Mental Health Accommodation in Tech Companies

Xinbei Yu

October 15 2022

1 Introduction

With the growing interest in mental health issues in every industry, tech companies are no exception. According to a World Health Organization report on mental health at work, by 2019, 15% of working-age adults are experiencing mental disorder, and an estimated $1 trillion US dollars are lost due to lacking working days due to mental disorder. According to BIMA Tech Inclusion and Diversity Report in 2019, 52% of tech employees suffered from depression or anxiety, and they are five times more depressed than UK average. There's no reason to suspect any difference from US and other demographics. Thus, it is important for employers to take actions in order to make a healthy and safe working environment to support their employees.

In this research, I will explore what factors affect the healthiness and safeness of mental health environment at work by using the easiness of taking a leave due to mental health conditions as the metrics, i.e. target variable. This problem is classification as the target variable is recorded in five category–very easy, somewhat easy, somewhat difficult, very difficult and don't know. There are in total of 1251 data points and 26 features before any dropping.

Except age, all other features are categorical. One third of the features are information-related, such as age, country, number of employee they work with, and if they are self-employed. The other two thirds of the features are subjective, including the target variable. These features, such as if they have observed negative consequence for co-workers taking a leave for mental health condition, if they are willing to discuss mental conditions with co-workers or during interviews, varies among people, depending on people's opinion.

The data set is from Kaggle, and it is collected from giving out surveys from a non-profit organization named Open Sourcing Mental Illness. Previous projects of this data set is mostly researching which factors affect a person's mental health condition. The most accurate model is XGBClassier, which gives an above 0.8 accuracy rate, and the lowest accuracy is from logistic regression, which gives an above 0.65 accuracy. However, with the different research question and target variable, I wouldn't say these results boost my expectation on the models' performance for my research.

The data is rather well documented in Kaggle description, and the feature names are mostly self-explanatory.

2 EDA

Target Variable Visualization:

As Figure 1 shows, almost half of the people who took the survey answered don't know if it is easy to take a leave for mental health condition. However, there are multiple explanation than people actually not knowing. One potential explanation is the anonymity for this survey isn't trusted, and people who answered 'Don't know' is worried about the negative consequence of answering 'very difficult.' Thus, we explore the correlation between people who observed negative consequence for taking a mental health leave, i.e. obs_consequence in this data set, and the target variable.

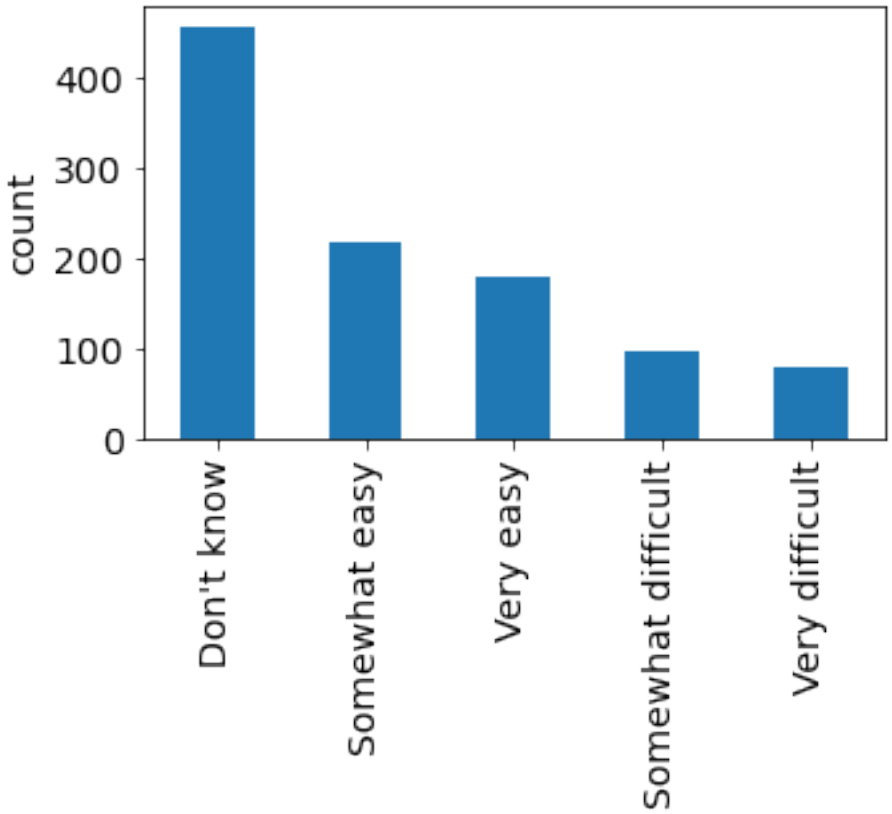

Figure 1: Easiness of taking a mental health leave [leave] bar plot visualization. Each bar represent the number of people who gave this answer.

obs_consequence vs. leave:

Unlike the assumption that most people answered 'Don't know' have observed negative consequence for taking a mental health leave, the stacked bar plot shows that majority of people who answered 'Don't know' have not seen a negative consequence.

However, although majority of people said they have not observed negative consequence for taking a mental health leave, within the people who answered 'Somewhat difficult' and 'Very difficult', the ratio of people who reported seen a negative consequence is more that two times than the rest of group.

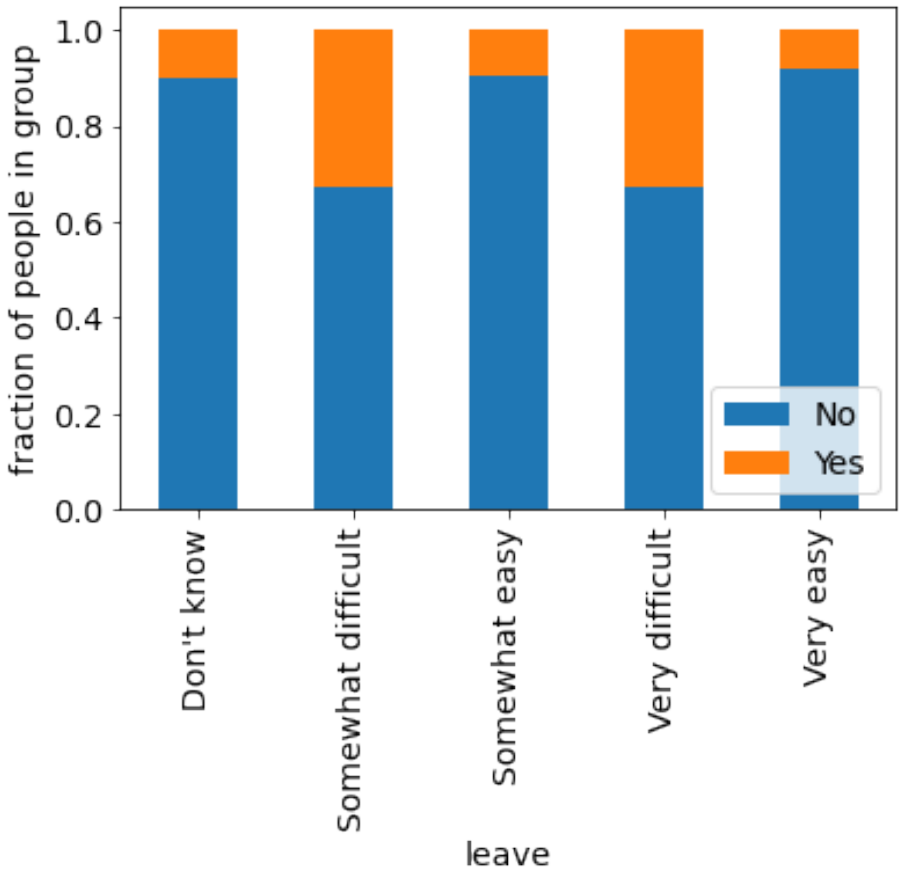

Figure 2: obs_consequence stands for if one has observed negative consequence for people who took a mental health leave. For each value in the target variable leave, the blue represents people who did not see a negative consequence, and orange represents people who saw a negative consequence.

anonymity vs. leave:

Another factor that have a rather high effect on easiness of taking a mental

health leave is the anonymity for people choosing to take advantage of the mental health programs. As Figure 3 shows, the result conforms with the result in Figure 2. People who answered 'Somewhat difficult' and 'Very difficult' are more likely to believe that the anonymity isn't protected. As the target variable level vary from very easy to very difficult, the number of people who believed in the protected anonymity is also rising.

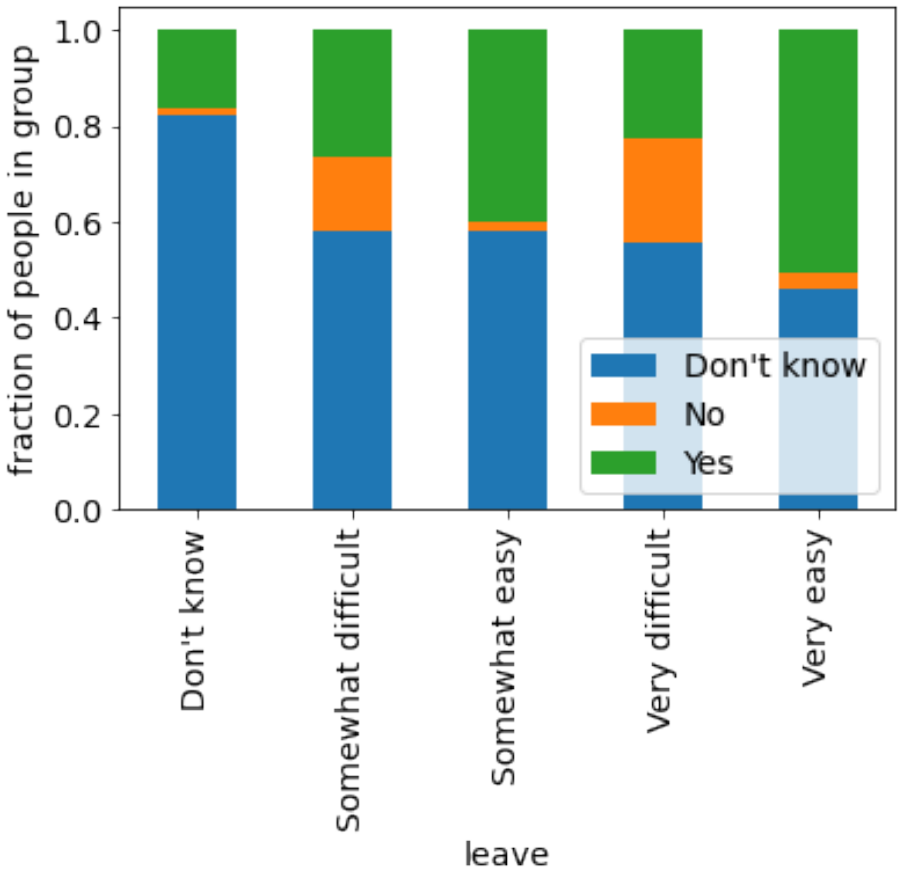

Figure 3: anonymity stands for if one believe the anonymity of taking a mental health leave is protected. Three values of this feature is 'Dont know', represented as blue, 'No', represented in orange, and 'Yes', represented in green.

age vs. leave:

Another potential factor is age, and contradicting with my assumption older people find it harder to take a mental health leave, there's no obvious variation as age changes.

Nevertheless, the distinction between people answered 'very difficult' and 'very easy' is still detectable as the average for people finding it difficult to take

a mental health leave to be older and skewed more to the older age.

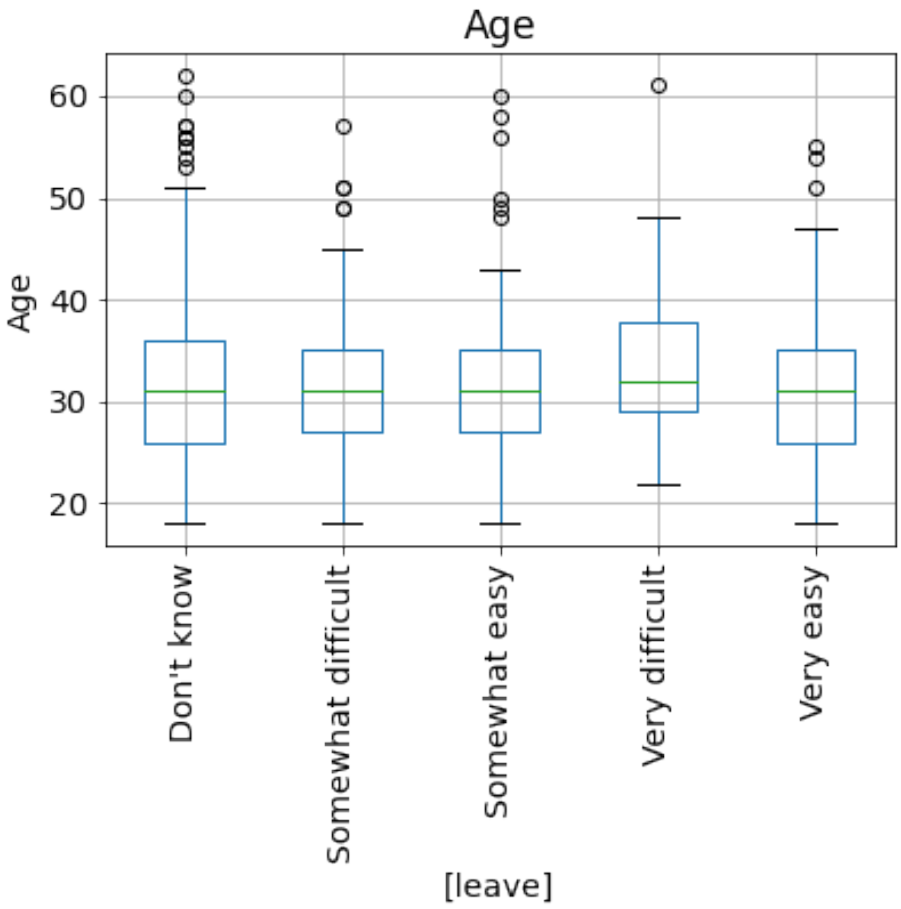

Figure 4: This is a box plot for different class in the target variable in terms of age.

3 Data Preprocessing

Missing value in feature state, which is only required for people answer US as country is imputed with new value 'other, ' missing value in feature self_employed, which stands for if one is self-employed is imputed with mode 'No' as the percentage of people answering 'No' is much higher than 'Yes, ' and the missing value in work_interfere, which stands for whether a person feel mental health condition will interfere their work, is imputed with a new value 'Don't know' as making an assumption might leads to the inaccuracy of the

model.

The data is filtered into 1025 data points with only people answered 'Yes' to the question whether they are employed in a tech company as we are only researching mental health leave easiness within the tech companies. This column is later dropped with comments and timestamp during preprocessing as they are irrelavant to the question of interest.

Because the balance of the target variable is $455 : 218 : 176 : 98 : 78$, which is rather imbalanced, the data is split using stratified k-fold. If we taken into account the possibility that many of the people taking the survey are possibility from the same company, then there will be an implicit group structure of company, which makes the data not i.i.d. However, this survey is distributed by putting on a website, and thus is less likely to be distributed among the same company. It is reasonable to make the assumption that this group structure is negligible in this question, and with this dataset not being a time series, allowing the subsequent assumption that this dataset can be treated as i.i.d. The dataset is split with a set random state for reproducibility, and it is split into a $6 : 2 : 2$ proportion for train, validation, and final test. There are in total 5 folds, stratified with respect to the target variable, and is shuffled to avoid the possibility that the data set is sorted by class.

Except the only continuous feature, age, all the other features are preprocessed using OneHotEncoder as there are no ordinal structure to their values. Age, however, is preprocessed using StandardScalar instead of MinMaxEncoder as the distribution is right-skewed when visualized. With the large number of OneHotEncoder usage, the number of features after preprocessing is 169 instead of the 23 after dropping.

4 Reference

Aditimulye. "Mental Health at Workplace." Kaggle, Kaggle, 19 Sept. 2021, https://www.kaggle.com/code/aditimulye/mental-health-at-workplace.

"Mental Health at Work." World Health Organization, World Health Organization, https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work.

The Voices of Our Industry - Bima. https://bima.co.uk/wp-content/uploads/2020/01/BIMA-Tech-Inclusion-and-Diversity-Report-2019.pdf.

5 GitHub Repository

https://github.com/XinbeiYu00/project-XinbeiYu.git