Mental Health in the Tech Workplace: What are the strongest predictors of mental health illness or certain attitudes towards mental health in the tech workplace?

Authors: Modupe Aggreh, Tobi Akinsiku, Adanna Ewuzie, Felicia Hessings, Wendy Jones, Davina McLaverty, Wafula Erick Mugoma, Victory Obed, Elizabeth Oduala
Group Number: Group 1

Introduction

According to the Mental Health Foundation one in four people in the UK will experience a mental health problem in any given year (Fundamental Facts About Mental Health, 2015). In the UK, mental health problems are responsible for the largest burden of disease compared to cancer and heart disease each representing 16% of the total disease burden, and the estimated costs of mental health problems are around £70-100 billion each year accounting for 4.5% of the GDP (Fundamental Facts About Mental Health, 2015). Due to the increasing demand for mental health services, many organisations are beginning to realise the importance of mental health in the workplace and maintaining the wellbeing of its employees. Many companies around the world are establishing that mental wellbeing is an essential component of a healthy and effective workplace, particularly in fast-paced and high-growth sectors of the economy.

The aim of the project was to determine what are the strongest predictors of mental health in the workplace by using various data science techniques to answer the following question 'What are the strongest predictors of mental health illness or certain attitudes towards mental health in the workplace?'. The survey was filled out by respondents who suffer from mental health disorders (diagnosed or un-diagnosed by medics) and work in tech companies. The data was collected by OSMI (Open Sourcing Mental Illness) a not-for-profit organisation dedicated to raising awareness, educating, and providing resources to support mental wellness in the tech industry (Open Sourcing Mental Illness, 2014).

Data Utility Framework

Firstly, the dataset utilised in this report was evaluated against the Health Data Research UK (HDR UK) Utility Framework. Within the framework the dataset is evaluated against various characteristics and then ranked from Bronze to Platinum across 5 categories.

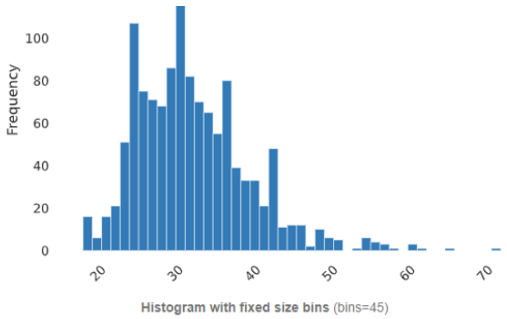
The data was obtained from OSMI who ran the largest survey exploring mental health within the tech industry in 2014. The survey received over 1200 responses collated within Google Forms which resulted in the dataset scoring very highly in the access and provision category. The data was collected in real time with 91.9% of the responses received within 1 week of the survey opening. Due to the digitalised format of the dataset, the original raw data is easily accessible on the OSMI website, and the data is covered by a Creative Commons Attribution License which allows free adapting and sharing of the survey results. The data opens opportunities to study trends over 8 years, as the study is repeated every year between 2014 – 2021. Limited information is provided about the technical quality of the data, but the dataset is supported with comprehensive data documentation. The data would have benefited from collating information regarding the type of tech companies the participants worked at – this could open the avenue to explore differences in the attitudes to mental health of healthcare tech companies in comparison to financial tech companies for example.

Data Exploration

Pandas-profiling is an open-source Python module used to generate interactive analytical reports based on a given dataset (Brugman, 2020). The module was initially used to examine our dataset (see below for an overview of the variables) to give us some insight and clarity into the data we were working with. We chose pandas-profiling instead of other platforms such as WhiteRabbit due to the team's greater familiarity of Python over Java (the programming language used for the WhiteRabbit application). It provided us with a variety of information, including the number of missing values within each feature. This helped to narrow down the categories of interest, by choosing the variables which were the most informative, and contained the least missing data.

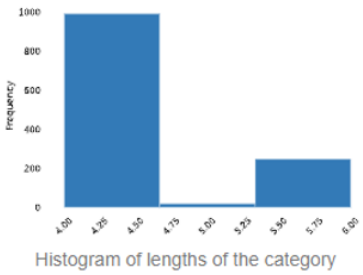
Pandas Profiling - Overview

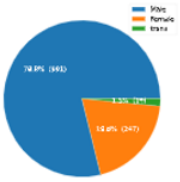
Age



Gender

Value	Count	Frequency (%)
Male	991	78.8%
Female	247	19.6%
trans	19	1.5%





Family History of Mental Health

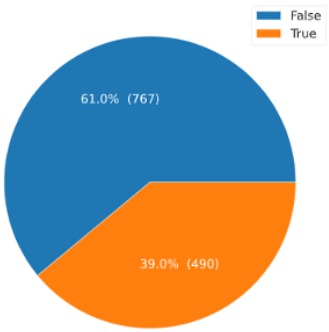


Figure 1. Pandas profiling

The dataset consisted of 1259 individuals from 48 unique countries. The majority (80.1%) of individuals were from the United States (60.0%), United Kingdom (14.7%) and Canada (5.7%). The remaining individuals (19.9%) include countries spanning multiple continents, but primarily from Europe.

The dataset contained over 25 variables, however we identified 10 main features for analysis which we believe best related to the question we aimed to tackle. These predictors are shown in the table below. In addition to this, p-values obtained from the Chi-squared test were evaluated to determine any associations between these 10 main features and the outcome of interest.

Variable	**Survey question**
Age	
Gender	
Self_employed	Are you self-employed?
Family history	Do you have a family history of mental illness?
No_employees	How many employees does your company or organization have?
Remote_work	Do you work remotely (outside of an office) at least 50% of the time?
Care_options	Do you know the options for mental health care your employer provides?
Seek_help	Does your employer provide resources to learn more about mental health issues and how to seek help?
anonymity	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
leave	How easy is it for you to take medical leave for a mental health condition?

Table 1. Overview of selected variables (Kaggle, 2014)

Two individuals were subsequently removed from analysis, due to insufficient data relating to gender. Most individuals were over 30 years of age (53.4%) and most individuals participating in the survey identified as male (78.8%).

Methods

Following the data exploration, we identified an outcome variable- “Treatment” and a number of potential predictors for mental health in the aforementioned section. Pre-processing was required before implementing the methodology. For the ‘gender’ variable, spelling mistakes were corrected, the gender observations were organised into three major groups (Male, Female and Trans) and the records without clear descriptions were dropped from the dataset. For the ‘age’ variable, we removed negative values, and values that were below 18 or over 120 were automatically populated using the median age. A new variable called ‘Age_Group’ with two categories, 18 - 30 yrs and Greater than 30 yrs, was created. The ‘no_employees’ variable was subdivided into three groups: 1-25, 26-1000 and More than 1000. For the ‘work-interfere’ variable, null values were populated with “Don’t Know” and, finally, null values under the ‘Self_employed’ variable were populated with “No”.

At the beginning, there were 1259 records and following the data cleansing there were 1257.

Following the cleaning, the outputs of data were used to perform a Chi-squared test to quantitatively display the difference between the result observed and the result expected. As the groups were subdivided, this allowed for greater precision. The variables that had a p value < 0.05 were gender, work-interfere, family history, care options, anonymity, benefits, and leave (Appendix 3). Therefore, these variables were noted to have an association with the outcome variable- (mental health) treatment. As a result, these variables were included in the final logistic regression model. [Please note that age was included in the model irrespective of its Chi-squared p value]. Our aim was to obtain odds ratios and confidence intervals from this model.

To begin the modelling process, we encoded all the variables before being inputted in the logistic regression model (Raoniar, 2020). This is because the model does not work with string values. Prior to this, the dataset was split into train and test, 80% train and 20% test. The 80% was first used to train the model, while the remaining 20% of the data was used to view how the model generalised an unseen dataset.

Following this, the model was evaluated using a confusion matrix and a Receiver Operator Characteristic (ROC) curve. A confusion matrix was used to evaluate how well the model fits the dataset. An ROC curve was used to highlight the diagnostic potential of the predictors by plotting the true positive rate against the false positive rate.

This analysis was done using Python Version 3 (Navlani, 2019).

Modelling and Results

Through the analysis of survey participants, we were able to determine various characteristics relating to the population. Most participants were male (approximately 78.8%) with the remainder identifying as female (19.6%) and trans (1.6%). In addition, most of the participants were aged over 30 years old (53.9%). Our analysis revealed of the 1257 participants, 50.5% received treatment for a mental health disorder. It was also revealed that of those who were receiving treatment, 56.9% were aged over 30, 70.9% were male, while those who identified as female and trans accounted for 26.8% and 2.4% respectively. Further analysis of the survey data allowed us to look more in-depth at the characteristics of the population and how they contributed to the incidence of mental health within the tech workspace.

Table 2 shows the logistic regression analysis of several variables within our model. The employment of our logistic regression enabled us to determine the odds ratios and the confidence intervals for all variables (Li, 2017). This analysis enabled us to determine that family history, gender and care options (highlighted within Table 2) had the greatest association with the incidence of mental health in comparison to all other factors including age and anonymity. This is exemplified by the increased odds ratios and confidence intervals when compared to the remaining variables.

Variable	**Odds Ratio**	**95% CI**
Family History (Yes)	**4.81**	**3.69-6.27**
Gender (Female)	**2.21**	**1.59-3.06**
Care Options (Yes)	**2.85**	**2.06-3.95**
Care Options (Not sure)	0.88	0.63-1.22

Age (greater than 30 years old)	1.25	0.97-1.61
Anonymity (Yes)	1.30	0.95-1.77
Anonymity (No)	1.19	0.65-2.15
Seek help (Yes)	0.82	0.57-1.19
Seek_help (Don't know)	1.02	0.75-1.39

Table 2. Logistic Regression Odds Ratios, 95% Confidence Intervals (CI) for participants for Mental Health Survey

By generating a confusion matrix, we were able to determine the performance of the model. We were able to assess the performance of the model using 252 testing samples. The blue coloured matrices represent the correct classifications, compared to other false classifications. From the figure, 34 data points were classed as false negatives and 13 as false positives. With most samples (81.3%) showing true positives and true negatives, it may be concluded that this model was effective in determining the association between mental health and employment within the tech sector.

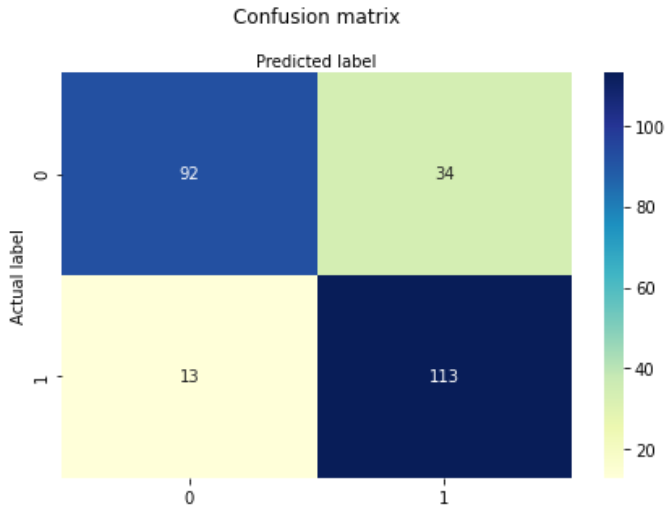


Figure 2. Confusion Matrix Heat Map. True Positives shown in the top left. False Negatives in the top right. False Positives in the bottom left. True Negatives in the bottom right. Accuracy: 0.8134920634920635. Precision: 0.7687074829931972. Recall: 0.8968253968253969

By plotting a ROC curve, we were able to visualise the trade-off between sensitivity and specificity of the model and its success. The x-axis represents the false positives for the model while the y-axis represents the true positive. The area under the curve (AUC) represents the ability of the model to accurately distinguish between positive and negative results, therefore the higher the AUC value, the better the performance. With an AUC of approximately 0.878, it suggests that this model can accurately determine predictors of mental health incidence in the workplace. As the AUC is above 0.7 we can assume that this model demonstrates excellent discrimination (Yang, 2017).

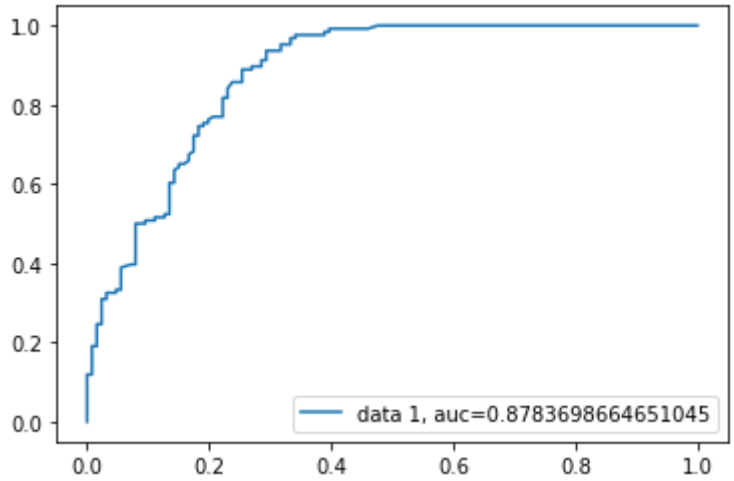


Figure 3. Receiver Operating Characteristic (ROC) curve. Demonstrating the trade-off between sensitivity and specificity. AUC = Area under the curve.

Conclusion

To conclude, the main insight we have drawn from this project is that there appears to be an association between mental health treatment and employees' family history, gender and workplace care options. This has provided evidence towards our research question through data exploration and modelling. Through the Data Utility Framework, we are confident that we have selected a high-quality dataset and our pre-processing techniques have successfully aided in building a model that can predict features in the dataset.

The main challenge that presented itself during the project was not being able to compare the logistic regression with other models such as K-nearest neighbours, Neural Networks, Decision Trees and Random Forest algorithms, due to the limited timeframe.

Recommendations from this project include employers should be encouraged to offer more mental health support packages to its employees. The industry should actively investigate what variables contribute to female mental health issues and how to address them. Finally, firms should prioritise psychological treatment for employees with a family history of mental illness. Firms must develop a dedicated mental health counselling department and genuinely care about its employees' mental health problems, such as through frequent psychological therapy, group building activities, and providing a right perspective of mental health problems

Appendix

1. **Code in GitHub**

https://github.com/Werick/HDRUK_Team_Challenge/blob/main/HDR_UK_Tech.ipynb

HDRUK
Health Data Research UK

2. Data Utility Framework

Category	Dimension	Definition	Bronze	Silver	Gold	Platinum	Please select the description that most closely matches the dataset	Notes (free text)
Data Documentation	Documentation Completeness	This element will be calculated separately						
	Availability of additional documentation and support	Available dataset documentation in addition to the data dictionary	Past journal articles demonstrate that knowledge of the data exists	Comprehensive ReadMe describing extracting and use of data, Dataset FAQS available, Visual data model provided	Dataset publication was supported with a journal article explaining the dataset in detail, or dataset training materials	As Gold, plus support personnel available to answer questions	Bronze	There are articles (although not published within scientific journals) which demostrate knowledge of the data.
	Data Model	Availability of clear, documented data model	Known and accepted data model but some key field un coded or free text	Key fields codified using a local standard	Key fields codified using a national or international standard	Data Model conforms to a national standard and key fields codified using a national / international standard	Not yet Bronze	
	Data Dictionary	Provided documented data dictionary and terminologies	Data definitions available	Definitions compiled into local data dictionary which is available online	Dictionary relates to national definitions	Dictionary is based on international standards and includes mapping	Silver	Data dictionary available on the website.
	Provenance	Clear description of source and history of the dataset, providing a "transparent data pipeline"	Source of the dataset is documented	Source of the dataset and any transformations, rules and exclusions documented	All original data items listed, all transformations, rules and exclusion listed and impact of these	Ability to view earlier versions, including versions before any transformations have been applied data (in line with deidentification and IG approval) and review the impact of each stage of data	Bronze	Data was collected via an electronic survery
Technical Quality	Data Quality Management Process	The level of maturity of the data quality management processes	A documented data management plan covering collection, auditing, and management is available for the dataset	Evidence that the data management plan has been implemented is available		Externally verified compliance with the data management plan, e.g. by ISO, COC, ICO or other body	Not yet Bronze	
	Data Management Association (DAMA) Quality Dimensions	These elements will be calculated separately						
Coverage	Pathway coverage	Representation of multi-disciplinary healthcare data	Contains data from a single speciality or area	Contains data from multiple specialties or services within a single tier of care	Contains multimodal data or data that is linked across two tiers (e.g. primary and secondary care)	Contains data across more than two tiers	Bronze	Collects data from people in who work in different tech companies/organizations
	Length of follow up	Average timeframe in which a patient appears in a dataset (follow up period)	Between 1 - 6 months	Between 6 - 12 months	Between 1 - 10 years	More than 10 years	Silver	The survey is repeated yearly from 2014 till 2020.
Access & Provision	Allowable uses	Allowable dataset usages as per the licencing agreement		Non-consented, aggregate data for specific academic uses (followine IG approval)	Aggregate data, for academic and specific commercial uses (following IG approval)	Fully consented for commercial uses (following IG approval)	Platinum	
	Time Lag	Lag between the data being collected and added to the dataset	Approximately 1 year	Approximately 1 month	Approximately 1 week	Effectively real-time data	Platinum	
	Timeliness	Average data access request timeframe	Less than 6 months	Less than 3 months	Less than 1 month	Less than 2 weeks	Platinum	
Value & Interest	Linkages	Ability to link with other datasets	Identifiers to demonstrate ability to link to other datasets	Available linkages outlined and/or List of datasets previously successfully linked provided	List of restrictions on the type of linkages detailed. List of previously successful dataset linkages performed, with navigable links to linked datasets via at DOI/URL	Existing linkage with reusable or downstream approvals	Bronze	
	Data Enrichments	Data sources enriched with annotations, image labels, phenomes, derivations, NLP derived data labels	The data include additional derived fields, or enriched data.	The data include additional derived fields, or enriched data used by other available data sources.	The derived fields or enriched data were generated from, or used by, a peer reviewed algorithm.	The data includes derived fields or enriched data from a national report.	Not yet Bronze	

HDRUK
Health Data Research UK

3. Chi squared summary table

Variable Name	Total Population N = 1257	Mental Health Treatment (No) N = 622	Mental Health Treatment (Yes) N = 635	Chi Square p value
Age Group				0.328
18-30	580 (46.1%)	306 (49.2%)	274 (43.1%)	
greater_than 30	677 (53.9%)	314 (56.8%)	361 (56.9%)	
Gender				<0.01
Male	991 (78.8%)	541 (87.0%)	450 (70.9%)	
Female	247 (19.6%)	77(12.4%)	170(26.8%)	
trans	19 (1.5%)	4 (0.6%)	15 (2.4%)	
self employed				0.988
No	1113 (88.5%)	559 (89.1%)	559 (88.0%)	
Yes	144 (11.5%)	68 (10.9%)	78 (12.0%)	
Work Interfere				<0.001
Often	142 (11.3%)	21 (3.4%)	121 (19.1%)	
Sometimes	465 (37.0%)	107 (17.2%)	358 (56.2%)	
Rarely	173 (13.8%)	51 (8.2%)	122 (19.2%)	
Never	213 (16.9%)	183 (29.4%)	30 (4.7%)	
Don't Know	264 (21.0%)	260 (41.8%)	4 (0.6%)	
family history				<0.001
No	767 (61.9%)	495 (79.6%)	272 (42.8%)	
Yes	490 (39.0%)	127 (20.4%)	363 (57.2%)	
no of employees				0.961
1 - 25	450 (35.8%)	233 (37.5%)	217 (34.2%)	
26 - 1000	525 (41.8%)	253 (40.7%)	272 (42.8%)	
More than 1000	282 (22.4%)	136 (21.9%)	146 (23.0%)	
remote work				0.944
No	883 (70.2%)	444 (71.4%)	439 (69.1%)	
Yes	374 (29.8%)	178 (28.6%)	196(30.9%)	
care options				<0.001
No	501 (39.9%)	294 (47.3%)	207 (32.6%)	
Yes	442 (35.2%)	137 (22.0%)	305 (48.0%)	
Not sure	314 (25.0%)	191 (30.7%)	123 (19.4%)	
seek help				0.112
No	646 (51.4%)	323 (51.9%)	323 (50.9%)	
Yes	248 (19.7%)	102 (16.4%)	146 (23.0%)	
Don't know	363 (28.9%)	197 (31.7%)	166 (26.1%)	
anonymity				<0.001
No	65 (5.2%)	27 (4.3%)	38 (6.0%)	
Yes	373 (29.7%)	147 (23.6%)	226(35.6%)	
Don't know	819 (65.1%)	448 (72.0%)	371 (58.4%)	
Benefits				<0.001
No	374 (29.8%)	193(31.0%)	181 (28.5%)	
Yes	475 (37.8%)	172 (27.7%)	303 (47.7%)	
Don't know	408 (32.5%)	257 (41.3%)	151 (23.8%)	
leave				0.001
Very easy	204 (16.2%)	103 (16.6%)	101 (15.9%)	
Somewhat easy	266 (21.2%)	135 (21.7%)	131 (20.6%)	
Somewhat difficult	126 (10.0%)	44 (7.1%)	82 (12.9%)	
Very difficult	98 (7.8%)	31 (5.0%)	67 (10.6%)	
Don't know	563 (44.7%)	309 (49.7%)	254 (40.0%)	

HDRUK
Health Data Research UK

References

Brugman, S., 2020. Pandas Profiling. [online]. Available at: https://github.com/pandas-profiling/pandas-profiling#types.

Kaggle, 2014. Mental Health in Tech Dataset. [online]. Available at: https://www.kaggle.com/osmi/mental-health-in-tech-survey.

Li, S., 2017. Building A Logistic Regression in Python, Step by Step. [online]. Available at: https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8.

Mental Health Foundation, 2015. Fundamental Facts About Mental Health 2015. Available at: https://www.mentalhealth.org.uk/publications/fundamental-facts-about-mental-health-2015

Navlani, A., 2019. Understanding Logistic Regression in Python. [online]. Available at: https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python.

Open Sourcing Mental Illness, 2014. OSMI Mental Health in Tech Survey. [online]. Available at: https://osmihelp.org/research.

Raoniar, R., 2020. Modelling Binary Logistic Regression Using Python . [online]. Available at: https://onezero.blog/modelling-binary-logistic-regression-using-python-research-oriented-modelling-and-interpretation/.

Risdal, M., 2018. Machine Learning for Mental Health. [online]. Available at: https://www.kaggle.com/kairosart/machine-learning-for-mental-health-1.

Yang, S. & Berdine, G., 2017. The receiver operating characteristic (ROC) curve. The Southwest Respiratory and Critical Care Chronicles. 5, .