NATIONAL UNIVERSITY OF SINGAPORE

FACULTY OF SCIENCE

AY 2021/2022, SEMESTER 2
ST4248: Statistical Learning II

Term Paper

Title:
Predicting Mental Health status amongst respondents

Matric Number: A0199513W

Summary:

In this term paper, various machine learning algorithms were used to accurately predict if an individual is likely to suffer from mental health problems based on socio-economic indicators and a mental health questionnaire. Additionally, factors that were significant in affecting mental health were selected and ranked. This will aid in the development of an early warning system, where individuals can utilise it to monitor their mental health and obtain timely treatment. This paper will cover the pre-processing of data, exploratory data analysis, models, feature importance, limitations, and conclusion.

Motivation

Mental health affects how we think, feel, and behave. It also determines how we communicate with others, handle stress, and make decisions. Thus, a decline in our mental health could lead to mental illnesses that would disrupt our daily activities. Mental illnesses such as depression increases the risk of many different types of physical health problems, such as diabetes and heart diseases. [1] Therefore, there is a need to identify potential individuals with poor mental health as early as possible, so that they can obtain timely treatment before complications arise.

Description of data

The dataset was obtained from Dryad, an international open-access repository of research data. [2] It contains responses of 5485 male and female rural-to-urban migrant workers who had worked in Shanghai for at least 6 months, regarding their health-related behaviours.

There are 4 main aspects in the dataset:

1. Socio-economic indicators (i.e. age, occupation, income, etc.)
2. Physical Health (i.e. height, weight, blood pressure, etc.)
3. Lifestyle behaviours (i.e. smoking, alcohol consumption, sleep, etc.)
4. Psychometrics

For the last aspect, a Chinese version of the Symptom Checklist-90-Revised (SCL-90-R)[3] was used to assess the mental health of the respondent. Respondents would answer a total of 90 questions on a scale of 0-4 (Not at All – Extremely) that covers a broad range of psychological problems. A total score of 160 and below indicates normal mental health, while a total score above 160 indicates abnormal mental health.

Problem Statement

This paper examines how an individual's profile (such as socio-economic indicators and psychometric instruments) would determine whether they are likely to have a good or poor mental health. The proposed algorithm aims to identify individuals suffering from poor mental health at early stages, so that there could be timely intervention to prevent their conditions from worsening and lead to severe repercussions.

Data Cleaning and Exploration

Feature Engineering and Response variable

The dataset contains 5484 rows and 120 features. It was discovered that there were quite a lot of NAs in one of the columns regarding smoking (Smoke at least once in the past 30 days). This was because respondent had answer No to the previous question of smoking at least 100 cigarettes in lifetime, thus leaving this question blank. As dropping these

NA values would result in a loss of 72% of the entire dataset, steps were taken to deal with it. Respondents were categorised into 3 categories by combining the `smokePast30` and `smoke100` columns into a new column called `smokeType`. The new categories are: Current smoker, Previous smoker, and Non-smoker.

A new column `healthStatus` was created, based on each respondent score in the SCL-90-R. It is the response variable containing binary values (0 signifying Good Mental Health, 1 signifying Poor Mental health).

Data Cleaning

Since the percentage of entries containing NA was not prevalent, all rows containing any missing values were removed. Next, correlation was checked between features to reduce multicollinearity in the dataset. Given the large feature space in this dataset, correlation between features or even with the response could result in erroneous predictions. Due to the mixture of categorical and continuous features in the dataset, different methods were used to calculate the correlation between each possible pair of variables and the results can be seen in Table 1.

Comparison	Method	Analysis	Result
Continuous variables	Pearson Correlation Coefficient	Moderate correlation found between: 1. `height` and `weight` (0.647) 2. `bphigh` and `bplow` (0.652)	1. Combine `height` and `weight` into a new continuous feature: `BMI` 2. Combine `bphigh` and `bplow` into a new categorical feature: `highBP`
Binary Categorical variables	Cramér's V	Most of the values < 0.5 and are close to 0. However, moderate correlation found for 54 unique pairs.	1. Remove `hasBlue` 2. Remove other 53 features (mainly from SCL-90-R) due to high correlation with response variable (`healthStatus`)
Non-Binary Categorical variables	Phi Coefficient	All values are either < 0.5 or > -0.5 and are close to 0	No correlation was found
Continuous & Binary Categorical variables	Point-Biserial Correlation Coefficient	Most of the values are either < 0.5 or > -0.5 and are close to 0. However, moderation correlation was found between `numSmoke` and `smokeType` (-0.7816529)	Removed `numSmoke`, since `smokeType` is a better representation of smoking

Table 1: Correlation analysis of features in the dataset

This resulted in a final data set of 5318 rows and 63 features.

Data Exploration

Figure 1 shows part of the exploratory data analysis conducted on some of the features in the dataset, which are likely to result in abnormal mental health.

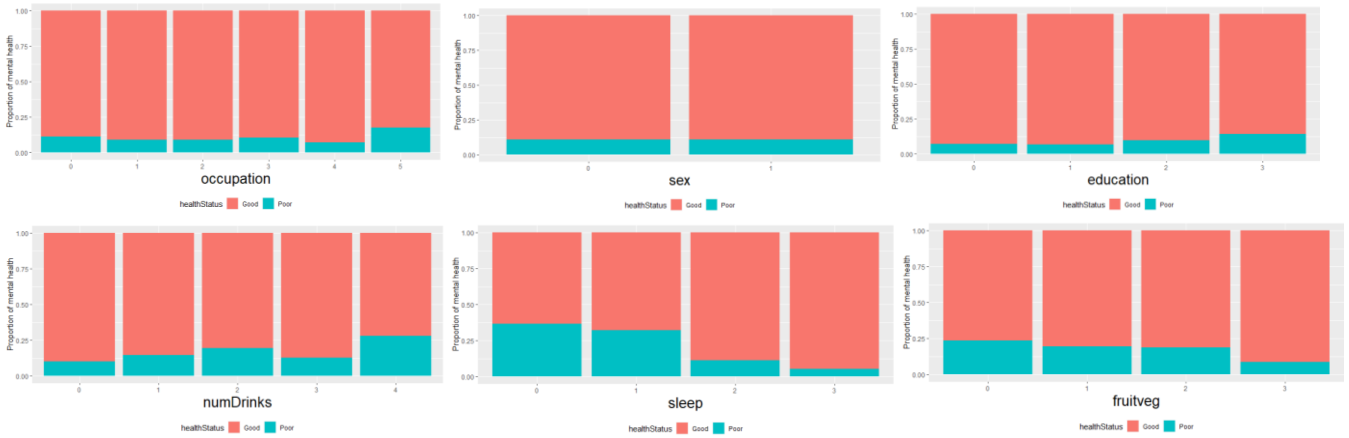


Figure 1: Plot of proportion of mental health in various categorical variables

From Figure 1, it is interesting to note that the proportion of poor mental health between gender is the same [4] and occupations in the Recreation/Leisure sector had a higher proportion of poor mental health as compared to occupations in the Construction and Manufacturing sector. For the rest of the variables, it is evident that there might be some relationship between them and mental health, such that the rate of decline in mental health differs in each category of the variable.

It was also discovered that there is a clear imbalance between the number of respondents having good mental health (89.1%) and poor mental health (10.9%). This would cause a problem in the modelling process and will be dealt with in the next section.

Models

Prior to building the models, a 70-30 train test split was carried out on the 5318 rows, resulting in 3723 rows of training data and 1595 rows of testing data.

Due to the imbalanced nature of the dataset, oversampling of the minority class was done by using the SMOTE() function from the smotefamily library on the training data. This was only implemented on the training data to prevent the introduction of bias to the distribution of the test set. The new train data has a total of 6572 rows, where there are 3316 rows for Class 0 and 3256 rows for Class 1.

L1 Regularised Logistic Regression model

The first model is a regularised variation of the logistic regression model to discover if the addition of regularisation can effectively reduce the large feature space and return good results. L1 was chosen over L2 due to the lasso's greater interpretability.

The cv.glmnet() function from the glmnet library was used for the fitting of the model, setting parameters `alpha=1` and `family="binomial"` to specify that it is a 2-class lasso regularised logistic regression model. No scaling was required since the function standardises variables by default. Also, this function uses cross-validation (cv) to tune the optimal value for the λ hyperparameter and the cv error returned for the different values of λ can be seen in Figure 2.

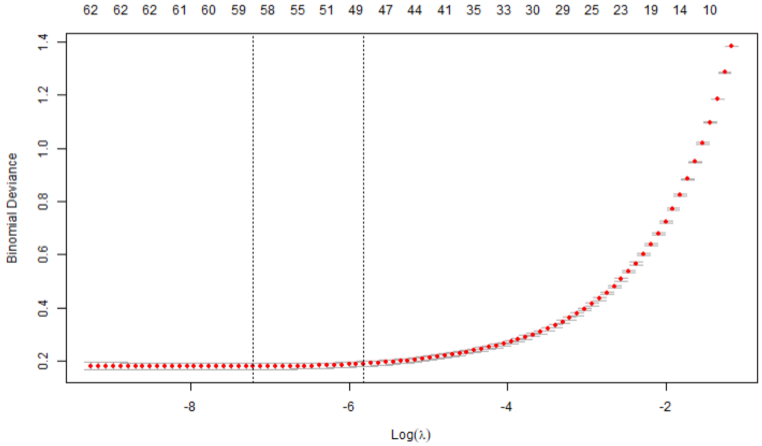
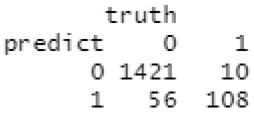


Figure 2: Cross-validation error for various log(λ) values

The optimal λ value returned was λ = 0.000740, which corresponds to log(λ) = -7.21. Using this value of λ, an L1 regularised logistic regression model was fitted, and its performance was evaluated on the test set. The resulting model suggested removing 3 variables: `salary`, `hasTroubleSleep` and `hasAwakening`. Setting the threshold to be 0.5, the predictions made on the test set can be seen in Figure 3 and the score obtained for the various metrics can be seen in Table 2.



```
        truth
predict    0    1
      0 1421   10
      1   56  108
```

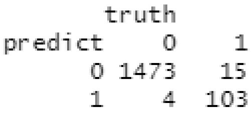
Figure 3: Confusion matrix

Accuracy	Precision	Recall	F1 score
0.959	0.659	0.915	0.766

Table 2: Test set performance

eXtreme Gradient Boosting model (XGBoost)

Since the precision score for the L1 Regularised Logistic Regression model was low as compared to the other metrics, this led to a hypothesis that perhaps there might be non-linearity present in the data that was not captured. Thus, an XGBoost model is proposed to better explain this non-linearity. Randomised search was used to find the best set of hyperparameters that has the lowest validation error within the large parameter space. Threshold of 0.5 was similarly set to obtain Figure 4 and Table 3.



```
        truth
predict    0    1
      0 1473   15
      1    4  103
```

Figure 4: Confusion matrix

Accuracy	Precision	Recall	F1 score
0.988	0.963	0.873	0.916

Table 3: Test set performance

Comparing Models

Between the 2 models, there is a decrease in Recall for the XGBoost model. This is justified by the huge increase in precision. Overall, the XGBoost model was chosen as the best model because of its high scores in other metrics and especially F1 score.

Feature Importance

From the XGBoost model, the feature importance can be plotted and analyse.

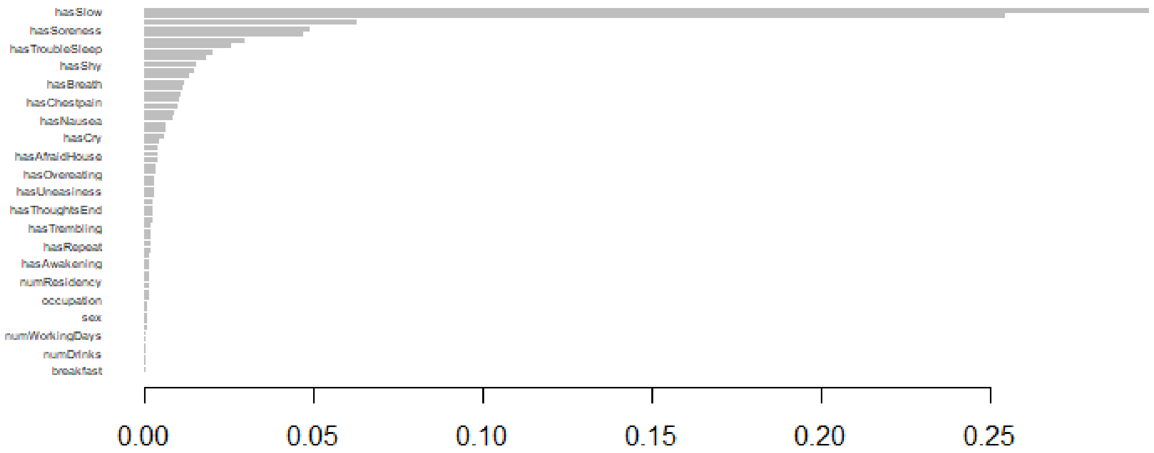


Figure 5: Feature Importance

From Figure 5, the top 3 features are `hasSlow` (Do things slowly to ensure correctness), `hasBlameOthers` (Feeling others are to blame for own's trouble), `hasPushed` (Feeling pushed to get things done). As compared to the psychometric test, most of the socio-economic indicators were ranked at the bottom of the feature importance ranking. This shows that their contribution the prediction of mental health is of a smaller impact.

Limitations & Conclusion

During the analysis, it was noted that there was loss of information in the `Salary` feature. It was converted to a categorical feature by the researchers, where the bins were scaled to cater at a lower level due to the low income received by the migrant workers. A stronger relationship between `Salary` & `healthStatus` could be derived if the `Salary` feature was continuous instead. As mental health is intangible (where it is harder to diagnose mental illnesses due the lack of clear symptoms that exhibit it), there is likely many mental health illnesses being misdiagnosed.[5] Thus, a highly accurate model would be able to make faster and accurate diagnosis of an individual's mental health status.

Despite the XGBoost model having stellar performance in all 4 metrics, this model is only applicable to migrant workers in Shanghai, where it was tailored towards a lower salary and tough working environment involving manual labour. Hence, a new point must be relatable to the dataset for the predictions to be accurate and sound.

However, this term paper can act as a starting point for the development of models to predict mental health status, where it can ultimately affect an individual's treatment method, medication, and treatment duration.

Appendix

[1] Centers for Disease Control and Prevention. (2021, June 28). *About mental health*. Centers for Disease Control and Prevention. Retrieved April 25, 2022, from https://www.cdc.gov/mentalhealth/learn/index.htm

[2] *Data from: Health-related lifestyle behaviors among male and female rural-to-urban migrant workers in Shanghai, China*. Dryad Data -- Health-related lifestyle behaviors among male and female rural-to-urban migrant workers in Shanghai, China. (n.d.). Retrieved April 25, 2022, from https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.61dg2

[3] *Symptom checklist 90-R - university of Pennsylvania*. (n.d.). Retrieved April 24, 2022, from https://dmu.trc.upenn.edu/dmumain/PDF_Files/scl.pdf

[4] World Health Organization. (2022, April 24). *Gender and Mental Health*. World Health Organization. Retrieved April 25, 2022, from https://www.euro.who.int/en/health-topics/health-determinants/gender/activities/gender-and-non-communicable-diseases/gender-and-mental-health

[5] Harrop, K. (2017, April 7). *How professionals diagnose mental health issues that have similar symptoms*. ATTN. Retrieved April 25, 2022, from https://archive.attn.com/stories/16231/why-mental-health-hard-diagnose