Predictive Modeling for the Identification of Depression and Suicidality

Nicholas K. Branigan

November 18, 2020

1 Introduction

Major Depressive Disorder (MDD) is a leading cause of disability in the United States and globally [6]. In this project, I attempt to predict depressive symptomatology using data from the National Health and Nutrition Examination Survey (NHANES). I apply three models to this effort: logistic, support vector machines, and a fully-connected neural network. My goal is to offer a predictive approach for identifying two critical health conditions: moderately severe Major Depressive Disorder and suicidal ideation.

One may wonder what the point of a model is in this context. What's wrong with simply asking people whether they're depressed, or having suicidal thoughts? After all, this is broadly the approach taken across all divisions of modern medicine. Doctors wait for patients to present with worrying or problematic symptoms, and then provide appropriate treatment. For most illnesses, this seems to be a sensible approach. However, for psychiatry, this strategy has major shortcomings. Campaigns to de-stigmatize Major Depressive Disorder still have a long way to go. Many people experiencing depressive crises are not eager to seek medical attention [2]. Moreover, a substantial number fail to represent their symptoms truthfully when asked about them by friends or family [4]. Predictive modeling of MDD and suicidal ideation could help identify patients in desperate need of medical intervention who are not seeking it.

I am aware of two papers that attempt to predict the presence of depressive symptoms using machine learning approaches. Sharma and Verbeke apply XGBOOST on a comprehensive dataset of biomarkers for over 10,000 Dutch citizens [7]. Closest to my work, Oh et al. apply deep learning and a logistic model to NHANES data covering 1999 to 2014 and K-NHANES data (the South Korean equivalent of the CDC's dataset) [5]. I am not aware of any publications that attempt to predict suicidality specifically from a large public health database or that try to predict depressive symptoms from only a standard blood screen. The focus of my paper is on these last two tasks, though, for comparison, I will also attempt to predict depressive symptomatology from the best set of predictors I can find.

2 Data

2.1 Overview

The data are from the 2015-2016 and 2017-2018 cycles of NHANES. NHANES is a vast public health dataset. Modules of variables are classified under demographics, dietary, examination, laboratory, questionnaire, and limited access data. In my analyses, I include predictors from `Demographic Variables and Sample Weights`, `Oral Health - Recommendation of Care`, `Standard Biochemistry Profile`, `Audiometry`, `Diet Behavior Nutrition`, `Drug Use`, `Physical Activity`, and `Smoking - Cigarette Use`. I construct my target variable from `Mental Health - Depression Screener`. Each cycle of NHANES includes data from roughly 10,000 subjects. However, missing data are ubiquitous. This is even the case for non-obscure variables that would seem to be of great interest to researchers.

2.2 Pre-Processing

For some difficult to fathom reason, CDC only releases NHANES data in the .XPT formal. .XPT is a transport file for the SAS statistical software. CDC claims that this presents no compatibility issues, since SAS provides a free software to view .XPT files and save them to .CSV files. Aside from this software being available only for Windows, I have learned that it has a critical flaw: it saves missing values as 0. For many NHANES modules, this contaminates and makes unusable the data. Apparently, this is a known issue and SAS has not solved it because they provide only minimal technical support for this program. I learned that Matlab is capable of importing .XPT files to tables from which export to .CSV is simple. However, I am astonished that CDC has placed this barrier between the public and one of the major Amer-

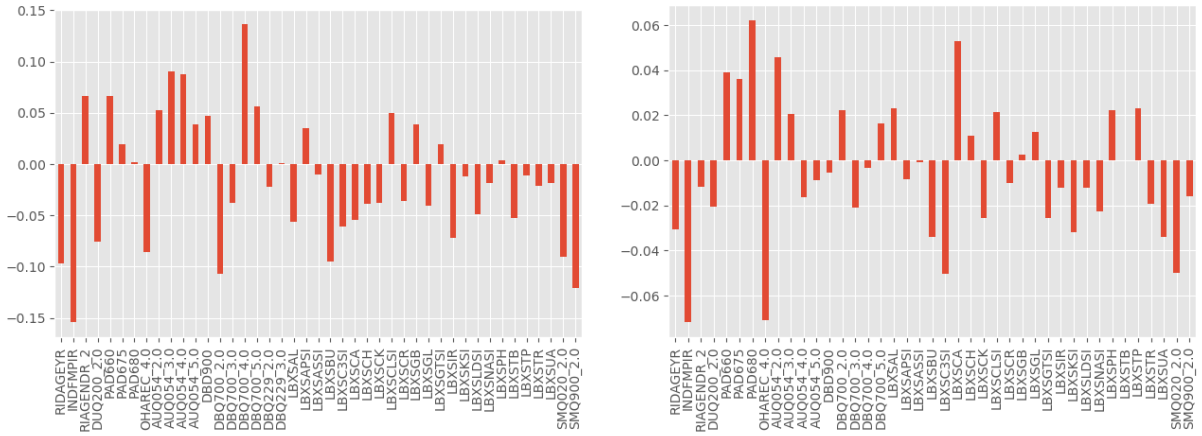

Figure 1: Pearson's correlation coefficient of predictors with moderate MDD (left) and suicidality (right).

ican public health databases for no apparent reason but their ignorance.

After pulling the modules into .CSV, I coded missing values, dropped variables that were not of interest, turned categorical variables into indicators, and merged the modules into one joint dataset. For predictors that had missing observations, I experimented with two approaches: filling the missing observations with the mean of the predictor, and dropping the missing observations entirely. I chose the latter strategy, since it led to consistently better model performance, though it meant discarding thousands of observations.

2.3 Features

I performed analysis on the total dataset as well as a subset of the total dataset consisting of the standard biochemistry panel. I made this choice for two reasons. First, a standard biochemistry panel is inexpensive to administer and is, essentially, objective. The other variables I examine are, for the most part, questionnaire results. Consequently, they are subject to many known biases. Respondents may forget events in systematic ways, and even if they accurately recall what they are being asked about, they may distort the truth or lie in response to sensitive questions. Moreover, even if we assume that NHANES subjects have excellent recall and are unusually honest, questionnaire data remain problematic. A primary goal of my model is to predict well on unseen data. Whether the model can successfully do that will depend on how accurately people not in the NHANES subject pool respond to these questions. Reasons for thinking that people outside of this group may answer more or less truthfully include varying cultural attitudes that determine whether a question is deemed sensitive or not and different protocols for asking questions (e.g., responding to prompts from a computer or from a person, the promise of anonymity). Second, since I chose to discard all observations that included any missing data, including fewer predictors substantially increased the size of the available data. The two datasets are described below:

Design matrix	Predictors	Component	Obs.
biochemistry only	22	train	7536
		test	2000
all predictors	43	train	834
		test	400

2.4 Targets

The NHANES `Mental Health – Depression Screener` includes subject responses to the nine questions of the ambiguously named Patient Health Questionnaire (PHQ). The PHQ is a standard screening device for MDD. It queries subjects on how often they experienced symptoms that fit the DSM-IV diagnosis of MDD. Results on the PHQ of 5, 10, 15, and 20 represent mild, moderate, moderately severe, and severe MDD respectively [3].

I have chosen as my target variables the presence or absence of moderately severe MDD and suicidal ideation. Following [3], I define moderately severe MDD as a PHQ score greater than 9. To define suicidal ideation, I look to the penultimate question on the PHQ: "Over the last 2 weeks, how often have you been bothered by thoughts that you would be better off dead or of hurting yourself in some way?" I define the presence of suicidal ideation to be any answer other than "not at all."

Figures 1 and 2 give a sense of the modeling problem. For no predictor is the absolute value of its correlation

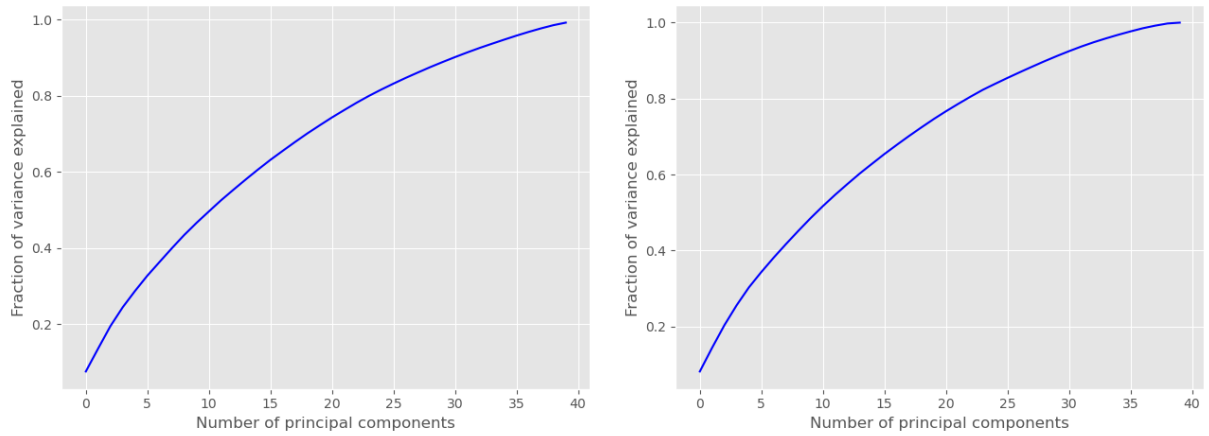

Figure 2: Using principal component analysis, variance of moderate MDD explained (left) and variance of suicidality explained (right).

with moderate MDD greater than 0.2 and with suicidal ideation greater than 0.1. The correlations are consistent with expectations and past literature examining demographic predictors of psychiatric illness [8], [1]. Predictors that are noticeably positively correlated with MDD include auditory deficits, fair and poor self-reported diet, and female gender. Those that are negatively correlated include higher age, higher ratio of family income to the poverty line, very good self-reported diet, having smoked fewer than 100 cigarettes during lifetime, having never used an e-cigarette, and having never used marijuana or hashish. The predictors tend to be similarly correlated with suicidality, with major correlates of suicidality including lower income, higher minutes sedentary activity, poor oral health on dental exam, lower bicarbonate, and higher calcium. The curves in Figure 2 are roughly identical, each with a gentle slope. For both targets, each principal component, even those that do the most explanatory work, contributes fairly little to the total fraction of variance in the targets explained.

From this choice of targets, the dataset is highly imbalanced. The table below summarizes the class constituents:

Design matrix	Component	Class	% Obs.
biochemistry only	train	depressed	8.54
		suicidal	3.73
	test	depressed	9.4
		suicidal	4
all predictors	train	depressed	3.60
		suicidal	2.40
	test	depressed	6.75
		suicidal	2

Both suicidality and moderate MDD are very uncommon in the dataset. Interestingly,the biochemistry only data has a uniformly higher prevalence of both suicidality and depression compared to all predictors. This is not entirely unsurprising, since in order for a subject to be included in the latter, he needs to have no missing data whatsoever for the 40 some variable I am considering. Compared with non-depressed and non-suicidal subjects, depressed and suicidal ones may be considerably less eager to participate in the NHANES data collection process.

3 Methods

3.1 Class Imbalance

I address the class imbalance problem in two equivalent ways. For the neural network model, I upsample the minority class by randomly selecting elements of the minority class to duplicate so that the upsampled training data consists of an equal number of positive and negative observations for the target variable. For the other two models, I use a standard re-weighting scheme. Let TP denote true positives, FP false positives, TN true negatives, and FN false negatives. Define

$$\rho = \frac{TP + FN}{TP + TN + FP + FN}. \tag{1}$$

Further, let $\lambda = \frac{\rho}{1-\rho}$ and n be the number of observations in the dataset. Then, the re-weighted loss function for logistic regression can be written

$$J(\theta) = -\frac{1+\lambda}{2n} \sum_{i=1}^{n} w^{(i)} \Big(y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \Big) \tag{2}$$

Prediction	Data source	Model	Parameters	Balanced accuracy	Accuracy	Sensitivity	Specificity
depression	all predictors	logistic	13 PC	55.52%	71.5%	37.04%	74.00 %
		SVM	42 PC	57.28%	78.00%	33.33%	81.23 %
		neural	(48, 14, 9)	58.70%	90.25%	22.22%	95.17 %
	biochem only	logistic	2 PC	59.02%	57.7%	60.64%	57.40%
		SVM	2 PC	58.88%	60.9%	56.38%	61.37%
		neural	(4, 9, 9)	60.07%	71.7%	45.74%	74.39%
suicidality	all predictors	logistic	4 PC	72.58%	58.25%	87.5%	57.65%
		logistic	37 PC	60.33%	70.25%	50%	70.66%
		SVM	12 PC	73.09%	59.25%	87.5%	58.67%
		SVM	16 PC	77.04%	55%	100%	54.08%
		neural	(2, 6, 9)	60.71%	95%	25%	96.43%
		neural	(25, 3, 9)	73.21%	71.5%	75%	71.43%
	biochem only	logistic	2 PC	53.02%	58.1%	47.5%	58.54%
		SVM	5 PC	48.3%	83.6%	10%	86.67%
		neural	(37, 9, 9)	60.31%	67.5%	52.5%	68.125%

Notes on table notation: balanced accuracy and accuracy are on the test set; PC stands for principal components; (i, j, k) refers to the number of hidden units in the neural network where i is the number of hidden units in the first layer, j is the number in the second layer, and k is the number in the third layer.

where $w^{(i)} = 1$ if $y^{(i)} = 0$ and $w^{(i)} = \frac{1}{\lambda}$ if $y^{(i)} = 1$. The support vector machines implementation via Scikit-Learn uses a similar re-weighting scheme.

To evaluate my models in light of the substantial class imbalance, I consider two metrics. The first is the standard measure of accuracy, which I denote A.

$$A = \frac{TN + TP}{TN + TP + FN + FP}. \tag{3}$$

In other words, A corresponds to the fraction obtained by dividing the number of correct predictions by the number of predictions. Clearly, this metric fails to convey the performance of a model well if one class is much more prevalent than another. The metric I put more emphasis on is the balanced accuracy, denote A_b.

$$A_b = \frac{1}{2}\left(\frac{TN}{TN + FP} + \frac{TP}{TP + FN}\right). \tag{4}$$

Observe that A_b is an equally-weighted sum of the sensitivity and specificity of the classifier.

3.2 Learning Approaches

I applied three learning algorithms to the datasets: logistic regression, support vector machines, and a neural network. I implemented all three algorithms using the library Scikit-Learn.

For the logistic and support vector machines models, I saw the best results after first standardizing my design matrix, then applying principle component analysis to reach a transformed design matrix, then fitting the model to some number of principle components determined by hyper-parameter tuning. For logistic regression, I experimented with different solving methods, types of regularization, and regularization strengths. I settled on the Newton conjugate-gradient algorithm with l2 regularization and the default regularization strength. As discussed in the previous section, I applied a re-weighting scheme to the data by setting the Scikit logistic argument class_weight to balanced. I set all other parameters to the Scikit defaults.

For the SVM model, I experimented with different kernels and regularization strengths. I saw, by far, the best results with the linear kernel. I also settled on default regularization strength. As with logistic, I set class_weight to balanced, and all other Scikit parameters to the default.

My neural network is a fully-connected system with ReLU activation functions for the hidden layers and a perceptron activation function for the output layer. I tuned the number of layers and the number of units per layer. I varied the first layer size between 1 and 50, the second layer size between 0 and 25, and the third layer size between 0 and 10. I used a random seed for reproducible results and all other Scikit parameters set to defaults.

4 Results

In the page 4 table, I report the best performing models from each model type for the various predictive tasks. There is a wide range of model performance, and no one

model consistently outperforms the other two for each task. Moreover, the models were broadly more successful in some classification contexts than in others. I report such a range of results since I want to give a sense of what worked and what did not, and since different models could be useful in different predictive contexts.

Now, I discuss a few of the highlights. Two neural network models had possibly useful specificity on the test data. For predicting moderate MDD, the neural model with 48 units in the first hidden layer, 14 in the second, and 9 in the third returned 95.17% specificity. Though sensitivity of this model was only 22.22%, of the 24 people it classified as depressed, 6 were in fact depressed. Thus, 25 % of the positive labeled group were depressed, though only 6.75 % of the overall test set was depressed. Similarly, the neural network with 2 units in the first hidden layer, 6 in the second, and 9 in the third reached 96.43 % specificity. Of the people it classified as experiencing suicidal ideation, 14.29% were in fact experiencing suicidal thoughts compared with 2% in the test set. If a school, a corporation, or a primary care provider had to choose some (small) subset of affiliates to reach out to, these models provide a dramatically better chance of capturing people in crisis than choosing by random chance.

Another interesting model is the SVM using 16 principal components which successfully identified every single respondent reporting suicidality. Only 4.44 % of the suicidal ideation group was correctly classified. However, if instead of trying to identify a small subset of a large sample for intervention, an organization were concerned with ruling out as many people as possible from an intervention, this model could safely rule out more than half of the group.

The models trained on the biochemical screen did not fare as well as those trained on all of the data. The neural network approach meaningfully outperformed the SVM and logistic models on this dataset. A bright spot was the neural network predicting suicidal ideation with balanced accuracy of 60.31% and accuracy of 67.5%. That result is not eye-popping, but it is clearly better than chance. Moreover, it suggests that from a standard blood test, the prosaic kind people receive at doctors' offices at annual physical exams, we can glean real information about whether someone is having suicidal thoughts. This is unlikely to shock any neuroscientist, but it may come as a surprise to lay persons, perhaps even a number of physicians.

It is worth noting that these models' reported accuracy could be a lower bound for their true accuracy. Some, if not many, of the people flagged as experiencing depression or suicidal ideation but who report no depression or suicidal ideation may not be telling the truth. It could be that the model is correctly classifying some of these shy respondents. Though it is being penalized for this behavior, this should actually be highly rewarded. These are the people who most need to be identified. Based on these NHANES data, I am unsure how to address this possible issue.

5 Conclusion and Future Work

If I continued working on this classification task, I would collect more data. I believe that my models faced a major headwind as a result of insufficient data on depressed and suicidal persons. Even with 800 observations in my all predictors training set, my models were looking for patterns in a small number of individuals. And with 400 observations in the test set, they were being evaluated on a tiny number of people. NHANES has data going back to 1999, so it would be sensible to follow [5] and include much of this. The reason I have not already done so is that, from previous experience with this dataset, I have learned that it is quite challenging and time consuming to include data from more than a few years. Variables change names within the modules while representing the same underlying data. Or they fail to change names when they represent different data. Only a few variables are consistently available from 1999 to 2018. Identifying the variables that I could use across this time period and consistency checking them is, unfortunately, on its own a project length chore.

Current approaches for identifying MDD and suicidality fall short. And for the indefinite future, barring a dramatic shift in attitudes toward psychiatric illness, many of those stricken with these brain diseases will continue not to seek treatment. Failing to identify these common, disabling, and often fatal illnesses is a substantial contributor to morbidity and mortality. Though none of the models presented in this report are prepared to revolutionize the identification of moderate MDD or suicidality, they hint at what could be achieved with even more sophisticated methods trained on far larger datasets.

References

[1] J. Angst, A. Gamma, M. Gastpar, J.-P. Lépine, J. Mendlewicz, and A. Tylee. Gender differences in depression. *European Archives of Psychiatry and Clinical Neuroscience*, 252(5):201–209, October 2002.

[2] Lisa J. Barney, Kathleen M. Griffiths, Anthony F. Jorm, and Helen Christensen. Stigma about Depression and its Impact on Help-Seeking Intentions. *Australian & New Zealand Journal of Psychiatry*, 40(1):51–54, January 2006.

[3] Kurt Kroenke, Robert L Spitzer, and Janet B W Williams. The PHQ-9. *Journal of General Internal Medicine*, 16(9):606–613, September 2001.

[4] B. H. Mulsant and M. Ganguli. Epidemiology and diagnosis of depression in late life. *The Journal of Clinical Psychiatry*, 60 Suppl 20:9–15, 1999.

[5] Jihoon Oh, Kyongsik Yun, Uri Maoz, Tae-Suk Kim, and Jeong-Ho Chae. Identifying depression in the National Health and Nutrition Examination Survey data using a deep learning algorithm. *Journal of Affective Disorders*, 257:623–631, October 2019.

[6] M. S. Reddy. Depression: The Disorder and the Burden. *Indian Journal of Psychological Medicine*, 32(1):1–2, 2010.

[7] Amita Sharma and Willem J. M. I. Verbeke. Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081). *Frontiers in Big Data*, 3, 2020.

[8] Frederick J. Zimmerman and Wayne Katon. Socioeconomic status, depression disparities, and financial strain: what lies behind the income-depression relationship? *Health Economics*, 14(12):1197–1215, 2005.