Mental Health in Tech Industry

CMPE 272 Group Project - **Team 4**

Alok Goyal
Software Engineering Department
San Jose State University
San Jose, USA
alok.goyal@sjsu.edu

Kunika Mittal
Software Engineering Department
San Jose State University
San Jose, USA
kunika.mittal@sjsu.edu

Pratyush Sharma
Software Engineering Department
San Jose State University
San Jose, USA
pratyush.sharma@sjsu.edu

Shalabh Neema
Software Engineering Department
San Jose State University
San Jose, USA
shalabh.neema@sjsu.edu

***Abstract*—Mental health is among the growing contributors to poor health and workplace accidents. Employees around the world are suffering from different mental health issues that are not being discussed and addressed by employers mostly. Employers need to develop a conducive environment in the workplace where mental health is considered as normal as the physical health and proper help is extended to the employees suffering from it. The aim of this project is to inform the employer about the state of mental health of their tech employees and provide them with actionable insights that can help them improve the work environment. Using the machine learning models, the surveys received from the tech employees in the company are analyzed to gauge their mental health status. The results of this project would not only help the employers but also benefits the employees by recognizing their mental needs.**

***Index Terms*— mental health in technology industry, machine learning in mental health, logistic regression, bagging, boosting, decision tree, random tree, knn, stacking, k means clustering**

I. INTRODUCTION

Mental health is the psychological wellness of a person. Messaging board app Blind recently conducted a survey of over 11,000 tech employees. 57% of the sample set reported significant feelings of stress and burnout as a result of their jobs.[1]

Eric Salvatierra, former CFO of Skype, VP at PayPal and one of the executives who helped build eBay from the scratch struggled with Bipolar Disorder and abruptly ended his life by stepping in front of the Caltrain in 2012. He could never talk about his health conditions with his co-workers [2]. Long working hours, high productivity and no sleep for long stretches are considered badge of honors in tech industry.

One of the biggest challenges in promoting mental wellness in the workplace stems from the fact mental illness is stigmatized [3]. Employers are not ready to take up the ugly conversation. HR executives just tip-toe around the issue by promoting social activities like team bonding exercises or weekly yoga classes to combat it but mental wellness is much beyond that. Direct resources need to be made available to people to help and support them.

According to a study, businesses around USA lose between $80 - $100 billion annually due to mental health and low productivity of the employees [4]. Depression is thought to count for up to 400 million lost work days annually [5]. The project is inclined to enlighten the employer with current mental state of their technical workers. Early detection of the mental health issues improve the chances of positive response to the medication and faster recovery. Measures such as mentor-mentee matching using machine learning technology are also provided to the employer. Proper supervision and support from the company can have a positive impact on the employee and accelerate his recovery.

II. DATA DESCRIPTION

Open Source Mental Illness (OSMI) is a non-profit organization that is working towards promoting awareness about the issues of mental illness and disorders in the tech industry. It aims to eradicate the stigma around mental health and find resources to help employees struggling with the issues.

The data for the project has been taken from the OSMI Mental Health in Tech survey 2016 [6]. The data set was trained on different machine learning algorithms to predict the probability of a person suffering with stress, anxiety and other mental disorders.

The OSMI Mental Health in Tech Survey 2016 consisted of a total of 1433 of responses. The survey is designed to cover all the professional and personal aspects of a given individual so that a complete picture of the environment the given personal dealt with can be understood comprehensively. The survey is taken by a wide range of tech employees working in various organizations across the globe. It also captures information about the work environment specific to a given country because they differ from country to country.

Data for the mentor mentee network has been taken from Kaggle. Kaggle is a crowd-sourced platform for the data scientists around the world [7]. It consists of datasets that are open for anyone to use. The data set that has been used for building the mentor mentee network has been contributed by the FSEV UK [8]. It is a data set of young people that explores their preferences, interests, fears and opinions about a wide spectrum of topics.

Data Cleaning

The original data set consisted of 1433 responses and 63 attributes. The data set was divided into train and test dataset so the accuracy of the models can be checked. 70% of the data was used for train the model and 30% was used to test the model. The data was then cleaned to fit into different machine learning models and a total of 8 attributes were considered for the design of the final model. These were the parameters that were carefully selected by the team to understand the mental well-being of the given individual.

III. FEATURE SELECTION

Feature Selection is the mechanism where you pick the features that most add to the function or value you are interested in automatically or manually. Feature based selection algorithm Pearson Correlation has been used for feature selection. The absolute value of the Pearson's correlation is checked between the target and numerical features in our dataset. The top n features based on this criterion are kept [9].

Figure 1: Correlation Matrix for Feature Selection

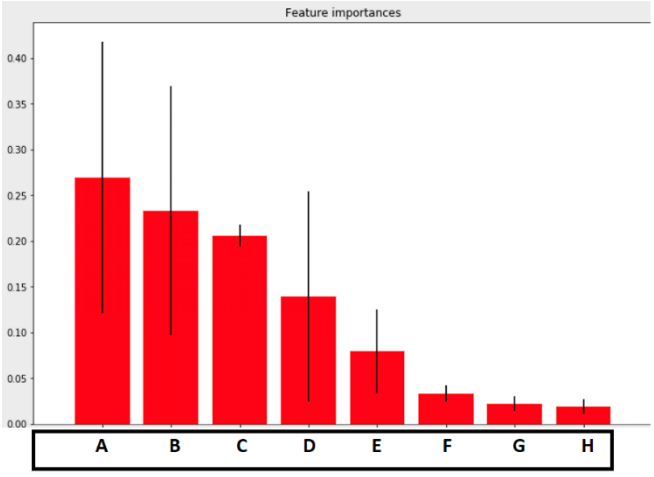

Figure 2: Important Features for predicting the medical assistance need of the employee

Fig Label	Feature Question
A	What is your age?
B	What is your gender?
C	Do you have a family history of mental illness ?
D	Have you had a mental disorder in the past?
E	Do you believe your productivity is ever affect ed by a mental health issue?
F	If yes what percentage of your work time (tim e performing primary or secondary job functi ons) is affected by a mental health issue?
G	If you have a mental health issue do you feel t hat it interferes with your work when being tr eated effectively?
H	If you have a mental health issue do you feel t hat it interferes with your work when NOT be ing treated effectively

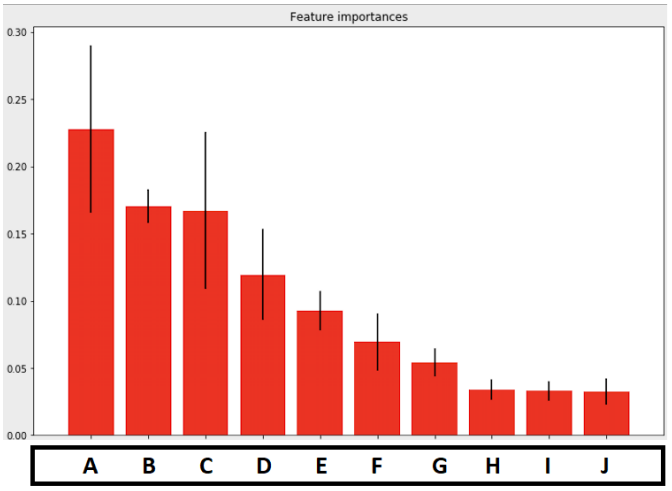

Figure 3: Important Features for predicting the comfort level of the employee discussing mental health issues with employer

Fig Label	Feature Question
A	What is your age?
B	What is your gender?
C	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?
D	If a mental health issue prompted you to requ est a medical leave from work asking for that leave would be?
E	Would you feel comfortable discussing a men tal health disorder with your coworkers?
F	Have you heard of or observed negative cons equences for co-workers who have been ope n about mental health issues in your workplac e?
G	Do you think that team members/co-workers would view you more negatively if they knew you suffered from a mental health issue?
H	If you have a mental health issue do you feel t hat it interferes with your work when NOT be ing treated effectively
I	Have you observed or experienced an unsupp ortive or badly handled response to a mental health issue in your current or previous work place?
J	Have your observations of how another indivi dual who discussed a mental health disorder made you less likely to reveal a mental health issue yourself in your current workplace?

IV. TECHNIQUES USED

The various machine learning techniques that were used for the data modeling as follow. The precision of each technique was

different. Post preliminary training, the data models were also tuned up to improve their accuracy.

1. Logistic Regression

Logistic Regression is a predictive analysis method and is used to describe data and explain the relationship between dependent and independent variables. The independent variables in our model were the attributes in the survey data and the dependent variable was the person's mental state, if he needed assistance or not.

2. KNN Classifier

K Nearest neighbor classifier is used in pattern recognition and statistical estimation. It stores the labelled data and classifies the new data on the basis of its previous data.

3. Decision Tree Classifier

Decision trees builds classification models in the form of a tree. The decision trees were used to find out the highest contributing factors so that due attention can be given to them.

4. Random Forest Classifier

Random forest is a supervised learning algorithm. It can be used for classification and regression problems. It builds multiple decision trees and merges them to get a stable and accurate prediction.

5. Bagging

Bagging is also called as Bootstrap Aggregation. It reduces variance and helps to avoid overfitting. It involves training the model on different subsets of data.

6. Boosting

Boosting is an ensemble method to improve the model predictions of any given learning algorithm. The main motive of boosting is to reduce the bias.

7. Stacking

Stacking is an ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta-regressor. The base level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model as features.

V. PERFORMANCE METRICS

Parameters considered to evaluate the accuracy of our trained models are as follow:

1. Cross Validation Score

Securing stability is a major aspect of creating a prediction model. We go for cross-validation in order to increase the same, in which we use a portion of the training samples as test data and then match our model.

2. Precision

The number of true positives divided by the number of true positives plus the number of false positives is defined as precision. False positives are cases where the model labels incorrectly as positive which are negative in fact [10].

3. Classification Accuracy

It is an indicator of the classification model's validity. Accuracy of classification is the percentage of successful predictions made out of total predictions. It can be used among different models as a performance rank.

4. False Positive Rate

It is basically a measure of how many instances a negative event was classified as positive by the model. In our case, it would mean the model classifying a person as needing help to cope up with stress and mental health although the person is perfectly alright. Lower the false positive rate, better the model.

5. AUC Score

AUC reflects the probability of putting a random positive example to the right of a random negative example. The value of AUC varies from 0 to 1. A model with 100 percent wrong predictions has an AUC of 0.0; one with 100 percent correct predictions has an AUC of 1.0.

VI. RESULTS AND ANALYSIS OF DIFFERENT METHODS

All the methods discussed previously have been implemented in Python to predict if a person needs mental health assistance or not. The results are visualized and tabulated as follow:

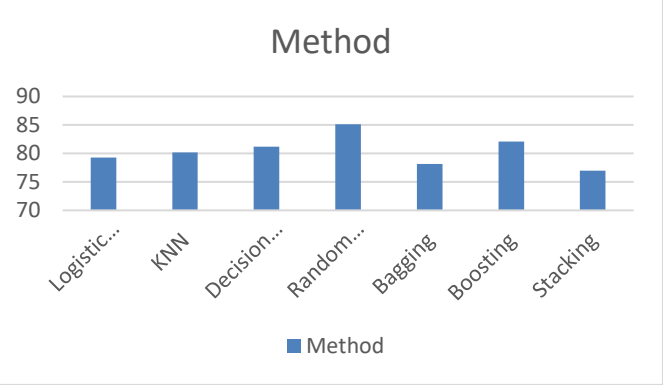

Figure 4: Accuracy Percentage for each method

It is evident from Figure 1 that *Random Forest Classifier* had the highest accuracy amongst all the other analysis algorithms. Boosting is also extremely close while stacking had the worst prediction rate.

VII. OUR SOLUTION

Post the analysis, the employer is suggested corrective measures to improve the workspace environment. The solutions suggested by us are as follow:

1. Mentor-Mentee Network

Mentors are supervisors in the company who are not directly in the management chain of the employee suffering with mental health issues. The mentors and the mentees are matched considering various factors such as technology skills, personality types of both the individuals. Depending on the case of the employee, the factors taken into consideration are given importance.

For example: If an employee is going through imposter syndrome, where the sufferer feels that he/she is not capable of doing the job and is not fit for the work assigned to him. In such cases, giving

priority to the technical skills of both the mentor and the mentee makes more sense. The mentor can help the employee by motivating him and this can boost the morale of the sufferer.

In other cases, the personality type of the mentor and the mentee is more important for both of them to work effectively together.

2. Medical Help Prediction

Based on the analysis of the employee survey, it would be predicted if the person needs professional medical help. Early detection of symptoms can help the person get medical help in time.

3. Daily Stress Indicator

Daily stress indicator of all the tech employees of the company will be shown on the dashboard of the employer. This will help the employer know the pattern when employees are most stressed out and need help. Actions such as workload distribution, meditation classes and other stress bursting activities can be planned around this time in advance.

4. Employee perception on mental health discussion

The employer will be shown the graph on his dashboard which estimates how many employees in the company are comfortable discussing their mental health issues with their managers or supervisors. This would help the employee take corrective measures to make a conducive work environment for his employees.

VIII. FUTURE SCOPE OF WORK

The project has a lot of scope for improvement and more solutions can be implemented. Following are the solutions that will be implemented in future to enhance the project.

1. 24X7 Helpline
2. Dog Therapy
3. One-on-one Counselling Sessions
4. Self Help Groups

IX. REFERENCES

1. Fagan Kaylee, “Employee Burnout Is Real. A Survey of More Than 11,000 Tech Workers Reveal Where It's Worst” published on Inc
2. Segall Laurie, “Silicon Valley’s Secret” on Mostly Human with Laurie Segall aired on CNN.
3. Kasbergen Nara, “Supporting Mental Health in Tech Workplace” published on InfoQ
4. Sime Carley, “The Cost Of Ignoring Mental Health In The Workplace”, published on Forbes
5. Sime Carley, “The Cost Of Ignoring Mental Health In The Workplace”, published on Forbes
6. OSMI Research: https://osmihelp.org/research
7. https://www.kaggle.com/
8. Young People Survey, https://www.kaggle.com/miroslavsabo/young-people-survey
9. Agarwal Rahul, “The 5 Feature Selection Algorithms every Data Scientist should know” published on Towards Data Science
10. Beyond Accuracy Precision: https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c