Mental Health Accommodation in Tech Companies

Xinbei Yu

Data Science Initiative, Brown University

Link to Project:
https://github.com/XinbeiYu00/project-XinbeiYu.git

December 4, 2022

1 Introduction

With 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. 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 healthiness and safeness of mental health environment at work by using easiness of taking a leave due to mental health conditions as metrics, i.e. target variable. This problem is classification as 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 features are information-related, such as age and country. The other two thirds of features are subjective, including target variable. These features, such as if they have observed negative consequence for co-workers taking a leave for mental health condition 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 XGBClassifier, 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 different research questions, the best performed model is SVC instead of XGB Classifier, and accuracy is around 0.5 for 5 different classes.

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 negative consequence of answering 'very difficult.' Thus, we explore correlation between people who observed negative consequence for taking a mental health leave, i.e. obs_consequence in this data set, and target variable.

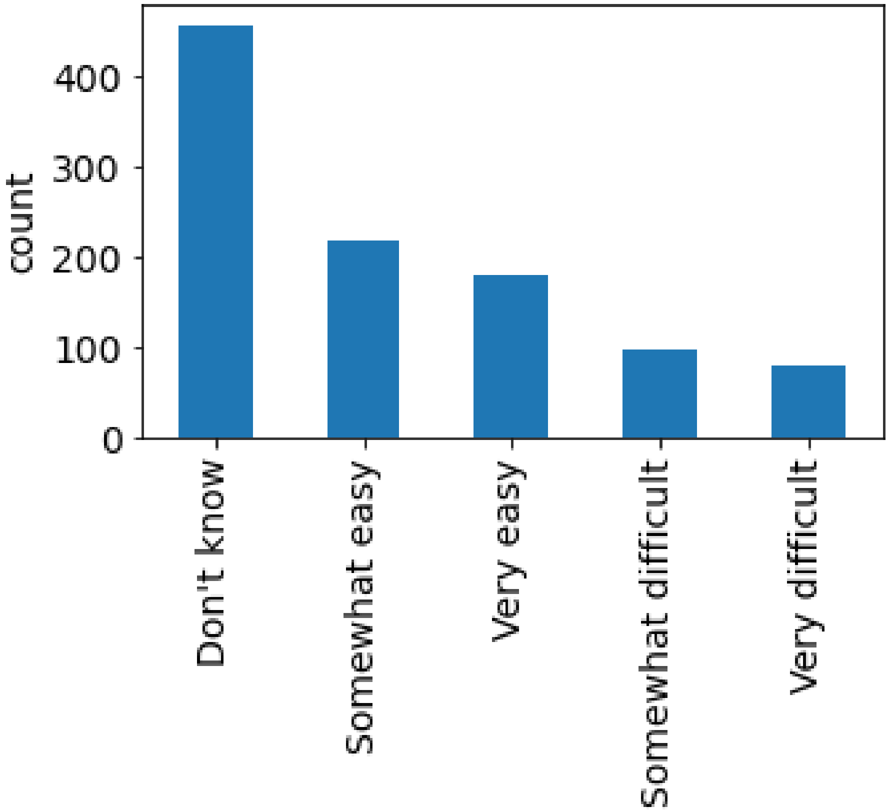


Figure 1: Easiness of taking a mental health leave [leave] bar plot visualization. Each bar represent number of people who gave this answer.

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.

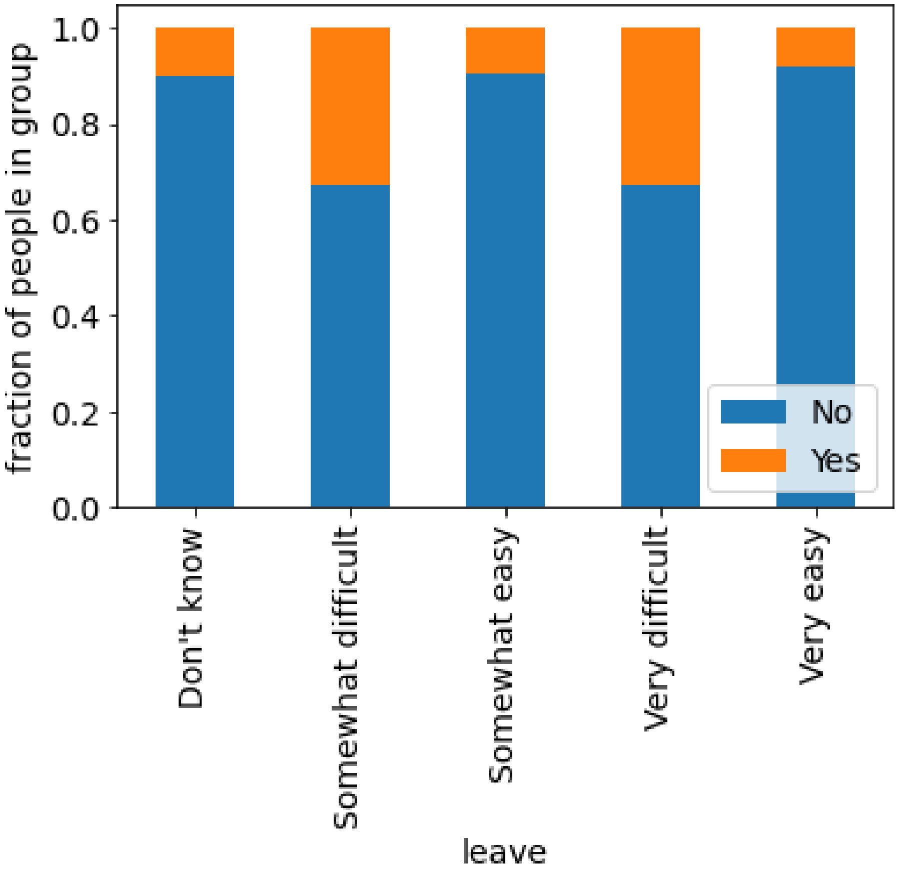


Figure 2: obs_consequence stands for if one has observed negative consequence for people who took a mental health leave. Blue represents people who did not see a negative consequence.

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 mental health programs. As Figure 3 shows, 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, number of people who believed in the protected anonymity is also rising.

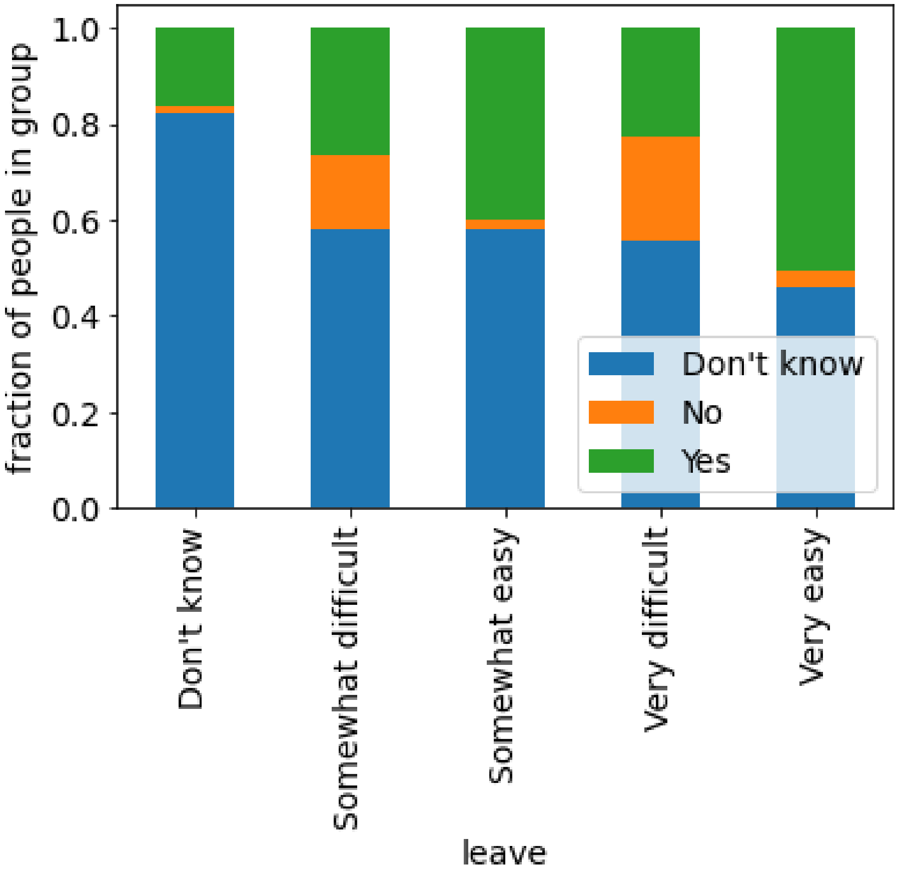


Figure 3: anonymity stands for if one believe the anonymity of taking a mental health leave is protected. Three values of this feature is ’Dont know’, represented as blue, ’No’, represented in orange, and ’Yes’, represented in green.

3 Methods

Because the balance of target variable is $455 : 218 : 176 : 98 : 78$, which is rather imbalanced, data is split using stratified k-fold. The dataset is split with a set random states 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 data set is sorted by class.

Except the only continuous feature, age, all the other features are preprocessed using OneHotEncoder. Age, however, is preprocessed using StandardScalar instead of MinMaxEncoder as the distribution is right-skewed when visualized. With large number of OneHotEncoder usage, the number of features after preprocessing is 169 instead of 23 after dropping.

All ML algorithms tried in this project are measured using accuracy score since assumed audiences are tech companies' health department, accuracy is the most understandable metrics. The uncertainties due to splitting are measured by embedding splitting as part of ML pipeline with 10 different random states. With collected test scores from all random states, standard deviation of the test scores are calculated to represent uncertainties in splitting.

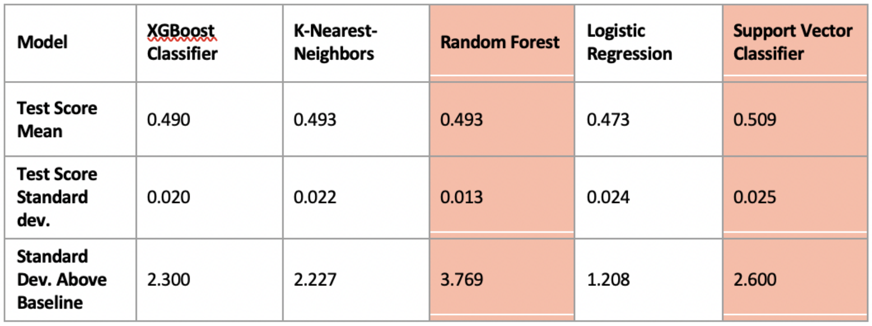
Except for XGBClassifier, all machine learning algorithm is implemented by using gridsearch to tune hyperparameters. The first one tried is Logistic Regression with three different penalties for regularization as hyperparameters: L1, L2, and elastic net. The second algorithm used is K-Nearest-Neighbors, with two hyperparameters tuned: weight for each neighbor is tuned between distance and uniform, and number of neighbors used for prediction are 1, 3, 10, 30, 100, 300, which is about 1/4 of total number of data. The next algorithm tried is random forest(RF). In this algorithm, the uncertainty due to pipeline itself is measured using standard deviation of all test scores as well and setting random state to a set value when initializing the algorithm. The hyperparameters tuned are number of features used for each decision tree, from half of features, three quarter of features to all features. Another hyperparameter tuned is depth of each decision tree, which is evenly spaced in logspace from 1 to 100. Last model using gridsearchcv pipeline is support vector machine(SVC). The first parameter tuned is regularization parameter: 1e-1, 1e0, and 1e1. The second parameter tuned is gamma, which is a parameter for non linear hyperplanes. This parameter is tuned to avoid overfitting, and is tested with 1e-3, 1e-1, 1e1, 1e3, 1e5.

For XGBClassifier, a parameter grid is used . So, unlike the other algorithms with grid search, split data is preprocessed by hand before passing down to paramter grid to find the best hyperparameters. Due to runtime of XGB, the only hyperparameter tuned is max depth of trees as it affects result the most. In addition, the number of subsample used is set to 0.5 to avoid overfitting.

4 Results

All the test scores are higher than the baseline score, which is about 0.444, calculated by assuming all prediction falls in the 'Don't know' class. If I were to use the standard deviation of test scores of SVC, most predictive model with an average accuracy around 0.509 and best hyperparameters as gamma = 1.0 and C = 0.1, to measure the distance between baseline score and the model' performance, it is about 2.5 standard deviation higher than the baseline, which makes it less likely for the result to be an uncertainty due to splitting. Although SVC is the best performing model, the result also suggest that Random forest has the highest standard deviation above the baseline, meaning it is more stable than SVC between random states.

The following figure shows each ML model's performance:



Model	XGBoost Classifier	K-Nearest-Neighbors	Random Forest	Logistic Regression	Support Vector Classifier
Test Score Mean	0.490	0.493	0.493	0.473	0.509
Test Score Standard dev.	0.020	0.022	0.013	0.024	0.025
Standard Dev. Above Baseline	2.300	2.227	3.769	1.208	2.600

Figure 4: This table summarizes all 5 models' performance

As figure 5 shows, the most predictive model is Support Vector Machine and the second predictive model is RF with best hyperparameter as max_depth = 10, and max_feature = 0.5.

To visualize the result of SVC's prediction, the following figure agree with the imbalanced class that most accurate predictions come from Don't know class.

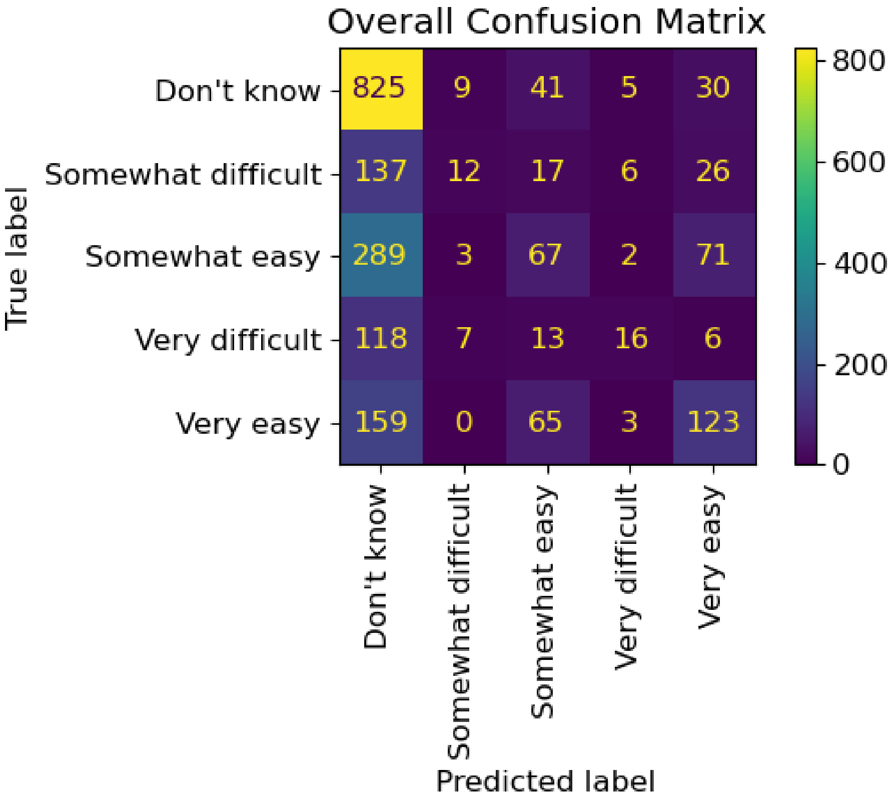


Figure 5: SVC's confusion matrix visualized the prediction

As figure 6 shows, using the permutation feature importance on SVC, the top three most important features are 'if people take mental health as seriously as physical', 'if people believe there's consequence when taking a mental health leave', and 'if people believe the anonymity of taking a mental health program is protected'. This method of feature importance is the one with most interpretability as feature used is before preprocessing.

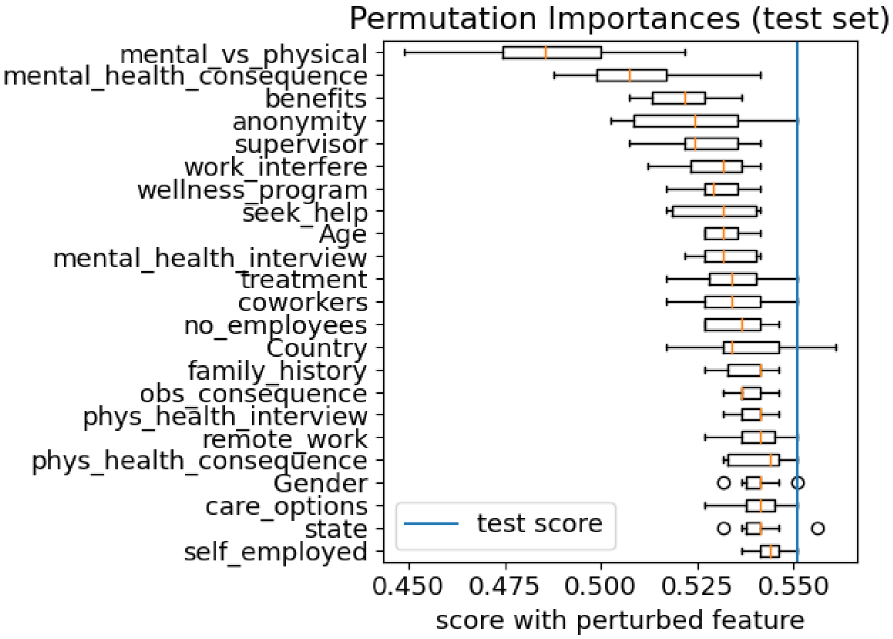


Figure 6: This figure shows the feature importance from highest to lowest

Figure 7 shows the feature importance using shap values, but the value of this result is less than that of permutation as the features are after preprocessing.

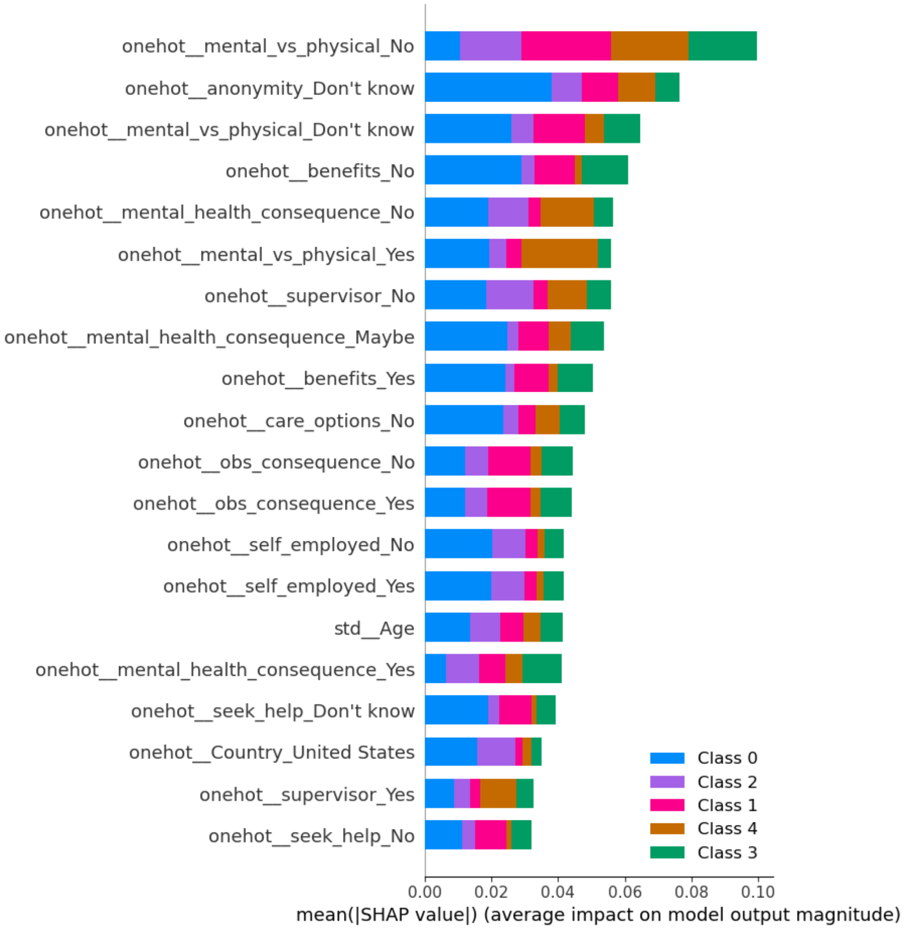


Figure 7: global feature importance calculated using SHAP

Lastly, figure 8 shows the feature importance from RF model as SVC has no third method to measure global feature importance. The method is based on built-in method for RF package in sklearn, and it is in a similar logic of permutation. As the figure shows, different models value different features.

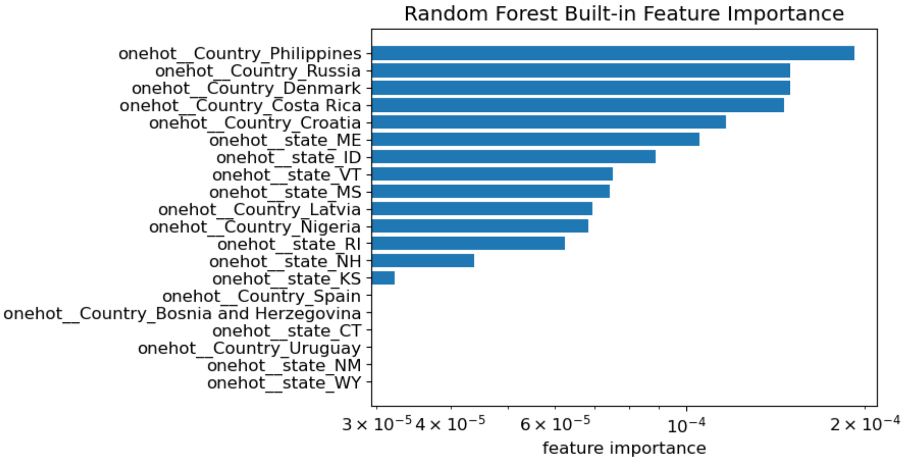


Figure 8: global feature importance calculated based on RF

The highest model score for local feature importance give is 0.4 among five classes, which is the 'Don't know' class. As figure 10 shows, the preprocessed feature that impact this score positively is 'if this person is willing to discuss his/her mental health issue with coworkers', and the feature that impact the score negatively the most is 'if this person is willing to discuss his/her mental health issue with supervisors.' This specific person, i.e. data point, then demonstrate a situation where the easiness of taking a mental health leave is largely affect by the working environment being a safe place to share. To better understand, figure 11 shows the force plot for 'Very Easy' class, which shows that the features that influence this person's decision the most are also 'if this person is willing to discuss his/her mental health issue with coworkers' and 'if this person is willing to discuss his/her mental health issue with supervisors.' Nevertheless, the force plot for 'Very Difficult' class is a bit different as figure 12. This result is reasonable by the fact that if the person finds mental health issue not as seriously as physical health, it would be hard for them to ask a mental health leave.

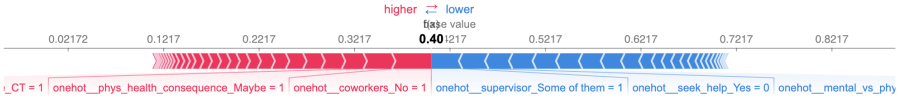


Figure 9: shap value force plot for 'Don't know' class visualization where the features in pink means positively influenced the score and feature sin blue means negatively influenced the score.

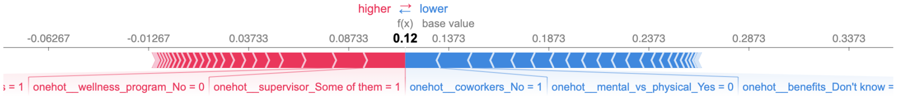


Figure 10: shap value force plot for 'Very Easy' class

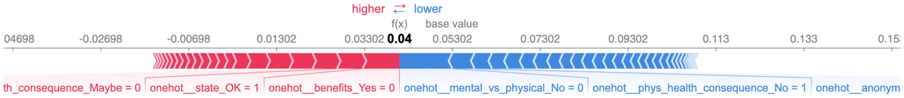


Figure 11: shap value force plot for 'Very Difficult' class

5 Outlook

To give my model more predictive power, Catboost might provide a better result as most feature in this dataset is categorical.

One weak spot of this modeling approach is the 'Don't know' class which is hard to understand as it is not related to the anonymity being protected. A potential solution is to re-classify data points into a binary class - "Easy" and "Hard", and drop the "Don't know" class. This could potentially both improve the predictive power and interpretability of the model.

Another technique that can improve the original 5 class data set is to add class weight parameter during analysis to minimize the effect of imbalanced data.

Lastly, if we could have more data on people answering 'Very easy' and 'Very hard, ' the model is likely to improve its predicative power as these two classes have least number of points but is very helpful in analyzing if a company's employees find it hard to take a mental health leave.

6 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.