pods project

by Shiva Taherzadehlari

Mental health condition and disruption at work

Shiva Taherzadehlari
Department of Computer Science
City, University of London
London, United Kingdom
Shiva.Taherzadehlari@city.ac.UK

Abstract —In this paper, we used a mental health survey in the tech dataset to find some of the factors that affect disruption at work as a result of having a mental health condition and also to see If we can find any trend between employee profile and this disruption. Decision tree, random forest, and xgboost models were constructed to check how accurately we can predict the target variable and to find the most important features. The main focus of this paper is on this question that was asked from employees: 'if you have a mental health condition, do you feel that it interferes with your work?'

I. INTRODUCTION

Employers risk suffering significant financial losses if they fail to take care of their staff members' mental well-being. If we focus only on one aspect of mental health, such as depression, we can see that it has a significant impact on workplace productivity. The fraction of people who experience depression are most likely to feel intervention at work and would not be productive enough during working hours and might even miss some working days every week. These consequences of depression, cost employers an estimated $44 billion each year in lost productivity [1]. We can see how important of a problem this is when you consider other aspects such as acute stress disorder, bipolar disorder, and anxiety disorders as well. This paper focuses on analyzing the profile of the employees that are more prone to feel work intervention as a result of having mental health problems and the factors that might reduce or increase feeling intervention at work to provide some insights for employers so that they can prevent such burnouts and improve work productivity.

II. DATA AND ANALYTICAL QUESTIONS

The dataset is from a 2014 Open Sourcing Mental Illness (OSMI) survey and was downloaded from the Kaggle website.

As mentioned before, the analytical questions are designed to provide insights about the employee profile being most at risk and the factors that affect work intervention. Apart from that, it is also important to see how accurately can we predict how often employees might feel intervention at work and what features have the most effect on that. If successful, employers can use the same approach to predict who is more at risk of work intervention and take more serious actions towards preventing it. our questions divide into two groups: observational and modeling.

Observational questions:

1. Which age group and gender (employee profile) are more prone to feel intervention at work as a result of mental health problems?
2. Does having heard of or observed negative consequences for coworkers with a mental health condition in the workplace affect feeling work intervention?
3. Does caring or not caring about an employee's mental health affect how often an employee with mental health conditions feels work intervention?

Modeling questions:

4. To what extent can we predict how often an employee feels work intervention?
5. what features does the prediction model consider to be the most important?

III. DATA (MATERIALS)

KEY CHARACTERISTICS

The dataset is the result of a survey taken by 1259 employees answering 27 questions, so each observation represents an employee, and features are the questions that were asked. All features have categorical types. The columns include different types of information from personal information to mental health care options and services. Here are mostly used and focused attributes in this paper:

Column name	*question*	*value categories*
Work_interfere	If you have a mental health condition, do you feel that it interferes with your work?	Sometimes, never, Rarely, often, no
Age	What is your age?	From 18 to 72
Gender	What is your gender?	Male, female, others
Obs_consequences	Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?	Yes, no
benefits	Does your employer provide mental health benefits?	Yes, no, don't know
leave	How easy is it for you to take medical leave for a mental health condition?	Easy, difficult, don't know
anonymity	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?	Yes, no, don't know
Seek_help	Does your employer provide resources to learn more about mental health issues and how to seek help?	Yes, no, don't know
Wellness_program	Has your employer ever discussed mental health as part of an employee wellness program?	Yes, no, don't know

Ta

Value categories are shown after data preprocessing

Is It Suitable for Our Study?

The data includes the 'work_interfer' feature which is our target variable and the answer to analytical questions are captured by analyzing this feature with other variables. It also contains over 1200 responses and was the largest survey done on mental health in the tech industry [2] which helps us to make our choices with more confidence.

Key Assumptions

In the target variable, 262 missing values were replaced with 'No' assuming people who did not answer this question did not have mental health conditions as the question starts with 'If you have a mental health condition…'

IV.ANALYSIS

A) *Data preparation*

here are the preparation steps taken on the most used features:

1. Dealing with missing values and imputation:

missing values of the 'work_interfere' feature were replaced with 'No' because as mentioned before, it is assumed that employees who did not answer this question did not have any mental issues.

2. Removing irrelevant features:

'state' and 'comments' features were dropped because approximately 50 % of the data were missing. The 'timestamp' and 'country' features were also removed because they were not relevant to our analysis.

3. Detecting outliers:

For removing the outliers of the 'Age' feature, we first tried the two standard deviation method. However, this method did not remove all of the outliers. Consequently, we set thresholds for removing them meaning that we removed all of the ages above 80 and below 15. Overall, 8 rows were deleted and the number of rows decreased to 1251 after outlier detection.

4. Correcting the value inconsistencies:

There were more than 30 categories for the 'Gender' feature. They were reduced to only 3 types: 'Male', 'Female', and 'Others'. To make the analysis easier the same approach was used for the 'leave' attribute. The number of categories in this feature was reduced from 5 to 3.

B) *Encoding*

Before performing correlation analysis, the data was encoded using the Label encoding method which is appropriate for categorical data.
One hot encoding was also considered, but since the number of categories in some features is quite large, using this method would result in having a large data frame, and would make the correlation matrix complex and unexplainable, so it was not applied.

C) *Feature selection*

Feature selection reduces the dimensions of features and efficiently speeds up the learning process.[3]
In this paper, three methods of feature selection were used.

1. Correlation analysis:

Correlation analysis is applied to determine the degree to which two variables are related. In this paper, the spearman correlation coefficient is used because apart from working with linear relationships between variables, it works with monotonic relationships as well.

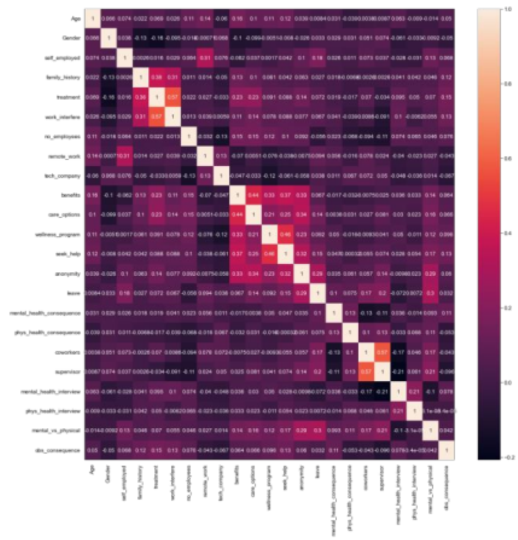

Figure 1: Heatmap showing the relationship between variables

In figure 1 we can see some of the variables have strong correlations, and there is not a very strong correlation between the target variable and other features.
To have a better understanding, the top ten positively and negatively correlated variables were extracted, here are the results:

Table2

Variable 1	Variable 2	Correlation
'supervisor'	'coworkers'	0.571820
'wellness_program'	'seek_help'	0.463808
'benefits'	'care_options'	0.438058

these three pairs of variables had the highest positive correlations (above 0.4) and the negative correlations between variables were not strong. After that, the top ten most correlated variables with the target variable were extracted and there was only one strong correlation: 'treatment' and 'work_interfere' with 0.565393 correlation.

For the next two steps, from the variable pairs that had high correlations, 'care_options', 'seek_help', and 'coworkers' w ere removed.

2. MCA:

mca method is an appropriate technique to gain a general understanding of how categorical variables are related and discover the underlying structure of the data. Figure 2 is the mca plot illustrating how categories spread out in certain ways along the components.

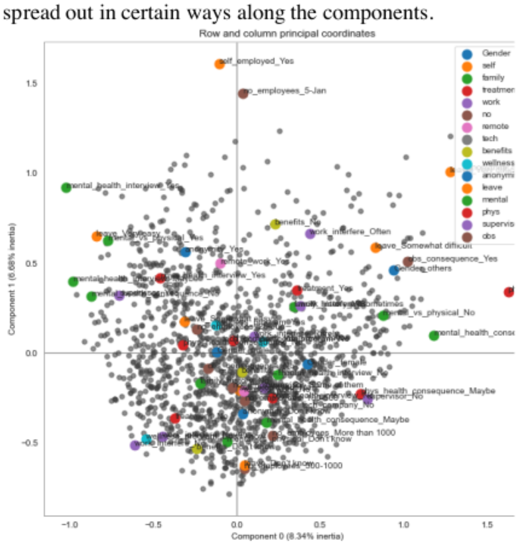

Figure 2

It is clear that in our case, mca is not a good representative of our data since the total variance of components 0 and 1 is 15.02% which is too small (it should be at least 80%).

3. Chi2 test for independence:

Chi-Square Test is another method to show the association b etween categorical variables.
It checks whether two variables are related or not.
Two categorical variables are dependent and more useful for classification when the chi2 statistic has a higher score.
In this paper we checked the association between 'work_interfer' and other features, figure 3 shows the results after getting the feature scores and plotting them.

Figure

Here also we can see that 'treatment' has the highest association with 'work _interfere' followed by 'family_history', 'obs_consequences', 'benefits', 'leave' etc. the top 5 most important features will be used to train our models.

D)construction of models:

D) construction of models:
Random forest, decision tree, and xgboost models were used to predict the target variable.
Since this dataset is not large, we assume random forest outperforms decision tree.
Random forest uses the bagging technique to take the average of many decision trees, while xgboost is built on boosting which is a strong alternative for bagging.
Moreover, xgboost includes a unique split-finding algorithm to optimize trees, along with built-in regularization that reduces overfitting.[4] as a result, xgboost might perform better than random forest.

validation of results:

Table 3

Decision tree Before tuning	model	Decision tree after tuning
0.48	F1	0.48
0.34	recall	0.34
0.484	accuracy	0.484
0.19	precision	0.19

Table 4

Random Forest Before tuning	model	Random Forest after tuning
0.47	F1	0.48
0.35	recall	0.34
0.468	accuracy	0.476
0.33	precision	0.26

Tables 3 and 4 illustrate after hyperparameter tuning using grid search, the accuracy for random forest had a slight improvement (from 0.47 to 0.48), but for the decision tree, it did not change (0.48). The accuracy for xgboost was 45.24%. Consequently, in our dataset, random forest and decision tree performed better.
The reason for gaining low accuracy in all of the models might be that as understood from figure 1, the prediction feature had a weak correlation with the training features.

Figure 4 shows the feature importance in random forest.

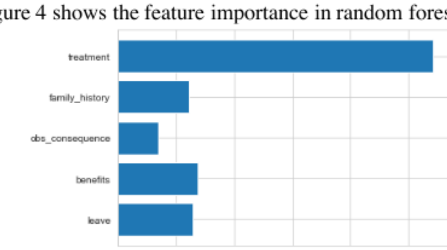

Figure 4

V. FINDINGS AND REFLECTIONS AND FURTHUR WORK

A) Findings And Reflections

1. Which age group and gender (employee profile) are more prone to feel intervention at work as a result of mental health problems?

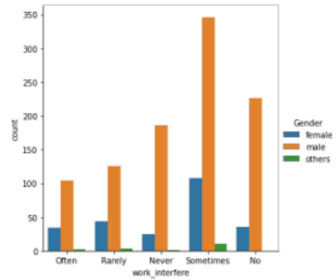

Figure 5

From figure 5 we understand that for men, the order of answers to this question: 'if you have a mental health condition, do you feel that it interferes with your work?*'* in the categories from most to least is:
'sometimes',' No',' Never',' rarely',' Often'
for women, it is: 'sometimes',' Rarely',' Often',' No', and' Never'.
Since 'often' appears in the third place for women and at the end for men and 'sometimes' is the first one for both groups, it can be concluded that women who have mental health issues are more likely to experience intervention at work. However, we should consider that as this survey was taken from tech company employees, women accounted for only 20% of the observations, so our data is small and this conclusion cannot be generalized to the whole female employees.
Moving on to age groups, figure 6 shows the distribution of 'Age' after removing outliers.

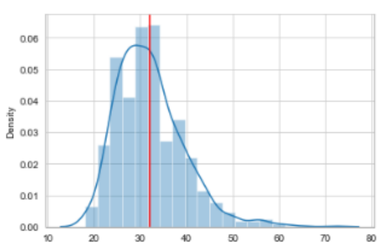

Figure 6

we can see that most of the employees conducting the survey were aged between 25 to 35 and the average age was 32.

Figure 7

From figure 7 we understand that since most employees who voted for 'often' work intervention were aged 30 and this is also true about the 'never' category and also because the maximum number of answers was given by employees who aged around 30, we cannot say which age group is more prone to feel work intervention.

In conclusion, the first research question could not be answered properly due to data limitations for females in the 'gender' column and the concentration of ages around 30 in the 'age' column.

2. Does having heard of or observed negative consequences for coworkers with a mental health condition in the workplace affect feeling work intervention?

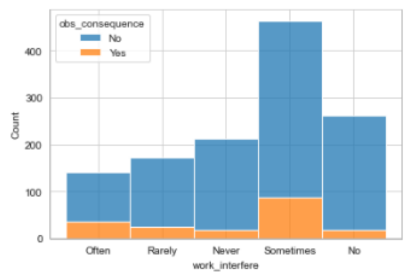

Figure 8

From figure 8 we understand that among those who 'often' and 'sometimes' felt work intervention, the proportion of people who had observed consequences is bigger than those who answered 'rarely',' never', and 'no' for 'work_interfer'. considering that the number of people who had not observed consequences was 6 times greater than those who had, we can still see that those who had observed, were more prone to feel intervention in work.
Consequently, if employers want to minimize the factors that might lead to work intervention, one of the things they should care about is to create a safe environment so that no one neither experiences nor observes negative consequences as a result of mental health problems.

3. does caring or not caring about an employee's mental health affect how often an employee with mental health conditions feels work intervention?

To answer this question, we considered the answers to 'benefit', 'leave', 'anonymity', 'seek_help', and 'wellness_program' questions(features) as representatives of whether an employer cares about the employee's mental health or not.

In other words, the answer 'yes' for 'benefits', 'anonymity', 'seek_help', and 'wellness_program' features and 'easy' for the 'leave' feature means that the employer pays enough attention to the mental health of the employees.

The next step would be visualizing and analyzing each of these features with the target variable, 'work_interfere':

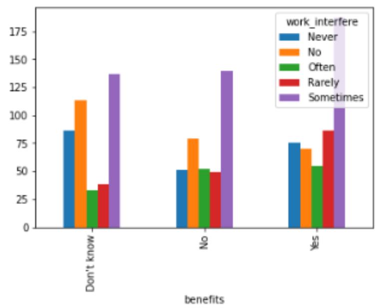

Figure 9

Figure 9 shows that among those employees whose employer provided mental health benefits, the proportion who voted for 'never' and 'rarely' for 'work_interfere' was higher than those who worked with companies without mental health benefits provided.

As a result, we can say providing mental health benefits for employees can result in less work intervention at work.

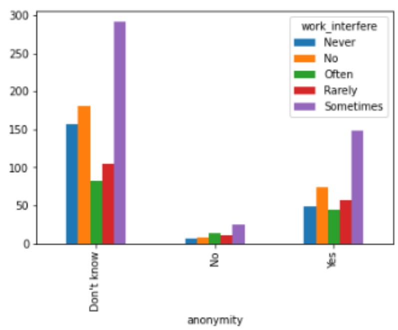

Figure 10

From figure 10 we understood that not protecting employees' anonymity can result in feeling more work intervention because among those who voted for 'No' for anonymity (meaning that their anonymity is not protected if they choose to take advantage of mental health resources), the proportion who 'often' felt work intervention is higher than those who voted for 'yes' for anonymity.

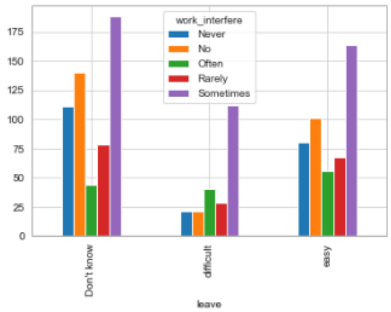

Figure 11

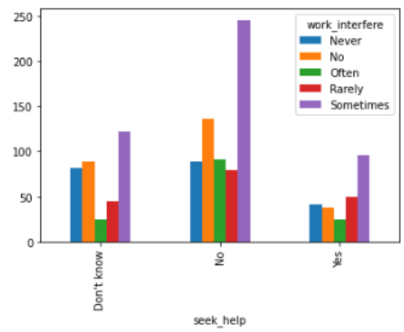

Figure 12

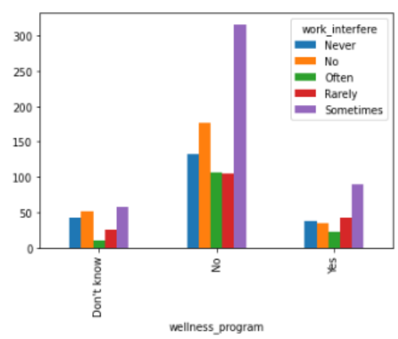

Figure 13

Analyzing the same way as we did for 'benefit' and 'leave' features for figures 11,12, and 13, we can understand being able to easily take a medical leave for a mental health condition, providing employees with resources to learn more about the mental health condition and how to seek help, and discussing mental health as a part of an employee wellness program can also lead to less intervention at work.

In simple words, caring about an employee's mental health can reduce how often an employee experiences work intervention. Lastly, this information can be insightful for employers as it shows how paying attention to simple and basic rules can positively affect the employers' well-being which increases productivity and prevention of financial loss.

4. To what extent can we predict how often an employee feels work intervention?
The accuracy of the models was not high. The best accuracy gained was 48% which means the models cannot predict the target variable properly.
5. what features does the prediction model consider to be the most important?
In figure 4, we can see the important features of the random forest prediction model with 'treatment' being the most important. However, since the accuracy is low, it is not logical to rely on the feature's importance produced by the model.

B) Further work:

1. Since 2014, this survey has been conducted each year and the data is available on the OSMI website. Collecting more data would help to make a more accurate analysis and find stronger trends for example for 'Age' and 'work_interfere'. It would also make the conclusions more reliable.

2. Tuning the hyperparameters of the xgboost model would be helpful to increase the accuracy of the model.

3. To improve the accuracy of the random forest and decision tree models, other hyperparameter tuning methods can be used such as random search.

4. other classification algorithms can be applied such as logistic regression and naïve Bayes to check the accuracy level.

5. As we observed in the correlation matrix, figure1, there was not a strong correlation between our target variable and other variables, so maybe by changing the target variable to another feature, we would get better results and higher accuracy.

VI. WORD COUNTS

SECTION	
abstract	105
introduction	174
Data and analytical questions	211
Data(materials)	173
analysis	850
Findings and reflections and further work	1029
overall	2542

VII. REFERENCES

[1]: Workplace Mental Health – Depression, December 18, 2022 (https://www.workplacementalhealth.org/mental-health-topics/depression)

[2]: Research, Open Sourcing Mental Health - Changing how we talk about mental health in the tech community – Stronger Than Fear December18,2022 (https://osmihelp.org/research)

[3]: Huan Liu and Rudy Setiono, Chi2: Feature Selection and Discretization of Numeric Attributes, (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=479783)

[4]: Corey Wade, 2020, Getting Started With XGBoost in scikit-learn,(https://towardsdatascience.com/getting-started-with-xgboost-in-scikit-learn-f69f5f470a97)