MIDTERM PROJECT REPORT

IDENTIFYING MENTAL HEALTH SUFFERERS IN THE TECH INDUSTRY

Haoran Zheng
Kevin Jiang
Austin Lutterbach

Mid-term Update Report: As of this mid-term report, we successfully preprocessed and encoded the raw data into a clean format we can work with. Furthermore, we constructed a suitable evaluation metric and a tree model to capture important features as well as a logistic regression model to get some preliminary results on our task. We finally propose future work we can pursue later this semester.

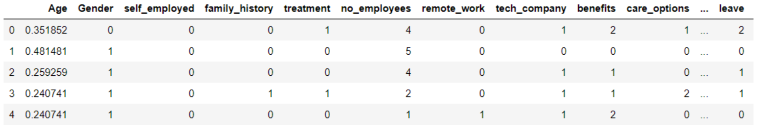
Original Dataset Description

The dataset is collected from a 2014 survey launched by Open Sourcing Mental Illness (OSMI) that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. [1] The dataset includes 1259 response records and 27 features in each record, describing respondents' demographic information, employment information and mental health status. A comprehensive introduction of 25 features can be seen in the [2] in the References section.

Data Preprocessing and Encoding

The purpose of the project is to build a classification model which can recognize people who suffer from mental health issues but do not seek treatment. Before using primary and advanced algorithms to build models, data preprocessing will be conducted to get dependent and independent variables.

For the independent variable, since no feature directly points out those who are suffering from mental issues but not seeking for help, features treatment and work_interfere would be used to get the independent variable. The definition of treatment is that "Have you sought treatment for a mental health condition". And the definition of work_interfere is that "If you have a mental health issue, do you feel that it interferes with your work?". There are five types of value within the work_interfere column, including "NA", "Rarely", "Sometimes", "Often", "Never". Therefore, respondents who are not reporting "NA" are actually the ones with mental health conditions. Through filtering the "NA" records, treatment columns can be served as independent variables directly. For the dependent variables, features country, comments, state and timestamp would be dropped because of data scarcity and model simplicity. All the other features left would serve as potential dependent variables. After initial data processing, the dataset includes 994 records and 23 features in each record. A sample of the encoded dataset which is ready for modeling is seen in Figure 1 .



	Age	Gender	self_employed	family_history	treatment	no_employees	remote_work	tech_company	benefits	care_options	...	leave
0	0.351852	0	0	0	1	4	0	1	2	1	...	2
1	0.481481	1	0	0	0	5	0	0	0	0	...	0
2	0.259259	1	0	0	0	4	0	1	1	0	...	1
3	0.240741	1	0	1	1	2	0	1	1	2	...	1
4	0.240741	1	0	0	0	1	1	1	2	0	...	0

Figure 1: Encoded Data

Data Visualization

We now provide a sample of the data visualizations conducted. Figure 2 details potential gender differences with care options (awareness of mental health care options provided by company). Figure 3 summarizes the age distribution between those who do and do not seek treatment. As we can see, we are likely to see gender play an important role in our analyses. Also, the average age appears to be more concentrated around late 20s to early 30s and is fairly comparable between treatment levels.

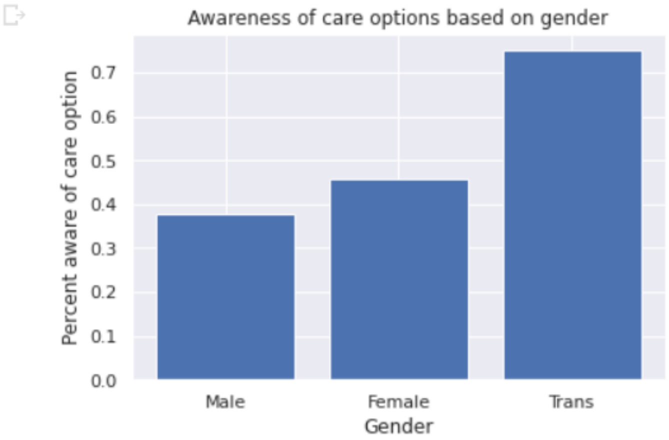


Figure 2: Care Option Seeking Among Genders

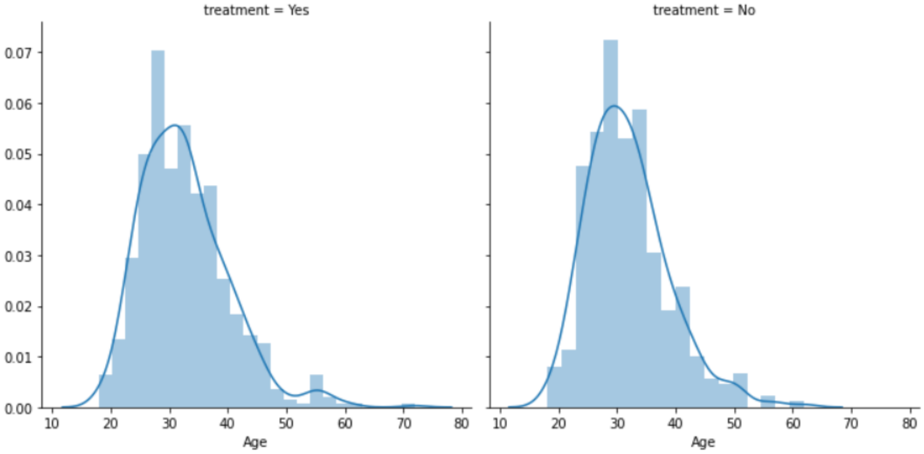


Figure 3: Distribution Among Ages Differentiated by Treatment

Evaluation Metrics

In order to evaluate the effectiveness of our models, we need an evaluation metric that is specifically designed to handle binary classification problems. To this effect, we used a 0-1 loss function. Furthermore, we used a confusion matrix to help understand our model's prediction

Preliminary Analysis

We built our current model using logistic regression and were able to predict correct classification with 70 percent accuracy. We also had roughly 70 percent precision as well as AUC score of 0.68. We also constructed a forest to highlight the defining features that can be see in Figure 4. We chose logistic regression since this is a classification problem and logistic loss function has nice properties like being fairly robust against outliers and is continuous and differentiable. A random forest was constructed as well since decision trees are more likely to overfit the data. We are sacrificing accuracy by doing random forest, however, we believe for our current purposes that 70 percent is sufficient.

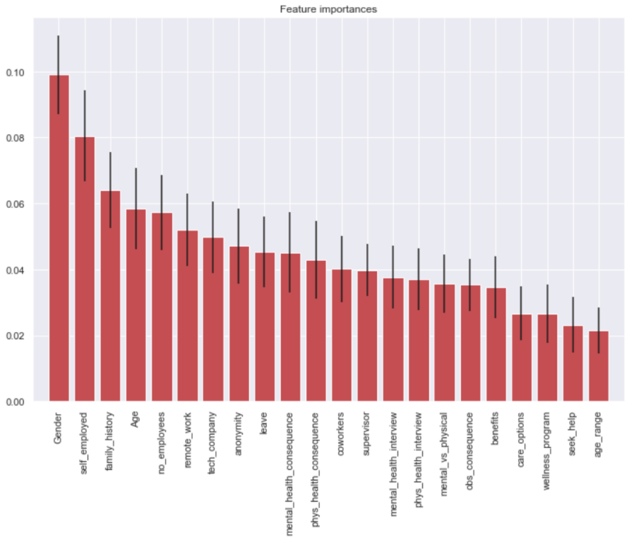


Figure 4: Random Forest

Future Work

We will ensure overfitting is not an issue by potentially reducing the amount of features we have and exploring some regularization techniques as well. Similarly, we will try other loss functions to produce different models and look into generalizing our model further.

REFERENCES

[1] Open Sourcing Mental Illness (OSMI) https://osmihelp.org/research

[2] https://www.kaggle.com/osmi/mental-health-in-tech-survey