Term Project

Yongshin Kim

Sensor Data Science (Fall 2020)

CONTENTS

Problem Statement

- K-EmoPhone (TA: Hansoo)
  - T1: Emotion recognition with mobile phone & wearable sensor data

- Replication paper
  - Bogomolov, Andrey, et al. "*Daily stress recognition from mobile phone data, weather conditions and individual traits*." Proceedings of the 22nd ACM international conference on Multimedia. 2014.

Problem Statement

- **What I want to predict? “Stress”**

  Using wearable devices and smartphones to track people's physical activities so that we can understand the relationship between physical activities and stress level.

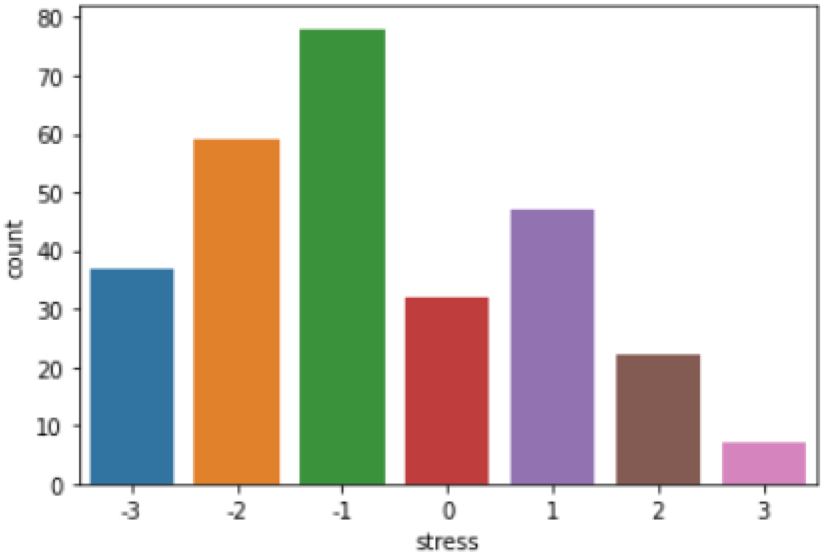
- **What others did in the past?**
  - Research on stress detection based on **voice analysis** considered different speech characteristics such as pitch, glottal pulse, spectral slope and phonetic variations.
    - ✓ However, these methods depend on sound quality, which is not granted in natural settings (e.g., crowded public places, noisy outdoor), and the correlation between speech and emotion is subjected to large individual differences.
  - Other studies focused on the **video analysis** of behavioural correlates of psychological stress.
    - ✓ These systems, despite providing an unobtrusive method for stress monitoring, cannot be employed in a large variety of real world and mobile environments and pose privacy concerns related to the recording of people’s behaviour.

Overview

- Dataset: K-Emophone
- Data preprocessing
  - eliminate missing values.
  - convert timestamp variable to a date.
  - set the time window three hours before and after ESM received.
  - combine ESM data with physical activity data.
  - devide stress data into two parts(non-stressed, stressed).
  - scale unification.
  - Outlier detection.
- Feature extraction
  - mean, median, min, max, variance, standard deviation.
  - use SHAP values to select important feature(feature selection).
- Classification
  - daily stress as a 2-class(binary) classification problem(non-stressed vs stressed).
  - decision tree(baseline model), support vector machines, **random forest**, Xgboost.
- Evaluation
  - K-fold cross validation (k = 10).

Approach

- Dependent variable



1. Stress level
   - To address the imbalance in stress data, the stress level was changed from 7 point likert scale to binary scale
     - ✓ -3, -2, -1 → -1 and 0, 1, 2, 3 → 1
   - stressed vs non-stressed

Approach

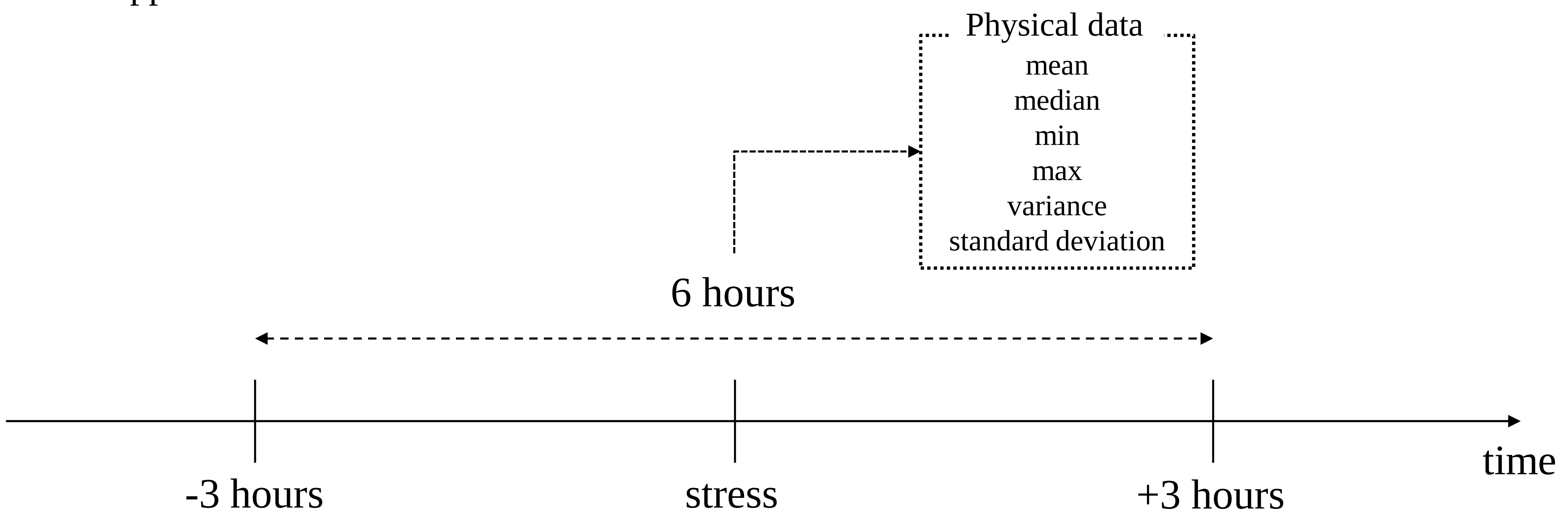
- Independent variables
  1. User info data
     - openness, conscientiousness, neuroticism, extraversion, agreeableness
  2. ESM data
     - valence, arousal, attention_level, emotion_duration, disturbance, emotion_change
  3. Physical data
     - heart rate, gsr, accelerometer-x, y, z

Approach

- Preprocessing
  - Eliminate missing values.
  - Convert timestamp variable to a date.
  - Set the time window three hours before and after ESM received.
  - Combine ESM data with physical activity data.
  - Devide stress data into two parts(non-stressed, stressed).
  - Scale unification.
  - Outlier detection.

Approach

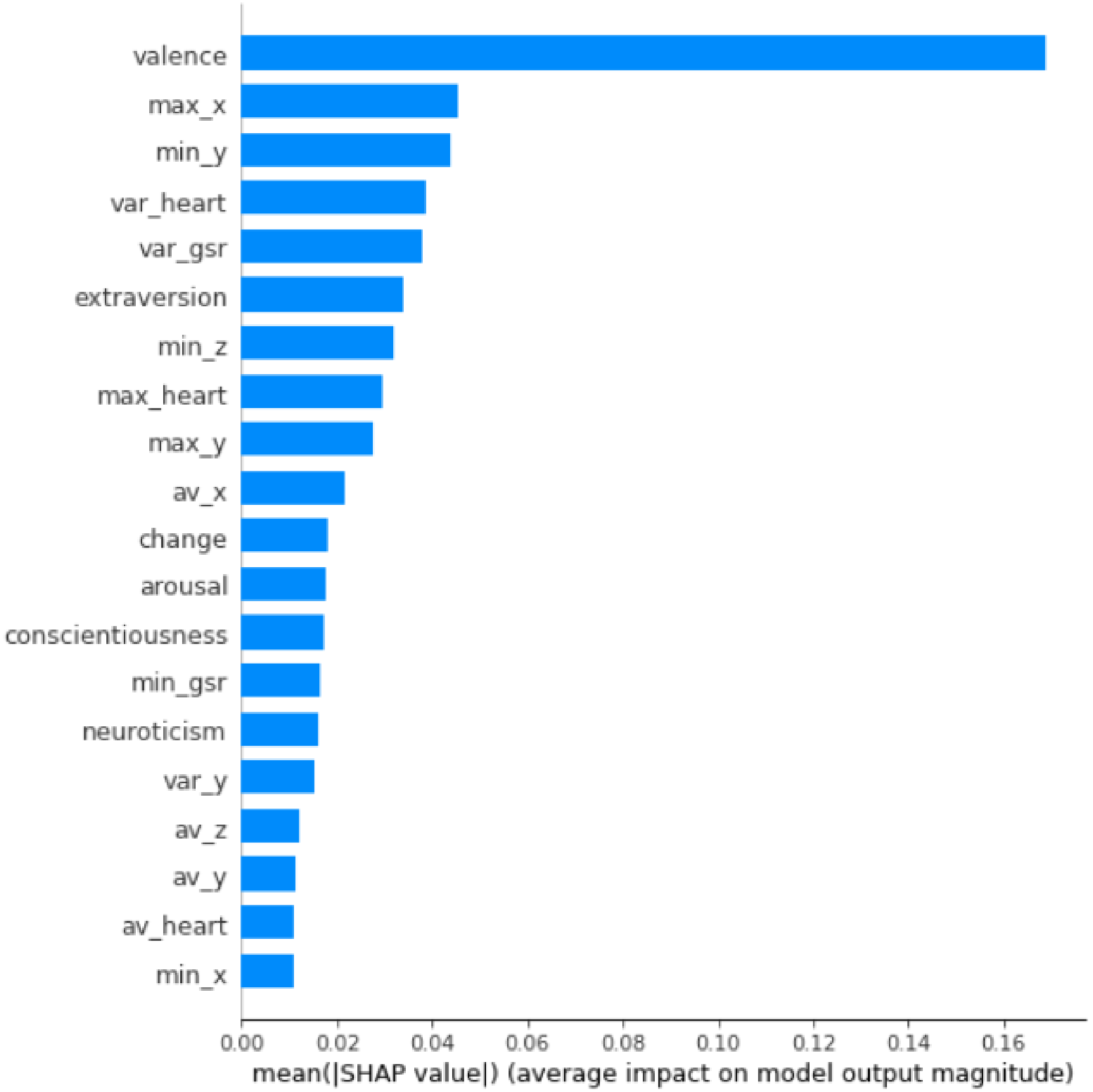
- Feature extraction
  - Because stress data is significantly less than physical data, features are extracted based on the time at which stress data is collected.
  - Feature will be extracted 3 hours before and after stress data is collected.
  - Mean, median, min, max, variance, standard deviation collected in a given time zone are applied to the stress data.



Approach

- Model building
  - Decision tree (baseline model)
  - Support vector machines
  - Random forest
  - XGBoost

- Feature selection
  - Use SHAP values to select important feature(feature selection).
  - The top 20 variables(out of total 41) were selected.



Evaluation

- K-fold cross validation (k = 10)
  - Referred to the replication paper.

- All features(n=41)

	Precision	Recall	Accuracy	F1-score
Decision Tree	0.583	0.586	0.599	0.583
SVM	0.824	0.712	0.761	0.716
Random Forest	0.686	0.660	0.702	0.665
XGBoost	0.730	0.667	0.723	0.672

- Selected features

	Precision	Recall	Accuracy	F1-score
Decision Tree	0.608	0.606	0.631	0.607
SVM	0.806	0.678	0.732	0.674
Random Forest	0.701	0.682	0.716	0.687
XGBoost	0.740	0.674	0.730	0.680

Conclusion

- The goal of the research was to predict people’s daily stress level from three different sets of data: 1) User info data; 2) ESM data; 3) Physical data.

- Taken together, the results provide evidence that individual daily stress can be predicted with about 70 percent accuracy.

- Among many features, SHAP values were used to find important features, which resulted in reduced model training time and improved evaluation metric.

Code review

- **Hansoo(TA)**

1. **Feature extraction and Evaluation**
    - The difference between feature extraction and feature selection, as well as evaluation and research goals, is not clear.
    - It is recommended to study how to grasp the feature importance by SHAP value for features extracted through SHAP values by rotating the entire model as taught in class.
2. **Why you use those models: decision tree, SVM, and XGboost**
    - There is no explanation as to why this model was used.
    - The baseline model doesn't exist in your code and explanation

- **Panyu(peer)**
    - recommend that more documentation about parameters are given.
    - for full dataset since it takes a lot of time, you may save it as a csv file.
    - compare with the baseline model

Reference

[1] Bogomolov, Andrey, et al. "Daily stress recognition from mobile phone data, weather conditions and individual traits." Proceedings of the 22nd ACM international conference on Multimedia. 2014.
[2] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury. Stresssense: detecting stress in unconstrained acoustic environments using smartphones. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp ’12, pages 351–360, New York, NY, USA, 2012. ACM.
[3] K. R. Scherer, D. Grandjean, T. Johnstone, G. Klasmeyer, and T. B¨anziger. Acoustic correlates of task load and stress. In INTERSPEECH, 2002.
[4] D. Giakoumis, A. Drosou, P. Cipresso, D. Tzovaras, G. Hassapis, A. Gaggioli, and G. Riva. Real-time monitoring of behavioural parameters related to psychological stress. Studies in health technology and informatics, 181:287, 2012.
[5] N. Bolger and E. A. Schilling. Personality and the problems of everyday life: The role of neuroticism in exposure and reactivity to daily stressors. Journal of Personality, 59(3):355–386, 1991.
[6] https://github.com/soyoungCf/shap
[7] https://colab.research.google.com/drive/1lwJC3S7On43OKuaECi1CS5pH8u9pPNJV

THANK YOU