Prediction for Mental Health Condition in Thirty Days

Group Member: Xinwei Wang, Xinyu Zhang, Neyan Deng

Introduction

Background

Mental health is fundamental to overall well-being and quality of life. According to the report of the National Institute of Mental Health (NIMH) in 2022, 8.4% of U.S. adults experienced a major depressive episode, reflecting the broad scope of these challenges[1]. With the pervasive impact of mental health issues like depression and anxiety, which affect more than 264 million individuals worldwide, there's a clear signal of a significant global mental health burden[2].Ample research underscores the influence of social and familial factors on mental health outcomes. Mental health struggles often coincide with socioeconomic and family dynamics, such as the family care availablity[3, 4,5]. The correlation between lower socioeconomic standing and a heightened risk of mental health problems, particularly in children and adolescents[6]. Beyond this, experiencing family hardship or a lack of emotional support can be traumatic, having enduring effects on one's mental health[7,8,9].

Motivation

In light of this knowledge, our study is particularly interested in delving deeper into this relationship. The study aims to construct predictive models that utilize familial and social indicators to forecast mental health status within a 30-day timeframe with heightened precision. We seek to identify accurate models and discern impactful predictors for mental health forecasting.

Data Description

The Behavioral Risk Factor Surveillance System (BRFSS) is utilized in our study. This is a state-based survey system in the United States, managed by the Centers for Disease Control and Prevention (CDC). The BRFSS data was chosen for its extensive coverage over a decade, providing a comprehensive dataset on American health behaviors. Despite it is not specifically used for investigating mental health issues and searching for alternative longitudinal sources, the BRFSS emerged as the most complete option available. It offers a wide array of variables crucial for exploring potential factors linked to bad mental conditions and building predictive models.

Method

Data Integration & Management

Our data integration process encompassed records spanning from 2013 to 2022. The initial challenge we faced was the sheer volume of the dataset, which exceeded 4GB. This vast size significantly hampered the speed of integration, especially when using basic R programming techniques. To address this bottleneck, we employed parallel computing methods. This approach markedly enhanced our processing efficiency, enabling us to handle the large dataset more effectively.

Key to this process was the identification and removal of outliers, which was crucial for maintaining the integrity and relevance of our analysis. Such exclusions were necessary to ensure the validity of our findings. Another significant aspect of our data management was the categorization of mental health scores into six distinct levels. This categorization allowed for a more nuanced and precise analysis of mental health trends within the dataset.

Our meticulous data cleaning and management efforts resulted in a substantial refinement of the dataset. From an initial volume of approximately 4 million rows, we streamlined the data to a

more focused and manageable set of 100,000 rows. This optimized dataset was then saved as a new data file. This reduction not only improved the manageability of the data but also significantly reduced processing time and enhanced the overall efficiency of the project.

Predictors Selection

The outcome variable of this study is the mental health condition within 30 days. Table 1 outlines the variables selected from the BRFSS data spanning 2013 to 2023, categorized by demographic, health, family, and socioeconomic types. We also created a new predictor IYEAR to provide a temporal context. The year 2020 was set as the pandemic outbreak of the baseline '0' year, with the dataset spanning from 2013 to 2023.

Table 1: decriptions of predictors of mental health.

Predictors	Types	Description
IYEAR	Pandemic	The time range from -7 to 3 relative to 2020 spans the years 2013 to 2023.
EDUCA	Demographic	The highest level of education completed by respondents, ranging from no formal education to college degrees or higher.
MARITAL	Demographic	The marital status of respondents, including categories such as married, divorced, widowed, separated, never married, or in an unmarried partnership.
SMOKE100	Health	A bivariate predictor recorded if respondents have smoked at least 100 cigarettes in their lifetime.
EXERANY2	Health	A bivariate predictor recorded if respondents have engaged in any physical activities or exercises outside of their regular job in the past 30 days.
RENTHOM1	Family	Housing status, indicating whether they own, rent, or have another arrangement for their home.
CHILDREN	Family	The number of children under 18 years of age living in respondents' households, ranging from none to 20.
NUMADULT	Family	The number of adults in a respondent's household, ranging from 1 to 5 adults, with an option to indicate 6 or more.
INCOME2	Social	Respondents' annual household income from all sources into ranges, starting from less than $10,000 to $75,000 or more.
EMPLOY1	Social	Respondents' current employment status, with options ranging from employed for wages to unable to work, including categories for those out of work or retired.
HLTHPLN1	Social	A bivariate predictor recorded if respondents have any form of health care coverage, including insurance, government plans, or other health services.
MEDCOST	Social	A bivariate predictor recorded if respondents were unable to see a doctor in the past 12 months due to cost.
PVTRESD1	Social	A bivariate predictor recorded if respondent's living arrangement is a private residence.

Sampling

The sampling process for our report involved a stratified approach to ensure proportional representation from the BRFSS dataset across different years. For 2013-2022, each year's data was grouped and a sample of 1200 records was selected due to the full year's data availability. In contrast, for 2023, where only January data was available, a smaller sample of 100 records was chosen. The samples from each year were then combined into a single dataset for analysis, balancing the representation across the years within the constraints of data availability.

Train & Test Set Split

After data processing, we split the data set into a training set with 70% of our sampled data, and a testing set with 30%, to construct our model to predict days that mental health does not feel good in 30 days, and check the precision of our prediction results.

Transformation

Before constructing our first model which is a multiple linear regression model, we checked the residual plots for each one of them. Residuals of four variables (SMOKE100, EXERANY2, MARITAL, and NUMADULT) showed higher correlations with our response variable. A log transformation helped with this problem.

Model & Results

Multiple Linear Regression

We have our first model through multiple linear regression using all 13 predictors, with four of them using log transformation. The following results has been obtained after fitting the model:

$$MENTHLTH = 5.813 + 0.032 \times IYEAR - 0.064 \times EDUCA - 0.583 \times log\,(SMOKE100) + 0.775 \times log\,(EXERANY2) + 0.060 \times log\,(MARITAL) - 0.086 \times log(NUMADULT) + 0.103 \times RENTHOM1 + 0.002 \times HLTHPLN1 + 0.063 \times EMPLOY1 - 1.707 \times PVTRESD1 - 0.308 \times MEDCOST - 0.038 \times CHILDREN - 0.102 \times INCOME2$$

We calculated MSE, MAE, MBD (Mean Biased Deviation), and R-squared of this model, and the results are MSE = 3.008, MAE = 1.445, MBD = 50.683%, R-squared = 0.127. The summary of the linear model shows that despite variables of MARITAL, NUMADULT, HLTHPLN1, and PVTRESD1 with p-values greater than 0.05, all other variables are significantly related to the response. We want to check if a better result could be achieved by other models.

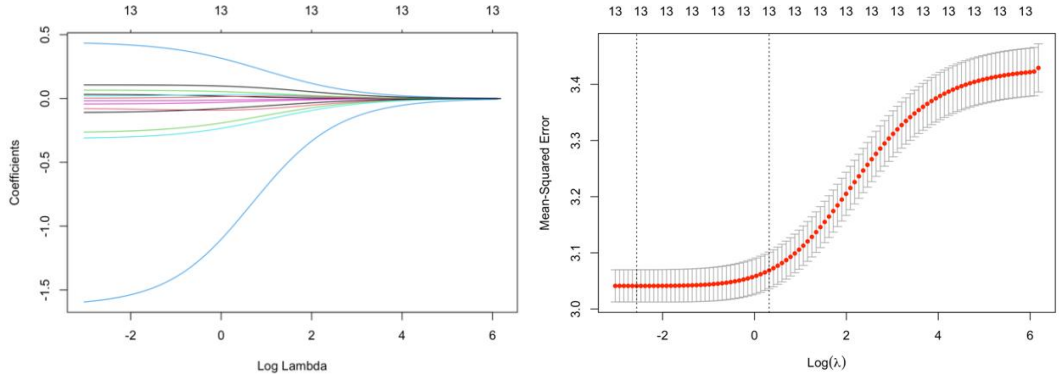
The correlation plot shows that the income variable is relatively highly correlated to a few variables including EDUCA (Corr = 0.49), MARITAL (Corr = -0.38), RENTHOM1 (Corr = 0.35), and EMPLOY1 (Corr = -0.41). The VIF results suggest that there is no significant multicollinearity among these predictors. But we still processed a ridge regression for a double check.

IYEAR	EDUCA	LOG(SMOKE100)	LOG(EXERANY2)	LOG(MARITAL)	LOG(NUADULT)
1.048	1.414	1.078	1.092	1.382	1.096

RENTHOM1	HLTHPLN1	EMPLOY1	PVTRESD1	MEDCOST	CHILDREN	INCOME2
1.225	1.027	1.247	1.003	1.060	1.012	1.909

Ridge Regression

With a cross-validation method to choose our parameter λ, we finally selected the minimum value λ.

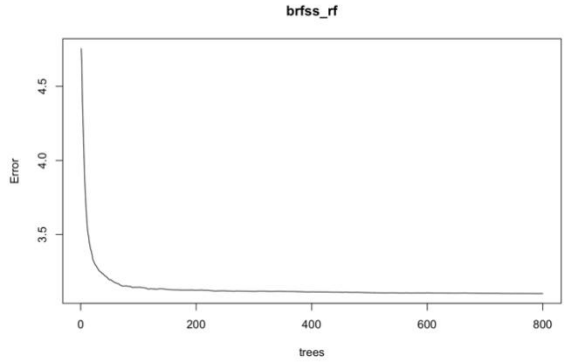


However, the performance of predicting did not show any improvement, with MSE = 3.024, MAE = 1.455, MBD = 51.567%, and R-squared = 0.125. It is not significantly useful for

building the prediction model. Also, there is no problematic multicollinearity that needs to be considered.

Random Forest

Random forest uses bootstrap and bagging to train the data, which are algorithms for improving the accuracy of machine learning models. It is less likely to overfit because it integrates the decisions of multiple trees, reducing the model's dependence on specific data. The graph of the Error vs. Trees shows that model error decreases as the number of trees increases.



We set the number of trees to 800. The results show a slight improvement in MAE and MBD compared to the multiple linear regression model, with MSE = 3.054, MAE = 1.442, MBD = 50.25%, and R-squared = 0.114.

We have checked the importance as well. Employment status (EMPLOY1) emerged as the most crucial predictor for our response variable, with the highest increase in Mean Squared Error (%IncMSE) when permuted. Exercise frequency (EXERANY2), marital status (MARITAL), and income level (INCOME2) also significantly influenced the model's predictions, indicating these factors are strongly associated with mental health outcomes. In contrast, variables like private residence (PVTRESD1) had no discernible impact on the model's performance, while year (IYEAR) and health plan coverage (HLTHPLN1) showed relatively minor importance. This highlights the model's reliance on employment, lifestyle, and socio-economic factors as key determinants of mental health within the dataset.

Discussion & Conclusion

Some Improvement has been tried during our process. We implement parallel computing and spark when doing the random forest since the original running time was extraordinarily long, especially with a larger dataset. The following table is a comparison of the spending time. Using parallel computing helped to speed up a little bit, but using Spark achieve the fastest way to compute.

RANDOM FOREST	BASE	PARALLEL	SPARK
TIME / *SECOND*	147.81	132.43	4.53

We have also tried to use Rcpp functions for the MLR part. It did speed up, but originally, it did not take much time to run.

By comparing the four indicators between the MLR model and RF model for predicting days with bad mental health in 30 days, each model exhibited strengths across different evaluations. The MLR model demonstrated a lower Mean Squared Error (MSE=3.008) and a higher R-squared value (0.127), indicating better overall fit and predictive accuracy when compared to the RF model (MSE=3.054, R-squared=0.114). On the other hand, the RF model showed slightly better results in terms of Mean Absolute Error (MAE=1.442) and Mean Bias Deviation

(MBD=50.25%), suggesting closer alignment with the actual data points in terms of average prediction error. Given the closeness in performance across these metrics, the selection of the best model for mental health prediction might involve additional considerations such as stability.
Moreover, both two models did not show high prediction accuracy for predicting days with bad mental health in 30 days. This may be due to the following limitations:
We should consider more about the problem of causal order, which is the chronological sequence of the events (variables), and the causal relationship as well.
Meanwhile, BRFSS data is a cross-sectional dataset that does not have a clear focus on mental health or related fields. It has more than 300 variables that involve various aspects of human life and behavior, while we only extracted 14 of them which are related to our prediction purpose.
Besides, this dataset captures information at specific points in time without tracking individual changes over the years. This aspect restricts our ability to establish temporal sequences between factors and mental health outcomes, potentially affecting the predictive accuracy of our model. Although the data spans a decade, each year's dataset is independent, limiting longitudinal analysis. Future research would benefit from longitudinal data to trace and analyze the dynamics between risk factors and mental health more accurately.

Reference

1. National Institute of Mental Health. (2022). Depression. (NIH Publication No. 21-MH-8079). U.S. Department of Health and Human Services, National Institutes of Health.
2. Jones, Grant M., and Matthew K. Nock. "Lifetime use of MDMA/ecstasy and psilocybin is associated with reduced odds of major depressive episodes." *Journal of Psychopharmacology* 36.1 (2022): 57-65.
3. Compton, M. T., & Shim, R. S. (2015). The social determinants of mental health. *Focus*, *13*(4), 419-425.
4. Alegría, M., Alvarez, K., Cheng, M., & Falgas-Bague, I. (2023). Recent advances on social determinants of mental health: looking fast forward. *American Journal of Psychiatry*, *180*(7), 473-482.
5. McKay, M. T., Cannon, M., Chambers, D., Conroy, R. M., Coughlan, H., Dodd, P., ... & Clarke, M. C. (2021). Childhood trauma and adult mental disorder: A systematic review and meta-analysis of longitudinal cohort studies. *Acta Psychiatrica Scandinavica*, *143*(3), 189-205.
6. Pang, S., Liu, J., Mahesh, M., Chua, B. Y., Shahwan, S., Lee, S. P., ... & Subramaniam, M. (2017). Stigma among Singaporean youth: a cross-sectional study on adolescent attitudes towards serious mental illness and social tolerance in a multiethnic population. *BMJ open*, *7*(10), e016432.
7. Lee, H., Kim, Y., & Terry, J. (2020). Adverse childhood experiences (ACEs) on mental disorders in young adulthood: Latent classes and community violence exposure. *Preventive medicine*, *134*, 106039.
8. Crandall, A., Miller, J. R., Cheung, A., Novilla, L. K., Glade, R., Novilla, M. L. B., ... & Hanson, C. L. (2019). ACEs and counter-ACEs: How positive and negative childhood experiences influence adult health. *Child abuse & neglect*, *96*, 104089.
9. McCutchen, C., Hyland, P., Shevlin, M., & Cloitre, M. (2022). The occurrence and co-occurrence of ACEs and their relationship to mental health in the United States and Ireland. *Child Abuse & Neglect*, *129*, 105681.