2019

Mental Health in Technology Industry

PREDICTING EMPLOYEE TREATMENT NEED TO IMPROVE PRODUCTIVITY AND RETENTION

NILAY DESAI

What is Mental Illness?

- Mental illnesses are health conditions involving changes in emotion, thinking or behavior (or a combination of these)
- In any given year, 1 in 5 adults in the U.S. experience mental illness, but only 41% receive any type of mental health service
- It can affect anyone regardless of age, gender, geography, income, social status, race/ethnicity, religion/spirituality, sexual orientation, background or other aspect of cultural identity

How does it affect businesses?

- Serious mental illness cost the U.S over $190 billion in lost earning every year
- Depression alone is estimated to result in over 200 million lost workdays yearly
- World Health Organization projects the world would lose 12 billion workdays to depression and anxiety disorders by 2030
- Affected employees showed lack of loyalty and productivity by more than 10% from their happier and healthier counterparts

Problem Statement:

The stigma attached to mental health raises concerns among employees for being judged or held back professionally if they discuss their illness openly. 56.4% of adults with a mental illness received no treatment due to lack of insurance coverage and lack of awareness about the benefits and care options provided by their employers. 90% of employees said that struggling with mental health issues stops them from thriving at work and performing to the best of their ability. 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.

Dataset:

The data set was obtained from Kaggle under the title of "Mental Health in Tech Survey". This survey was conducted by Open Source Mental Illness (OSMI) in 2014 and it was the largest of its kind in the tech industry with almost 1300 participants. This survey contains employee attitudes and frequency of mental health treatment sought by employees. 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.

Data Wrangling:

Data Exploration

The csv file was imported into a pandas dataframe first. Upon exploration, the gender category contained non binary & open ended responses along with nan values. Self-employed category contained a small number of nan's. Work interference category a sizeable number of nan's. The comments category was completely open ended with more than 50% nan values. Timestamp category values were all from the same day in 2014. Age category contained negative values and values ranging over 120.

Dealing with incomplete data

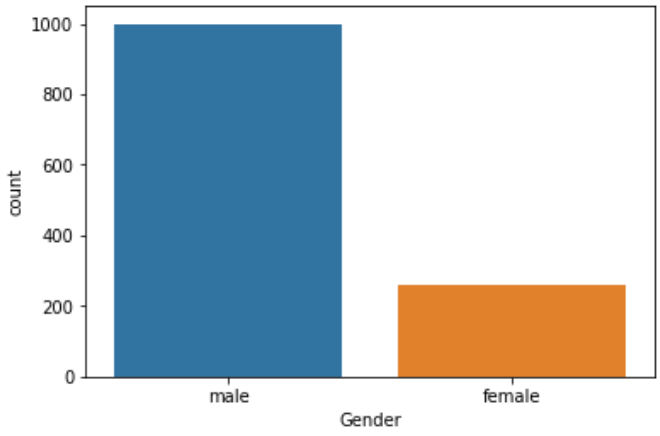
I used the replace function along with regex to convert male and female values to 0 and 1. I replaced all nan and open ended values with a random values of 0 or 1. Replaced all Self-employed nan values to the No category. Created a new category of Don't Know created for Work Interfere nan's. Replaced all age values less than 18 and greater than 75 to the median value of the feature

Finalizing the data

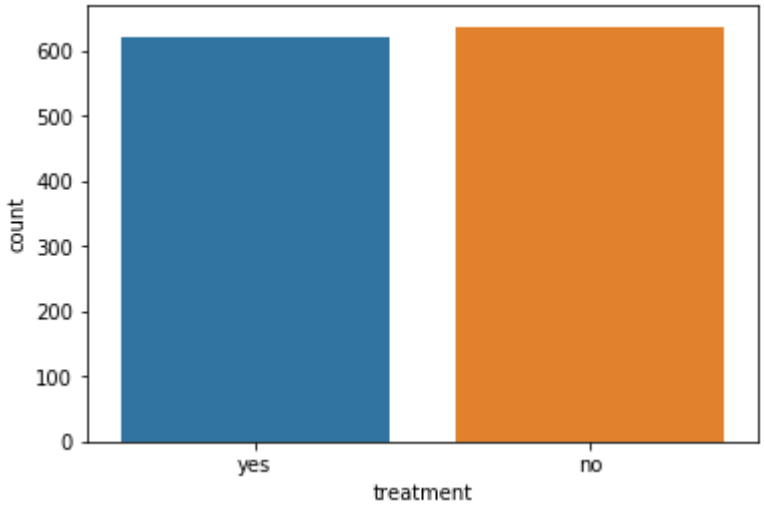
I dropped the comments, state and timestamp features from the final data frame. I also applied the label encoder method on all categorical features to convert them to numerical variables for analysis. Created new Age range column to view the ranges of the ages of the survey respondents. Age range contains 4 category of age ranges denoted by numerical values

Exploratory Data Analysis (EDA):

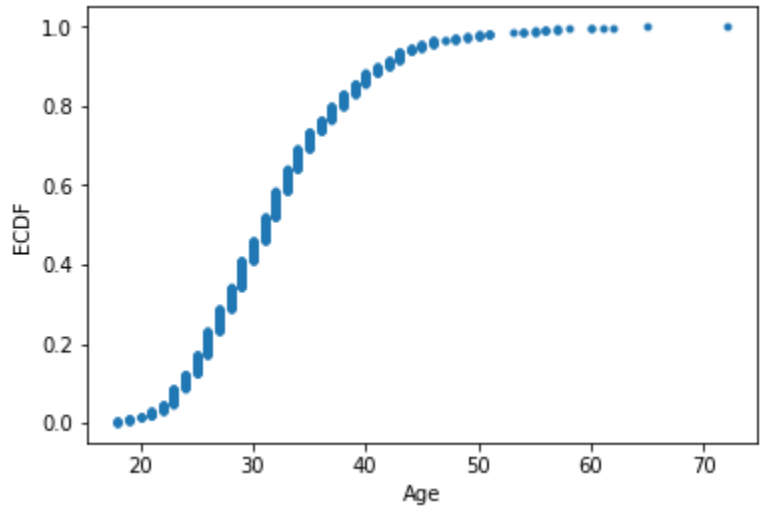
To start out we will explore the number of respondents to the survey based on gender. Below we can clearly see a huge discrepancy between the male and female respondent. Males comprise nearly 80% whereas females comprise only 20%. This is representative of how uneven the representation of gender is in the current tech industry.



Next we will take a look at the distribution of our target variable called 'treatment'. Treatment with a value of 1 signifies whether the respondent has sought treatment and a value of 0 signifies whether they haven't. Below we can see a nice and equal distribution between the two categories giving us a balanced target variable.

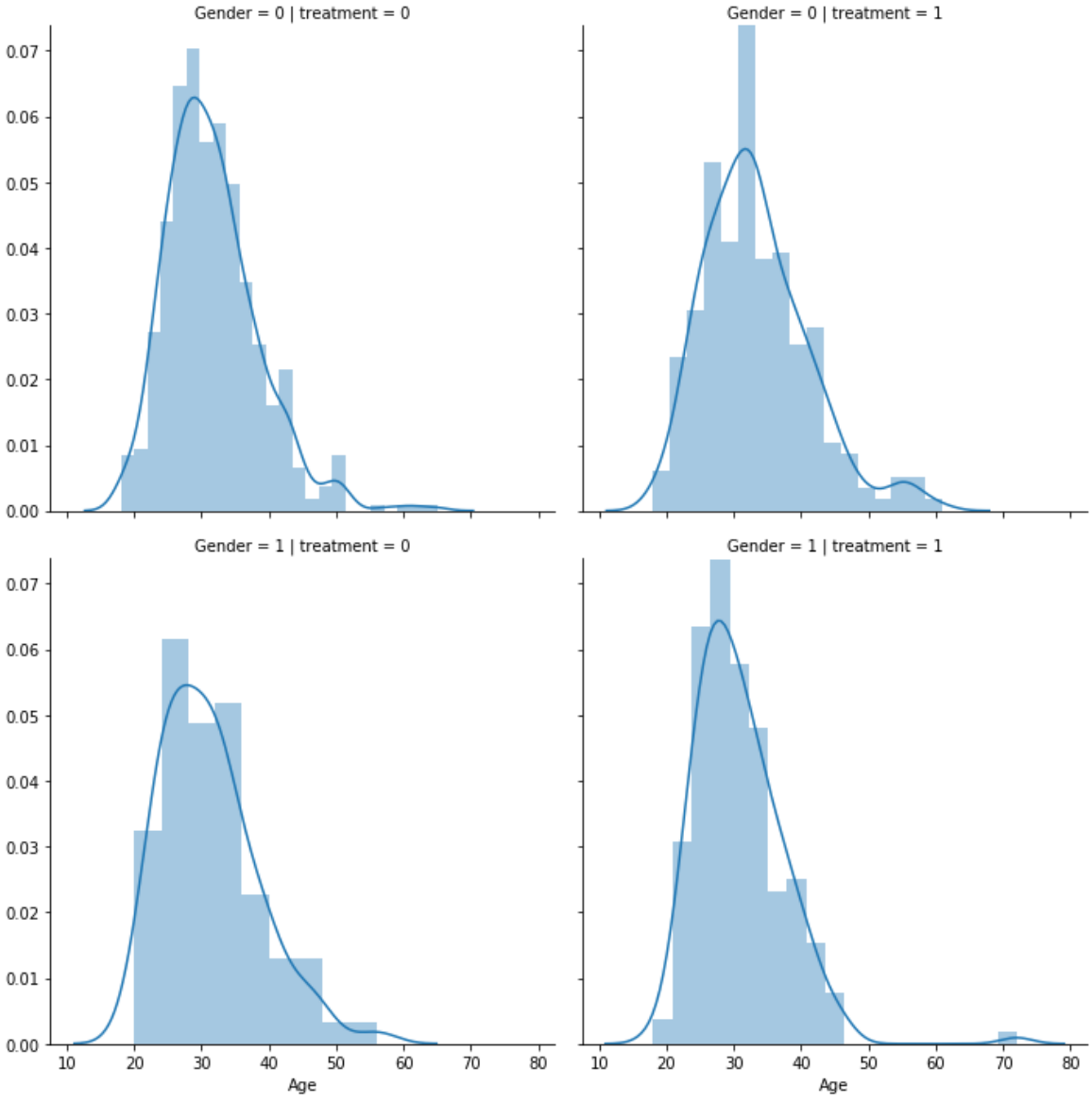


We will now explore the normality of the distribution of the survey respondents based on their age. We want to explore whether employees of a wide range of ages have responded or are they bunched among a certain age group. Based on the Empirical Cumulative Distribution Function (ECDF) plot we learn that the data is indeed normal with 1 or 2 outliers aged over 70 years.

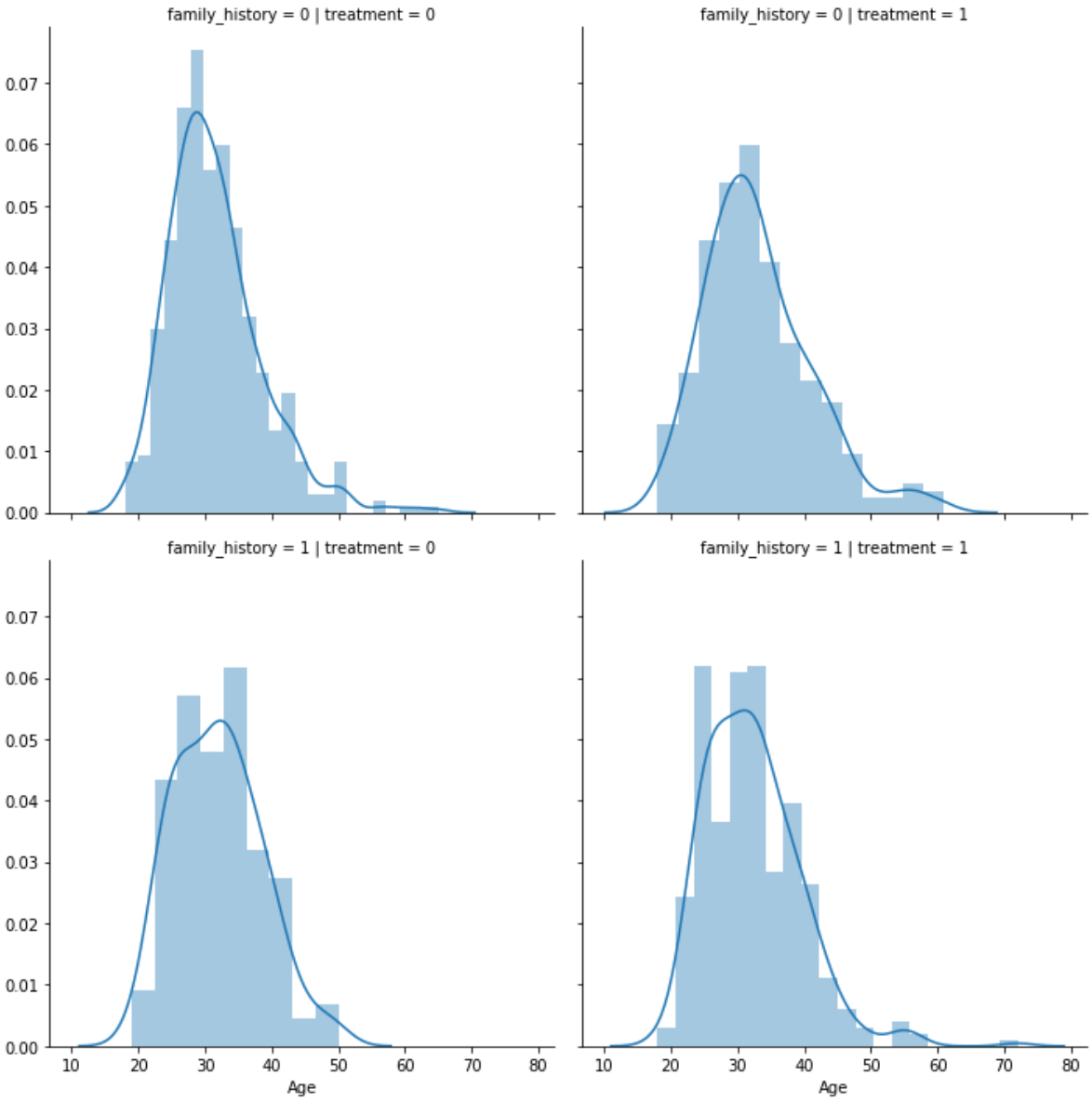


Another distribution I wanted to test for normality was the split on the basis of gender and treatment over age. This was done to see if there are any interesting

discoveries, outliers etc when we look at gender and role treatment plays over their age. Below we find that all four scenarios are indeed normal which means we will need to further investigate our data to find insights. Gender value of 0 corresponds to males and 1 to females.



We repeat the step above to investigate normality of distribution of family history of mental illness and treatment over age to see the irregularities and the role family history plays for seeking treatment. Again we find that the distributions are normal and we will require further exploration.



Performing Statistical tests on the data

Since all the features in the dataset are turned to numeric variables using label encoder I performed the chi-square test against the target variable where the threshold for my alpha value was .05. After performing the Chi-square test the following variables were found to be 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 performed a two-sample Z test on family history vs treatment. This clearly shows us that family history plays some role in employees seeking treatment as the proportion is not the same.

Null hypothesis - The proportion of people with family history seeking treatment is the same as those without family history seeking treatment
Alternate hypothesis - The proportion of people with family history seeking treatment is not the same as those without family history seeking treatment
P-value ~ 0 therefore we will reject Null hypothesis

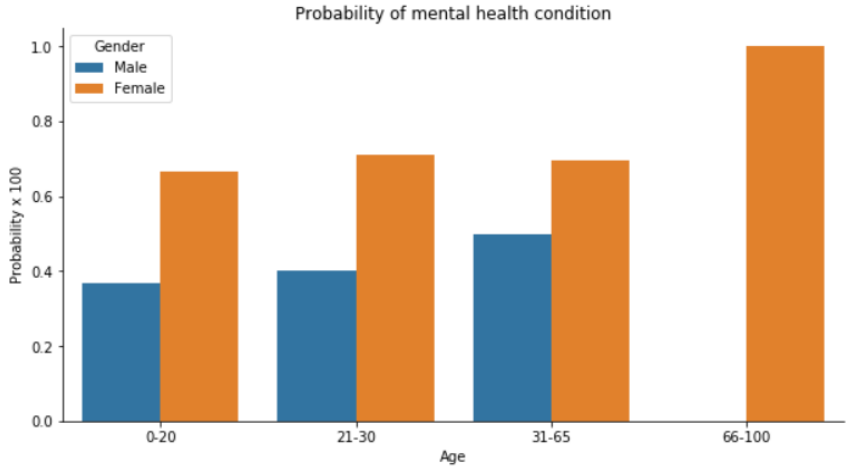
I performed a similar test on gender vs treatment to determine the role of gender and whether both gender seek the treatment at the same proportion. By rejecting our null hypothesis we can see that the treatment is not the same and there is a clear difference in the proportion of treatments sought out among the genders.

Null hypothesis - The proportion of males seeking treatment is the same as females seeking treatment
Alternate hypothesis - The proportion of males seeking treatment is not the same as females seeking treatment
P-value ~ 0 therefore reject Null hypothesis

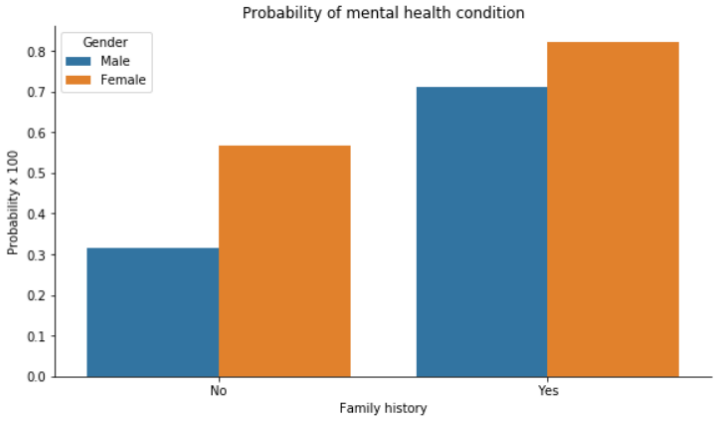
Data Story:

Having seen that there's a clear difference in how the genders approach treatment let's take a closer look into how the probabilities look like among the different age ranges.

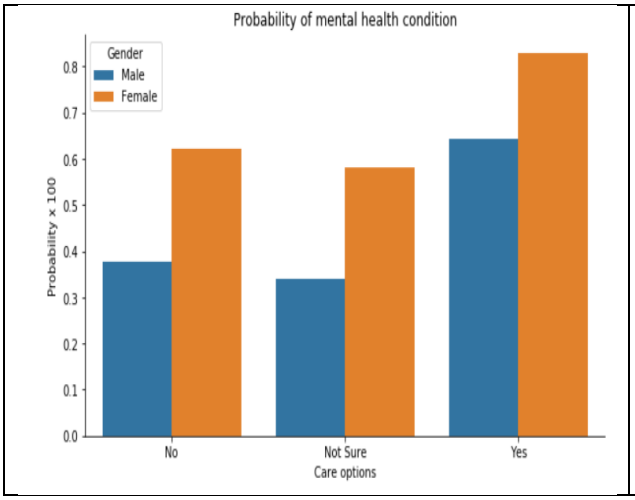
To start out let's plot the treatments sought out split between genders over the age ranges. This reveals a few interesting insights to us. Females in the first age range sought treatment at almost double the proportion. Females among all ages sought treatment at a higher proportion than males. This might indicate to us that females are more likely to suffer from mental health issues or they seek treatment more than their male counterparts or a combination or both.

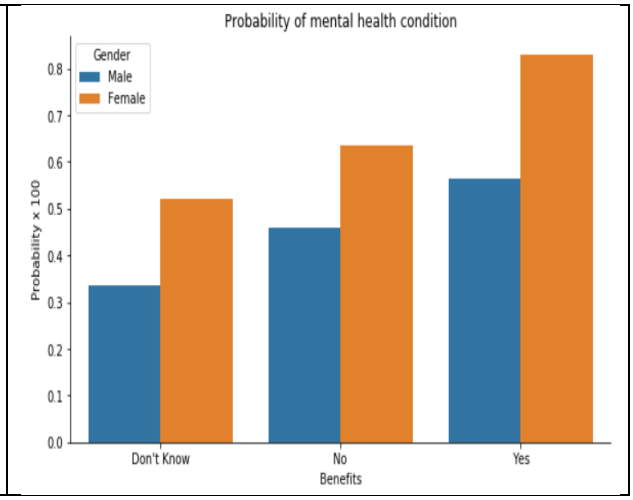


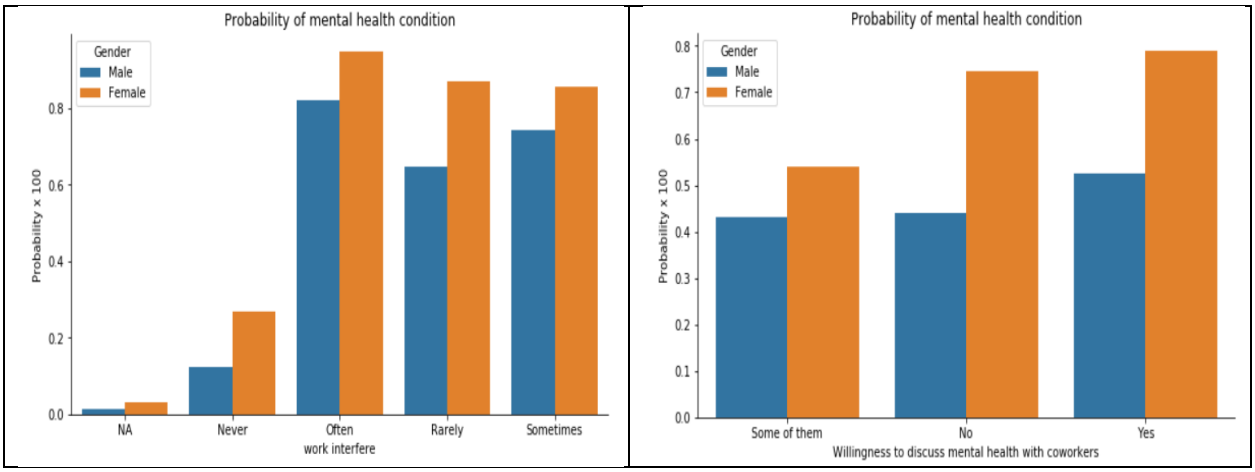
Next let's look at the role of family history. We see that the proportion of males and females seeking treatment with a family history is only different by 10%. However, there is a significant difference in females seeking treatment without a family history. We should not assume that the proportion of males with mental illness is low based on this because there is a huge stigma around males admitting to having Mental illness and seek treatment for it. Females on the other hand are more open and approach it as any other illness and seek the treatment without stigma.



Continuing further with our exploration we see a clear difference in the proportion of females seeking treatment than males in every category. Some of which stand out to us are when the employees know about care options, benefits, willingness to discuss their illness with colleagues and the amount the work interferes with their illness. All of these suggest quite clearly that even if males know about their treatment options they seem less likely to seek it out because of the stigma attached to discussing and publicizing their illness.







Machine Learning – Creating Predictive Models:

One of the most important tasks in trying to improve productivity and retention among employees affected by mental illness is identifying those who are seeking treatment. We can do so by leveraging the 22 other features in our dataset. Since the outcome of the question we are trying to answer is either a yes or no (1 or 0) it is a binary classification problem. Therefore we will start with a Logistic regression classifier. To start, I created a numpy array of the target variable y for treatment and created a numpy array X for the rest of the features. I then split X and y into a 70%-30% split between training and testing data.

Logistic Regression Classifier

I created and applied a base model of logistic regression to the training data and trained it with default parameters. Next, I evaluated the results such as accuracy, confusion matrix, classification report and ROC curve. After establishing baseline results, I created another model to tune the hyperparameters to improve the results. To do this I focused on the GridSearch CV method and applied it over a wide range of C and Penalty parameter values. I then evaluated the results the best values of the hyperparameters in the pipeline gave. The evaluation showed us that the base model performed better overall and the hyperparameter tuning in this specific case didn't improve our model.

	Base Model	Tuned Model
Accuracy	83.3%	83.06%
Precision	84%	83%
Recall	83%	83%
F-1 score	83%	83%

Support Vector Machine Classifier

I created and applied a base model of SVM to the training data and trained it with default parameters. Next, I evaluated the results such as accuracy, confusion

matrix, classification report and ROC curve. After establishing baseline results, I created another model to tune the hyperparameters to improve the results. To do this I focused on the GridSearch CV method and applied it over a wide range of C and gamma parameter values. I then evaluated the results the best values of the hyperparameters in the pipeline gave. The evaluation showed us that the base model performed the same as the tuned model and the hyperparameter tuning in this specific case didn't improve our model.

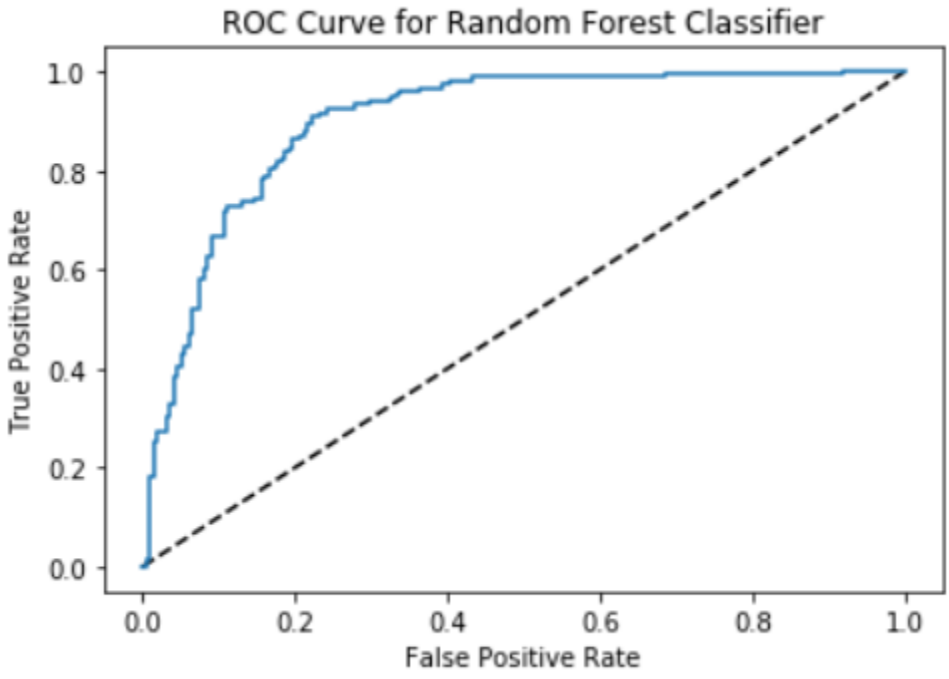
	Base Model	Tuned Model
Accuracy	83.06%	83.06%
Precision	83%	83%
Recall	83%	83%
F-1 score	83%	83%

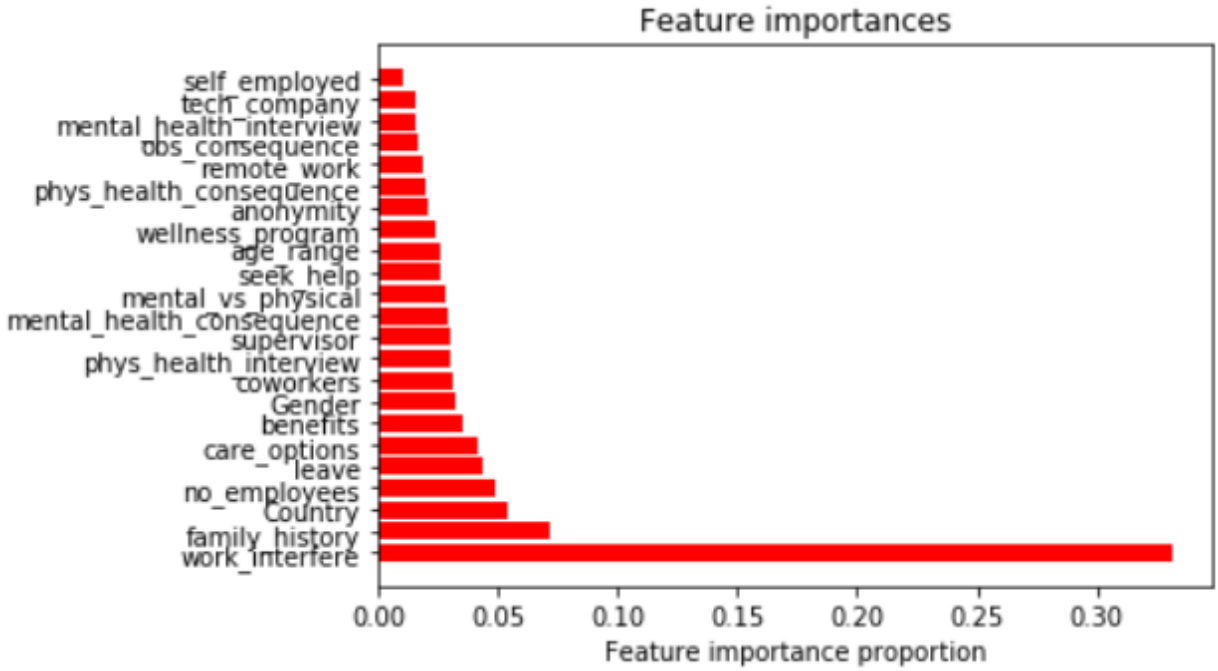
Random Forest Classifier

I created and applied a base model of Random Forest Classifier to the training data and trained it with default parameters. Next, I evaluated the results on the test data with metrics such as accuracy, confusion matrix, classification report and ROC curve. After establishing baseline results, I created another model to tune the hyperparameters to improve the results. To do this I focused on the GridSearch CV method and applied it over a wide range of n_estimators, max_depth, bootstrap, max_features values. I then evaluated the results the best values of the hyperparameters in the pipeline gave. The evaluation showed us that the tuned model performed better than the base model and the hyperparameter tuning in this specific case did improve our model.

	Base Model	Tuned Model
Accuracy	83.6%	84.6%
Precision	84%	85%
Recall	84%	85%
F-1 score	84%	85%

The Random Forest Classifier has given the best results in all category to us therefore we will take a closer look at its results. Below is the ROC curve for the model and the goal of the ROC curve is to have the largest area under the curve possible. In our case the area is 85% which suggests that if we get a completely new set of data we should be able to predict the correct results 85% of the time which is significantly better than guessing which will only give a 50% result. The final plot is that of feature importance. This tells which feature played what percent role in order to predict results for out model. The feature which jumps out the most is work interference. This right away suggests to us whether an employee has some sort of mental illness or not and whether it affects their work or not. For our clients it can play an important role to track this feature to identify individuals who have mental illnesses playing a role in their work but are still not seeking out treatment for it. The other important features are family history, care options, gender and leave etc which are as expected since these factors can play an important role in people mental health. The surprising features are country and number of employees.





Recommendations:

Conduct periodic anonymous surveys and seek permission from the users to use their results to be modeled. Use the Random Forest Classifier model to predict the results of whether an employee has sought treatment for mental illness. Train the model with new data to improve the tuned model further. Develop programs focusing on raising awareness for the following. Pervasive nature of the issue and

importance of treatment. Removing the stigma attached to discussing it in workplace. Existing programs to promote healthy lifestyle (Ex: yoga, fitness subsidy etc). Existing benefits and care options available for mental health. Bridging the gap for seeking treatment among genders. Collect happiness and productivity index data as part of the survey to assess the impact of the programs. Track the work interference column to see if the NA and never columns increases over time and the often, rarely and sometimes answer proportion decreases over time. This would suggest the people are either getting cured for their illnesses or getting the right treatment where it doesn't hinder their work productivity. Having happy and healthy employees mean they will most like lead happier and healthier lives outside of work too.