Data Science in the Wild Project Milestone Report: Mental Health in Tech

Mia Casey, Sara Wang, Madison Ramsey

Abstract

For this project, we explored mental health in the tech industry by examining data from the OSMI Mental Health in Tech survey. Our goal was to determine if we could predict whether employees felt comfortable discussing mental health with their direct supervisor(s) given their other responses to other questions in the survey. We compared the performance of K-Nearest Neighbors, Logistic Regression, Random Forest, Boosting, and Support Vector Machine classification models. We also used word vectorizer models to analyze free text responses to extract key words and sentiments corresponding to our target variable.

Background

We aim to bring awareness to mental health in the workforce, specifically tech, and reduce the social stigma surrounding mental health. Mental health can impact employee productivity, happiness, and retention/turnover. As individuals looking to enter the tech industry workforce, we are personally invested in this topic as well. We believe that by understanding how employees feel about the culture surrounding mental health at their companies, companies can begin to identify areas for improvement and action.

Rather than trying to predict the individuals' mental health status, as previous Kaggle notebooks associated with OSMI datasets have done, we are more interested in understanding the company's culture around mental health based on participants' survey responses. We will predict whether employees feel comfortable discussing mental health with their direct supervisor. The target variable is '*Would you feel comfortable discussing a mental health issue with your direct supervisor(s)?*', with possible values of Yes, No, or Maybe. We will treat this target variable as a proxy for the company's culture surrounding mental health. Evaluating feature importance will allow us to understand the particular characteristics of a company that contribute to employees' comfort bringing this topic up with their supervisor.

Dataset

We will use the 2017, 2018, and 2019 OSMI datasets. Combined, these datasets have 1525 data points and 59 features. This data can be stored on our computers and downloaded directly from OSMI. The 2017-2019 data is richer than the 2014 and 2016 data because it has more text responses. We can thus implement more advanced NLP methods and build comprehensive word vectorizers. With our data exploration, we have found that the predictor and features appear to be reasonably balanced (no bias).

Analysis

Survey Questions Analysis

We began our analysis by identifying columns that were shared across each dataset. We found that the survey questions asked in 2014 and 2016 varied greatly with the 2017-2019 datasets, so we decided to use only data from 2017-2019 for a richer set of features. Since we plan to aggregate the data into one dataset, we identified which columns were in all three datasets, including columns whose names were exact matches and those that were different only in spelling, punctuation, or formatting.

From this column subset, we identified the unique values of each column for each dataset and evaluated if the columns had the exact same possible values, if the values were different but logically equivalent, or if the values had different meanings. We found that the possible values for most columns in the 2017 and 2018 data were equivalent, but some of the 2019 data, though logically equivalent, would need to be transformed. For example, the question '*Has being identified as a person with a mental health issue affected your career?*' had possible values of *0* or *1* in the 2017 and 2018 datasets while the 2019 data had possible values of *True* or *False*. In addition to identifying transformations required to align the data across each year, we identified transformations that would be required for that column in general. This primarily consisted of bucketing columns as nominal categorical or ordinal categorical where applicable.

Once we had an understanding of the meaning of each column and its possible values, we determined columns that we could safely remove. Because our research goal is to understand the attitudes of respondents' current employers towards mental health, we removed all questions related to previous employers. We also removed columns containing survey metadata (e.g. survey start and submit dates). Finally, we removed columns with no responses since these would not be meaningful.

Text Responses

We found that the text responses were fairly sparse, but still interesting. We decided to explore them separately and extract the following five text response questions from our columns: '*Describe the conversation you had with your employer about your mental health, including their reactions and what actions were taken to address your mental health issue/questions.*', '*Describe the conversation with coworkers you had about your mental health including their reactions.*', '*Describe the conversation your coworker had with you about their mental health (please do not use names)*.', '*Why or why not?*' (This question is a follow up to '*Would you be willing to bring up a physical health issue with a potential employer in an interview*?'), '*Briefly describe what you think the industry as a whole and/or employers could do to improve mental health support for employees.*' We pre-processed the text data to remove noise from the unprocessed text in four steps: convert text to lowercase, remove punctuation, tokenize words, and strip the stop words (e.g. "the", "and"). We created five text vectors corresponding to one of the five questions, and performed text analysis to find the most common words in predicting our target variable, *y*.

Merging Datasets

First, we made sure that all categorical response values were consistent across all three datasets. We converted all True/False responses in the 2019 dataset to 1/0 to align with the 2017 and 2018 datasets. We confirmed that all remaining questions had the same categorical response values and formats, with two exceptions related to the following questions: *'If possibly, what disorder(s) do you believe you have?' and 'If so, what disorder(s) were you diagnosed with?*'. In the 2019 dataset, these questions likely had dropdown menus from which multiple responses could be selected, as the values consisted of lists of mental health disorders. The 2017 and 2018 datasets instead had a column for each disorder with responses in the form of binary values. Despite the different formatting, the mental health disorders were the same across all three datasets. This allowed us to transform each of the two question columns in the 2019 dataset into multiple

columns (one for each mental health disorder) with 1/0 values and align formatting with that of the 2017 and 2018 datasets.

Next, we added a '*Year*' column onto each dataset so that we could keep track of which year the data came from after merging into one table. We renamed columns that were logically equivalent across all three datasets so that we could merge the tables together. With equivalent column names and response values, we joined the tables together by vertically stacking their rows.

Finally, '*What is your gender?*' had user-inputed responses that varied in meaning, spelling, and formatting, so we decided to address this column after merging. We created six response levels for this question: 'cis male', 'cis female,' 'trans male', 'trans female', 'genderqueer/non-binary', and 'other'. Using regular expressions and some case-by-case mapping, we grouped the survey responses into these six categories so that this feature would have consistent labels for modeling.

Missing Data

For each column remaining after our initial analysis of survey questions, we identified columns containing missing or NA values and investigated reasons for the missingness. We identified a number of columns that consistently had missing data when respondents indicated that they were self-employed, as well as columns that only had data for respondents that were self-employed. The systematic relationship between these columns and '*Are you self-employed?*' suggests that they are MAR. It is likely that the original surveys made these questions available based on respondents' self-employment statuses.

Since our target variable '*Would you feel comfortable discussing a mental health issue with your direct supervisor(s)?*' is a question applicable only to respondents who are *not* self-employed, we decided to remove all columns with questions that only applied to self-employed individuals and drop all rows for self-employed respondents. This should not introduce bias in our dataset since our target variable only exists for participants who are not self-employed. Outside of this, we also identified 2 rows in the 2017 data where all responses are missing, and dropped these rows.

After merging our three datasets, we noticed that many missing data points could be explained by survey structure. Many of these questions were follow-up questions to previous questions in the survey. For example, all missing responses to '*Has being identified as a person with a mental health issue affected your career?*', corresponded to individuals who answered 'No' to '*Are you openly identified at work as a person with a mental health issue?*'. In this case and all similar cases, we decided to replace missing data with a 'Not Applicable' category, as these questions simply did not apply to the individuals who had missing responses.

All missing data for questions asking about where individuals lived or worked in the United States corresponded to individuals who did not live or work in the United States, and was similarly replaced with a 'Not Applicable' category. We also saw that '*What is your race?*' was missing a large proportion of values. Without domain knowledge on how to proceed or a clear methodology on how to replace this data, we decided to drop this column altogether.

Imputation

The remaining columns with missing values were: '*Have you had a mental health disorder in the past?*', '*Have you ever discussed your mental health with coworkers?*', and '*Have you ever had*

a coworker discuss their or another coworker's mental health with you?' To determine if these were MAR or MCAR, we performed chi-square tests against all other feature columns in the dataset to determine if there were possible relationships between columns. Because we found possible significant relationships with other variables for these columns, we diagnosed all as MAR. For example, chi-square tests returned p-values<0.05 when comparing '*Have you had a mental health disorder in the past?*' against '*How has it affected your career?*', '*Do you currently have a mental health disorder?*', '*Have you ever been diagnosed with a mental health disorder*', '*What US state or territory do you live in?*', '*What country do you live in?*', '*What country do you work in?*', and '*What is your gender?*'

We then used KNN imputation to impute values for our MAR-identified columns ('*Have you had a mental health disorder in the past?*', '*Have you ever discussed your mental health with coworkers?*', and '*Have you ever had a coworker discuss their or another coworker's mental health with you?*').

We first created subsets of the merged data for each column that excluded the other two columns with missing values. For example, for '*Have you ever discussed your mental health with coworkers?*', we created a dataset excluding '*Have you had a mental health disorder in the past?*' and '*Have you ever had a coworker discuss their or another coworker's mental health with you?*' From there, we made '*Have you ever discussed your mental health with coworkers?*' our target variable, since we would effectively be predicting the missing values in this column using KNN. We then split out any rows missing values for our target variable into a test dataset, while the remaining acted as our training dataset.

For each training subset, we created a 70/30 training/validation split. We then created KNN models and evaluated their accuracy for up to k=20 neighbors. The table below describes the best accuracy for each.

Target Column	*Optimal K*	*Prediction Accuracy*
'Have you had a mental health disorder in the past?'	k=7	0.6567656765676567
'Have you ever discussed your mental health with coworkers?'	k=9	0.6957928802588996
'Have you ever had a coworker discuss their or another coworker's mental health with you?'	k=17	0.6440129449838188

Finally, we used the best performing models for each target variable and predicted the missing values on their respective test sets. The predicted values were then imputed into the original feature set.

Encoding and Transforming Data

Before beginning feature selection or fitting classification models to our data, we needed to scale and encode data accordingly. Before encoding different column values, we renamed our column names from survey questions to abbreviated column names using Camel Case naming

convention, for example "*Do you have previous employers?*" was changed to *prevEmployers*. After renaming, we noticed that *selfEmployed* and *possCurrMhd_Psychotic* had all uniform values, so we dropped them from the dataset.

After renaming, we were ready to encode our variables. We began by encoding the target variable, *comfortDiscussMhSupervisor*, to labels (1, 2, and 3) using sklearn's LabelEncoder. Then, we looked at each of the survey questions and split them into nominal and ordinal variables. For our 31 nominal variables, such as "*Does your employer provide mental health benefits as part of healthcare coverage?*", we dummy-encoded them using Pandas' get_dummies built-in function. Then, for our 12 ordinal variables, such as "*If a mental health issue prompted you to request a medical leave from work, how easy or difficult would it be to ask for that leave?*" with five levels ranging from "Very Easy" to "Difficult", we created a dictionary mapping for each of the variables and used Pandas' map function to recode the values. Lastly, for the 23 columns corresponding to a specific mental health disorder, such as *diagMhd_Anxiety*, we binary encoded the variables to 0 (no disorder) or 1 (disorder).

Feature Selection

To decrease multicollinearity in the case of logistic regression and generally increase model interpretability, we decided to test whether a subset of our features would increase model performance. We used multiple methods for feature selection, including incremental PCA, kernel PCA, k best subset selection, recursive feature elimination, removing features with low variance, selecting the top k percentile of features with the highest ANOVA f-value, and tree-based feature selection.

Each set of features was used to train a logistic regression, SVM, random forest, boosting, and K-nearest neighbors model. The classification accuracy of each model-type was compared across feature subsets using 5-fold cross validation. Some feature selection methods worked better for certain algorithms, but we chose the method of selecting the top k percentile of features with the highest ANOVA f-value because it improved model performance across the board.

To tune the hyperparameter k for how many features we select, we again used 5-fold cross validation to compare accuracy results for each algorithm across different values of k. We found that values of k=40 and k=5 resulted in feature subsets that most significantly improved model performance compared to the full set of features. Thus, we decided to carry out our modeling by training on three different sets of features: 1) the full feature set, 2) the features in the top 40th percentile with the highest ANOVA f-value, and 3) the features in the top 5th percentile with the highest ANOVA f-value.

Results

Text Responses

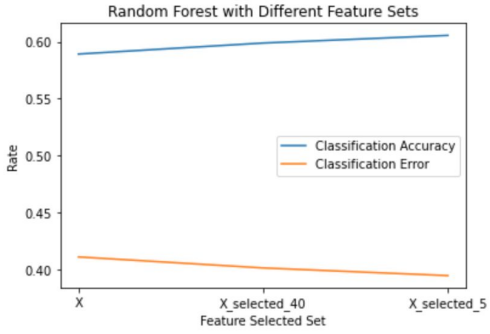
We used the processed data from the text response processing and created two word vectorizer models using sklearn CountVectorizer for each of the five questions: a Bag of Words (1-gram) model and a 2-gram model. For our text analysis, we trained a Logistic Regression model using both feature vectors. We inspected the weight vector from the model to identify the most important words for deciding whether an individual feels comfortable discussing mental health

with a direct supervisor. For each question, we list the ten words with the largest weight vector with both the 1-gram and 2-gram model:

1. *'Describe the conversation you had with your employer about your mental health…'*
    a. 1-gram: therapy, better, time, sick, office, diagnosis, wanted, anxiety, problems, talk
    b. 2-gram: take time, health issues, anxiety disorder, time needed, anxiety depression, let know, discussed anxiety, mental health, depression anxiety, lot anxiety
2. *'Describe the conversation with coworkers you had about your mental health…'*
    a. 1-gram: could, don't, much, stress, slack, therapy, struggles, always, talking, go
    b. 2-gram: mental health, slack channel, ive talked, health issues, ive discussed, discuss mental, coworkers mental, mental illness, health coworkers, depression anxiety
3. *'Describe the conversation your coworker had with you about their mental health…'*
    a. 1-gram: another, feel, support, problems, lot, ive, general, need, direct, would
    b. 2-gram: depression anxiety, one coworker, coworkers tell, another coworker, mental health, health issue, anxiety depression, direct report, coworker told, panic attacks
4. *'Why or why not?'*
    a. 1-gram: potential, employers, though, order, ive, bias, someone, well, otherwise, feel
    b. 2-gram: potential employer, need know, feel like, affect chances, depend whether, might need, hurt chances, relevant interview, even though, relevant job
5. *'Briefly describe what you think the industry as a whole and/or employers could do to improve mental health support for employees.'*
    a. 1-gram: flexible, environment, clear, outside, safe, encouraging, encourage, developers, good, use
    b. 2-gram: health days, support employees, long hours, would help, health disorders, take care, employers need, make clear, health benefits, less stigma

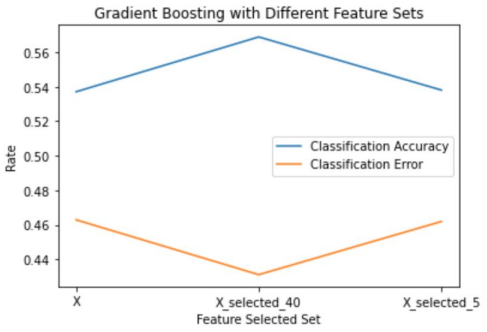
We underlined a few of the 1-gram and 2-gram words that gave us insight into how people feel that their conversations and feelings about mental health may contribute to whether or not they would feel comfortable discussing a mental health issue with their direct supervisor.

Random Forest Classifier



Scikitlearn's RandomForestClassifer was used to train a random forest model on each of the three feature sets. With the default hyperparameter settings, the feature subset consisting of the top 5th percentile of features with the highest ANOVA f-value yielded the best performance. Moving forward with this feature subset, the 5-fold cross validation classification accuracy was assessed for different maximum depth values. We found that a max_depth value of 7 yielded the model with the highest performance. After tuning the number of estimators, minimum samples per leaf, and minimum samples split hyperparameters, we found that values of

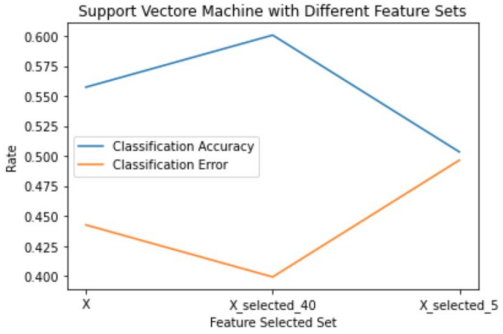
n_estimators=100 and min_samples_leaf=9 returned the highest average CV accuracies. Our final random forest model yielded a cross validation accuracy of 0.607.



Gradient Boosting Classifier

The Scikitlearn GradientBoostingClassifier was used to train a gradient boosting model on each feature set. With the default hyperparameter settings, the feature subset consisting of the top 40th percentile of features with the highest ANOVA f-value yielded the best performance. With this feature subset, the 5-fold CV classification accuracy was assessed for different maximum depth, number of estimators, and learning rate values. Our best performing model had hyperparameter values of max_depth=7, n_estimators=100, and learning_rate =1.0 and obtained an average cross validation accuracy of 0.569.

Support Vector Machine

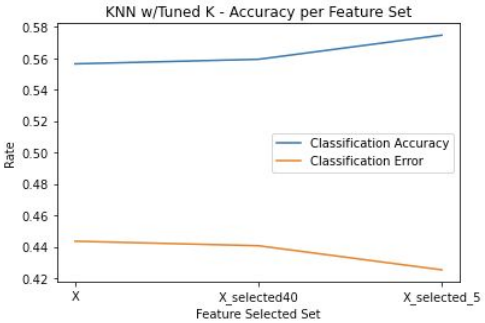


We used Scikitlearn's SVC classifier to develop a Support Vector Machine model. First, we standardized our data using Scikitlearn's StandardScaler(). Then, we trained the SVC classifier with default hyperparameter settings on each of the three feature sets. The feature subset consisting of the top 40th percentile of features with the highest ANOVA f-value yielded the best performance. The 5-fold CV classification accuracy was assessed for different kernel types using this feature set, and a 'sigmoid' kernel was found to yield the best model performance. Using the sigmoid kernel, our best SVM model obtained an average cross validation accuracy of 0.601.

Logistic Regression

Scikitlearn's LogisticRegression was used to train a random forest model on each of the three feature sets. The feature subset consisting of the top 5th percentile of features with the highest ANOVA f-value yielded the best performance. This Logistic Regression model trained on the best feature subset achieved an accuracy of 0.615. In the figure, we can see that the accuracy and error increased and decreased accordingly when the feature subset size decreased.

K-Nearest Neighbors



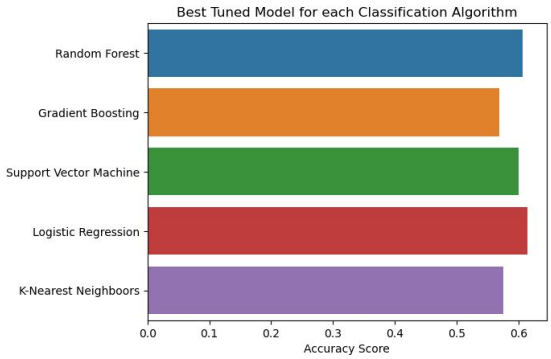
We used Scikitlearn's Neighbors and GridSearchCV to train and tune KNN models on each of the three feature sets using 5-fold CV. We found that the best performance occurred using the top 5th percentile of features with the highest ANOVA f-value and 17 neighbors. This had a 5-fold CV mean accuracy of 0.575.

Accuracy and error increased and decreased respectively for the tuned KNN models when the feature subset size decreased.

Conclusion

Findings

We ultimately found that Logistic Regression on the feature subset consisting of the top 5th percentile of features with the highest ANOVA f-value best predicted the target variable '*Would you feel comfortable discussing a mental health issue with your direct supervisor(s)?*' with a classification accuracy of 0.615. The figure compares the accuracy of the best tuned model for each classification algorithm we evaluated.



The relatively low classification accuracies suggest that the target variable we selected may not have been as representative of the company's culture surrounding mental health as we had initially hypothesized. It may instead be the case that an individual's relationship with their direct supervisor is distinct from the individual's relationship with the company at large, or how the individual perceives the company's culture with regard to mental health.

Despite this, our text analysis revealed the most important words respondents used in their free text survey responses impacting their willingness to discuss mental health with a direct supervisor. Individuals who referenced therapy, taking time off, anxiety, and depression in conversations with employers about mental health, and those who mentioned the concepts of struggle and support in conversations with coworkers were more likely to be willing to discuss mental health with a direct supervisor. This suggests that those who are generally more open to these conversations in a work environment are more likely to discuss mental health with their supervisor. Values such as flexibility, safe and encouraging environments, health benefits, and less stigma in relationship to how the tech industry and tech employers could improve mental health support for employees also impacted individuals' willingness to discuss mental health with their supervisor.

Reflection and Future Work

The vast majority of our work was spent on data exploration, cleaning, and feature engineering. Since our data was generated from surveys but we did not have access to the original surveys, our column and missing value analysis required us to reverse-engineer the survey structure in order to understand the meaning of the data in each column. Through this we found that our chosen target variable was applicable only to non-self-employed individuals. While we chose to exclude self-employed individuals from the scope of our project, the sentiment of self-employed individuals in the tech industry around mental health could be a future area of analysis. Additionally, given the results of our classification analysis, future work could explore other columns as potential target variables. Ultimately, this project has reinforced the concept that in order to perform meaningful analysis on data, it is critical to incorporate domain knowledge, understand the content of the data, and understand how the data was collected.

Contributions

We all contributed to column analysis, reformatting and cleaning data, and completing project deliverables. Mia completed categorical encoding, text analysis, and Logistic Regression model analysis. Madison took point on feature engineering and Random Forest, Boosting, and SVM model analyses. Madison and Sara worked together on missing value analysis. Sara and Mia renamed feature columns. Sara completed missing value imputations and KNN model analysis.

References

- OSMI Mental Health in Tech Survey (2017, 2018, 2019)
- Kaggle Competition, Mental Health in Tech Survey (2014)
- Prompt (software to facilitate conversations about mental health in tech)
- Mental Health in the Workplace, CDC
- Mental Health in the Workplace, WHO
- American Psychiatric Association's Center for Workplace and Mental Health
- Kaggle notebooks:
    - Predictors of mental health treatment (Notebook, Notebook) on 2014 data
    - Looking at the mental health in tech 2016 data (Notebook), predict whether someone currently has a mental health disorder
    - Exploratory data analysis on survey data from 2014 - 2018 (Notebook)