Springboard-DSC Program Capstone Project 2

Predicting Mental Health Support Needs

By Jude M. Santos

October, 2024

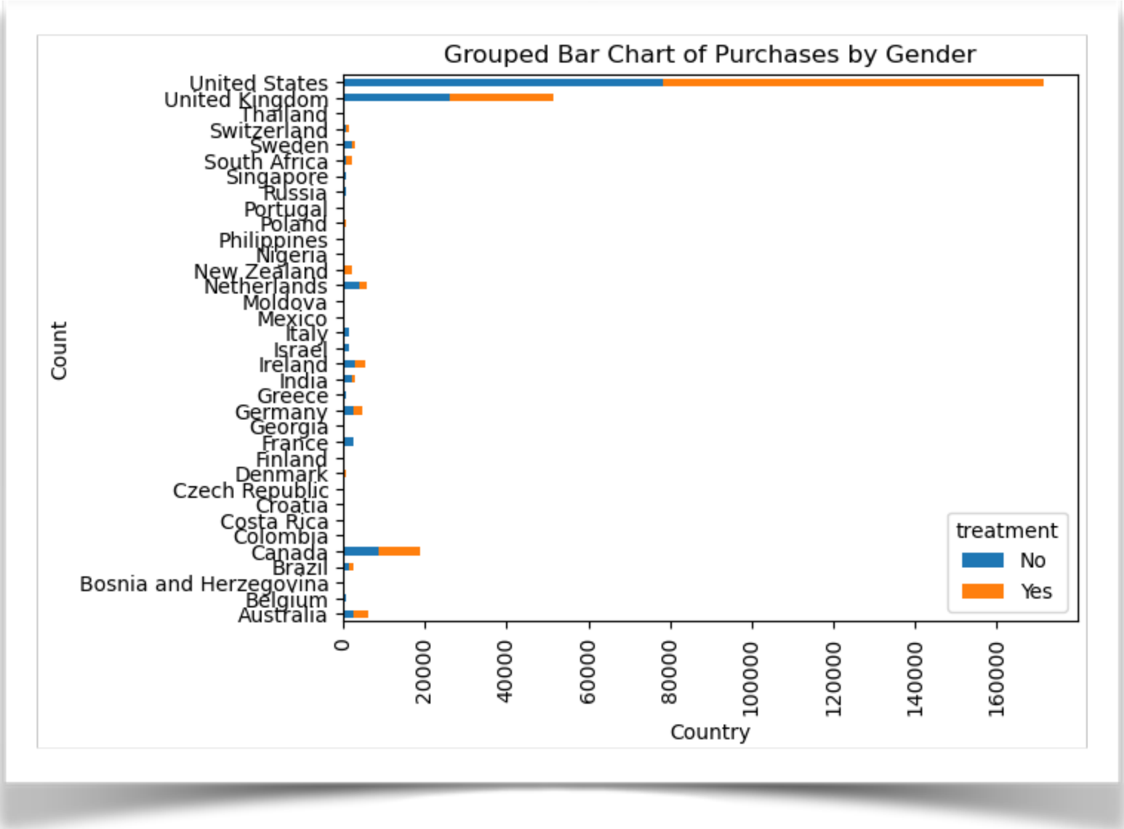
Introduction

This project aims to identify groups likely to require mental health support by analyzing demographic data, mental health conditions, sentiment analysis, and psychological indicators. Stakeholders include mental health organizations and policymakers who can use these findings for effective resource allocation.

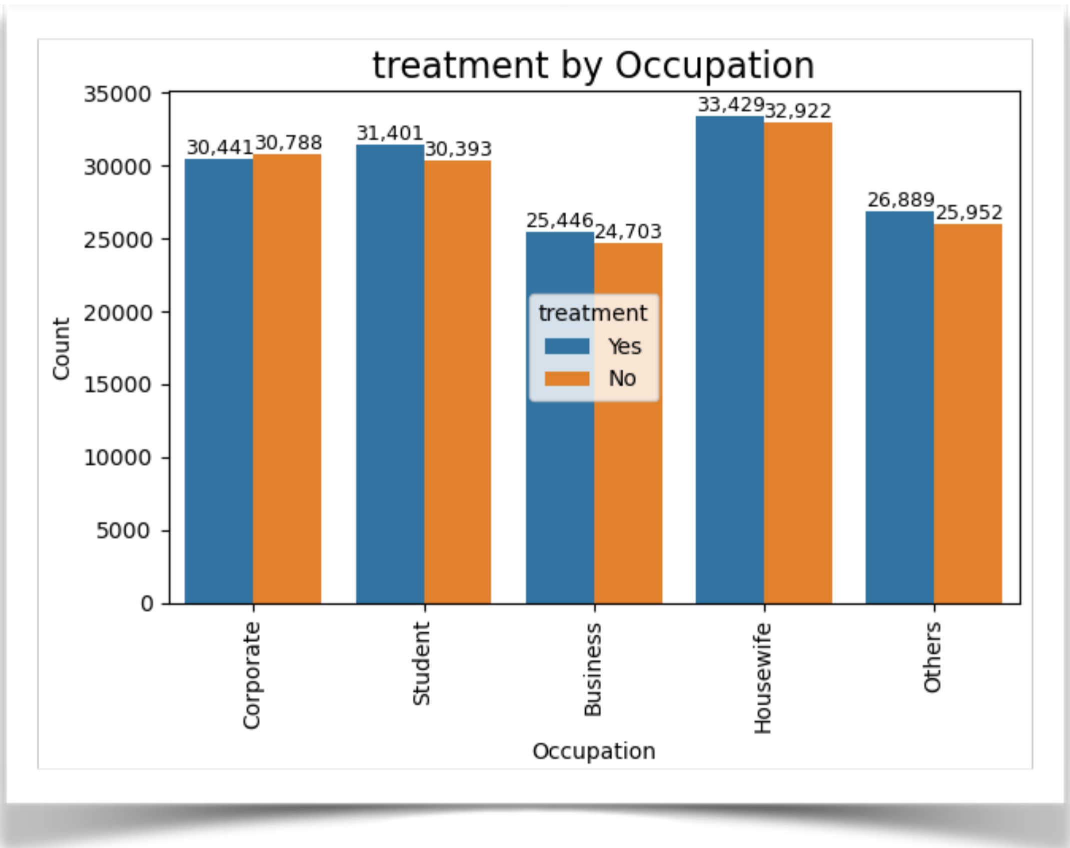
Approach

- *Data Source: Dataset sourced from Kaggle, combining online surveys and publicly available data.*
- *Data Wrangling: Removed irrelevant features, created new features from 'Timestamp', and handled missing values with appropriate imputation methods.*
- *Exploratory Analysis: Visualizations illustrated trends and key demographic patterns related to mental health.*
- *Modeling: Baseline modeling with* **Logistic Regression***, followed by advanced modeling using* **XGBoost, Random Forest** *and* **LightGBM***.*

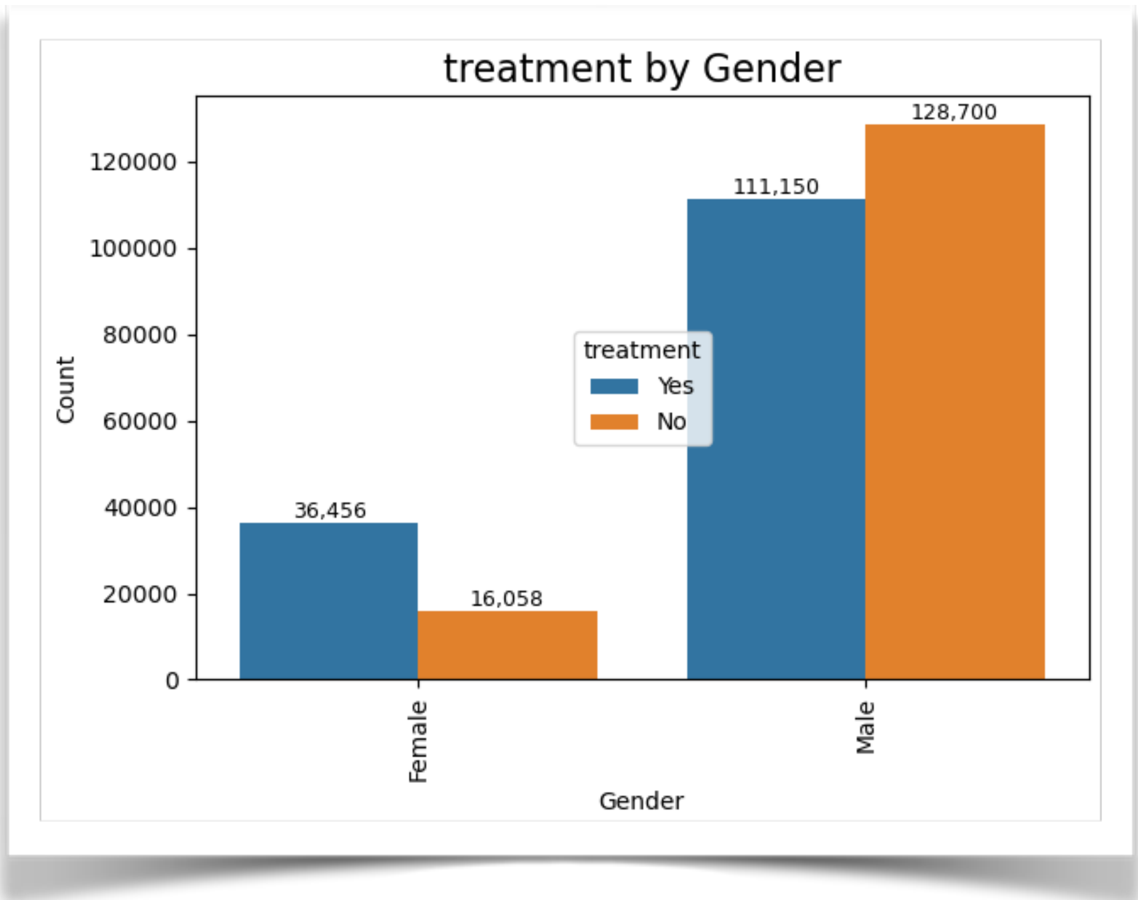
Data Analysis and Visualization



The U.S. had the highest representation, affecting insights on healthcare access and cultural attitudes.



Higher treatment entries among housewives, indicating unique stressors in domestic roles.



Gender disparities: Males less likely to seek treatment compared to females, highlighting social stigma.

Key Findings

- *XGBoost achieved highest scores in precision, recall, and* **F1** *for class 1 predictions, indicating its effectiveness.*
- *LightGBM and Random Forest maintained strong performance but fell short of XGBoost.*
- *Important predictors included* **access to care**, **family history** *of mental health issues, and* **demographic** *factors.*

Model Performance

	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 highest scores in precision, recall, and* ***F1*** *for class 1 predictions, indicating its effectiveness.*
- *LightGBM and Random Forest maintained strong performance but fell short of XGBoost.*

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 higher recall (**0.997** at threshold **0.684**) but increases false negatives.*
- *If false negatives are costly (e.g., missed diagnoses), a lower threshold is preferable.*
- *If false positives are more costly (e.g., unnecessary treatments), a higher threshold may be favored.*

Conclusions & Future Work

- *The project successfully identified key predictors of mental health support needs, utilizing various data science techniques. Future work could include:*
- *Incorporating additional datasets for deeper insights.*
- *Testing sophisticated modeling techniques like deep learning.*
- *Expanding the feature set to include temporal factors influencing mental health.*

Recommendations

- *Adopt a threshold of* **0.849** *for optimal model deployment, balancing precision and recall.*
- *Implement outreach programs for high-risk demographic groups.*
- *Utilize predictive models for dynamic allocation of mental health resources.*
- *Regularly update the model with new data and enhance it with real-time monitoring systems.*