**Data Source and Relevancy**

As the rates of those diagnosed with depression continue to rise in the United States,[1](https://news.gallup.com/poll/505745/depression-rates-reach-new-highs.aspx) depression is expected to rank first in disease burden among high-income countries by 2030.[5](https://doi.org/10.1371/journal.pmed.0030442) New intervention methods are critically needed. About 60% of individuals with one episode of major depression will experience another in their lifetime.[2](https://doi.org/10.1016/s0013-7006(09)73481-5) Some research suggests that early intervention can prevent the onset of depressive disorders. [3](https://doi.org/10.1176/appi.ajp.2008.07091422)I suggest using machine learning to find potential trends in the National Longitudinal Study of Adolescent to Adult Health (Add Health)[4](https://www.icpsr.umich.edu/web/ICPSR/studies/21600/datadocumentation) that could predict the diagnosis of depressive disorder, to provide targeted, early intervention.

If there is sufficient correlation between the blood panels and the diagnosis of depression, I aim to explore the feasibility of creating a model that could predict the most likely change—based on a small number of questions, demographic factors, and a blood panel—in depression symptoms as eight years later. This model could have clinical relevance in directing early intervention, preventing the onset of depression, and reducing the disease burden on society.

The Add Health study follows a cohort of subjects from adolescence into adulthood. During Wave IV, when the subjects are 24 to 32 years old, blood samples are taken and tested for inflammation, immune function, glucose homeostasis, and lipids, along with a questionnaire on lifestyle habits, demographic factors, and mental and physical health. In Wave IV and eight years later in Wave V, the same subjects are asked a series of four questions that align with an abbreviated version of the CES-D scale for depression. These four questions also align with four out of the five questions that were validated by *Perreira et al.* (2005) for comparison purposes and are the same questions used by Boardman et al. (2011), in their analysis of the same data source. They found that using that scale “gave virtually identical results” to the nine-item version in wave IV.

From the questionnaire portion, my model uses four demographic factors, two questions related to income and military status, eight questions related to health, 11 questions related to physical activity, 10 questions related to community, religious, and political involvement, five questions related to addictive behavior, one question related to a memory task. It also included 31 variables from the biometric section, for a total of 72 variables. I combined two variables related to diabetes to create one new variable coded as 0 = no diabetes, 1 = prediabetes, and 2 = diabetes. I also applied one-hot encoding to one variable (political identification) to produce six separate binary variables. At this point I had 3713 entries and after removing entries that were missing seven or more data points I was left with 3254 entries.

|  |  |
| --- | --- |
| **y** | **Number of samples** |
| -3 | 0 |
| -2.75 | 1 |
| -2.5 | 4 |
| -2.25 | 5 |
| -2 | 10 |
| -1.75 | 13 |
| -1.5 | 24 |
| -1.25 | 54 |
| -1 | 92 |
| -0.75 | 157 |
| -0.5 | 246 |
| -0.25 | 455 |
| 0 | 789 |
| 0.25 | 569 |
| 0.5 | 341 |
| 0.75 | 164 |
| 1 | 101 |
| 1.25 | 64 |
| 1.5 | 32 |
| 1.75 | 28 |
| 2 | 18 |
| 2.25 | 14 |
| 2.5 | 6 |
| 2.75 | 4 |
| 3 | 1 |

For the response variable, I took the average score from the four mental health questions (each coded from 0 to 3) in Wave IV and subtracted it from the average in Wave V to create a new variable that indicates the change in depression symptoms eight years after Wave IV. I removed any entries that didn’t have data for either Wave IV or Wave V on any of these questions. After all of this, there were 3,193 samples left.

The output represents changes in depression symptoms over eight years and is pseudo-continuous. This is due to the nature of the data and averaging used on it, where only a discrete number (24) of outcomes are possible. (fig 1)

**Methods**

I applied a 70%, 15%, 15% split of the data for parameter tuning, training, and a final testing of the model.

Next, I took all the explanatory variables and put them through a preprocessing pipeline. I applied a power transform and log transform on non-binary data with obvious skews based on histograms (fig 2) to create more normal distributions. I then imputed missing data points using a median initial strategy and created polynomial features of degree two to prevent underfitting the data. I thought some of the data, like “fast time”—the amount of time participants had fasted before giving blood—was more relevant in combination with other biometric results. After that, the pipeline applied a standard scaler to mean-center the data and scale it to a variance of 1, as required for PCA. For these steps, I made sure to train the model on only the training data to prevent data leakage.

(Fig. 1) distribution of y

Since I hypothesized that there would be a lot of multicollinearity in my dataset, I applied PCA for dimensionality reduction to prevent overfitting, retaining 95% of variance. I also trained this on only the training dataset and used it to transform the others. It reduced the number of features from 2,700 down to 505.

Finally, I fed the test set into a linear regression model with stochastic gradient descent (SGD) and an ElasticNet penalty. Using a custom loop, I searched for the optimal parameters by comparing the RMSE on 40 random parameters sets within a parameter space and running the model for 5,000 epochs for each parameter set or until the RMSE did not improve for five straight epochs. I then saved all the different parameters and their associated RMSE to a CSV so I could compare them later. I searched for an regularization strength (alpha), an initial learning rate (eta0), and an l1 ratio. I chose the Elasticnet so that I could find the best ratio of l1 and l2 regularization.

**Results**

The best parameters are: (fig 3)

• **l1\_ratio** = 0.556

• **eta0** = 0.001

• **alpha** = 0.373

Using these optimal parameters, I trained a new model with a 50,000 iteration limit on a combination of the training and validation sets. I used this model to predict the final test set and recorded the mean squared error for the final validation of the model. The final resulting RMSE equals **0.605** and the R² equals **0.174**. So, my features explain 17% of the variance in the data, which is a modest amount.

I didn’t have enough time to dig into the parameters and backwards engineer the PCA analysis to figure out which features were most important. My model could explain around 17% of the variance, so it isn’t highly effective at predicting, but it would be interesting to understand more about which features were most important in explaining that, and I just ran out of time.

**Discussion**

Initially, I set it up as a classification problem, sorting my data into two groups: depressed or not depressed. I eliminated everyone from the dataset who, in Wave IV, had said they had been diagnosed with depression at any point in their life. This eliminated a significant portion of my data. Next, I used the question in Wave V: “Has a doctor, nurse, or other health care provider ever told you that you have or had depression?” I combined this with another question that asked when this diagnosis occurred, and if it was less than eight years ago, I included them as a positive data point—no depression in Wave IV but depression in Wave V. Due to the unreliability of this question, this removed a number of people from Wave V who either didn’t answer this question or whose follow-up answers were inconsistent with their Wave IV data. Moreover, I felt this question was a poor way to assess if someone was actually experiencing depression.

Although I couldn’t find a clear mapping of the mood-related questions in Wave IV and V to a well-validated diagnostic questionnaire, I considered doing some kind of clustering analysis to see if those who said they were diagnosed with depression in Wave V clustered with a group that didn’t. I could then use that model to classify between depression and not depression. When I built this model, I found the ROC score was only 0.51—basically useless. But after some thinking and a literature review, I found a paper that validated a five-question diagnostic questionnaire appropriate for comparative purposes. Four of these questions were asked in both Wave IV and Wave V. I also read some papers that measured changes in depression, and I decided to take the average answer of those four questions for Wave IV and subtract it from the average of the same questions for the same participant in Wave V to find the change in depression symptoms. This allowed me to keep many more participants in my dataset and seemed like a cleaner and more objective way of assessing changes in depressive-like symptoms. This meant I had to change my model from a logistic regression used for classification to a linear regression. It also led me to realize I should be using polynomial features—an oversight previously. I wonder what the previous model’s ROC score would have been if I had used polynomial features.

If given more time, on top of working backwards through the model to figure out which variables explained the most variance, I would also like try using a random forest, which would allow me to capture nonlinear relationships. This may hurt the interpretability of the model, but I wonder if it would improve the prediction accuracy. I would also be interested in applying a neural network for much the same reasons. Additionally, applying more or fewer features, doing more mindful feature engineering and skipping PCA, may have helped create more meaningful and interpretable data. Finally, I would have loved to apply t-SNE to the dataset to visualize it and see if I could identify any clustering between the group that would develop depression and the ones that would not. I also could have explored other aspects of the data, seeing if there was clustering between depressed and not depressed at a single point in time. I also would have liked to, after applying the response variable discussed above, create four separate groups to explore: those who were depressed at Wave IV and Wave V, those that were not depressed in both waves, those that got more depressed, and those that became less depressed.

That also points to an issue with the analysis I did. Those that started depressed would be less likely to get more depressed, and those that started not depressed would have been less likely to become less depressed, which may introduce some bias into my data, as those two groups were heavily skewed—most people were not depressed to start off with. I wonder if this would cause some noise in the data and make it harder to pick out features that may be contributing to depression or alleviating it. Another problem is that the SGD algorithm, as the name implies and as I realized while running this model multiple times, isn’t 100% reproducible. I used it because I thought it would help my model converge faster, and I was having issues with that. But if I wanted to make inferences about which features are most important and publish that, I might need a model that could be reproduced more reliably.

A group of graphs showing different types of data

Description automatically generated

A chart with yellow and blue dots

Description automatically generatedA screenshot of a graph

Description automatically generated

(Fig. 3) Scatterplot of L1 ratio vs alpha for parameter search. (purple is better)

(Fig. 2) Histograms of variables in X before preprocessing