Week 1: Mental Health at Work

Authors: Tony, Navya, Kunaal, and Jaya (Group 4)
Course: DS 4002 - Data Science Final Project Course
Professor: Brian Wright
Date: January 8th, 2021

Day 1: Background, Dataset Overview, & Initial Hypotheses

Background & Motivation

- In 2017, WHO estimated that 1 in 17 adults experience a serious mental illness each year[1]
  - More than 44 million adults are affected annually by mental illnesses, many of whom are also active within the workforce[2]
- Poor mental health and stress can negatively affect employee job performance, work engagement, communication with coworkers, physical capability, and other day-to-day functions[3]
- Only 57% of employees who report moderate depression and 40% of those who report severe depression receive treatment to control symptoms[3]
- Due to COVID, mental health is increasingly affecting work life
  - 55% of employees feel uncomfortable confiding in anyone at work[4]
  - Remote work can either be an alleviator or exacerbator of a mental illness

Dataset Overview: Mental Health in Tech

This dataset originates from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. The original dataset is from Open Sourcing Mental Illness (OSMI).

Link: https://www.kaggle.com/osmi/mental-health-in-tech-survey[5]

- **Timestamp**
- **Age**
- **Gender**
- **Country and state (if United States)**
- **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?

Dataset Overview: Mental Health in Tech (Continued)

- **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 employer provide resources to learn more about mental health issues and how to seek help?
- **anonymity**: Is anonymity protected if employee takes advantage of mental health or substance abuse treatment?
- **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

Relevant Research

- CDC Workplace Health Guide - Mental Health in the Workplace[3]
  - Tracks mental health solutions, awareness frameworks, and strategies
  - Encourages employers to monitor indicators and risk factors of mental health such as **stigma**, **lack of health care,** and **lack of social connections**
- Predictors of repeated sick leave in the workplace because of mental disorders (Sado et al.)[6]
  - Analyzed Return to Work (RTW) and repeated sick leave rates among 194 subjects employed at a manufacturing company
    - Exploratory Variables: RTW, sex, age at time of employment, job tenure, diagnosis, etc.
  - Methods: Univariate Analyses using log-rank test and multivariate analysis using Cox proportional hazard model
  - Results: Strongest predictors of repeated sick leave were found to be **age** and **previous sick-leave episodes**

Initial Hypothesis

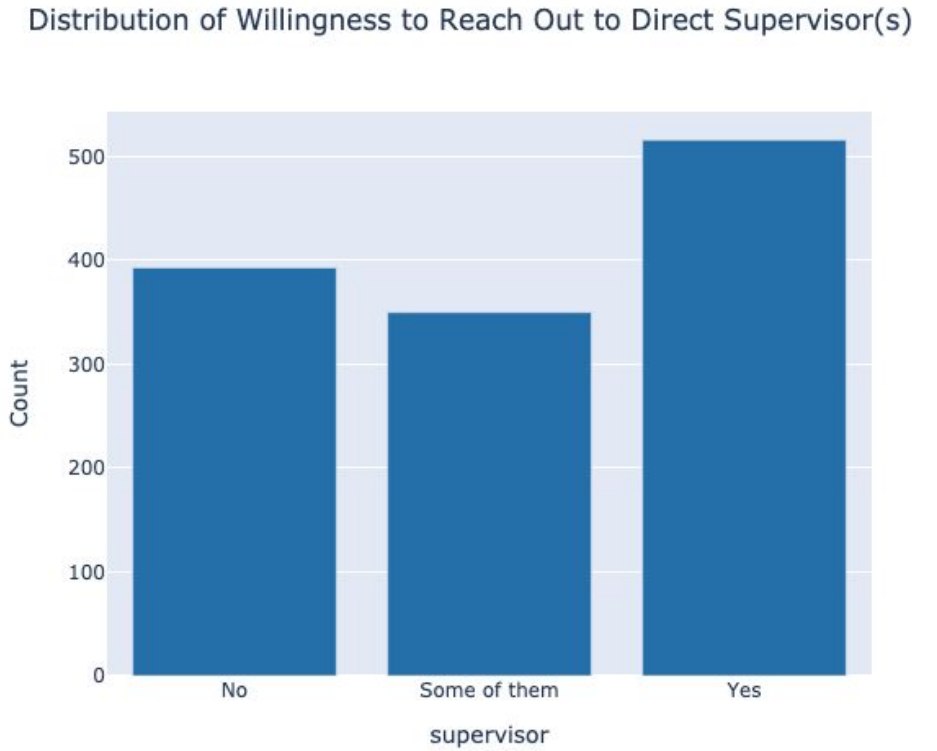
- Target predictors: Would you be willing to discuss a mental health issue with your direct supervisor(s) {e.g., **supervisor** in the dataset}?
  - ***Obs_consequence***: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
  - ***No_employees:*** 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?
  - ***Benefits:*** Does your employer provide mental health benefits?
- **Null Hypothesis:** The 4 target predictors do not constitute the majority (50%) of Random Forest feature importance when predicting whether employees are willing to discuss mental health issues with supervisors
- **Alternative Hypothesis:** The 4 target predictors constitute the majority (50%) of Random Forest feature importance when predicting whether employees are willing to discuss mental health issues with supervisors

Day 2: Exploratory Data Analysis, Finalized Hypotheses, & Initial Model Plan

Exploratory Data Analysis

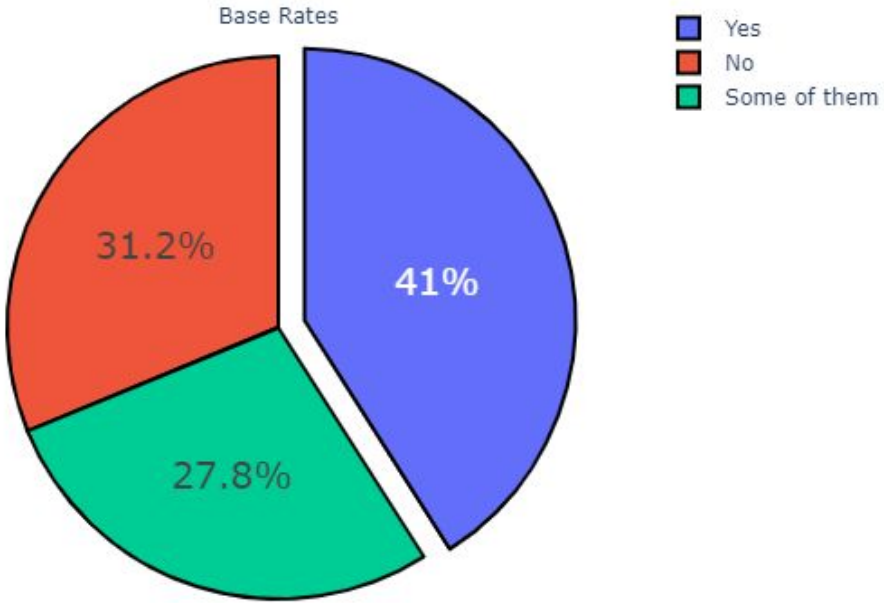
- Questions of interest
  - Generally, how willing are employees to reach out to their direct supervisor(s) about a mental health issue?
  - Do demographic indicators such as **age**, **gender**, and **country** play a role in how willing employees are to reach out to their direct supervisor(s)?
  - How much of a role do our 4 initial target predictors play a role in the willingness of employees to reach out to their direct supervisor(s)?
    - **no_employees, obs_consequence, remote_work, benefits**
  - How much of a role do two target predictors suggested by our peers play a role in the willingness of employees to reach out to their direct supervisor(s)?
    - **leave, seek_help**
  - What is the multicollinearity between factors in our dataset? Which factors are highly correlated with our target label (**supervisor**)?
- Based on the answers to these questions, we can change the way we think about our general question, as well as restate our hypotheses to that question

QUESTION 1: Generally, how willing are employees to reach out to their direct supervisor(s) about a mental health issue?



Visual Representation of the distribution of responses for our label

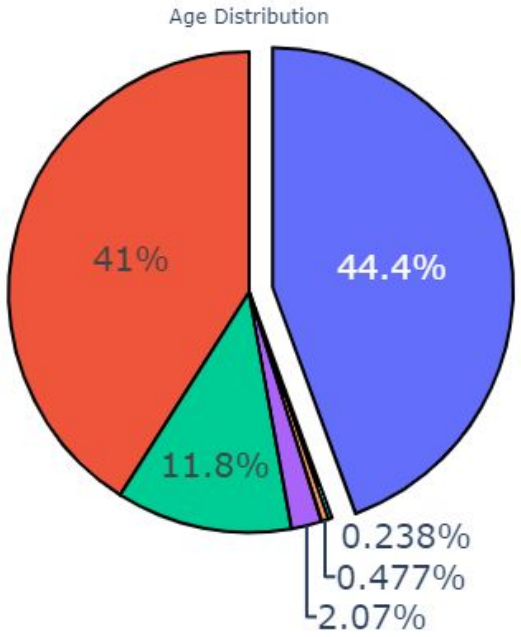
QUESTION 1: Generally, how willing are employees to reach out to their direct supervisor(s) about a mental health issue?

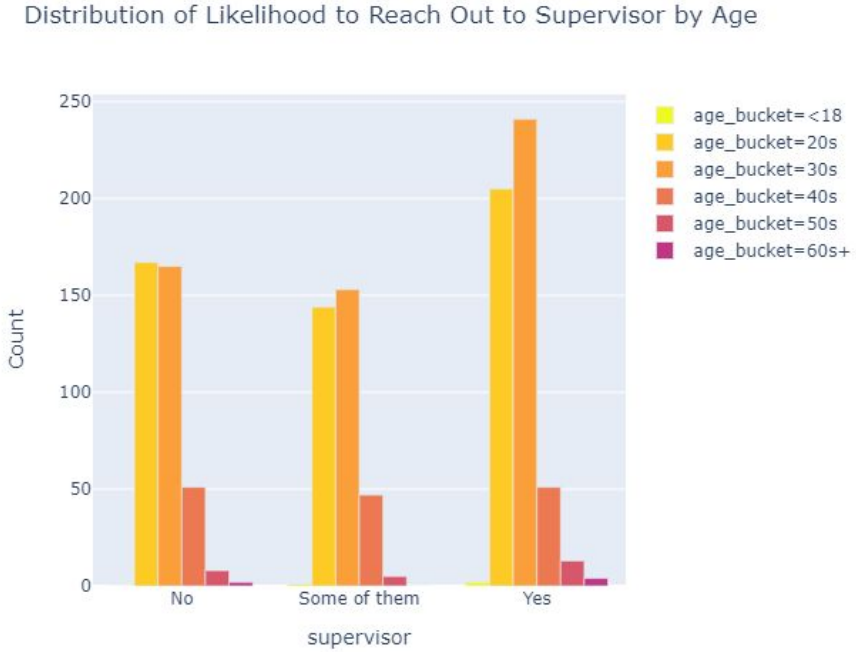


Response	Base Rate Percentage
Yes	41%
Some of them	27.8%
No	31.2%

QUESTION 2: Do demographic indicators such as *age*, *gender*, and *country* play a role in how willing employees are to reach out to their direct supervisor(s)?

AGE



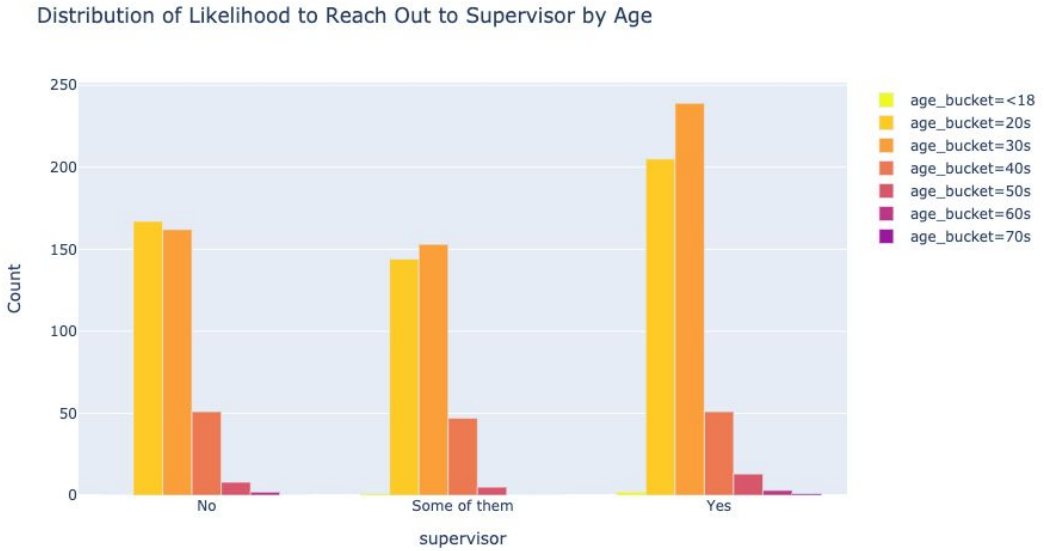


QUESTION 2: Do demographic indicators such as *age*, *gender*, and *country* play a role in how willing employees are to reach out to their direct supervisor(s)?

AGE

Outcome: There was not much variation in response from the created age groups, though the younger age groups (20s and 30s) showed an increase in 'Yes' responses.

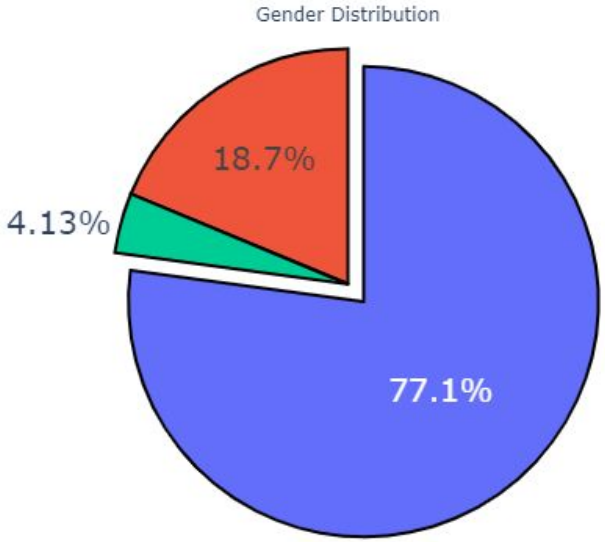
Conclusion: Due to only two groups showing noticeable variation, we decided that we would not add age into our group of target predictors.

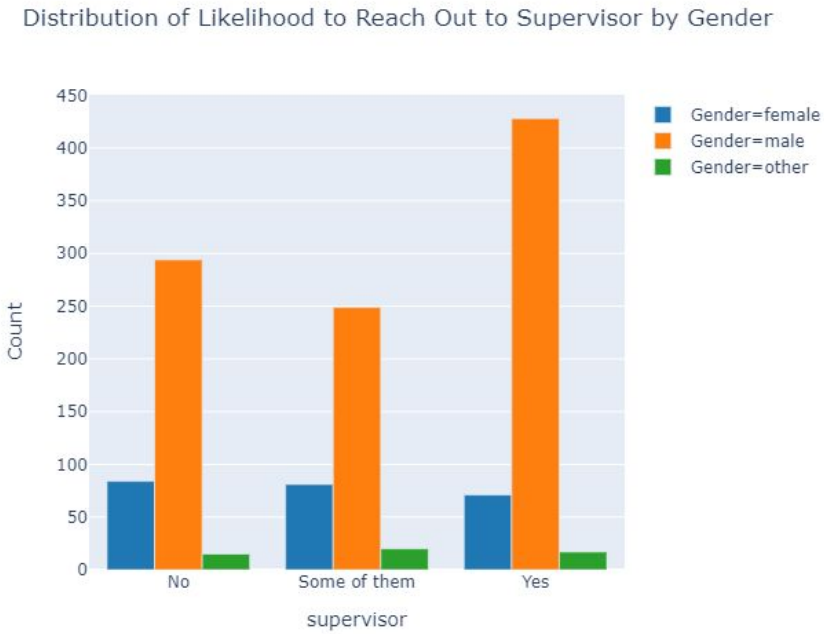


Feature	chi2	p-value
Age	0.3302	0.8478

QUESTION 2: Do demographic indicators such as *age*, *gender*, and *country* play a role in how willing employees are to reach out to their direct supervisor(s)?

GENDER



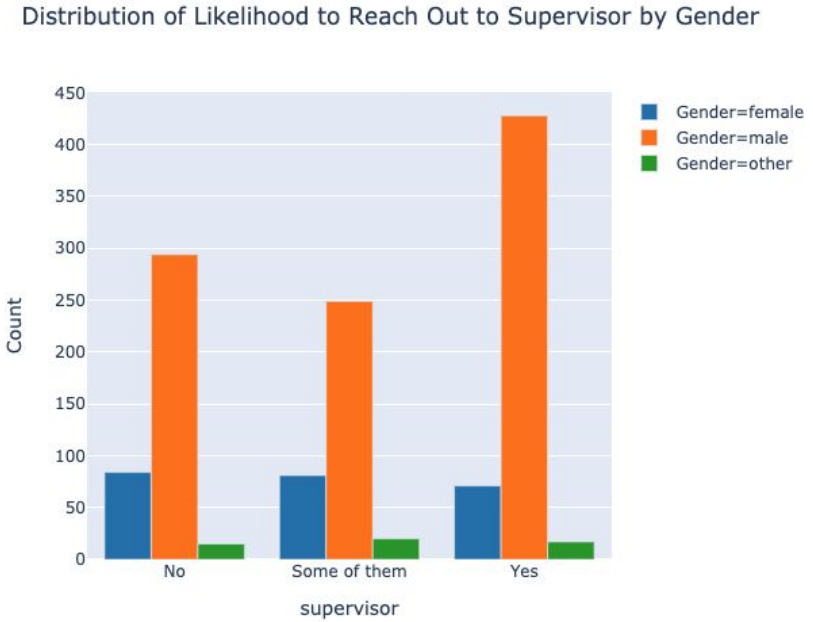


QUESTION 2: Do demographic indicators such as *age*, *gender*, and *country* play a role in how willing employees are to reach out to their direct supervisor(s)?

GENDER

Outcome: There was not much variation in response from the female and other gender groups. The men did see a higher amount that felt comfortable reaching out to a supervisor.

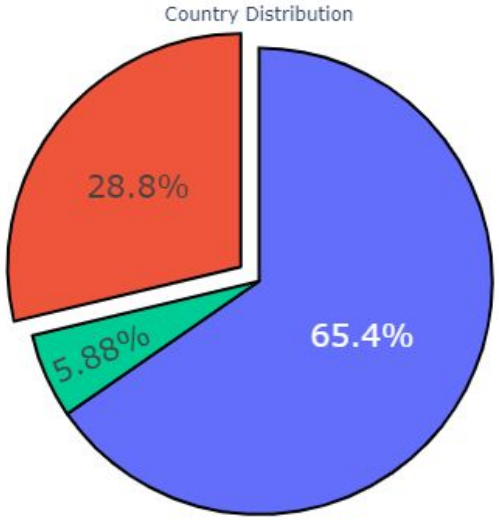
Conclusion: We believe this rise in response from men was only due to the fact that there is a larger amount of men who took the survey. Therefore, we will not consider the gender variable in our hypothesis.

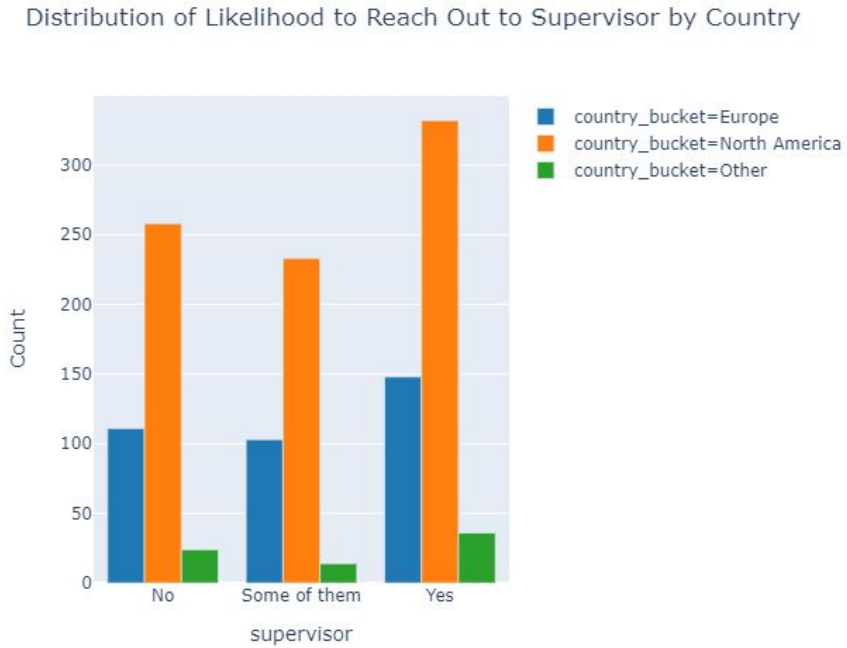


Feature	chi2	p-value
Age	1.7637	0.4140

QUESTION 2: Do demographic indicators such as *age*, *gender*, and *country* play a role in how willing employees are to reach out to their direct supervisor(s)?

COUNTRY



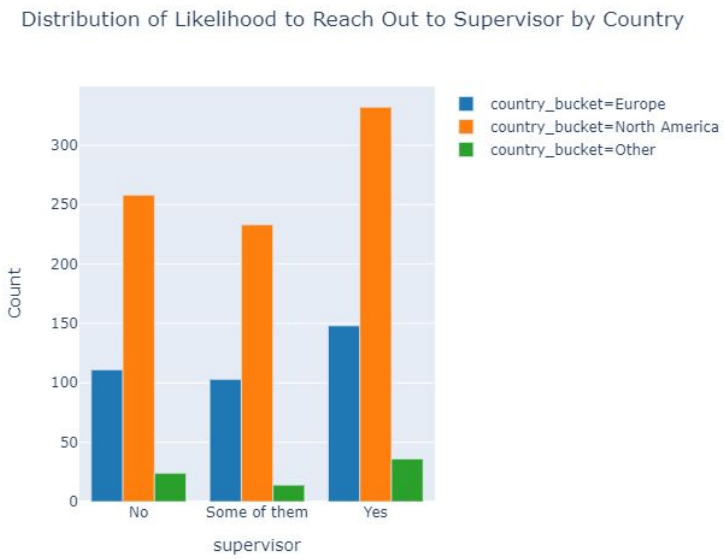


QUESTION 2: Do demographic indicators such as *age*, *gender*, and *country* play a role in how willing employees are to reach out to their direct supervisor(s)?

COUNTRY

Outcome: Qualitatively, it can be observed that each Country bucket's trend across different instances of the response variable is quite indistinct. Furthermore, the p-value as a result of a chi-square analysis is 0.8126, which does not allow us to reject the null hypothesis.

Conclusion: The combination of qualitative and quantitative results from the EDA suggested that country does not have a strong impact on the likelihood of reaching out to a direct supervisor.



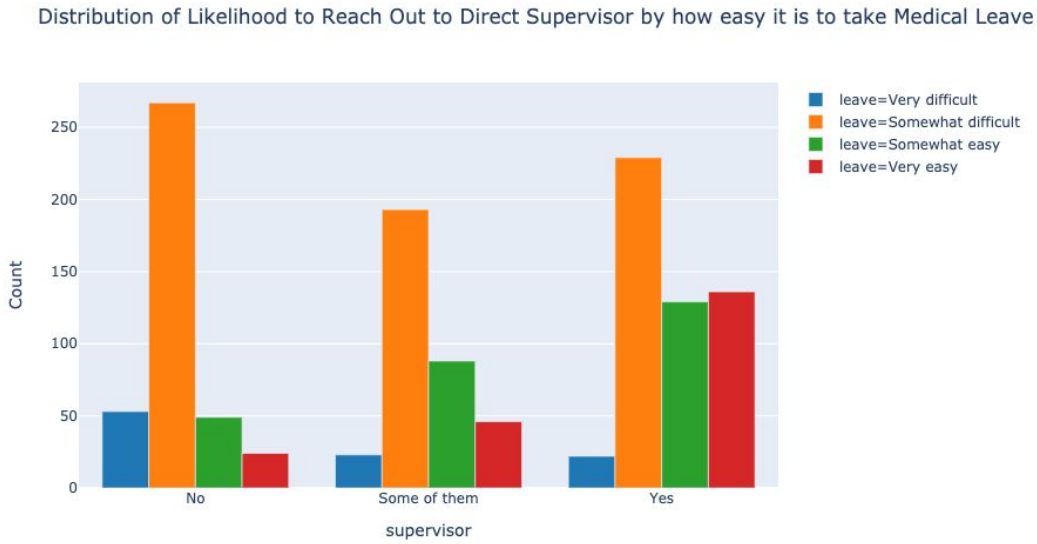
Feature	chi2	p-value
Country	0.4152	8.1255e-01

QUESTION 4: How much of a role do two target predictors suggested by our peers play a role in the willingness of employees to reach out to their direct supervisor(s)?

leave

Outcome: As the difficulty to take medical leave for a mental health condition decreased, the likelihood of communicating with a supervisor increased. The inverse was also true.

Conclusion: Originally, we did not think leave would be a main predictor of willingness of an employee to speak with a supervisor. Due to the outcome, we decided the variation in responses was high enough to add the variable to our target predictors.



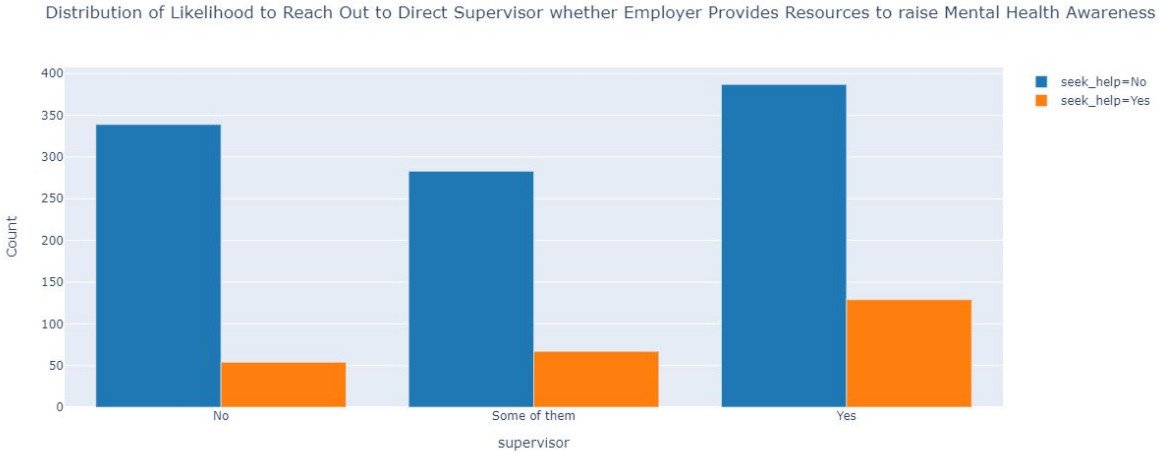
Feature	chi2	p-value
leave	81.8797	1.6598e-18

QUESTION 4: How much of a role do two target predictors suggested by our peers play a role in the willingness of employees to reach out to their direct supervisor(s)?

seek_help

Outcome: The qualitative distribution of **supervisor** as a function of **seek_help** depicted no distinct trend. Furthermore, the p-value as a result of performing a chi-square analysis in relation to the **supervisor** response variable was found to be 7.5876e-04.

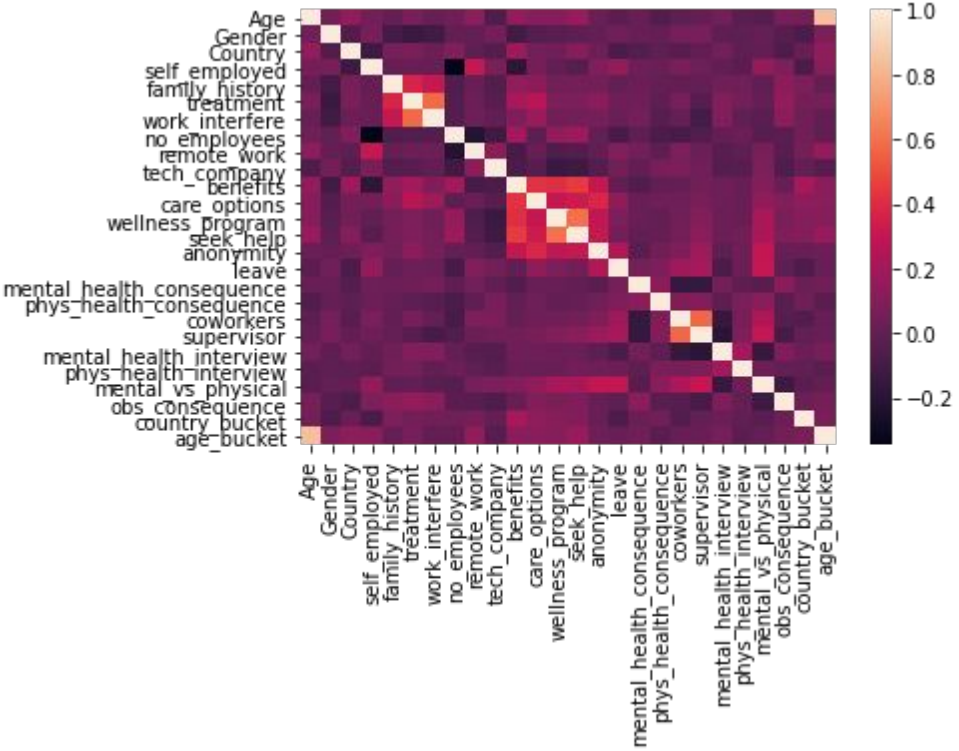
Conclusion: The combination of the qualitative and quantitative observations listed above suggested that **seek_help** would not sufficiently predict the class of **supervisor**. We decided to omit it from our target predictor set.



Feature	chi2	p-value
seek_help	14.3677	7.5876e-04

QUESTION 5: What is the multicollinearity between factors in our dataset? Which factors are highly correlated with our target label (**supervisor**)?

- The highest correlations with the **supervisor** variable (e.g. the label we are trying to predict):
  - **coworkers -** willingness to speak to a coworker
    - 0.57
  - **mental_vs_physical** - mental health is taken as serious as physical
    - 0.31



QUESTION 5: What is the multicollinearity between factors in our dataset? Which factors are highly correlated with our target label (**supervisor**)?

FEATURE	Correlation	FEATURE	Correlation
age_bucket	0.0163	*wellness_program*	0.1040
Gender	0.0681	*seek_help*	0.1193
country_bucket	-0.0054	*anonymity*	0.1798
self_employed	0.0374	*leave*	0.2084
family_history	0.0037	*mental_health_consequence*	-0.1531
treatment	-0.0361	*phys_health_consequence*	0.1038
work_interfere	-0.0927	*coworkers*	0.5743
no_employees	-0.0527	*supervisor*	1.0000
remote_work	0.0252	*mental_health_interview*	-0.1895
tech_company	0.0495	*phys_health_interview*	0.0828
benefits	0.0396	*mental_vs_physical*	0.3117
care_options	0.0702	*obs_consequence*	-0.0905

Finalized Hypothesis (Based on EDA)

- Target predictors: Would you be willing to discuss a mental health issue with your direct supervisor(s) {e.g., **supervisor** in the dataset}?
  - ***Obs_consequence***: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
  - ***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?
  - ***Benefits:*** Does your employer provide mental health benefits?
  - ***Leave:*** How easy is it for you to take medical leave for a mental health condition?
- **Null Hypothesis:** Using Random Forest and feature importance from sklearn (i.e., mean decrease in Gini Index), the 5 target predictors, in no particular order, will not be the most important when predicting whether employees are willing to discuss mental health issues with supervisors.
- **Alternative Hypothesis:** Using Random Forest the gini index and feature importance (i.e., mean decrease in Gini Index) from sklearn, the 5 target predictors, in no particular no order, will be the most important when predicting whether employees are willing to discuss mental health issues with supervisors

Model Plan

Model:

- Sklearn Library
- Random Forest Model
    - Bootstrapping to create multiple different models from the same dataset

Justification:

- Random forest does not require thorough hyper-parameter tuning to produce an accurate result. Using the random forest algorithm makes it very easy to measure the relative importance of each feature on the prediction, which is what our hypothesis is about.
- Sklearn allows us to directly view feature importance by looking at how much the tree nodes that use that feature reduce impurity across all trees in the forest. This is computed automatically for each feature after training, and the results are scaled so the sum of all importance is equal to one
- We will use Bootstrapping in the RandomForestClassifier from sklearn.ensemble to improve predictive accuracy and prevent overfitting (e.g., reduce variance of the overall classifier)
- RandomForestClassifier will allow us to call permutation feature importance (as well as related methods such as feature importance and tree feature importance) so that the predictive ability of the target categories (i.e., leave, no_employees, obs_consequence, remote_work, & benefits) is more easily able to be determined

Model Plan (Continued)

Optimization:

- A parameter search algorithm (GridSearchCV) will be used to tune parameters related to tree size, maximum and minimum node size, and number estimators

Training and Testing:

- Data will be randomly split 80% towards training and 20% towards testing

Metrics:

- To test our hypothesis, we will call the permutation feature importance function and measure if the normalized importances for our target predictors sum to greater than or less than 50%
    - Defined to be the decrease in a model score when a single feature value is randomly shuffled Procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature
- We can also compare permutations feature importance with other feature importance metrics if the answer to our hypothesis is unclear

Day 3: Feature Engineering, & Results of Initial Model Performance

Preprocessing

- Dropped **timestamp**, **state**, **coworkers**, **mental_health_interview**, **phys_health_interview** and **comments**
  a. We did not think **timestamp** that the survey was submitted was relevant to determining our question
  b. We dropped **state** because we are looking at global data, and have **country** already
  c. We dropped **coworkers** because it has the exact same question as supervisor, our label, but for coworkers
     i. We thought that including this category would introduce overfitting to our model (0.57)
  d. We dropped **mental_health_interview** because we thought it was too similar to our label
     i. we thought including this category would introduce overfitting
  e. We dropped **physical_health_interview** because we thought it was too similar to our label
     i. we thought including this category would introduce overfitting
  f. We dropped **comments** because the it contains many null values, and the string responses are all very different, so it would be very hard to format this data to actually benefit our model

Initial Model Before Feature Engineering

```
#Instantiate and fit Entropy on Full Dataset
RF_entropy = RandomForestClassifier(n_estimators = 100,
                                    criterion = 'entropy',
                                    max_depth = None,
                                    min_samples_leaf = 1,
                                    bootstrap = False,
                                    warm_start = False,
                                    random_state = 508)
RF_entropy_fit = RF_entropy.fit(X_train_full, y_train_full)
```

```
rfc_cv_score = cross_val_score(RF_entropy, X, y, cv=10, scoring='roc_auc')
print(rfc_cv_score.mean())

rfc_predict = RF_entropy.predict(X_test_full)
print(classification_report(y_test_full, rfc_predict))
```

```
0.7897531456796163
```

F1 score:

- Micro avg: 0.84
- Macro avg: 0.03
- Weighted avg: 0.76
- Samples avg: 0.83

Training score: 1.0

Our initial model before feature engineering completely overfit and slightly outperformed our model with feature engineering, which is likely because the features that are very similar to our label were included in this model

Feature Engineering

- We had to clean the **gender**, **wellness_program**, **anonymity**, **seek_help**, **leave**, **mental_vs_physical**, **benefits**, **care_options**, **work_interfere**, **self_employed**, and **age** categories
  a. Many of these neede
- We bucketed **age** as a factor for plotting, thereby additionally making it a discrete variable
  a. We also imputed **age** outliers with their mean (32)
  b. We ended up using age as a continuous variable instead of a bucketed variable (see Model Weaknesses)
- We used **get_dummies** (pandas library) to turn all of our categorical features into dummy variables
  a. One-hot Encoding
- We put separated **country** into **North America**, **Europe** and **other**
  a. Many of the options for country had only one survey respondent, and hence was not useful for the model in its original format

String Processing

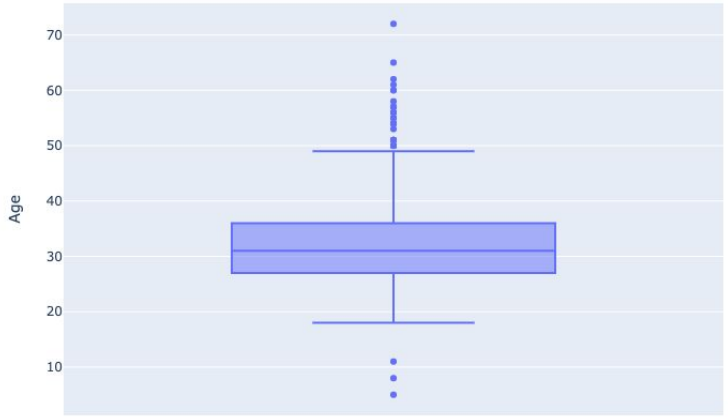
```
df.Gender.value_counts()
```

```
Male                                                615
male                                                206
Female                                              121
M                                                   116
female                                               62
F                                                    38
m                                                    34
f                                                    15
Make                                                  4
Male                                                  3
Woman                                                 3
Female (trans)                                        2
Man                                                   2
Female                                                2
Cis Male                                              2
Trans-female                                          1
Mail                                                  1
fluid                                                 1
queer                                                 1
Enby                                                  1
cis male                                              1
maile                                                 1
Genderqueer                                           1
msle                                                  1
Femake                                                1
queer/she/they                                        1
Trans woman                                           1
male leaning androgynous                              1
Agender                                               1
ostensibly male, unsure what that really means        1
All                                                   1
Nah                                                   1
Female (cis)                                          1
Malr                                                  1
non-binary                                            1
Androgyne                                             1
Cis Female                                            1
Cis Man                                               1
cis-female/femme                                      1
Neuter                                                1
woman                                                 1
Mal                                                   1
femail                                                1
```

- The original format of the **gender** column was unprocessed, so we cleaned **gender** into male, female, and other to retain the possibility of using all the data
- Mapped misspelled string columns back to standardized answers (ie: Yes, No)

```
femail                                                1
Male (CIS)                                            1
p                                                     1
A little about you                                    1
Male-ish                                              1
something kinda male?                                 1
Guy (-ish) ^_^                                        1
Name: Gender, dtype: int64
```

Age Processing



Imputed clear outliers, for instance had a value of nearly 100 billion and other data points were negative, to the mean (32). Bar graph on the left shows the age before data was cleaned and on the right is the data post cleaning.

Initial Model: Random Forest Using Gini/OHE

```
#Instantiate and fit Gini
RF_gini = RandomForestClassifier(n_estimators = 100,
                                 criterion = 'gini',
                                 max_depth = None,
                                 min_samples_leaf = 1,
                                 bootstrap = False,
                                 warm_start = False,
                                 random_state = 508)

#Fitting model
RF_gini_fit = RF_gini.fit(X_train, y_train)

#Printing model scores
print('Training Score', RF_gini_fit.score(X_train, y_train).round(7))
print('Testing Score:', RF_gini_fit.score(X_test, y_test).round(7))
```

```
Training Score 0.9970209
Testing Score: 0.7460317
```

- Tested both gini and entropy criterions with the same parameters
- Our testing score is already 0.746 on our original model before hyperparameter tuning

Initial Model: Random Forest Using Entropy/OHE

```
#Instantiate and fit Entropy
RF_entropy = RandomForestClassifier(n_estimators = 100,
                                    criterion = 'entropy',
                                    max_depth = None,
                                    min_samples_leaf = 1,
                                    bootstrap = False,
                                    warm_start = False,
                                    random_state = 508)
RF_entropy_fit = RF_entropy.fit(X_train, y_train)

#Printing model scores
print('Training Score', RF_entropy_fit.score(X_train, y_train).round(7))
print('Testing Score:', RF_entropy_fit.score(X_test, y_test).round(7))
```

```
Training Score 0.9970209
Testing Score: 0.7579365
```

- This Entropy testing score is slightly better (0.03) than the previous one before hyperparameter tuning

Initial Model: Random Forest Using Entropy/LE

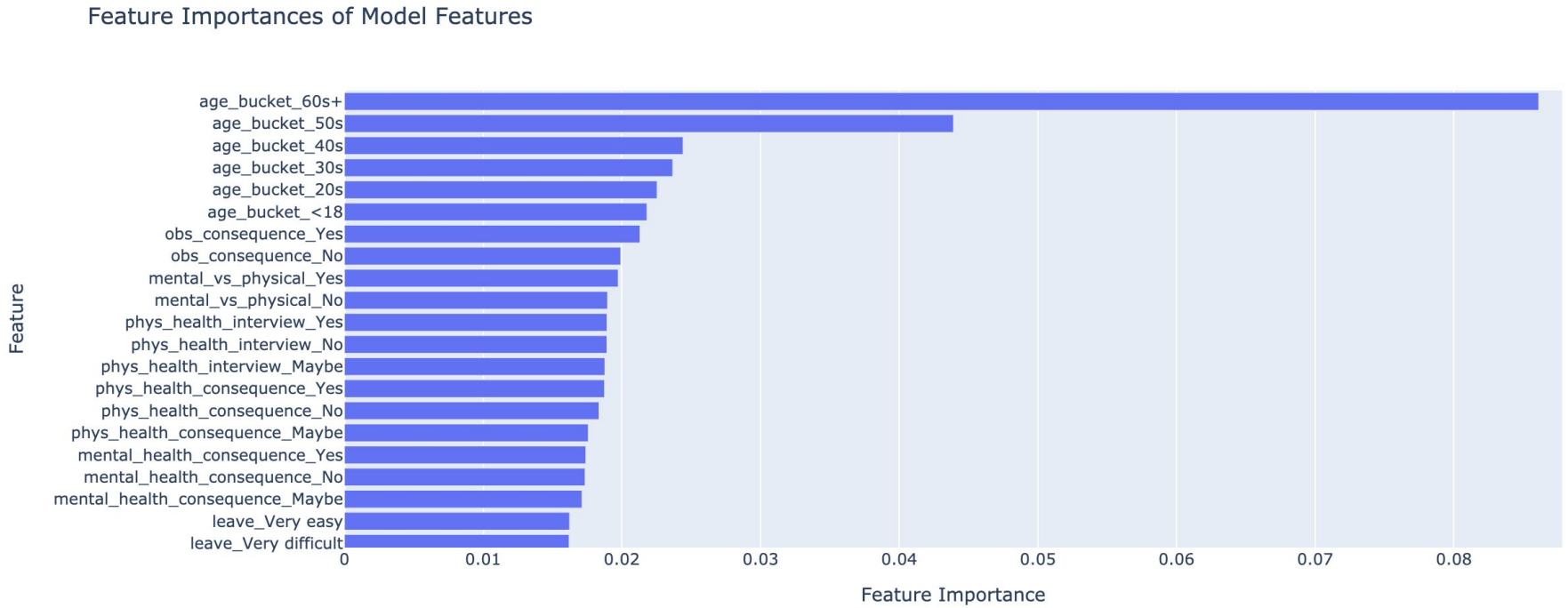
```
#Instantiate and fit Entropy
RF_entropy = RandomForestClassifier(n_estimators = 100,
                                    criterion = 'entropy',
                                    max_depth = None,
                                    min_samples_leaf = 1,
                                    bootstrap = False,
                                    warm_start = False,
                                    random_state = 508)
RF_entropy_fit = RF_entropy.fit(X_train, y_train)

#Printing model scores
print('Training Score', RF_entropy_fit.score(X_train, y_train).round(7))
print('Testing Score:', RF_entropy_fit.score(X_test, y_test).round(7))
```

```
Training Score 0.9970209
Testing Score: 0.5833333
```

- Also testing effect of encoding on the performance
- The label encoded testing score is lower than the label encoded and also displays signs of overfitting

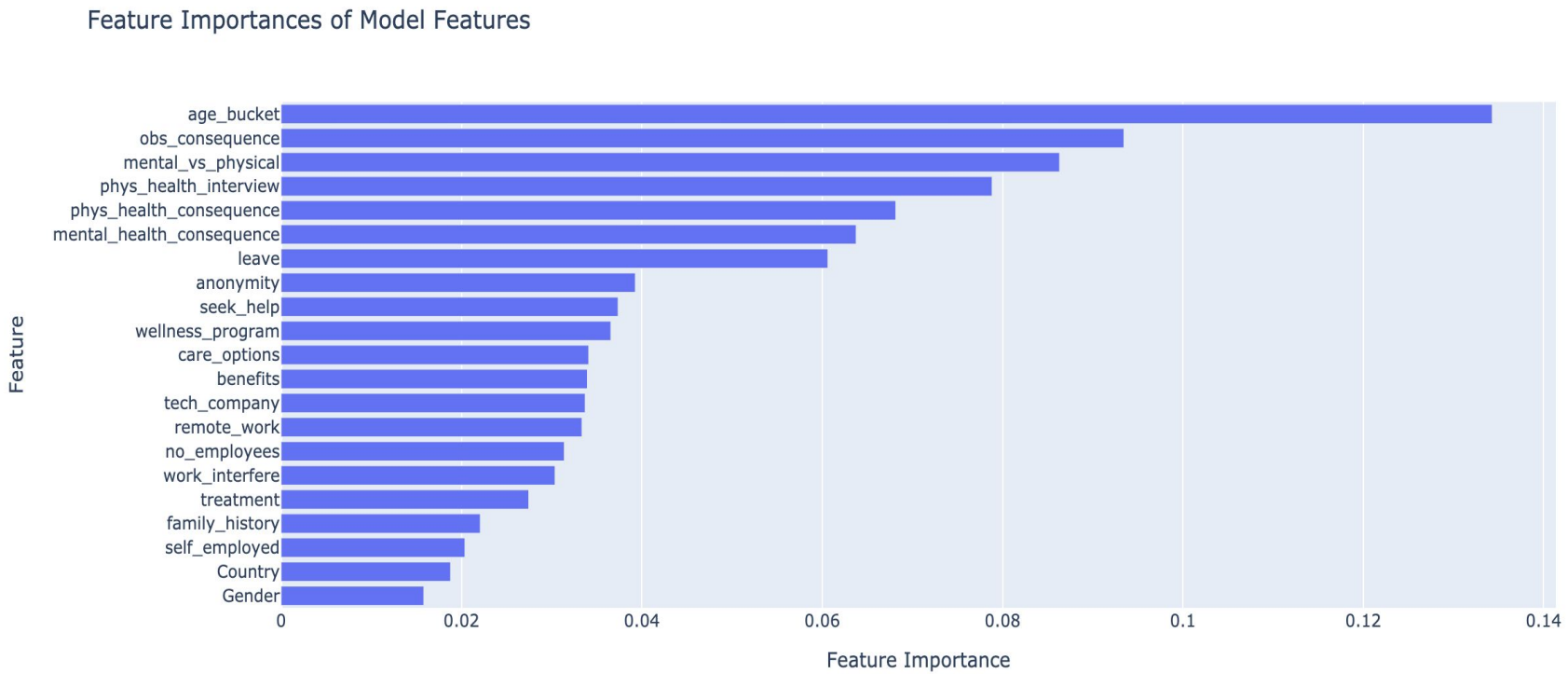
Feature Importance with One-Hot Encoding

Random Forest using Entropy



Feature Importance with Label Encoding

Random Forest using Entropy



Initial Model Response

- Average **cross validation score** was 0.79
- **age** played a bigger role than we expected, but we think this is because we ended up creating **age** buckets (discrete) instead of using exact ages (continuous) as we had originally planned when writing our hypothesis
- **observed_consequence** is the 2nd most important feature of our initial model
- **leave** is 7th most important feature of our initial model
- **benefits** is the 12th most important feature of our initial model
- **remote_work** is the 14th most important feature of our initial model
- **no_employees** is the 15th most important feature of our initial model
- Thus, we cannot reject our null hypothesis yet!

Next Steps

- We plan on running our test to compare models using age as a numerical feature instead of the creating age buckets
- We are going to hypertune our model using a GridSearch in order to get a higher test score and optimal parameters
- We have overfitting in our training data, so we need to work on decreasing that
  - Will implement more extensive cross-validation
  - Bootstrapping

Day 4: Finalized Model Results, Discussion, & Conclusion

Preliminary Results

	LE pre feature engineering	OHE pre feature engineering	LE feature engineered	OHE feature engineered	OHE feature engineered w/o age buckets
Accuracy	0.60	0.77	0.57	0.74	0.75
F1	0.59	0.77	0.54	0.74	0.75
Recall	0.60	0.77	0.57	0.74	0.75
Precision	0.59	0.77	0.54	0.74	0.75
Training	1.0000	1.0000	0.9921	0.9940	1.0
Testing	0.6032	0.7698	0.5714	0.7421	0.7854

Hyperparameter Tuning Approach

1) Optimizing **n_estimators** and **min_samples_leaf** using GridSearchCV (cv=5)
   a) These parameters were tuned first because they are most integral to creating an accurate model
      i) **n_estimators** → Number of trees utilized in the model
         (1) Values: [1, 20, 50, 75, 100, 200, 300]
      ii) **min_samples_leaf** → The minimum number of samples required to be at a leaf node
         (1) Values: [1, 5, 10, 50, 100]
   b) Baseline Model Accuracy: **0.7637**
2) **max_features** → Number of features in dataset to consider when looking for the best split
   a) This tuning step did not increase the accuracy of our model, so we ended up leaving this at the default value (None) in our final model in order to avoid unnecessary complexity
3) Tuning **max_depth** and **max_samples**
   a) Tuning these parameters increased the accuracy of our model
      i) **max_depth** → The maximum depth of the tree
         (1) Values: [5, 10, 15, 20, 50, 100]
      ii) **max_samples** → What fraction of the original dataset is given to create the tree (bootstrapping)
         (1) Values: [0.2, 0.4, 0.6, 0.8]
   b) Improved Model Accuracy: **0.7676**

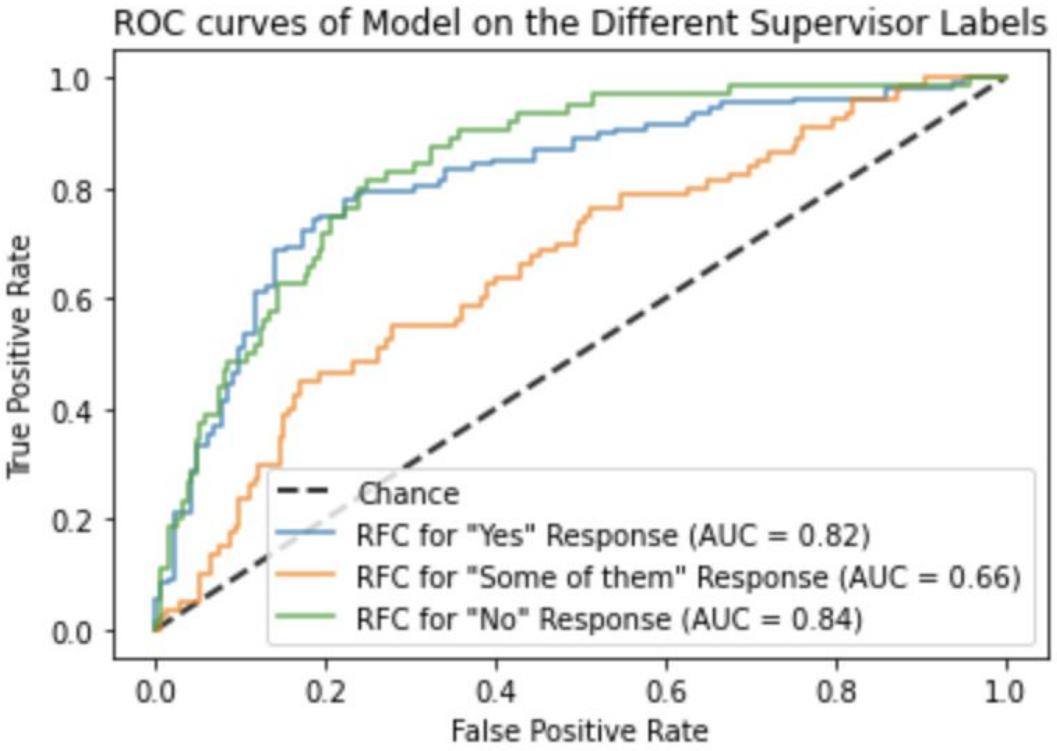
Hyperparameter tuning: Initial vs Final Model

	Initial Model	**Final Model**	**Implication**
criterion	Entropy/Gini	Gini	Gini Impurity measure better
max_depth	None	10	Limit on length of nodes beneficial
min_samples_leaf	1	10	Multiple samples inform every decision in tree
n_estimators	100	75	Limit to # trees to improve performance
max_samples	None	0.6	Higher # samples to train each estimator
bootstrap	True	True	Random sampling with replacement beneficial to performance

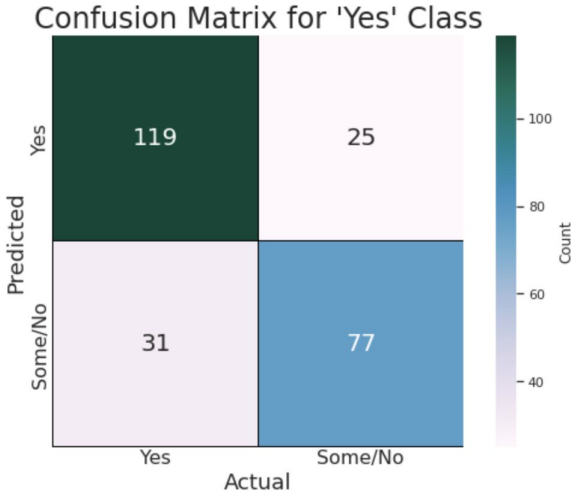
Final Model: Comparing the Three Response Variable Classes

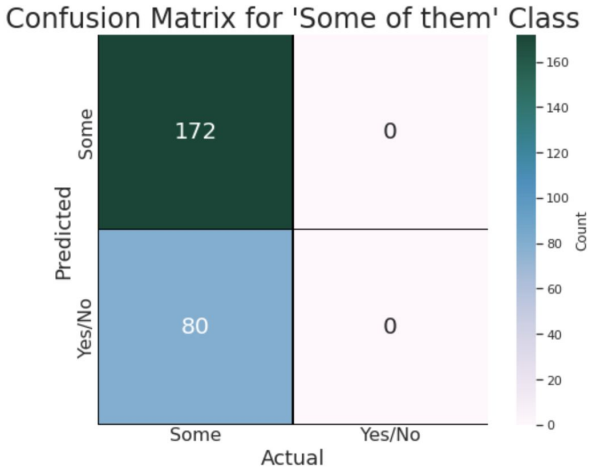
	Yes	**Some of Them**	**No**
Testing Accuracy	0.78	0.68	0.79
F1 Score	0.81	0.81	0.87
Recall	0.79	0.68	0.82
Precision	0.83	**1.0**	0.93

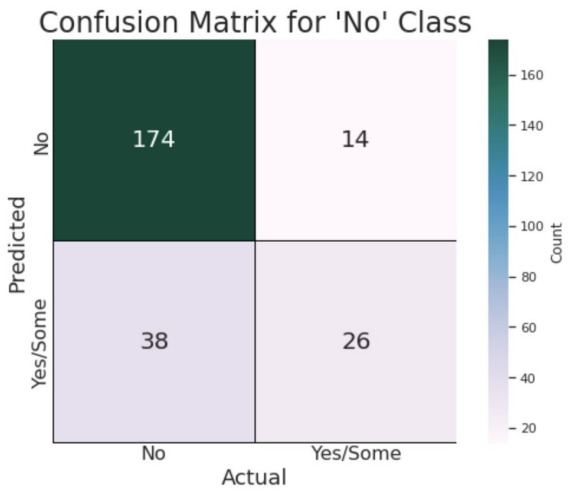
ROC Curves


ROC curves of Model on the Different Supervisor Labels
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Chance
RFC for "Yes" Response (AUC = 0.82)
RFC for "Some of them" Response (AUC = 0.66)
RFC for "No" Response (AUC = 0.84)

Confusion Matrices & Summary Statistics

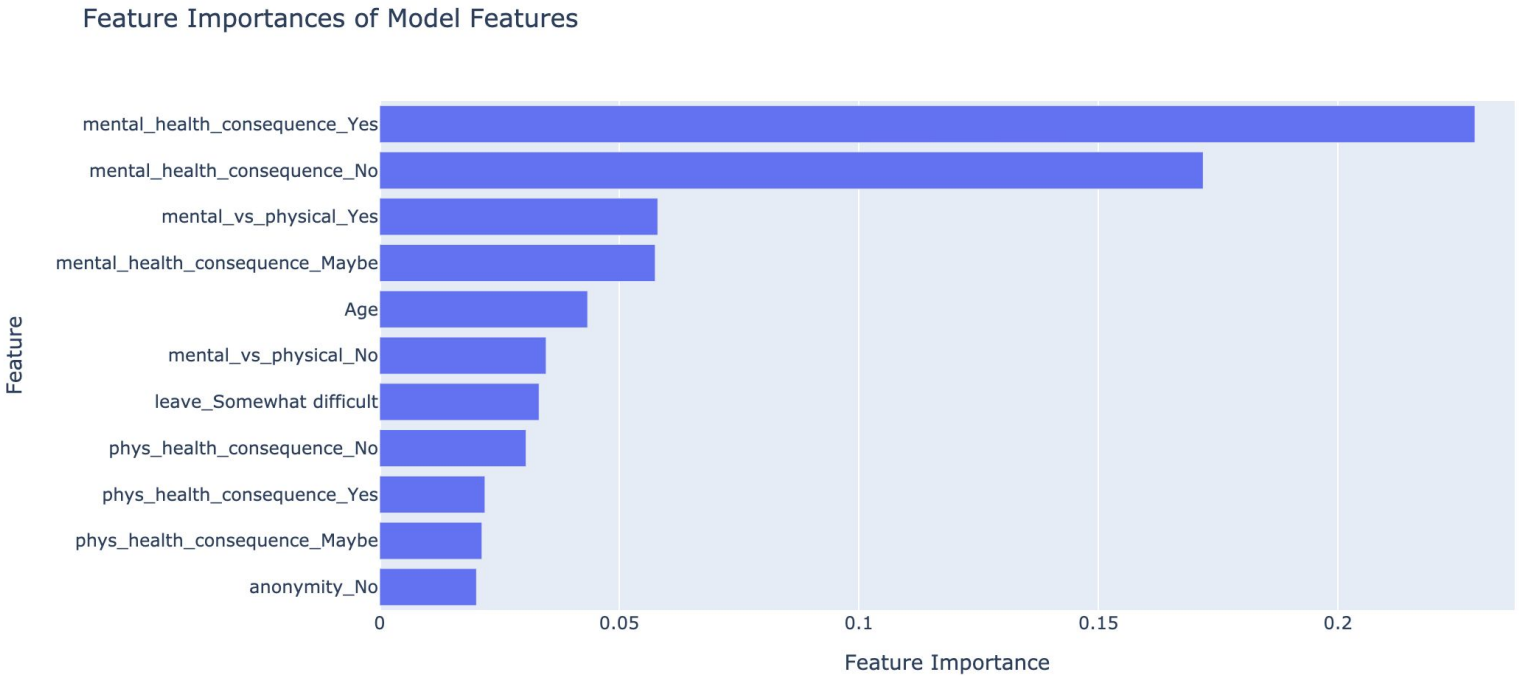






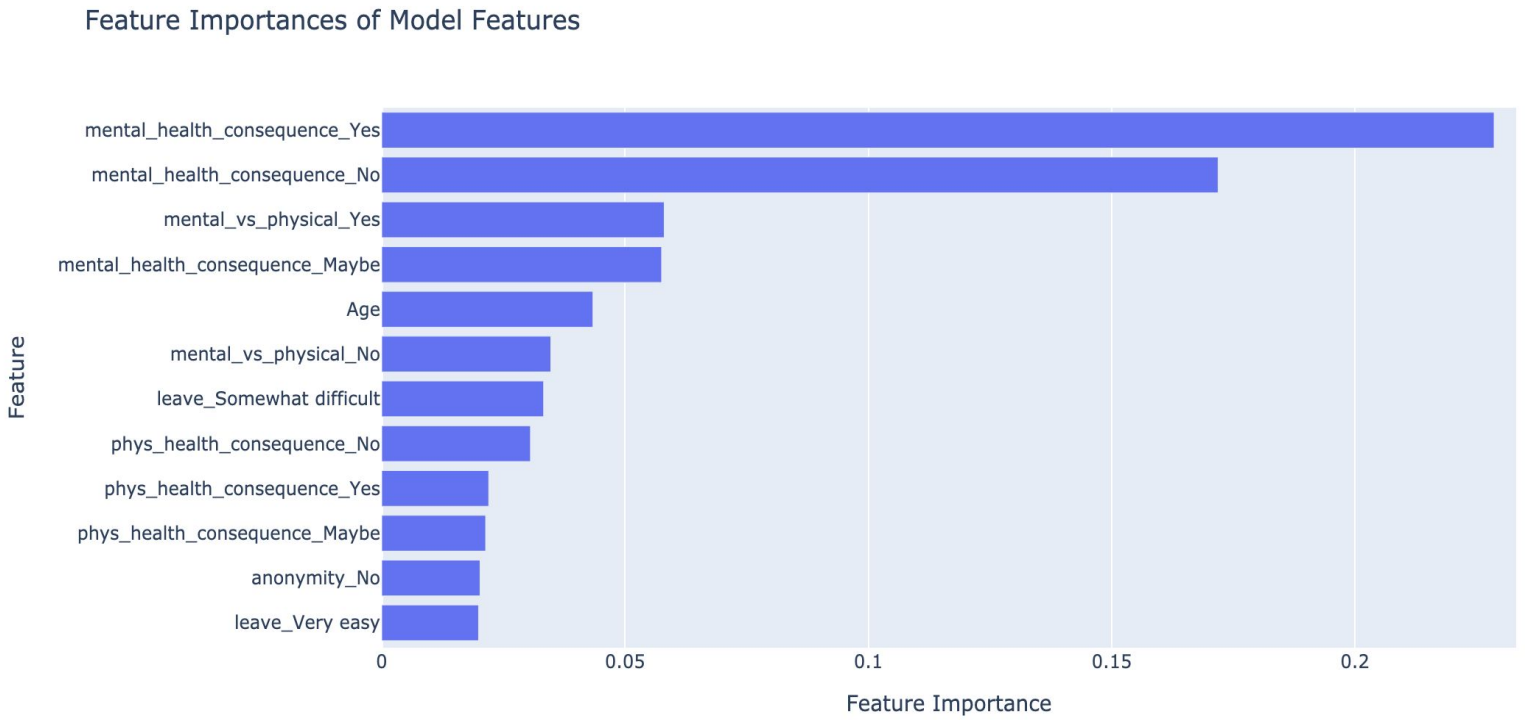
Feature Importance for Yes

Random Forest using Gini



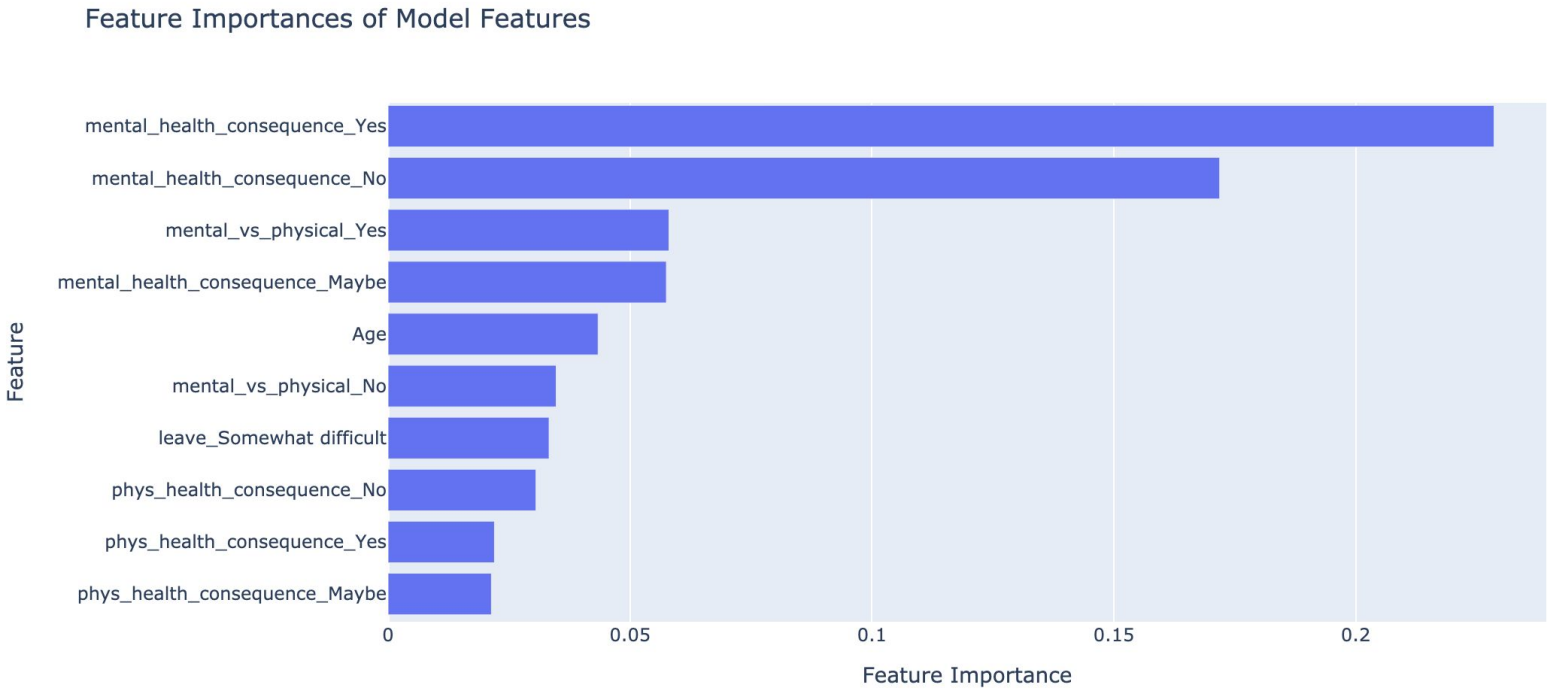
Feature Importance for ‘Some of Them’

Random Forest using Gini



Feature Importance for ‘No’

Random Forest using Gini



Feature Importance Magnitude

Top 5 Greatest Importances

Original_Column	Feature_Importance
mental_health_consequence	0.457977
mental_vs_physical	0.092797
leave	0.080582
Age	0.043406
anonymity	0.028990

Top 5 Smallest Importances

	Original_Column	Feature_Importance
4	care_options	0.024788
6	family_history	0.016607
5	country_bucket	0.015575
1	Gender	0.012911
3	benefits	0.011813

Interpreting the Results

- We cannot reject the null hypothesis based on mean decrease in gini
  - Only **leave** was within the top five features ranked by importance
  - The top five features we found with our best model:
    - **mental_health_consequence**
    - **mental_vs_physical**
    - **leave**
    - **age**
    - **anonymity**
  - **benefits** was actually found to be the least most important feature
  - **no_employees**, **obs_consequence**, and **remote_work** were not near the top five strongest predictors

Interpreting the Results (Continued)

- **Mental_vs_physical**: Do you feel that your employer takes mental health as seriously as physical health?
  - Intuitively it is clear to see why this would be the one of the top predictors of feeling comfortable with talking to your supervisor about mental health issues
    - If you feel that your supervisor takes mental health as serious as physical health, you are more likely to feel comfortable reaching out to them about mental health
- **Mental_health_consequence:** Do you think that discussing a mental health issue with your employer would have negative consequences?
  - **Mental_health_consequences** inherently tracks the substantial stigma associated with having a mental illness in the world today, so it is not surprising that this would be a top predictor as well
    - Ultimately, we thought the **obs_consequence** feature would be sufficient to track this, but **mental_health_consequence** may be better because it is what an employee thinks will happen to themselves
    - Not what they observed to happen to someone else (**obs_consequence**)

Interpreting the Results (Continued)

- **anonymity**: Is anonymity protected if an employee takes advantage of mental health or substance abuse treatment?
  - Also an intuitive explanation, as if an identity is protected, an employee will more likely to be comfortable reaching out to others, including a supervisor, with mental health concerns
- **age** {continuous}
  - This is surprising, but there are two explanations for this:
    - There was a large unimodal distribution of age highly in favor of employees in their 20s and 30s
    - Furthermore, stigma may be more debilitating to older members of the workforce who are accustomed to mental health not being discussed at all
- **leave :** How easy is it for you to take medical leave for a mental health condition?
  - This was the only one present in our hypothesized target set!

Interpreting the Results (Continued)

Response Variable Class	Base Rate	Model Accuracy
Yes	0.4100 = 41.00%	**0.7800 = 78.00%**
Some of Them	0.2780 = 27.80%	**0.6800 = 68.00%**
No	0.3120 = 31.20%	**0.7900 = 79.00%**

Model Limitations

- Generalizability
  - Low samples sizes from “Other” Countries
  - Healthcare structures not identified, U.S Healthcare highly variable
  - Gender/cultural differences per country
  - Small dataset (1,259 survey responses across several different countries)
    - Far from representative of the tech workforce population (strongest predictors unlikely to be the same if conducting analysis with a more representative dataset)
- Continuous variables have greater statistical power than categorical
  - Continuous variables have higher sensitivity while categorical ones have low sensitivity
    - Categorical variables also generally increase variance without regard to sample size
  - Binning country made model more interpretable but was not all inclusive
- Class imbalance
  - Model tends to perform better for the majority
    - Respondents: U.S
    - Target variable: Yes (41%)
  - Feature engineering and hypertuning was only performed through analysis on ‘Yes’ class

Future Avenues

- Explore options to improve class imbalance
  - Downsample ‘Yes’ (41%) compared to ‘Some’ (27%) or ‘No’ (32%)
  - Relevel the class to make ‘some’ equal to either ‘yes’ or ‘no’, or try both and observe differences
- Explore more specific health policies in the workplace & their effect on predictability
  - Maternity/paternity leave
  - Subsidized child care
- Explore more specific health structures & their effect on the response variable
  - Privatized insurance, universal healthcare, public option, etc.
- Include a time series analysis
  - OSMI, organization that conducted the survey, have data from similar surveys taken between 2016 and 2020 (can observe how predictors change over time; can even make specific to COVID-19 remote work)
- Expand dataset features and make them more specific
  - Specific types of mental health + coverage
- Perform feature engineering & hypertuning with respect to each of the label’s classes
- Explore different machine learning models through utilizing multi-class classification

Conclusion

- Dataset: 2014 OSMI survey measuring attitudes towards mental health and frequency of mental health disorders in the tech workplace.
- Null Hypothesis:
  - Using Random Forest and feature importance from scikit-learn (i.e., mean decrease in Gini Index), the 5 target predictors, in no particular order, will not be the most important when predicting whether employees are willing to discuss mental health issues with supervisors.
  - **obs_consequence**, **benefits**, **no_employees**, **leave**, & **remote_work**
- EDA:
  - Analyzed feature distributions with respect to response variable & removed highly correlated features
- Feature Engineering:
  - Tradeoff between categorical and continuous variables, bias vs. variance, & hypertuning model
- Results:
  - Only one out of five of our target predictors was in the top five most important features (measured by mean decrease in Gini Index) for our model
    - **leave**
- Conclusion:
  - Although model resulted in substantially higher accuracies than base rates with respect to response variable class, we failed to reject the Null Hypothesis

References

1. Depression. Accessed January 6, 2021. https://www.who.int/en/news-room/fact-sheets/detail/depression
2. Disability in the Workplace: A Unique and Variable Identity - Alecia M. Santuzzi, Pamela R. Waltz, 2016. Accessed January 6, 2021. https://journals.sagepub.com/doi/full/10.1177/0149206315626269
3. Mental Health in the Workplace. Published April 26, 2019. Accessed January 6, 2021. https://www.cdc.gov/workplacehealthpromotion/tools-resources/workplace-health/mental-health/index.html
4. COVID-19's Impact on Mental Health and Workplace Well-being. NIHCM. Accessed January 6, 2021. https://nihcm.org/publications/covid-19s-impact-on-mental-health-and-workplace-well-being
5. Mental Health in Tech Survey. Accessed January 6, 2021. https://kaggle.com/osmi/mental-health-in-tech-survey
6. Sado M, Shirahase J, Yoshimura K, et al. Predictors of repeated sick leave in the workplace because of mental disorders. *Neuropsychiatr Dis Treat*. 2014;10:193-200. doi:10.2147/NDT.S55490

Predictors of Mental Health Climate in the Tech Industry

Authors: Tony, Navya, Kunaal, and Jaya (Group 4)
Date: January 8th, 2021

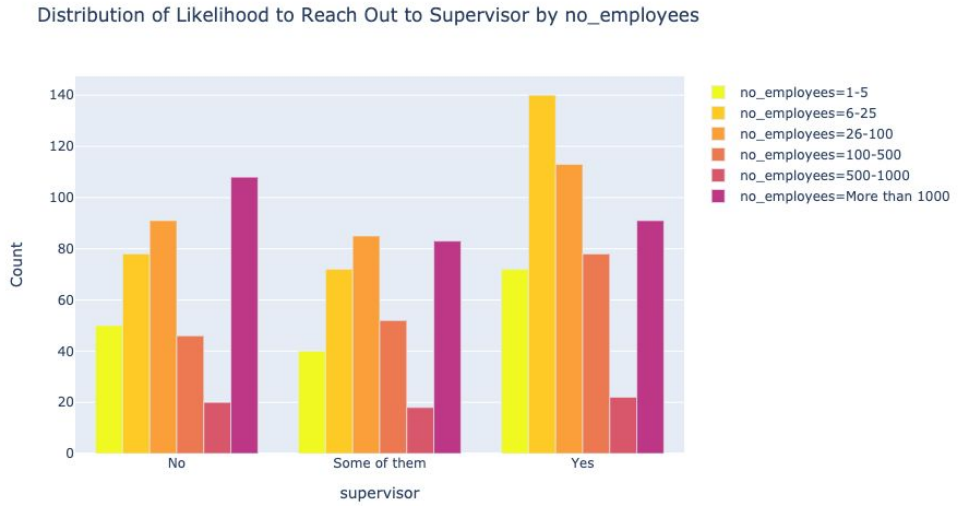
Background & Motivation

- In 2017, WHO estimated that 1 in 17 adults experience a serious mental illness each year[1]
  - More than 44 million adults are affected annually by mental illnesses, many of whom are also active within the workforce[2]
- Only 57% of employees who report moderate depression and 40% of those who report severe depression receive treatment to control symptoms[3]
- Due to COVID, mental health is increasingly affecting work life
  - 55% of employees feel uncomfortable confiding in anyone at work[4]
  - Remote work can either be an alleviator or exacerbator of a mental illness
- 2014 survey Open Sourcing Mental Illness (OSMI) survey on attitudes towards mental health specifically in the technology field

Initial target predictors' relationship to target: no_employees

Outcome: People who worked in smaller companies more commonly responded that they would reach out to a supervisor compared those in larger companies.

Conclusion: As companies get larger, employees may have less interaction with their supervisor and feel uncomfortable reaching out. Due to the expected outcome, we will keep the **no_employees** variable as a predictor in our hypothesis.

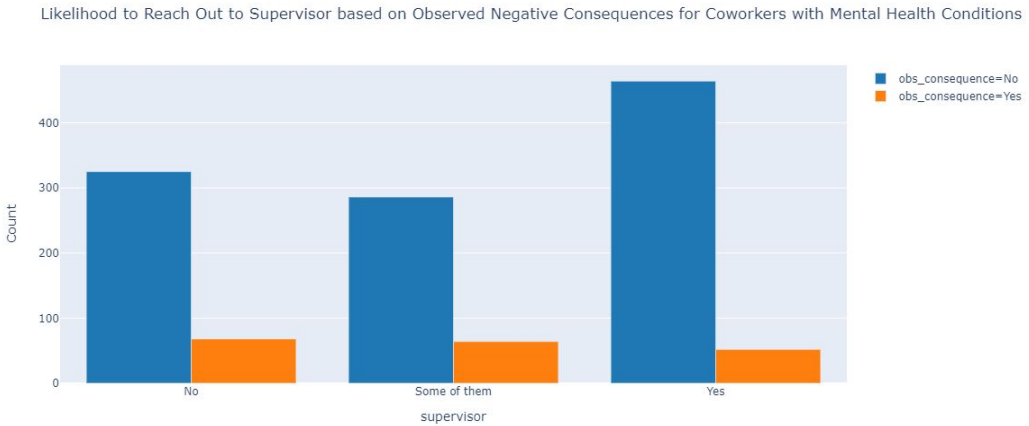


Feature	chi2	p-value
no_employees	3.7966	0.1498

Initial target predictors' relationship to target: consequences

Outcome: Employees who had not observed consequences were more likely to feel comfortable. Those who had seen negative consequences, were more likely to not reach out to a supervisor.

Conclusion: We expected that those who observed consequences would be uncomfortable bringing up their own issues. Thus, **obs_consequences** was included in the model

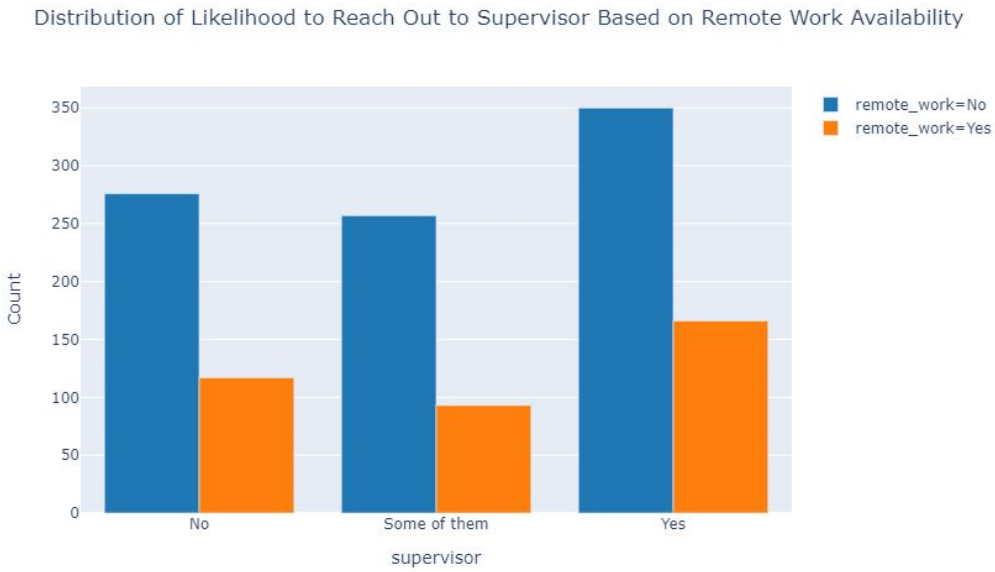


Feature	chi2	p-value
obs_consequence	12.4387	0.0020

Initial target predictors' relationship to target: remote_work

Outcome: Qualitatively, employees who did not work remotely were more likely to reach out to their supervisor(s). However, those who worked remotely were also more likely to reach out to their supervisor.

Conclusion: The first part of the outcome matched our expectations. Although the second portion was surprising, we decided to leave **remote_work** in our target predictor set based on the literature we reviewed.

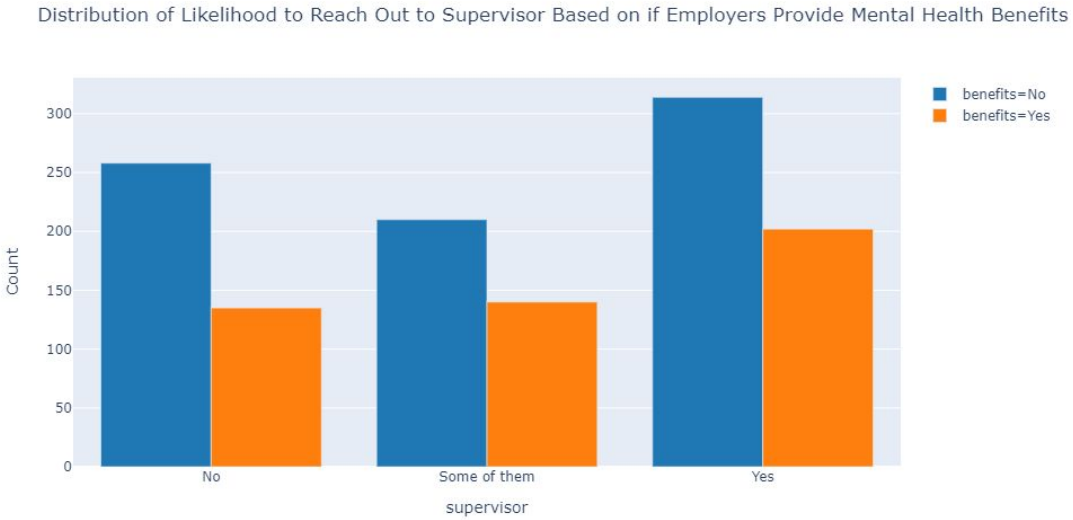


Feature	chi2	p-value
remote_work	2.1908	0.3344

Initial target predictors' relationship to target: benefits

Outcome: As seen qualitatively, employees who have mental health benefits were more likely to reach out. However, there was no distinct trend for employees who did not receive mental health benefits.

Conclusion: The first portion of the outcome listed above matches our expectations. Considering that mental health benefits was also heavily cited in the literature, we will keep **benefits** in our target predictor set.



Feature	chi2	p-value
benefits	1.9256	0.3818

Initial Hypothesis

- Target predictors: Would you be willing to discuss a mental health issue with your direct supervisor(s) {e.g., **supervisor** in the dataset}?
  - ***Obs_consequence***: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
  - ***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?
  - ***Benefits:*** Does your employer provide mental health benefits?

Null Hypothesis	The 4 target predictors do not constitute the majority (50%) of Random Forest feature importance when predicting whether employees are willing to discuss mental health issues with supervisors
Alternative Hypothesis	The 4 target predictors constitute the majority (50%) of Random Forest feature importance when predicting whether employees are willing to discuss mental health issues with supervisors

Potential target predictors' relationship to target: leave

Outcome: As the difficulty to take medical leave for a mental health condition decreased, the likelihood of communicating with a supervisor increased. The inverse was also true.

Conclusion: Originally, we did not think leave would be a main predictor. Due to the outcome, we decided the variation in responses was high enough to add **leave** to our target predictors and remove remote_work like our feedback suggested

Feature	chi2	p-value
leave	81.8797	1.6598e-18

Finalized Hypothesis

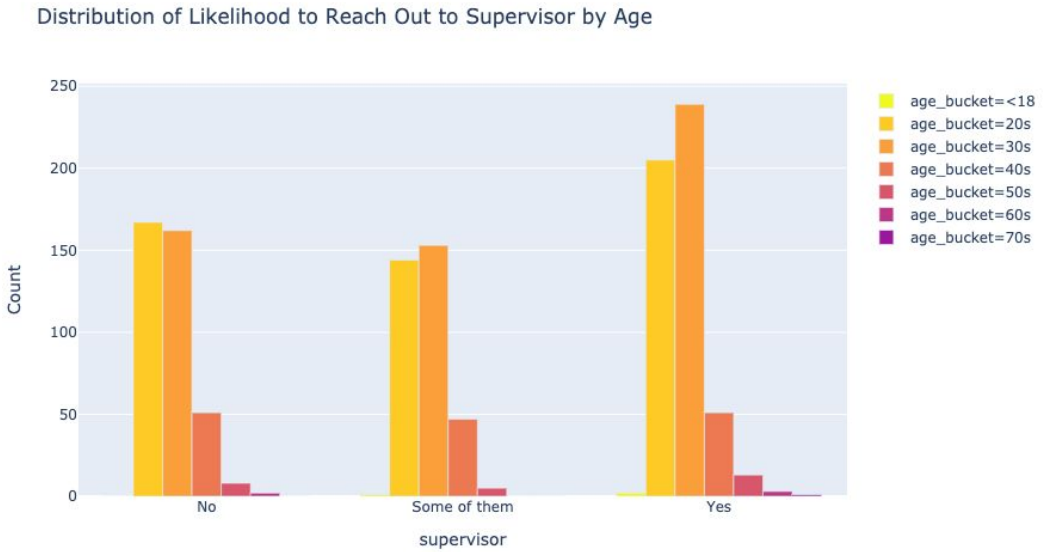
- Target predictors: Would you be willing to discuss a mental health issue with your direct supervisor(s) {e.g., **supervisor** in the dataset}?
  - ***Obs_consequence***, ***No_employees, Remote_work, Benefits, and Leave***

Null Hypothesis	Using Random Forest and mean decrease in Gini Index the 5 target predictors, not considering order, will not be the most important when predicting whether employees are willing to discuss mental health issues with supervisors
Alternative Hypothesis	Using Random Forest the gini index and mean decrease in Gini Index, the 5 target predictors, not considering order, will be most important to predict whether employees are willing to discuss mental health issues with supervisors

Potential predictors' relationship to target: age

Outcome: There was not much variation in response from the created age groups, though the younger age groups (20s and 30s) showed an increase in 'Yes' responses.

Conclusion: Due to only two groups showing noticeable variation, we decided that we would remove age bucketing



Feature	chi2	p-value
Age	0.3302	0.8478

Preliminary Results

	LE pre feature engineering	OHE pre feature engineering	LE feature engineered	OHE feature engineered	OHE feature engineered w/o age buckets
Accuracy	0.60	0.77	0.57	0.74	0.75
F1	0.59	0.77	0.54	0.74	0.75
Recall	0.60	0.77	0.57	0.74	0.75
Precision	0.59	0.77	0.54	0.74	0.75
Training	1.0000	1.0000	0.9921	0.9940	1.0
Testing	0.6032	0.7698	0.5714	0.7421	0.7854

Hyperparameter Tuning Approach

1) Optimizing **n_estimators** and **min_samples_leaf** using GridSearchCV (cv=5)
   a) These parameters were tuned first because they are most integral to creating an accurate model
      i) **n_estimators** → Number of trees utilized in the model
         (1) Values: [1, 20, 50, 75, 100, 200, 300]
      ii) **min_samples_leaf** → The minimum number of samples required to be at a leaf node
         (1) Values: [1, 5, 10, 50, 100]
   b) Baseline Model Accuracy: **0.7637**
2) **max_features** → Number of features in dataset to consider when looking for the best split
   a) This tuning step did not increase the accuracy of our model, so we ended up leaving this at the default value (None) in our final model in order to avoid unnecessary complexity
3) Tuning **max_depth** and **max_samples**
   a) Tuning these parameters increased the accuracy of our model
      i) **max_depth** → The maximum depth of the tree
         (1) Values: [5, 10, 15, 20, 50, 100]
      ii) **max_samples** → What fraction of the original dataset is given to create the tree (bootstrapping)
         (1) Values: [0.2, 0.4, 0.6, 0.8]
   b) Improved Model Accuracy: **0.7676**

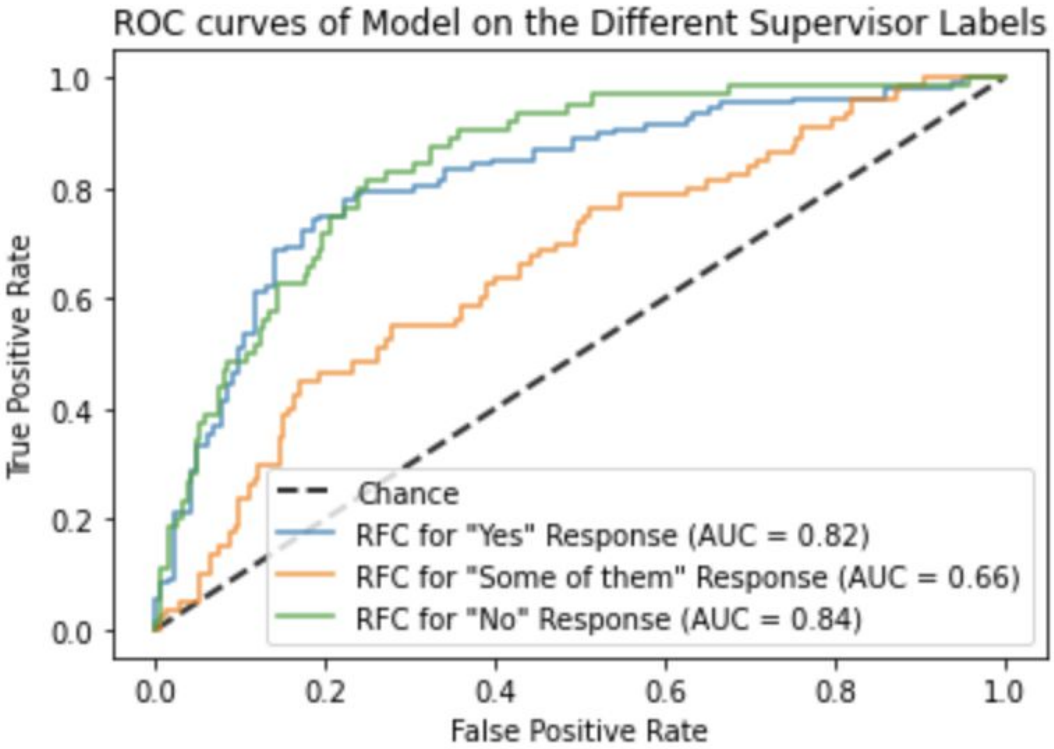
Hyperparameter tuning: Initial vs Final Model

	Initial Model	**Final Model**	**Implication**
criterion	Entropy/Gini	Gini	Gini Impurity measure better
max_depth	None	10	Limit on length of nodes beneficial
min_samples_leaf	1	10	Multiple samples inform every decision in tree
n_estimators	100	75	Limit to # trees to improve performance
max_samples	None	0.6	Higher # samples to train each estimator
bootstrap	True	True	Random sampling with replacement beneficial to performance

Final Model: Three Response Variable Classes

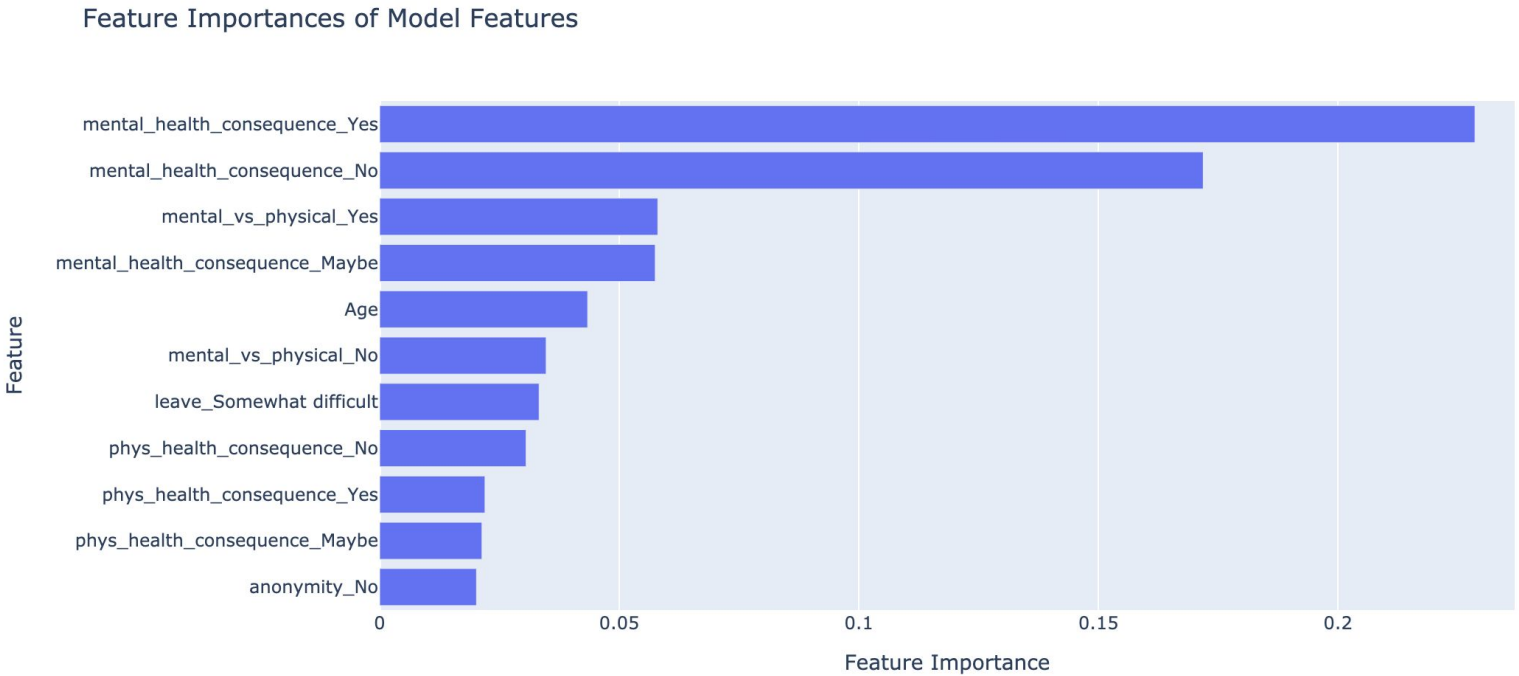
	Yes	Some of Them	No
Base Rate	41%	27%	31%
Accuracy	78%	68%	79%
F1 Score	0.81	0.81	0.87
Recall	0.79	0.68	0.82
Precision	0.83	**1.0**	0.93

ROC Curves


ROC curves of Model on the Different Supervisor Labels
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Chance
RFC for "Yes" Response (AUC = 0.82)
RFC for "Some of them" Response (AUC = 0.66)
RFC for "No" Response (AUC = 0.84)

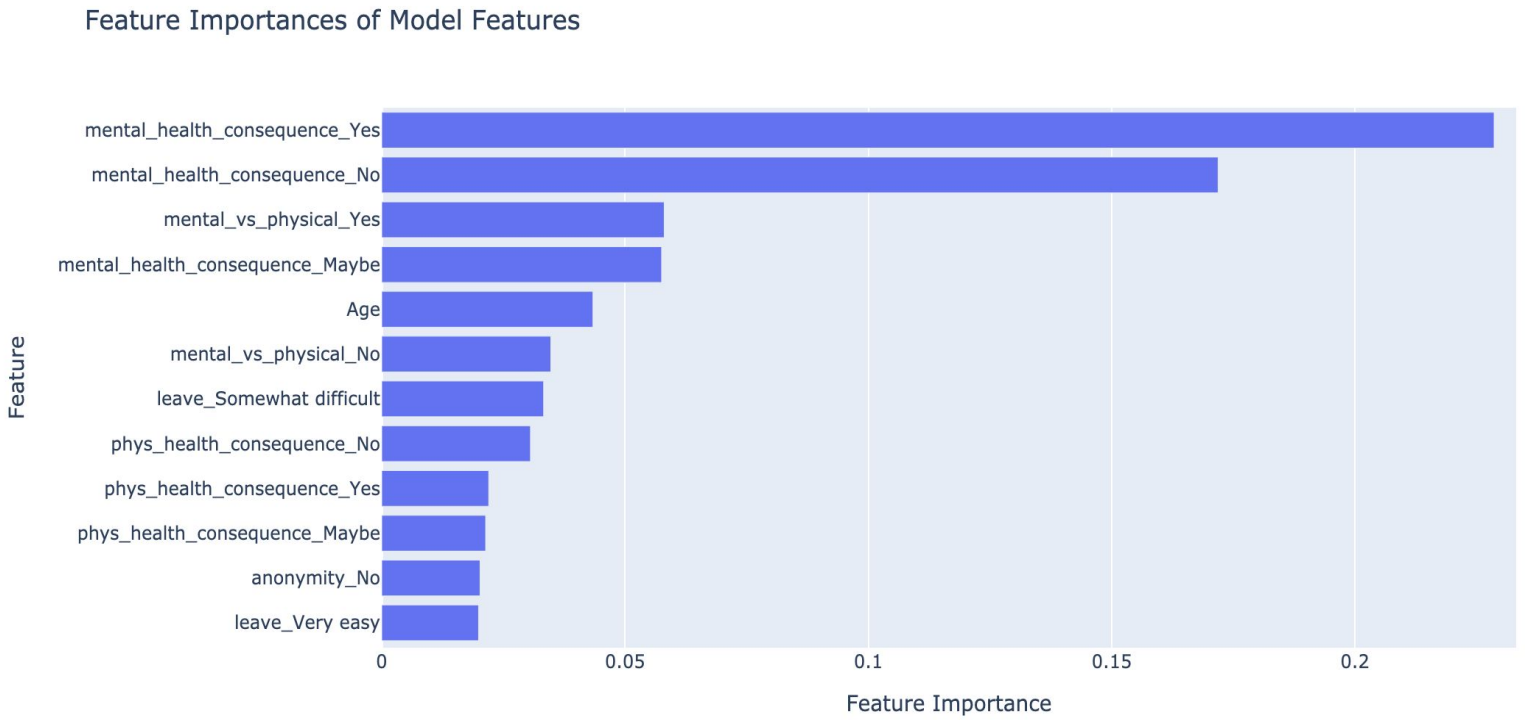
Feature Importance for Yes

Random Forest using Gini



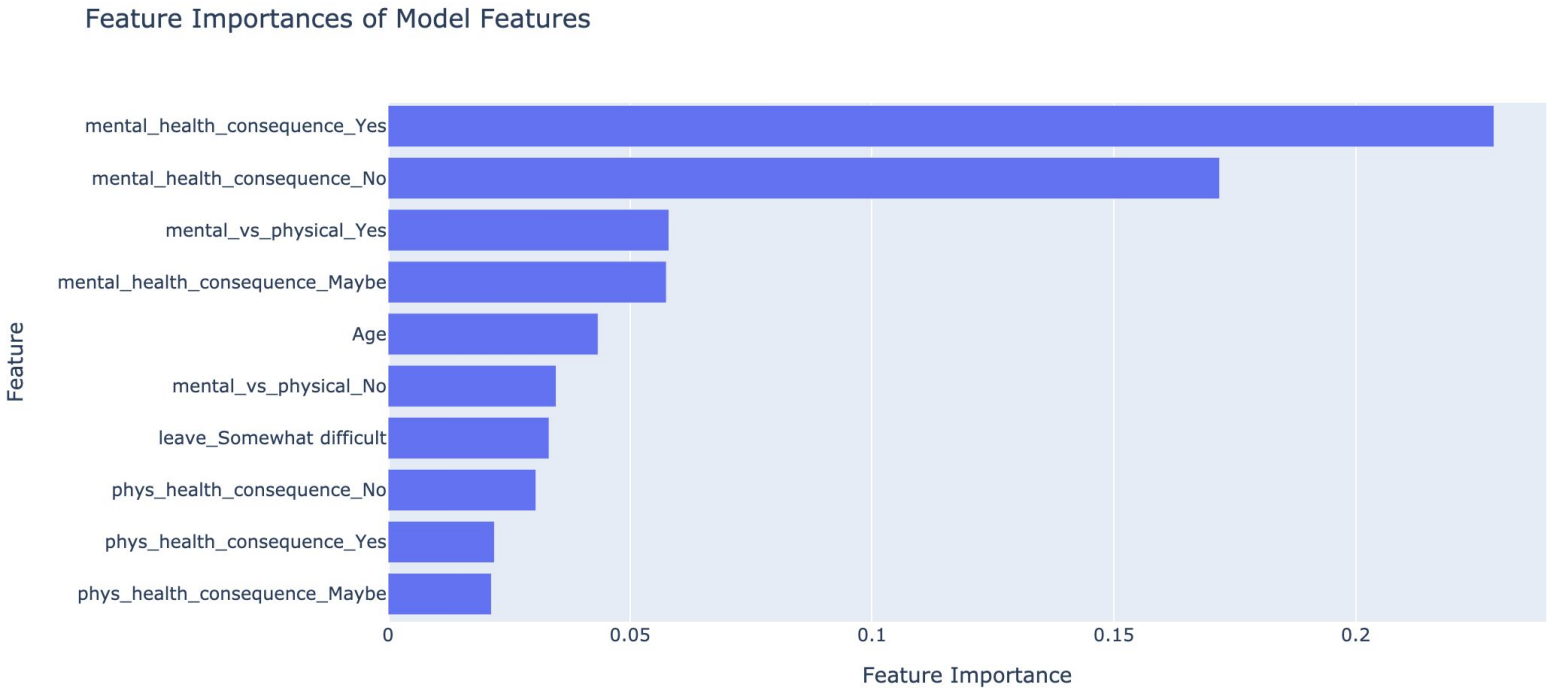
Feature Importance for ‘Some of Them’

Random Forest using Gini

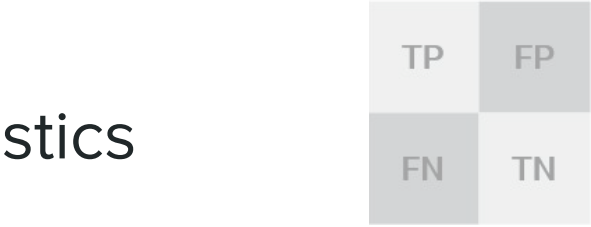


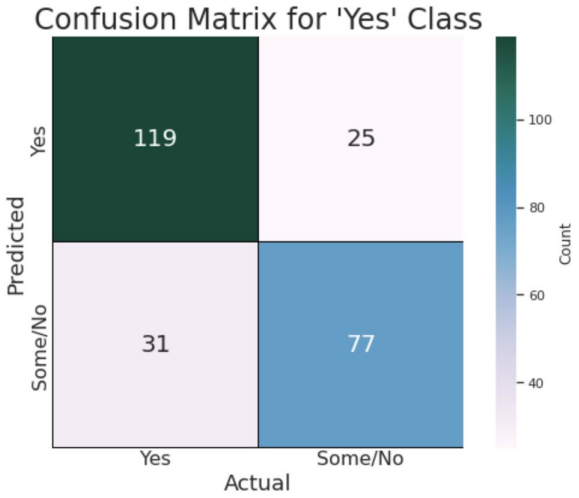
Feature Importance for ‘No’

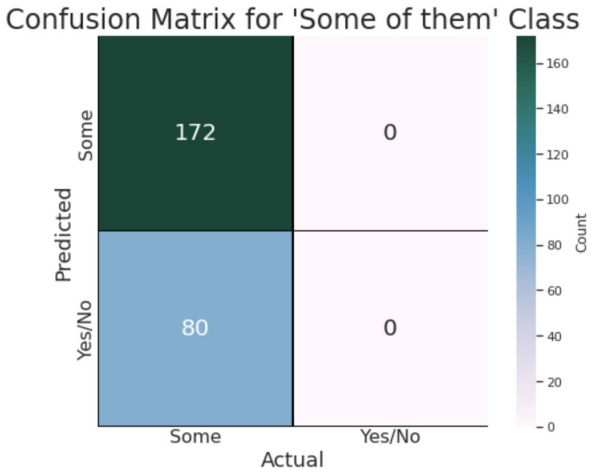
Random Forest using Gini

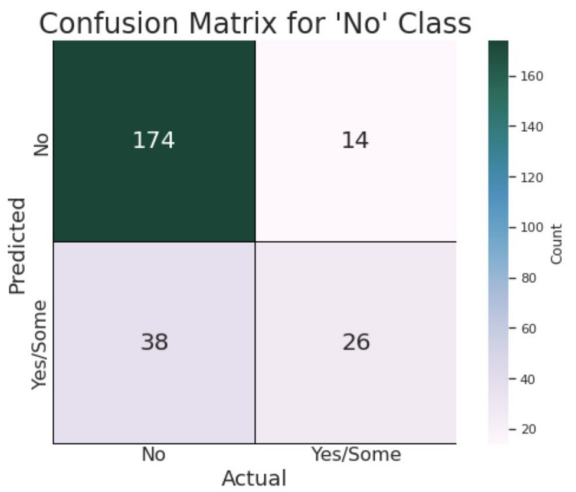


Confusion Matrices & Summary Statistics









Feature Importance Magnitude

Top 5 Greatest Importances

Original_Column	Feature_Importance
mental_health_consequence	0.457977
mental_vs_physical	0.092797
leave	0.080582
Age	0.043406
anonymity	0.028990

Top 5 Smallest Importances

	Original_Column	Feature_Importance
4	care_options	0.024788
6	family_history	0.016607
5	country_bucket	0.015575
1	Gender	0.012911
3	benefits	0.011813

Interpreting the Results

- We cannot reject the null hypothesis based on mean decrease in gini
  - Only **leave** was within the top five features ranked by importance
  - The top five features we found with our best model:
    - **mental_health_consequence**
    - **mental_vs_physical**
    - **leave**
    - **age**
    - **anonymity**
  - **benefits** was actually found to be the least most important feature
  - **no_employees**, **obs_consequence**, and **remote_work** were not near the top five strongest predictors

Interpreting the Results (Continued)

- **Mental_vs_physical**: Do you feel that your employer takes mental health as seriously as physical health?
  - Intuitively it is clear to see why this would be the one of the top predictors of feeling comfortable with talking to your supervisor about mental health issues
    - If you feel that your supervisor takes mental health as serious as physical health, you are more likely to feel comfortable reaching out to them about mental health
- **Mental_health_consequence:** Do you think that discussing a mental health issue with your employer would have negative consequences?
  - **Mental_health_consequences** inherently tracks the substantial stigma associated with having a mental illness in the world today, so it is not surprising that this would be a top predictor as well
    - Ultimately, we thought the **obs_consequence** feature would be sufficient to track this, but **mental_health_consequence** may be better because it is what an employee thinks will happen to themselves
    - Not what they observed to happen to someone else (**obs_consequence**)

Interpreting the Results (Continued)

- **anonymity**: Is anonymity protected if an employee takes advantage of mental health or substance abuse treatment?
  - Also an intuitive explanation, as if an identity is protected, an employee will more likely to be comfortable reaching out to others, including a supervisor, with mental health concerns
- **age** {continuous}
  - This is surprising, but there are two explanations for this:
    - There was a large unimodal distribution of age highly in favor of employees in their 20s and 30s
    - Furthermore, stigma may be more debilitating to older members of the workforce who are accustomed to mental health not being discussed at all
- **leave :** How easy is it for you to take medical leave for a mental health condition?
  - This was the only one present in our hypothesized target set!

Interpreting the Results (Continued)

	Yes	Some of Them	No
Base Rate	41%	27%	31%
Accuracy	78%	68%	79%
F1 Score	0.81	0.81	0.87
Recall	0.79	0.68	0.82
Precision	0.83	**1.0**	0.93

Model Limitations

- Generalizability
  - Low samples sizes from “Other” Countries
  - Healthcare structures not identified, U.S Healthcare highly variable
  - Gender/cultural differences per country
  - Small dataset (1,259 survey responses across several different countries)
    - Far from representative of the tech workforce population (strongest predictors unlikely to be the same if conducting analysis with a more representative dataset)
- Continuous variables have greater statistical power than categorical
  - Continuous variables have higher sensitivity while categorical ones have low sensitivity
    - Categorical variables also generally increase variance without regard to sample size
  - Binning country made model more interpretable but was not all inclusive
- Class imbalance
  - Model tends to perform better for the majority
    - Respondents: U.S
    - Target variable: Yes (41%)
  - Feature engineering and hypertuning was only performed through analysis on ‘Yes’ class

Future Avenues

- Explore options to improve class imbalance
  - Downsample ‘Yes’ (41%) compared to ‘Some’ (27%) or ‘No’ (32%)
  - Relevel the class to make ‘some’ equal to either ‘yes’ or ‘no’, or try both and observe differences
- Explore more specific health policies in the workplace & their effect on predictability
  - Maternity/paternity leave
  - Subsidized child care
- Explore more specific health structures & their effect on the response variable
  - Privatized insurance, universal healthcare, public option, etc.

Future Avenues

- Include a time series analysis
  - OSMI, organization that conducted the survey, have data from similar surveys taken between 2016 and 2020 (can observe how predictors change over time; can even make specific to COVID-19 remote work)
- Utilize all or expand dataset features
  - Specific types of mental health + coverage
- Perform feature engineering & hypertuning with respect to each of the label's classes
- Explore different machine learning models through utilizing multi-class classification

Conclusion

- Dataset: 2014 OSMI survey measuring attitudes towards mental health and frequency of mental health disorders in the tech workplace.
- Null Hypothesis:
  - Using Random Forest and feature importance from scikit-learn (i.e., mean decrease in Gini Index), the 5 target predictors, in no particular order, will not be the most important when predicting whether employees are willing to discuss mental health issues with supervisors.
  - **obs_consequence**, **benefits**, **no_employees**, **leave**, & **remote_work**
- EDA:
  - Analyzed feature distributions with respect to response variable & removed highly correlated features

Conclusion

- Feature Engineering:
  - Tradeoff between categorical and continuous variables, bias vs. variance, & hypertuning model
- Results:
  - Only one of our target predictors was in the top five most important features (measured by mean decrease in Gini Index) for our model (**leave**)
- Conclusion:
  - Model resulted in substantially higher accuracies than base rates with respect to response variable class
  - We failed to reject the Null Hypothesis
  - Negative reinforcement effect (leave)
  - Discrepancy between offered and taken benefits
  - Mental and physical health

References

1. Depression. Accessed January 6, 2021. https://www.who.int/en/news-room/fact-sheets/detail/depression
2. Disability in the Workplace: A Unique and Variable Identity - Alecia M. Santuzzi, Pamela R. Waltz, 2016. Accessed January 6, 2021. https://journals.sagepub.com/doi/full/10.1177/0149206315626269
3. Mental Health in the Workplace. Published April 26, 2019. Accessed January 6, 2021. https://www.cdc.gov/workplacehealthpromotion/tools-resources/workplace-health/mental-health/index.html
4. COVID-19's Impact on Mental Health and Workplace Well-being. NIHCM. Accessed January 6, 2021. https://nihcm.org/publications/covid-19s-impact-on-mental-health-and-workplace-well-being
5. Mental Health in Tech Survey. Accessed January 6, 2021. https://kaggle.com/osmi/mental-health-in-tech-survey
6. Sado M, Shirahase J, Yoshimura K, et al. Predictors of repeated sick leave in the workplace because of mental disorders. *Neuropsychiatr Dis Treat*. 2014;10:193-200. doi:10.2147/NDT.S55490