Sleepiness Classification Using Empatica E4

—

By Litian Zhou, Bangyao Zhao
Chenyi Yu, Ningyuan Wang, Qingzhi Liu

Can we use the Empatica E4 to predict whether the student is sleeping?

E4 & Data

From all five of us!

- 3-axis Acceleration (ACC)
- Electrodermal Activity (EDA)
- Heart Rate (HR)
- Temperature (TEMP)

Expectation

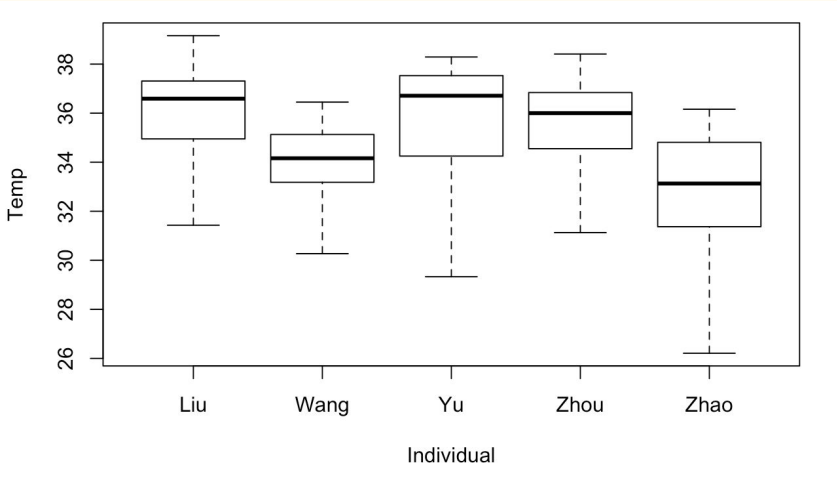
During Sleep time, relatively

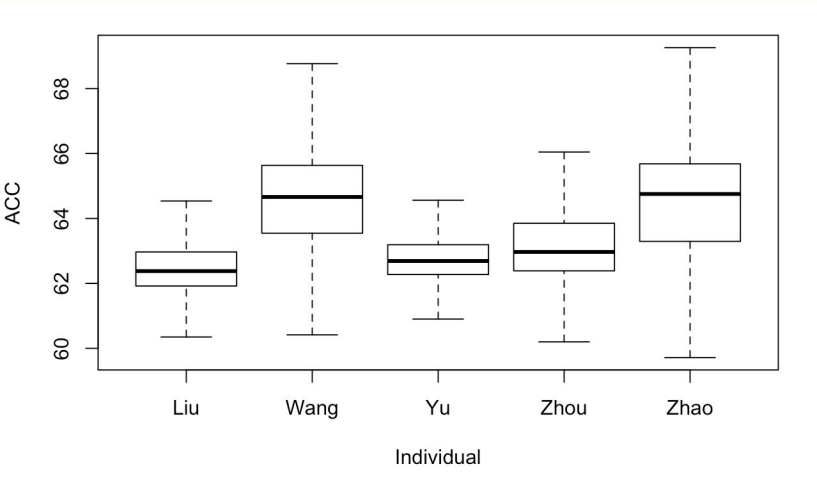
- Lower ACC
- Lower EDA
- Lower HR
- Lower TEMP

Compared to the one during daytime

Descriptive statistic

	Duration Total (sleep) Hours	HR Range (mean) bpm	TEMP Range (mean) °C	EDA Range (mean) μS	ACC Range (mean) 1/64 g
Qingzhi Liu	34.30 (6.00)	[45.12, 201.65] (74.29)	[34.95, 39.16] (35.85)	[0.00267, 9.61] (0.27)	[1.41, 221.13] (62.74)
Ningyuan Wang	45.24 (13.33)	[50.56, 159.28] (71.73)	[26.41, 36.45] (33.99)	[0.001281, 7.39] (0.40)	[2.45, 220.55] (65.10)
Chenyi Yu	43.27 (15.05)	[52.00, 166.3] (73.61)	[27.57, 38.29] (35.66)	[0.01565, 11.46] (0.81)	[3.74, 193.38] (63.29)
Bangyao Zhao	46.50 (17.83)	[44.42, 186.48] (78.24)	[31.37, 36.16] (32.66)	[0.000267, 8.05] (0.78)	[2.24, 221.7] (65.10)
Litian Zhou	39.60 (8.26)	[41.53, 196.83] (76.89)	[34.55, 38.41] (35.23)	[0.00027, 9.61] (0.68)	[2.83, 221.70] (65.65)





Preparation Steps

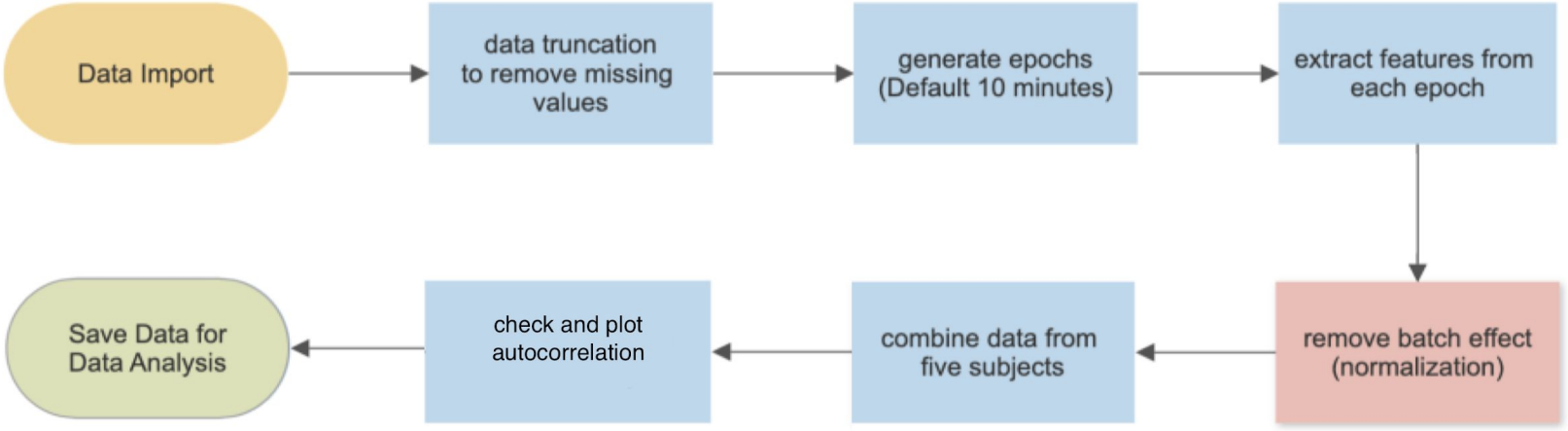


Figure 1. *Project Workflow*

Data structure

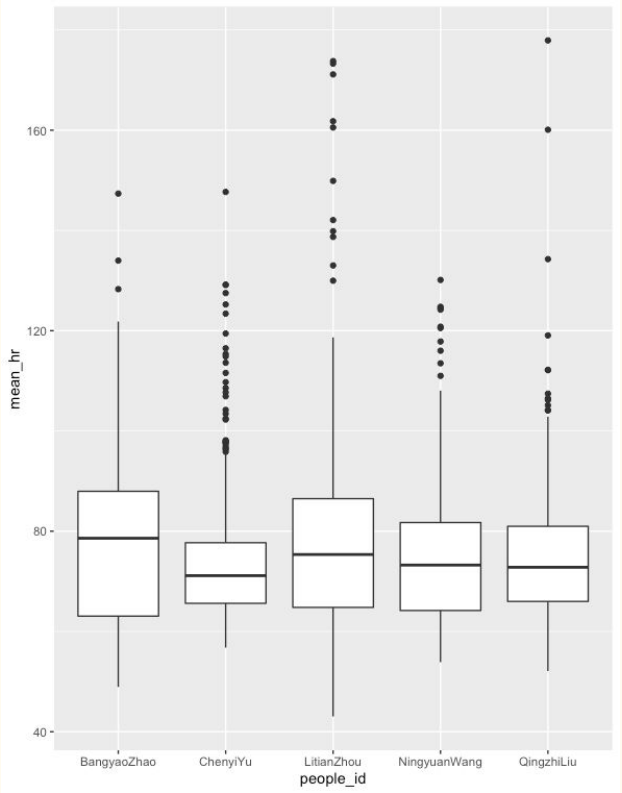
Outcome:

Sleep, not sleep

Covariate:

mean_acc, sd_acc, mean_eda, sd_eda, mean_temp, sd_temp, mean_hr and sd_hr, people_id

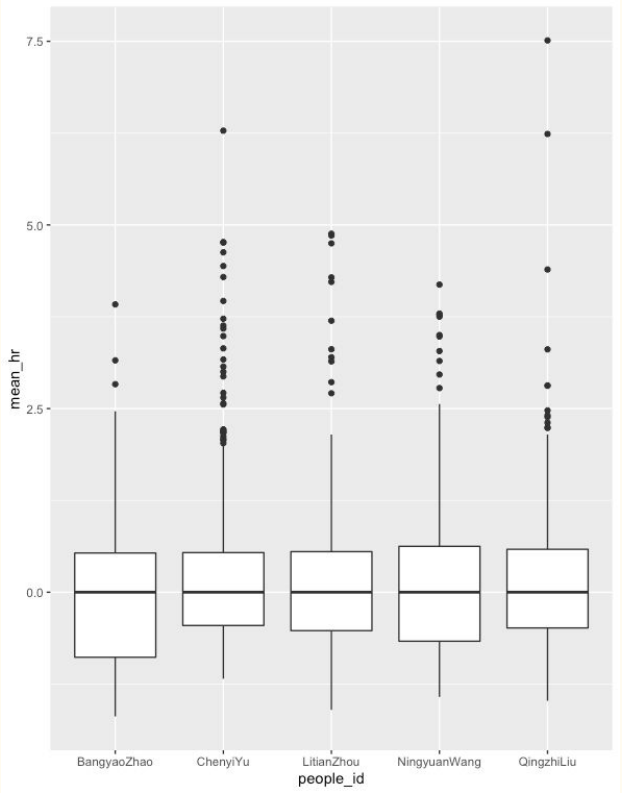
Box plot showing batch effect



<before

Normalized by centering at mean and divided by standard error

after>

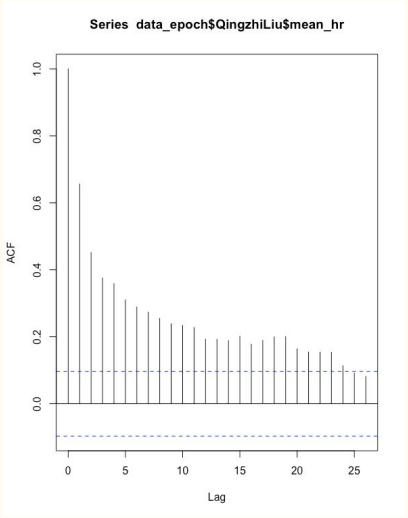


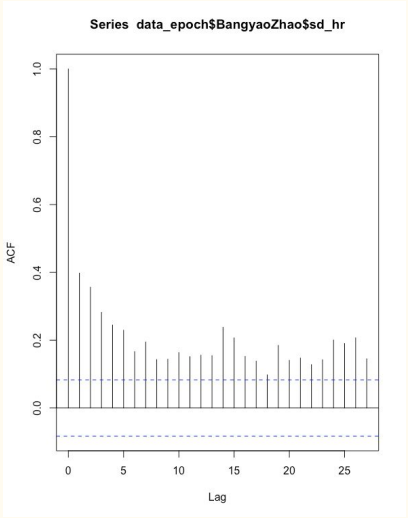
Autocorrelation plots

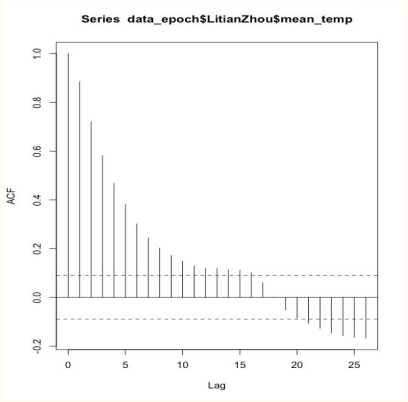
1. All five members' mean heart rate autocorrelation is lower than 0.4 after Lag = 5.

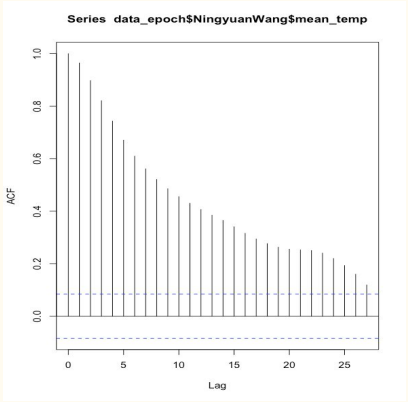
2. The heart rate standard error has even lower autocorrelation.

3. Other features, like temperature, varies from person to person.









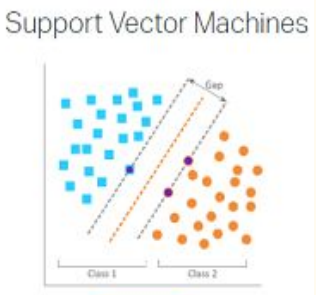
Methods: Models for prediction

Machine learning methods

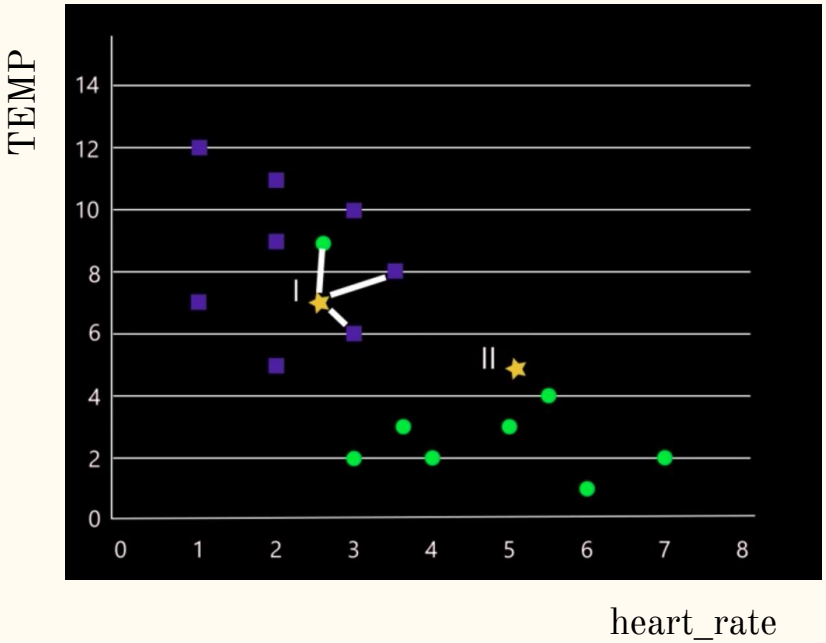
- KNN
- Tree-based methods
- SVM

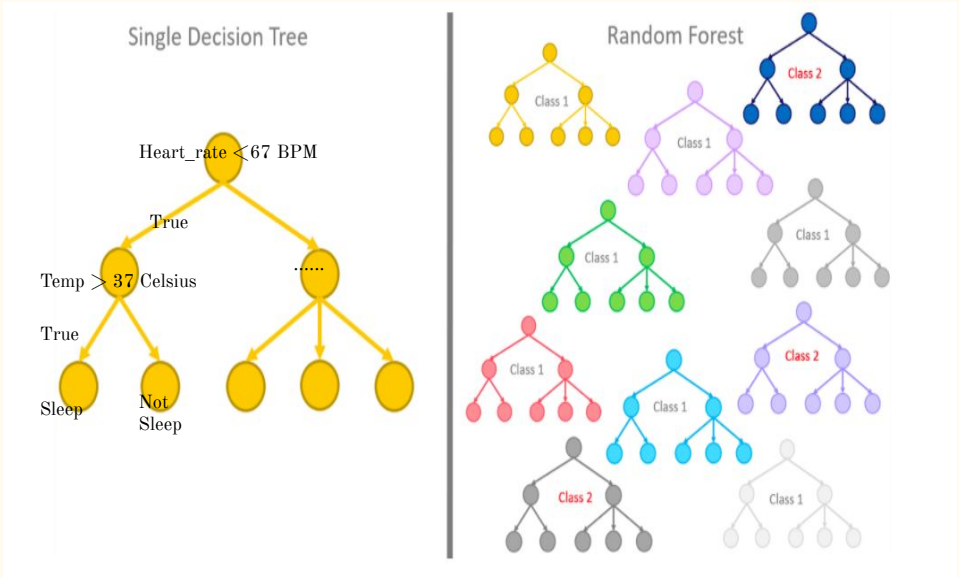
Classical statistical model

- Logistic Regression



KNN & Tree Methods

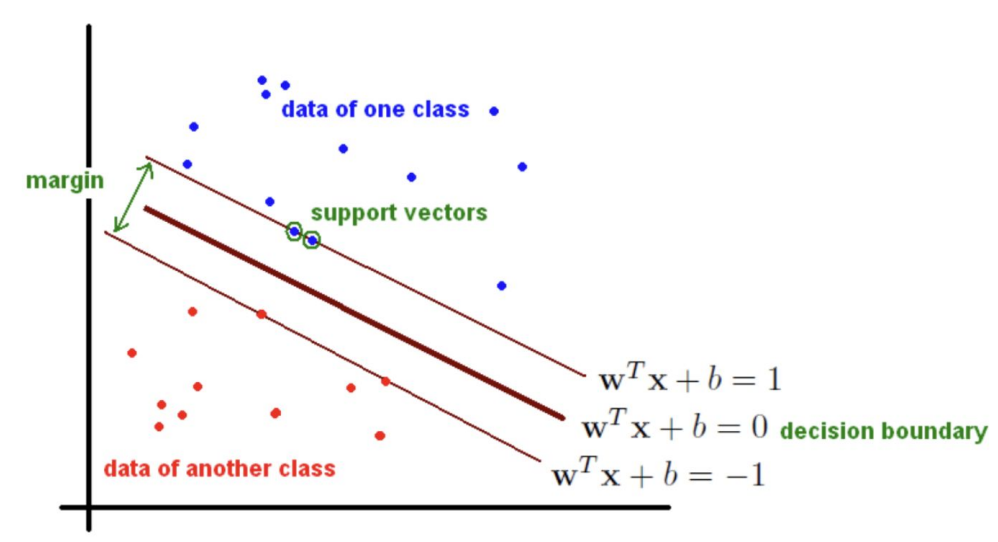




https://www.geeksforgeeks.org/k-nearest-neighbours/
https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147

SVM

- SVM for separable data



- SVM for nonseparable data

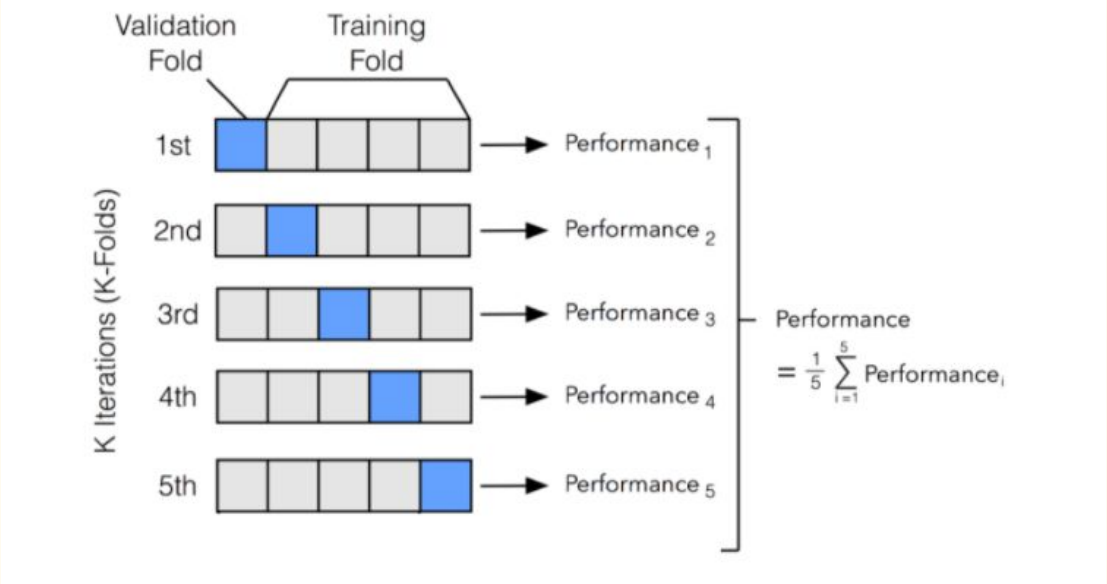
Logistic regression

- A generalized linear model (GLM)
- Response variable is binary
- Predict Sleep_Status = 1 if p(x) >= 0.5; Sleep_Status = 0 if p(x) < 0.5.

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + x \cdot \beta$$

Methods: training and validation

- Implement 5-fold cross validation for each model
- In each iteration, split data into training (4 subjects) and testing (1 subject) parts
- For the 4 different models, the accuracy on testing data is calculated and compared.



http://ethen8181.github.io/machine-learning/model_selection/model_selection.html

Methods: statistical inference

- Fit the final model by using 5 subjects' data and the best algorithm.

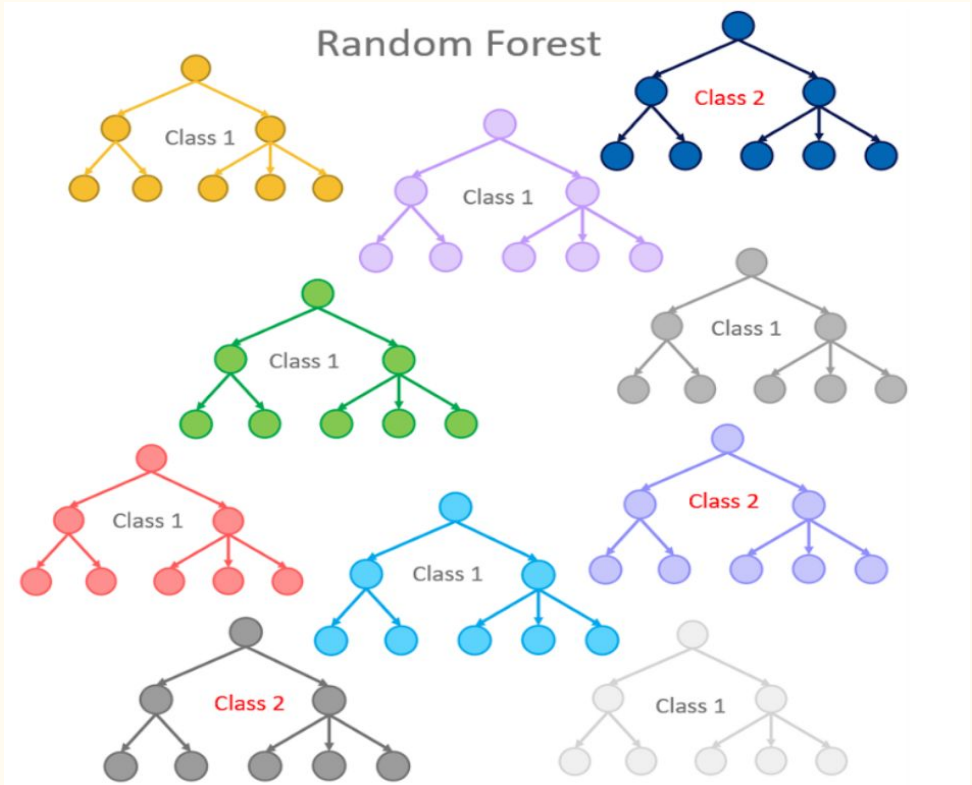
- Select the variable that is highly correlated with sleep status, based on certain criterion.

Results: model performance

Accuracy Table

	Random Forest	Logistic Regression	KNN	SVM
Litian Zhou	0.95	0.95	0.94	0.94
Bangyao Zhao	0.91	0.87	0.89	0.87
Qingzhi Liu	0.85	0.87	0.87	0.87
Ningyuan Wang	0.96	0.96	0.86	0.88
Chenyi Yu	0.84	0.78	0.82	0.7
Total	0.9	0.89	0.87	0.85

Winner: Random Forest Model



https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147

Remarks

- Bootstrap resampling
- Select features from a random **subset** of features → Decorelating

Example

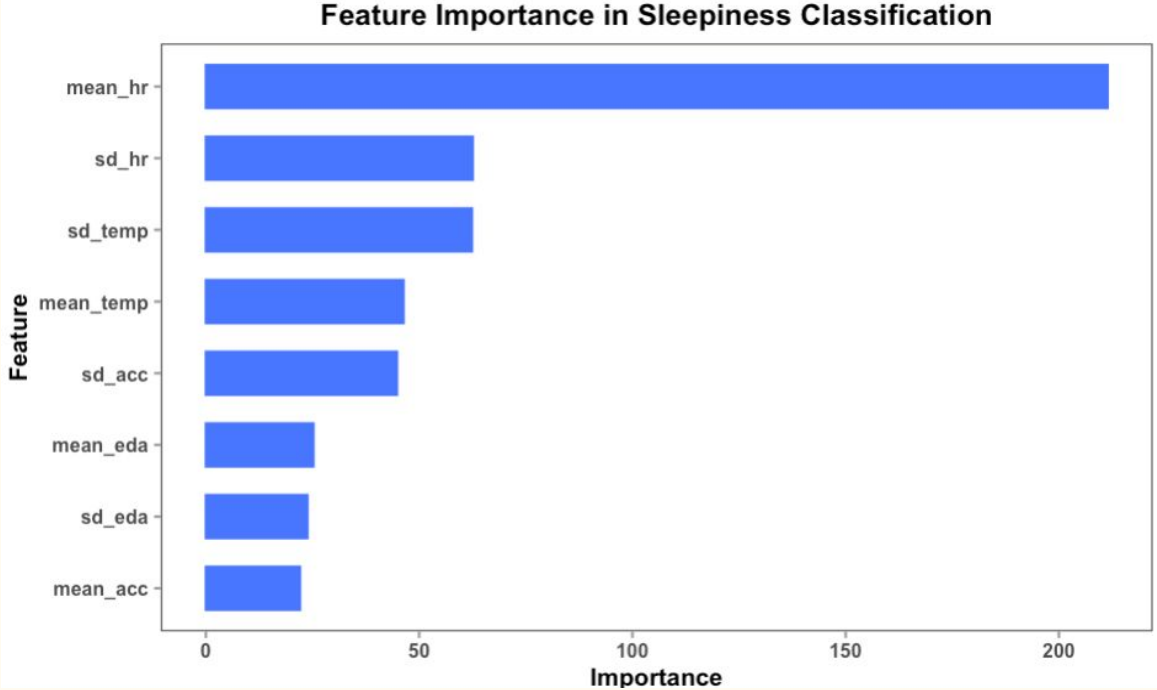
- 8 features in total
- Feature candidates in each split is 3

Tune Random Forest Parameters

- Mtry
  - Number of features sampled as candidates at each split
  - Mtry = 2
- Ntree
  - Number of trees to grow
  - Ntree = 480

Important Features in Random Forest

Based on mean decrease Gini



Important features

Features in Random Forest [ranked by **Gini**]

- Mean_hr
- Sd_hr
- Sd_Temp
- Mean_temp
- Sd_acc

Features in Logistic Regression [by **Significance**]

- Mean_hr
- Sd_temp
- Mean_eda
- Mean_temp
- Sd_eda

Logistic Coefficients

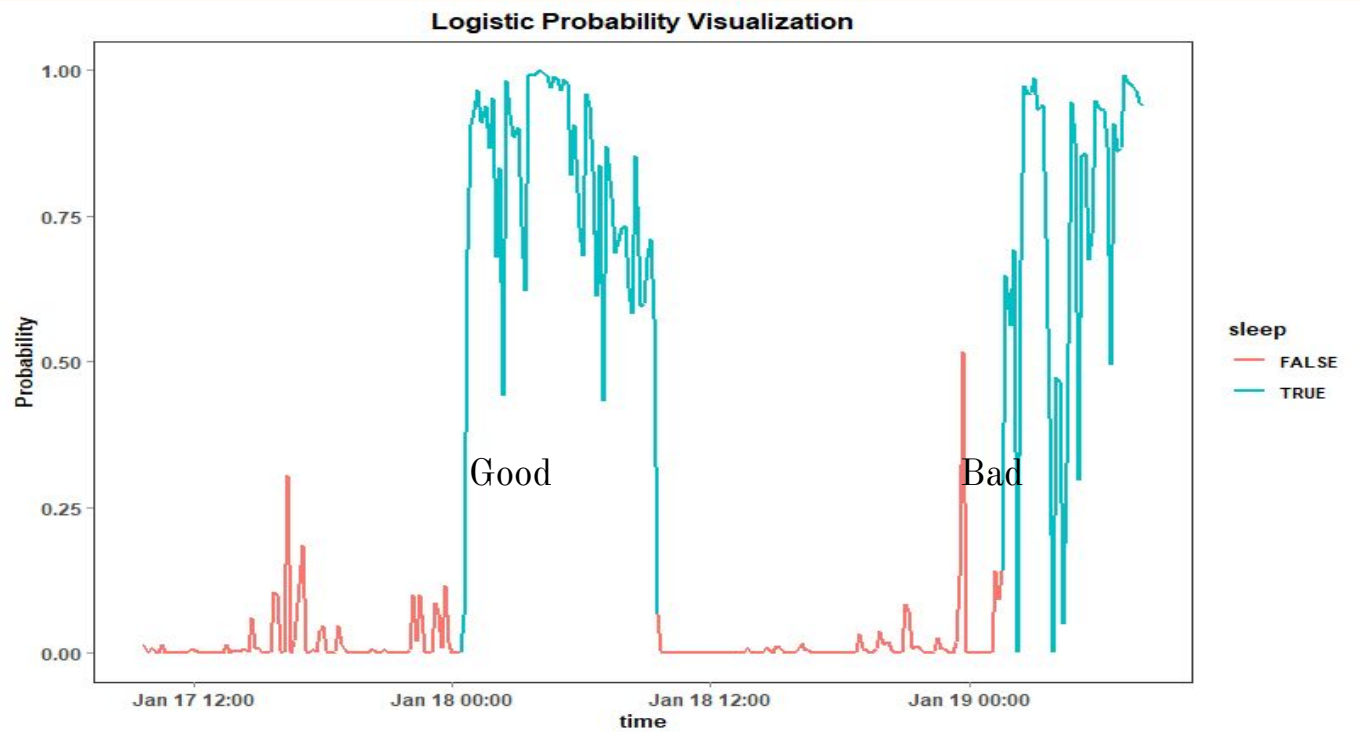
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  21.97856    6.09554   3.606 0.000311 ***
mean_hr     -13.83536    0.77768 -17.791  < 2e-16 ***
sd_hr        -0.09175    0.04768  -1.924 0.054327 .
mean_eda      0.20465    0.04114   4.974 6.56e-07 ***
sd_eda       -0.04303    0.01216  -3.539 0.000402 ***
mean_acc    -12.35554    5.81318  -2.125 0.033550 *
sd_acc        0.01145    0.01539   0.744 0.456997
mean_temp     0.16899    0.04507   3.749 0.000177 ***
sd_temp      -5.96742    1.00075  -5.963 2.48e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2962.3  on 2501  degrees of freedom
Residual deviance: 1225.3  on 2493  degrees of freedom
AIC: 1243.3

Number of Fisher Scoring iterations: 9
```

Logistic probability as sleep quality score?

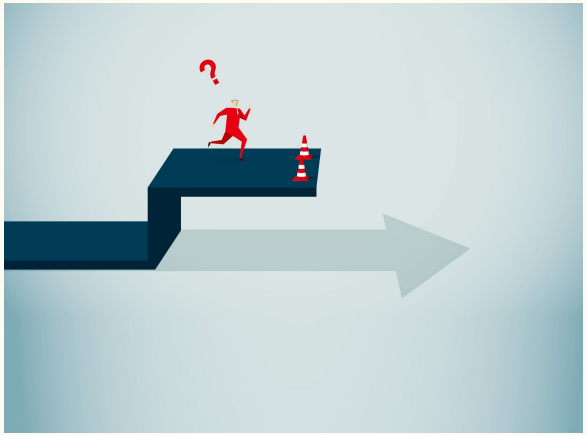


Diary:
The second night sleep was bad due to the effect of coffee

Limitation of this sleep quality score

1. Prediction accuracy of logistic model is not the best

2. Insufficient data to justify the interpretation



Strength

1. High model accuracy
2. Able to identify significant variables
3. Data normalization

Weakness

1. Unable to evaluate sleep quality
2. Independence assumption

1. Explore the possibility of sleep quality evaluation with more data
2. Further justify the model assumptions (e.g. independence)
    a. Try different epoch length
    b. Autocorrelation diagnosis

We enjoy the data analysis of our own data, and we believe that wearable devices have high potential to enhance personal health.

By one of our group member

Inspiration


QA


THANK
YOU!