Mental Health in Technology Industry

Data Analysis Report

Tianyi Tan

Introduction

Given the rapid growth of technology industry, management has realized the importance of mental health and started to examine the accessibility to mental health care options in the company. According to an article written by Snobar (2018), individuals in tech industry, especially start-up founders, serves multiple roles in small teams with long working hours, worry about success constantly. Based on the statistics mentioned in this article (Snobar, 2018), 72% of the entrepreneurs surveyed by University of California self-reported mental health conditions which revealed the prevalence of concerns for mental health and the consequences of neglecting.

As the mental health concerns growing in the industry, the treatment seeking process and available options provided in the workplace have become vital to the healthy growth of the industry. Therefore, two research questions were motivated and aimed to be answered by the data analysis report:

1. How does the treatment seeking and attitudes toward mental health illness vary by geographic location?
2. What are the strong predictors of seeking mental health treatment in the workplace?

The report detailed the data collection and preparation in the **Data Preparation** section, introduced the methodology of four main machine learning techniques used in the **Method** section, revealed the results and analysis through visualization, comparison of performances of different techniques and important predictors in the **Analysis and Results** section, concluded the important predictors from both employees as well as companies and future work in the **Conclusion** section.

The report concluded that employees who had sought treatment in their past for mental health conditions tended to work in developed countries and states with big technology firms in the United States. They also might have family history and held the belief that mental health conditions interfered with their work. Their companies provided resources, options and benefits for mental health care. Therefore, management of technology firms should provide various options and benefits of mental health care. A better environment for employees to express their feeling while protecting their anonymity should be offered.

Data Preparation

The dataset contained results of a 2014 survey from Open Sourcing Mental Illness (OSMI) measuring attitudes towards mental health and mental health status in the tech workplace. The dataset was directly downloaded via Kaggle:
(https://www.kaggle.com/osmi/mental-health-in-tech-survey).
The dataset provided 27 variables with 1 numeric variable and 26 text variables recording the survey response. This dataset included views from employee's perspectives. It contained their personal information (e.g., age, gender, country, state), feedbacks about the available resources in the workforce, insights related to whether an employer recognized the importance of mental

health care perceived by employees and available information about their mental health status (e.g., family history, past treatment). Appendix Table 1 listed more detailed descriptions for all features. The feature “treatment” which recorded the response of “Have you sought treatment for a mental health condition?” were used as a target class label for classification and prediction.

Filling missing value

There were two features containing missing values. 'self_employed' contained 18 missing records and work_interfere contained 264 missing record (Appendix Table 2).
The missing values for self_employed were only 0.014% of the data. After carefully inspecting the records with missing value, the missing values were replaced by ‘No’ based on the answers to other survey questions.

The missing values for work_interefere were only 0.20% of the data. The missing values were replaced by ‘Don’t know’ because it might provide valuable information about this feature. Practically, people might use N/A to indicate that they did not know the answer to whether mental health condition interfered with their work.

Smooth the noisy data

Outliers were observed in five point summary of Age (Appendix Table-3) . The minimum and maximum of the age were below the legal age for working or clearly an impossible age value. For example, 5, 8 and 11 were considered to be illegal age for working. -1726, -29,-1, 329 and 99999999999 were not possible or beyond the range of human life expectancy. The outliers were smoothed with the median age. After smoothing, the distribution of age skewed slightly to the right and had a unimodal distribution.

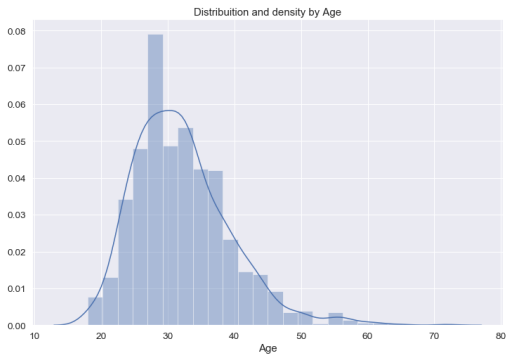


Figure 1 Distribution of Age (smoothed)

Resolve inconsistency

Due to the fact that the data was collected via survey, the responses for the same questions were in a large variety of different format. 43 unique values were provided for gender. For example, for the responder to identify himself as male in gender feature, 15 different kinds of response were recorded (e.g. “m”, “male”, “male-ish”). To resolve this issue, all the observations will be categorized and transformed into new variables. Two records were deleted during the process with value that cannot be categorized.

Data Transformation

Age as the only numeric data which was different from other features, binning and discretization for age were conducted. The cut points were chosen based on age ranges that might indicate different career status.

To prepare for machine learning process, text variables will be assigned number to represent different categories using one hot encoding which transformed the categorical variables into zeros and ones. A separate analysis with text variable encoded by label encoding that transformed the different levels of a feature into numbers was also conducted for comparison

Methods

Application of Machines Learning techniques

1. *Logistic Regression*

Logistic regression predicts the dependent variable with two categories. It transforms the dependent variable into logit function to obtain the maximum likelihood estimation and predict the odd ratio for the dependent variable.

2. *K-Nearest-Neighbors*

K-nearest-neighbors is a supervised learning for classification. To predict the label for the instance in the test set, distances between the instance and all instances in the trainings set are calculated. The labels of K instances with nearest distance will predict the label.

3. *Decision Tree*

Decision tree is a supervised learning for classification. In decision tree, instances were put in different node based on different splitting criteria. Gini index and entropy are two criteria to evaluate the splitting. Each parent node represents a splitting test on a feature, each branch represents the splitting outcome and each leaf node represents a class label. The path from root to leaf represents classification rules.

4. *Naïve Bayes*

Naïve Bayes is a supervised classification technique based on Bayes' Theorem with an assumption of independence among features. Probabilities are estimated based on observed frequencies in the training data. Predictions are made based on the probabilities.

Parameter Tuning

The hyper parameters of different machine learning algorithm were tuned by GridSearchCV and RandomizedSearchCV. They provided the best performance by search for the best combinations of parameters. The score function was leveraged to determine the model performance. 'Accuracy' score was used for measuring classification performance.

Model Evaluation

Best parameters were selected based on accuracy via parameter tuning methods.
Accuracy, false positive rate, precision, confusion matrix, ROC curve (receiver operating characteristic curve), AUC (Area under the ROC Curve) score and cross-validated AUC score were evaluated for the model with best parameters of different machine learning techniques.

Analysis and Results

Data Exploration and Visualization

For this part, all the text variables are label encoded to indicate different categorical levels using label encoding. Appendix Table-4 provided the data description for all features.

For this analysis, the target variable was "treatment" which recorded the response of "Have you sought treatment for a mental health condition?". The boxplot (Appendix Figure-1) of 'treatment' showed that it had an approximately equal 'Yes' and 'No' labels.

From the correlation matrix (Appendix Figure-1) for all features, it can be observed that most of the features only have weak correlations. The correlation between 'treatment' and features with moderate to high correlation to it had been analyzed.

Treatment has high correlation with work_interfere ("If you have a mental health condition, do you feel that it interferes with your work?", correlation: 0.62). People who had experience mental health issue tend to believe that the mental health issue would interfere with their work. The two features seek_help ("Does your employer provide resources to learn more about mental health issues and how to seek help?") and wellness program ('Has your employer ever discussed mental health as part of an employee wellness program?') has moderate to high correlation. The correlation seemed to be reflecting the fact that when employer provided resources, they might also take a step further and discuss mental health as a part of an employee wellness program.

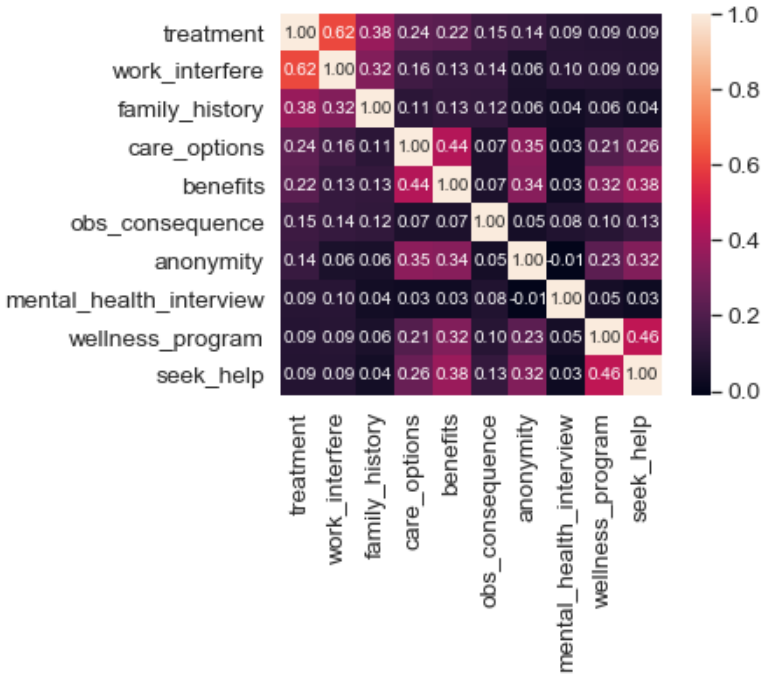
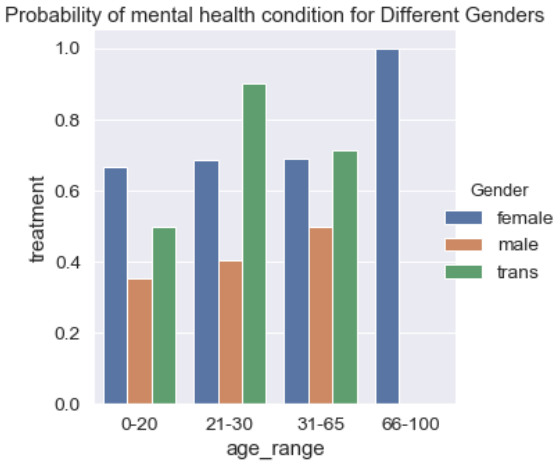


Figure 2 Correlation Matrix for the Top 10 Strong Associations

The median of the distribution of age for people having treatment was slightly higher for people without treatment. The distribution of age for people having treatment peaked at around 28 to 30. Both distribution was skewed to the right.

Figure 3 Distributions of Age by Treatment Value

For age range from 66-100. 100% of the people who had sought treatment for a medical condition are female. There were less male had sought treatment in the past.
People with family history of mental issue tended to seek for treatment. People with family history of mental issue and also trans gender tended to have higher probability.



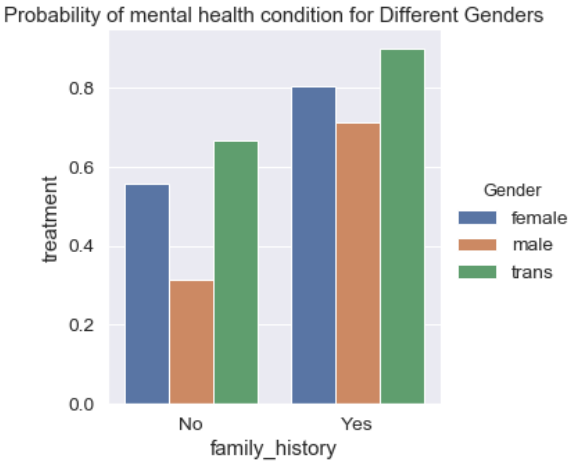


Figure 4 Treatment Probability by Family History and Gender *Figure 5 Treatment Probability by Age Ranges and Genders*

Employees who know the options for mental health care provided by the employers tended to seek treatment for mental health issue more. More of the employees who believed mental health issues often or sometimes interfered with their work sought treatment.

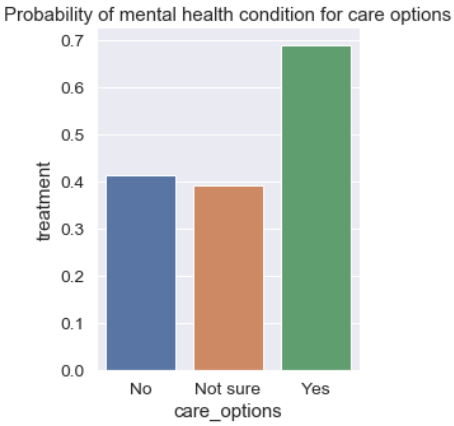


Figure 6 Probability of Treatment by Care Options

Figure 7 Probability of Treatment by Work Interference

Geographic data

A separate analysis on the geographic data was conducted to visualize the geographic difference on mental health in technology industry. The bar chart showed that most of the people filling the survey were from the United States. The proportions of people had sought treatment were high in the United States, United Kingdom, Canada and Germany.

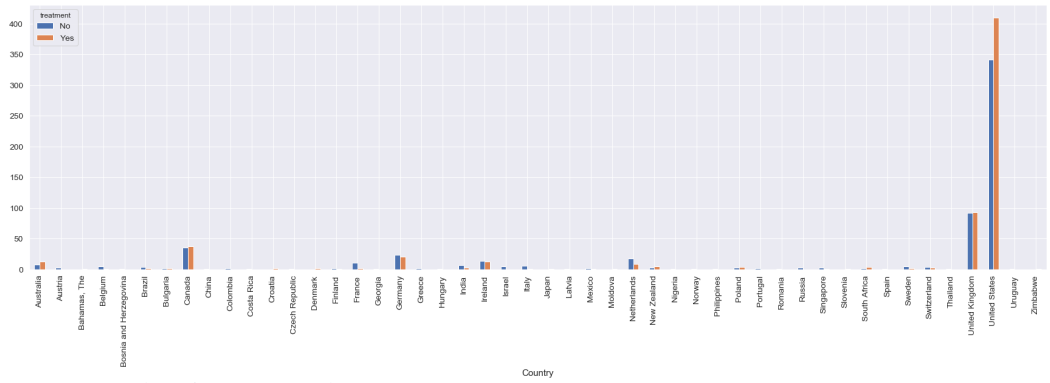


Figure 8 Barchart for Treatment by Countries

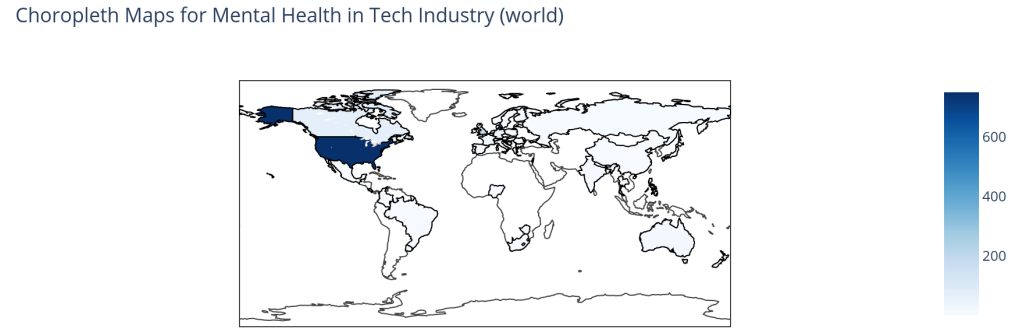


Figure 9 Choropleth Maps for Treatment (Countries) (generated by Plotly)

In United States, the geographic differences in different states were also analyzed. California (138) had the highest count for people had treatment, followed by Washington (70) and New York (57).

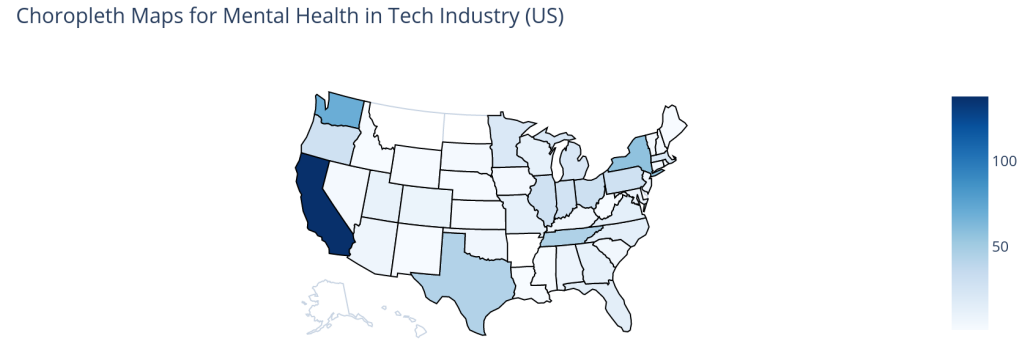


Figure 10 Choropleth Maps for Treatment (States in the US) (generated by Plotly)

Pattern Discovery:

The dataset was split into training (80%) and testing set (20%). 10-fold cross validation was conducted in grid search and randomized grid search. Four machine learning techniques, two encoding techniques and four different distance metrics were used for comparison. The results and analysis of one-hot-encoding data were the main focus of this report as a demonstration of the analysis process.

1. *Logistic Regression*

Logistic regression was used together with ridge regularization which added square magnitude of coefficient as penalty term to the loss function. The best model explained 25.35% of the variance of the dependent variable. Appendix Table 5 listed the top 10 important features ranked by the standardized coefficients. Work_interefere (Sometimes, Often and Rarely), seek_help (Don't know), family_history(Yes), care_options (Yes), coworkers (Yes), benefits(Yes), supervisor(No) and anonymity(Yes) were the top 10 important features. The model provided 81.35% accuracy.

2. *K-Nearest-Neighbors*

Different distance metrics (Jaccard, Dice, Matching and Euclidean, Appendix Note -1 for definition in sklearn), weights and K (from 1 to 30) had been evaluated. Best parameters were chose based on accuracy after 10-fold cross-validation. The best model selected 'Matching' and 'distance' weights with K equaled 15, which provided 77.01% accuracy for training set and 72.62% accuracy for test set.

3. *Decision Tree*

Different splitting criteria (Gini index or entropy), minimum parents or child nodes, maximum feature and depth were evaluated using randomized grid search. Best parameters were chose based on accuracy after 10-fold cross-validation. Appendix Table 6 listed the top 10 important features ranked by the feature importance. Work_interefere (Sometimes and often), family_history(No), Work_interefere (Rarely), care_options (Yes), Work_interefere (Never), benefits(No), Obs_consequence(No), anonymity(Don't know) and supervisor (No) were the top 10 important features. The best model with 11 depth selected entropy as the splitting criteria, 21 for maximum features, 7 for minimum leaf nodes and 7 for minimum parent nodes, which provided 78.36% accuracy for training set and 77.38% accuracy for test set.

4. *Naïve Bayes*

Both Gaussian Naïve Bayes and Bernoulli Naïve Bayes were utilized. Gaussian Naïve Bayes model had 80.95% accuracy and Bernoulli Naïve Bayes had 77.78% accuracy.

The results were summarized in the Appendix Table 7. Logistic regression and Gaussian Naïve Bayes provided higher accuracy, precision and cross-validated AUC scores compared to other techniques.

Discussion

One-hot-encoding (OHE) versus Label Encoding (LE):
Training data transformed by OHE in fact provided better results for logistic regression and Naïve Bayes techniques. Decision tree and KNN performed better with data transformed by LE.

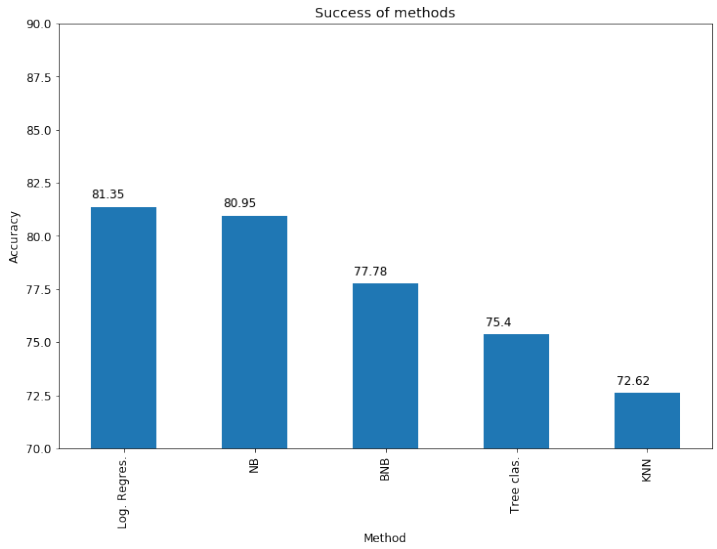


Figure 11 Machine Learning Performance for One-Hot-Encoding Data

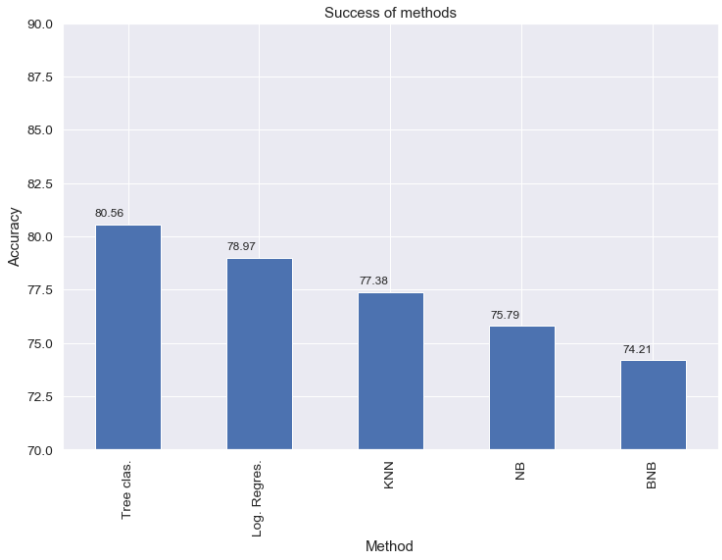


Figure 12 Machine Learning Performance for Label Encoding Data

The reasons why OHE had bad performance for some of the techniques were that it induced sparsity into the dataset. Dummy variables increased the dimensionality of the dataset and the curse of dimensionality impaired the prediction through distance calculation for KNN and splitting process in Decision Tree.

However, for better interpretation of the results, OHE was chosen due to the fact that LE assumed the data to be in order but it might be misleading for nominal features.

Feature importance from Logistic Regression (LR) and Decision Trees (DT)

By computing the standardized coefficient from LR and obtaining the feature importance from DR, the top 10 important features for both were shown in the bar charts.
Work_interfere ("If you have a mental health condition, do you feel that it interferes with your work?") seemed to be the top indicator agreed by both models. Family history ("Do you have a family history of mental illness? ") were also on the top list. Seek_help ("Does your employer provide resources to learn more about mental health issues and how to seek help?
") and Care_options ("Do you know the options for mental health care your employer provides?") were the fourth and sixth most important features for LR.

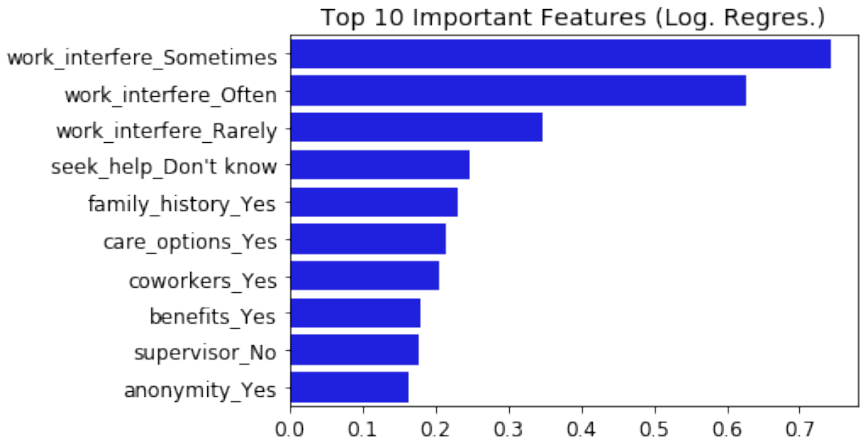


Figure 13 Top 10 Important Features for Logistic Regression

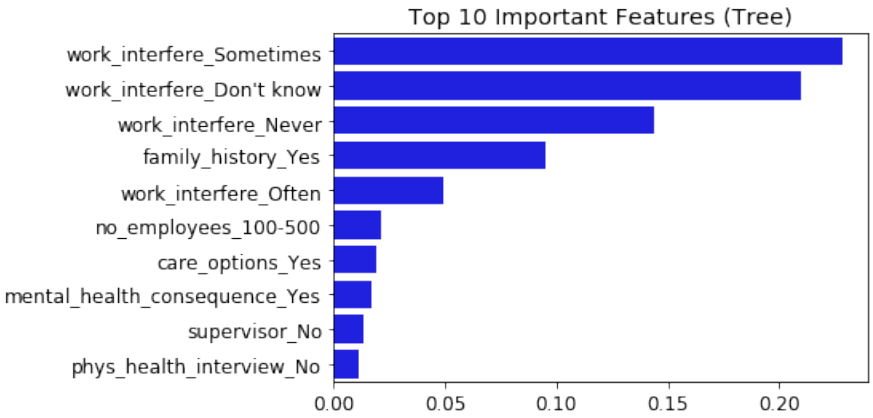


Figure 14 Top 10 Important Features for Decision Tree

For employees to seek treatment for mental health issue, they seemed to believe that the mental health condition might interfered with their work. Family history of mental illness also made them more likely to seek for help. Companies also played an important role by providing resources to learn more about mental health, options for mental health care and mental health benefit (benefits_Yes). They were willing to discuss the issue with their co-workers (co-worker_Yes) instead of direct manager (supervisor_No) and preferred to have their anonymity protected (anonymity_Yes).

Conclusion

By utilizing data visualization, Logistic Regression, K-Nearest-Neighbors, Decision Tree and Naïve Bayes, the report answered the two questions listed in the introduction.

Developed countries such as United States and United Kingdom, certain states in the US with big technology firms such as California and Washington tended to have more employees in technology industry sought for treatment for mental health issue. It might be because people were more aware of their mental health issues, provided more options for treatment seeking or more population in the technology industry.

Through pattern discovery process, the treatment seeking was found to be associated factors from employees themselves and also the companies. For employees, whether the mental health issue interfered with their work and whether they had family history was associated with treatment seeking the most. In companies with resources, options and benefits for mental health care, treatment seeking was more common for employees.

Thus, the implication for the technology firms would be providing more mental health care options and educated their employees with their options and benefits. Also, they should offer a safe environment for them to express their feelings and protected their privacy. It is important to remember that wellbeing takes different forms for different people. Managers in the tech industry need to create a culture in their teams where employees can unplug from their device and seek help anytime when they feel their work is interfered by mental health conditions.

Future work

The dependent variable 'treatment' ("Have you sought treatment for a mental health condition?") could be interpreted in several ways. First, the employees had concerned about their mental health issue. Also, they were willing to and were able to seek treatment. However, it also provided some ambiguity because the treatment might be sought before employment. To know more about the mental health issue in the industry, it would be good if the answers about their current mental health status could be collected. More conclusive and in-depth analysis would be conducted.

Reference

Snobar, A. (2018). Getting Honest About Mental Health In The World of Tech Starups. Forbes. Retrieved from: https://www.forbes.com/sites/forbestechcouncil/2018/08/08/getting-honest-about-mental-health-in-the-world-of-tech-startups/#24f2851a641a

Appendix

Table-1

Data Description:

	Variables	Description
1	Timestamp	Time the survey was submitted
2	Age	Respondent Age
3	Gender	Respondent Gender
4	Country	Respondent Country
5	State	If you live in the United States, which state or territory do you live in?
6	Self-employed	Are you self-employed?
7	family_history	Do you have a family history of mental illness?
8	treatment	Have you sought treatment for a mental health condition?
9	work_interfere	If you have a mental health condition, do you feel that it interferes with your work?
10	no_employees	How many employees does your company or organization have?
11	remote_work	Do you work remotely (outside of an office) at least 50% of the time?
12	tech_company	Is your employer primarily a tech company/organization?
13	benefits	Does your employer provide mental health benefits?

14	care_options	Do you know the options for mental health care your employer provides?
15	wellness_program	Has your employer ever discussed mental health as part of an employee wellness program?
16	seek_help	Does your employer provide resources to learn more about mental health issues and how to seek help?
17	anonymity	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
18	leave	How easy is it for you to take medical leave for a mental health condition?
19	mental_health_consequence	Do you think that discussing a mental health issue with your employer would have negative consequences?
20	phys_health_consequence	Do you think that discussing a physical health issue with your employer would have negative consequences?
21	coworkers	Would you be willing to discuss a mental health issue with your coworkers?
22	Supervisor	Would you be willing to discuss a mental health issue with your direct supervisor(s)?
23	mental_health_interview	Would you bring up a mental health issue with a potential employer in an interview?
24	phys_health_interview	Would you bring up a physical health issue with a potential employer in an interview?

25	mental_vs_physical	Do you feel that your employer takes mental health as seriously as physical health?
26	obs_consequence	Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
27	comments	Any additional notes or comments

Table - 2

Missing Value

Age	0
Gender	0
Country	0
self_employed	18
family_history	0
treatment	0
work_interfere	264
no_employees	0
remote_work	0
tech_company	0
benefits	0
care_options	0
wellness_program	0
seek_help	0
anonymity	0
leave	0
mental_health_consequence	0
phys_health_consequence	0
coworkers	0
supervisor	0
mental_health_interview	0
phys_health_interview	0
mental_vs_physical	0
obs_consequence	0

Table – 3

Statistical Summary for Age

	Age
count	1259.000000
mean	79428148.311358
std	2818299442.981968
min	-1726.000000
25%	27.000000
50%	31.000000
75%	36.000000
max	99999999999.000000

Table – 4

Data Description

	Age	Gender	self_employed	family_history	treatment	work_interfere	no_ employees	remote_work
count	1257	1257	1257	1257	1257	1257	1257	1257
mean	32.07	0.82	0.11	0.39	0.51	2.29	2.79	0.30
std	7.27	0.42	0.32	0.49	0.50	1.60	1.74	0.46
min	18	0	0	0	0	0	0	0
25%	27	1	0	0	0	1	1	0
50%	31	1	0	0	1	3	3	0
75%	36	1	0	1	1	4	4	1
max	72	2	1	1	1	4	5	1
	tech_company	benefits	care_options	wellness_program	seek_help	anonymity	leave	mental_health _consequence
count	1257	1257	1257	1257	1257	1257	1257	1257
mean	0.82	1.05	0.95	1.03	0.91	0.65	1.41	0.85
std	0.39	0.84	0.87	0.57	0.69	0.91	1.51	0.77
min	0	0	0	0	0	0	0	0
25%	1	0	0	1	0	0	0	0
50%	1	1	1	1	1	0	1	1
75%	1	2	2	1	1	2	2	1
max	1	2	2	2	2	2	4	2
	phys_health _consequence	coworkers	supervisor	mental_health _interview	phys_health _interview	mental_ vs_physical	obs_ consequence	age_range

count	1257	1257	1257	1257	1257	1257	1257	1257
mean	0.83	0.97	1.10	0.87	0.72	0.81	0.14	1.52
std	0.49	0.62	0.84	0.43	0.72	0.83	0.35	0.54
min	0	0	0	0	0	0	0	0
25%	1	1	0	1	0	0	0	1
50%	1	1	1	1	1	1	0	2
75%	1	1	2	1	1	2	0	2
max	2	2	2	2	2	2	1	3

Table – 5
Logistic Regression
Top-10 Most Important Features

Features	Std_Coef
work_interfere_Sometimes	0.7427
work_interfere_Often	0.6275
work_interfere_Rarely	0.3467
seek_help_Don't know	0.246
family_history_Yes	0.2315
care_options_Yes	0.214
coworkers_Yes	0.2042
benefits_Yes	0.179
supervisor_No	0.1767
anonymity_Yes	0.1637

Table – 6
Decision Tree
Top-10 Most Important Features

Features	Std_Coef
work_interfere_Sometimes	0.7427
work_interfere_Often	0.6275
work_interfere_Rarely	0.3467
seek_help_Don't know	0.246
family_history_Yes	0.2315
care_options_Yes	0.214
coworkers_Yes	0.2042
benefits_Yes	0.179
supervisor_No	0.1767
anonymity_Yes	0.1637

Table – 7

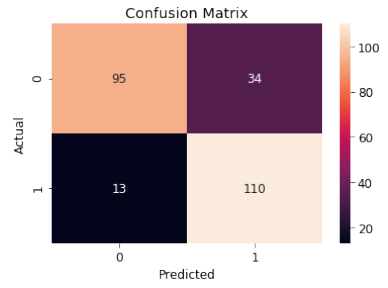
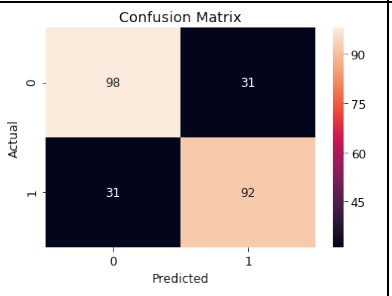
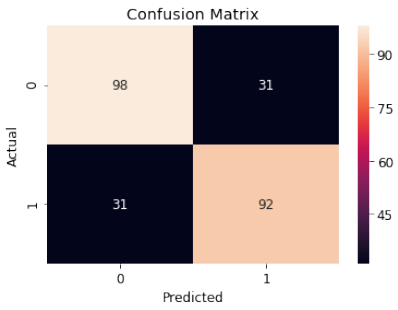
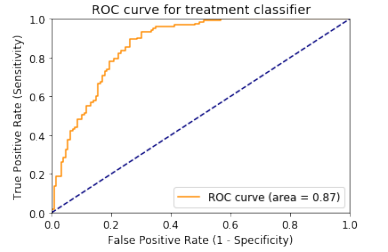
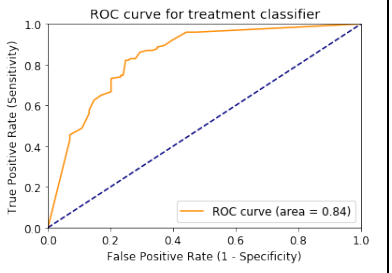
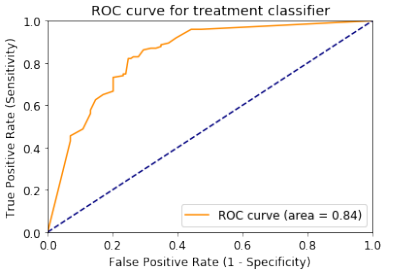
General Modeling results:

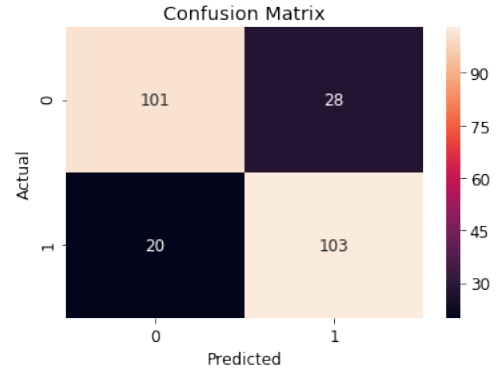
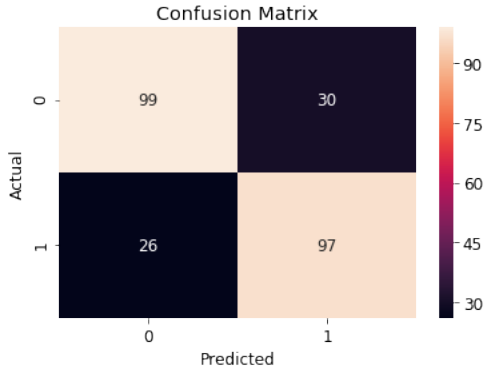
```
########### Logistic Regression ###############
Null accuracy:
 0    129
1    123
Name: Yes, dtype: int64
Actual Class ('Yes') : 0.4880952380952381
Actual Class ('No') : 0.5119047619047619
```

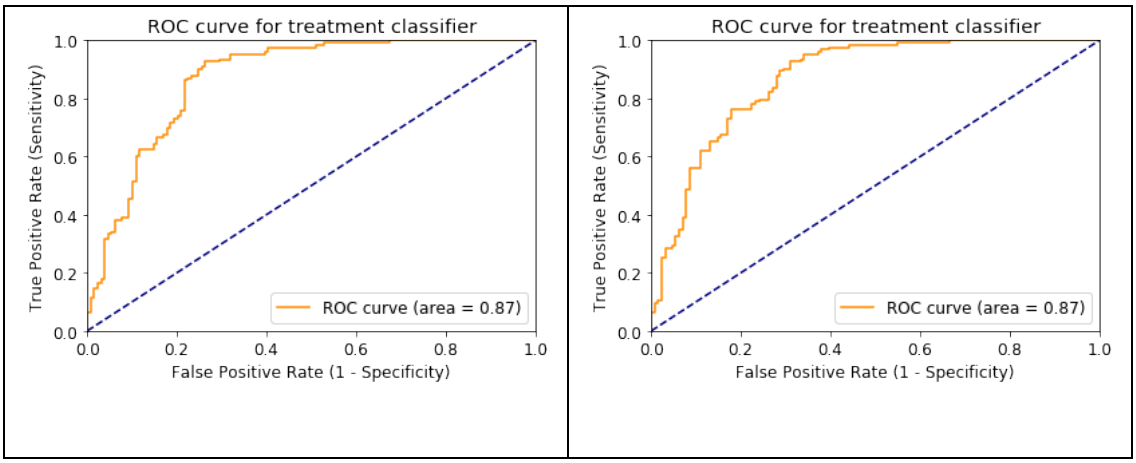
```
Classification Accuracy: 0.8134920634920635
Classification Error: 0.1865079365079365
False Positive Rate: 0.26356589147286824
Precision: 0.7638888888888888
AUC Score: 0.8153715258082813
Cross-validated AUC: 0.89235231054787

########### KNeighborsClassifier ###############
Null accuracy:
 0    129
1    123
Name: Yes, dtype: int64
Actual Class ('Yes') : 0.4880952380952381
Actual Class ('No') : 0.5119047619047619
Classification Accuracy: 0.7261904761904762
Classification Error: 0.27380952380952384
False Positive Rate: 0.26356589147286824
Precision: 0.7213114754098361
AUC Score: 0.7259406314993382
Cross-validated AUC: 0.8441024225550434

########### Tree classifier ###############
Null accuracy:
 0    129
1    123
Name: Yes, dtype: int64
Actual Class ('Yes') : 0.4880952380952381
Actual Class ('No') : 0.5119047619047619
Classification Accuracy: 0.753968253968254
Classification Error: 0.24603174603174605
False Positive Rate: 0.24031007751937986
Precision: 0.7479674796747967
AUC Score: 0.7538287010777085
Cross-validated AUC: 0.8403361815156171

########### GaussianNBClassifier ###############
Null accuracy:
 0    129
1    123
Name: Yes, dtype: int64
Actual Class ('Yes') : 0.4880952380952381
Actual Class ('No') : 0.5119047619047619
Classification Accuracy: 0.8095238095238095
Classification Error: 0.19047619047619047
False Positive Rate: 0.21705426356589147
Precision: 0.7862595419847328
AUC Score: 0.8101720552089242
Cross-validated AUC: 0.874020737327189
```

Table 8

Logistic Regression	KNN	Decision Tree
Accuracy: 0.8135	Accuracy: 0.7262	Accuracy: 0.7540
Precision: 0.7639	Precision: 0.7213	Precision: 0.7480

CV AUC: 0.8924	CV AUC: 0.8441	CV AUC: 0.8403
		
		

Gaussian Naïve Bayes	Bernoulli Naïve Bayes
Accuracy: 0.8095	Accuracy: 0.7777
Precision: 0.7863	Precision: 0.7637
CV AUC: 0.8740	CV AUC: 0.8732
	



Tree plot

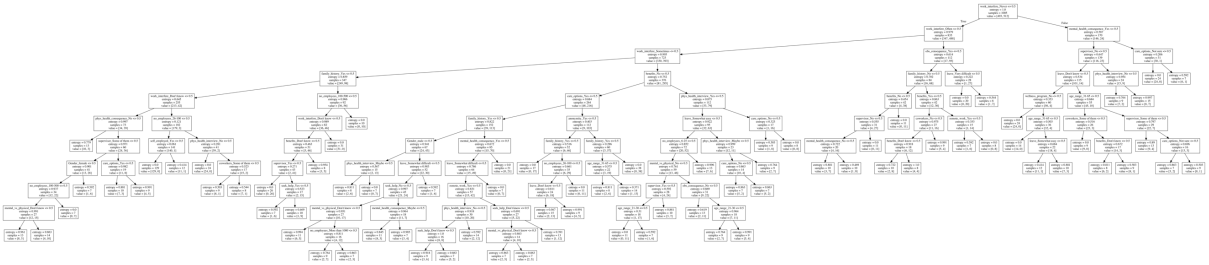


Figure – 1

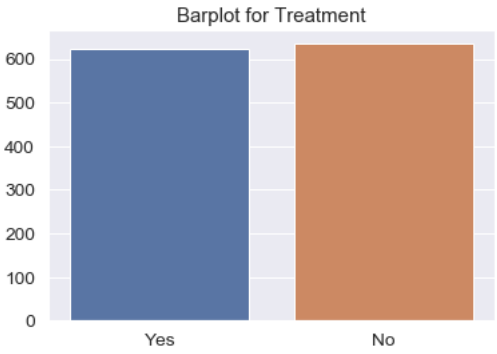
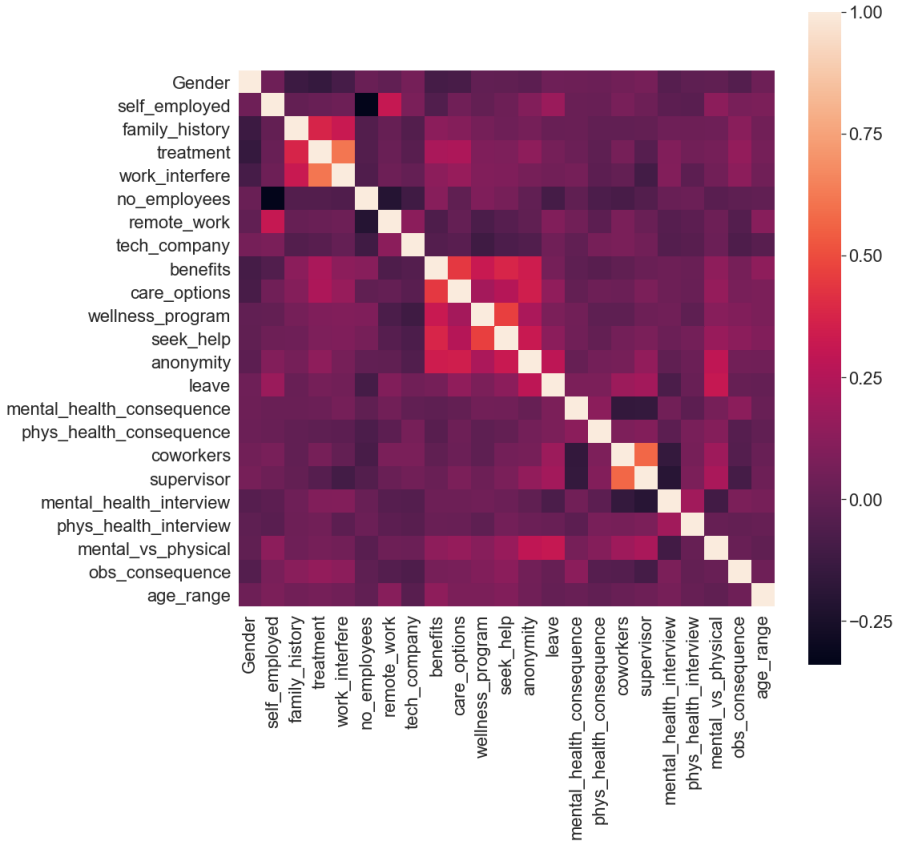


Figure - 2

Correlation Matric for All Features



Note -1
Distance metrics:

1. Jaccard: number of non-equal dimensions divided by number of nonzero dimensions.
2. Dice: number of non-equal dimensions divided by sum of number of dimensions in which both values are True and number of nonzero dimensions
3. Matching: number of non-equal dimensions divided by number of dimensions
4. Euclidean: the square root of the sum of square of the difference in each dimension

Code:

Code for Data Pre-processing, Data Visualization (geographic data and all other features) and Machine Learning process (Label Encoded Data) can be found in the following files:

Preprocessing Visualization and LE.html
Project Preprocessing Visualization and LE.ipynb

Code for Data Pre-processing and Machine Learning process (One-Hot-Encoding Data) can be found in the following files:

Final Models (One-Hot-Encoding) and Model Comparison.html
Final Models (One-Hot-Encoding) and Model Comparison.ipynb