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 5th, 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

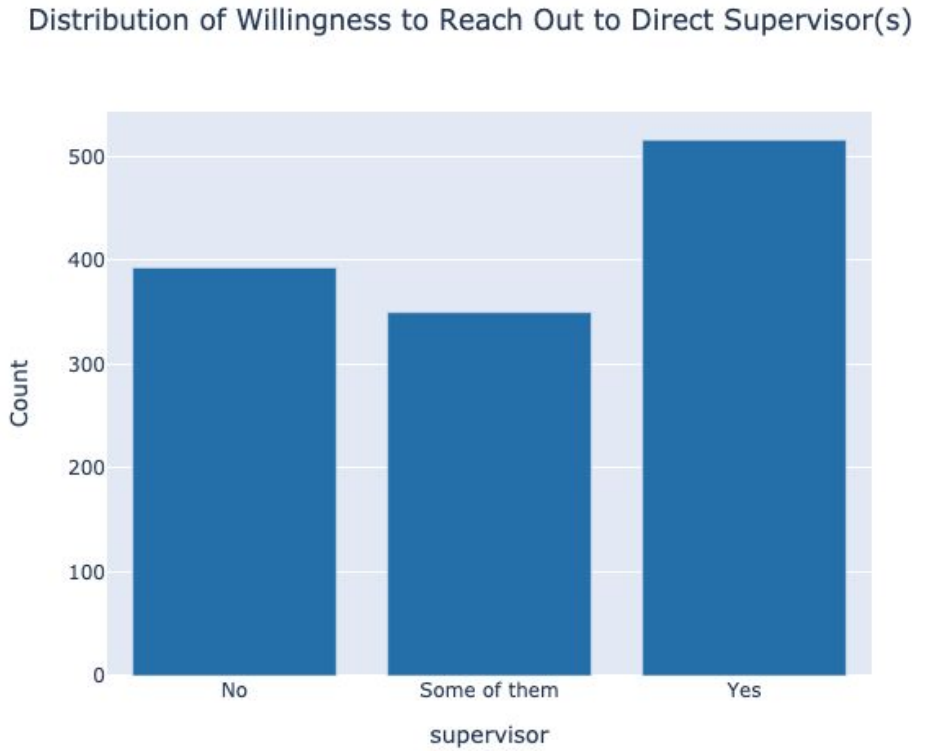
- Target predictors: Would you be willing to discuss a mental health issue with your direct supervisor(s) {e.g., **supervisor** in the dataset}?
  - ***Obs_consequence***: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
  - ***No_employees:***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

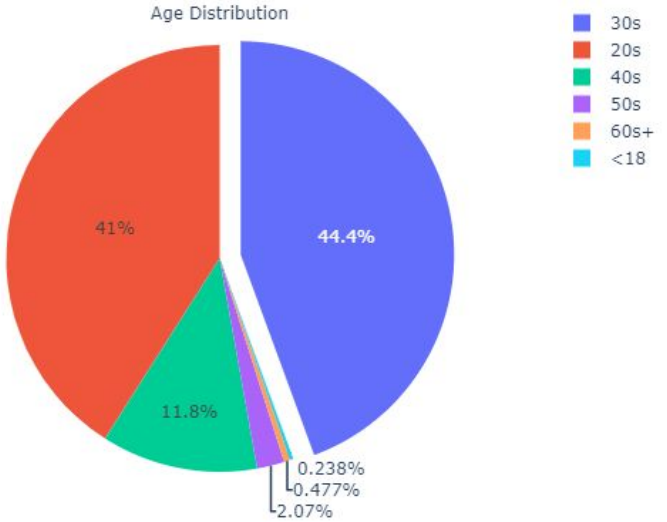
QUESTION 1: Generally, how willing are employees to reach out to their direct supervisor(s) about a mental health issue?

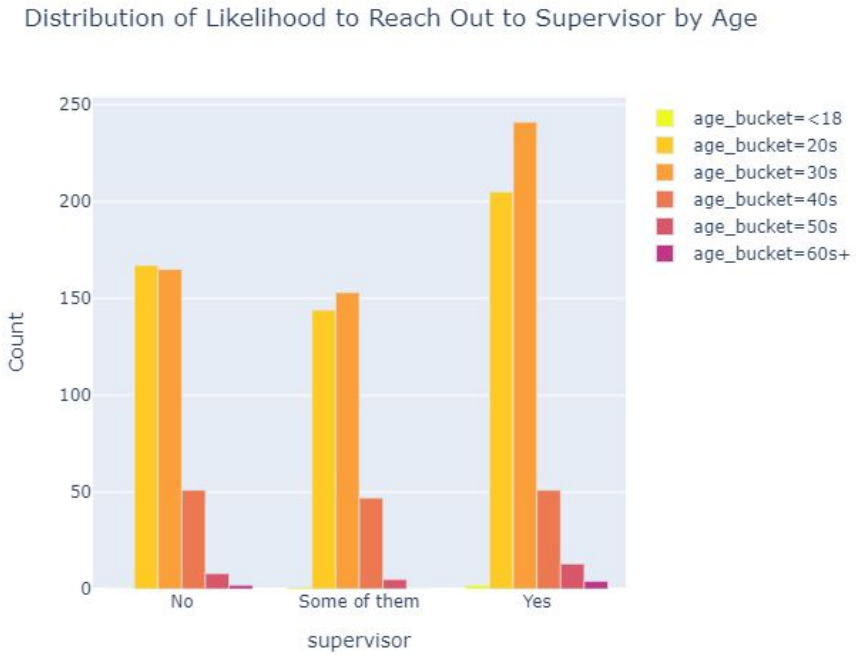


Visual Representation of the distribution of responses for our label

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

AGE



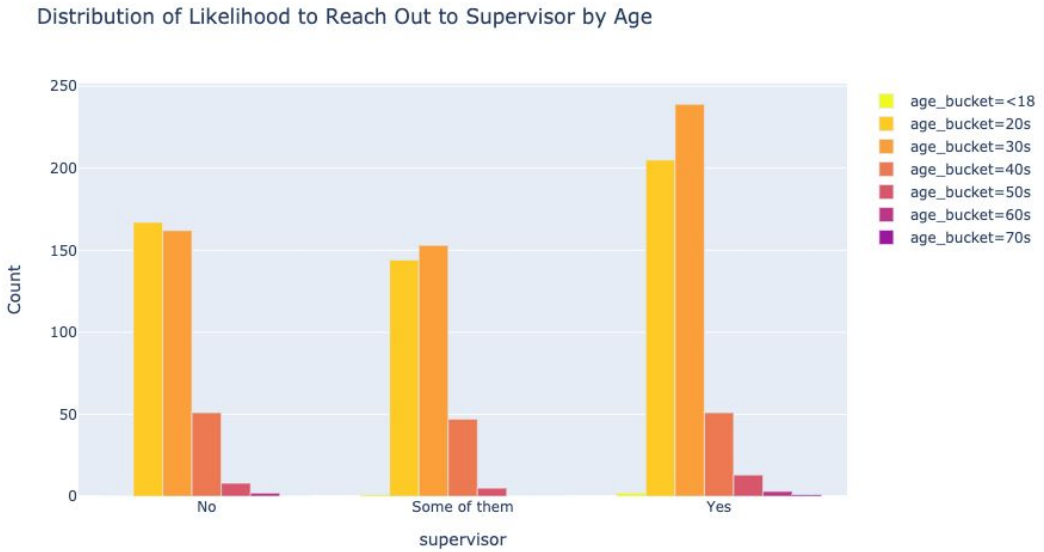


QUESTION 2: Do demographic indicators such as *age* and *gender* play a role in how willing employees are to reach out to their direct supervisor(s)?

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.

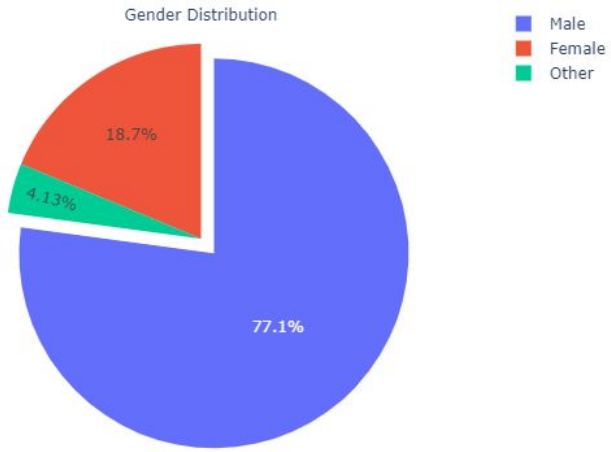
Conclusion: Due to only two groups showing noticeable variation, we decided that we would not add age into our group of target predictors.

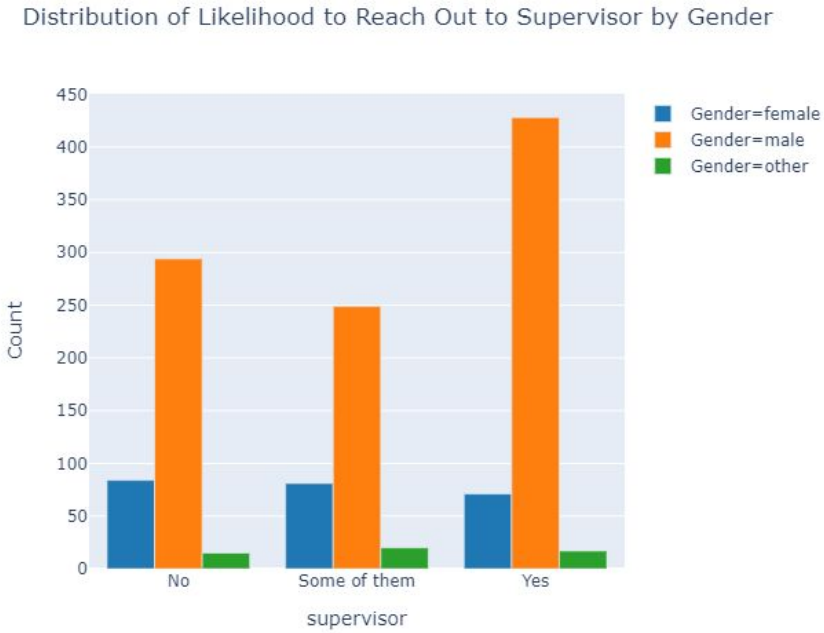


Feature	chi2	p-value
Age	0.3302	0.8478

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

GENDER



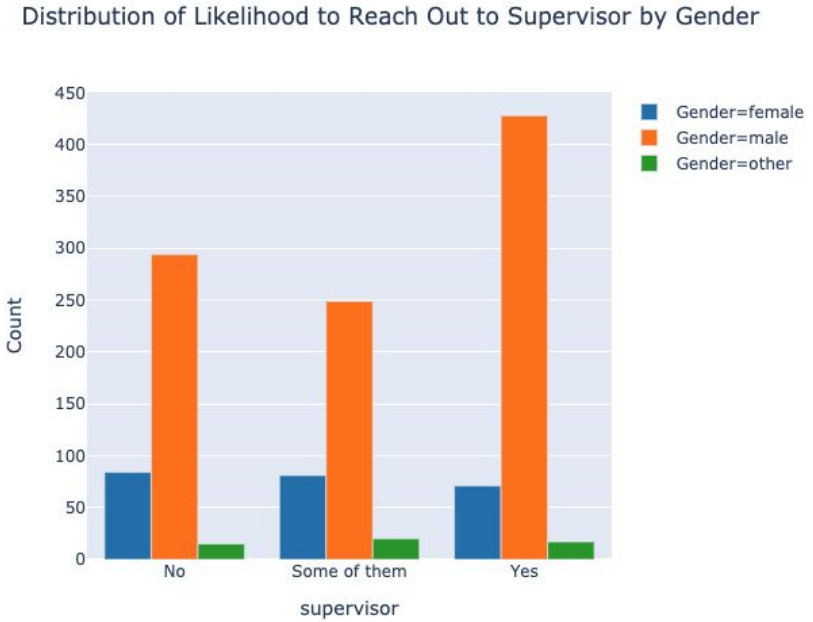


QUESTION 2: Do demographic indicators such as *age* and *gender* play a role in how willing employees are to reach out to their direct supervisor(s)?

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.



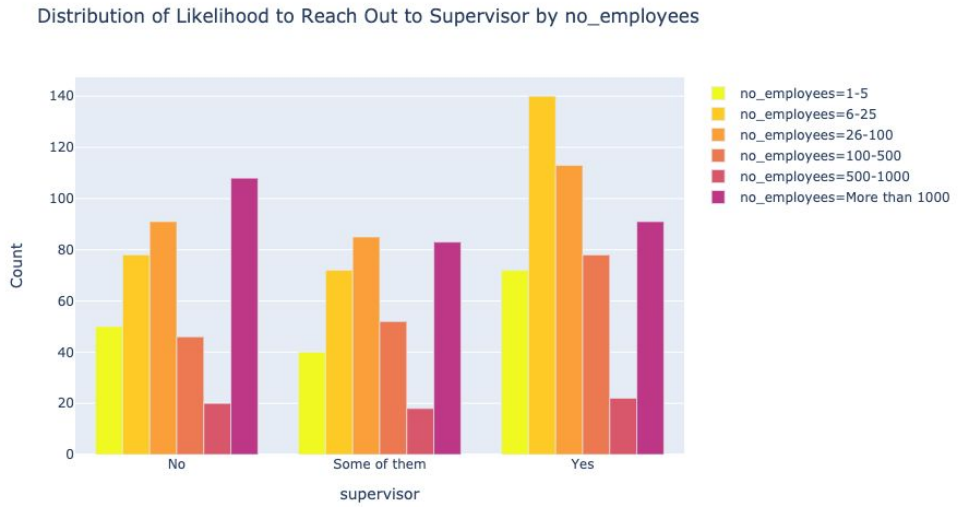
Feature	chi2	p-value
Age	1.7637	0.4140

QUESTION 3: 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

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

Conclusion: The outcome achieved is what was expected. 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.



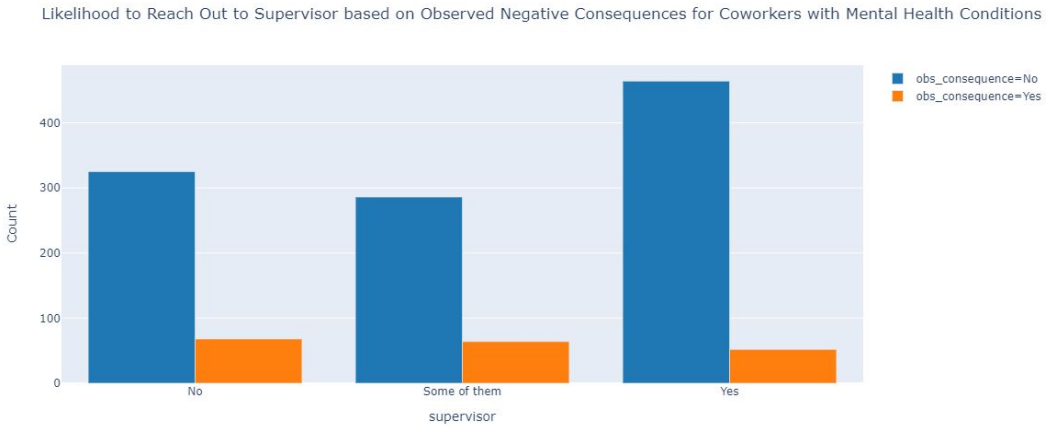
Feature	chi2	p-value
no_employees	3.7966	0.1498

QUESTION 3: 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)?

obs_consequence

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 had expected that people who had observed negative consequences were going to be less likely to feel comfortable bringing up their own issues with a supervisor.



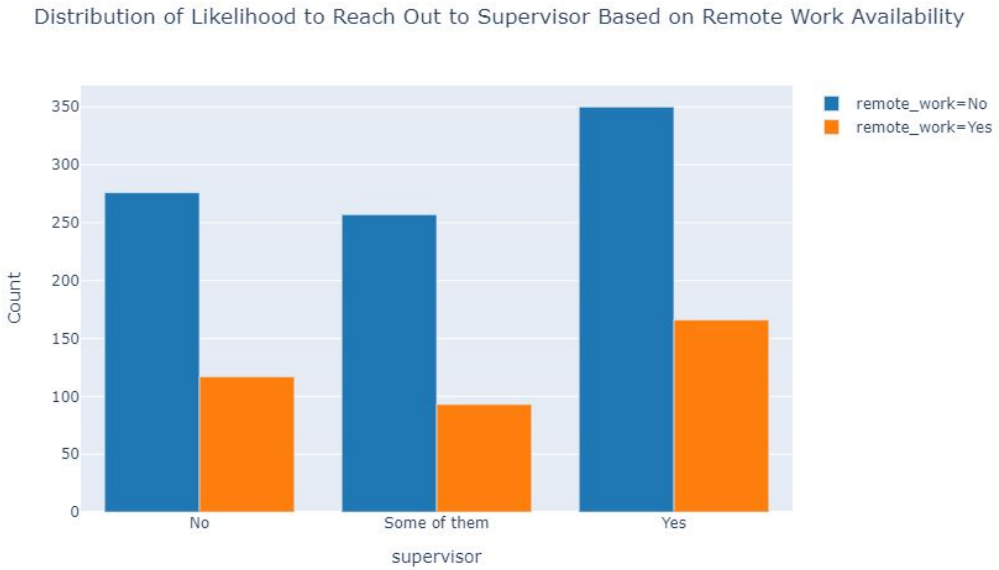
Feature	chi2	p-value
obs_consequence	12.4387	0.0020

QUESTION 3: 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)?

remote_work

Outcome:

Conclusion:



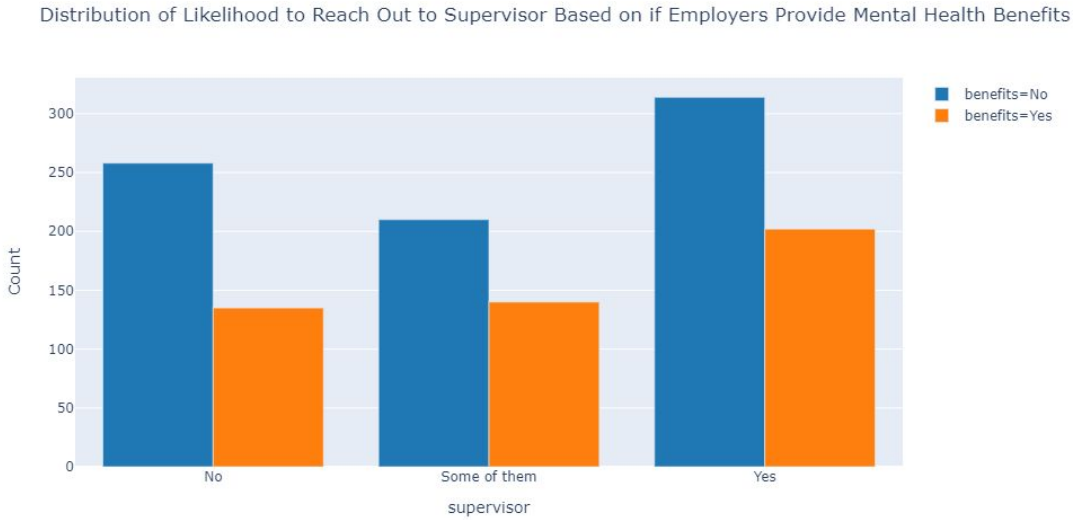
Feature	chi2	p-value
remote_work	2.1908	0.3344

QUESTION 3: 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)?

benefits

Outcome:

Conclusion:



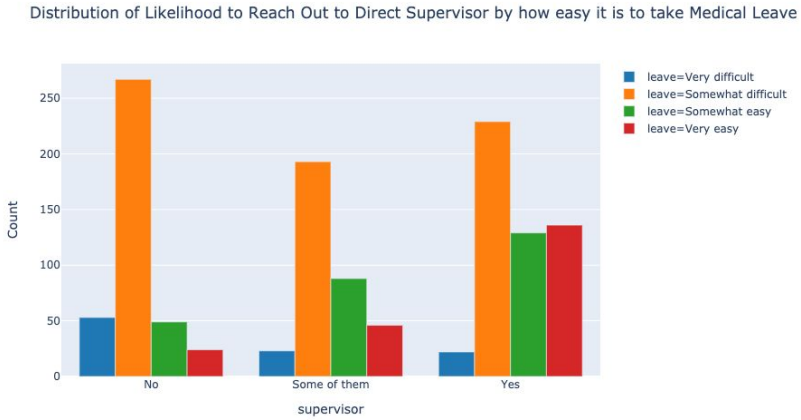
Feature	chi2	p-value
benefits	1.9256	0.3818

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.



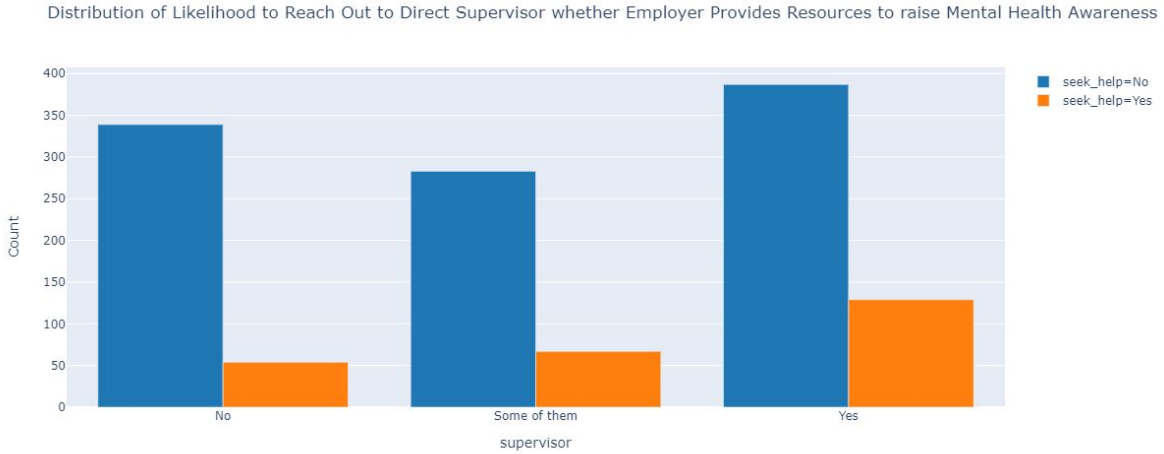
Feature	chi2	p-value
leave	81.8797	1.6598e-18

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:

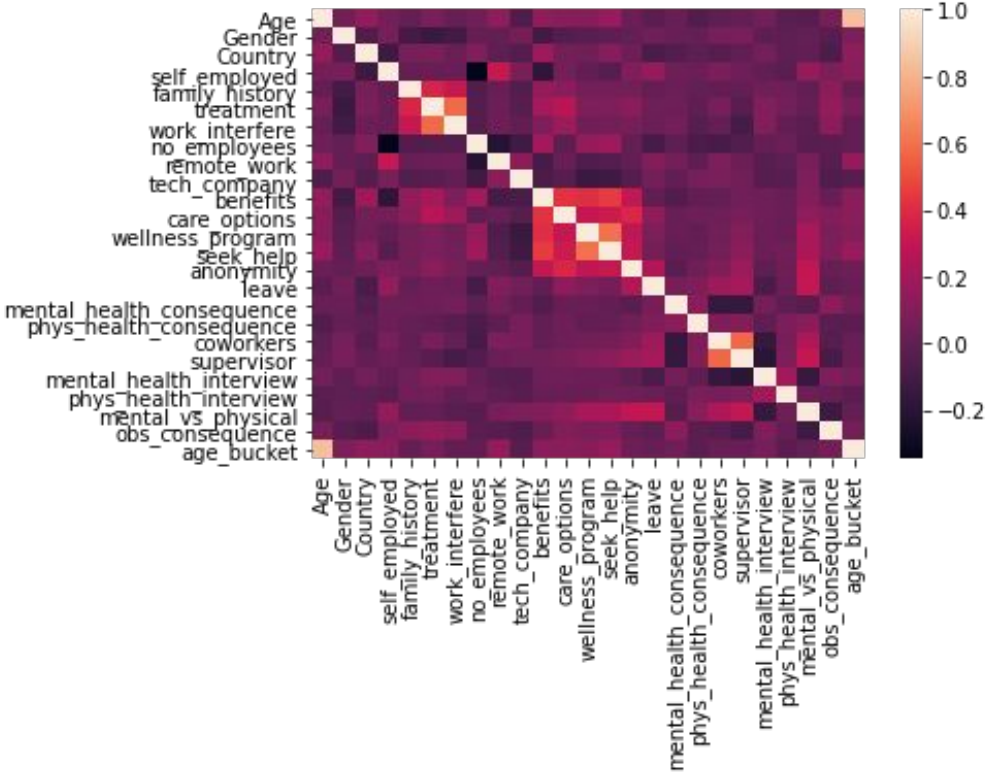
Conclusion:



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
  - **leave** - Availability of mental health leave
    - 0.21
  - **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	-0.0013	*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 normalized permutation importance from sklearn, 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 and normalized permutation importance 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

Feature Engineering

- Dropped **timestamp**, **state**, **coworkers**, **mental_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 a country variable that we kept
  c. We dropped **coworkers** because it has the exact same question as supervisor, our label, but for coworkers
     i. We thought including this category would introduce overfitting to our model
  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 **comments** because the it contains many null values, and the values with responses are all very different, so it would be very hard to format this data to actually benefit our model
- 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
- We bucketed **age** as a factor for plotting, thereby additionally making it a discrete variable
- We imputed **age** outliers with their mean (32)
- We used **get_dummies** (pandas library) to turn all of our categorical features into dummy variables

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
```

Initial Model: Random Forest Using Gini/Encoding

```
#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
```

Our testing score is already 0.746 on our original model before hyperparameter tuning

Initial Model: Random Forest Using Entropy/Encoding

```
#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 testing score is slightly better than the previous one 0.7579 before hyperparameter tuning

Initial Model: Random Forest Using Entropy/Categories

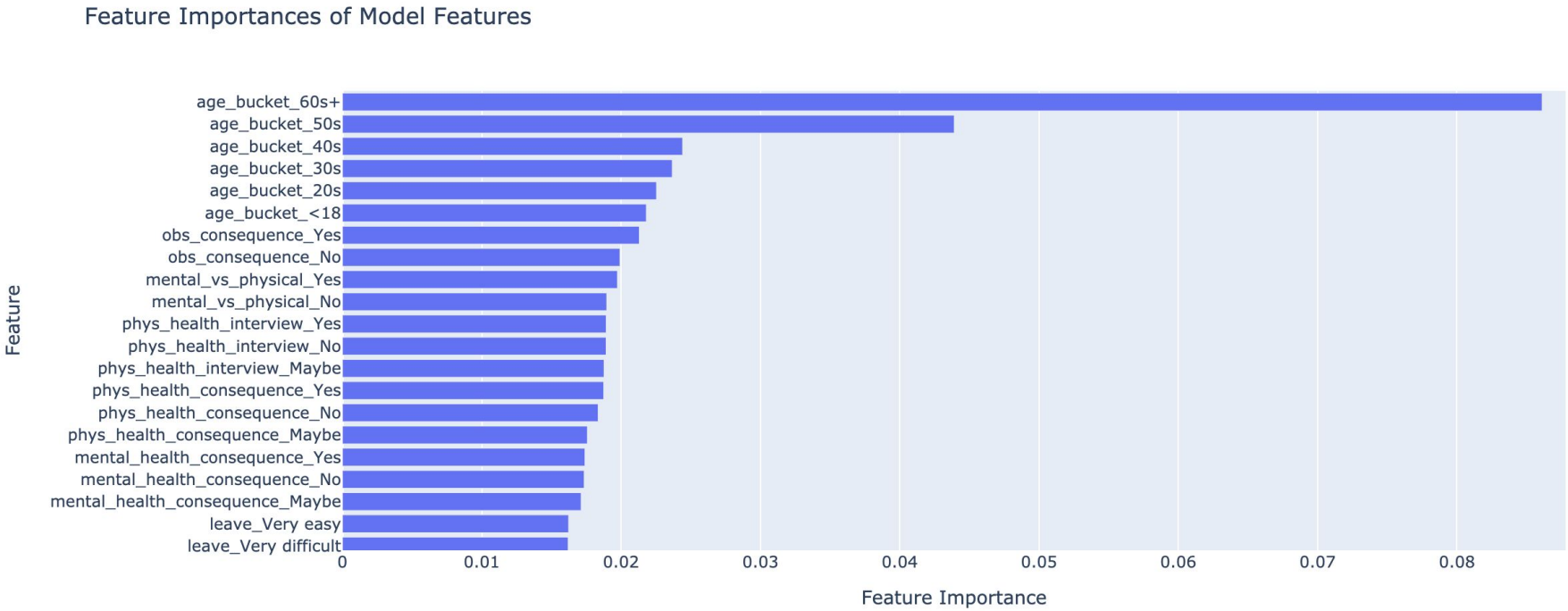
```
#Instantiate and fit Entropy
RF_entropy = RandomForestClassifier(n_estimators = 100,
                                    criterion = 'entropy',
                                    max_depth = None,
                                    min_samples_leaf = 1,
                                    bootstrap = 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
```

This testing score is lower than the data encoded and also displays signs of overfitting

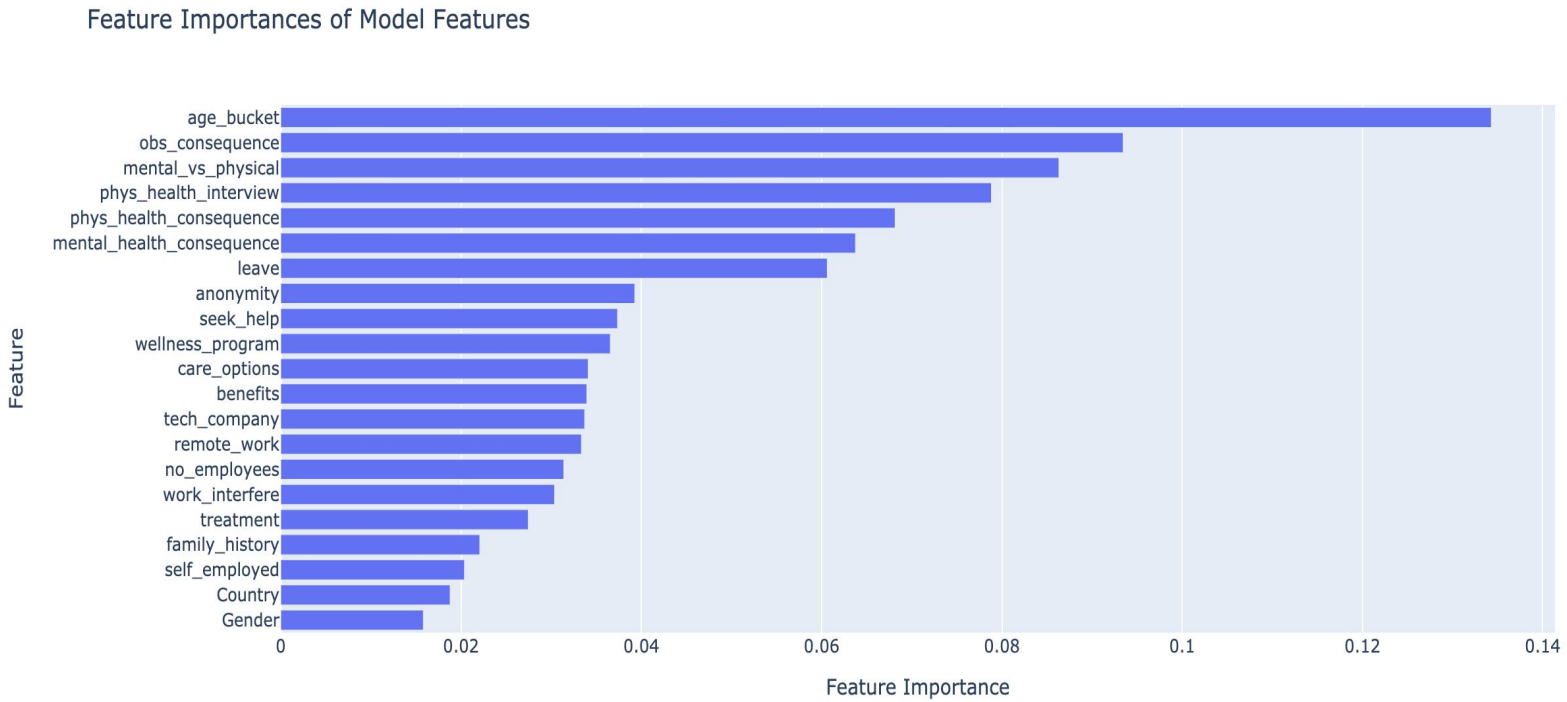
Feature Importance with Encoded Labels

Random Forest using Entropy



Feature Importance with Categorical Labels

Random Forest using Entropy



Initial Model Response

- **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 cross-validation
  - Bootstrapping

Day 4: Finalized Model Results, Discussion, & Conclusion

Day four stuff

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