Mental Health Analysis in Tech

Members: Priyanka Mutha, Pallavi Chandrashekar, Mayank Phadke

Mental health issues continue to affect tens of millions of people each year. In 2016, the Substance Abuse and Mental Health Services Administration (SAMHSA) surveyed on drug usage and mental health, found that one in five adults suffers from a mental health issue, with a higher proportion of women (22%) suffering from it relative to men (15%).

The tech industry's foundation lies in the bright minds developing new solutions to create economic or social impact. This fast-paced industry has high stakes, which require people to meet even higher expectations. This is especially true for startup founders. There is this constant strain of "making it". Researchers from the University of California found that 72% of entrepreneurs surveyed self-reported mental health concerns and about 49% disclosed they deal with ADHD, bipolar disorder, addiction, depression or anxiety.

While data science is not widely used in the mental health industry currently, there are encouraging signs showing how data science could make a significant impact in the not too distant future.

Problem Statement:

Mental health in the technology industry is a known problem, with discussions over anxiety, depression and regular work pressure that is spread among employees in a wide range of sectors.
Mental health issues in the technology industry have grown to such an extent that people working in these fields are just as stressed as health service workers and are up to five times more depressed than the average worker.

Our main goal is to find out some of the main factors that affect the mental health of tech employees and provide suggestions to reduce their issues based on the results. We would also predict based on the main factors found, whether an employee is likely to reach out for help or not.

Some questions that aid this analysis can be:

1. How does the size of a company relate to an employer formally discussing mental health?
2. How does the age of an employee relate to their comfort in discussing mental health issues with their peers?
3. Which parts of the world are most affected?

4. Is the company’s attitude towards the health of an employee the main factor?
5. Does the employer provide mental health benefits?
6. Are the employees comfortable talking about mental health with their coworkers?
7. What kind of job has more mental health issues?
8. Does the employer provide help for mental health issues?

Data Sources:

The data has been sourced from Open Source Mental Illness (OSMI), which is a non-profit organization dedicated to raising awareness and educating about mental health in tech and open source communities. The data has been published on their website; please find the link below:
OSMI: Research

Primary Dataset: The 2014 dataset will be the main data set because it is the largest survey to date and thus, will give a bigger sample to work with. Results from the 2018 and 2019 surveys will be used to validate the findings.
The dataset comprises information about the participants pertaining to their workplace environment, their family history with mental wellness and as well as their own experience with it.
https://www.kaggle.com/osmi/mental-health-in-tech-survey

Secondary Dataset: By using a secondary dataset that describes the crime statistics in the USA, we will analyze whether the crime rates in the various states affect the mental health of the employees.
https://data.world/ucr/crime-in-us-2014-offenses

Project Outline:

The first step of the project would be to obtain the data from the source. Since datasets from three different surveys are being analyzed, it is vast and detailed enough for the analysis that is intended. An initial examination of the data will be done to understand the type of information that is collected and it’s meaning. This will help determine the type of data analysis that needs to be done. The next step would be to identify important facts, relationships, anomalies, and trends in the data. This will involve looking at the data in many ways, using a combination of data visualization, data analysis, and data mining methods.

Statistical Modeling Techniques:

Statistical Modeling techniques that can be used are:

1. **Logistic Regression**: Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.
2. **Decision Tree**: A decision tree can be used to visually and explicitly represent decisions and decision making.
3. **Clustering**: Task of grouping a set of objects in such a way that objects in the same group are similar.

Validation & Evaluation Technique:

K-fold cross-validation

The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias. It also gives insights on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set).

Confusion Matrix

Describes the complete performance of the model.
Since we are carrying out binary classification, the Confusion Matrix will be a matrix between actual and predicted Positives and Negatives.

KPIs:

We will be predicting whether an employee will seek treatment for their mental health issues

1. **Cost of False Positive**: Medium (If the model predicts that the employee may seek help, but in reality, the employee may not seek help)
2. **Cost of False Negative**: High (If the model predicts that the employee may not seek for help but in reality, the employee may seek help)

   Recall: What proportion of actual positives is correctly classified?
   True Positive/(True Positive + False Negative)

Precision: What proportion of predicted positives is truly positive?

True Positive/ (True Positive + False Positive)

F1 score: Used when we want a model with both good Precision and Recall.

2*[(Precision * Recall) / (Precision + Recall)

If by further investigation, it is found that our evaluation may need to give more weightage to either Precision or Recall, Weighted F1 score may b used for evaluation

Example:

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}.$$

This gives β times more weightage to precision than recall.

Success Metric: Considering the above example,

If F1 is used, the result will be a high value of F1

KPI	Threshold
Recall	0.85
Precision	0.90
F1 score	0.87

Deliverables:

The deliverables will be an interactive application, graphs, and technical reports
The model will be deployed in Shiny.

Resources:

- “Using Data Science to Help Tackle Mental Health Issues.” *DiscoverDataScience.org*, www.discoverdatascience.org/social-good/mental-health/.
- “About OSMI:: Open Sourcing Mental Illness - Changing How We Talk about Mental Health in the Tech Community - Stronger Than Fear.” *OSMI Home*, osmihelp.org/about/about-osmi.
- Myatt, Glenn J., and Wayne P. Johnson. *Making Sense of Data I: a Practical Guide to Exploratory Data Analysis and Data Mining*. John Wiley & Sons, Inc., 2014.
- “Getting Honest About Mental Health In The World Of Tech Startups.” https://www.forbes.com/sites/forbestechcouncil/2018/08/08/getting-honest-about-mental-health-in-the-world-of-tech-startups/#39c5c203641a