Stress Detection

Sara Bay, Ben Butler, Ian Franda, Jack Grosskreuz, Annika Kennerhed

Data

We will analyze a dataset focused on stress detection, consisting of 3000 daily entries for 100 participants over 30 days. The dataset includes various physiological, behavioral, and physiological attributes.

Key variables include personality traits (such as Openness, Extraversion, Agreeableness etc.), daily perceived stress scale (PSS) scores, sleep metrics, and various phone usage indicators (e.g. screen-on-time, number of calls etc.) as well as skin conductance and physical activity.

Data source: https://www.kaggle.com/datasets/swadeshi/stress-detection-dataset

Questions

Our analysis will focus on the following questions:

a. Which factors are most correlated with stress levels?
    i. Which personality traits are most correlated with high stress?
    ii. How can stress levels vary throughout the week?
    iii. Is phone usage correlated with stress level?
    iv. How is sleep/physical activity correlated with stress levels?

We will use the results from our data to address these questions:

a. Are these determining factors easy to control or manage?
b. What would be potential ways to mitigate high stress levels?

Methods

We will be using the PSS_score (Perceived Stress Scale score, indicating stress level) as the dependent variable in our analysis. The PSS_score is an integer in the range 10 to 40, which is why we will be using a regression algorithm to determine the impact of other variables on the PSS_score.

We will train several models and perform model selection with cross validation. We will use 80% of the data for training and 10% for validation and model selection, and we will reserve the last 10% for testing the predictive power of our selected models. Part of our process will be to determine how we split the dataset. One option is to assign each row of the dataset to the training, validation, or test set at random. However, each patient occupies 30 rows representing 30 days. If stress levels between patients vary more than stress levels of a specific patient during the 30 day time period, it would be advantageous to split along patients. (80% of patients are training, 10% are validation, 10% are test). Choosing how we split our dataset will be the first step of our analysis.

One model of interest will be a Lasso regression model to identify the most significant predictors of stress levels. Lasso regression performs feature selection, penalizing less important variables, which is very suitable for this dataset with more than 15 variables that could potentially be predictors of the PSS_score. It would be especially interesting to compare how such a Lasso model performs when compared against ridge regression and gradient descent models. We will also look into correlation between the independent variables, and make sure to only include one variable from potential sets of collinear features to improve feature importance.

Reading in Data

The dataset can be read with Pandas as follows:

```
import pandas as pd
df = pd.read_csv("stress_detection.csv", index_col=0)
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
participant_id	3000.0	50.500000	28.870882	1.000000	25.750000	50.500000	75.250000	100.000000
day	3000.0	15.500000	8.656884	1.000000	8.000000	15.500000	23.000000	30.000000
PSS_score	3000.0	24.701000	8.615781	10.000000	17.000000	25.000000	32.000000	39.000000
Openness	3000.0	3.020663	1.159310	1.005003	2.024510	3.050115	4.029171	4.997405
Conscientiousness	3000.0	3.007883	1.140511	1.000982	2.055579	3.022064	3.981639	4.999137
Extraversion	3000.0	3.002101	1.143507	1.000584	2.018888	2.985548	4.005973	4.997642
Agreeableness	3000.0	3.047659	1.161074	1.002206	2.043813	3.091778	4.050223	4.999881
Neuroticism	3000.0	2.963589	1.158624	1.000173	1.974606	2.940948	3.955566	4.996408
sleep_time	3000.0	7.002145	1.160442	5.003291	6.017217	6.978221	8.029503	8.999948
wake_time	3000.0	6.990567	1.161225	5.001927	5.992499	6.982260	7.999644	8.998371
sleep_duration	3000.0	7.477953	0.867602	6.000561	6.710527	7.463421	8.222883	8.999061
PSQI_score	3000.0	2.490333	1.121454	1.000000	1.000000	2.000000	3.250000	4.000000
call_duration	3000.0	29.717149	17.563027	0.002886	14.544219	29.465154	44.831582	59.983074
num_calls	3000.0	9.362000	5.755055	0.000000	4.000000	9.000000	14.000000	19.000000
num_sms	3000.0	24.472333	14.553940	0.000000	12.000000	24.000000	37.000000	49.000000
screen_on_time	3000.0	6.624776	3.145666	1.006874	3.909183	6.700975	9.339519	11.997871
skin_conductance	3000.0	2.762022	1.282752	0.501595	1.670218	2.769001	3.843533	4.999104
accelerometer	3000.0	1.317038	0.688541	0.100791	0.731260	1.320422	1.910750	2.499946
mobility_radius	3000.0	0.803164	0.402527	0.100041	0.455688	0.793928	1.148284	1.499890
mobility_distance	3000.0	2.801677	1.306818	0.501622	1.682666	2.801065	3.971290	4.999929