Case study:

Mental Health in Technology-related Jobs (Task 1)

Course: DLBDSMLUSL01 – Machine Learning – Unsupervised Learning and Feature Engineering

By Dmitrii Shevchuk
Matriculation: 9213737
22.08.2024

Tutor: Christian Müller-Kett

Introduction

Our company's Human Resources (HR) department is aiming to design a pre-emptive program towards mitigation of issues with mental health conditions (MHC’s) amongst the company staff. It is assumed that the company can not influence instances of MHC due to their very complex nature. However, the HR department expects that more information about MHC, benefits and more MHC friendly corporate culture may increase the share of people with MHC seeking treatment and by this - mitigate productivity loss.

To seek treatment a person should first recognise symptoms, and here both information and culture are important. After diagnostics some people get an official diagnosis and can start treatment covered by insurance or not. If they don’t get an official diagnosis, they can do nothing and continue to suffer or seek treatment with their own means.

Treated MHC on average supports higher productivity compared to untreated. However some MHC’s with low severity can produce low effect on productivity and it can be compensated by longer working hours or reduced salary. Hence seeking treatment is a complex decision involving type and severity of MHC, previous experience if any, available information, corporate benefits and culture related to MHC.

Goal of this case study is to learn what possible interventions by the company may produce what effects on productivity of its staff with MHC. This is not that obvious. For example, easy sick leaves with MHC symptoms may actually decrease an employee's motivation to seek treatment and productivity will suffer. So investment in such intervention will have negative returns.

For other interventions, returns may be close to zero or too small to make sense. Thus the model we are going to build, should help to design an optimal MHC program.

We will use data from “OSMI Mental Health in Tech Survey”, and on the way learn more about MHC in the IT industry.

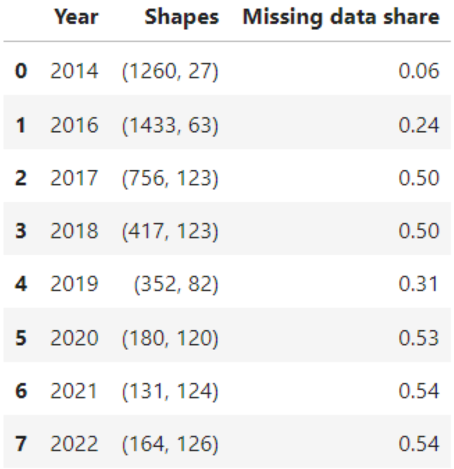
We start with estimating the scope of the problem: why should a company care about people with MHC? On the way we learn more about the dataset and types of MHC. In section 2 we outline the model and describe data preparation. In section 3 we describe the model and its implementation. Conclusion is about key findings and results.

The code involved in this project is possible to access by the permalink: https://github.com/shevchukum/OSMI-Mental-Health-in-Tech-Survey/blob/92e01652aed4251e4724c28cb37901e2f1b78837/MHC%20in%20IT%20analysis.ipynb

1. Data understanding

One of the goals of “OSMI Mental Health in Tech Survey” was to show how many people with MHC are actually working in the IT industry. The survey has 8 ways from 2014 till 2023. In 2023 organizers got only 6 respondents, so we skipped this wave.

Table 1: Shapes and Missing Data Share for Each Year in “OSMI Mental Health in Tech Survey”



	Year	Shapes	Missing data share
0	2014	(1260, 27)	0.06
1	2016	(1433, 63)	0.24
2	2017	(756, 123)	0.50
3	2018	(417, 123)	0.50
4	2019	(352, 82)	0.31
5	2020	(180, 120)	0.53
6	2021	(131, 124)	0.54
7	2022	(164, 126)	0.54

Shape is showing the number of rows (observations) and number of columns (questions). As can be seen, the number of questions is changing from wave to wave. Some questions appear, some disappear in later waves, some change wording. We will focus on questions of our interest and try to stack the answers from as many waves as possible.

1.1. Scope of the problem

To estimate the prevalence of MHC in the IT industry we can use the question 'Do you currently have a mental health disorder?' (DIS). It can be found in 2016-2022 waves. 40% of respondents answered ‘Yes’, and another 26% ‘Maybe’ (or ‘Possibly’).

Should we conclude that at least 40% of our staff are affected by MHC issues? Probably not. According to the National Institute of Mental Health (n.d.) in 2021 only 22.8% of adults experienced some form of diagnosed mental illness in the US. Higher levels can be found in prisons and jails - about 40% (Substance Abuse and Mental Health Services Administration, n.d.).

It is hard to believe that the IT industry has the same MHC levels as the criminal one. It is more likely that it has the same levels as the general population. In section 3.1 we estimate that for the given age, gender and country structure OSMI survey was roughly 3.2 times more interesting for people with MHC than without one. It is a self-selection bias.

Hence a company in the USA with similar to this dataset gender and age structure should have 64.6 / 3.2 = 20.2% of staff with diagnosed MHC and another 13.1% with some subclinical symptoms (‘Maybe’). Of course, for better estimates we should ask people in our own company.

How big is it in productivity terms? This is a more complicated question.

Starting from 2016 there are two questions we can use:

- 'If you have a mental health issue, do you feel that it interferes with your work when being treated effectively?' (ProTE)
- 'If you have a mental health issue, do you feel that it interferes with your work when NOT being treated effectively?' (ProNTE)

We need to encode the answers [Often, Sometimes, Rarely, Never] into the change of original (non MHC) productivity. Hence ‘Never’ should be 1 (no change of productivity), but should ‘Often’ be 0.5 or even 0?

In research of untreated MHC impact on productivity we find numbers in the range of 20-30%. For example Adler et al. (2006) found “... monotonic relationship between depression symptom severity and productivity loss: with every 1-point increase in PHQ-9 score, patients experienced an additional mean productivity loss of 1.65% (P <.001)”. Maximum possible PHQ-9 score is 27, hence the worst depression possible leads to 45% productivity loss, and the most common moderate one (10-14) leads to 20% productivity loss.

If we encode ProNTE (and ProTE) as [0.75, 0.9, 0.95, 1] we get average ProNTE in our dataset for people with diagnosis 0.79, which is 21% productivity loss. Treated MHC, according to this dataset, leads to a 9% reduction in productivity on average. Hence we get 12% of productivity back from treatment.

Actually for IT people loss of productivity due to MHC may be somewhat higher than average as this job demands long and complex brain activity. But let’s keep this encoding as a conservative one.

For people without diagnosis average productivity loss is 16% if untreated and 9% if treated, hence productivity gain from treatment is lower - only 7%.

So making people to get treatment can save up to 0.202 * 0.12 + 0.131 * 0.07 = 3.3% of total company productivity.

Next question: to what extent can the employer influence the decision to get treatment? Unfortunately there is no question in our dataset if a person is receiving MHC treatment, but there is a question in the 2016 wave: 'Have you sought treatment for a mental health condition?' (and similar in the following waves).

Of course, seeking treatment may not necessarily mean receiving it. For the general US population only 47% with diagnosed MHC received treatment in 2021 (65% with serious mental illness) (National Alliance on Mental Illness, n.d.). About 11% had no insurance and probably could not afford the treatment. Another reason is a lack of access to care (long waiting times, long distances to mental institutions in rural areas). And of course many people choose to avoid treatment even if they can afford it, meaning they do not seek treatment.

Let’s assume that financial constraints and lack of access to care is not significant for IT people, and about all seeking treatment get it both in case of official diagnosis and without it. Hence we can check out the effect of seeking treatment as an effect of receiving it.

To gauge company influence we construct an MHC friendliness index reflecting how well a company provides employees with MHC information, benefits and keep friendly to MHC corporate culture. Index is from 0 (unfriendly) to 1 (friendly). There are seven questions we can use for it (see section 2.2.2).

Table 2 shows that shifting from being MHC unfriendly (index<0.3) to MHC friendly (index>0.7) may increase the share of people seeking help by 10% points for those with diagnosis and by 25% for those who have not.

Table 2: Share of People with MHC Seeking Treatment in MHC Friendly and Unfriendly companies

	MHC unfriendly company (index < 0.25)	MHC friendly company (index > 0.75)
Share of people with official diagnosis seeking treatment	0.890	0.976
Share of people with NO official diagnosis seeking treatment	0.391	0.710

Welch's t-test p value for both differences is below 1% (see notebook).

This changes our productivity gain calculations to 0.202 * 0.12 * 0.086 + 0.131 * 0.07 * 0.319 = 0.5%. Impact is getting even lower if we move from medium friendly companies (index from 0.25 to 0.75) to highly friendly ones (index > 0.75), productivity gain shrinks to 0.25%.

To conclude: rough calculations show that MHC programs probably have to have low cost and be well targeted to have positive returns for the company. This underlines the importance of the model we are going to develop.

1.2. MHC types

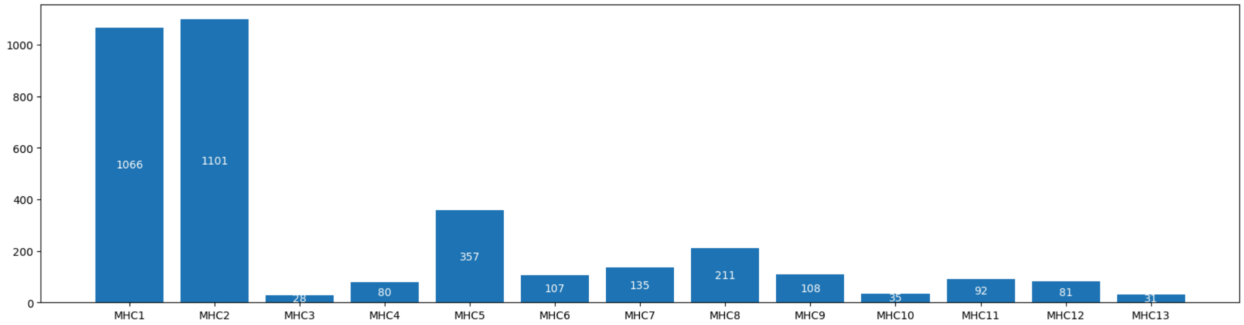
It is reasonable to expect that productivity effects will be different for different types of MHC. And our dataset allows us to include MHC types into the modeling, but only if we have productivity loss estimates for each disorder type. To get those we would need to do extensive literature research or ask some experts. On this stage we decided to skip the types. However we can give a brief overview of how the distribution of types looks like in the dataset.

In 2016 wave MHC type was an open question, but from 2017 the set was fixed to 13 types:

1. 'Anxiety Disorder (Generalized, Social, Phobia, etc)': MHC1
2. 'Mood Disorder (Depression, Bipolar Disorder, etc)': MHC2
3. 'Psychotic Disorder (Schizophrenia, Schizoaffective, etc)': MHC3
4. 'Eating Disorder (Anorexia, Bulimia, etc)': MHC4
5. 'Attention Deficit Hyperactivity Disorder': MHC5
6. 'Personality Disorder (Borderline, Antisocial, Paranoid, etc)': MHC6
7. 'Obsessive-Compulsive Disorder': MHC7
8. 'Post-Traumatic Stress Disorder': MHC8
9. 'Stress Response Syndromes': MHC9
10. 'Dissociative Disorder': MHC10
11. 'Substance Use Disorder': MHC11
12. 'Addictive Disorder': MHC12
13. 'Other': MHC13

After classifying open answers from 2016 into those 13 types, we get the Figure 1.

Figure 1: Bar Chart of 13 MHC Types for 2016-2022



Comparing these levels with the US general public ones (Table 3) we can see that not only is the total MHC prevalence too high in the survey, but also the type structure is different from the general population. It might be due to different age and gender structure.

Table 3: Comparison of MHC Type Structures in OSMI Survey and General US Public

	Prevalence in the survey	Prevalence in general US public
Mood Disorder (Depression, Bipolar Disorder, etc) (MHC2)	32%	10%
Anxiety Disorder (Generalized, Social, Phobia, etc) (MHC1)	31%	19%
Attention Deficit Hyperactivity Disorder (MHC5)	10%	4.4%
Post-Traumatic Stress Disorder (MHC8)	6%	3.6%

Note: one person may have two and more MHC’s.

2. Data preparation

We start with target variables for our model and complete the section with defying and studying features.

2.1. Target variables

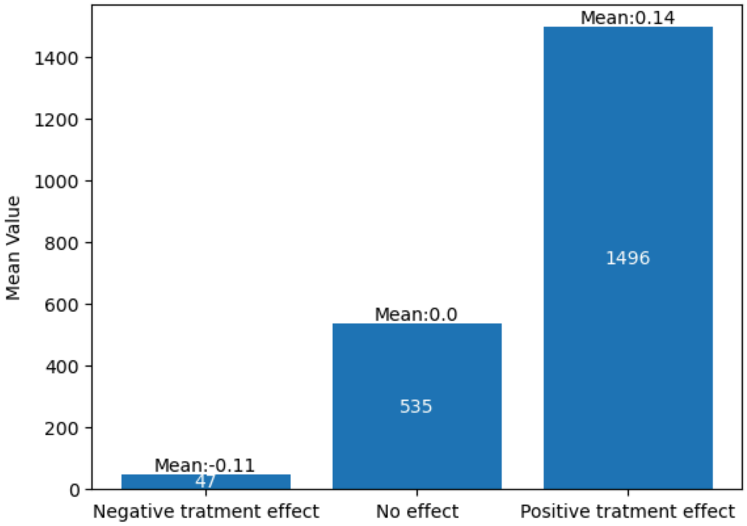
Our goal is to forecast productivity improvement, so we need two questions from 2016-2022 waves (with our short names):

- 'If you have a mental health issue, do you feel that it interferes with your work when being treated effectively?': ‘ProTE’

- 'If you have a mental health issue, do you feel that it interferes with your work when NOT being treated effectively?' : ‘ProNTE’

So our first target variable is ‘ProD’ := the difference between ‘ProTE’ and ‘ProNTE’. In the Figure 2 we can see that treatment does not always improve productivity, actually it can make it worse or may have no effect. This is another reason why some people choose not to seek treatment.

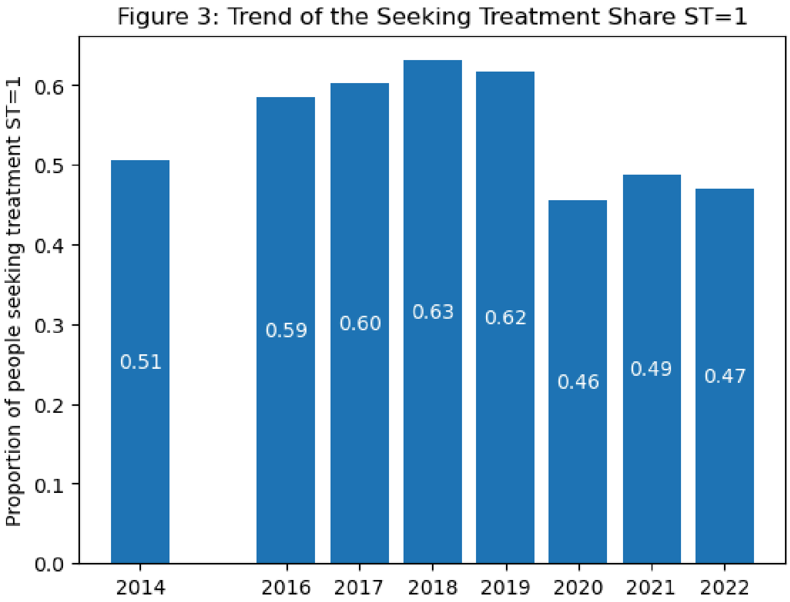
Figure 2: ProD Frequency and Means for Negative, Zero and Positive Effect of Treatment



Of course, the treatment effect can be estimated only by people who got (sought) treatment. So our second target variable is coming from the question (2014-2022 waves):

- 'Have you sought treatment for a mental health condition?': 'ST' (2014)
- 'Have you ever sought treatment for a mental health issue from a mental health professional?': 'ST' (2016-2022)

Of course the word “ever” in the second version: “have you ever sought treatment …”, changes the context. If a person sought treatment 10 years ago, we should ask them about conditions at that time, not now. If we look at Figure 3, we can see that adding the word “ever” may indeed increase the share of people who sought treatment in 2016-2019 waves from 50 to 60%, but after that the share drops below 50%.



Chi2 test for independence of ‘ST’ from time (years) gives p-value way below 0.01 (see notebook). So we can ignore the word ‘ever’ effect and move on.

On average for all years 56% from all people in the survey have sought treatment, which is well balanced for modeling the seeking treatment decision.

In real life to seek treatment we need to get or at least suspect some diagnosis. Hence comes two more target variables: (1) if a person has MHC with official diagnosis - MHCD (1 if yes, 0 if no), (2) if a person has any MHC (with diagnosis or without) - MHCA. To construct these variables we need two questions:

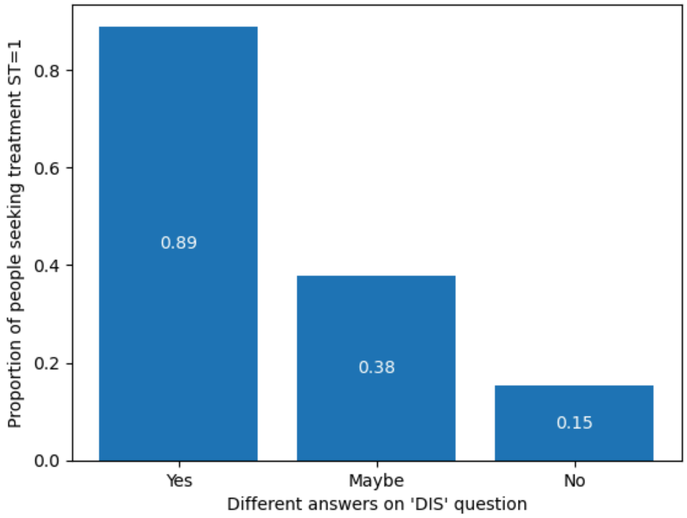
- 'Have you ever been diagnosed with a mental health disorder?': 'DIAG'
- 'Do you currently have a mental health disorder?': 'DIS'

But DIAG is a very strange variable. In 2016 it had about 50/50 of yes/no answers. In 2017 (and following years) the ratio was something like 30/1, and the share of nan’s increased from zero to 57%. It seems like lots of no’s became nan’s somehow. Something went wrong, and we can not use this variable.

Instead we use DIS to create MHCD:
MHCD = 1 if DIS is ‘Yes’
MHCD = 0 if DIS is ‘Maybe’

Figure 4 shows that in 2016 89% of people who chose in DIS ‘Yes’ had been diagnosed by a medical professional, assuming it’s 100% for the whole dataset is a strong assumption. But alternatively we need to assume that all nan's in DIAG equal to 'No'. Which is even stronger, as it is not clear what happened with DIAG after all.

Figure 4: Share of People With and Without MHC Diagnosed by a Medical Professional



Interestingly, 38% of people who answered on 'DIS' question 'Maybe' actually had been diagnosed, probably in the past, but they are not sure about their current condition, or they do not trust the diagnosis. And only 15% who answered on 'DIS' question 'No' say they were diagnosed. These might be people who recovered, but also people who decide for themselves that they do not have MHC, and this might be wrong leading to underestimation of the true MHC levels. Definitely questions should be designed after the models, not vice versa.

Next target variable (MHCA) is constructed in the similar way:
MHCA = 1 if DIS = 'Yes' or 'Maybe' or 'Possibly'
MHCA = 0 if DIS = 'Don't know'

Hence in total we have 4 target variables. They can be predicted by one DL model or we can build 4 separate regular models and connect them to make the final prediction. We choose the last option because we have limited data.

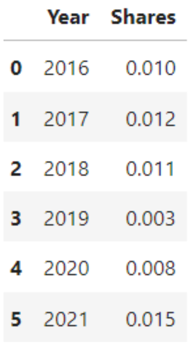
2.2. Features

For each of our 4 models we presumably need demographics (age, gender, country, race) and employer MHC friendliness features - possible interventions or components of a MHC program HR dept is considering we split into three topics: information, benefits, culture.

2.2.1. Demographics

Since 2016 the country has been asked with two questions: where do you work and where do you live? But the share of people living and working in different countries is below 2% (table 4).

Table 4: Share of People Working and Living in Different Countries



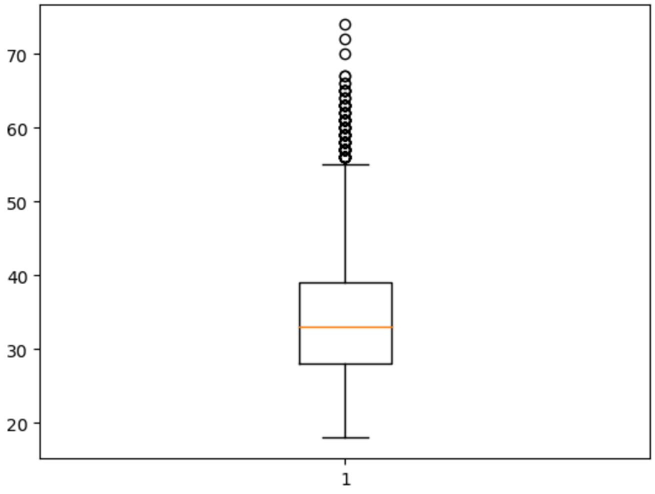
	Year	Shares
0	2016	0.010
1	2017	0.012
2	2018	0.011
3	2019	0.003
4	2020	0.008
5	2021	0.015

Hence we can use only one question, let's take the country of living. In 2014 there was one question. People from different countries might have different MHC levels, so we want to keep the country among features. But there are about 50 countries in each wave and for many countries there is only one observation. Let's aggregate **countries** to five values: 'USA (59%)', 'UK' (11%), 'Canada' (5%), 'Germany' (3%) and 'Other' (22%). Feature name: CTRY.

Race questions are missing in 2014 and 2016 waves, and starting from 2017 we can see that about 90% of respondents are white, so we can skip this feature because of low variability and too many missing values.

Age has 22 errors (higher than 75 and lower than 18) and missing values, so we replaced them with median. Boxplot shows 137 outliers for aggregated over 8 waves dataset.

Figure 5: Age Boxplot for 8 Survey Waves



Median age is 33, below the US general population median of 39.

Gender is messy due to no response options being given, hence there are up to 70 unique values in each wave. To clean it we replace all values with ‘Male’, ‘Female’ and ‘Other’ (trans, queer, not sure, nan and the like). After cleaning and filling nan’s (34) with ‘Male’ as the most frequent answer, we get in total counts: 3384 males, 1186 females and 123 others.

2.2.2. Employer MHC friendliness features

Let’s divide MHC friendliness feature or possible interventions on 3 topics:

1. better information about MHC: ‘INT1’ (information)
2. better insurance and benefits: ‘INT2’ (benefits)
3. better social attitudes towards people with MHC: ‘INT3’ (culture)

Good news is that information and benefits questions remain similar in all waves:

‘INT1’ (information):

- 'Has your employer ever discussed mental health as part of an employee wellness program?': ‘INT1_1’
- 'Does your employer provide resources to learn more about mental health issues and how to seek help?': ‘INT1_2’
- 'Do you know the options for mental health care your employer provides?': ‘INT1_3’

‘INT2’ (benefits):

- 'Does your employer provide mental health benefits?': ‘INT2_1’

Bad news is that only 4 culture questions remain the same in all waves:

‘INT3’ (culture):

- 'Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?': ‘INT3_1’
- 'How easy is it for you to take medical leave for a mental health condition?' : ‘INT3_2’
- 'Would you be willing to discuss a mental health issue with your coworkers?': ‘INT3_3’
- 'Would you be willing to discuss a mental health issue with your direct supervisor(s)?': ‘INT3_4’

In the 2016-2022 waves about 10 more culture questions appear, some old questions substantially change wording or disappear. And only 4 questions remain the same, so we can use only them if we want to have a maximum number of observations.

After unifying feature names we get 579 lines with missing values (in all 8 questions). Dropping rows with all missing values, we have only ‘INT1_3’ with 7% nan’s. We replace them randomly with True and False according to frequencies for this variable.

Encoding is tricky. Let’s take the ‘INT1_1’ question: 'Has your employer ever discussed mental health as part of an employee wellness program?'. Answer options: ‘Yes’, ‘No’, ‘Don’t know’ (and similar). We are going to model the impact of our interventions, like ‘stat to discuss mental health as part of an employee wellness program’ on people with WHC (diagnosed or not). These people are supposed to be interested in any MHC information. Yet 8% of them answered ‘Don’t know’. Probably it means ‘maybe my employer has discussed it, but not recently or not clearly’. Hence we rather replace it with ‘No’. We could also try one-hot encoding, but it is harder to interpret in the context of our model: what intervention should change the proportion of ‘Don’t know’ people?

With the same logic we replace ‘Don’t know’ with ‘No’ in ‘INT1_2’ (25%).

Question ‘INT1_3’ 'Do you know the options for mental health care your employer provides?' has an interesting interplay with the question ‘INT2_1’ 'Does your employer provide mental health benefits?' It seems that we should expect that people who are not sure about the options may only answer ‘I don’t know’ on ‘INT2_1’. But there were 87 people who answered ‘Yes’. Maybe they do not know all the options, but they know about benefits for sure. There were also 218 people who knew nothing about options, but were sure that their employer provided benefits.
To avoid these inconsistencies we drop ‘INT1_3’.

Coming back to ‘INT2_1’, we can see that 22% of respondents with MHC answered ‘Don’t know’ if their employer provided benefits. What kind of person with MHC does not check about possible benefits? Even for people with official diagnosis this share is 17%! At the same time 70% of people with MHC who don’t know about the benefits are seeking treatment, why do they never check with their employer? The only answer we could guess is that information is not openly available, and they are scared to ask and show interest, prefer solving the problem on their own. This is saying something about the corporate culture around MHC.

Hence we can create a new information variable (with a culture component in it):
‘INT1_4’ = True if ‘INT2_1’ = ‘Yes’ or ‘No’

‘INT1_4’ = False otherwise.

‘Don’t know’ in ‘INT2_1’ we replace with nan.

Question 'INT3_1' 'Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?' also have options ‘Yes’, ‘No’, ‘Don’t know’. 70% of people with MHC don’t know the answer. Practically it means that they can not fully trust the employer, because anonymity is something to be expected in all health issues. So we encode ‘Don’t know’ as False.

Question 'INT3_2' 'How easy is it for you to take medical leave for a mental health condition?' has several options, we choose following encoding: 'Very difficult': 0, 'Difficult': 1/6, 'Somewhat difficult': 1/3, 'Neither easy nor difficult': 1/2, 'Somewhat easy': 3/4, 'Very easy': 1.

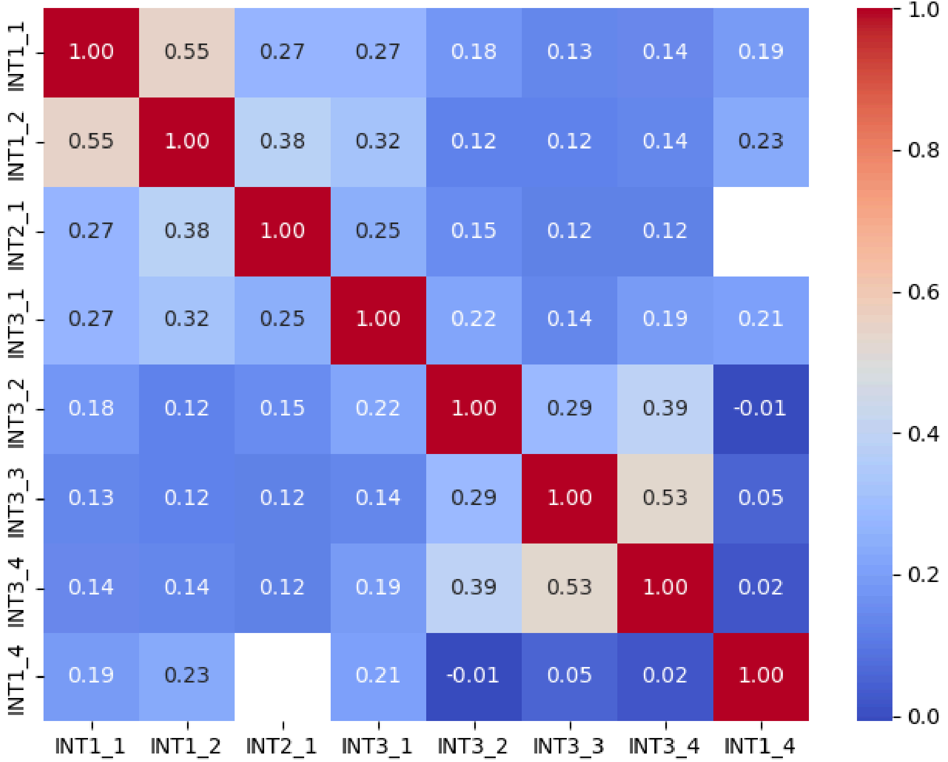
Questions ‘INT3_3’ and ‘INT3_4’ in 2014 sound like 'Would you be willing to discuss a mental health issue with your coworkers?' and the same about supervisors. Both questions had option ‘Some of them’. In 2016-2022 waves questions changed to 'Would you feel comfortable discussing a mental health disorder with your coworkers?' and 'Would you feel comfortable discussing a mental health disorder with your direct supervisor(s)?' getting the answer option ‘Maybe’ instead of ‘Some of them’.

General meaning stays the same, but options change a lot. ‘Some of them’ usually means 20-30% of possible connections. ‘Maybe’ may mean something from 30% to 70% probability. So we encode ‘Some of them’ as 0.25 and ‘Maybe’ as 0.5.

Now we can look at nan’s. There are 586 rows with all INT variables being nan. On top of that we have nan’s in ‘INT2_1’ and ‘INT3_2’ we chose to replace with predictions from Logistic and Linear regressions of corresponding variables on the rest of INT’s (see notebook). After this we only have 586 rows with all nan’s we can drop when we need INT’s.

Checking out the correlation matrix below shows that there are features with low to medium correlations, meaning it could be good to reduce the feature space down to several orthogonal features.

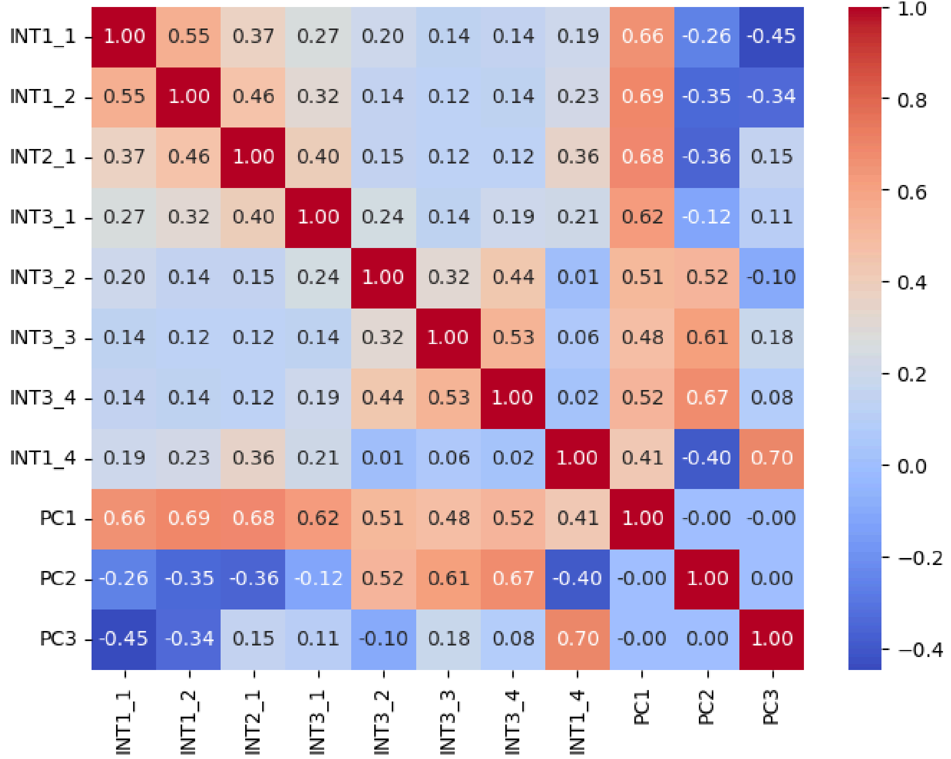
Figure 6: Rang Correlation Matrix of MHC Friendliness Features



We start with **PCA** for 3 components explaining 65% of total variance. Correlation matrix below shows that there is one vector (PC1) with 34% of variance positively correlated with all features. This means there is a strong tendency in the data points to be good or bad in all features.

Second vector (PC2) with 20% of variance corresponds to a tendency to be relatively good/bad in MHC culture and bad/good in information and benefits. And the third vector (PC3) with 11% of variance picks out INT4_1.

Figure 7. Correlation matrix between original INT features and PCA components



	INT1_1	INT1_2	INT2_1	INT3_1	INT3_2	INT3_3	INT3_4	INT1_4	PC1	PC2	PC3
INT1_1	1.00	0.55	0.37	0.27	0.20	0.14	0.14	0.19	0.66	-0.26	-0.45
INT1_2	0.55	1.00	0.46	0.32	0.14	0.12	0.14	0.23	0.69	-0.35	-0.34
INT2_1	0.37	0.46	1.00	0.40	0.15	0.12	0.12	0.36	0.68	-0.36	0.15
INT3_1	0.27	0.32	0.40	1.00	0.24	0.14	0.19	0.21	0.62	-0.12	0.11
INT3_2	0.20	0.14	0.15	0.24	1.00	0.32	0.44	0.01	0.51	0.52	-0.10
INT3_3	0.14	0.12	0.12	0.14	0.32	1.00	0.53	0.06	0.48	0.61	0.18
INT3_4	0.14	0.14	0.12	0.19	0.44	0.53	1.00	0.02	0.52	0.67	0.08
INT1_4	0.19	0.23	0.36	0.21	0.01	0.06	0.02	1.00	0.41	-0.40	0.70
PC1	0.66	0.69	0.68	0.62	0.51	0.48	0.52	0.41	1.00	-0.00	-0.00
PC2	-0.26	-0.35	-0.36	-0.12	0.52	0.61	0.67	-0.40	-0.00	1.00	0.00
PC3	-0.45	-0.34	0.15	0.11	-0.10	0.18	0.08	0.70	-0.00	0.00	1.00

Figure 8: Scatter Plot of PC1 vs PC2 with PC3 Being the Color

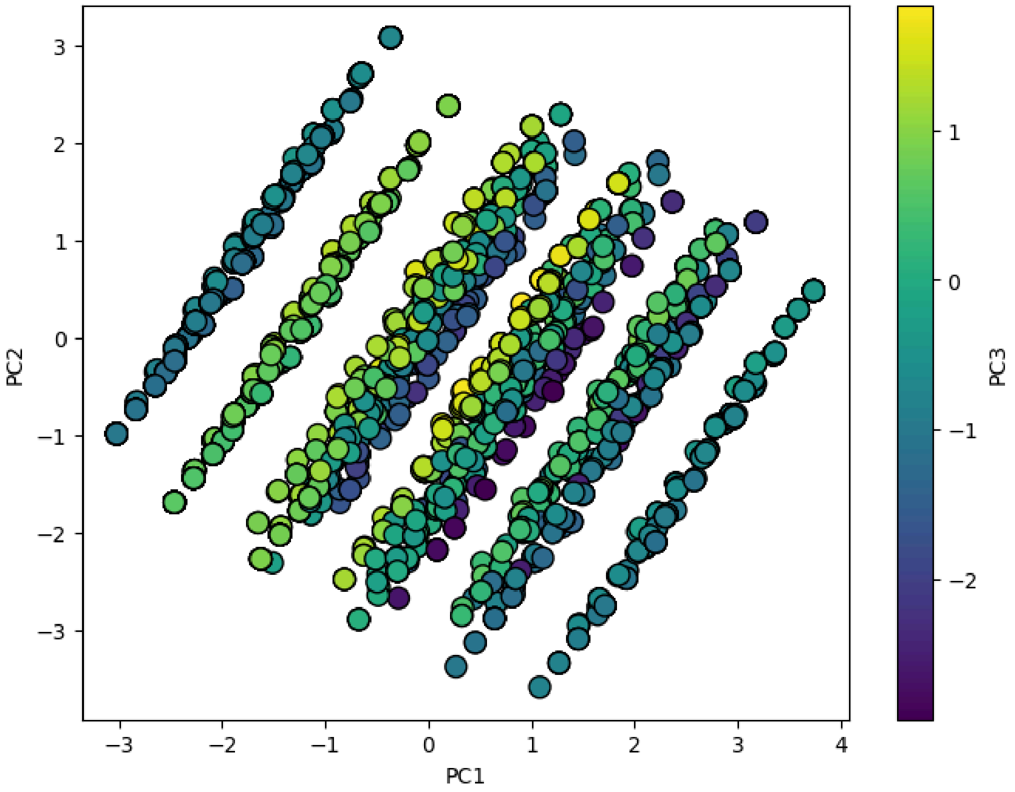
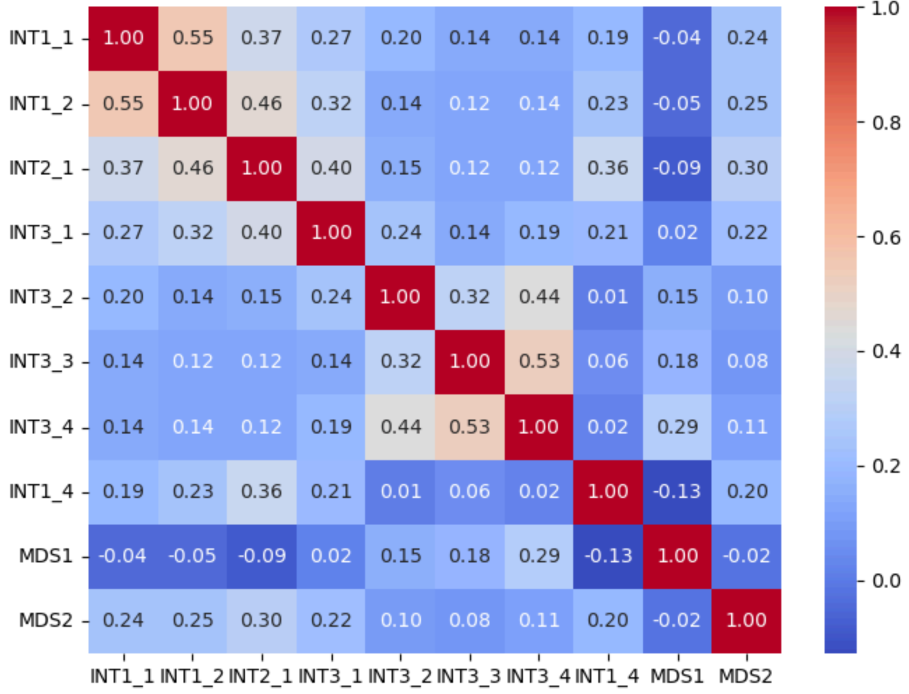


Figure 8 shows several possible clusters with very linear shapes of increasing overall MHC friendliness (PC1) with culture (PC2) within each cluster.

Multi-Dimensional Scaling (MDS) - another method to project data from multidimensional space into a few dimensions. It can preserve more complicated patterns in the data, but take significantly more time to run, than PCA. With 2 components we got explained variation - 97%.

As can be seen in Figure 9, MDS1 can be more related to culture, and DMS2 - to information and benefits.

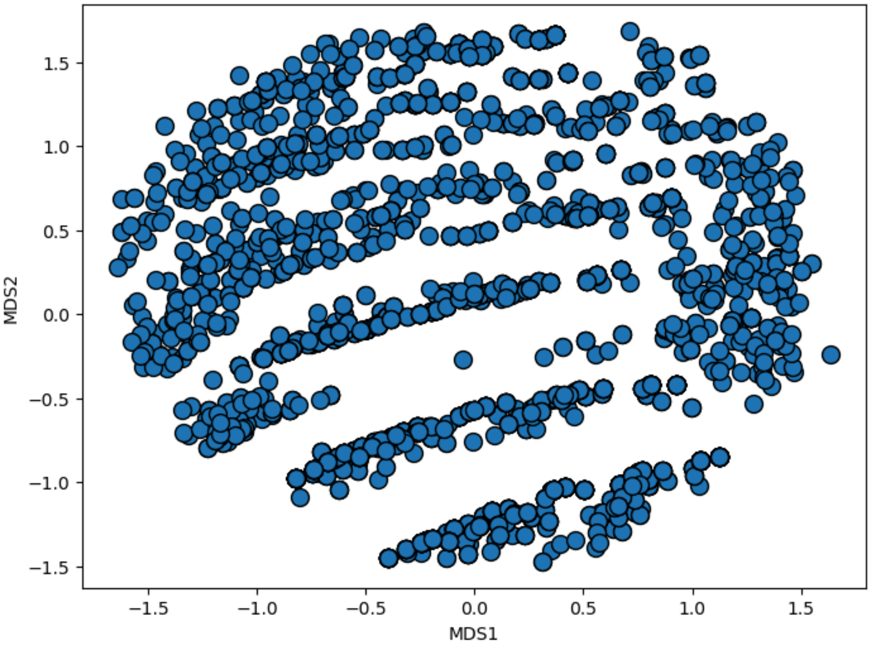
Figure 9. Correlation Matrix between Original Features and MDS Dimensions.



	INT1_1	INT1_2	INT2_1	INT3_1	INT3_2	INT3_3	INT3_4	INT1_4	MDS1	MDS2
INT1_1	1.00	0.55	0.37	0.27	0.20	0.14	0.14	0.19	-0.04	0.24
INT1_2	0.55	1.00	0.46	0.32	0.14	0.12	0.14	0.23	-0.05	0.25
INT2_1	0.37	0.46	1.00	0.40	0.15	0.12	0.12	0.36	-0.09	0.30
INT3_1	0.27	0.32	0.40	1.00	0.24	0.14	0.19	0.21	0.02	0.22
INT3_2	0.20	0.14	0.15	0.24	1.00	0.32	0.44	0.01	0.15	0.10
INT3_3	0.14	0.12	0.12	0.14	0.32	1.00	0.53	0.06	0.18	0.08
INT3_4	0.14	0.14	0.12	0.19	0.44	0.53	1.00	0.02	0.29	0.11
INT1_4	0.19	0.23	0.36	0.21	0.01	0.06	0.02	1.00	-0.13	0.20
MDS1	-0.04	-0.05	-0.09	0.02	0.15	0.18	0.29	-0.13	1.00	-0.02
MDS2	0.24	0.25	0.30	0.22	0.10	0.08	0.11	0.20	-0.02	1.00

Figure 10 shows the scatter plot of MDS1 vs MDS2. There may again be several clusters with linear and nonlinear shapes. And again culture and information with benefits tend to increase together within clusters, but not always.

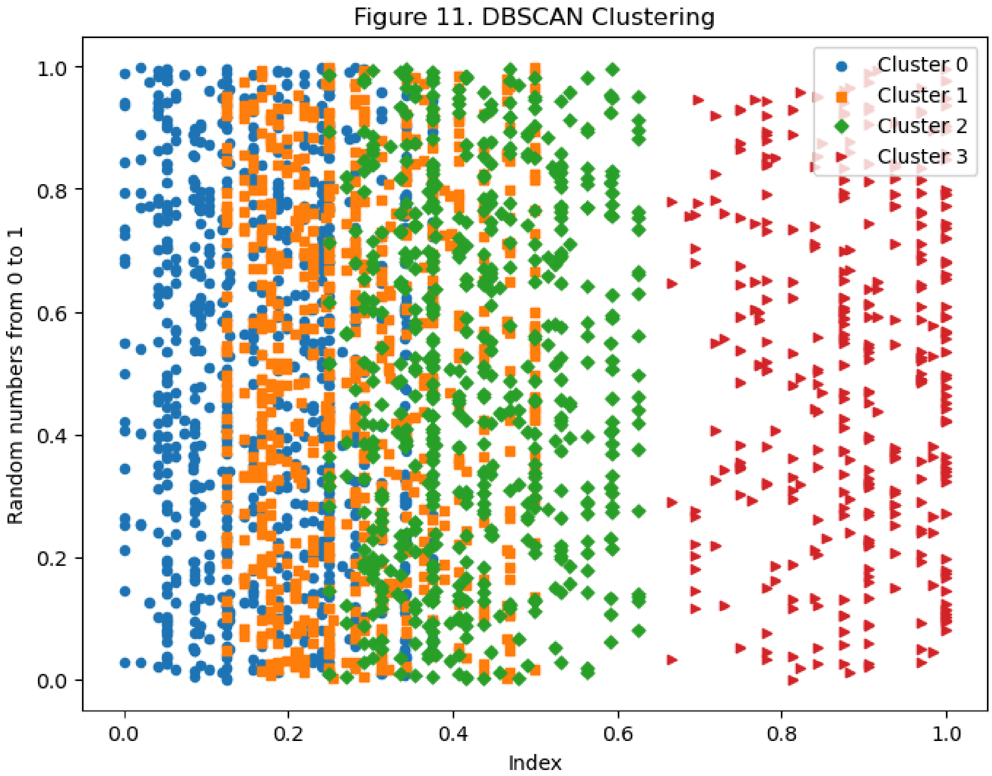
Figure 10. Scatter Plot of MDS1 vs MDS2



Let’s try a clustering algorithm to see if we can find some reasonable number of clusters. **DBSCAN** builds fuzzy clusters by collecting together close enough data points. By tuning ‘eps’ and ‘min_samples’ parameters we change distance between points and min cluster size. This changes the number of clusters.

To print a scatter plot we aggregated all the INT features into an index of MHC friendliness and for y axis used random numbers to make it viewable.

Figure 11 shows 1 cluster of champions (8% of observations) and 3 fuzzy clusters of followers (50% in total). The rest is noise.

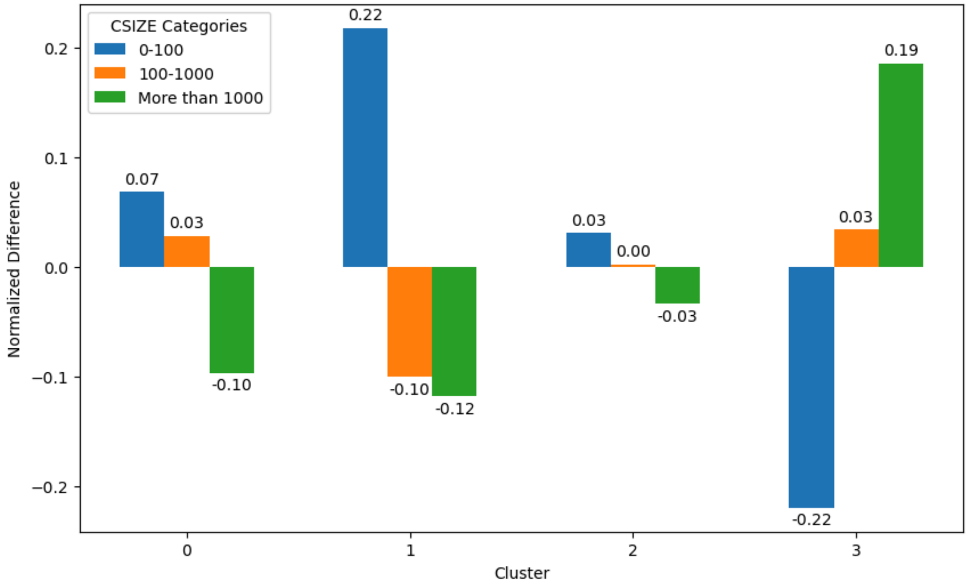


It is interesting to check what size of the company is typical for each cluster. From the 2016 wave there is a question: 'How many employees does your company or organization have?' with a range of possible answers. Figure 12 is the normalized size structure of each cluster minus average structure for the whole dataset.

Interestingly, cluster 1 and 3 are very close to the average cluster. Champions (cluster 4) have much more people from large companies and much less from the small ones. Cluster 2 is the mirror of cluster 4, but it’s not the worst in the MHC friendliness index.

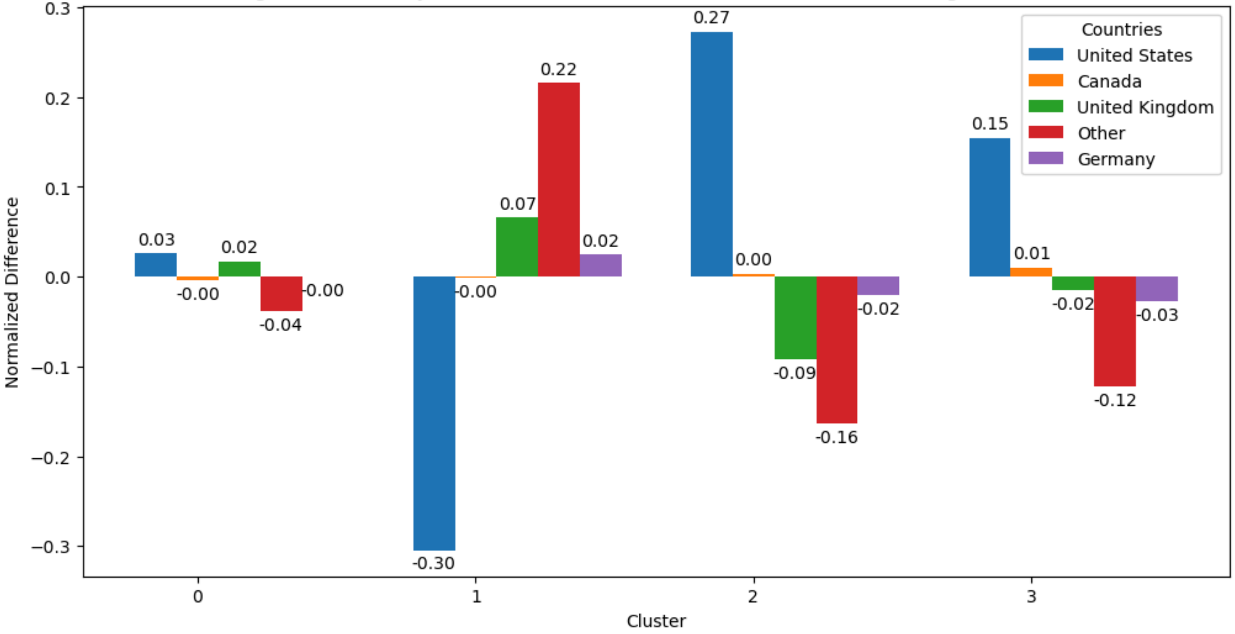
So being large probably helps to invest more into MHC friendliness, although it is also possible for medium and even small companies to have a decent level of friendliness. At the same time cluster with worst companies in MHC friendliness has close to average company size structure.

Figure 12: Company Size Structures for Different Clusters Relative to Average Structure



Another characteristic is county. In Figure 13 we can see that the zero cluster is country average, first is very skewed into Other countries, second and third - into USA. It is possible that legal and cultural norms in the USA are in favor of MHC friendliness, at least in big and medium companies.

Figure 13: Country Structures for Different Clusters Relative to Average Structure



3. Modeling

To assemble our model predicting the productivity effect of MHC interventions we need to build 4 smaller supervised regular (non-DL) models for our 4 target variables:

1. MHCA (if the person has MHC),
2. MHCD (if the person with MHC get diagnosis or not),
3. ST (if the person is seeking treatment having MHC),
4. ProD (treatment effect on productivity).

Each model we make first in linear form with Linear or Logistic Regression and second in nonlinear form with XGBoost. Then we compare metrics and choose the best model for deployment.

3.1. Modeling MHCA and MHCD

We will model MHCA by demographics (age, gender, country) only. Of course, the psychological environment in a company may affect the prevalence of MHC's, but we consider these effects second order to demographics and much more complex.

Gender and country are one-hot encoded and one of the vectors dropped (Female and Canada) to avoid multicollinearity.

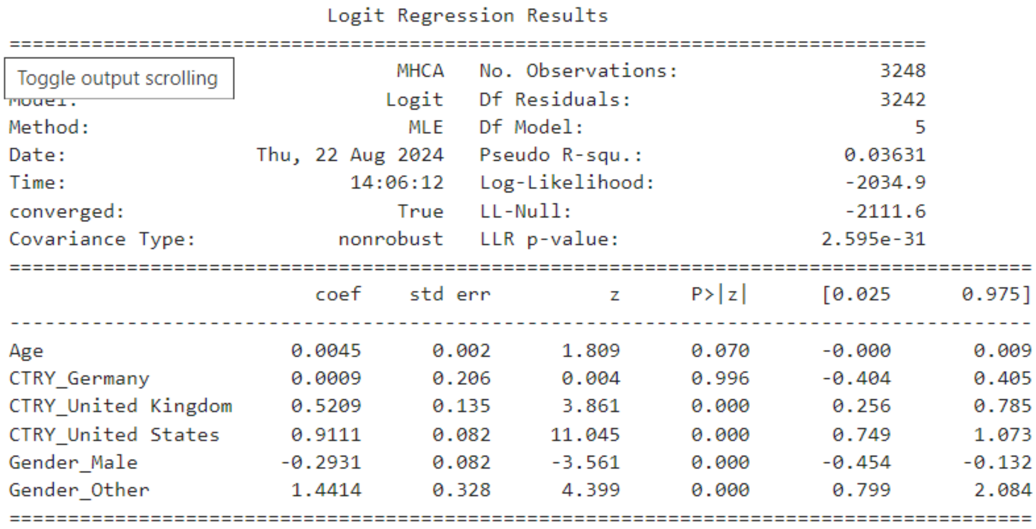
Chi2 test shows that MHCA is significantly not independent from 6 features, meaning there are significantly different MHCA means for different ages, genders and countries.

We use EFS search to find a combination of features minimizing log_loss function. This metric measures the accuracy of the predicted probabilities, which is most important for us because we need to predict the share of population with MHC, not MHC status of particular individuals (in which case we could target just accuracy or similar).

The best Logit model includes 7 features. Accuracy on the test set is 0.6523 and Log Loss - 0.6110.

To see the signs of the feature effects we can build Logit from `statsmodels`, results are in Table 5. As can be seen, MHC level is higher in the USA and UK, lower among males and much higher among people with other genders, than among females.

Table 5: Logistic Regression Using Best Features to Predict MHCA



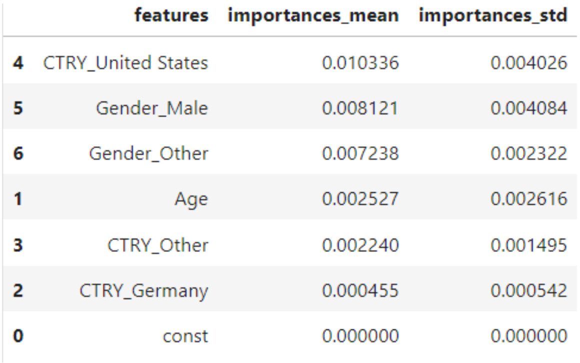
```
                           Logit Regression Results
==============================================================================
                                  MHCA   No. Observations:                 3248
Model:                           Logit   Df Residuals:                     3242
Method:                            MLE   Df Model:                            5
Date:                 Thu, 22 Aug 2024   Pseudo R-squ.:                 0.03631
Time:                         14:06:12   Log-Likelihood:                -2034.9
converged:                        True   LL-Null:                       -2111.6
Covariance Type:             nonrobust   LLR p-value:                 2.595e-31
=========================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Age                   0.0045      0.002      1.809      0.070      -0.000       0.009
CTRY_Germany          0.0009      0.206      0.004      0.996      -0.404       0.405
CTRY_United Kingdom   0.5209      0.135      3.861      0.000       0.256       0.785
CTRY_United States    0.9111      0.082     11.045      0.000       0.749       1.073
Gender_Male          -0.2931      0.082     -3.561      0.000      -0.454      -0.132
Gender_Other          1.4414      0.328      4.399      0.000       0.799       2.084
=========================================================================================
```

For **XGBoost** we divide the training set into small training and validation sets. Small training set is used for parameters grid search, and validation - for permutation of features to skip features reducing the score of the model. Score is set to neg_log_loss.

On the test set we get accuracy 0.6508 and Log Loss: 0.6163 which are slightly worse than for the Logit model.

All features have a positive effect on score (see Table 6).

Table 6: The Importance of Features by Their Effect on Total Log Loss in XGBoost MHCA Model



	features	importances_mean	importances_std
4	CTRY_United States	0.010336	0.004026
5	Gender_Male	0.008121	0.004084
6	Gender_Other	0.007238	0.002322
1	Age	0.002527	0.002616
3	CTRY_Other	0.002240	0.001495
2	CTRY_Germany	0.000455	0.000542
0	const	0.000000	0.000000

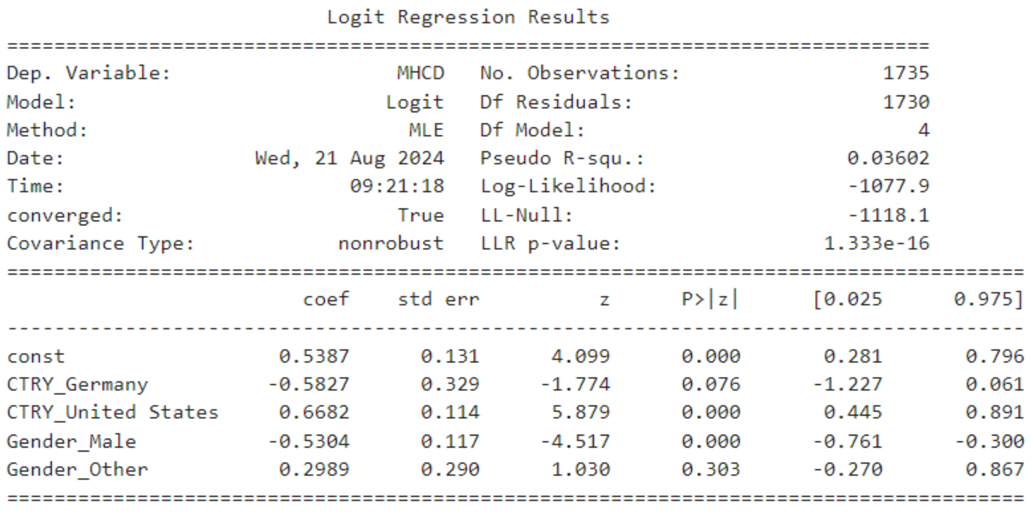
Based on formal criteria of accuracy and Log Loss we should choose the Logit model for the deployment.

And before we move on, we need to estimate the correction coefficient for MHCA, as it is much higher in the dataset than generally in the US population, probably due to self-selection of people participating in the poll.

All we need is to supply the model with calibration data point:
USA = 1, all other countries = 0, Age = 39 (median age in USA), Gender_Male = .49, Gender_Other = 0.0036. Logit predicts 70% and XGBoost 74.8%. As we know that MHCA for the US is 22.8%, hence we get correction coefficients 3.09 and 3.28. Taking average we get 3.2. Meaning the dataset MHCA is 3.2 times higher than it would be if the sampling was random, without self-selection.

For **MCHD** we do the same way Logit and XGBoost models and the same features. Logit delivered Accuracy: 0.6571 and Log Loss: 0.6317.

Table 7: Logistic Regression Using Best Features to Predict MHCD



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                   MHCD   No. Observations:                 1735
Model:                          Logit   Df Residuals:                     1730
Method:                           MLE   Df Model:                            4
Date:                Wed, 21 Aug 2024   Pseudo R-squ.:                 0.03602
Time:                        09:21:18   Log-Likelihood:                -1077.9
converged:                       True   LL-Null:                       -1118.1
Covariance Type:            nonrobust   LLR p-value:                 1.333e-16
======================================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                  0.5387      0.131      4.099      0.000       0.281       0.796
CTRY_Germany          -0.5827      0.329     -1.774      0.076      -1.227       0.061
CTRY_United States     0.6682      0.114      5.879      0.000       0.445       0.891
Gender_Male           -0.5304      0.117     -4.517      0.000      -0.761      -0.300
Gender_Other           0.2989      0.290      1.030      0.303      -0.270       0.867
======================================================================================
```

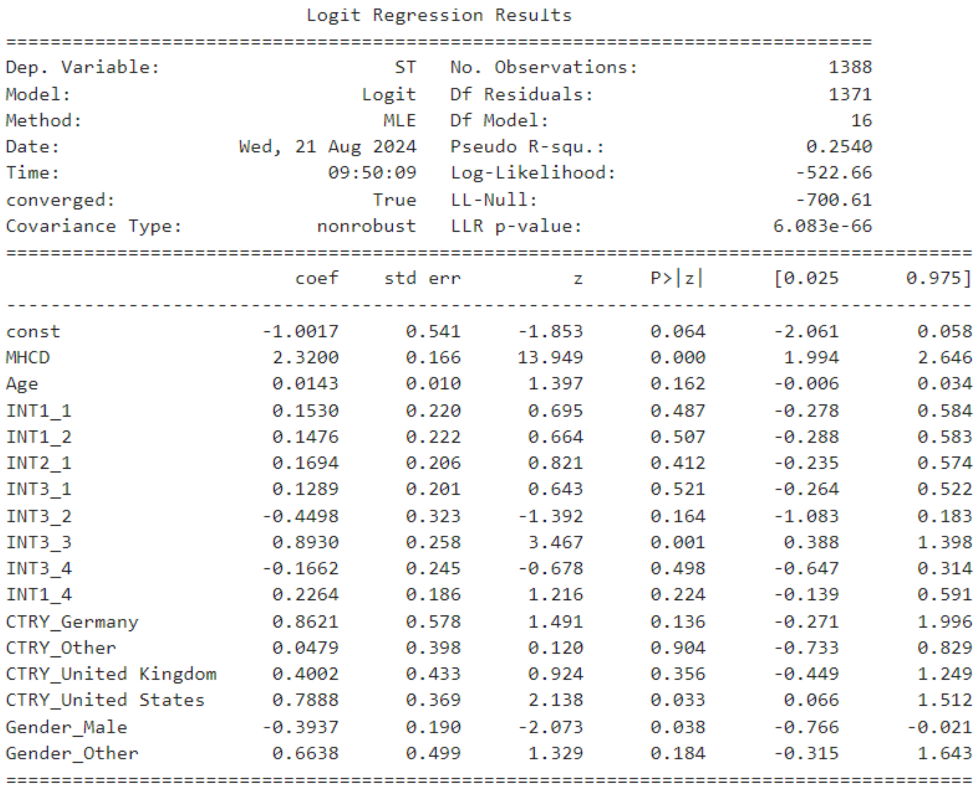
Probability of getting a diagnosis while having MHC is higher for people from the USA and lower for males.

XGBoost this time is slightly worse than Logit in accuracy 0.6484 vs 0.6571, but better in Log Loss: 0.6307 vs 0.6317. So we keep XGBoost for deployment. Most important features in XGBoost are: ‘CTRY_United States’, ‘Gender_Male’ and ‘Gender_Other’.

3.2. Modeling decision to seek treatment

Now we add our intervention features and MCHD as a feature to the working dataset. This increases the number of possible features to 16, time to run feature selection becomes too long. Hence we start with building Logistic Regression with all features using `statsmodels` (Table 8).

Table 8: Logistic Regression Using All Available Features to Predict ST

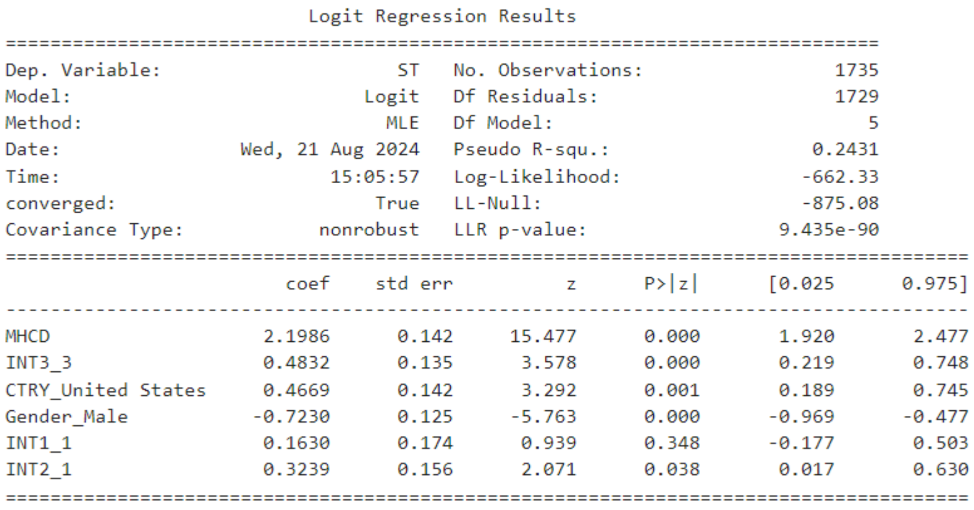


```
                           Logit Regression Results
==============================================================================
Dep. Variable:                     ST   No. Observations:                 1388
Model:                          Logit   Df Residuals:                     1371
Method:                           MLE   Df Model:                           16
Date:                Wed, 21 Aug 2024   Pseudo R-squ.:                  0.2540
Time:                        09:50:09   Log-Likelihood:                -522.66
converged:                       True   LL-Null:                       -700.61
Covariance Type:            nonrobust   LLR p-value:                 6.083e-66
========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                   -1.0017      0.541     -1.853      0.064      -2.061       0.058
MHCD                     2.3200      0.166     13.949      0.000       1.994       2.646
Age                      0.0143      0.010      1.397      0.162      -0.006       0.034
INT1_1                   0.1530      0.220      0.695      0.487      -0.278       0.584
INT1_2                   0.1476      0.222      0.664      0.507      -0.288       0.583
INT2_1                   0.1694      0.206      0.821      0.412      -0.235       0.574
INT3_1                   0.1289      0.201      0.643      0.521      -0.264       0.522
INT3_2                  -0.4498      0.323     -1.392      0.164      -1.083       0.183
INT3_3                   0.8930      0.258      3.467      0.001       0.388       1.398
INT3_4                  -0.1662      0.245     -0.678      0.498      -0.647       0.314
INT1_4                   0.2264      0.186      1.216      0.224      -0.139       0.591
CTRY_Germany             0.8621      0.578      1.491      0.136      -0.271       1.996
CTRY_Other               0.0479      0.398      0.120      0.904      -0.733       0.829
CTRY_United Kingdom      0.4002      0.433      0.924      0.356      -0.449       1.249
CTRY_United States       0.7888      0.369      2.138      0.033       0.066       1.512
Gender_Male             -0.3937      0.190     -2.073      0.038      -0.766      -0.021
Gender_Other             0.6638      0.499      1.329      0.184      -0.315       1.643
========================================================================================
```

Then we choose the most significant 9 features and run the EFS feature selection process, drop the non-best features and add new ones up to 9 pcs until we converge to the best set.

Best model on the test dataset delivered accuracy 0.8329 and Log Loss: 0.3743. Feature effects can be seen in Table 9.

Table 9: Logistic Regression Using All Best Features to Predict ST



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                     ST   No. Observations:                 1735
Model:                          Logit   Df Residuals:                     1729
Method:                           MLE   Df Model:                            5
Date:                Wed, 21 Aug 2024   Pseudo R-squ.:                  0.2431
Time:                        15:05:57   Log-Likelihood:                -662.33
converged:                       True   LL-Null:                       -875.08
Covariance Type:            nonrobust   LLR p-value:                 9.435e-90
======================================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------
MHCD                   2.1986      0.142     15.477      0.000       1.920       2.477
INT3_3                 0.4832      0.135      3.578      0.000       0.219       0.748
CTRY_United States     0.4669      0.142      3.292      0.001       0.189       0.745
Gender_Male           -0.7230      0.125     -5.763      0.000      -0.969      -0.477
INT1_1                 0.1630      0.174      0.939      0.348      -0.177       0.503
INT2_1                 0.3239      0.156      2.071      0.038       0.017       0.630
======================================================================================
```

As can be seen from Table 9, having a diagnosis (MHCD=1) greatly increases the probability to seek treatment. Also Americans seek treatment more often than people from other countries. Males seek less often than females and other genders.

Two interventions have significant positive impact:

- 'Would you be willing to discuss a mental health issue with your coworkers?': 'INT3_3'
- 'Does your employer provide mental health benefits?': 'INT2_1'

Interestingly, the fact of benefits has relatively low coefficient, meaning that benefits are not that important compared to being accepted by coworkers. Effect of benefits might be more precise if we asked about their size.

XGBoost shows worse accuracy: 0.8213 and log loss: 0.3779. So we keep the Logit model for deployment.

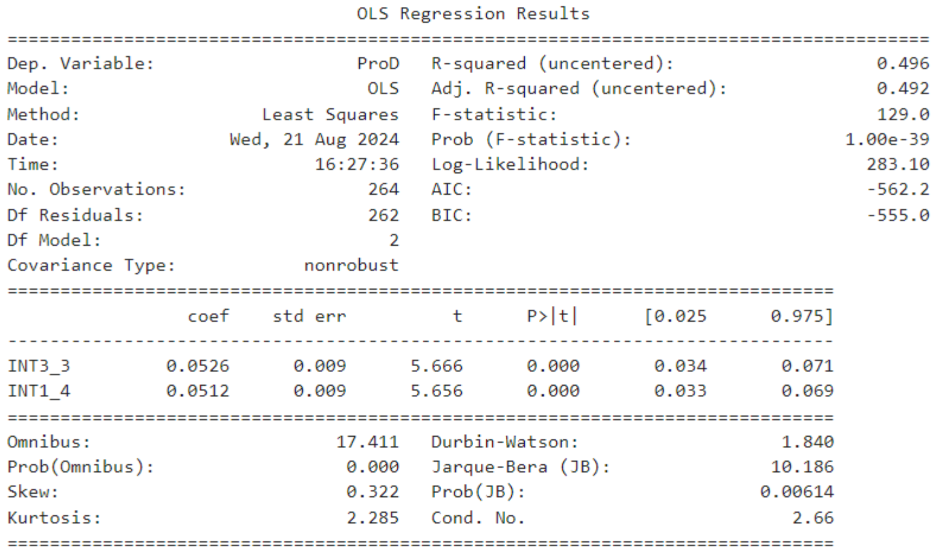
3.3. Productivity among people with MHC

For modeling productivity we switch to OLS and XG Boost Regressions with the aim to minimize MSE. We also build two models: (1) for people with diagnosis MHCD =1, and (2) people without MHCD = 0. We filter out people who have not sought treatment.

3.3.1. Productivity model for people without diagnosis

Total number of such observations is 264. ANOVA test shows only one significant feature. Little bit more we get after running EFS, see Table 10.

Table 10: OLS Regression Using Best Features to Predict ProD for People Without Diagnosis



```
                            OLS Regression Results
=======================================================================================
Dep. Variable:                   ProD   R-squared (uncentered):                   0.496
Model:                            OLS   Adj. R-squared (uncentered):              0.492
Method:                 Least Squares   F-statistic:                              129.0
Date:                Wed, 21 Aug 2024   Prob (F-statistic):                    1.00e-39
Time:                        16:27:36   Log-Likelihood:                          283.10
No. Observations:                 264   AIC:                                     -562.2
Df Residuals:                     262   BIC:                                     -555.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
INT3_3         0.0526      0.009      5.666      0.000       0.034       0.071
INT1_4         0.0512      0.009      5.656      0.000       0.033       0.069
==============================================================================
Omnibus:                       17.411   Durbin-Watson:                   1.840
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               10.186
Skew:                           0.322   Prob(JB):                      0.00614
Kurtosis:                       2.285   Cond. No.                         2.66
==============================================================================
```

Two significant interventions are:

- 'Would you be willing to discuss a mental health issue with your coworkers?': 'INT3_3'
- Information and cultural intervention ‘INT1_4’ = True if ‘INT2_1’ ('Does your employer provide mental health benefits?') = ‘Yes’ or ‘No’. Basically True means that both information is delivered well and employees are fine with requesting such information if needed.

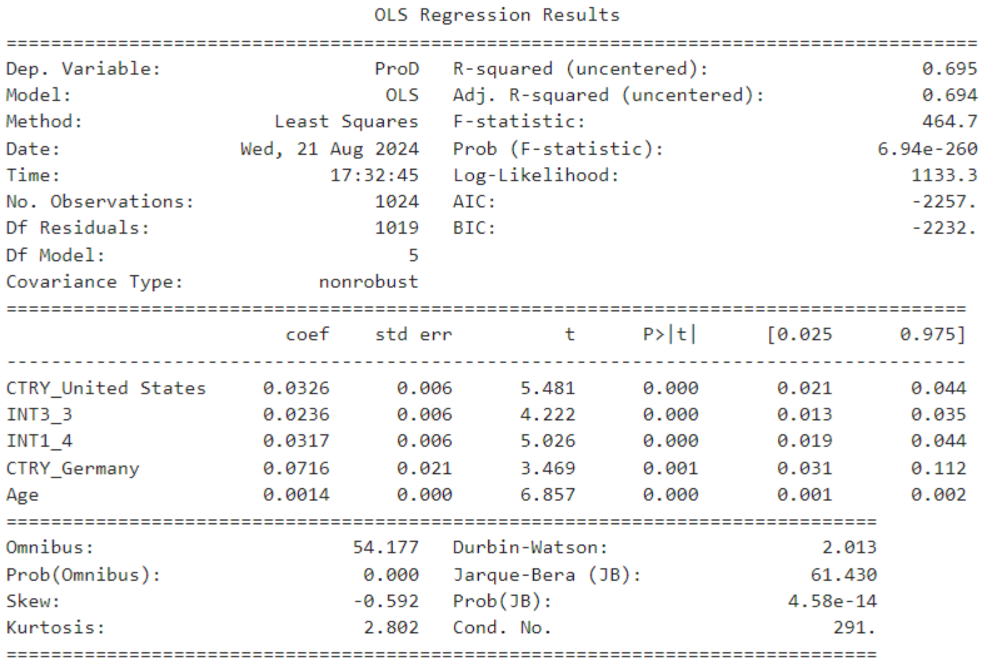
Both features have rather big effects. As it is mentioned in section 1.1 people without diagnosis taking treatment may expect back on average 7% out of 16% of their productivity loss due to MHC. And the two features above may provide together 10%.

XGBoost finds three important features: ‘INT1_4’, ‘CTRY_United States’ and ‘CTRY_United Kingdom’. OLS delivers MSE 0.0064 and XGBoost 0.0061, but we keep the OLS as it is better at forecasting continuous variables in the sparse regions.

3.3.2. Productivity model for people with diagnosis

Total number of such observations is 1024. ANOVA test shows no significant features. After running EFS we can see a much more interesting picture (Table 11).

Table 11: OLS Regression Using Best Features to Predict ProD for People With Diagnosis



```
                                 OLS Regression Results                                
=======================================================================================
Dep. Variable:                   ProD   R-squared (uncentered):                   0.695
Model:                            OLS   Adj. R-squared (uncentered):              0.694
Method:                 Least Squares   F-statistic:                              464.7
Date:                Wed, 21 Aug 2024   Prob (F-statistic):                   6.94e-260
Time:                        17:32:45   Log-Likelihood:                          1133.3
No. Observations:                1024   AIC:                                     -2257.
Df Residuals:                    1019   BIC:                                     -2232.
Df Model:                           5                                                  
Covariance Type:            nonrobust                                                  
======================================================================================
                          coef    std err          t      P>|t|      [0.025     0.975]
--------------------------------------------------------------------------------------
CTRY_United States      0.0326      0.006      5.481      0.000       0.021      0.044
INT3_3                  0.0236      0.006      4.222      0.000       0.013      0.035
INT1_4                  0.0317      0.006      5.026      0.000       0.019      0.044
CTRY_Germany            0.0716      0.021      3.469      0.001       0.031      0.112
Age                     0.0014      0.000      6.857      0.000       0.001      0.002
==============================================================================
Omnibus:                       54.177   Durbin-Watson:                   2.013
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               61.430
Skew:                          -0.592   Prob(JB):                     4.58e-14
Kurtosis:                       2.802   Cond. No.                         291.
==============================================================================
```

Again ‘INT3_3’ and ‘INT1_4’ are significant, but not that strong. As it is mentioned in section 1.1 people with diagnosis taking treatment may expect back on average 12% out of 21% of their productivity loss due to MHC. Two mentioned features may deliver 5%. Living in Germany delivers 7%, meaning the treatment effect all else equal on average in Germany is bigger than in other countries. As well as it is in the USA on 3%. Age is also positive and significant, older people somehow get more out of treatment.

Test MSE is 0.00556, higher than XGBoost 0.00577, so we keep Logit for deployment.

4. Deployment

Our initial goal was to build a model predicting the effect of some MHC programs on company productivity. Hence the input is a vector of features values representing the current company demographics, MHC information, benefits and culture. Demographics we know, intervention features we can estimate from some kind of a poll or by experts in HR dept.

And we need a vector representing values we expect to be reached after the program is implemented and its effect is realized. This can be done only by experts from HR dept.

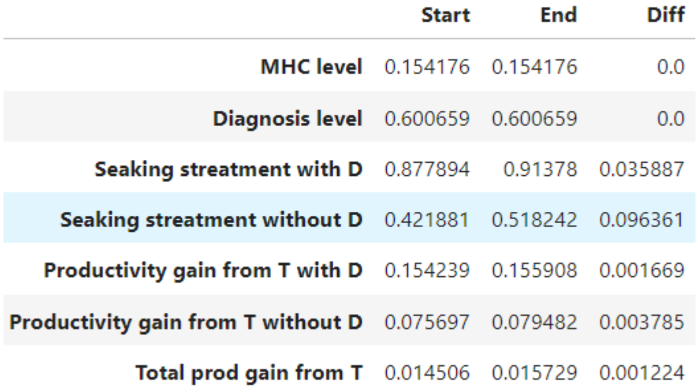
For example, we start with a company in Germany, 70% male and 1% of other genders and median age of 33. These will not change. Suppose also that our company is average in intervention features for Germany, see Table 12.

Let there be two possible programs on the agenda. Case 1 is to improve MHC information delivery to employees, increase MHC benefits, modestly improve anonymity protection, keep the same medical leave options, slightly improve MHC openness with coworkers and significantly with supervisors. Case 2 is the cost redistribution scenario: decrease benefits, keep the same medical leave, but invest more in corporate culture. Possible values of intervention features for both cases are in Table 12. The outcomes are in Tables 13 and 14.

Table 12: Current and Target States for Intervention Features

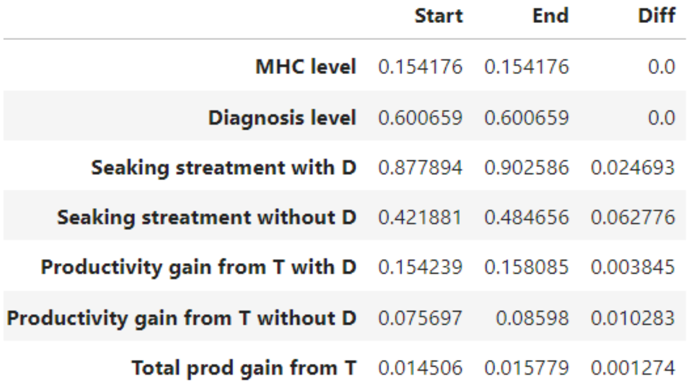
Feature	Current state	New state Case 1	New state Case 2
'Has your employer ever discussed mental health as part of an employee wellness program?': ‘INT1_1’	0.175573	0.9	0.9
'Does your employer provide resources to learn more about mental health issues and how to seek help?': ‘INT1_2’	0.137405	0.9	0.9
'Does your employer provide mental health benefits?': ‘INT2_1’	0.335878	0.9	0.1
'Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?': ‘INT3_1’	0.290076	0.5	0.7
'How easy is it for you to take medical leave for a mental health condition?' : ‘INT3_2’	0.664436	0.67	0.67
'Would you be willing to discuss a mental health issue with your coworkers?': ‘INT3_3’	0.437023	0.5	0.7
'Would you be willing to discuss a mental health issue with your direct supervisor(s)?': ‘INT3_4’	0.473282	0.7	0.7
‘INT1_4’ = True if ‘INT2_1’ = ‘Yes’ or ‘No’, False otherwise.	0.679389	0.8	0.8

Table 13: Case 1 Results



	Start	End	Diff
MHC level	0.154176	0.154176	0.0
Diagnosis level	0.600659	0.600659	0.0
Seaking streatment with D	0.877894	0.91378	0.035887
Seaking streatment without D	0.421881	0.518242	0.096361
Productivity gain from T with D	0.154239	0.155908	0.001669
Productivity gain from T without D	0.075697	0.079482	0.003785
Total prod gain from T	0.014506	0.015729	0.001224

Table 13: Case 2 Results



	Start	End	Diff
MHC level	0.154176	0.154176	0.0
Diagnosis level	0.600659	0.600659	0.0
Seaking streatment with D	0.877894	0.902586	0.024693
Seaking streatment without D	0.421881	0.484656	0.062776
Productivity gain from T with D	0.154239	0.158085	0.003845
Productivity gain from T without D	0.075697	0.08598	0.010283
Total prod gain from T	0.014506	0.015779	0.001274

As can be seen from both tables both cases lead to about the same result in total productivity gain from treatment of MHC - 0.12%. For a company with annual turnover 10M EUR such productivity increase would generate up to 12K EUR. This can be the maximum MHC program budget in the case 1. In case 2 reduction of MHC benefits would generate savings the company could use to increase the program budget.

From experimenting with the model we can find that only 4 out of 8 INT features do affect final productivity gain. ‘INT1_2’ is a redundant feature when ‘INT1_1’ is present. Indeed in Figure 6 we find that correlation of these two features is 55%.

Another redundant feature is not that obvious: ‘INT3_1’ 'Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?'. It does not mean we can stop protecting anonymity, but in this model it has no effect on productivity gain from taking treatment from MHC.

'INT3_2' 'How easy is it for you to take medical leave for a mental health condition?' is also redundant. Easiness of medical leave may have positive and negative effects which on average may cancel each other. For example if it’s easy to take a leave, there might be less motivation to take treatment. But on the other hand, if it’s hard, people might have no time to take care of their conditions.

And 'INT3_4' 'Would you be willing to discuss a mental health issue with your direct supervisor(s)?' is also redundant in the presence of ‘INT3_3’ due to high correlation, see Figure 6. At the same time we know that increasing vertical trust is as important as horizontal, and in some sense for a company management it is easier to control. And 'INT3_3' has the biggest effect on productivity, if it is dropped to zero in case 2, productivity gain decreases by a huge 0.2% points, and becomes negative.

Though technically speaking 4 redundant features can be dropped or merged with their high correlation neighbors, it is important to keep them in mind while designing an effective MHC program.

Conclusion

The goal of this project was to build a model for prediction of productivity gain from an MHC program HR dept is willing to design. We have used data from “OSMI Mental Health in Tech Survey” 2014-2022 waves to build the model and investigate two possible cases for an MHC program.

For this project we used Logit and OLS regressions as well as XGBoost models to make a four step prediction process:

1. Modeling MHC level in the company.
2. Modeling share of people with official diagnosis.
3. Modeling share of people seeking treatment.
4. Modeling productivity gain.

We used the EFS process and feature permutations to select the best features for each model. As well as grid search for the best parameters of XGBoost. We applied PCA and MDS to see that reduction of feature space can be meaningful for at least visualization of the patterns we have in the data. We did not build our model on reduced feature space, although at the end we found that 4 intervention features out of 8 turned out to be statistically redundant. As a next step we could re-build our models in the reduced space and see if we get better metrics.

We also used DBSCAN to produce clusters of companies in intervention feature space to find out that more friendly to MHC companies usually are bigger and more often located in the USA.

The final model can be used for productivity gain forecasts based on different MHC program designs.

Other key findings of the project:

1. MHC prevalence in IT is probably 3.2 times lower than it is in the dataset.
2. Companies can support people with MHC with better information, MHC benefits and corporate culture change.
3. Such interventions may impact the share of people seeking treatment and also their productivity loss due to MHC. The impact can be around 0.1-0.5% of total company productivity.
4. People with official MHC diagnosis probably experience 21% productivity loss if MHC is not treated. Treatment on average returns 12% of productivity back. Not full 21%.
5. For people without diagnosis average productivity loss is 16% if untreated and 9% if treated, hence productivity gain from treatment is only 7%.

References

Adler, D. A., McLaughlin, T. J., Rogers, W. H., Chang, H., Lapitsky, L., & Lerner, D. (2006). *Severity of depression and magnitude of productivity loss*. Psychosomatic Medicine, 68(4), 586-593. https://doi.org/10.1097/01.psy.0000221272.33234.21

National Alliance on Mental Illness. (n.d.). *Mental health by the numbers*. https://www.nami.org/About-Mental-Illness/Mental-Health-By-the-Numbers

National Institute of Mental Health. (n.d.). *Any anxiety disorder*. U.S. Department of Health and Human Services. https://www.nimh.nih.gov/health/statistics/any-anxiety-disorder

National Institute of Mental Health. (n.d.). *Attention deficit hyperactivity disorder*. U.S. Department of Health and Human Services. https://www.nimh.nih.gov/health/statistics/attention-deficit-hyperactivity-disorder-adhd

National Institute of Mental Health. (n.d.). *Any mood disorder*. U.S. Department of Health and Human Services. https://www.nimh.nih.gov/health/statistics/any-mood-disorder

National Institute of Mental Health. (n.d.). *Mental illness*. U.S. Department of Health and Human Services. https://www.nimh.nih.gov/health/statistics/mental-illness

National Institute of Mental Health. (n.d.). *Post traumatic stress disorder*. U.S. Department of Health and Human Services. https://www.nimh.nih.gov/health/statistics/post-traumatic-stress-disorder-ptsd

Substance Abuse and Mental Health Services Administration. (n.d.). *Criminal and juvenile justice*. U.S. Department of Health and Human Services. https://www.samhsa.gov/criminal-juvenile-justice/about

Tables and Figures