****

**Mental Health in Technology Industry**

Casey Patel

**1) Introduction**

What is Mental Health Illness?

Mental health is the foundation for emotions, thinking, communication, learning, resilience, hope and self-esteem. Mental health is also key to relationships, personal and emotional well-being and contributing to community or society.

According to Mental Health America, over half (56%) of adults with a mental illness in the US don't receive treatment, which amounts to over 27 million people.

How does it affect businesses?

Employees experiencing mental health issues may struggle with concentration, decision-making, and completing tasks efficiently.

Decreased productivity can result from increased absenteeism (taking sick leave) and presenteeism (being physically present but mentally unwell and unproductive).

Companies that neglect mental health can suffer reputational damage, affecting their attractiveness to potential employees and customers.

**2)** **Problem Statement:**

Mental health issues among employees represent a critical challenge for businesses, affecting productivity, employee retention, and overall organizational effectiveness. Despite increased awareness, many companies struggle to adequately address the mental health needs of their workforce, leading to significant operational and financial repercussions. The goal of this project is to classify employees who have sought treatment for mental illnesses. Using these predictions and insights from the data we want to help our clients determine creation of programs and awareness campaigns to promote a healthier work culture, boost productivity and retain talent.

**3) Stakeholders:**

The client for a mental health business-related issue could be any organization or company that is concerned about the well-being of their employees and wants to improve their work environment. This can include Corporate HR Departments, Healthcare Providers, Insurance Companies, Government Agencies and Non-profit Organizations.

**4) Dataset:**

The dataset is obtained from Kaggle called [Mental Health in Tech Survey](https://d.docs.live.net/230e9210bcc2194a/Documents/Capstone%202-%20Project%20Proposal.docx). This dataset is from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. There are a total of 27 features, 22 categorical and 5 numerical. The key features include age, gender, employee type, benefits, care options, work interference, seek help, coworkers, family history, wellness programs etc.

**5) Data Wrangling:**

Upon examining the imported CSV file, several data inconsistencies were identified. The gender field included diverse and non-traditional responses along with some blanks. The self-employment field had a few missing entries, while the work interference field had numerous missing values. The comments section was entirely free-form with more than half of the entries missing. The timestamp data showed that all entries were recorded on the same day in 2014. The age field had some implausible values, including negative numbers and ages over 120.

Dealing with Incomplete Data

To address the incomplete data, I used the replace function with regex to convert male and female entries to 0 and 1, respectively. All NaN and open-ended responses in the gender column were randomly replaced with 0 or 1. NaN values in the self-employed column were replaced with "No." For the work interference column, a new category, "Don't Know," was created to replace NaN values. Age values that were less than 18 or greater than 75 were replaced with the median age value.

Finalizing the Data

I removed the comments, state, and timestamp columns from the final dataframe. The label encoder was applied to all categorical features to convert them into numerical values for analysis. Additionally, an age range column was created, categorizing respondents' ages into four numerical age ranges.

**6) Exploratory Data Analysis (EDA):**

We examined the distribution of the target variable, "treatment." A value of 1 indicates that the respondent has sought treatment, while a value of 0 indicates they have not. The distribution between these two categories is well-balanced, providing us with an evenly distributed target variable.

A blue and orange rectangular bars

Description automatically generated

We will examine the number of survey respondents based on gender. The data reveals a significant disparity, with males comprising nearly 80% of the respondents and females only 20%. This imbalance reflects the uneven gender representation in the current tech industry.

A blue and orange rectangular bar graph

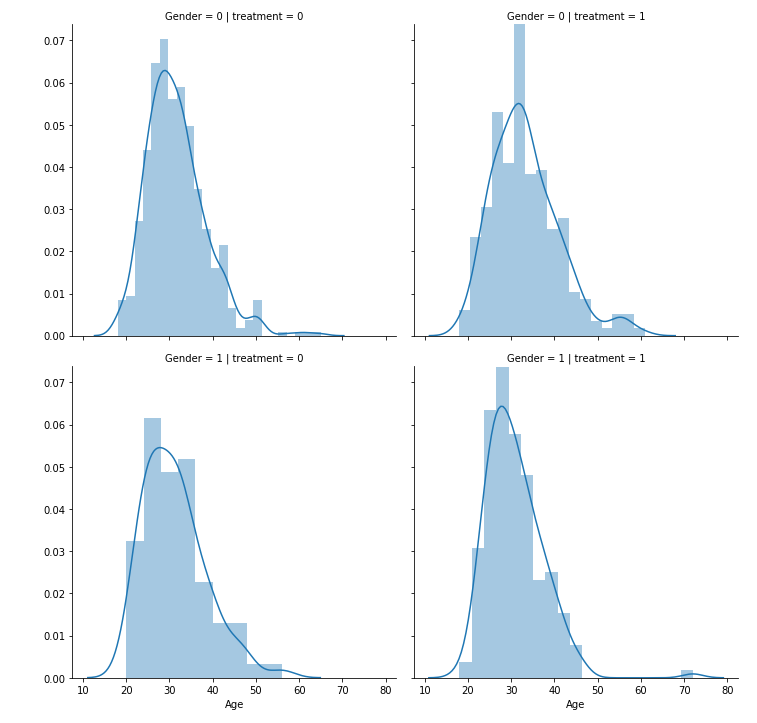
Description automatically generated

Next, we'll assess the age distribution of the survey respondents to see if the participants represent a wide range of ages or are clustered within a particular age group. The Empirical Cumulative Distribution Function (ECDF) plot indicates that the age data is normally distributed, with just one or two outliers over the age of 70.

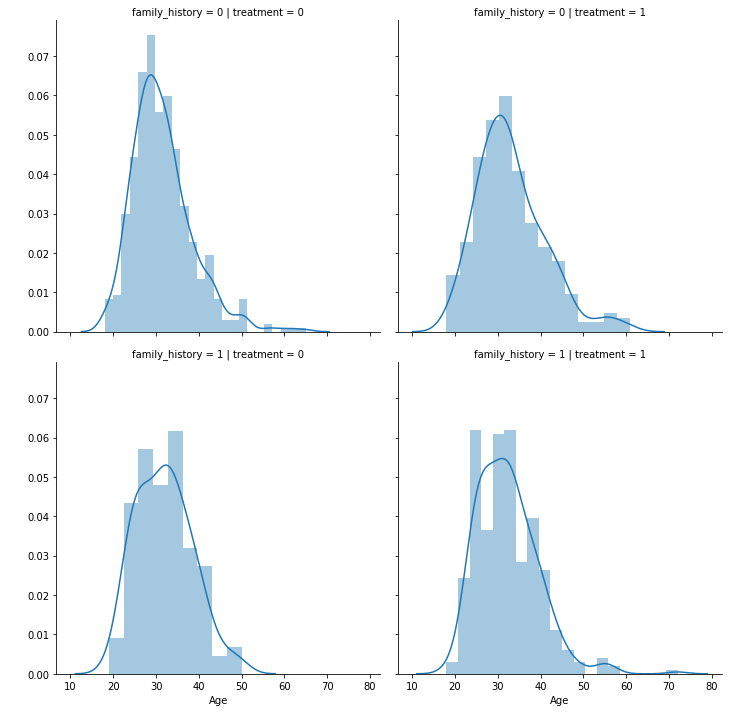
A graph of age and age

Description automatically generated

I also examined the normality of age distribution based on gender and treatment categories. This analysis aims to uncover any noteworthy patterns or outliers related to how gender and treatment affect age groups. The findings indicate that all four scenarios exhibit a normal distribution, prompting further exploration of the data to uncover insights. In this analysis, a gender value of 0 represents males, and 1 represents females.



The investigation was repeated to assess the normality of the distribution of family history of mental illness and its relationship with treatment across different age groups. This step aimed to identify any anomalies and understand the impact of family history on treatment-seeking behavior. The results indicated that the distributions were normal, suggesting the need for additional exploration to uncover insights into this relationship.



**7) Statistical Testing:**

After converting all dataset features to numeric variables using a label encoder, I applied a chi-square test against the target variable, setting the alpha threshold at 0.05. The chi-square test revealed that the following variables are statistically significant.

|  |  |
| --- | --- |
| Gender | Wellness programs |
| Family history | Seek help |
| Mental health interview | Anonymity |
| Country | Leave |
| Work interference | Coworkers |
| Benefits | Mental vs physical |
| Care options | Obs consequences |

Next, I conducted a two-sample Z test on family history versus treatment. The results clearly indicate that family history influences employees' likelihood of seeking treatment, as the proportions between the groups are significantly different.

**Family History vs. Treatment**

**Null Hypothesis:** The proportion of individuals with a family history of mental illness seeking treatment is the same as those without a family history seeking treatment.

**Alternate Hypothesis:** The proportion of individuals with a family history of mental illness seeking treatment is different from those without a family history seeking treatment.

P-value: ~0, leading us to reject the null hypothesis.

The results of the two-sample Z test demonstrate that family history significantly affects whether employees seek treatment, as the proportions are markedly different.

**Gender vs. Treatment**

**Null Hypothesis:** The proportion of males seeking treatment is the same as the proportion of females seeking treatment.

**Alternate Hypothesis:** The proportion of males seeking treatment is different from the proportion of females seeking treatment.

P-value: ~0, leading us to reject the null hypothesis.

The two-sample Z test results reveal a significant difference in the proportions of males and females seeking treatment, indicating that gender significantly impacts treatment-seeking behavior.

**8) Data Story:**

Recognizing the distinct differences in how genders approach treatment, we can delve deeper into the probabilities across various age ranges. By plotting the treatments sought by gender within each age range, we uncover some intriguing insights. Notably, females in the youngest age range sought treatment at nearly twice the proportion of males. Additionally, females across all age ranges sought treatment more frequently than males. This pattern may indicate that females are more likely to experience mental health issues, seek treatment more often than their male counterparts, or a combination of both factors.

A graph of a number of people with mental health condition

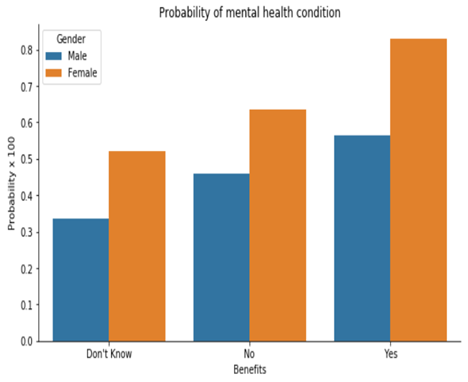
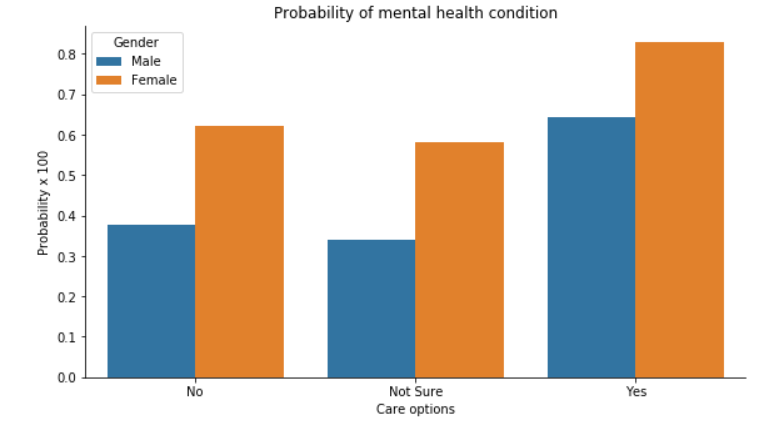
Description automatically generated

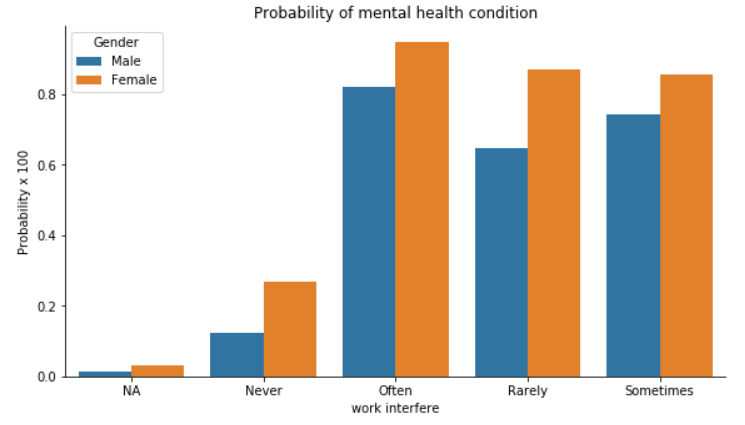
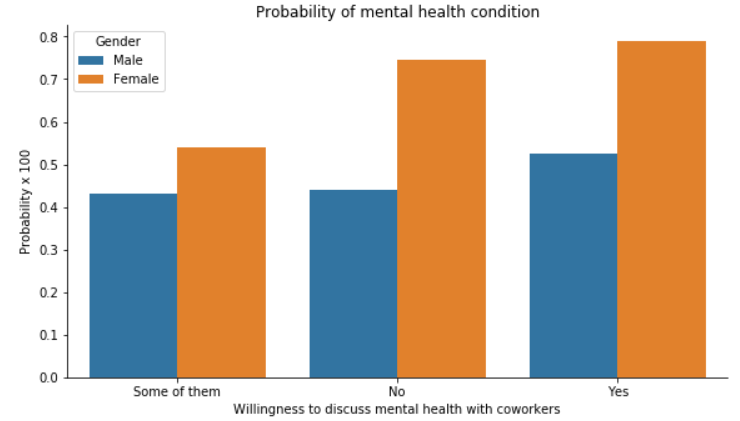
Next, let's explore the role of family history. We find that the proportion of males and females seeking treatment with a family history of mental illness differs by only 10%. However, there is a significant difference in the proportion of females seeking treatment without a family history. This does not necessarily mean that fewer males suffer from mental illness. Instead, it highlights the substantial stigma males face in admitting to mental health issues and seeking treatment. In contrast, females are generally more open about mental health and approach treatment without the same level of stigma.

A graph of a person and person

Description automatically generated

Continuing our analysis, we notice a consistent trend: females are more likely to seek treatment compared to males across all categories. This disparity becomes particularly pronounced when examining factors such as awareness of care options, benefits, willingness to discuss their illness with colleagues, and how work impacts their ability to manage their illness. These findings strongly indicate that while males may be aware of treatment options, they are less inclined to pursue them due to the stigma surrounding open discussion and acknowledgment of mental health challenges.



**9) Machine Learning: Predicting Models:**

Identifying employees who seek treatment for mental health issues is pivotal for improving productivity and retention strategies. Utilizing the 22 other features in our dataset, we aim to accomplish this objective. Given that the task involves predicting whether treatment is sought (binary outcome of 1 or 0), it falls into the realm of binary classification. Firstly, I created a numpy array y to represent the target variable for treatment and an array X comprising the remaining features. Subsequently, I divided X and y into training and testing datasets, adhering to a 70%-30% split.

Support Vector Machine

I began by creating and applying a base Support Vector Machine (SVM) model to the training data, using default parameters. Following this, I evaluated the model's performance metrics, including accuracy, confusion matrix, classification report, and ROC curve.

After establishing the baseline results, I proceeded to improve the model by tuning its hyperparameters. Utilizing the GridSearchCV method, I explored a wide range of values for the C and gamma parameters. Subsequently, I evaluated the model's performance using the best hyperparameter values obtained from the tuning process.

Surprisingly, the evaluation indicated that the base model performed equivalently to the tuned model. In this specific case, hyperparameter tuning did not yield improved results for our model.

|  |  |  |
| --- | --- | --- |
|  | Base Model | Tuned Model |
| Accuracy | 83% | 83% |
| Precision | 83% | 83% |
| Recall | 83% | 83% |
| F-1 score | 83% | 83% |

Random Forest Classifier

I initially developed and trained a Random Forest Classifier on the training data using default parameters. To evaluate the model's performance, I assessed metrics such as accuracy, confusion matrix, classification report, and ROC curve on the test data. These results established a baseline for comparison.

Next, I aimed to enhance the model's performance by tuning its hyperparameters. Utilizing the GridSearchCV method, I explored a wide range of values for parameters including n\_estimators, max\_depth, bootstrap, and max\_features. After identifying the best hyperparameter values, I trained a new model using these optimized settings.

Upon evaluation, the tuned model demonstrated improved performance compared to the baseline model, confirming that hyperparameter tuning effectively enhanced the model's accuracy and overall performance.

|  |  |  |
| --- | --- | --- |
|  | Base Model | Tuned Model |
| Accuracy | 83% | 83% |
| Precision | 83% | 84% |
| Recall | 83% | 83% |
| F-1 score | 83% | 83% |

The Random Forest Classifier has provided the best results across all categories, so we will examine its performance more closely. Below is the ROC curve for the model. The goal of the ROC curve is to maximize the area under the curve (AUC). In our case, the AUC is 83%, indicating that the model can accurately predict outcomes 83% of the time with a new set of data, which is significantly better than the 50% accuracy of random guessing.

Additionally, we analyzed the feature importance plot, which highlights the contribution of each feature to the model's predictions. The most prominent feature is "work interference," suggesting that whether an employee's mental illness affects their work significantly influences the prediction. This insight can help our clients identify employees whose mental health issues impact their work but who may not be seeking treatment.

Other important features include family history, care options, gender, and leave, which are expected to influence mental health. Surprisingly, features such as country and the number of employees also play a role. This information can guide our clients in understanding and addressing the factors affecting their employees' mental health.

A graph of a curve

Description automatically generated

A graph with blue and white text

Description automatically generated

**Conclusion:**

Random Forest Classifier has a slightly higher accuracy (0.84) compared to the second model SVM (0.83). For both classes (0 and 1), the precision(0.88, 0.80), recall(0.80, 0.88), and F1-score(0.84 , 0.84) are marginally higher in the first report(Random Forest Classifier) compared to the SVM report. Also, its macro and weighted averages are consistently higher.

Conduct regular anonymous surveys and request permission from employees to use their responses for modeling purposes. Utilize the Random Forest Classifier model to predict whether an employee has sought treatment for mental illness.

Continuously train the model with new data to enhance its accuracy.

**Awareness Programs:** Highlighting the Issue and Importance of Treatment: Raise awareness about the widespread nature of mental health issues and the critical importance of seeking treatment.

**Reducing Stigma in the Workplace:** Develop initiatives to remove the stigma associated with discussing mental health at work.

**Promoting a Healthy Lifestyle:** Encourage participation in existing programs that support a healthy lifestyle, such as yoga classes and fitness subsidies.

**Mental Health Benefits and Care Options:** Inform employees about existing mental health benefits and care options available to them.

**Gender Gap in Treatment Seeking:** Focus on bridging the gap in seeking treatment between genders.

* Data Collection and Impact Assessment

**Happiness and Productivity Index:** Collect data on employee happiness and productivity as part of the survey to assess the impact of these programs.

**Monitoring Work Interference:** Track responses in the work interference column to see if the proportion of "NA" and "Never" increases over time while "Often," "Rarely," and "Sometimes" decrease. This would indicate that employees are either recovering from their illnesses or receiving effective treatment that minimizes work interference.

* Goal

Fostering a happy and healthy workforce will likely lead to happier and healthier lives outside of work as well.