Predictors of Global Mental Illness Rates

G4-MentalMentats

Data Science Capstone Project
Data Acquisition and Pre-Processing Report

6/12/2024

Team Members

Name: Kelsey Fox

Name: Greg Savage

Name: Tommy Pennington

Table of Contents

Introduction

This report summarizes our efforts to use machine learning models to forecast the prevalence rates for depressive disorders for countries using past socioeconomic and mental health data found in online research databases. We selected this topic because mental health conditions are a growing economic problem that many nations are slow to support with government funding or programs.

Because we were interested in investigating global patterns over time, we focused on datasets that covered a range of years for multiple countries. These datasets and our reasons for selecting each are summarized below.

About the Datasets

To measure a country's socioeconomic status, we found GDP per capita from 1990 to 2019 by DataBank. The per capita metric was important for our work since it accounts for population size and represents how a country's wealth is distributed amongst its population. For historical mental health data, we found prevalence rates from the 2019 Global Burden of Disease Study by the Institute for Health Metrics and Evaluation (IHME) from 1990 to 2019 for males and females of ages 5 to 100+ for more than 200 countries. This source had information that covered a range of disorders, but we focused on five disorders specifically: depressive, anxiety, eating, schizophrenia, and bipolar, and selected the prevalence rate, which measured the number of people diagnosed with the condition per 100,000 persons of that country. Last, to measure government interest or support in mental health, we used the World Health Organization's governance dataset, which recorded information about how much a country's government supported mental health care (e.g., do they have a state policy; is that policy still active; what year was the policy adopted). We also found survey data that captured an individual's perception of mental health care and disorders (as well as if the survey respondent identified as having any disorders), but we had to exclude it from our model since the results were only for one year and not a range of time.

Acquisition and Preprocessing Summary

We consolidated our sources to a single dataset by merging with 'country' and 'year' in stages (e.g., first merging GDP per capita to the prevalence data, then merging that result to the governance dataset, etc.). We examined null value counts as well as distributions before and after merging to validate each merge was done correctly. This merging is summarized in Table 1, shown right.

The acquisition and preprocessing steps are detailed in separate notebooks, see Appendix Table 3, and available as supplemental materials.

Table 1. *Dataset features and merging summary. These three datasets were merged to make the 'predicting_prevalence' dataset.*

dataset name	features
prevalence	country year sex age prevalence_rate_[disorder] share_[disorder]
gdp	country year gdp
governance	country year has_mental_health_legislation has_mental_health_policy
predicting_prevalence	country year sex age prevalence_rate_depressive_disorder gdp has_mental_health_legislation has_mental_health_policy

Exploratory Data Analysis

The exploratory data analysis (EDA) established that the two most prevalent mental disorders were anxiety and depressive disorders. Of these two, we selected depressive disorders as the focus for our modeling work. (Note: when establishing which disorders were most prevalent, we aggregated the prevalence rates using the median due to its large range.) For context, the depressive disorder prevalence rates range from 201.93 to 14,611.92 (Poland in 2016 and Uganda in 2014, respectively). This skew is significant as it will affect our modeling efforts, which assume the data have normal distributions. This brings us to the most important finding from our EDA work: skewed distributions.

EDA also revealed many of our features (columns) had skewed distributions, beyond just the prevalence rates. To combat the skewed distribution, we transformed the data with a logarithmic function. By taking the logarithm of the values, we compressed the range between larger and smaller values, effectively making the distribution more symmetric.

Comparing Depressive Disorder Prevalence Rate Distributions: Before and After Log Transformations

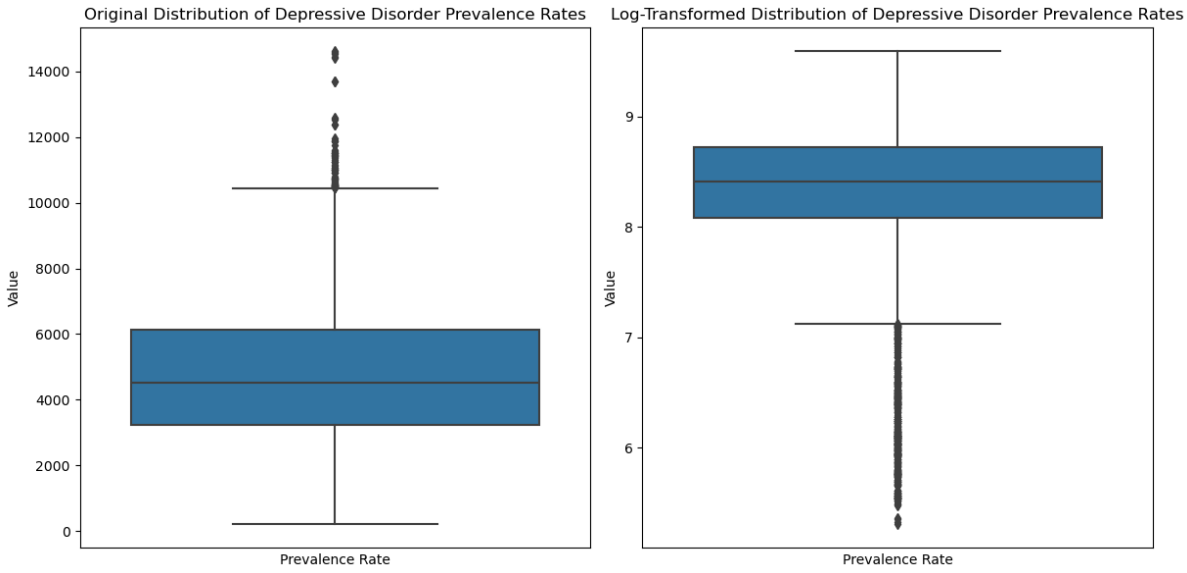

Figure 1. A comparison of the distribution for depressive disorder prevalence rates before (right) and after (left) log transformation.

As Figure 1 shows, the original distribution (left-most figure) is right-skewed, indicating that a few high-value outliers are significantly affecting the distribution. These outliers are shown in the figure's left boxplot (the points that extend beyond the whisker). The median (the line inside the box) is closer to the lower quartile, which tells us that most of the data points are concentrated toward the lower values.

Looking at the figure's right boxplot, we see the effect of the log transformation. The distribution now appears more symmetrical and less skewed. While the median (line inside the box) is more centrally located, suggesting a more balanced distribution of data points around the center, the number of outliers has increased, shown by the greater number of points extending below the whisker. While there are more outliers after the log transformation (78 vs 541), they are not as extreme. For the purposes of this project, we will use the log-transformed data for our modeling.

Predictive Modeling

Implementation

We implemented five regressor models for a range of complexity: linear regression, decision tree, gradient boosting, random forest, and support vector machine (SVM). The least complex was the linear regression. To capture non-linear relationships between the features and target variable, we used the decision tree, gradient boosting, random forest, and SVM models. We expected the random forest and SVM models to be the best options for our data, as they are both known to work with high-dimensional datasets and are robust models that reduce overfitting.

We first tested how the linear regression model performed on data before the log transformation and then compared its performance after applying the log transformation. It greatly improved the results, so we used the transformed data for all other models. (Detailed in notebook 'D1.predictive_modeling.ipynb'). We also performed a 50/25/25 split on the data to form the train, validation, and test sets. (The intent is for the validation data to be used for hyperparameter tuning to avoid model bias but we did not complete hyperparameter tuning).

The performance metrics for all models are summarized in Table 2. We used the mean squared error (MSE) and coefficient of determination (R^2) to evaluate the test, train, and validation datasets. The coefficient of determination measures how well a model performs relative to a simple mean of the target values, where $R^2 = 1$ means a perfect match and $R^2 = 0$ means the model does not perform better than taking the mean of the data, and a negative value means it performed worse than that. Mean squared error (MSE) is the average of the squared differences between the actual and predicted values. The closer this score is to 0, the more accurate the prediction. The cross-validation results are measured with the root mean squared error (RMSE) mean and standard deviation, which are used to measure how consistent the models perform.

Table 2. *Model results summary table to evaluate model performance using log-transformed 'predicting_prevlaence' dataset. Metrics include the mean squared error (MSE) and coefficient of determination (R^2) for test, train, and validation data and the cross validation RMSE mean and standard deviation. Bold indicates best score for each metric, excluding the training results (due to overfitting by the Decision Tree model). Random forest had the best overall performance.*

Metric	Linear Regression	Decision Tree	Gradient Boosting	Random Forest	SVM
Train MSE	0.01902	0.00000	0.02766	0.00232	0.01655
Train R^2	0.95846	1.00000	0.93961	0.99492	0.96387
Validation MSE	0.02081	0.02725	0.03185	**0.01415**	0.02004
Validation R^2	0.95808	0.94511	0.93584	**0.97150**	0.95963
Test MSE	0.02127	0.02918	0.03190	**0.01553**	0.02042
Test R^2	0.95134	0.93324	0.92701	**0.96448**	0.95328
Cross Validation RMSE: Mean	0.14439	0.17648	0.17832	**0.13582**	0.14392
Cross Validation RMSE: Std. Dev.	0.00625	0.01193	0.00762	0.00820	**0.00357**

Conclusion

The random forest model had the best overall performance with the lowest validation and test MSE, the highest validation and test R^2, and the lowest mean cross-validation-RMSE of the four models. The SVM model had the lowest standard deviation in the cross-validation results, suggesting it is the most consistent (since it had the most consistent performance across the different subsets of data). The decision tree model showed overfitting with the training R^2 score of 1.

Future Work

There are several areas of future work that could provide valuable insight. We would have liked to include data on social perceptions around mental illness but only had one year of survey data and were, therefore, unable to conduct an analysis over time. Further research looking at changes in perception over time could uncover potential relationships with government intervention and, ultimately, rates of mental illness. Changing social perceptions may influence government intervention or vice versa. Knowing which source is most effective to address first would prove valuable for mental health advocates.

More robust analysis of government intervention would also be beneficial for future research. Our data only covered a four-year time span with gaps in reporting from many countries. It would be ideal to have a more complete dataset as well as a better understanding of the legislation or policy implemented. Without knowing the validity of the intervention applied, one could draw the incorrect conclusion that government intervention in general is not effective in reducing rates of mental illness when, in fact, it may simply be that the particular intervention chosen was poorly designed or poorly executed. An evaluation of the legislation or policy implemented would provide more significance to an analysis of its effect on prevalence rates.

We could also increase the size of our dataset if we spent time cleaning the country names before merging. Conflicts in country names (e.g., “United States” vs “United States of America”) were resolved lazily, via merging.

For the models, future work would include hyperparameter tuning and evaluating how other methods to combat outliers and skewed distributions affect model performance. For example, winsorizing or winsorizationation.

Appendix

Data Sources

This project used three datasets, but identified and processed four, each described below. Access instructions are detailed in each preprocessing notebook file.

1. Prevalence dataset
    - *Institute for Health Metrics and Evaluation (IHME), Global Burden of Disease Study (2019):* prevalence rates of mental health conditions from countries across the globe for male, female, and both genders, dating from 1990 – 2019.
        - Accessed using the owid-catalog Python package, detailed in this report's Data Acquisition section. [1]
2. GDP dataset
    - DataBank's *World Development Indicators*: provides GDP per capita (current US$, accessed 2024) for countries from 1990 – 2019.
3. Survey dataset (for public perceptions of mental health)
    - Excluded
    - *Wellcome Global Monitor (WGM) 2020: Mental Health—Wave 2:* the world's largest survey about how people think and feel about science and health challenges in 2020. [1]
4. Governance dataset
    - *World Health Organization (WHO)*'s Mental Health Governance dataset, which captures information about a country's level of government support for mental health care (e.g., do they have a state policy; is that policy still active; what year was the policy adopted).

Table 3. Summary table of project files. While the table excludes original datasets, these are still included in the supplemental files.

type	filename
processed dataset	A1.cleaned_governance A2.final_merged A3.gdp_per_capita_melted A4.gbd_mental_health_prevalence_rate_flattened
pre-processing notebook	B1.preprocessing_gdp_per_capita_dataset B2.preprocessing_governance_dataset B3.preprocessing_merging_datasets B4.preprocessing_prevalence_dataset
analysis notebook	C1.data_processing_and_analysis
modeling notebook	D1.predictive_modeling