CIS 520 Spring 2019, Project Report

Modeling Suicide Rates at the Country Level

Team Members:

Monal Garg (PennKey: `mgarg`; Email: `mgarg@sas.upenn.edu`)

Amit Gupta (PennKey: `akgupta`; Email: `akgupta@seas.upenn.edu`)

Aashish Jain (PennKey: `aashj99`; Email: `aashj99@seas.upenn.edu`)

Moksh Jawa (PennKey: `moksh`; Email: `moksh@seas.upenn.edu`)

Assigned Project Mentor:

Jane Lee

Team Member Contributions:

Team Member	Contributions
Monal Garg	Past Work, Solution Approach, Machine Learning Algorithms, Visualizations, Report
Amit Gupta	Problem Formulation, Algorithms Plan for Use, Machine Learning Algorithms, Visualizations, Report
Aashish Jain	Data Cleaning, Data Stitching, Machine Learning Algorithms, Visualizations, Report
Moksh Jawa	Data Cleaning, Data Stitching, Machine Learning Algorithms, Visualizations, Report

Code Submission:

Public GitHub Repo can be found at https://github.com/akgpta/CIS520-Final-Project

Contents

1 Abstract

Suicide is a major public health concern worldwide. Because living in a different countries and therefore different environment can greatly impact one's lifestyle, we wanted to explore suicide rates over various countries. Specifically, we considered 78 countries over the five years from 2010 to 2014. For each country, we collected data on ground truth suicide rates, Youth Unemployment, Alcohol Consumption, Life Expectancy, Country Population, and Fertility Rate. We were interested in understanding how these factors could help predict suicide rates. To do this, we implemented k-Nearest Neighbors, Linear Regression, and Neural Networks. Results can be best summarized as the data fully characterize the problem. What we mean is that the countries seem to have an endogenous suicide rate, but that is determined outside of our input features. However, our features are complex enough to identify countries with this information and map it to the suicide rates. This gives an essentially overfit model which actually performs well due to high time correlation of suicide rates per country through time.

2 Introduction

2.1 Problem Motivation

Unfortunately, it is a sad reality that many individuals each year commit suicide. We in the U.S. are slowly becoming aware of this issue, but we wanted to understand suicide rates internationally. Particularly, we wanted to look at these rates as they related to common features of each country, such as GDP, unemployment, and alcohol consumption. At a high-level, we wanted to draw out any meaningful relationship between these macro features and suicide rates.

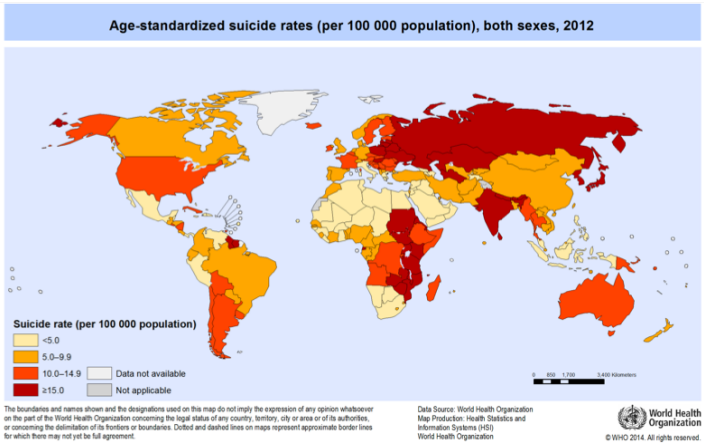

Figure 1: Map Depicting Suicide Rates (per 100,000 population)

2.2 Data Set and Source

Our project grew from an intriguing World Health Organization data set with well over 35 years of suicide rates by country broken up by demographics [1]. In an effort to extend existing analyses, we sourced global data from a variety of sources, including a FiveThirtyEight data set on alcohol consumption [2], a World Bank Data Set on Youth Unemployment [3], and another from the World Bank on population, fertility, and life expectancy [4]. A lot of the work done was with respect to time series and things like that. What we wanted to do, though, was take a look at combining data from multiple sources. Our project team sought to understand a little bit more about the intricacies of sourcing data from multiple places. This proved to be a difficult challenge.

3 Related Work

3.1 Related to Models

Among the papers that analzyed economic factors, a regression model seemed to be the popular choice. For example, a Pandey and Kaur paper used a Auto-Regressive Distributed Lag (ARDL) model [5].

Most articles concerning the relationship betweem suicide rate and alcohol consumption related trends from a biological perspective utilized relatively simple statistical analysis [6]. On the other hand, medical research that involved more complex data such as fMRI brain scans utilized algorithms such as Gaussian Naive Bayes [7].

3.2 Related to Problem

During our research, we came across multiple research papers that analyzed the correlation between suicide rates and either economical, behavioral, medical, or psychological factors. For example, the Pandey and Kaur paper investigates suicidal trends and economic determinants in an Indian population[8]. Whereas articles published in medical journals explored suicide with alochol/drug consumption, psychotic behavior, brain scans, etc [9].

However, most papers tended to focus on either the economic or behavioral/medical factors. As suicide is likely governed by multiple factors, we sought to investigate this problem through a more holistic approach by simultaneously analyzing both medical, behavioral, and economic factors. Through this application of machine learning, we hoped to gain insight on the relative role each factor plays in suicide and about the interaction of these factors.

Moreover, most papers perform a time series analysis. We provide a different approach by analyzing the data by country. Papers that do focus on certain regions usually provide analyses on one country or a specific region, whereas we analyzed global statistics.

4 Problem Formulation and Data

4.1 Data Processing Needed

Our data came in a variety of comma-separated formats. A lot of countries were labelled in differing manner and years were often a column and other times a point in a row. We used Excel Pivot Tables to make the CSVs easier to parse and standardized. Later, we used Pandas and Python to remove blanks and resolve issues with inconsistencies to finish with an aggregated set of data. Other pre-processing included filling in missing values, reconciling some country names such as (Trinidad & Tobago vs. Trinidad and Tobago), and dropping some values if a country was not universally present across all datasets. Our final product was a single CSV with 78 rows (countries) with columns for each of the 8 features. The relatively small number of countries is a testament to the variance in countries in each dataset as well as the missing values.

4.2 Inputs, Outputs

Inputs are a variety of real-valued data points tied to the country. They are described above. Most are also measured for a specific year, however some are just duplicated across all years (e.g. Alcohol Consumption). Such data points were treated as constant for a country through time. Outputs are a model, which given a countries characteristic features, can predict suicide rates. Note that the prediction is the number of suicides per 100,000 people for any given country. Thus, our problem becomes a regression problem from $\mathbb{R}^8 \to \mathbb{R}$.

4.3 Performance Measures

Given the regression nature of this problem, we used a squared loss model to measure performance. This is because with positive real valued data with explicit bounds of $[0, 10^5]$, but more reasonable empirical bounds of around $[0, 10^3]$. The following image is a representation of squared loss, but in a one dimensional feature space. [10]

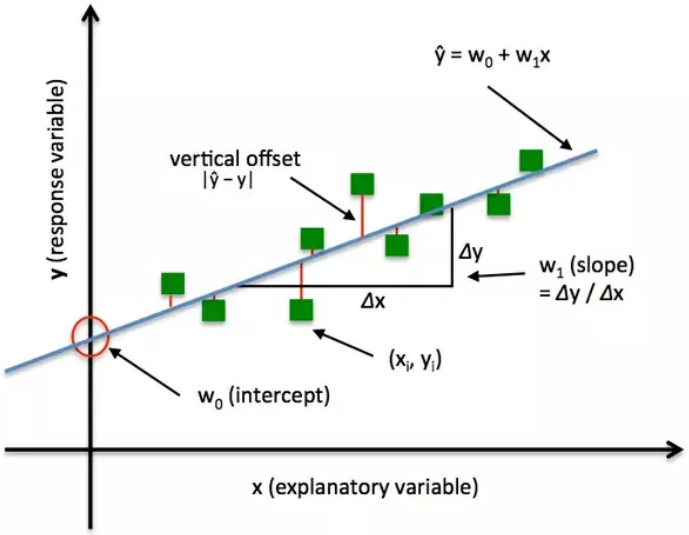

Figure 2: Squared Loss Model

5 Methods and Results

5.1 Experiment Set Up

Altogether, we had five years worth of data on 78 countries. We maintained the same training data, test data, and folds when implementing all three models. The data from 2014 data was reserved as the test data and we trained our models on the data from 2010 to 2013. Four folds were set up that such that each one reserved a different year from 2010 to 2013 to test with.

5.2 k Nearest-Neighbors

We first implemented a k-nearest neighbor algorithm using Euclidean distance as we suspected suicide rates between similar countries would be comparable. We cross validated across the following values of $k = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$ and found that $k = 1$ produced the best model with a cross validation squared error of $4.697 * 10^4$, training error of $8.160 * 10^6$ and test error of $1.720 * 10^6$. Although this error appears to be high, we remind the reader that this squared error is taken

over 78 test data points per fold and each of the suicide rate values can be as large as 10000. In fact, if you average this cross validation error over these points and square root you get an average error of $\sqrt{\frac{4.697*10^4}{78}} \approx 24.5$ on each sample.

Figure 3 shows the results when cross-validating over different k values. For now, focus only on the orange $q = 2$ line, which shows square errors when run with Euclidean distance. Here we have omitted the training and test error to highlight an interesting trend in cross-validation error. Recall that our each of our folds have data of from each country of 3 years. Say we are looking for the nearest neighbors for country x. We suspect that the error is very low for k values of $1, 2, 3$ because the model is utilizing the country x's 3 data points from different years to predict suicide rates for the test year. Errors then pick up for values of k greater than 3 because the model is forced to use the data from countries other than country x, which may not be as informative when making decision about suicide rates for country x.

To take this a step further, we also built a nearest neighbor model using Minkowski's distance, a more general distance formula, shown below.

$$D(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^q \right)^{1/q}$$

We cross-validated across the following values of $q = [1, 2, 3, 4, 5]$ and for each of these q values, found the best k value. Interestingly, we found for each value of q, $k = 1$ produced the minimum squared error. Moreover, the same phenomenon of small errors for k values from one to three and relatively larger errors for every subsequent value of k was observed for all values of q, as can be seen in Figure 3. Overall, $k = 1$ and $q = 2$ (which is equivalent to Euclidean distance) produced the best model.

From these results, we began to suspect that nearest neighbors may be memorizing the countries suicide history. Indeed when we looked further into the data, we realized that the countries very distinct, causing the countries to be most similar to themselves.

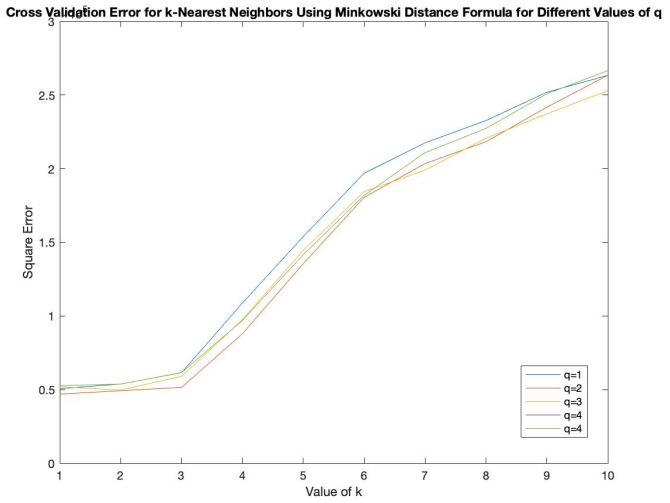

Figure 3: Cross Validation Error for k-Nearest Neighbors Using Minkowski Distance.
Note that $q = 2$ refers to a model built off of a Euclidean distance formula.

5.3 Neural Network

We ran a neural network model with one hidden layer and a single output neuron to try to accurately classify suicide rates. To accurately determine the number of hidden nodes that would produce the best model, we cross validated over the number of nodes: {1, 5, 10, 15, 25, 50, 100}.

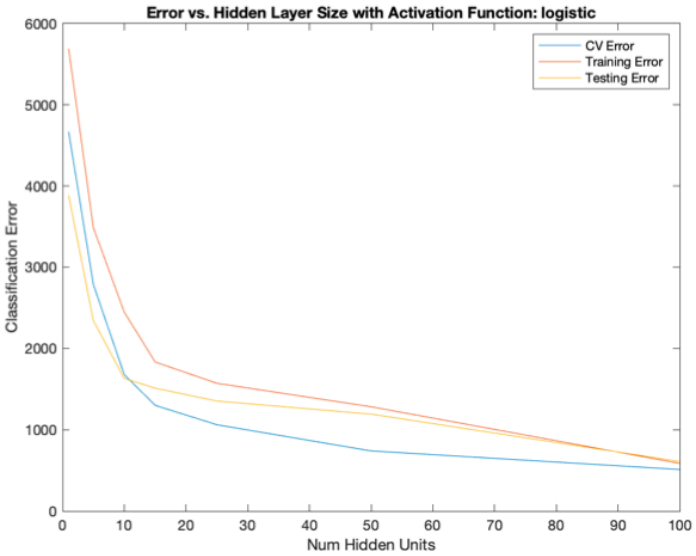

Figure 4: Classification Error of Neural Network Model: Logistic

However, we observed that our model performed better and

better as the number of nodes increased. This furthers the hypothesis that our data favors a complex model because it a complex model can over fit the data and essentially functions as a memorizer.

We found that a learning rate of 0.001 and 5000 iterations was more than enough to fit the training data. Increasing the learning rate and iterations further only resulted in arithmetic underflow.

5.4 Linear Regression

As our previous models grossly over fit the data, we thought implementing a regularized linear regression may provide us more insight on how to deal with the data. However, we quickly realized that this was not the case.

We found that $\lambda = 10^{-2}$ worked the best and gave a cross validation error of $3.607 * 10^5$, training error of $1.431 * 10^6$, and test error of $2.524 * 10^5$. The low λ value and incredibly high errors suggests that the data did not respond well to regularization. In fact, the data doesn't seem well suited to a linear model. This likely stems from the fact that the suicide rates are not particularly correlated with our input features.

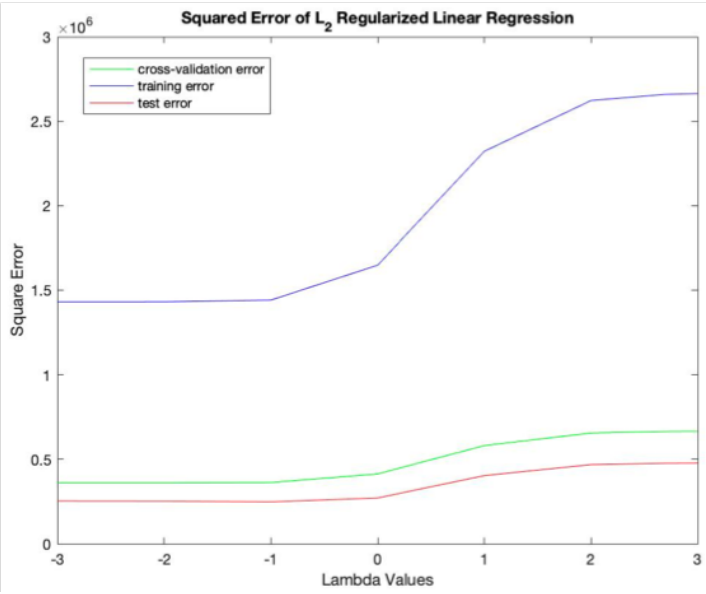

Figure 5: Linear Regression Model

5.5 Overall Results

We ended up with memorizers on neural net and nearest neighbors. Meaning that our models would be particularly well suited for something like a geography bee.

The best predictor of suicide rates is the country label itself. Suicide rates in Ghana in 2013 are going to look like suicide rates in Ghana in 2012 most. Modeling this regression outcome as independent draws disregards the time series data and correlation of examples. As such, nearest neighbors performs exceptionally well, since it just recognized the country through these 8 input features.

6 Discussion and Reflections

We see a lot of what would canonically be overfitting in the data. However, in the context of this problem, overfitting is not necessarily the worst. We have such complete data that we can actually begin to drive at the underlying distribution at a per instance level. Note, when we say instance here, we are referring to a unique country (not a specific $\mathbb{R}^8$ vector in the input space). This means our classifiers actually easily approach a Bayes Optimal Classifier.

We recall the handwriting recognition problem from class. Consider fitting a unique model to an individual with fairly comprehensive handwriting samples. Handwriting for an individual evolves minimally over time and there are only 26 letters. Overfitting and understanding what exactly these 26 letters look like in some sort of training set for the individual means that predictions of future handwriting, *for that individual*, will be fairly reliable.

This project was certainly an adventure in machine learning. Let us recall some of the lines from what we set out to do. Note, these were established a priori in our proposal: *"at a high-level we wanted to draw out any meaningful relationship between these macro features and suicide rates"* and *"we hope to gain insight on the relative role each factor plays in suicide and about the interaction of these factors"*. We were thinking that a model would illimuniate which features were coupled with suicide rates. As the project continued though, as noted in our Results section, models were using the features to identify individual countries. We are careful in this report to not conflate correlation with causation and, as such, have left out implications of this country-recognizer model built.

Reflecting on our goal, it seems the problem we have chosen to tackle doesn't make sense from a machine learning perspective. Sure, the factors we have chosen can be used to generate

a model which fairly reliably outputs suicide rates. However, this correlation should not be taken to imply causation or anything close to that. From a model story perspective, suicide rates are affected by many factors exogenous to our world of data but endogenous to a country.

When you consider some sort of underlying sampling distribution assumptions on the data, it doesn't make sense that suicide rates for countries on different continents with different socioeconomic statuses are coming from one distribution. Although this seems to say that this project wasn't an effective learning opportunity, we, as a team, feel quite the opposite.

There are a finite number (195) of countries in the world. As such, worries about reasonable time computation, tractability, and other aspects of problems centered on larger data (think users of Netflix or game states) does not factor in here. Had we had more experience at the start of this project, we could have foreseen the uniqueness of countries leading to a model which could basically be a geography bee champion. However, we feel that the machine learning exercises carried out during the course of this project deepened our understanding of course algorithms, implications, and relevant considerations.

7 Next Steps

Although we discovered that are models largely functioned as memorizers, we believe there are still possible approaches that have not been exhausted. For example, as suicide rates are highly time correlated, we could try implementing a Auto-Regressive Distributed Lag (ARDL) model, as a related paper had tried [11].

8 Acknowledgements

First and foremost, we would like to acknowledge our project mentor, Jane Lee, for being very helpful throughout the project and Professor Agarwal for a fantastic class. Additionally, we acknowledge the Kaggle users who provided the datasets.

9 References

Notes

[1]https://www.kaggle.com/szamil/who-suicide-statistics

[2]https://www.kaggle.com/fivethirtyeight/fivethirtyeight-alcohol-consumption-dataset

[3]https://www.kaggle.com/sovannt/world-bank-youth-unemployment

[4]https://www.kaggle.com/gemartin/world-bank-data-1960-to-2016

[5]https://crawford.anu.edu.au/acde/asarc/pdf/papers/2009/WP2009/_08.pdf

[6]https://www.ncbi.nlm.nih.gov/pubmed/9229027

[7]https://nocklab.fas.harvard.edu/files/nocklab/files/just_2017_machlearn_suicide_emotion_youth.pdf

[8]https://crawford.anu.edu.au/acde/asarc/pdf/papers/2009/WP2009/_08.pdf

[9]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2872355/

[10]https://blog.algorithmia.com/introduction-to-loss-functions/.

[11]https://www.projectguru.in/publications/auto-regressive-distributed-lag-model-ardl/

10 Appendix

Dataset Summary Statistics:

Statistic	Average liters/person/year
Min	0
Max	14.4
Mean	4.71
Median	4.2
Standard Deviation	3.77

Table 0: Alcohol Consumption by Country

	Youth Unemployed				
Statistic	% 2010	% 2011	% 2012	% 2013	% 2014
Min	0.6999	0.6999	0.5	0.6999	0.6999
Max	57.2	57.1	61.7	58	57.9
Mean	17.89	17.9	18.15	18.1	17.94
Median	14.9	14.52	14.4	14.1	14.12
Standard Deviation	10.54	10.88	11.43	11.67	11.55

Table 1: Youth Unemployment by Country

	Fertility Rates				
Statistic	% 2010	% 2011	% 2012	% 2013	% 2014
Min	1.06	1.11	1.16	1.13	1.21
Max	7.49	7.46	7.42	7.38	7.34
Mean	2.91	2.88	2.84	2.81	2.79
Median	2.41	2.37	2.34	2.34	2.33
Standard Deviation	1.45	1.42	1.39	1.37	1.34

Table 2: Fertility Rates by Country

	Life Expectancy				
Statistic	% 2010	% 2011	% 2012	% 2013	% 2014
Min	47.56	48.284	49.041	49.825	50.621
Max	82.9780488	83.4219512	85.4170732	83.8317073	83.9804878
Median	72.2783848	72.4719847	72.657	72.786	72.9707317
Standard Deviation	8.34929356	8.19102272	8.04819537	7.89104622	7.80484223

Table 3: Life Expectancy by Country