Springboard-DSC Program
Capstone Project 2

Predicting Mental Health Support Needs

By Jude M. Santos
October, 2024

Table of Contents

1. Introduction

Mental health issues are increasingly prevalent in society, with significant implications for individuals and communities. This project aims to identify groups most likely to require mental health support by analyzing demographic data, mental health conditions, sentiment analysis scores, and psychological indicators.

Stakeholders include mental health organizations, policymakers, and community support services, who can utilize the findings to allocate resources effectively.

The data science results, including model performance metrics and insights, will guide strategic interventions in mental health support.

Detailed implementation can be found in the notebooks developed throughout the project, available at GitHub.

2. Approach

2.1. Data Acquisition and Wrangling

The dataset for this project was sourced from Kaggle, combining online survey data and publicly available mental health datasets.
After initial acquisition, the data underwent exploration and wrangling processes, which included:

- **Target Identification**: The target variable is 'treatment,' indicating whether an individual receives mental health treatment. This is a categorical variable, making it suitable for classification modeling.
- **Feature Transformation**: All features are categorical and required transformation for modeling. Binary features were converted into numerical values, and categorical variables were one-hot encoded.
- **Irrelevant Feature Removal**: Features such as the 'Timestamp' column were removed for better model focus.
- **Imputation**: Missing values in categorical variables were filled with 'Unknown' to maintain consistency.

Addressing Duplicates

During the exploratory data analysis phase, we discovered a large number of duplicate records. Initial modeling with Logistic Regression and XGBoost showed high accuracy, with XGBoost reaching nearly 100%. However, this raised concerns about potential overfitting due to duplicate observations. To test this, we compared the performance of models trained on datasets with duplicates (Dataset D) versus those without duplicates (Dataset ND).

- **Dataset D**: Contained 194,000 duplicates out of 292,364 observations.
- **Dataset ND**: No duplicates.

The model using Dataset D showed higher performance, while Dataset ND showed a noticeable decline. These results suggested that duplicates might have introduced bias, leading to overfitting. To investigate further, we applied a Chi-Squared test to assess differences in categorical feature distributions between the two datasets.

Chi-Squared Test Results

The Chi-Squared test confirmed that the distributions of several features differed significantly between the duplicated and deduplicated datasets. Key findings include:

- **Country_United_States**: Overrepresented (60% of the dataset), leading to overfitting on U.S.-specific patterns.
- **self_employed_Yes**: Imbalanced with 29,168 records for self-employed individuals versus 257,994 for non-self-employed, causing bias toward non-self-employed patterns.
- **mental_health_interview**: Imbalance with 232,166 “No” responses versus 8,624 “Yes,” skewing the model toward predicting “No.”
- **Country_Canada**: Representing 6.3% of the dataset, it still had less influence than the United States, potentially leading to underrepresentation of Canadian patterns.

The Chi-Squared test indicated significant bias due to duplicates. Removing duplicates should help reduce bias, leading to improved model generalization across diverse populations. As a result, the dataset without duplicates (ND) was used in the modeling phase, ensuring more accurate and reliable results.

Countries with single-class representation in the target feature.

The single-class representation of the target variable in some countries, combined with duplicated records, contributed to the overfitting of the initial model. The model learned patterns specific to the dominant class, resulting in inflated performance metrics and poor generalization. Removing the single-class instances ensures a more balanced dataset, reducing overfitting and improving the model's ability to generalize across all classes.

With the removal of single-class representations from certain countries, the main model now lacks representation from those countries, which could reduce its ability to make predictions for those specific groups. This might lead to biased or incomplete predictions, especially if the model is intended to be used globally or across diverse regions.

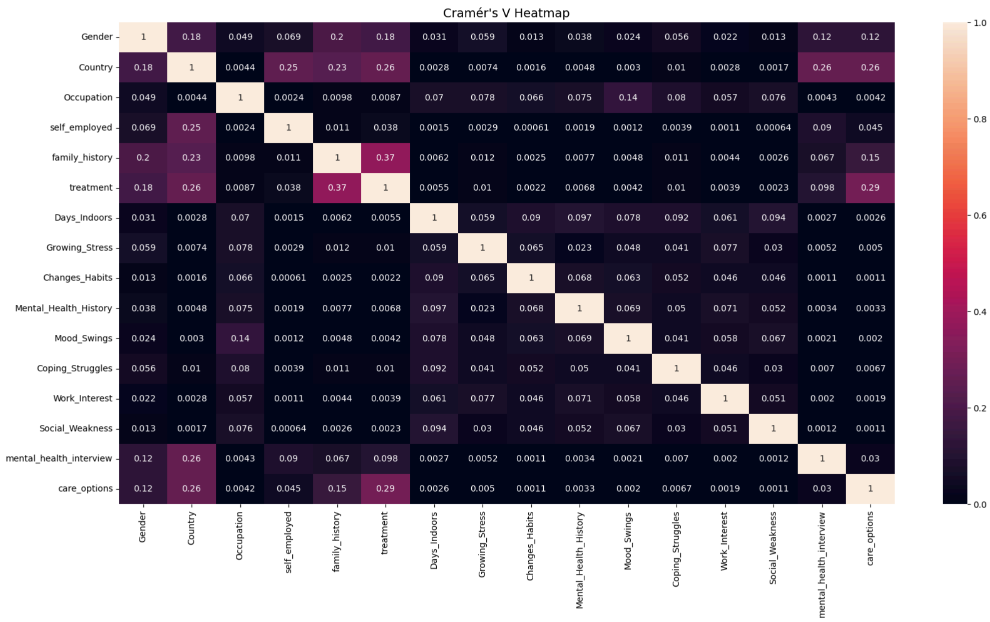
2.2. Storytelling and Inferential Statistics

The Data Structure

Gender	Country	Occupation	self_employed	family_history	treatment	Days_Indoors	Growing_Stress	Changes_Habits	Mental_Health_History	Mood_Swings	Coping_Struggles
Female	United States	Corporate	NaN	No	Yes	1-14 days	Yes	No	Yes	Medium	No
Female	United States	Corporate	NaN	Yes	Yes	1-14 days	Yes	No	Yes	Medium	No
Female	United States	Corporate	NaN	Yes	Yes	1-14 days	Yes	No	Yes	Medium	No
Female	United States	Corporate	No	Yes	Yes	1-14 days	Yes	No	Yes	Medium	No
Female	United States	Corporate	No	Yes	Yes	1-14 days	Yes	No	Yes	Medium	No

The dataset comprises **292,364** observations and includes features that provide insights into the demographics and mental health conditions of individuals.This analysis explores key factors affecting mental health treatment, including mental health history, treatment, coping strategies, and social factors. The aim is to identify relationships that could help improve mental health care outcomes.

Feature Relationships



1. **Strong Associations**:
   - **Family History & Treatment**: A significant correlation (Cramér's V: 0.37) between family history and receiving treatment suggests that individuals with a family history of mental health issues are more likely to seek treatment.
   - **Care Options & Treatment**: Moderate correlation (Cramér's V: 0.29) between access to care options and treatment suggests that improving access could increase treatment uptake.
2. **Weak Associations**:
   - **Days Indoors**: No strong correlation with mental health outcomes, suggesting that time indoors isn't a major factor.
   - **Mood Swings & Social Weakness**: Very weak correlation (Cramér's V: 0.06), indicating that social factors may not directly impact mood fluctuations.
3. **Care Options & Family History**: A slight positive correlation (Cramér's V: 0.15) between access to care options and Family History. While there is a slight tendency for individuals with a family history of mental health issues to have access to care options, the relationship is not very strong.

The analysis highlights that family history and care options play a major role in treatment outcomes. By focusing on early interventions, better follow-up care, and expanding access to services, we can improve mental health support for individuals in need.

Descriptive Statistics

	count	unique	top	freq
Timestamp	292364	580	8/27/2014 11:43	2384
Gender	292364	2	Male	239850
Country	292364	35	United States	171308
Occupation	292364	5	Housewife	66351
self_employed	287162	2	No	257994
family_history	292364	2	No	176832
treatment	292364	2	Yes	147606
Days_Indoors	292364	5	1-14 days	63548
Growing_Stress	292364	3	Maybe	99985
Changes_Habits	292364	3	Yes	109523
Mental_Health_History	292364	3	No	104018
Mood_Swings	292364	3	Medium	101064
Coping_Struggles	292364	2	No	154328
Work_Interest	292364	3	No	105843
Social_Weakness	292364	3	Maybe	103393
mental_health_interview	292364	3	No	232166
care_options	292364	3	No	118886

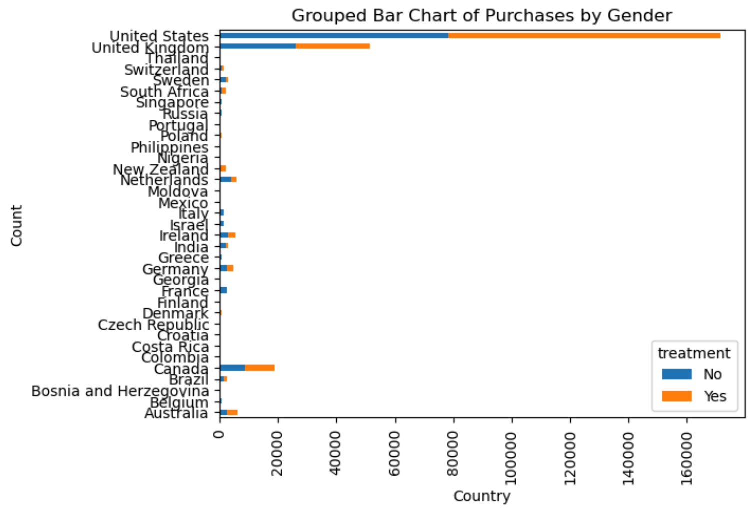
- The dataset is predominantly composed of male respondents from the United States, with a significant number indicating no family history of mental health issues.
- The majority of respondents did not undergo mental health interviews and lack access to care options.
- A number of participants report having received treatment, many indicate struggles with coping and changes in habits.
- The distribution of days spent indoors and stress levels suggests a variety of living conditions and mental health experiences among respondents.

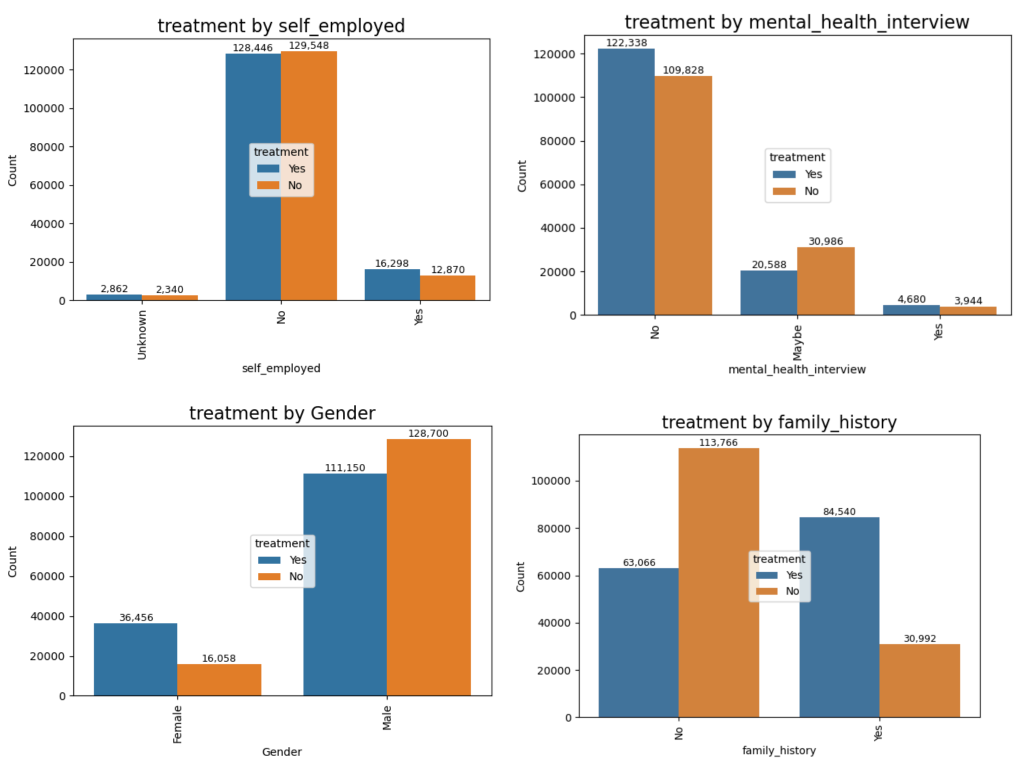
- Namely, the features **Gender**, **self_employed**, **family_history**, and **Country** show significant imbalance in the distribution of class labels. Let us explore this imbalance further.

Exploring Target and Feature Distribution Imbalance

Countries: The US has the highest population, with **171,308 entries**. This suggests that the data may trend and show issues specific only to the U.S. population, this may influence factors like healthcare access, cultural attitudes toward mental health, and social pressures.

While the dataset shows mental health trends from different countries, the unbalanced representation of the U.S. data could limit effective comparative analysis. Future studies might benefit from ensuring a more balanced representation to capture a wider range of mental health issues globally.





Gender: Males have a significantly higher trend of **No treatment** compared to **Females**, who show responsiveness towards seeking mental health care. This trend reflects social norms where men are not likely to seek help due to the stigma of vulnerability.

Females are underrepresented in the dataset.

Mental Health Interview Response: "No" has the highest count, suggesting that a large proportion of respondents did not undergo a mental health interview. However, all three categories show a balanced distribution of individuals who received treatment and those who did not. This suggests that treatment is adequately represented, regardless of whether the person had a mental health interview.

Self-employed: For those not self-employed, the treatment is almost equally divided between "Yes" (128,446) and "No" (129,548), showing a nearly even split between those who have received treatment and those who have not. Self-employed individuals are naturally less common than privately employed individuals, and this reflects reality. Over-sampling or artificially balancing this category might distort the representation of self-employed individuals in the analysis and potentially harm the model's generalizability.

Family History: Treatment distribution in both categories, the number of untreated individuals significantly outweighs the treated ones, but the **difference in proportions** between treated and untreated is higher in the "No Family History" group. The imbalance is notable in that the treatment rates differ across the family history groups, which could be an important factor to address in understanding mental health support and access disparities. The imbalance seems to represent a **natural occurrence** rather than a dataset-specific issue that needs balancing

Imbalances in the dataset can lead to biased models that overfit to the majority class, reducing generalization and accuracy. To address this, **SMOTENC (Synthetic Minority Over-sampling Technique for Nominal/Continues Features)** generates synthetic examples for the underrepresented class, creating a more balanced dataset. This improves the model's ability to recognize patterns across all classes, reducing overfitting and leading to more accurate, reliable predictions.

Based on the analysis, it appears that only the **Gender** and **Country** features require balancing due to dataset issues.

2.3. Baseline Modeling

The strategy for modeling involves both **global** and **per-country modeling** approaches:

- **Global Modeling** will be used to create a single model that captures overall patterns across all countries, providing a broad, high-level understanding of the data.
- **Per-Country Modeling** will be applied next, where individual models will be created for each country to capture specific trends and variations unique to each region.

This dual approach allows for general insights from the global model while enhancing prediction accuracy through more localized, country-specific models.

For the baseline model, we implemented

Logistic Regression - *A linear classifier; interpretable and quick, best for simple data. Limited to linear relationship*,

which serves as a standard approach for binary classification tasks. This initial performance set a benchmark against more complex models. Logistic Regression allowed us to gain insights into feature importance, aiding in the understanding of which factors contributed significantly to predicting mental health needs.

The model was evaluated using **precision** and **recall** metrics, which are crucial for understanding the balance between identifying individuals in need of support and minimizing false positives. Achieving optimal precision and recall scores in this base model will serve as a baseline for performance analysis, which can then inform **cost-benefit trade-off** decisions.

Model Evaluation and Validation Metrics

The following methods are applied to rigorously assess and validate the model's performance, ensuring accuracy and reliability across different models. The process includes **stratified cross-validation**, detailed **performance metrics**, and **visualization** techniques for a comprehensive understanding of model behavior.

1. **Cross-Validation:** A 15-fold Stratified K-Fold Cross-Validation approach is implemented to evaluate the model's consistency. By stratifying the data, each fold retains balanced class distributions, allowing a robust assessment of model performance on diverse

subsets. Averaging predictions across folds provides a generalized view, reducing potential overfitting and enhancing the model's predictive stability.

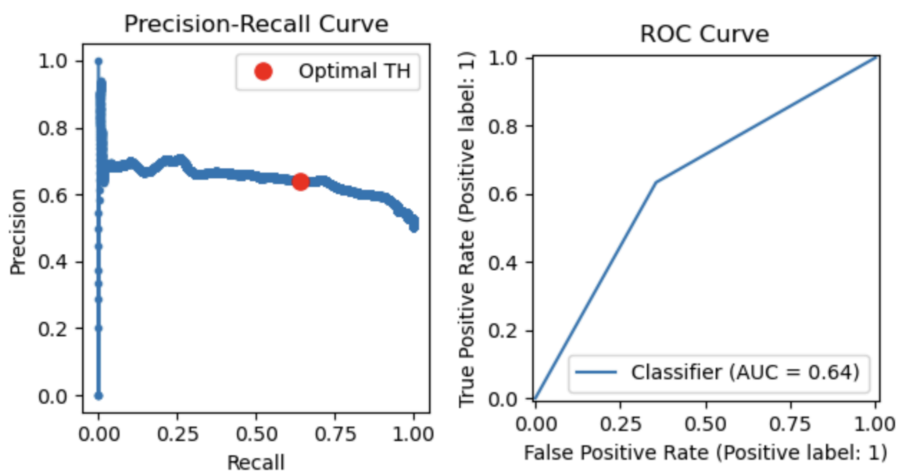
2. **Performance Metrics:**
    - **Confusion Matrix:** The confusion matrix summarizes the model's prediction accuracy across classes, showing correct and incorrect predictions. This matrix offers immediate insight into any class-specific imbalances or misclassification tendencies.
    - **Classification Report:** A comprehensive classification report includes **precision**, **recall**, and **F1-score** for each class. These metrics reveal how well the model distinguishes between different classes, highlighting any areas that may require further optimization.
    - **Cross-Validation Scores:** The model's performance across all 15 folds is displayed, with the **average score** providing an overarching measure of predictive accuracy and consistency. This helps verify the model's stability and resilience across different data samples.
3. **Visualizations:**
    - **Precision-Recall Curve:** This curve shows the trade-off between precision and recall for different thresholds, with an optimal threshold marked by a red dot for easy identification. It helps in understanding the model's ability to distinguish positive cases, especially in imbalanced datasets.
    - **ROC Curve:** By plotting the ROC curve and calculating the ROC-AUC score, the model's ability to discriminate between classes is evaluated. The area under the curve (AUC) provides a single-value metric of discrimination, helping to gauge model effectiveness in classification tasks.

This evaluation approach provides clear and actionable insights into the model's performance on various metrics, ensuring both individual metric clarity and cross-validation stability for a well-rounded analysis.

The evaluation strategy discussed here is applied to **LogisticRegression**, as well as the other **extended models**.

1. LogisticRegression Model Performance:

```
Confusion Matrix (Treatment Required):
 No: [8767 4822]
Yes: [4997 8642]

Classification Report:
              precision    recall  f1-score   support

           0       0.64      0.65      0.64     13589
           1       0.64      0.63      0.64     13639

    accuracy                           0.64     27228
   macro avg       0.64      0.64      0.64     27228
weighted avg       0.64      0.64      0.64     27228

Cross-V Mean Score: 0.6424
Cross-V Scores:
[0.63860304 0.6421497  0.64771528 0.64386013 0.63950801]
Opt. Precision: 0.6394101275766789
Opt. Recall: 0.6393785808726311
```



The Logistic Regression model achieved an accuracy of **64%**, with balanced **precision**, **recall**, and **F1**-scores of around **0.64** for both classes. The cross-validation scores were consistent, with a mean of **0.6424**. The model shows stable but moderate performance across different splits. In summary, the model shows moderate performance based on both the **Precision-Recall** and **ROC curve**, with an **AUC** of **0.64** indicating ***room for improvement***.

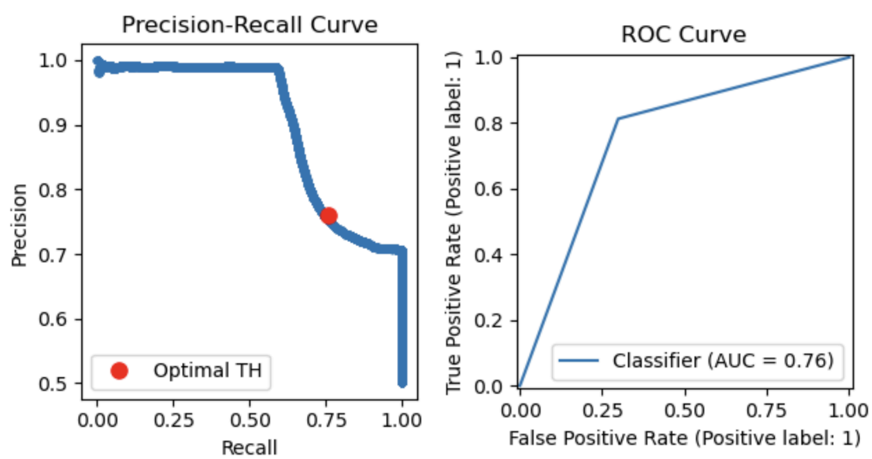
2.4. Extended Modeling

The motivation for the selected models stemmed from their ability to handle complex relationships and interactions within the data. The implementation of these models also included hyperparameter tuning, which further optimized their performance:

Random Forest: *Ensemble of trees; captures complexity, reduces overfitting, shows feature importance. Slower on large datasets; can be outperformed by boostin*g.
LightGBM: *A fast gradient boosting; highly efficient, accurate, great for large data. Sensitive to hyperparameters, risk of overfittin*g.
XGBoost: A *regularized boosting; fine-tuning options, excels in complex data accuracy. *High memory/computation cost, slower than LightGBM*,

2. Random Forest Model Performance:

```
Confusion Matrix (Treatment Required):
 No: [9534 4055]
Yes: [ 2555 11084]

Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.70      0.74     13589
           1       0.73      0.81      0.77     13639

    accuracy                           0.76     27228
   macro avg       0.76      0.76      0.76     27228
weighted avg       0.76      0.76      0.76     27228

Cross-V Mean Score: 0.7521
Cross-V Scores:
[0.75374451 0.75182655 0.75058203 0.75167323 0.75242946]
```



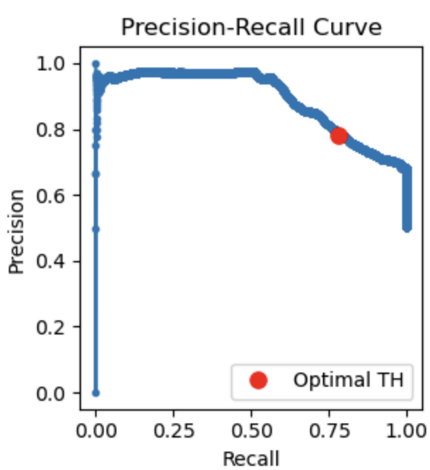
The **Precision-Recall** curve for the **Random Forest model** shows high precision at lower recall values, which decreases as recall increases, with the optimal threshold marked by a red dot. The **ROC curve** indicates a solid ability to distinguish between the positive and negative classes, with an **AUC** of **0.76**, reflecting **moderate** model performance.

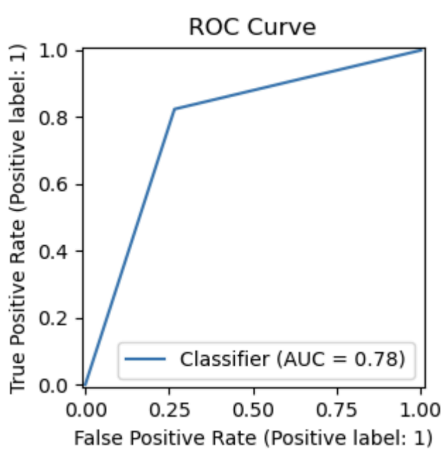
In terms of metrics, the Random Forest model achieved an **accuracy** of **76%**, with **precision** values of 0.79 for class 0 and 0.73 for class 1, and recall values of 0.70 for class 0 and 0.81 for class 1. The **F1**-scores were 0.74 for class 0 and 0.77 for class 1. The model's cross-validation mean score was **0.7521**, demonstrating consistent performance across different splits.

Overall, the Random Forest model outperforms the Logistic Regression model, showing **better** precision, recall, F1-score, and AUC.

3. LightGBM Model Performance:

```
Confusion Matrix (Treatment Required):
 No: [9978 3611]
Yes: [ 2402 11237]

Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.73      0.77     13589
           1       0.76      0.82      0.79     13639

    accuracy                           0.78     27228
   macro avg       0.78      0.78      0.78     27228
weighted avg       0.78      0.78      0.78     27228

Cross-V Mean Score: 0.7843
Cross-V Scores:
[0.78273804 0.77788299 0.78547137 0.78912751 0.78648561]
```





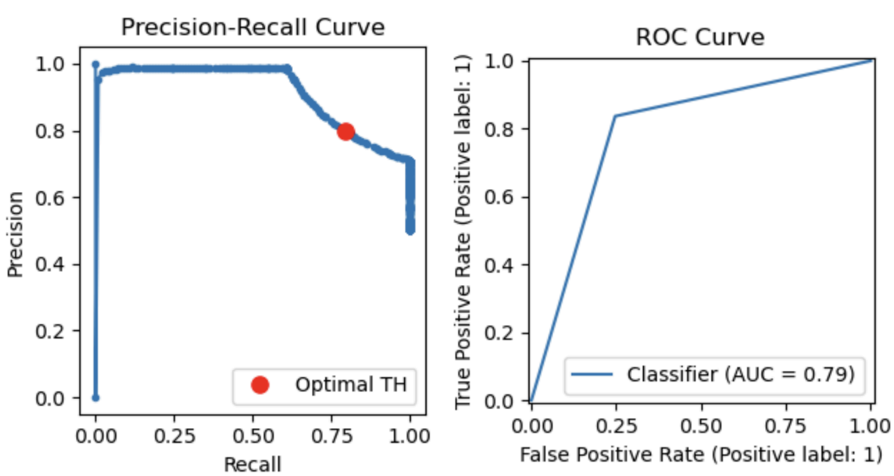
The **Precision-Recall** Curve for **LightGBM** shows a high precision at lower recall values and decreases as recall increases. The optimal threshold is marked by the red dot.

The model achieved an **accuracy** of **78%**, with **precision** of 0.81 for class 0 and 0.76 for class 1, and recall of 0.73 for class 0 and 0.82 for class 1.

The **AUC** is **0.78**, showing improved performance compared to Logistic Regression (AUC = 0.64) and Random Forest (AUC = 0.76). Overall, LightGBM outperforms the previous models in accuracy, precision, recall, F1-score, and AUC.

4. XGBoost Model Performance:

```
Confusion Matrix (Treatment Required):
 No: [10235  3354]
Yes: [ 2227 11412]

Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.75      0.79     13589
           1       0.77      0.84      0.80     13639

    accuracy                           0.80     27228
   macro avg       0.80      0.79      0.79     27228
weighted avg       0.80      0.80      0.79     27228

Cross-V Mean Score: 0.7947
Cross-V Scores:
[0.79643566 0.79477163 0.79194984 0.79710357 0.79304825]
```



The **XGBoost** model achieved an accuracy of **80%**, with **precision** values of 0.82 for class 0 and 0.77 for class 1. The **recall** was 0.75 for class 0 and 0.84 for class 1, with F1-scores of 0.79 and 0.80, respectively.

The model demonstrated strong performance with an **AUC** of **0.79**, indicating a good ability to distinguish between classes. **Cross-validation** showed a mean score of **0.7947**, with individual scores ranging from 0.7919 to 0.7971.

When compared to previous models, XGBoost outperformed Logistic Regression, Random Forest, and LightGBM in terms of accuracy (80% vs. 64%, 76%, and 78%, respectively) and precision.

The F1-scores and AUC were also higher than those of the other models, highlighting XGBoost's superior performance in classification tasks. The AUC of 0.79 is slightly better than LightGBM (0.78) and significantly higher than Logistic Regression (0.64) and Random Forest (0.76).

3. Findings

3.1 Model Performance Comparison

Model	Precision	Recall	F1-Score	AUC	Accuracy
Logistic Regression	0.64	0.63	0.64	0.64	0.64
Random Forest	0.73	0.81	0.77	0.76	0.76
LightGBM	0.76	0.82	0.79	0.78	0.78
XGBoost	0.77	0.84	0.8	0.79	0.8

The **Model Comparison** table summarizes the performance of the four models: **Logistic Regression**, **Random Forest**, **LightGBM**, and **XGBoost**, based on key metrics including **Precision**, **Recall**, **F1-Score**, **AUC**, and **Accuracy**.

- **XGBoost** emerges as the best-performing model with the highest **precision** (0.77), **recall** (0.84), **F1-score** (0.80), **AUC** (0.79), and **accuracy** (0.80).
- **LightGBM** follows closely, with strong performance in **precision** (0.76) and **recall** (0.82), achieving **0.78** in both **AUC** and **accuracy**.
- **Random Forest** also performs well, showing the highest **recall** (0.81) and a solid **F1-score** of 0.77, but lags slightly in **precision** (0.73) and **AUC** (0.76).
- **Logistic Regression**, while providing a reasonable performance, trails behind in all metrics with **precision** and **recall** both at 0.64, resulting in lower **F1-score** and **accuracy** (0.64).

Overall, **XGBoost** stands out as the most robust model, offering the best overall balance of precision, recall, and accuracy, followed closely by **LightGBM**. Both models show better potential for deployment in real-world applications.

3.1. Threshold-tradeoff Analysis

The threshold trade-off analysis is crucial for aligning the model's performance with the objective of providing timely and efficient mental health support. A well-chosen threshold ensures that we meet the need for intervention while optimizing resource allocation.

Threshold	TP	FP	FN	Precision	Recall	Cost
0.1179	13639	6329	0	0.6830	1.0000	6329
0.1455	13639	5906	0	0.6978	1.0000	5906
0.1850	13639	5735	0	0.7040	1.0000	5735
0.2632	13639	5674	0	0.7062	1.0000	5674
0.3337	13639	5615	0	0.7084	1.0000	5615
0.4371	13639	5569	0	0.7101	1.0000	5569
0.5026	11053	2979	2586	0.7877	0.8104	8151
0.6054	8573	407	5066	0.9547	0.6286	10539
0.8495	7568	106	6071	0.9862	0.5549	12248
0.9316	4720	61	8919	0.9872	0.3461	17899
0.9610	1222	16	12417	0.9871	0.0896	24850

The table presents performance metrics evaluated at various decision thresholds. The definitions of each column are as follows:

- **Threshold**: The decision threshold used to classify instances as positive.
- **TP (True Positives)**: The number of correctly predicted positive instances.
- **FP (False Positives)**: The number of incorrectly predicted positive instances, where negative instances are predicted as positive.
- **FN (False Negatives)**: The number of incorrectly predicted negative instances, where positive instances are predicted as negative.
- **Precision**: The proportion of true positives (TP) out of all predicted positives (TP + FP), indicating the accuracy of positive predictions.
- **Recall**: The proportion of true positives (TP) out of all actual positives (TP + FN), indicating how many of the actual positives were correctly identified.
- **Cost**: A measure of the penalty for misclassifications, calculated using the formula `cost = fp + 2 * fn`. This formula assigns a higher weight to false negatives (multiplied by 2) to reflect that missing a positive case (e.g., failing to identify someone in need of support) is considered more costly than a false positive. We use the cost function to prioritize minimizing false negatives in **mental health prediction**, as missing someone

in need of support can have **severe consequences**. The formula for cost may change depending on the specific requirements, as the cost of misclassifications can vary based on context.

Key Observations:

- **Precision and Recall Trade-off**: Higher thresholds increase precision (0.6830 to 0.9871) but decrease recall. Lower thresholds boost recall (1.0000) but reduce precision.
- **Cost**: Cost decreases as the threshold rises (6329 to 24850), correlating with fewer false positives. However, at very high thresholds (above 0.5026), cost increases due to more false negatives.
- **False Positives vs. False Negatives**: Lower thresholds result in more false positives but no false negatives, while higher thresholds reduce false positives but increase false negatives.
- **Optimal Threshold**: A threshold around 0.5026 balances precision (0.7877) and recall (0.8104) with moderate cost (8151).

Summary: **Low thresholds** have **high recall** but **low precision** and **higher costs**; **high thresholds** favor **precision** but **increase false negatives** and **costs**. A **threshold** of **0.5026** offers a good **balance**, with the choice of threshold depending on the application's priority (e.g., favoring recall for critical tasks like medical testing).

3.3 Country Models

The country model in mental health modeling can be useful for several reasons:

1. **Cultural and Societal Differences**: Mental health issues and their recognition or treatment can vary significantly across countries. A country-specific model can account for these differences in diagnosis, treatment availability, stigma, and social support, which may impact the outcomes.
2. **Data Distribution Variance**: The distribution of features and mental health conditions may differ from one country to another. A country-specific model can help capture regional nuances in the data, leading to more accurate predictions.
3. **Policy and Healthcare System Influence**: Mental health services, policies, and healthcare accessibility vary by country. These factors can affect the likelihood of individuals seeking help or reporting mental health conditions, which a country-specific model can better incorporate.
4. **Segmentation**: If the global model performs suboptimally in certain countries or regions due to these factors, a country-specific model allows for improved segmentation and targeted predictions for those specific regions.

The country model can improve the overall performance by addressing regional variances and ensuring that the model is contextually relevant to each country's unique challenges and healthcare systems.

This section details the model performance evaluation for each country using the dataset that excludes countries with single-class targets, meaning we have 13 countries for modeling in this test.

Evaluating Performance using XGBoost

The evaluation process for XGBoost with country-specific data included the following steps:

- Create country-specific subsets using `df_country = df[df['Country'] == country_name]`.
- Pre-process the data (transform, feature-engineer).
- Split the data into train/test sets.
- Run predictions and compute metrics:
    - Accuracy
    - AUC
    - Cross-validation mean and standard deviation (weighted F1, cv=5).

XGBoost was selected for its strong performance in this classification task. Cross-validation with **StratifiedKFold** preserved class distribution, and **cross_val_predict** ensured robust evaluation. Metrics like precision, recall, accuracy, and F1 score were used to assess the model's generalization.

Evaluation Result:

Country	Threshold	Precision	Recall	Accuracy	F1
United States	0.473	0.691	0.772	0.681	0.729
Australia	0.477	1.000	0.861	0.919	0.925
Canada	0.496	0.700	0.884	0.756	0.781
United Kingdom	0.490	0.705	0.978	0.733	0.820
South Africa	0.997	1.000	1.000	1.000	1.000
Sweden	0.493	0.710	0.778	0.817	0.743
New Zealand	0.998	1.000	1.000	1.000	1.000
Netherlands	0.987	1.000	1.000	1.000	1.000
India	0.972	1.000	1.000	1.000	1.000
Ireland	0.481	0.937	0.794	0.899	0.860
Brazil	0.484	0.844	0.581	0.824	0.688
Germany	0.983	1.000	1.000	1.000	1.000
Switzerland	0.491	0.718	0.785	0.738	0.750

Key Observations:

1. High Performance:
   - **South Africa**, **New Zealand**, **Netherlands**, **India**, **Germany** have perfect scores (1.000), indicating excellent model performance.
2. Strong but Balanced:
   - **Australia** (precision: 1.000, recall: 0.861), **Ireland** (precision: 0.937, recall: 0.794) show high precision but varying recall.
3. Moderate Performance:
   - **Canada**, **United Kingdom**, **Sweden** have balanced but moderate precision and recall (around 0.7-0.8).
4. Room for Improvement:
   - **United States** and **Brazil** have lower precision (0.691 and 0.844), suggesting missed positives or false negatives.

Other Considerations:

5. **One-Class Issue**: Several countries, including Poland, Belgium, and France, are with only one class being present in the dataset. This prevents meaningful modeling. This indicates a potential data imbalance, or a lack of diverse cases in those regions.
6. **Requires Further Analysis**: Countries with perfect scores needs a closer examination of the dataset to understand if the models are overly fitted or if the data truly represents a distinct population without variability.
7. **Next Steps**: Consider enhancing datasets for countries facing the one-class issue to include more diverse cases.

4. Conclusions and Future Work

In conclusion, the project successfully identified key predictors of mental health support needs, utilizing various data science techniques. The advanced models demonstrated superior performance compared to the baseline, offering valuable insights for stakeholders.

The findings highlight patterns, such as the **correlation** between **family history**, **gender**, and the likelihood of seeking mental health treatment.

Feature importance analysis revealed that the most **significant** factors influencing the prediction of mental health needs included **access** to **care option**s, **family history of mental health issues**, and **gender**. Specifically, respondents with access to care options were more **likely** to have a **higher prediction** score, while being **male** and having **no family history** of mental health issues had less positive impact on the model's predictions. This insight is important as it highlights the factors that could be targeted to improve mental health interventions.

Although XGBoost achieves the highest performance with an accuracy of 80%, it may not yet be fully optimized for a critical application like mental health prediction.

While it shows promising results, further improvements in precision, recall, and F1-score are needed to ensure more reliable predictions. Consider enhancing the model with feature engineering, model tuning, and exploring ensemble methods or neural networks.

Since **XGBoost** is one of several models being compared, optimizing its performance alongside the others will help create a stronger, more robust solution before deploying it in production.

Future work could explore:

- Incorporating additional datasets, such as social media sentiment analysis, to enhance model performance.
- Future research should prioritize obtaining a more balanced dataset that adequately represents various demographic groups across different countries.
- Testing more sophisticated modeling techniques, such as deep learning algorithms.
- Expanding the feature set to include temporal factors, which may influence mental health outcomes.

5. Recommendations

- Based on the **Threshold-Trade Off** insights, it is recommended to adopt the **0.5026 threshold** for the global model development, as it achieves an **optimal** balance between precision and recall. This approach **minimizes** the risk of overlooking true positives while ensuring that all predicted positive cases are indeed **correct**.
- Implement outreach programs for high-risk demographic groups identified in the modeling process.
- Utilize the predictive model to allocate mental health resources dynamically, ensuring timely support for those in need.
- Regularly update the model with new data to enhance its accuracy and relevance over time.
- To improve proactive mental health support, the development of real-time monitoring systems utilizing predictive models could be beneficial. These systems can help identify at-risk populations promptly, facilitating timely interventions.

6. Consulted Resources

- https://ourworldindata.org/mental-health
- Python libraries: pandas, scikit-learn's RandomForest, LogisticRegression, scipy.stats, matplotlib, xgboost's XGBoostClassifier, lightgbm's LightGBMClassifier, seaborn, matplotlib, bayesopt