Mental Health In the Tech Industry

Bernard Jiang, Dylan Limnavong, Paula Moya Nieto

btj29, dtl53, pm445

Abstract

The goal of this project is to evaluate the mental health condition of an employee, according to certain parameters of workplaces, and to determine whether working in the tech industry, usually associated with a stressful environment, plays a role. To better measure the mental health of an individual, we classified each employee as either having mental health issues or not. The predictions were derived from Logistic Regression, K-Nearest Neighbours, and Tree-Based Models. The overall classification accuracy and balanced accuracy were used to compare the different models. Logistic Regression and Random Forest were found to be appropriate models. They identified *mental_health_consequence*, *phys_health_consequence*, and *treatment* as being the most important predictors in the workplace. There was not a clear relationship between mental health and working in the tech industry.

Introduction

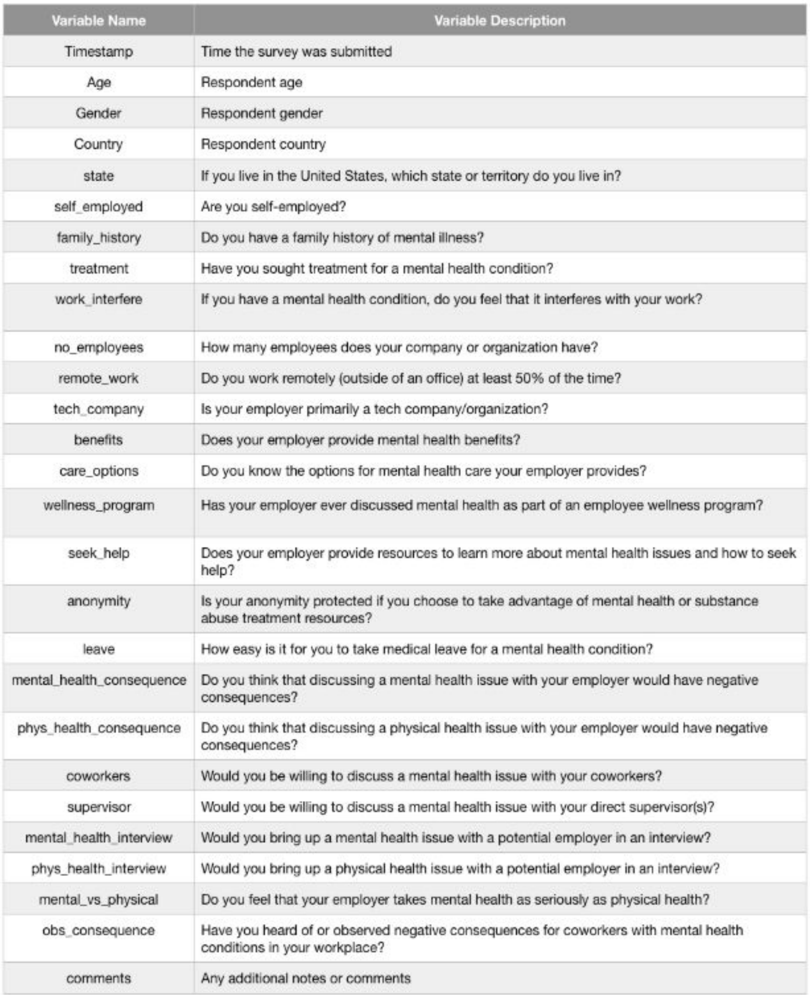
Mental Health is one if the most important issues affecting people today, especially the youth. More than 44 million Americans suffer from a mental illness, and the rate of youth living with mental health conditions is continually rising every year [4]. People living with mental health problems are more likely to live with chronic medical conditions and to die twenty-five years earlier than people without these problems [3]. Apart from the medical and general health problems that poor mental health could cause, it can also induce low levels of productivity in employees, which costs hundreds of millions of dollars in lost earnings every year [3]. Therefore, employees should care about their mental health, and their employers should attend to it as well. However, there is an enormous stigma surrounding mental health in places of employment (and society in general), and the tech industry is not treating this problem as carefully as it should [5].

In this project we will aim to answer two questions: 1) "Which predictors are the most important in determining whether an employee has mental health problems?", and 2) "Does tech has a disproportionate number of mental health cases?". To answer these questions we will use a dataset gathered by *Open Sourcing Mental Illness, LTD*, a 2014 survey that "measures attitudes towards mental health and frequency of mental health disorders in the tech workplace" [1]. The dataset contains 1259 responses and 27 predictors, which are detailed in *Table 1.*

Our analysis will be categorical and will use 3 data science approaches: Logistic Regression, K-Nearest Neighbors, and Decision Trees and Random Forest.

Data Preprocessing

Table 1: Dataset Variables



Variable Name	Variable Description
Timestamp	Time the survey was submitted
Age	Respondent age
Gender	Respondent gender
Country	Respondent country
state	If you live in the United States, which state or territory do you live in?
self_employed	Are you self-employed?
family_history	Do you have a family history of mental illness?
treatment	Have you sought treatment for a mental health condition?
work_interfere	If you have a mental health condition, do you feel that it interferes with your work?
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?
tech_company	Is your employer primarily a tech company/organization?
benefits	Does your employer provide mental health benefits?
care_options	Do you know the options for mental health care your employer provides?
wellness_program	Has your employer ever discussed mental health as part of an employee wellness program?
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?
mental_health_consequence	Do you think that discussing a mental health issue with your employer would have negative consequences?
phys_health_consequence	Do you think that discussing a physical health issue with your employer would have negative consequences?
coworkers	Would you be willing to discuss a mental health issue with your coworkers?
supervisor	Would you be willing to discuss a mental health issue with your direct supervisor(s)?
mental_health_interview	Would you bring up a mental health issue with a potential employer in an interview?
phys_health_interview	Would you bring up a physical health issue with a potential employer in an interview?
mental_vs_physical	Do you feel that your employer takes mental health as seriously as physical health?
obs_consequence	Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
comments	Any additional notes or comments

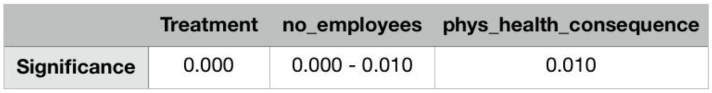
Given the nature of the data collection, several variables were malformed. Before the data provided could be used, we needed to do some preprocessing to clean up the data. The first thing done was to condense the categories in the "Gender" column into "Male", "Female", or "Other". The original data set contained multiple values like "Male", "M", "cis-male", etc., that all represent the same category. Secondly, several entries needed to be removed as the data in the "Age" column was infeasible (e.g. negative values and values in the hundreds).

Logistic Regression

We used the values of the predictor *work_interfere* to create another variable called *mental_health_issues*, with binary values "*Yes*" and "*No*". This new predictor was used to determine which individual suffered from a mental health issue. We then used a Logistic Regression model to 1) determine which predictors are the most important to determine the mental health condition of people, and 2) if the tech industry has a disproportionate number of mental health cases.

To answer the questions, we first ran a Logistic Regression on the whole model with all the predictors. *Table 2* shows the predictors that had a significance of 0.01 and below.

Table 2: Most Significant Predictors in Determining Mental Health Conditions (Entire Model)

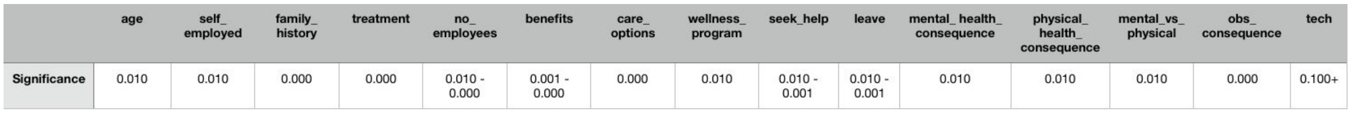


	Treatment	no_employees	phys_health_consequence
Significance	0.000	0.000 - 0.010	0.010

When using the entire model, only three predictors are significant. The model is clearly overfitting the data and using many variables do not get us closer to the real model.

We then ran a Logistic Regression for every individual predictor to see if each predictor was related to *mental_health_issues.* E.g. we created a model for only *mental_health_issues* and *age*, and so on for the rest of the predictors. *Table 3* shows the predictors that had a significance of 0.01 and below.

Table 3: Most Significant Predictors (and tech) in Determining Mental Health Conditions for Each Predictor



	age	self_ employed	family_ history	treatment	no_ employees	benefits	care_ options	wellness_ program	seek_help	leave	mental_ health_ consequence	physical_ health_ consequence	mental_vs_ physical	obs_ consequence	tech
Significance	0.010	0.010	0.000	0.000	0.010 - 0.000	0.001 - 0.000	0.000	0.010	0.010 - 0.001	0.010 - 0.001	0.010	0.010	0.010	0.000	0.100+

As we can see in the table above, there are many predictors that are related to a person's mental health condition. The predictors with the least significance (the most important predictors with a significance of 0.000) are *family_history, treatment, no_employees* (specifically more than 1000 employees), *benefits* (employer provides mental health benefits), *care_options* (knowledge of mental health options of employer), and *obs_consequence* (observation of negative consequences for coworkers). Additionally, working in a tech company was not found to be salient in the suffering of a mental health condition.

Furthermore, we used the individual Logistic Regression models to predict whether a person would suffer from a mental health issue according to each individual predictor. For each predictor we calculated which threshold would provide the highest overall accuracy. *Figure 1* is an example of the overall and balanced accuracies of the variable *age* at different thresholds.

Figure 1: Age's Overall and Balanced Accuracy Plots

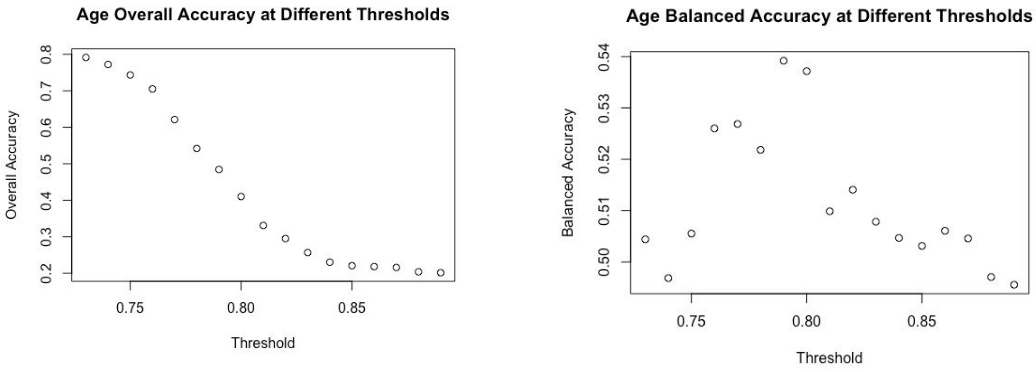
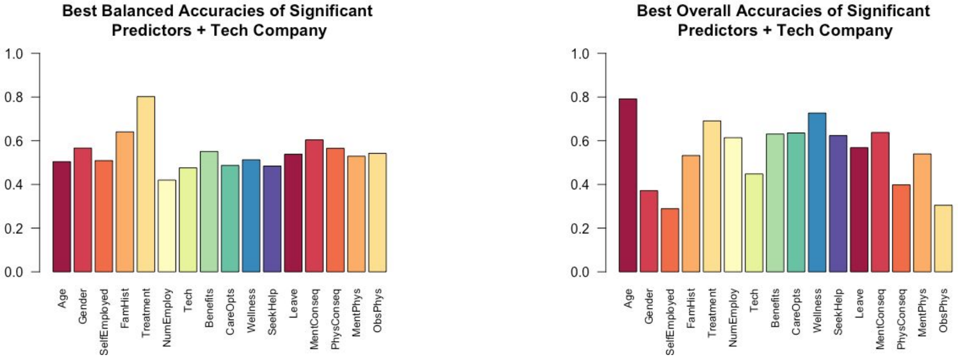


Figure 2 displays the optimal overall and balanced accuracies for all the significant predictors (and *tech* as well, although it is not significant):

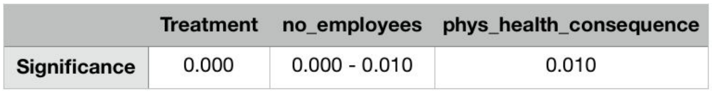
Figure 2: Optimal Overall and Balanced Accuracies for All Significant Predictors



As we can see in *Figure 2*, some variables achieve a maximum overall and balanced accuracy of ~80% (like *treatment* and *age*, respectively). However, most accuracies are not very high, and many of them are below 50%. Although the model does not have high accuracies it is sufficient for real world applications. From the perspective of an employer, it is useful to classify someone as having a mental health condition, even if their probability of having one is not high, because giving appropriate treatment to that person if they are actually suffering will ultimately increase their productivity and work performance. From the perspective of of individuals, it may also be good to be classified as having an illness because if they do have an illness then they will get treatment, and if they do not have an illness then a psychiatrist will probably notice and not treat them or they might even stop the illness before it fully develops.

Next, we ran another Logistic Regression, this time using only the predictors that were significant in the previous regression (which used individual regressions for each predictor). *Table 4* shows the most significant results.

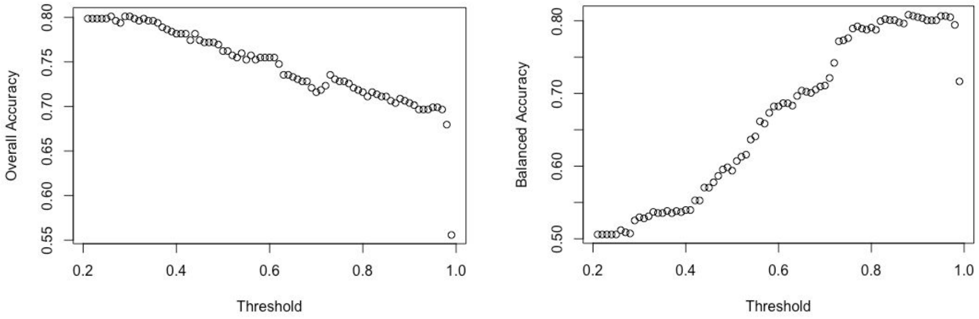
Table 4: Most Significant Predictors in Determining Mental Health Conditions (subset)



	Treatment	no_employees	phys_health_consequence
Significance	0.000	0.000 - 0.010	0.010

This time, we got the same results as we did when we ran the entire model, which suggests that *treatment*, *no_employees*, and *phys_health_consequence* really are the most important predictors of the dataset when determining someone's mental health condition in the workplace.

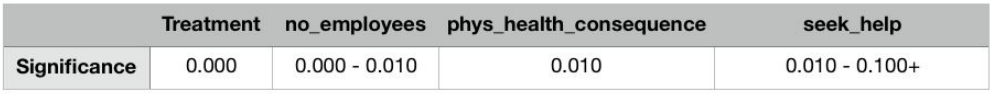
Figure 3: Subset Model Overall and Balanced Accuracies



For this model we again calculated the accuracies, and the results can be seen in *Figure 3.* As overall accuracy went up, balanced accuracy went down. Using the threshold for optimal overall accuracy (threshold = 0.26) gave us an overall accuracy of 0.801 and a balanced accuracy of 0.512; using the threshold for optimal balanced accuracy gave us an overall accuracy of 0.709 and a balanced accuracy of 0.808 (threshold = 0.88). When using a large threshold the increase in balanced accuracy is large and the decrease in overall accuracy is not that high (although significant). However, it is still better to choose a lower threshold due to our problem's context. As mentioned earlier, it is better to classify someone as having a mental illness with a higher probability because the illness might be caught in time, or even before it even fully develops, and even if someone does not have an illness then the psychiatrist will evaluate them and come to a correct conclusion. Thus, using this subset model with a low threshold would be the optimal way to apply the Logistic Regression in the real world.

Finally, we performed Stepwise Subset Selection on the full model. We used the MASS R package's "stepAIC" function. When we did a stepwise selection in both the forward and backward directions we got a subset with the predictors *treatment, no_employees, seek_help, mental_health_consequence*, *phys_health_interview, obs_consequence,* and *phys_health_consequence*. *Table 5* shows the most significant predictors of this model along with their significance codes. Then, we did forward stepwise selection, which resulted in the entire model once again, which is clearly not a good subset (or even a subset). After, we did backward stepwise selection, and we got the same model as in the model with a selection in both directions. Again, the most significant predictors along with their significant codes are in *Table 5*. Therefore, according to Stepwise Subset Selection, the most important predictors in determining an employee's mental health condition are *treatment, no_employees, phys_health_consequence,* and *seek_help*, with the most important being *treatment* and *no_employees.* It is worth noting that *treatment, no_employees,* and *phys_health_consequence* were also the most significant predictors when using the entire model and not just a subset. Additionally, the significance code of the stepwise subset models, suggest the *seek_help* might not be as important as the other predictors in determining an employee's mental health condition. This means that the Logistic Regression of the entire model was was sufficient to answer the two questions presented in this paper.

Table 5: Most Significant Predictors in Determining Mental Health Conditions ("Forward" and "Both Directions" Stepwise Subsets)



	Treatment	no_employees	phys_health_consequence	seek_help
Significance	0.000	0.000 - 0.010	0.010	0.010 - 0.100+

K-Nearest Neighbors

As with previous models, we used the values of the *work_interfere* variable to determine whether an individual suffered from a mental health issue. We then attempted a model using the K-Nearest Neighbors classifier, testing various values of K to try and determine whether an individual had a mental health issue. The first attempt was to create a model using all predictors in the dataset.

Table 6: Classification Error at Different K

K	1	3	5	10	20	50	100	200
error	0.296	0.252	0.224	0.220	0.220	0.204	0.208	0.208

At an initial glance, it would appear that the model does a relatively good job based on the low error rates across multiple values of K. However, at a closer inspection, we found that the K-NN classifier was not performing that well. The reason for this can be seen quite clearly when the confusion table is examined.

Table 7: Confusion Matrix When K=1

K = 1	Actual No	Actual Yes
Classified No	17	39
Classified Yes	35	159

Table 8: Confusion Matrix When K=200

K = 200	Actual No	Actual Yes
Classified No	0	0
Classified Yes	52	198

As we can see in the confusion matrices above, at larger values of K, the classifier simply classified every test case as positive (i.e. having a mental health issue). This is due to the fact that the dataset is highly imbalanced, as almost 80% of the entries in the dataset are positive. This means the balanced error will provide a more accurate insight into the performance of the classifier over the true error.

Table 9: Balanced Error at Different K

K	1	3	5	10	20	50	100	200
error	0.565	0.564	0.575	0.556	0.514	0.517	0.500	0.500

When we evaluate the balanced error, we see that the classifier does not perform well at all. In fact, the model never exceeds a 50% accuracy rate, meaning at best the K-NN classifier performs as well as simply guessing. Evaluating each predictor individually yielded similar results.

Decision Tree and Random Forest

Decision tree has the advantage of achieving zero training error on the dataset. However, due to its large flexibility, the variance of the model is very high and the predictions would differ between different training data sets. A simple way to overcome that issue, without increasing the bias, is to take the average predictions of a large number of trees. Hence, we decided to use Random Forest Classifier to predict whether an individual suffered from mental health issues.

First, we split the data set into training and test set, with a ratio of 4:1. The Random Forest Classifier was fitted in the training set. The following table is a confusion matrix of the predictions on the test set.

Table 10: Confusion Matrix on Test set

RandomForest	Actual No	Actual Yes
Predicted No	6	7
Predicted Yes	45	189

From the confusion matrix, we could see that the overall accuracy was equal to 0.789 (1 - (FP+FN) / (FP+FN+TP+TN)). This result might suggest that our model was robust and doing a good job in classifying the individuals. However, as we saw previously, the data set was highly unbalanced. Hence, we noticed that just by predicting all the individuals as positive, we would achieve an accuracy of 0.794. When we looked at the balanced accuracy (0.5*True Positive Rate + 0*5 True Negative Rate), the model scored 0.541, which was almost as bad as a random classifier.

We came up with two ways to improve our model:

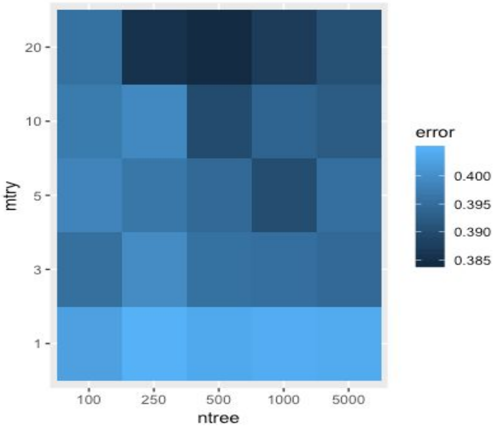
First, Random Forest is very popular because it is a model with low bias and low variance. Even though it is not very sensitive to its hyperparameters, we were able to improve the accuracy of the model by choosing the right number of trees and number of features. Also, one Random Forest' advantage is that not all the training data points are used during the fitting. Hence, the out-of-bagged (OOB) data points were used as validation data set for tuning the parameters.

We ran a grid search and used the overall accuracy to evaluate the two parameters.

Table 11: Random Forest Parameter Directory

Number of Trees	100	250	500	1000	5000
Number of Features	1	3	5	10	20

Figure 4: Heat Map Graph of the parameters' evaluation

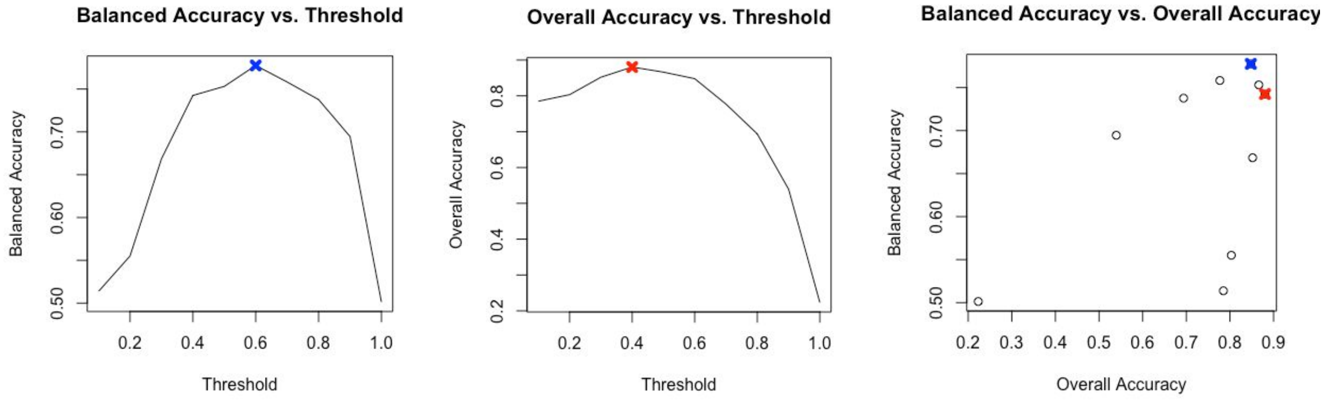


From the grid search, we obtained 500 for the number of trees and 20 for the number of features as the best candidates for the model.

Second, Random Forest also has the ability to provide the probability that a test point belongs to a certain class. Hence, similarly to a logistic regression, we ran a cross-validation to select the probability threshold for a data point to be classified as positive. Since the threshold was introduced to improve the balanced accuracy, we also kept track of that metric in addition to the overall accuracy.

Ideally, we would have used the OOB data as the validation data to identify the right threshold. However, the indexes of the OOB data points were difficult to obtain. (For the GridSearch, we were interested in the cross-validation error which was given by RandomForest()). Therefore, we split the training data into training and validation set. That way, we still had the test set to evaluate our final model.

Figure 5: Threshold selection for Overall and Balanced Accuracies.



We observed that the overall accuracy tends to decrease as the threshold increases. This was expected since the data set was largely composed of positive labels. Hence, setting a higher threshold reduced the number of individuals identified as positive and reduced the accuracy. On the other hand, the balanced accuracy had an inverse U-shaped plot. Setting a higher threshold, reduced the number of false positive and led to an increase in the true negative rate. The right threshold would balance the false positive and false negative rate.

Table 12: Accuracies of the predictions on the Validation set

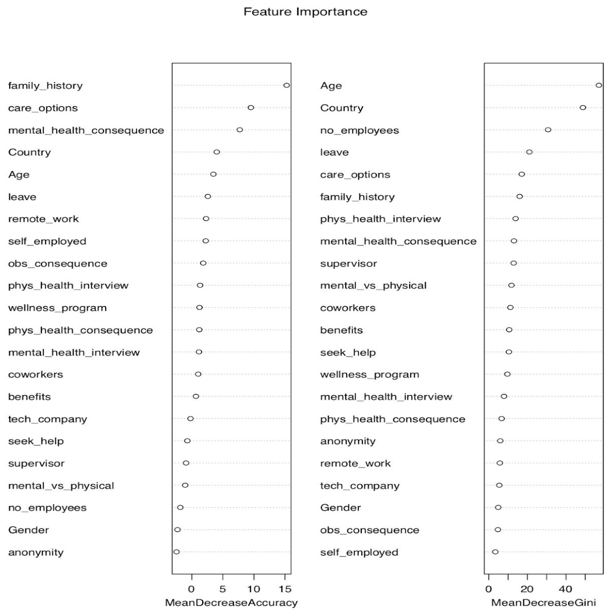
Threshold	Overall Accuracy	Balanced Accuracy
max(ov_acc): 0.4	0.880	0.742
max(bal_acc): 0.6	0.847	0.778

Table 13: Accuracies of the predictions on the Test set

Threshold	Overall Accuracy	Balanced Accuracy
max(ov_acc): 0.4	0.794	0.529
max(bal_acc): 0.6	0.745	0.665

Setting a threshold at 0.4 gave us the best accuracy on the test set, but we still obtained a low score on the balanced accuracy. Increasing the threshold solved the issue of unbalanced data. So in our situation, we obtained the best results with a threshold of 0.6.

Figure 6: Feature Importance in Mean Decrease Accuracy and Gini Index.



Finally, a Random Forest Classifier gave us the possibility to determine which features play an important role in our prediction.

Since we were basing our predictions on the accuracy of a label being positive or negative, we were interested in the decrease of the Accuracy after a split in a specific feature.

The features could be separated into two categories: individual and company characteristics. In the first category, the important features were *family_history*, *country*, and, *age*. In the second category, the features were *care_options*, *mental_health_consequence*, and *leave*. However, *tech_company* is an insignificant feature.

Conclusions

We converted the problem into a binary classification for this project. Three approaches were used: Logistic Regression, K-Nearest Neighbours, and Random Forest.

First, we noticed that deriving the classification labels from the predictor *work-interfere* resulted in a highly unbalanced data set. Almost 80% of the employees were classified as having mental health issues. Looking at our confusion matrices in the beginning, we observed that our models were predicting positive labels for almost all the entries. The overall accuracies were high, but because of the large number of false positives, the balanced accuracies were low.

Unlike the KNN approach, Logistic Regression and Random Forest methods allowed us to set a threshold to increase our performance in balanced accuracy. In Logistic Regression, using the optimal threshold for the balanced accuracy, we obtained an overall accuracy of 70.9% and a balanced accuracy of 80.8%, and using the optimal threshold for overall accuracy, we obtained 80.1% for overall accuracy and 51.2% for balanced accuracy. We achieved 74.5% and 66.5% with Random Forest. Nonetheless, it is important to note that for our problem, we were interested in the true positive rate, also called sensitivity. Indeed, misclassifying a healthy individual is less detrimental than missing an individual who suffers from mental health issues. In both models, performances were increased by setting a higher threshold (0.88 and 0.6 for Logistic Regression and Random Forest, respectively). Increasing the threshold tends to increase the false negative rate and consequently reduce the sensitivity. For this problem, Logistic Regression seems to achieve the best result. However, we need to favor the sensitivity when setting the threshold to maximize the balanced accuracy.

Moreover, individual Logistic Regression models identified *treatment, no_employees* and *phys_health_consequence (phc)* as the most important features in determining someone's mental health condition in the workplace. *care_options* and *mental_health_consequence (mhc)* were the ones from our random forest model. *care_options* represents an employee's knowledge concerning the options for mental health care from the company, and *treatment* represents whether an employee has sought treatment for mental health ailings. It is most likely that someone suffering from mental health issues would know more about their care options and they would also be more likely to actively seek treatment. However, *phc, mhc* and *no_employees* suggest that the culture of a company and its structure might have an effect on an employee's mental health. Indeed, the first two features express the concern of an employee to discuss a physical or mental health issue with the employer, and the number of employees can reflect the company's organisation. It is worth noting that none of our models suggest a relation between mental health issues and the tech industry.

Bibliography

[1] Open Sourcing Mental Illness, LTD. "Mental Health In Tech Survey." *Kaggle.com*. N. p., 2018. Web. 7 Dec. 2018.

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2017). *An introduction to statistical learning : with applications in R*. New York: Springer, E-book.

[3] "Mental Health By The Numbers | NAMI: National Alliance On Mental Illness." *Nami.org*. N. p., 2018. Web. 9 Dec. 2018.

[4] "The State Of Mental Health In America." *Mental Health America*. N. p., 2015. Web. 9 Dec. 2018.

[5] Snobar, Abdullah. "Getting Honest About Mental Health In The World Of Tech Startups." *Forbes*. N. p., 2018. Web. 9 Dec. 2018.