Assess your daily health

BIS634 Final project:Health Assessment Webpage

Yujin Han

dept.biostatistics, GSAS of Yale Univeristy, New Haven, U.S.A
Email: yujin.han@yale.edu

The final project of BIS634 is a health assessment website. This website helps users to assess the daily health by predicting a health score on 5 levels (excellent, very good, good, fair, poor) and give the user health advice based on their final score. The specific solution is that we chose the BRFSS-2013 as the raw dataset, which meets the FAIR principles and rich in health information. After data pre-processing, we first trained the Xgboost multi-classification model using a 5-fold cross-validation method, and then trained the K-means algorithm to help find more similar groups as the user and the best health level , thus giving more reasonable health advice by comparison. We finally deployed models and built the web page by using Flask, Bootstrap, Chart.js.

Keywords: BIS 634; health assessment; flask; Xgboost; k-means; Chart.js; Bootstrap4

1. BACKGROUND

Health assessment is the evaluation of the health status by performing a physical exam after taking a health history. To help people monitor their health more efficiently and in real time, this final project has been designed using data analysis, machine learning, flask, etc. to complete a webpage that can provide a comprehensive prediction and data visualization of the user's health.

Specificallyusing a unique ID, the user of the website can combine the corresponding personal health information stored by healthcare institutions (which are stable in the short term, e.g. height, presence of high blood pressure) with the entered health data (which are unstable in the short term, e.g. average daily sleep time, exercise time) to obtain a health assessment score. The website also provides data visual information and health advice based on their final health scores.

2. DATA SOURCES

The BRFSS -2013 Data was used in this project and it can be downloaded from the official website of the Centers for disease control and prevention (CDC). BRFSS means the Behavioral Risk Factor Surveillance Systemwhich is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.

The BRFSS -2013 was collected in 2013 and it meet the FAIR (Findable, Accessible, Interoperable and Reusable) principles. It also is described with rich metadata, includes Codebook Report, Survey report. BRFSS -2013 is open, free and public-use.

In detailBRFSS -2013 measures behavioural risk factors for the non-institutionalized adult population (18 years of age and older) residing in the US by Land-Line and Cell-Phone. There are 491775 samples and 330 variables related to health information, like diet, exercise, medical history, etc, in the data. The BRFSS -2013 also includes the general health (genhlth) feature to assess the health status of respondents, which fully meets the project's objective of predicting health level. The BRFSS -2013 is therefore an ideal dataset because of its large amount of data, the wealth of health information, its free and authority.

3. PREPROCESSING

The pre-processing consists of 3 steps.

3.1. Feature selection

As discussed in the background section, the data will be split into two parts. Features in part one are not change in the short term (e.g. ever have been told diabetes etc.) and stored by healthcare institutions. Part two is the data user input, which changes daily. Therefore, based on the principles, we had a total of 30 features from raw dataset .1. 23 of which are historical health variables, such as Ever Diagnosed with Heart Attack, Height, and 5 of which are real-time variables, such as How Much Time Do You Sleep, Minutes or Hours exercise. 1 of which is general health with 5 Health Levels, and we will use it as a label for training multinominal model. 1 of which feature is the State FIPS Code, which reflects the geographical distribution of the sample (only be used for visualization, not for modelling).

3.2. Missing Value

We further processed the missing values. Firstly we removed the samples without labels (general health). Secondly, we calculated the missing rate for each feature and drew a missing rate chart 1.

The missing value plot indicated that the missing rate of features is not severe, only feature somokeday2 with a missing rate higher than 20% . Somokeday2 is a categorical variable that reflects the frequency of smoking (never smoked, some days, every day) and is an important variable to measure the health habits of respondents. After consideration, somokeday2 was not removed, although its missing rate was 56.3%. After processing the missing values, the sample size became (489791,30)

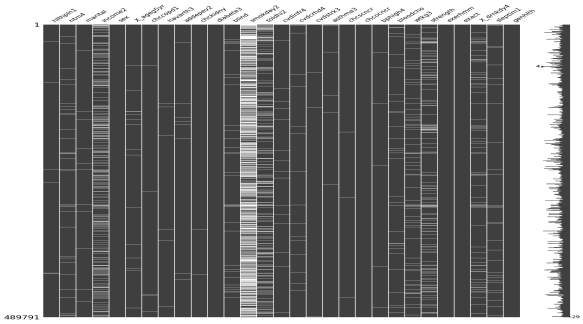

FIGURE 1. Missing rate

3.3. More Processing

More operations were performed to clean up the data

- Replace 999, 777 in the dataset used to indicate don't know, not asked with NA
- Replace weekly exercise/sleep time with daily and change their units to hours or minutes
- Convert a specific type of sport e.g. (running, swimming) to a number of sport types,for example, 2 types of exercise and the value is 2
- Convert weight and height to KG and meters for BMI calculation
- Add unique IDs. Additionally, a column of unique IDs was added for subsequent matching of historical health variables with real-time data variables entered by the user

4. DATA VISUALIZATION

4.1. Descriptive statistics

4.1.1. Geographical distribution

The redder the colour in the above graph2, the larger the sample size, and the greener the sample size, the smaller the sample size. And Florida has the largest population in the chart, followed by Kansas.

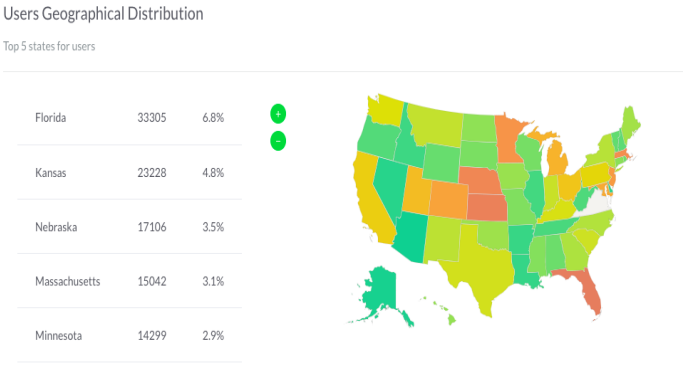

FIGURE 2. Geographical distribution

4.1.2. Numbers in each health level

The doughnut chart 3 above showed that people in Very Good (32.5%) and Good (30.7%) health levels accounted for 63.2% of the total sample. The Poor level had the smallest group, only accounting for 5.7%.

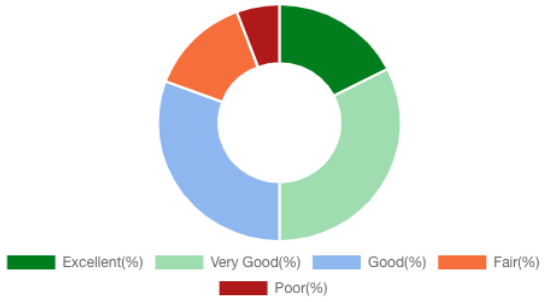

FIGURE 3. Numbers by health level

4.1.3. Sex ratio by health level

We can notice that 4 there are more women than men in each health level, with the highest proportion of women in the poor health class at 61.7%. This is likely due to the fact that women answer the telephone at home more often than do men and are more likely to be willing to cooperate [1]. This phenomenon illustrates the side effects that the BRFSS data set suffers from survivor bias due to the survey method.

4.1.4. Average exercise and sleep time by health level

We can tell from the diagram 5 that as health status deteriorates, average sleep time, aerobic exercise time and anaerobic exercise time decrease. Excellent groups is almost three times longer than Poor in average aerobic exercise time. The above linear graph may reflect a possible correlation between average sleep and exercise time and health status.

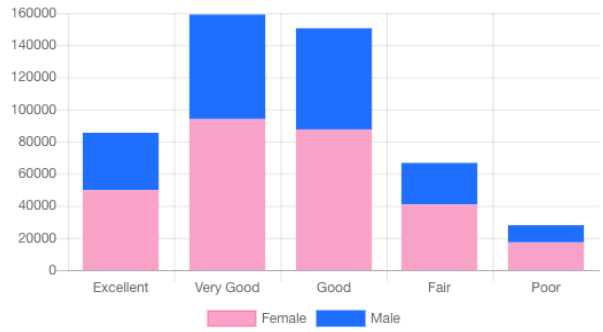

FIGURE 4. Sex ratio by health level

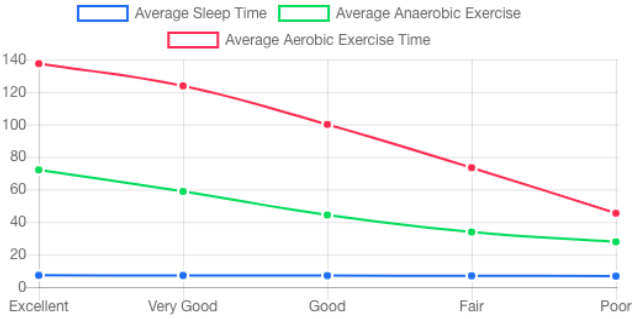

FIGURE 5. Average exercise and sleep time by health level

4.1.5. Typical Persona by health level

The figure 6 shows typical persona of these five levels and we can get a lot of information from it. Most of people in each level are married woman, which is consistent with the results from the previous linear plot analysis

Second, as health deteriorates, group income decreases and the number of pre-existing diseases begins to increase.

We further explored how smoking and drinking differed across the health level groups. Specifically, if one smokes every day and the scores is -10, if one smokes some days and the score is -2 and if one never smokes and the score is 2. Finally, the average score for each class and the average number of drinks per day were calculated.

We can find that the healthier the group the lower the smoking score, but the higher the average daily alcohol consumption. We cannot conclude from this result that the more alcohol is consumed, the healthier people is. Because the poor group generally has a lower income and the low average daily drinking may also be effected by the low income that prevents them from purchasing alcohol. More complex causal inference models and more confounding factors need to be controlled for to explain the relationship between alcohol consumption and health.

4.2. Exploring the relationship between variables and labels

The Pearson coefficient and significance (p¡0.01) for each feature and label(see my python file) was calculated, and only 22 of the 28 features were significantly correlated with the label 7 .2. However, the correlation coefficients for each feature and label are very small. And the largest significant correlation coefficient is only 0.05 coming from the feature strength.

5. TRAINING MODEL AND RESULTS

5.1. Training model

Two models were trained. The first one, Xgboost, was used to make final predictions on the user's data, and the second one, K-means, was used to find groups with similar health conditions to the user and groups with the best health conditions thus we can give better health advice based on comparisons between groups.

We chose Xgboost as a multi-classification model because of its high tolerance for missing values, its good interpretability, making it easy to explain to users what varaibles can influence their health and Xgboost's consistently good performance on classification problems. After dividing the dataset into training and test sets in the ratio of 7:3, we selected the parameters using the 5-fold cross-validation method. After obtaining the health level predicted by the model, we converted it into a prediction score by using the formula 1.

$$healthscore = A * healthlevel + B \tag{1}$$

where B is a random integer from the interval [0, 19).

For the K-means algorithm, we set the initial number of clusters to 5

5.2. Results

We finally obtained the best Xgboost and drew the top 10 features in terms of importance 8 .2.

We can note that 8 of the 10 characteristics are variables used to describe the pre-existing disease, with the remaining two variables being income and average aerobic exercise time. This illustrates the important impact of pre-existing disease on health assessment.

6. DEPLOYMENT

In the last part, we deployed our trained Xgoost and K-means models on the webpage via flask. What's more, to make the webpage more beautiful, we used Bootstrap4 for the Html framework and chart.js for the graphics. Bootstap4 is the fourth version of Bootstrap which is a set of open source front-end frameworks for web site and web application development, including HTML, CSS and JavaScript frameworks and many components. Chart.JS is a popular and powerful data visualization library. We can create flexibility plots just by feeding certain parameters.

FIGURE 6. Typical Persona by health level

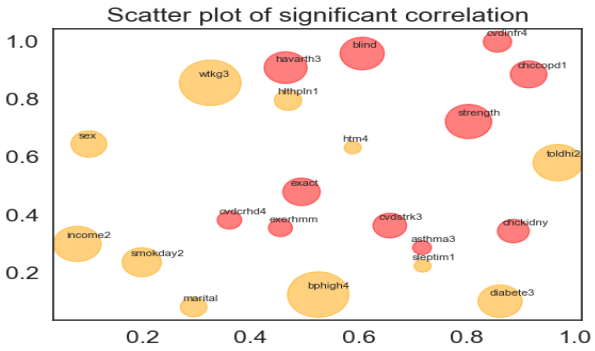

FIGURE 7. Pearson's correlation coefficient bubble chart

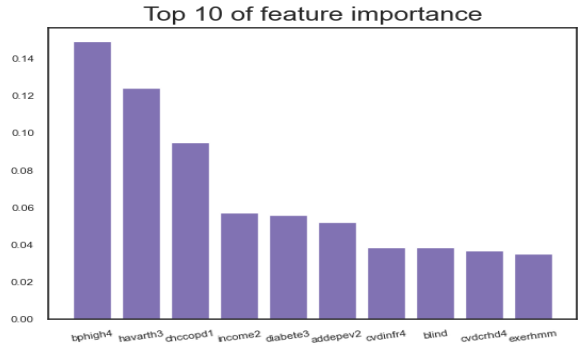

FIGURE 8. Feature importance

Three websits, Atlantis Lite Flask [2], Chart.js [3] and Boostrap-flask [4], are mainly referred to in the creation of webpages, and they all allow the user to create single personal website/app.

REFERENCES

[1] Bean J M, Johnstone B. Workplace reasons for saying you're sorry: Discourse task management and apology in telephone interviews[J]. Discourse processes, 1994, 17(1): 59-81.

[2] generator-atlantis-2021,Atlantis-Lite-Flask, https://github.com/app-generator/flask-dashboard-atlantis-dark

[3] noauthor-chartjs-nodate,Chart.js | Open source HTML5 Charts for your website, https://www.chartjs.org/

[4] noauthor-bootstrap-flask-nodate,Bootstrap-Flask-1.0.4-documentation,https://bootstrap-flask.readthedocs.io/en/stable/

Appendix .1.

TABLE .1. Feature Description

Feature	Definition
hlthpln1	Have any health care coverage
marital	Marital Status
income2	Income Level
sex	Respondents Sex
ageg5yr	Age
bphigh4	Ever Told Blood Pressure High
bloodcho	Ever Had Blood Cholesterol Checked
toldhi2	Ever Told Blood Cholesterol High
cvdinfr4	Ever Diagnosed with Heart Attack
cvdcrhd4	Ever Diagnosed with Angina or Coronary Heart Disease
cvdstrk3	Ever Diagnosed with a Stroke
asthma3	Ever Told Had Asthma
chcscncr	ever told you had skin cancer?
chcocncr	Ever told you had any other types of cancer?
chccopd1	Ever told you have chronic obstructive pulmonary disease, emphysema or chronic bronchitis?
havarth3	Told Have Arthritis servings per week 1-yes, 0-no
addepev2	Ever told you had a depressive disorder
chckidny	Ever told you have kidney disease?
diabete3	Ever told you have diabetes
blind	Blind or Difficulty seeing
smokday2	Frequency of Days Now Smoking
drnkdy4	Computed number of drinks of alcohol beverages per day
htm4	Reported Height in meters
strength	How many times did you do physical activities or exercises to STRENGTHEN your muscles?
exerhmm	average exercise time
exact	Number of sport types
sleptim1	average sleep time
wkg3	Reported Height in kg
state	State FIPS Code
genhlth	health level

TABLE .2. Pearson's correlation for features and labels

Feature	coefficient ($p < 0.01$
strength	0.0542
blind	0.0491
havarth3	0.0465
exact	0.0352
chccopd1	0.0338
cvdstrk3	0.0283
chckidny	0.0254
cvdinfr4	0.0201
cvdcrhd4	0.0150
exerhmm	0.0143
asthma3	0.0089
sleptim1	-0.0072
htm4	-0.0073
marital	-0.0177
hlthpln1	-0.0188
sex	-0.03242
smokday2	-0.03882
diabete3	-0.0488
income2	-0.0575
toldhi2	-0.0615
bphigh4	-0.0956
wtkg3	-0.0958

TABLE .3. Top 10 feature importance

Feature	value ($p < 0.01$
bphigh4	0.1493
havarth3	0.1244
chccopd1	0.0950
income2	0.0573
diabete3	0.0561
addepev2	0.0524
cvdinfr4	0.0389
blind4	0.0385
cvdcrhd4	0.0370
exerhmm	0.0351