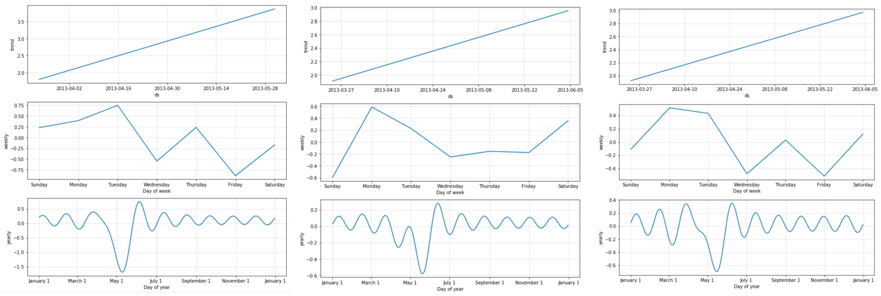

Figure 2: Components of time series data for IDs 10 (left) and 16 (middle) and the joined data set (right).

Activity sensor data for ID 10 contained 1062318 entries comprising information on time point and level of inferred activity. The table for ID 16 contained 981498 entries. For IDs 10 and 16 time stamps and activity inferences were missing for the last half of each data set when ordered by date, such that existing time points

Correlation Matrix	daily mean stress level
daily mean conversation duration (during morning)	-0.028
daily conversation frequency (during morning)	0.027
daily mean conversation duration (during day)	0.008
daily conversation frequency (during day)	0.054
daily mean conversation duration (during evening)	0.089
daily conversation frequency (during evening)	-0.03
daily mean conversation duration (during night)	0.036
daily conversation frequency (during night)	0.009

Table 3: Correlation of the daily conversation frequency and mean duraion during different times of the day with the daily means of the stress level.

ranged from march 27th to june 1st and march 27th to May 29th, respectively. The missing activity inferences were imputed with the ones from the last known time point. Since we did not come up with an idea how to impute the missing time points without skewing the time series analysis, we had to stop at this point.

4 Task 2: What else is stress correlated to?

To analyze the correlation of the stress level with the conversations of the students, we used the daily EMA stress level json files instead of the perceived stress scale (PSS) to produce a table similar to the table 6 from the Wang et al. [1] paper. Additionally the conversation csv files were used.

Each file for both conversation and stress contained only the information of one subject, so the files were merged on import with an added column with the user ID from the file name. Firstly the json file with the stress levels included some null-labeled values that belonged to either the location or stress level column, so they were moved to the appropriate column. Other reprocessing steps included removing the null value entries and grouping the data by user and date. Then the mean stress level was calculated per user and date.

For the conversation data the timestamps of conversation start and end were converted to a time format and all the conversation durations were computed. Then the conversations were grouped into four 'time of day' groups: morning, day, evening, night. The mean conversation duration as well as the frequency of conversations were computed per user, day and 'time of day'. Lastly a time period that was present in both data sets was chosen to perform a correlation analysis on.

The correlation was analyzed between the daily mean stress level and both the daily mean conversation duration as well as the daily conversation frequency, for all four 'time of day' groups (visible in Tab. 3). None of the correlations was particularly high, neither positive nor negative. Most of them stayed close to 0. One slightly higher correlation can be seen between the mean conversation duration during the evening and the stress level, with a correlation of 0.089. No underlying connections could be derived from these low correlations, using the shown data grouping and features.

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

We tried to predict the stress level of a student based on some of the processed features from the EMA sleep data set, as well as the previously used activity and conversations automatic sensing data. The stress level is once again derived from the mean EMA stress level values per day.

The EMA sleep set is preprocessed similarly to the EMA stress data set, because it also contained a null labeled column, that could be distributed between the other columns. But for all values <= 4 we could not determine if they should belong into hours of sleep, sleep rate or the social aspect, so we did not use the sleep data set for our classifier because of the time constraint and we were unable to solve the issue of these missing values. This will yield a limited model that does not capture a complete picture of student's lifestyle and stress causes. The activity data set was processed to show the mean activity inference per user and date. Additionally the user information was added to both the sleep set as well as the activity set. The stress and conversation daily averages were taken from the preprocessing of the previous step.

For classification the following features were used: daily mean activity inference, day/evening/night daily conversation frequency, day/evening/night daily mean conversation duration. These 7 features were classified for the target value stress level using a linear regression model on a 33%/67% test/train split for cross validation. The resulting classification yielded an R-squared score of 0.0237. The R-squared score assesses the amount of the response data's variability explained by the model. The score is defined as (1 -"residual sum of squares"/"total sum of squares") and ranges from the best possible score of 1 to below 0. A model always predicting the expected outcome value regardless of input would have score 0. The model's R-squared score showed that the selected features did not qualify for the stress level prediction using the linear regression classifier, because the input variable did not sufficiently explain the output variable of the regression.

6 Discussion

Tracking devices and the analysis and evaluation of all collected information as done in the StudentLife study provides multiple possibilities to change the communication between students, professors and university organisation unit and make the academic routine more dynamic and personalized. A model inspired by StudentLife could predict a student's overall or course-dependent performance based on his or her physical activty, social behaviour, sleep and study routine etc.. It could help students and other faculty members likewise achieving the optimal performance by recommending changes in lifestyle and discipline, but could also improve mental health by reminding of resting phases, social contact or exercise in order to prevent depression or burn-out. It could offer some kind of feedback for professors, e.g. if the workload was to big/small in comparison to workloads and stress levels of past terms or when to schedule exams in order to get the highest performance quota from the students.

Feedback

The StudentLife study was especially interesting, since the topic reflected our own lifestyle situation and issues what created a different relationship to the data than previous projects. Due to time constraints and the size of the data set, we could only do the required tasks and not fully explore the data. There was no particular bigger problem or difficulty with this project. It just took us more time this week to somehow organize the data and do the analyzes.

References

[1] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones." In Proceedings of the ACM Conference on Ubiquitous Computing. 2014.