Mental Health and Employment

Bradley Caldwell, Ariana Davis, and Stacey Marotta

April 2023

List of Technologies

- QuickDB
- pgAdmin with postgresSQL
- Jupyter Notebooks with Python
- Visual Studio Code
- Tableau
- Github

Communication Plan

- Slack Group
- Zoom Meetings during class, and additional meetings outside of class.

Part 1 - Project Intro + Data Exploration

Project Statement

The purpose of this analysis is to use Supervised Machine Learning to understand whether or not employers are providing healthcare benefits that includes mental healthcare in the workplace.

Our project would like to examine any relationships between mental health and employment. The dataset selected was survey data that asked participants to talk about their experiences with healthcare and mental health in the workplace.

Questions

- Does your employer provide healthcare that includes mental health care?
- Do you find it possible to be productive even with mental health restraints?
- Do you find your employer is supportive of you and your mental health needs?
- Does having a mental health disorder have negative consequences in the workforce?
- Do companies support employees’ mental health when working from home?
- Do all self-employed people work from home?
- Do companies foster an environment to speak openly about mental health disorders?

Questions chose not to go with after all:

- Does working remotely have a positive or negative impact on one’s employment?
- What support/best practices do employers provide for mental health when their employees work from home?
- Do people with mental health disorders work from home more than those who do not have a mental health disorder?

Data Exploration

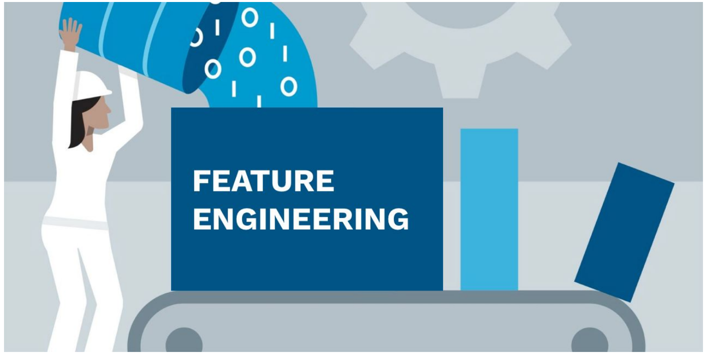
Preliminary Data Preprocessing

1. Correlation Matrix
2. Heat Map
3. Plots

Feature Selection

- Target
- Interest
- Future Analyses & Recommendations

Feature Engineering



There were a number of things we had to do in order to transform the given data into a form that was easier to interpret. First, we started off by renaming the columns so they could be more concise and easier to read while coding.

When choosing features we had columns with:

- Categorical Features; since algorithms are not designed to process textual data we decided to use one-hot encoding
- Multiple missing values; we used fillna() to replace all null values with NA.
- Our target was a categorical value as well, so we converted it to an integer to avoid being encoded

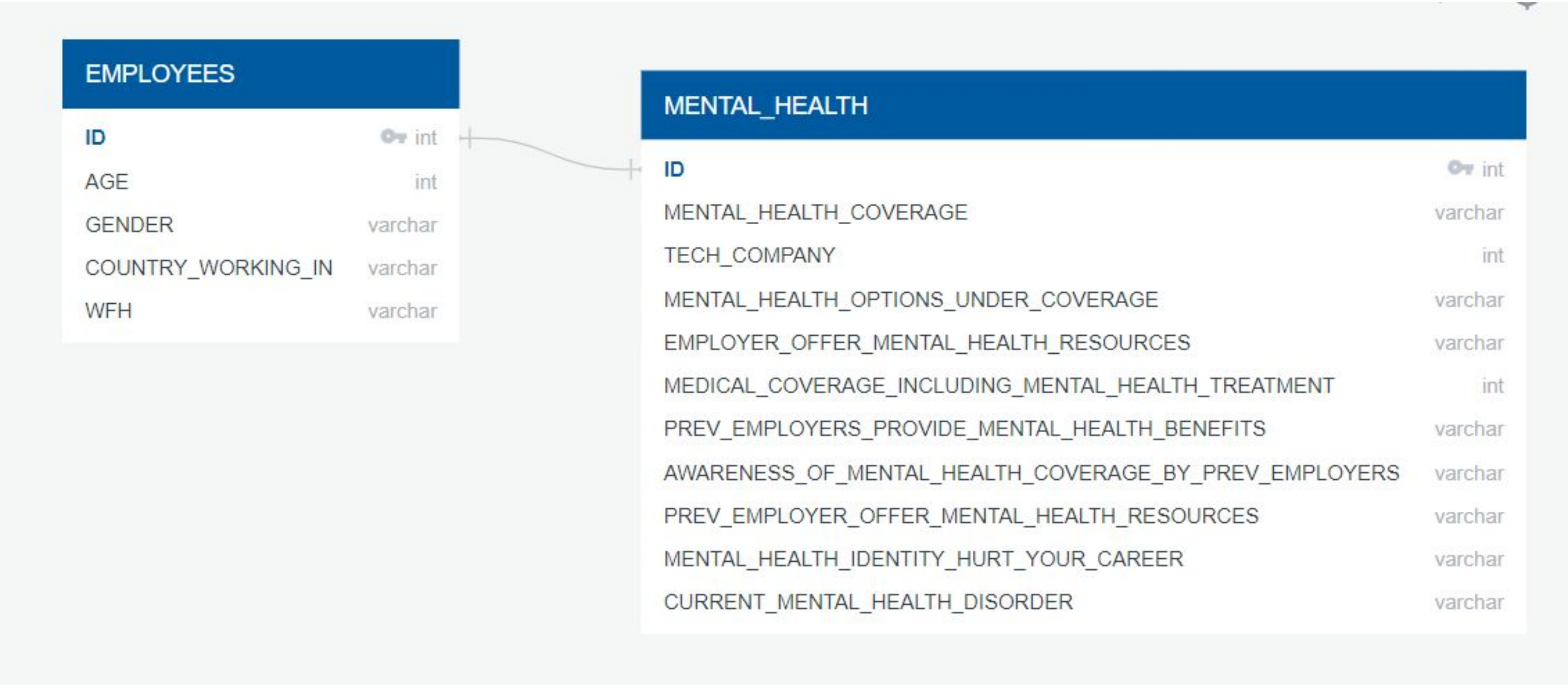
Features

Features				
‘TECH_COMPANY’	'MENTAL_HEALTH_OPTIONS_UNDER_COVERAGE'	'EMPLOYER_OFFER_MENTAL_HEALTH_RESOURCES'	'MEDICAL_COVERAGE_INCLUDING_MENTAL_HEALTH_TREATMENT'	'PREV_EMPLOYERS_PROVIDE_MENTAL_HEALTH_BENEFITS'
'AWARENESS_OF_MENTAL_HEALTH_COVERAGE_BY_PREV_EMPLOYERS'	'PREV_EMPLOYER_OFFER_MENTAL_HEALTH_RESOURCES'	‘MENTAL_HEALTH_IDENTITY_HURT_CAREER’	‘CURRENT_MENTAL_HEALTH_DISORDER’	‘AGE’
‘GENDER’	‘COUNTRY_WORKING_IN’	‘WFH’		

Target

Target
'MENTAL_HEALTH_COVERAGE?'

Database



EMPLOYEES	
ID	🔑 int
AGE	int
GENDER	varchar
COUNTRY_WORKING_IN	varchar
WFH	varchar

MENTAL_HEALTH	
ID	🔑 int
MENTAL_HEALTH_COVERAGE	varchar
TECH_COMPANY	int
MENTAL_HEALTH_OPTIONS_UNDER_COVERAGE	varchar
EMPLOYER_OFFER_MENTAL_HEALTH_RESOURCES	varchar
MEDICAL_COVERAGE_INCLUDING_MENTAL_HEALTH_TREATMENT	int
PREV_EMPLOYERS_PROVIDE_MENTAL_HEALTH_BENEFITS	varchar
AWARENESS_OF_MENTAL_HEALTH_COVERAGE_BY_PREV_EMPLOYERS	varchar
PREV_EMPLOYER_OFFER_MENTAL_HEALTH_RESOURCES	varchar
MENTAL_HEALTH_IDENTITY_HURT_YOUR_CAREER	varchar
CURRENT_MENTAL_HEALTH_DISORDER	varchar

Part 2 - Model Preparation + Analysis

Training and Testing Sets

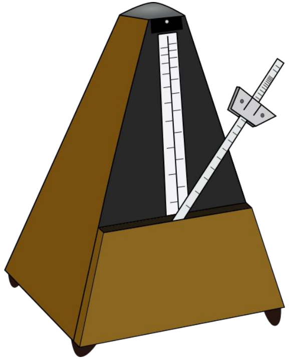
The training and testing sets were selected using scikit learn’s train_test_split().

Random seed 42 was selected as it’s the answer to the “Great Question” of “Life, the Universe and Everything”.

Machine Learning Model

- Logistic Regression
  - Accuracy Score ≈ 0.74
- Random Forests
  - Accuracy Score ≈ 0.81
- Gradient Boosting Classifier
  - Accuracy Score ≈ 0.83

Hyper Parameter Tuning

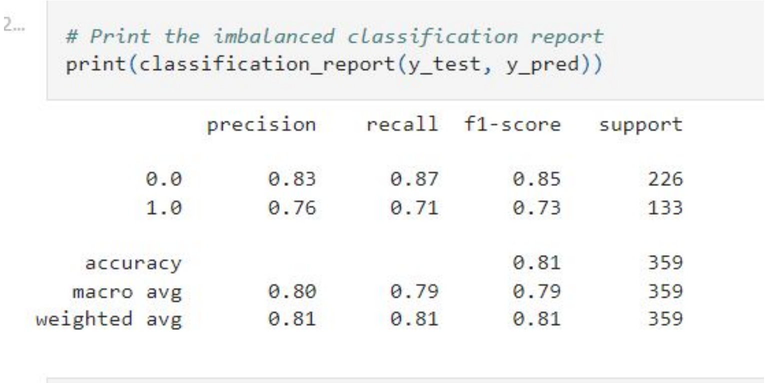


Analysis

How accurate the model is at predicting mental health coverage?

Gradient Boosting Classifier gave the best results

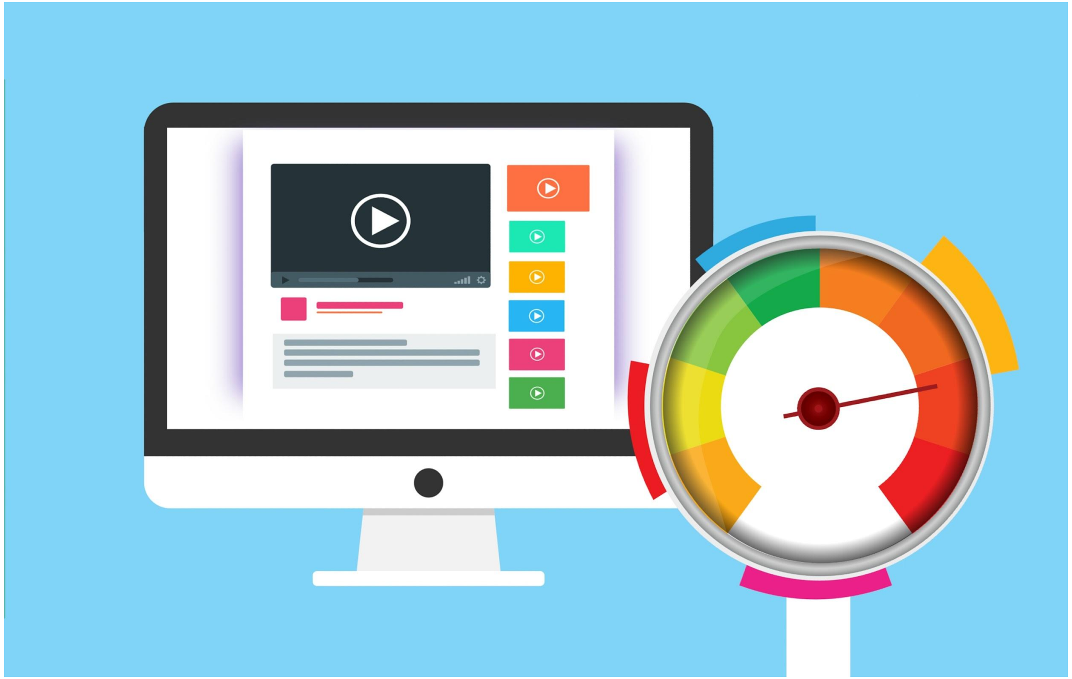
Accuracy Score = .83



```
# Print the imbalanced classification report
print(classification_report(y_test, y_pred))
```

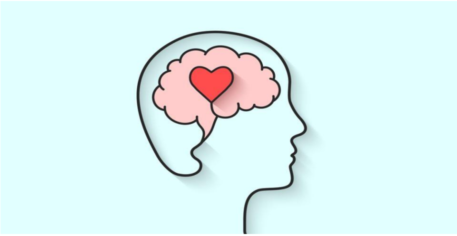
```
              precision    recall  f1-score   support

         0.0       0.83      0.87      0.85       226
         1.0       0.76      0.71      0.73       133

    accuracy                           0.81       359
   macro avg       0.80      0.79      0.79       359
weighted avg       0.81      0.81      0.81       359
```

Part 3 - Dashboard Preparation + Demo



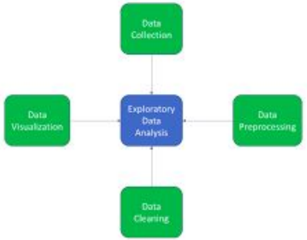
Storyboard

Question->DataBase->EDA->ML->Question Answered?



Will your potential employer offer mental health coverage in your healthcare plan?

Our database was created with PostgreSQL. 2 Tables: EMPLOYEES & MENTAL_HEALTH



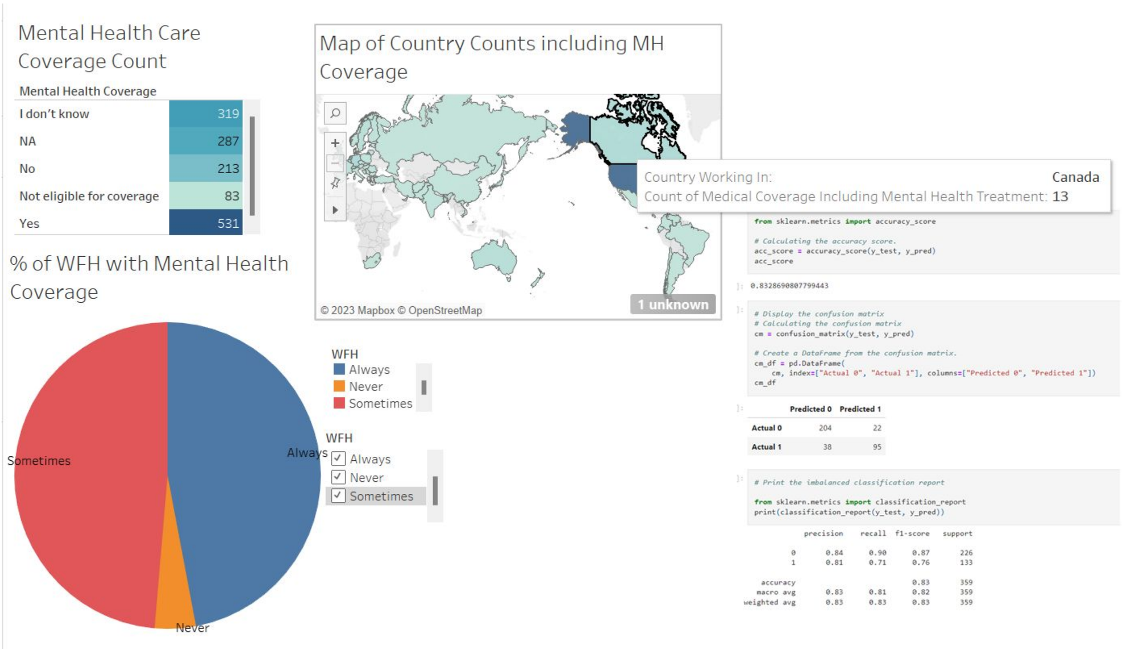
EDA was performed by cleaning & encoding data, and creating a heat map, and plots to visualize the data.

Meaningful Compression
Structure Discovery
Big data Visualization
Dimensionality Reduction
Feature Elicitation
Image Classification
Customer Retention
Identity Fraud Detection
Classification
Diagnostics
Recommender Systems
Unsupervised Learning
Supervised Learning
Advertising Popularity Prediction
Weather Forecasting
Clustering
Machine Learning
Regression
Targeted Marketing
Population Growth Prediction
Market Forecasting
Customer Segmentation
Estimating life expectancy
Real-time decisions
Game AI
Reinforcement Learning
Robot Navigation
Skill Acquisition
Learning Tasks

Gradient Boosting Classifier algorithm was used to achieve 83% accuracy

Dashboard Demo

The Dashboard will be use images from the jupyter notebook showing our phases of the machine learning process. Along with that, we will be using Tableau in order to create the interactive elements. To the right is our beginning stages showing the total counts of respondents for Mental Health Care coverage, and of those covered do they work from home? There is also a map of the countries showing the count of respondents covered by country.



Mental Health Care Coverage Count

Mental Health Coverage	
I don't know	319
NA	287
No	213
Not eligible for coverage	83
Yes	531

% of WFH with Mental Health Coverage

WFH
☑ Always
☑ Never
☑ Sometimes

Map of Country Counts including MH Coverage

Country Working In: Canada
Count of Medical Coverage Including Mental Health Treatment: 13

```
from sklearn.metrics import accuracy_score

# Calculating the accuracy score.
acc_score = accuracy_score(y_test, y_pred)
acc_score
```

```
0.8328690807799443
```

```
# Display the confusion matrix
# Calculating the confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Create a DataFrame from the confusion matrix.
cm_df = pd.DataFrame(
    cm, index=["Actual 0", "Actual 1"], columns=["Predicted 0", "Predicted 1"])
cm_df
```

	Predicted 0	Predicted 1
Actual 0	204	22
Actual 1	38	95

```
# Print the imbalanced classification report

from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.84      0.90      0.87       226
           1       0.81      0.71      0.76       133

    accuracy                           0.83       359
   macro avg       0.83      0.81      0.82       359
weighted avg       0.83      0.83      0.83       359
```

Recommendations for Future Analysis

Do Overs

- Dataset (combine multiple datasets to get a bigger sample)

Reset

Do you want to start over?

Presentation Rubric Slide [Remove for final submission]

To meet the requirements, the presentation must tell a cohesive story about the project and include the following:

- The selected topic and the reasoning for that selection. (6 points)
- A description of the data source. (6 points)
- The questions that the team planned to answer with the data. (6 points)
- A description of the data exploration phase of the project. (6 points)
- A description of the analysis phase of the project. (6 points)
- The technologies, languages, tools, and algorithms that the team used throughout the project. (10 points)
- The results of the analysis. (10 points)
- Any recommendations for a future analysis. (10 points)
- Anything that the team would have done differently if they had more time. (10 points)
- Live Action Demo of Dashboard