Springboard-DSC Program
Capstone Project 2

Predicting Mental Health Support Needs

By Jude M. Santos

October, 2024

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 as a combination of online surveys and publicly available mental health datasets.
After initial acquisition, the data underwent several wrangling processes, including:

- Removal of irrelevant features, such as the 'Timestamp' column.

- The creation of additional features from ‘Timestamp’ components: year, month, day, hour, and minute.

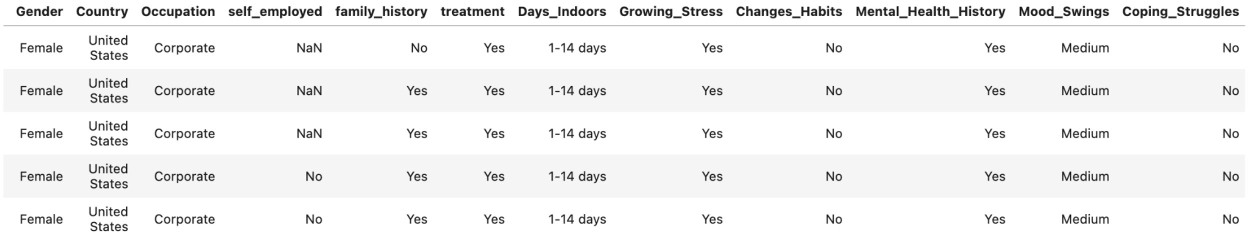
- Imputation of missing values in categorical variables with 'Unknown' and conversion of binary categorical columns to numerical values.

- Transformation of categorical features into one-hot encoded variables, ensuring that they were appropriately formatted for model training.

2.2. Storytelling and Inferential Statistics

The initial exploration of the dataset revealed key demographic patterns, such as variations in mental health conditions across different countries. Using visualizations, we illustrated these trends, highlighting the correlation between certain demographic factors and mental health outcomes. For instance, bar plots were used to display the prevalence of mental health issues by gender and occupation. Inferential statistics were applied to assess the significance of these relationships, providing a foundation for further analysis.

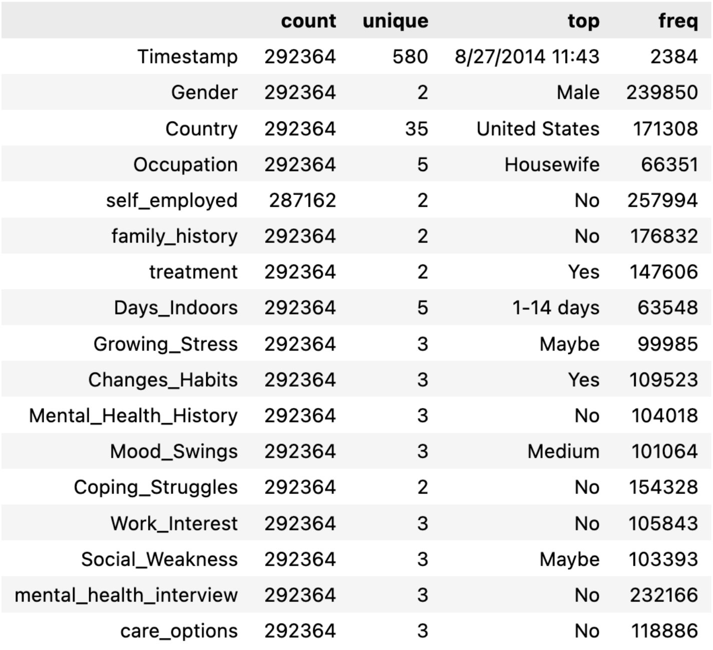
Sample Data:



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 consists of **292,364** observations and includes features that provide insights into the demographics and mental health conditions of individuals.

Descriptive Statistics:

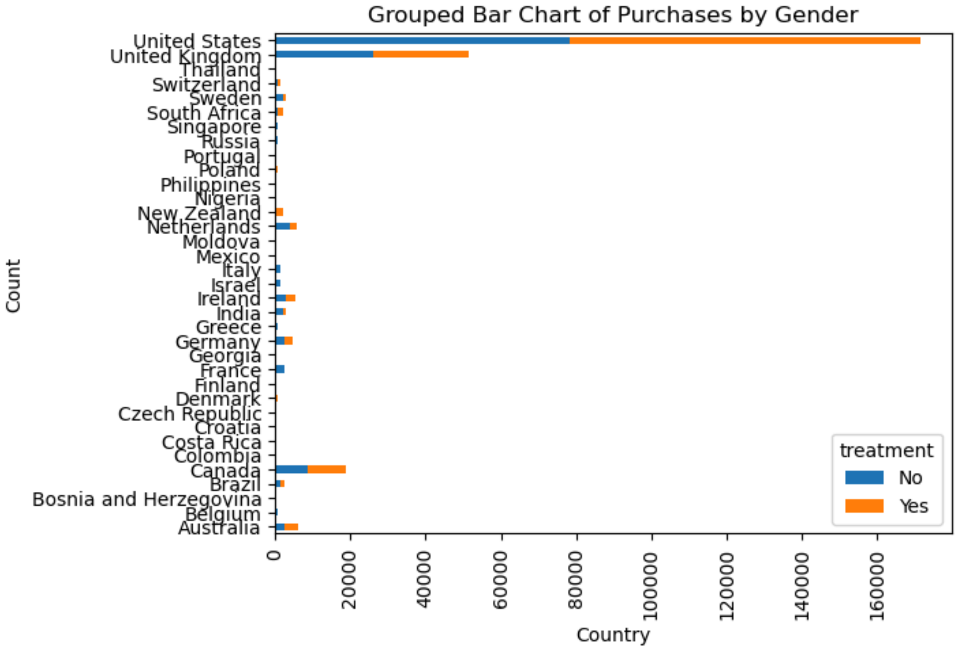


	count	unique	top	freq
Timestamp	292364	580	8/27/2014 11:43	2384
Gender	292364	2	Male	239850
Country	292364	35	United States	171308
Occupation	292364	5	Housewife	66351
self_employed	287162	2	No	257994
family_history	292364	2	No	176832
treatment	292364	2	Yes	147606
Days_Indoors	292364	5	1-14 days	63548
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.

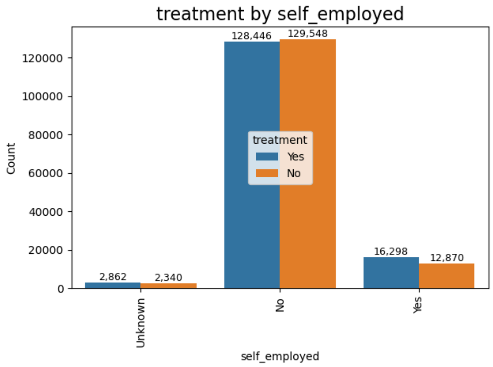
- The majority of respondents did not undergo mental health interviews and lack access to care options.
- A number of participants report having received treatment, many indicate struggles with coping and changes in habits.
- The distribution of days spent indoors and stress levels suggests a variety of living conditions and mental health experiences among respondents.

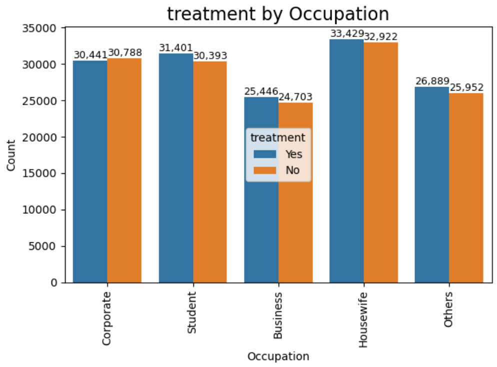
Visualizations:



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.

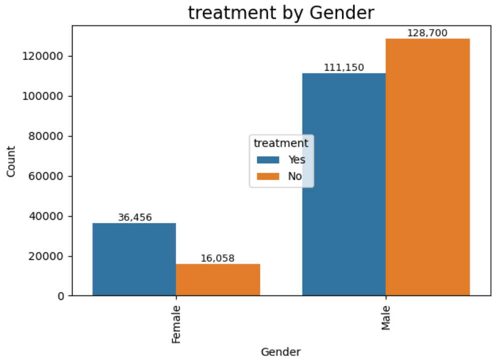
While the dataset shows mental health trends from different countries, the unbalanced representation of the U.S. data could limit effective comparative analysis. Future studies might benefit from ensuring a more balanced representation to capture a wider range of mental health issues globally.





Housewives show a higher prevalence of treatment entries, which suggests a trend in mental health issues in domestic responsibilities and the stress associated with them - these unique challenges may cause the likelihood of seeking treatment.

Other Occupations: The distribution of treatment entries with other occupations, while less prominent than housewives, still reflect varying levels of activities with mental health services.



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.

2.3. Baseline Modeling

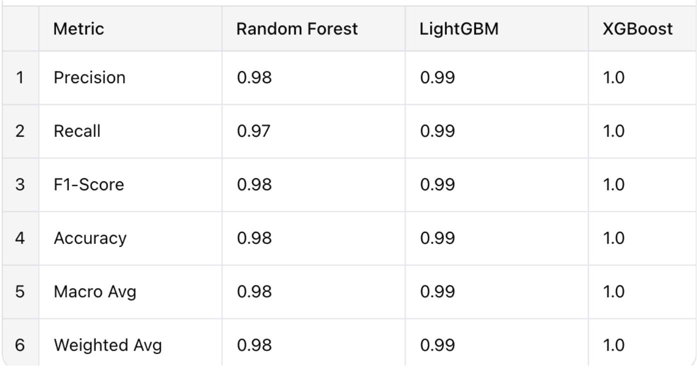
For baseline modeling, we implemented **Logistic Regression**, which serves as a standard approach for binary classification tasks. This initial performance set a benchmark against more complex models. Logistic Regression allowed us to gain insights into feature importance, aiding in the understanding 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 cost trade-off analysis.

2.4. Extended Modeling

To enhance prediction performance and accuracy, we explored more advanced models, including **XGBoost**, **LightGBM**, and **Random Forest**. The motivation for these 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.

3. Findings



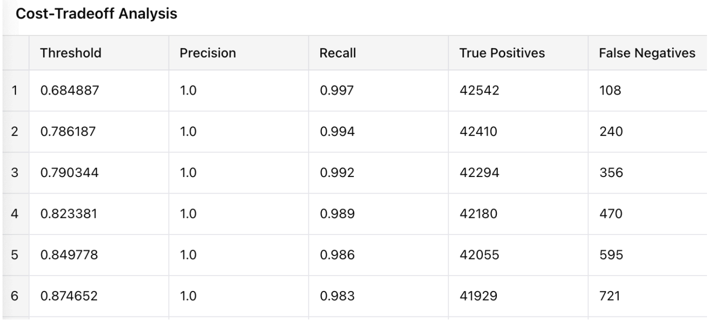
	Metric	Random Forest	LightGBM	XGBoost
1	Precision	0.98	0.99	1.0
2	Recall	0.97	0.99	1.0
3	F1-Score	0.98	0.99	1.0
4	Accuracy	0.98	0.99	1.0
5	Macro Avg	0.98	0.99	1.0
6	Weighted Avg	0.98	0.99	1.0

XGBoost achieved the highest performance across all metrics, boasting scores of 1.00 in precision, recall, and F1-score for class 1 predictions. This indicates that it perfectly predicted both classes without any false positives or negatives.

The results indicate that both **LightGBM** and **Random Forest** can maintain performance, though **XGBoost** is superior in this scenario. This could suggest that XGBoost is more robust against potential data variations and outliers in this particular dataset.

For tasks requiring **optimal accuracy**, **XGBoost** should be prioritized. However, if model **training speed and interpretability** are significant factors, **LightGBM** may also be a strong contender due to its efficiency and comparable performance.

3.1. Cost-tradeoff Analysis



Cost-Tradeoff Analysis

	Threshold	Precision	Recall	True Positives	False Negatives
1	0.684887	1.0	0.997	42542	108
2	0.786187	1.0	0.994	42410	240
3	0.790344	1.0	0.992	42294	356
4	0.823381	1.0	0.989	42180	470
5	0.849778	1.0	0.986	42055	595
6	0.874652	1.0	0.983	41929	721

A lower threshold yields a **higher recall** 0.997 at a threshold of 0.684, but increases the number of **false negatives** as the threshold increases.

If **false negatives** are **costly -** like missing a disease diagnosis, maintaining a lower threshold might be necessary. However, if the **cost** of **false positives** is **higher** - like unnecessary treatments, a **higher threshold** might be preferred.

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 **gender**, **occupation**, 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.

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 **cost-tradeoff** insights, it is recommended to adopt the **0.849778 threshold** for model deployment, 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