For office use only	Team Control Number	For office use only
T1 _____	**50193**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**C**	F4 _____

2018

MCM/ICM

Summary Sheet

Effects of Sleep on Human Body

Our basic model has two parts: firstly using Decision Tree to subdivide the data of the sample then get the weight of the factors that influence sleep quality, Secondly using Logistic Regression to classify the sample about the diagnosis and sleep features. Thence we analyze the importance of each factor that may influence peopleąŕs sleep quality and the relation between diagnosis and sleep features, finally the results can provide some reference to the process of actual medical treatment.

Our major assumptions toward the model are to suppose the target variable is only related to the variables contained in the raw data, and to suppose that each sample value is independent.

First, we preprocess the raw data and plot a chart of sample distribution then find out the uneven property of the sample. We adopt non-linear algorithm Decision Tree with C5.0 to subdivide the data of the node and CHAID as an extended model which optimize the result. By analyzing the two trees we draw a conclusion that age is the most important factor to influence peopleąŕs sleep quality, among each age range the weight of factors are different.

Contents

1 Introduction

1.1 Statement of the Problem

Sleep quality affect peopleą́s daily life in a large degree for the mental states and daily energy depend on oneą́s sleep quality last night. However, the number of insomnia is large with the rate of insomnia rising now. Therefore, researching and analyzing the factors influenced sleep quality and the relationship between diagnosis and sleep feature are very important, for they can provide effective reference to the medical treatment.

1.2 Our work

2 Detailed Definitions and Assumptions

2.1 Detailed Definitions

1. We do not take into account other factors that may influence sleep quality and diagnosis but the factors listed in the raw data.
2. We suppose that all the diagnosis are only decided according to the sleep features listed in the data,neglect the probability that other sleep features may affect the diagnosis.
3. To measure the importance of each attribute,we suppose that all the factors are independent.

2.2 Assumptions

3 The Models

3.1 Question1

In combination with Linear Regression and Logical Regression, we regard question 1 as a mathematical classification problem.

First, we try to use correlation coefficient to indicate the Contribution Rate of the factors. However, the three correlation coefficient in science of statistics need premise and assumptions. Pearson Correlation Coefficient can be effectual when the parameters are linearly dependent, in addition Spearman Correlation

Table 1: three correlation coefficient

Symbol	value
$Spearman$	0.004135
$Pearson$	1.363992
$Kendel$	0.003177

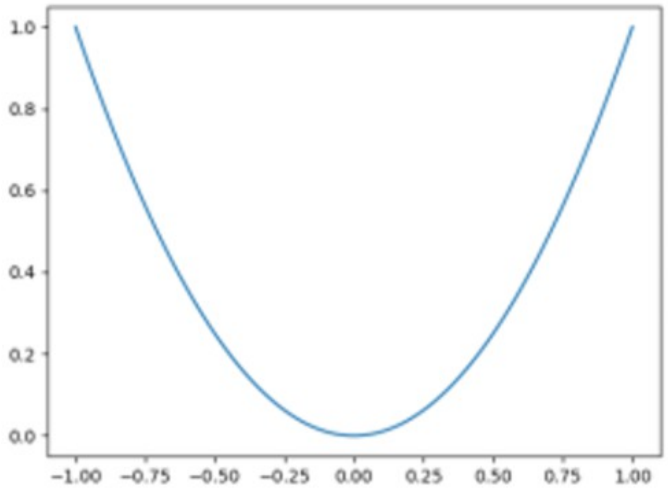

Figure 1: no name

Coefficient and Kendall Correlation Coefficient should satisfy the condition that curve monotonic increase.Then we obtain one counter example below.

Above that, sex is the first weighting correlation. But the result is inaccuracy according to our actual experience, so we decide to adopt non-linear classification algorithm.

We build the model as followed step:

Step1.Data cleaning: Based on Annex I, We preprocess the raw data to clean the incorrect and irrelevant data. We adopt two methods to clean the raw data:one is to digitize the variables, and the other is to delete the uncorrelated data. For the one, We use 0,1,2... to transfer the classification into number aiming at the category. For the other,we delete the uncorrelated data such as the features 'source' and 'number'. Finally, aiming at the missing data, we delete them directly for it has little weight among the whole data. The process and result are listed as Table 2:

Step2.Train data: We use Decision Tree to automatically detect the classification value of each candidate argument, and select the most meaningful factor and classification tree, therefore the divided groups has a higher consistency to reflect the attributes. We apply Information Gain Ratio to split the data set for the basic model and chi-square value and probability for the extended model, and compare the two model results to draw a more available and optimal conclusion.

Table 2: Data Cleaning

Feature	Type	Method	Old Values	New Values
Number	Continuous Numerical	Uncorrelated Deletion	151001001	-
Sex	Classification, String	Digitization	female	1
Sex	Classification, String	Digitization	male	0
Source	String	Uncorrelated Deletion	'Outpatient'	-

Basic model: First, define the cleaned data, for which we select Sleep quality as the target variable and other features as classification variables.Here we adopt C5.0 classifier to build the decision tree. The definitions of variables are listed in Table 3 as follow:

Table 3: Data Cleaning

Name	Type	Measure	Role
Age	Numeric	Scale	Input
Sex	Numeric	Nominal	Input
Sleep quality	Numeric	Nominal	Target
Reliability	Numeric	Scale	Input
Psychoticism	Numeric	Scale	Input
Nervousness	Numeric	Scale	Input
Character	Numeric	Scale	Input

In this method, the meaning of the variables in the formula are listed as Table3. The information entropy can be calculated as :

$$Ent(D) = -\sum_{k=1}^{|y|} p_k log_2 p_k \tag{1}$$

Among this, the smaller is $Ent_R(D)$, the higher the purity of D. It illustrates that we can subdivide the data of sleep quality according to feature R. In this way, feature R make the maximum impact to sleep quality when $Ent_R(D)$ reach the minimum.

Then we suppose that sleep quality may be influenced by not only one feature but the different combinations of the features. Upon this, we suppose that feature R have V values $R^1, R^2, ...R^V$, then we can divide D into V child classes for each child class contains all the sample values R^v according to feature in D. Because of the different amount between different child classes, thus we give different weight $\frac{|D^V|}{|D|}$ to different child class. That is, the larger sample size of R^V is, the greater the impact that feature R influence on D.Then we have:

$$Gain(D, R) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{2}$$

Among them, when $Gain(D, R)$ reach a lager value, it illustrates that we can obtain a purer sample by using feature R to subdivide the data. However, when use information entropy to subdivide the data set, it may tend to select the characteristics with more values. In order to solve this problem, we optimize the model with Information Gain Ratio. Then we have:

$$Gain_{ratio(D,R)} = \frac{Gain(D, R)}{IV(R)} \tag{3}$$

$$IV(R) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} log_2 \frac{|D^v|}{|D|} \tag{4}$$

Finally, we build a iteration tree depended on features. We know that different combination of the features have different effect on sleep quality, thus we can draw a conclusion that the feature which is closer to the root node affect sleep quality more, for the feature which is farther to the root node affect sleep quality less. The feature that isnąŕt contained in the combinations can be neglected.

1. We do ...

2. We do ...

3. We do ...

4 Sensitivity Analysis

The primary notations used in this paper are listed in **Table 4**.

Table 4: Notations

Symbol	Definition
A	the first one
b	the second one
α	the last one

5 Strengths and Weaknesses

5.1 Strengths

- Only one ...

5.2 Weaknesses

- First one ...
- Second one ...

6 Conclusions and Discussion

References

[1] Elisa T. Lee, Oscar T. Survival Analysis in Public Health Research. *Go.College of Public Health*, 1997(18):105-134.

[2] Wikipedia: Proportional hazards model. 2017.11.26. `https://en.wikipedia.org/wiki/Proportional_hazards_model`