Predictors of Mental Health Illness

Ronit Kumar
California State University, Long Beach

Abstract

This is a simple supervised learning model trained on OSMI Mental Health in Tech Survey 2017 and 2018 [2]. It measures attitudes towards mental health and probability of mental health disorders in the tech workplace.

The dataset has various fields exploring the conditions workers face everyday. This dataset is used to train various models and see how different factors effect a worker's well being. The models use these factors to predict how likely an individual is to seek medical attention for mental health issues.

1. Introduction

Understanding how circumstances influence mental health in a workplace is one of the goals in every organization. It has been seen that organizations benefit when workers feel happy, motivated and valued. When employees love their jobs, it shows in their work.

In this project, the open source OSMI dataset[2] is used to see how conditions at work influence the well being of an employee. What are the most important reasons that contribute to mental health? The survey has data on worker's age, gender, if they had problems with mental health conditions before, etc.

Finally, it is determined if factors not directly relating to place of work, like age and gender, also affect mental health.

The report is divided in sections that cover exploratory data analysis, data encoding, scaling, fitting and tuning. The purpose is to try to see if there is some relationship between different factors affecting mental health and if they can be used to successfully predict a mental health condition that may require further care.

2. Background

Mental illness is a severe issue that affects many individuals in the society. In this project we use the surveys provided by the non-profit organization Open Sourcing Mental Illness. The organization also works to create awareness about mental health and what can be done to help people who are having mental health issues.

The programs for this project have been run and tested on Spyder3. The project involves the use of several supervised learning techniques like KNeighborsClassifier, Decision Tree classifier, Random Forests, Bagging, Boosting and Stacking functions [1] .

3. Project

The project has been implemented on a system with the following specifications -
OS : Linux 5.1.14 kernel-current
IDE: Spyder 3.3.4
CPU: AMD Ryzen 5 2500U with Radeon Vega Mobile Gfx (8) @ 2.000GHz
RAM: 6930MiB
GPU: Radeon Vega Mobile Series

3.1. Design

The model design is constructed in several steps. The first of which is examining the dataset. The aim here is to determine the principal components that are affecting the results the most. Once an idea about what factors are influencing predictions the most has been ascertained, the observations are used to train the models. The models are also tested for precision accuracy.

3.2. Evaluation

The design of the project can be broadly divided into :

- Cleaning the dataset.
- Exploratory Data Analysis
- Training the model
- Results

The original dataset had many fields that could not be used, like subjective feedbacks on work environment, etc. These columns were dropped because they could not be effectively classified or labeled for training.

3.3. Cleaning the dataset

Variables like 'Network ID' were dropped because they were not really useful in gathering any meaningful information. Also, columns with missing data or nearly empty columns were dropped. Fig 2 shows the columns with the maximum amount of missing values. The columns of the dataset were checked for their datatype and processed through a label encoder.

3.4. Exploratory data analysis

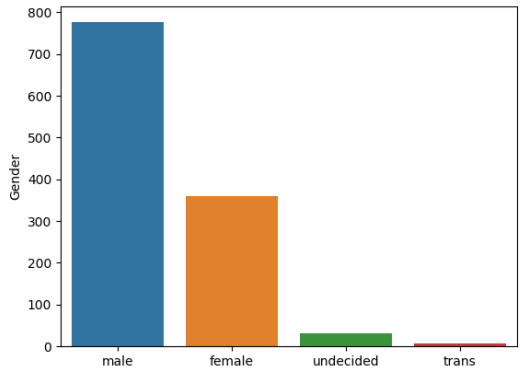

Figure 1. Gender distribution in dataset

Fig. 3 shows us some factors that positively effect mental health. The ability to share your concerns with co-workers or talk anonymously for help can go a long way in helping people with mental health issues.

The data has been scaled to age. The top features that show most covariance with actual results have been selected for further analysis.

3.5. Training the model

The model is tuned to parameters carefully chosen after testing with many values.The values of knn are tuned using multi-parameter tuning with weight-options 'uniform' or 'distance'.
The range for n-neighbors was 1 to 31.
It gives the following results :
Multiparam. Best Params: 'n-neighbors': 27, 'weights': 'distance'

OSMI Mental Health in Tech Survey 2017.csv
OSMI Mental Health in Tech Survey 2018.csv

	Total	Percent
Psychotic Disorder (Schizophrenia, Schizoaffect...	1173	100.000000
Obsessive-Compulsive Disorder	1173	100.000000
Anxiety Disorder (Generalized, Social, Phobia, ...	1173	100.000000
Attention Deficit Hyperactivity Disorder	1173	100.000000
Post-Traumatic Stress Disorder	1173	100.000000
Describe the circumstances of the supportive or...	1173	100.000000
Dissociative Disorder	1173	100.000000
Other	1173	100.000000
Addictive Disorder	1173	100.000000
Eating Disorder (Anorexia, Bulimia, etc)	1173	100.000000
Substance Use Disorder	1173	100.000000
Personality Disorder (Borderline, Antisocial, P...	1173	100.000000
Stress Response Syndromes	1173	100.000000
Psychotic Disorder (Schizophrenia, Schizoaffect...	1173	100.000000
Mood Disorder (Depression, Bipolar Disorder, etc)	1173	100.000000
Other.1	1168	99.573743
Eating Disorder (Anorexia, Bulimia, etc).1	1168	99.573743
Dissociative Disorder.1	1167	99.488491
Dissociative Disorder.2	1165	99.317988
Substance Use Disorder.1	1163	99.147485

Figure 2. Columns with missing values in percentage.

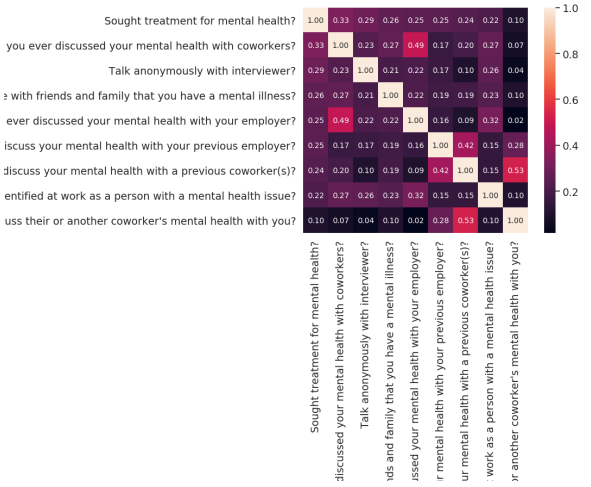

Figure 3. Covariance matrix with probabilities.

Further data analysis was performed by using various ensemble methods. The purpose here is to hit and try several techniques and see which provides the best results. 30% of the data has been reserved for cross validation.

The following functions were used to make predictions. As can be seen, not all methods perform with similar accuracy. The results obtained have been described below.

- **Logistic Regression**

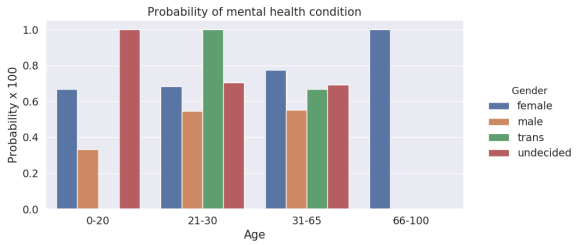

Figure 4. Probability of mental health conditions.

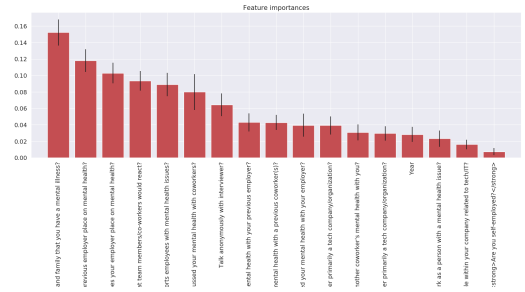

Figure 5. Most important features of dataset.

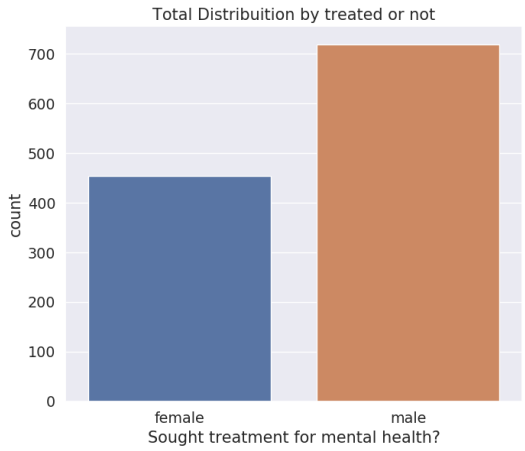

Figure 6. Sought Treatment - males and females.

Classification Accuracy: 0.7159090909090909
Classification Error: 0.28409090909090906
False Positive Rate: 0.34285714285714286
Precision: 0.7692307692307693
AUC Score: 0.7059299191374664
Cross-validated AUC: 0.7877543697504403

- **KNeighborsClassifier**

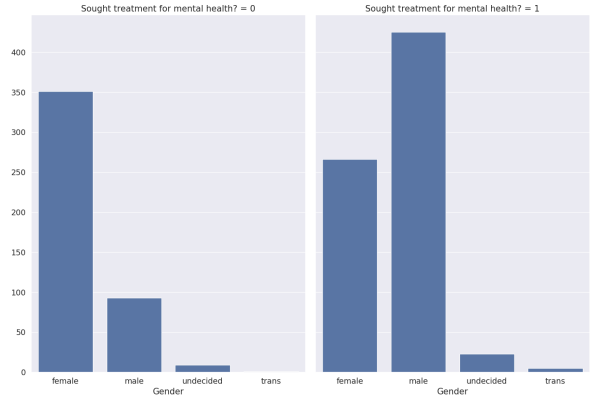

Figure 7. Sought treatment for mental health condition.

Classification Accuracy: 0.6278409090909091
Classification Error: 0.37215909090909094
False Positive Rate: 0.5142857142857142
Precision: 0.68
AUC Score: 0.6037061994609165
Cross-validated AUC: 0.69356350124364

- **Decision Tree classifier**
  Classification Accuracy: 0.6704545454545454
  Classification Error: 0.3295454545454546
  False Positive Rate: 0.34285714285714286
  Precision: 0.75
  AUC Score: 0.6681940700808624
  Cross-validated AUC: 0.7402981617563674

- **Random Forests**
  Classification Accuracy: 0.6704545454545454
  Classification Error: 0.3295454545454546
  False Positive Rate: 0.34285714285714286
  Precision: 0.7068965517241379
  AUC Score: 0.6439353099730458
  Cross-validated AUC: 0.7753109997883165

- **Bagging**
  Classification Accuracy:0.6647727272727273
  Classification Error: 0.3352272727272727
  False Positive Rate: 0.37857142857142856
  Precision: 0.735
  AUC Score: 0.6574123989218329
  Cross-validated AUC: 0.7120609175719912

- **Boosting**
  Classification Accuracy: 0.6960227272727273

Classification Error: 0.3039772727272727
False Positive Rate: 0.35714285714285715
Precision: 0.7560975609756098
AUC Score: 0.6869946091644205
Cross-validated AUC: 0.7787664243647607

- **Stacking**
  Classification Accuracy: 0.6818181818181818
  Classification Error: 0.31818181818181823
  False Positive Rate: 0.4142857142857143
  Precision: 0.7314814814814815
  AUC Score: 0.665498652291105
  Cross-validated AUC: 0.7438053873428438

4. Results

The system takes a few minutes to train the models and produce the results. Some further adjustments were made to the tuning parameters to use more features and try more cross-validation patterns. Although it improved the overall accuracy of some models, it did not provide any significant improvements in any model. The results achieved for the different models can be seen in Fig 8.

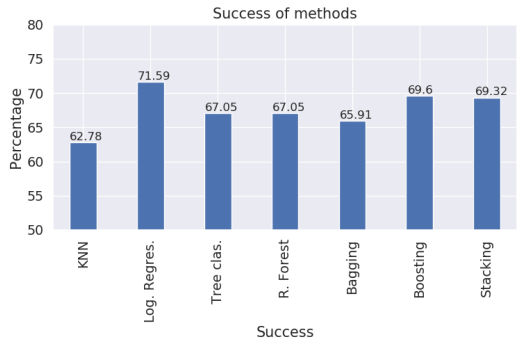

Figure 8. Prediction accuracy.

5. Summary

The OSMI datasets were imported and cleaned. Some exploratory data analysis was performed and the most important features affecting the results were chosen to build the models. This information was used to train the models to see if it can be successfully predicted that a person may require help with a mental health condition. This is a supervised learning model. Various ensemble and regression techniques were used to build the evaluation models. These models showed good accuracy. The best prediction accuracy reached was of 71.59% with logistic regression.

6. Conclusions

It was seen that tuning the parameters plays a very important role in how a model builds its predictions.

For the models used in this project, the Logistic regression gave the maximum prediction accuracy. Although an accuracy of about 72% is not considered sufficient. The future scope is to reach an accuracy of around 90%.

References

[1] https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/
[2] https://osmihelp.org/research