BAIT 509 Final Project
Mind Matters: Predicting Mental Health Challenges in Tech
Group 12: Shiwei Li, Yuchen Zhang, Young Ji Tuen
17 February 2024

1. Introduction

1.1. Background

The technology sector, known for its high-performing individuals, has recently faced significant challenges, including widespread layoffs after the COVID-19 pandemic (Gold & Trueman, 2024). Additionally, tech workers are vulnerable to burnout, characterized by exhaustion, cynicism, and negative self-evaluation (ACG Technical Editors Team, 2023).

Burnout can lead to various negative outcomes such as increased sick days, reduced productivity, and adverse effects on mental and physical health (University of Utah, 2023). In addition, it can also contribute to employee turnover as individuals lose motivation to continue in their roles.

1.2. Motivation

Tech employers have a vested interest in their employees' mental well-being. Burnout and mental health challenges not only impact individual employees but also affect workforce morale, retention rates, and organizational culture. Identifying at-risk employees and providing appropriate support can mitigate these challenges, fostering a healthier and more productive workplace environment.

1.3. Business Question

The central business question of this project is to help tech companies identify employees at risk of mental health challenges and propose strategies for supporting their mental well-being effectively.

2. Data and Statistical Question

2.1. Data

The dataset used in this project is from a 2014 survey measuring attitudes towards mental health in the tech workplace (Open Sourcing Mental Illness, 2017). The dataset has twenty-seven columns containing information on each respondent's demographics, workplace details, as well as their experience around and attitudes towards mental health.

2.2. Statistical question

The statistical objective of this project is to design a model that, given a set of data about tech employees, can predict if an employee has had experience with mental health challenges, measured by whether they have sought mental health treatment before.

3. Exploratory Data Analysis

3.1. Data cleaning

Dropping columns

After loading the dataset, we dropped the 'Timestamp' and 'comments' columns. The 'Timestamp' column contained the date and time respondents completed the survey, and the 'comments' column contained string values that were provided by a small minority of respondents. It was decided that these columns would not provide sufficient meaningful information for the classification task.

Gender

The 'Gender' feature consisted of 43 unique values, which represent the spectrum of genders respondents self-identified as. Some of the responses were clearly spelling mistakes (e.g., 'Mail', 'femail'), while others were variations of the same gender identity (e.g., 'F', 'f', 'female', 'Female'). To assist in the classification task, a function was designed to map unique responses into three groups: 'Male, 'Female', and 'Non-binary', the latter of which represents individuals who feel their gender identity cannot be defined by a gender binary (LGBT Foundation, 2024). This function was applied to the entire dataset.

Positive and negative labels

We reassigned the target column values to 1 or 0. 'Yes' responses were reassigned to the value of 1, and 'No' responses were reassigned to the value of 0. Since the business question for this project is to identify employees for whom tech companies can do more to support, it was decided that the primary statistical objective would be to capture employees who have sought treatment for mental health challenges as well as possible.

Separating US and non-US data

It was determined that the dataset should be split into observations from the United States (US) and observations not from the US. While observations from the US had 'state' information, not all observations from other countries had a 'state' value. To prevent loss of potential information by dropping the 'state' column, the dataset was separated into two. The 'state' column was then dropped from the non-US dataset, while the 'Country' column was dropped from the US dataset.

Splitting train and test sets

Next, the datasets were split into training and test sets, where the test sets contained 20% of the total number of observations. A random state of 123 was applied to ensure replicability.

3.2. Data exploration

In the US training data, there were 7 null values for 'state', 8 for 'self_employed', and 15 for 'work_interfere'. As for the non-US training data, there were 6 null values for 'self_employed', and 94 for 'work_interfere'. This suggests that there should be imputation when features are preprocessed to fill in missing values.

The training datasets for both US and non-US observations were relatively balanced in their distribution of target classes. 54.5% and 54.9% of observations had a positive label in the US and non-US training sets, respectively.

The distributions of features by target class were plotted, and displayed in Sections 7.1 and 7.2. Features seemed to demonstrate similar distributions between classes, except for the 'work_interfere' variable which asks respondents if their mental health condition interferes with their work. Employees in the 'no treatment' class mostly responded 'Never', or had no response, while employees in the 'treatment' class mostly responded 'Sometimes'. In addition, there was some differentiation between classes in the 'family_history' feature, which asks respondents if they have a family history of mental illness. Most respondents in the 'no treatment' class indicated 'No' while those in the 'treatment' class mostly answered 'Yes'.

4. Method and Results

4.1. Feature pre-processing

A pipeline was built to preprocess numerical features, first by imputation using median values, and then by scaling features using a standard scaler. Another pipeline was built to preprocess categorical features, first by imputation by filling all null values with 'missing', then with a one hot encoder that ignores unknown values in test set data. A third pipeline was built to preprocess binary features in a similar way to categorical features, with the additional step of dropping columns for the second class (e.g., dropping a 'No' column if a feature only has 'Yes' and 'No' values).

Finally, a column transformer combined the three transformers into a main preprocessor pipeline. One preprocessor pipeline was built for the US models, and another for the non-US models since the US and non-US datasets have different categorical features. This column transformer will be used in the subsequent models (section 4.4) to preprocess the data during cross validation. At this point, the preprocessors were fitted to the US and non-US training data to obtain a count of the number of total features to be included in model development. Thirty percent of these features were to be included in the feature selection (see Section 4.4.2 for more details).

4.2. Selected performance metrics

In evaluating the model we selected, our assessment focused on the f1 scoring metrics due to our goal of balancing precision and recall. This section outlines the rationale behind choosing this specific metric and its importance in the context of our project.

The goal of our analysis was to accurately identify employees who need mental treatment while also making judicious use of limited company resources. For the first goal, recall may appear to be the best metric as it measures a model's ability to identify positive instances (employees who have sought mental health treatment). For the second goal, precision may be more appropriate as it measures the accuracy of positive predictions made by the model, and a higher precision score can allow for more parsimonious resource utilization. Thus, the f1 scoring metric was determined to be the best balance between the two goals.

The f1 score is the harmonic mean of precision and recall, providing a single metric to assess the model's accuracy in cases where finding a balance between precision and recall is essential. This

metric is crucial for applications, such as this project, where both the costs of missing a positive case (people who need mental treatment) and the costs of false positives (people who don't need treatment but are identified as having mental issues) are significant.

4.3. Dummy model

A dummy classifier was built to serve as a baseline for comparison. The 'most_frequent' strategy was used, and the model was fit with the training data. This produced a training score of 0.545, and a test score of 0.550 when trained with the USA dataset. The cross validation f1 score was 0.706.

When trained with the non-USA observations, the dummy classifier produced a training score of 0.549 and a test score of 0.569. The cross validation f1 score was 0.

4.4. Hyperparameter tuning, feature selection, and fairness optimization

To identify the optimal model for this classification task, four classification models were developed and tuned. The f1 cross validation scores of these models were then compared.

4.4.1. Decision tree classifier

For both the US and non-US models, a pipeline was built, incorporating the preprocessors from section 4.1, as well as a decision tree classifier model.

A randomized search cross validation was conducted with this pipeline to find the best values of *max_depth* between values of 5, 7, and 9, 11 and *min_impurity_decrease* between values of 0.001, 0.01, 0.1, and 0.5. Tuning *max_depth* will prevent the classifier from creating infinite splits until the model overfits the training data. Optimizing *min_impurity_decrease* will ensure that a node will be split only if the split induces a decrease of the impurity greater than the specified value (Scikit Learn, n.d.). This ensures that the model does not become too complex by fitting perfectly to the training data.

For the US model, the best *max_depth* value was 5, and the best *min_impurity_decrease* value was 0.1. The model scored a cross validation f1 score of 0.874. For the non-US model, the best *max_depth* value was 7, and the best *min_impurity_decrease* value was 0.01. The model scored a cross validation recall score of 0.778.

4.4.2. kNN

We created pipelines for both the US and non-US datasets using the k-Nearest Neighbors (kNN) model, which included preprocessors and used RFECV for feature selection. Two folds were specified for the RFECV validation process, and the 'step' parameter was set to four, to remove only four features at each iteration of the elimination process. This was decided to be a compromise between efficiency and thoroughness of model optimization. We decided on selecting 30% of the processed features during RFECV, and this number was stored as the 'feature_number_select' object under the 'Preprocessing' section of the project code. Of note, the subsequent models (SVC and logistic regression) utilized the same feature selection method.

Then, we implemented a grid search cross-validation to find the best number of neighbors ('n_neighbors'), specifically 3, 5, 7, and 9, aiming to maximize the f1 score.

For the US model, the best *n_neighbors* parameter found was 7 neighbors, yielding an f1 score of 0.846. The non-US model's optimal *n_neighbors* was 7 neighbors, with an f1 score of 0.720.

4.4.3. SVC

A Support Vector Classifier (SVC) model was built using a pipeline that included data preprocessing and feature selection steps equivalent to that in the kNN model. Then, an SVC classifier was incorporated into the pipeline. A grid search cross validation was also conducted to identify the best combination of *gamma* and *C* hyperparameters for the SVC model.

The best hyperparameters found for both datasets were *C* = 1.0 and *gamma* = 0.1. The model achieved a cross-validation F1 score of 0.874 for the US dataset and 0.799 for the non-US dataset.

4.4.4. Logistic regression

US and non-US pipelines were also built for a logistic regression model, incorporating the preprocessors and an RFECV optimizer to select the best features for inclusion.

A randomized search cross validation was conducted with this pipeline to find the best value of *C* from a range of 100 equally spaced numbers between 0.1 and 100, inclusive.

For the US model, the best hyperparameter obtained from cross validation was *C* = 39.45 and the best cross validation f1 score was 0.869. For the non-US model, the best parameter obtained from cross validation was *C* = 67.71, and the best cross validation f1 score was 0.775.

4.5. Model selection

Model	Dataset	f1 cross validation score
Decision tree classifier	US	0.874
	Non-US	0.779
K nearest neighbours	US	0.846
	Non-US	0.720
Support vector classifier	US	0.874

	Non-US	0.799
Logistic regression	US	0.869
	Non-US	0.775

The best scoring model was the Support Vector Classifier, with the best f1 cross validation score of 0.874 for the US model, and 0.799 for the non-US model.

The US model achieved a final training score of 0.876 and test score of 0.886. The weighted average f1-score of this model was 0.87. In addition, the recall score for the 'treatment' class was 0.94, indicating that the model was doing a good job identifying employees who had sought treatment for mental health challenges. These scores are markedly higher than the dummy model.

The non-US model achieved a final training score of 0.817 and test score of 0.792. The weighted average f1-score of this model was 0.79. The recall score for the 'treatment' class was also high, at 0.91. Again, the non-US model scored better than the dummy classifier.

For the confusion matrix of the US and non-US models, see Sections 7.3 and 7.4.

4.6. Fairness optimization

Fairness in machine learning models is important to ensure that decisions do not disproportionately disadvantage any subgroup within the population, particularly when sensitive attributes such as gender, race, or age are considered. We selected False Negative Rate (FNR) Parity as our fairness metric to evaluate the equity of our model's predictions across different genders.

FNR parity refers to the condition where the FNRs across subgroups defined by sensitive attributes are equal or within an acceptable margin. A false negative occurs when the model incorrectly predicts a negative outcome for a positive case. High FNRs can be especially harmful in contexts where failing to identify positive cases can deny individuals critical opportunities or services. In our case, failing to identify employees who have had experience with mental health challenges can be costly for companies whose staff may feel unsupported by their employers, potentially compromising morale and contributing to turnover.

Achieving FNR parity ensures that our model does not systematically fail to identify positive cases more frequently in one gender than others. This focus on minimizing and equalizing false negatives can ensure that our model contributes to equitable support for all genders.

After fairness optimization, the FNR parity of both the US and non-US models did not demonstrate significant improvement. In addition, the demographic disparity (0.68 to 0.65 for the US model and 0.56 to 0.52 for the non-US model) and equal odds ratios (stayed constant at 0 for the US model and dropped from 0.36 to 0.32 for the non-US model) dropped slightly. It is possible that

an imbalance in the gender feature contributed to this phenomenon. Since the 'Non-binary' gender group is much smaller in size than the 'Female' or 'Male' groups, and the 'Male' group is much bigger than the 'Female' group, the fairness optimization process may have struggled to achieve fairness without compromising model performance. Additional data that allows for a more balanced distribution between genders may be needed to optimize fairness.

5. Communication of Results and Advice to a Non-expert

5.1. Project approach and results

Using machine learning techniques, we tested several predictive models and compared their performance. The aim of this project was to develop the best model to predict whether tech employees have sought treatment for mental health challenges before or not. This can, thus, serve as a proxy for identifying employees more at risk for facing mental health struggles whilst working. The evaluation method we used to choose the final model helps us to achieve high predication accuracy, as well as to avoid using up limited resources on employees who may not actually require mental health support. According to our results, the Support Vector Classifier (SVC) model performed the best.

The SVC model can be seen as a predictive tool. To develop this model, we cleaned and organized survey data from tech employees. Then, we trained the SVC model to recognize data patterns, akin to training a detective to find clues. We also used a method called grid search to find the best settings for our tool to work, identified as $C = 1.0$ and $gamma = 0.1$. These settings have been optimized to ensure that the tool performs well not only on the data provided in this project, but also on new data of other employees tech companies may want to classify.

Our final model scored the best among all the tested models in a metric called an f1 score, which combines how precise the detective is (how often they're right when they say someone needs help) with how good they are at finding everyone who actually needs help. Of note, these comparisons were conducted on data reserved for training the models. Twenty percent of the original data was set aside for a final round of evaluation of the chosen model. For the US, our chosen model scored a strong 87.4 out of 100, meaning that it is good at both spotting people who need help and not missing too many who do. For the rest of the world, the score was a bit lower at 79.9, but it still shows that the model is quite capable, just slightly less so than in the US. This model was, thus, chosen as the final predictive tool to be used.

On data that was reserved for final evaluation, in the US, the final predictive tool was right 88.6% of the time, successfully spotting individuals needing support. For people outside the US, the tool also performed quite well, with a success rate of 79.2%.

Having ensured that the comparison metric and final evaluation scores were optimal relative to the other tested models, this final tool can be applied to help tech companies identify employees who may need enhanced mental health support.

5.2. Recommendations for use

The final tool can be used to identify the employees in tech companies who have sought mental health treatment. The company can then use this information to decide who to target for enhanced mental health support. Support can come in the form of an on-site counselor, or through online or one-on-one therapy that is covered through a firm-wide plan, or covered. Therapy can be an effective way for employees to navigate work- and personal-related mental health challenges, thus improving productivity and work relationships.

Furthermore, if the company identifies a disproportionately high number of employees who are at risk of developing mental health challenges whilst at work, management may want to consider structural changes to foster an environment more conducive to mental health and wellbeing. Some examples of this may be to emphasize role clarity, as well as to improve communication and support from direct managers. A positive work environment that promotes collaboration, transparency and rewards for hard work is also critical to making sure employees feel valued. In addition, work flexibility has become a growing priority for today's workforce, with most millennials indicating more motivation to stay at a company who provides flexible work arrangements (Petrova, n.d.).

5.3. Possible areas for improvement

As the data used to develop this predictive tool were from 2014, it should be updated regularly with more recent data to reflect evolving workplace and emerging stressors. This keeps our approach to identifying at-risk employees and supporting mental health both accurate and up-to-date.

Increasing the amount of data can also significantly improve a model's reliability by providing it with more examples to learn from, which can enhance its ability to work for a more diverse population of employees. Furthermore, collecting data over time can capture changes and trends which can provide insight into whether the treatment provided by the company is effective to the employees over time.

In addition, the model only considers male, female, and non-binary genders. Thus, more granular and valuable information for genderqueer, genderfluid, agender, bigender, and other gender identities may be overlooked. With the increase of data in the future, we may be able to incorporate a more inclusive consideration of gender and gain more insight for non-binary employees, who may face a unique set of mental health challenges.

6. References

Gold, J., & Trueman, C. (2024, January 26). *Tech layoffs in 2024: A timeline.* Computerworld. https://www.computerworld.com/article/3685936/tech-layoffs-in-2023-a-timeline.html

ACG Technical Editors Team. (2023, June 8). *Tech worker burnout: What it is, and how to deal with it.* Pluralsight. https://www.pluralsight.com/resources/blog/cloud/tech-it-worker-burnout#:~:text=In%20a%20research%20report%20by,see%20the%20value%20in%20it

University of Utah. (2023, March 28). *How burnout impacts your mental health.* Healthcare. https://healthcare.utah.edu/healthfeed/2021/06/are-you-burned-out

LGBT Foundation. (2024, January 9). *What it means to be non-binary.* LGBT Foundation. https://lgbt.foundation/help/what-it-means-to-be-non-binary/

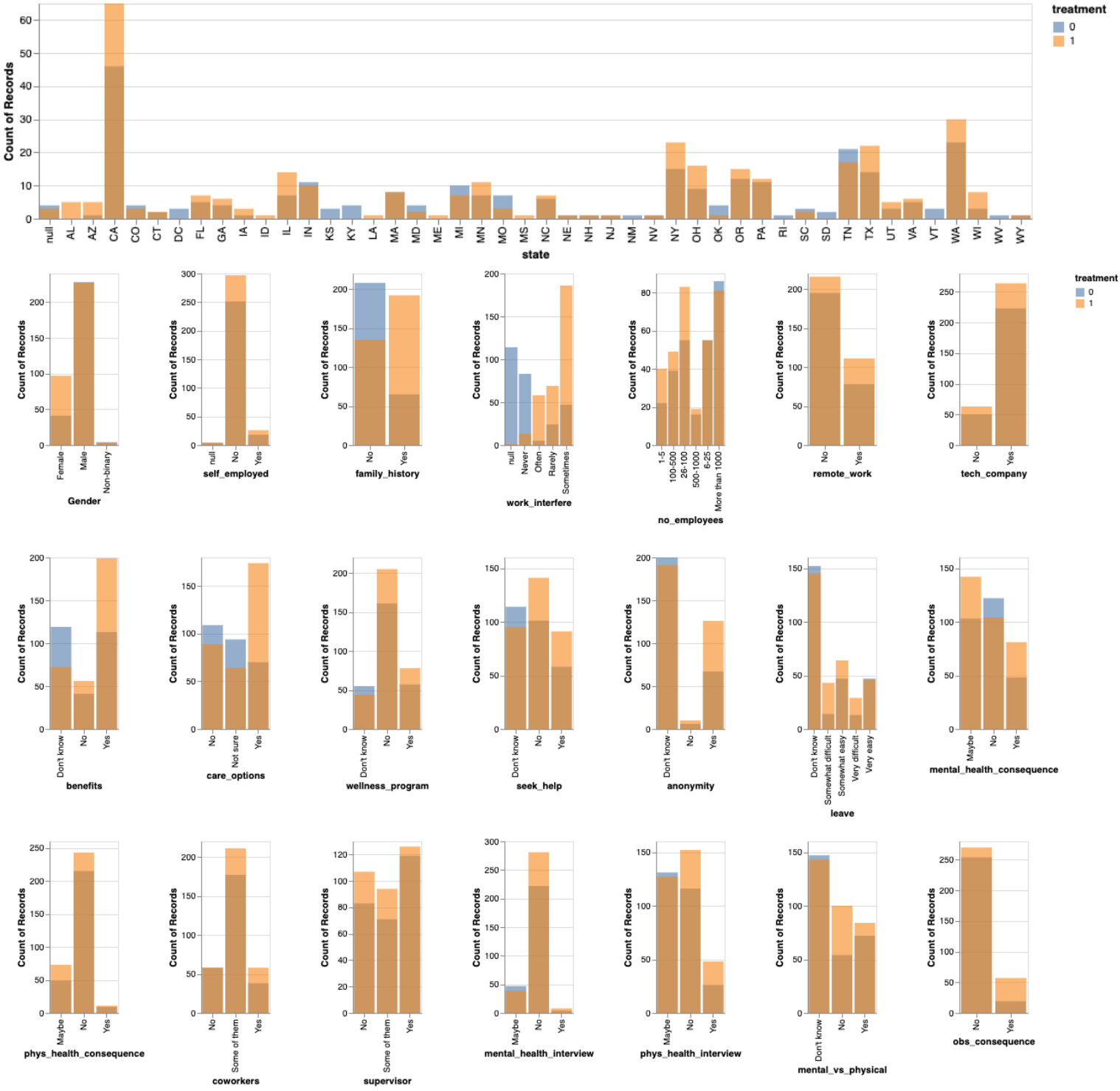
Open Sourcing Mental Illness, Ltd. (2017). *Mental health in tech survey.* [Data set]. Kaggle. https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey

Scikit Learn. (n.d.). *sklearn.tree.DecisionTreeClassifier.* Scikit Learn. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

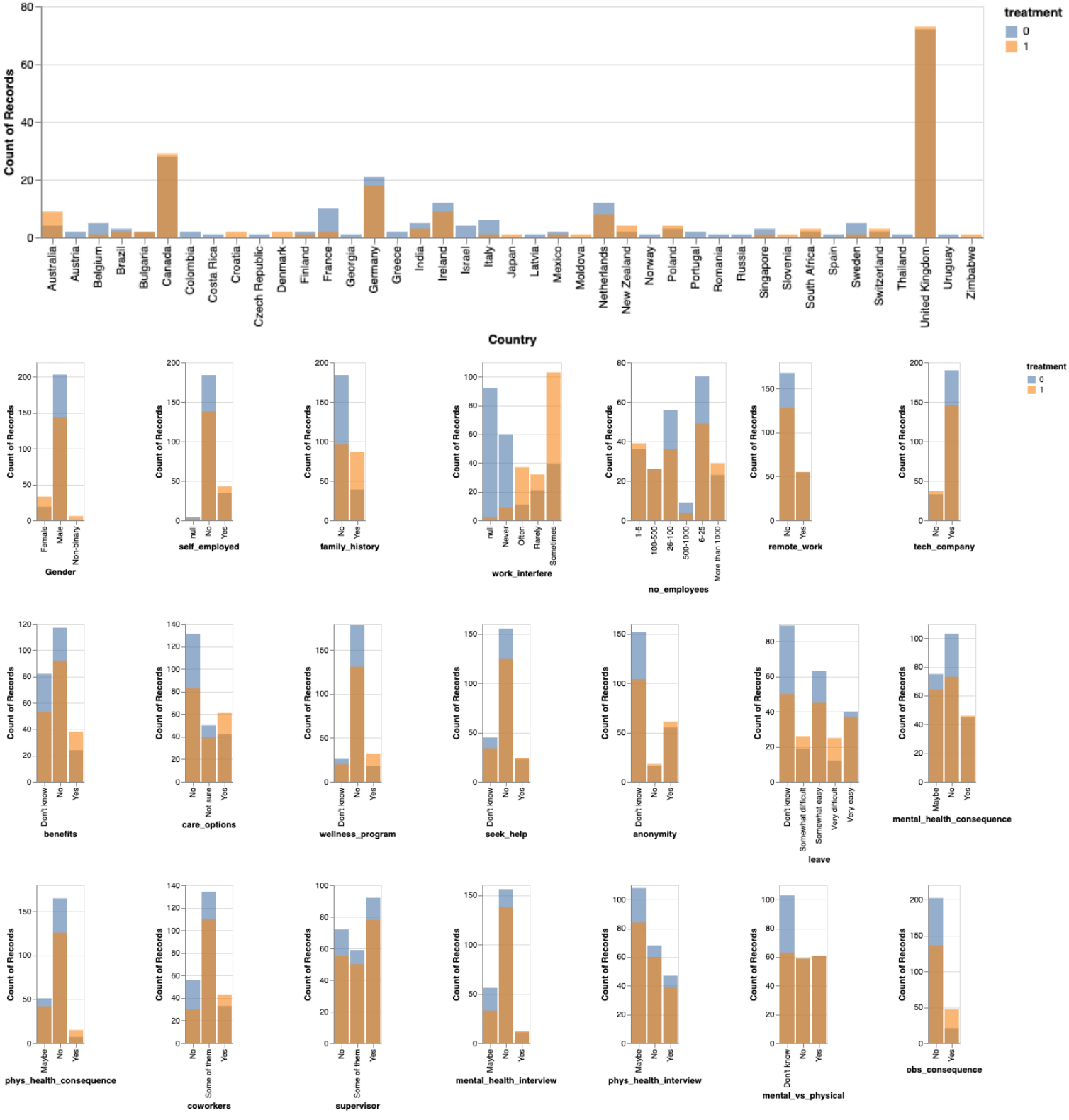
Petrova, S. (n.d.). *Burnout in the tech industry and what to do about it.* Adeva. https://adevait.com/blog/workplace/burnout-tech-industry#how-do-you-reduce-employee-burnout-in-software-development

7. Appendices

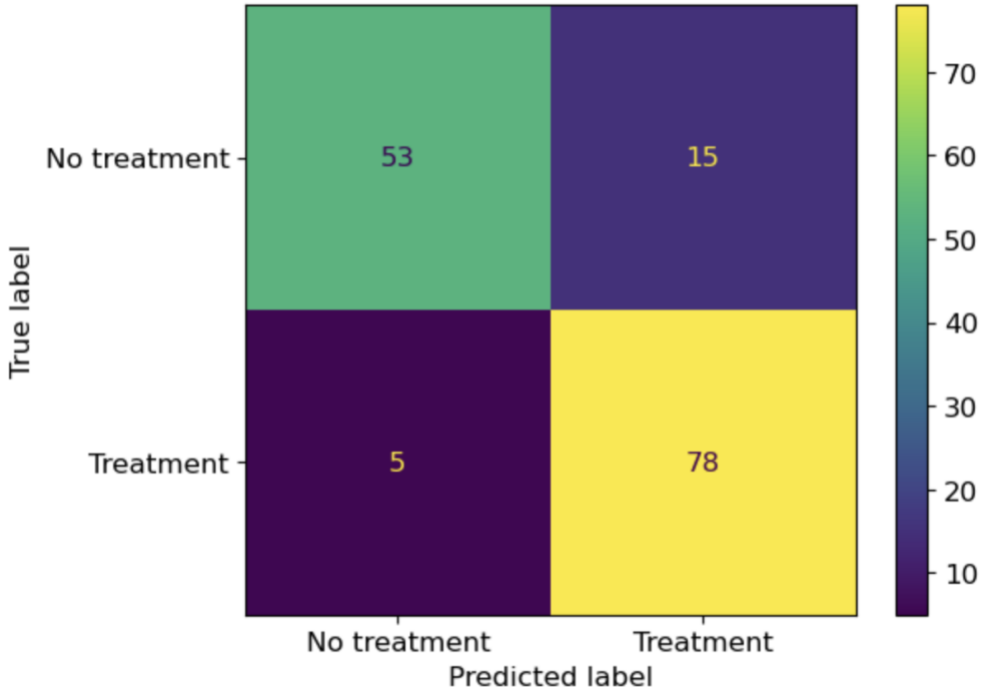
7.1. Distribution of features in the US training dataset



7.2. Distribution of features in the non-USA training dataset

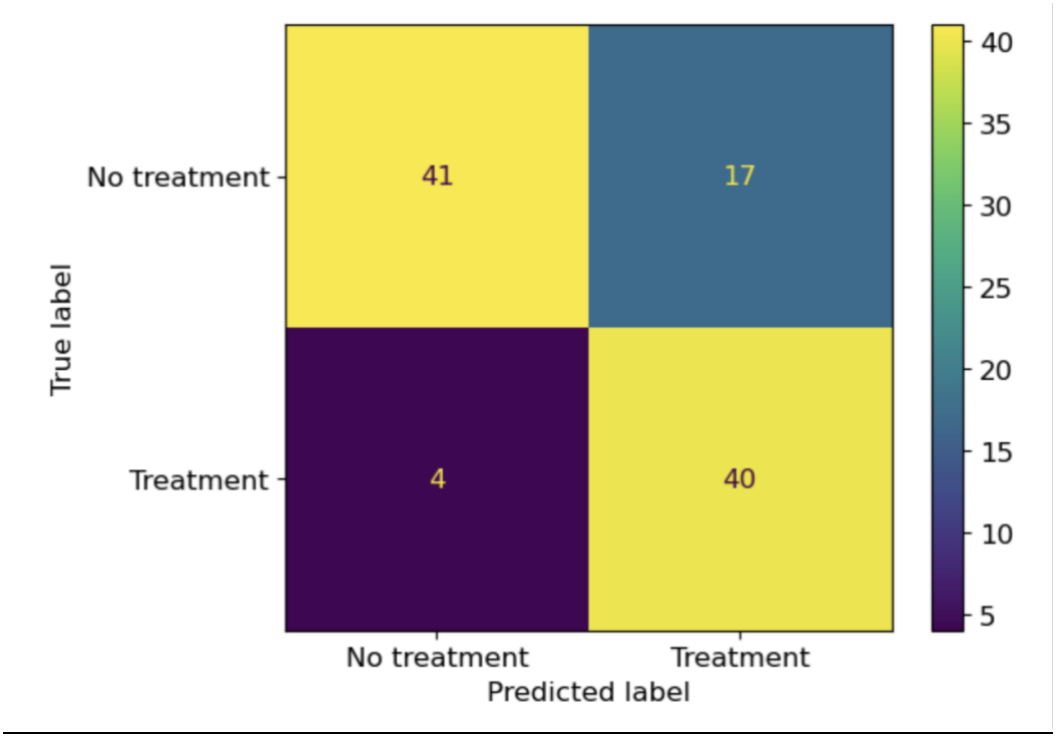


7.3. Evaluation metrics for the final US model



	Precision	Recall	f1-score	Support
No treatment	0.91	0.78	0.84	68
Treatment	0.84	0.94	0.89	83
Accuracy	-	-	0.87	151
Macro average	0.88	0.86	0.86	151
Weighted average	0.87	0.87	0.87	151

7.4. Evaluation metrics for the final non-USA model



	Precision	**Recall**	**f1-score**	**Support**
No treatment	0.91	0.71	0.80	58
Treatment	0.70	0.91	0.79	44
Accuracy	-	-	0.79	102
Macro average	0.81	0.81	0.79	102
Weighted average	0.82	0.79	0.79	102