Springboard–DSC Program
Capstone Project 2 Proposal

Predicting Mental Health Support Needs

Mental Health and Well-being: Predicting the Need for Mental Health Support

Prepared by:

Jude M. Santos
Data Scientist

September, 2024

Business Problem

The rising demand for mental health services is growing, but resources are unevenly distributed. One significant challenge is the lack of insight into who is most likely to seek mental health support. As a result, care may be delayed or inefficiently allocated, and patients may not receive timely support. This project seeks to address this challenge by developing predictive and descriptive models to identify individuals most likely to seek mental health services, enabling targeted outreach and resource allocation.

The primary objective of this project is to predict whether an individual will seek mental health support. The model will output either:

- A binary classification (1 = seeks support, 0 = does not seek support), and
- A probabilistic score indicating the likelihood of seeking support.

Stakeholders and Relevance

Healthcare Providers: By predicting which individuals are likely to seek mental health support, healthcare providers can better allocate resources, plan staffing, and provide timely care, leading to improved patient outcomes.

Insurance Companies: Insurers can leverage these insights to design preventative mental health programs, optimizing their services, and reducing long-term costs related to untreated mental health conditions.

Dataset

The dataset for this project will be sourced from Kaggle's Mental Health Dataset: https://www.kaggle.com/datasets/bhavikjikadara/mental-health-dataset

Data Science Approaches

We will approach this problem using supervised and unsupervised learning techniques to model the likelihood of individuals seeking mental health support. Some of the methods we plan to implement include:

Feature Selection and Engineering

Critical features like age, gender, socioeconomic status, and mental health conditions will be selected and engineered to refine the model and enhance its accuracy.

Logistic Regression (Baseline)

We will use logistic regression as our baseline model for predicting mental health service utilization.

Random Forest Classifiers

To explore feature interactions and improve the accuracy of predictions, we will implement a Random Forest classifier.

Unsupervised Learning (Clustering)

In addition to supervised models, we will also explore unsupervised learning techniques like clustering to group individuals with similar characteristics who might require mental health support. This provides an alternative perspective and may complement the supervised models. We can also explore Clustering as a way to estimate the probability of a person needing mental health care using approaches such as majority voting, or estimation of the probability based on cluster-based proportions.

Performance Evaluation

We will evaluate the performance of our models using various metrics and validation techniques:

Confusion Matrix and Classification Report

- Provide a comprehensive overview of the model's performance.
- Calculate metrics such as accuracy, precision, recall, and F1-score to assess overall correctness, positive predictive value, sensitivity, and harmonic mean of precision and recall.

Precision-Recall Curves

- Visualize the trade-off between precision and recall.
- Identify the threshold that represents the best compromise between precision and recall.

Cross-validation

- Estimates model hyper-parameters over various partitions of the training dataset.
- Prevents overfitting.
- Uses techniques like k-fold and stratified k-fold.
- Provides reliable performance estimates.

For interpretability

Feature Importance Analysis ranks features with respect to how much their impact in computing predictions.

SHAP (SHapley Additive exPlanations) will provide detailed insights into how individual features contribute to specific predictions, to the model's global performance, and feature interactions.