Machine Learning

Suicide Prediction

By Jonathan Lei

June 13, 2019

Table Of Contents

Abstract

In the last 45 years, suicide rates have increased by 60% worldwide. In the United States alone, a suicide is committed every twelve minutes, and it is the 10th leading cause of death in the United States. What factors contribute to suicide rates, and are these rates able to be predicted? My hypothesis is that a few variables are strongly correlated with suicide rates, which can then be predicted. This project uses Python machine learning algorithms to create a model from thousands of data points and proceeds to test the machine learning model. It analyzes the correlation between variables and suicide rates. In addition, the program creates predictions that are often close to the true results. While the results are not perfect, the program gives an insightful view into the statistics behind suicides.

Introduction

Problem

Suicides are becoming ever more common around the world. In the past 45 years, rates have increased by 60% worldwide. Inside the United States, every single state except for Nevada reported an increase in suicide rates from 1999 to 2016, with half reporting an increase of over 30%. The Center for Disease Control states, “Suicide is rarely caused by a single factor.” Nevertheless, one may find correlations that relate to suicide rates, and an understanding of these correlations may help slow the trend.

Hypothesis

Certain factors correlate with the likeliness of a suicide, and these may be used to predict suicide rates for one’s demographic. These factors include age, generation, GDP per capita, Human Development Index, gender, year, and country.

Purpose

This project aims to analyze suicides rates per 100 thousand people by scanning individuals’ characteristics. In addition, it attempts to combine all of these factors to predict the self-harm rate per 100 thousand people in the same demographic as a given individual. By doing this, social workers will hopefully be able to pinpoint and help individuals that the program estimates to have a higher-than-normal rate of suicide.

Methodology

Independent Variables:

- Country: Country of the citizen. Every country can have different awareness and prevention programs to limit suicides, so this should help determine the likeliness of a suicide.
- Year: Year of the suicide in question. Huge events that occur in a single year might impact self-harm rates. The program should also be able to see general trends and make predictions accordingly.
- Sex: Gender of the suicide victim. There may be slight differences in suicide rates depending on the gender of the victim.
- Age: Age of the victim during suicide. There may be varying self-harm rates between people of different ages.
- HDI: Human Development Index. This is a number used to assess the development of a country, not economic growth alone. Factors that impact an HDI include but are not limited to: the life expectancy at birth, government policy choices, and the expected years of schooling. The HDI of a country should have an impact on suicide rates.
- GDP per capita. GDP per capita is the total trade within a country. Thus, it correlates with the wealth of the average citizen and should correlate with peoples' general happiness as well as suicide rates.
- Generation: Generation of the victim. Examples would be: Generation X, Baby Boomers, etc. These are converted into a numerical format that the program is able to analyze. People who

are born at different times may have different suicide likeliness, even when other factors are the same.

Dependent Variable

- Suicide Rate Per 100K: Number of suicides per 100k people within the same demographics. The machine learning algorithm will attempt to see how much each independent variable affects the suicide rate per 100 thousand people within the same demographics.

Data Source

Russell Yates with the username, “Rusty,” on the data collection website, “Kaggle,” compiled a list of 27,800 suicides from 1985 to 2016. This list’s information is derived from sources including the United Nations Development Program (data from 2018), the World Bank (data from 2018), and the World Health Organization (data from 2018).

Decision Trees

Decision trees use a tree-like model to show possibilities and make decisions by going from branch to branch. I use the DecisionTreeRegressor to show the importance of some factors to analyze variable reliance.

RandomForestRegressor

RandomForestRegressor is the machine learning algorithm used in this project to estimate suicide rates based on a variety of independent factors. RandomForestRegressor uses a technique called Bootstrap Aggregation to train different trees. The results are combined so that not a single decision tree is relied upon.

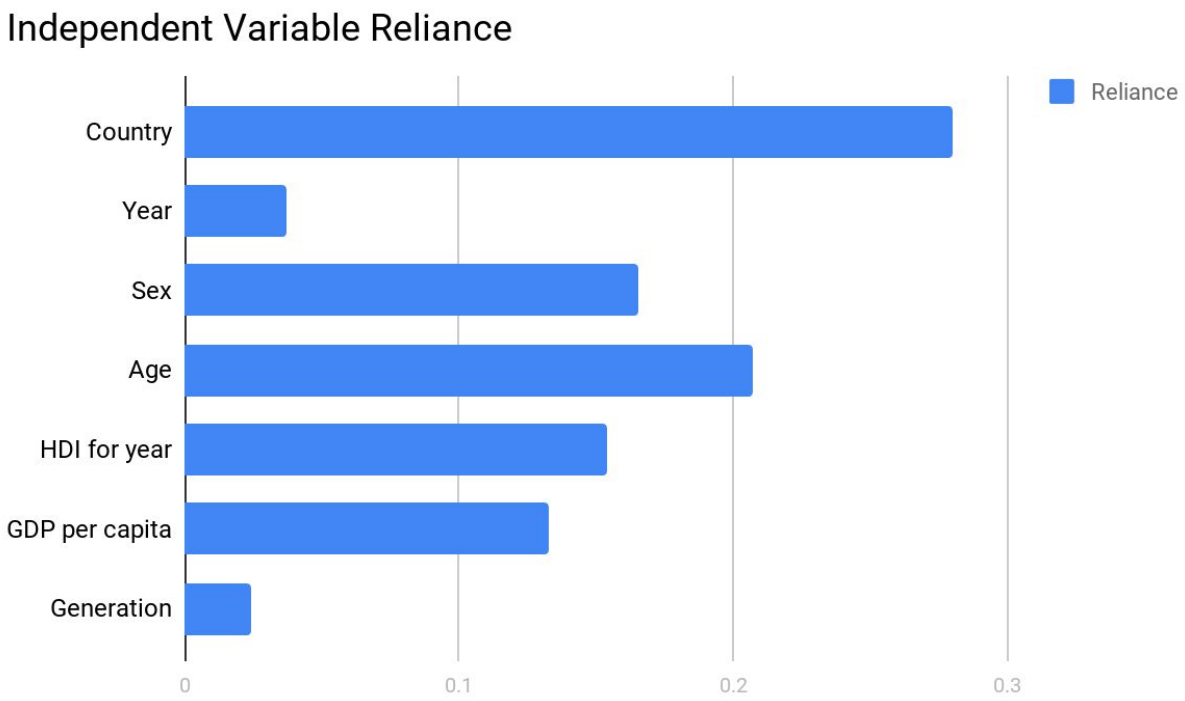
Steps

1. Import Python libraries - This project uses Python machine learning algorithms that are not automatically imported, so it must do so.
2. Read CSV file - The program extracts data from the following data set: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016. This data set covers tens of thousands of previous suicides. The Python program reads the CSV file by parsing the file into a DataFrame.
3. Classify non-numerical numbers into an integer - The program uses the LabelEncoder from the SKLearn library to encode and classify data. For example, “Baby Boomer” would be changed to “1,” while “Generation X” would be changed to “2.” Because Python cannot handle text, the program converts it into a numerical form first.
4. Drop null inputs - The data set contains records of some suicides that do not include all of the information. It drops the suicides that lack any independent variables.
5. Data processing
    a. The program uses the RandomForestRegressor to create its own model. When given the independent variables, it predicts the dependent variable.
    b. Other regressors have not worked as well. These include XXXXXXXXXXXXXXXXXXXXXXXXXXXX
6. Decision Tree
    a. The program uses a decision tree to measure the importance of each independent variable.

Results

Regression Prediction Score (Suicide Prediction):

0.9502331822986496



Variable Reliance:

Country	0.28017803
Year	0.03679934
Sex	0.16512986
Age	0.20688536
HDI for year	0.15415589
GDP per capita	0.13295611

Generation	0.02389541

Analysis

Regression Prediction Score (Suicide Prediction):

The program creates a model/hypothesis linking the independent variables to the suicide rates per 100 thousand people within the victim in question's demographic group. The program feeds every data point back into the algorithm. The algorithm uses its trained model/hypothesis to create its prediction for the suicide rates per 100 thousand people, which is compared to the real data. In a regression scenario, one aims for the highest possible regression score out of one, and in this case, the program gives itself a score of 0.95, which shows that this is relatively close to the real suicide rates per 100 thousand people for the demographics of the victim. Based on these results, one determines that the algorithm may be used in real life to predict suicide rates.

Variable Reliance:

Based on the results in the table, one can see that the program attributed the country that the victim was in to have nearly a 30% impact on the suicide rate per 100 thousand people within the same demographic. This is not surprising, given that a country with proactive suicide prevention would have lower overall rates than a country without proactive suicide prevention. We can also see that the program only attributed 3% of cases to have the year affect the suicide rate, which is surprising. Given the upward trend of suicides, one would expect a higher suicide per 100 thousand people ratio. While the age of the victim, human development index, and GDP per capita played an expectedly high role in determining the dependent variable, the sex of the victim was unexpectedly attributed to over fifteen percent of potential suicide rates.

Discussion

Future

In the future, the project could be improved by implementing additional values. These could include government aid systems that are active in the area to see if specific programs have worked, although this would require a brand-new dataset. Currently, the program uses about 8,500 rows of values, using more of the 27,800 rows could enlarge the sample size, giving the project more credibility. The reason why a majority of the data points are not used is because rows that contain nonexistent values are removed, but it would be better to still use any data, as more data would help the model more. From now on, improving the model requires a better data set, as many code enhances have already been implemented.

Application

After implementing numerous improvements, an implementation of the algorithm in the real world could be to aid social workers; they could put in data to a calculator to find where people would most likely commit suicide and help citizens there.

Conclusion

In conclusion, this project's findings show that a few factors can help predict suicides, as some factors correlate with the likeliness of one to commit suicide, and these may be used to predict rates for one's demographic. These factors include age, generation, GDP per capita, Human Development Index, gender, year, and country. Thus, the hypothesis is supported. It is important to note that the regression score was trained by using the entire data set, and was also tested using the entire data set. Thus, it was only being tested on what was being trained.

Nevertheless, the model was tested, and the program has no way of remembering sets of data. Thus, the program performed better than it would have if random data that it wasn't trained for was fed, but the predictions still show a strong correlation of specific factors to suicide rates nonetheless. Replicating the outputs are easy, as steps are detailed in the "Steps" section. All in all, suicides are more common than they should be, and this project has proved that specific factors correlate to the suicide rate of a demographic and that suicides can be predicted.

Acknowledgments

1. Danny from KTBYTE for teaching me about machine learning and Python and helping with the project
2. Michael Schultz for helping me fix a grammatical issues
3. Russell Yates for creating the "Suicide Rates Overview 1985 to 2016" dataset with over 25,000 data points

Works Cited

"Human Development Reports." *Human Development Index (HDI) | Human Development Reports*, hdr.undp.org/en/content/human-development-index-hdi.

Rusty. "Suicide Rates Overview 1985 to 2016." *Kaggle*, 1 Dec. 2018, www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016.

"Suicide Statistics." *AFSP*, 16 Apr. 2019, afsp.org/about-suicide/suicide-statistics/.

"Suicide Statistics and Facts." *SAVE*, save.org/about-suicide/suicide-facts/.

"Suicide Rates Rising across the U.S. | CDC Online Newsroom | CDC." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, www.cdc.gov/media/releases/2018/p0607-suicide-prevention.html.

Personal notes:

1. Good Classification / Regression Score
2. Writing of your report - Objective of the report is Science Fair / Publish on Github / Personal Blog/ Public blog
    a. Introduction
        Explain the problem 3 paragraphs
    b. Methodology
        Explain the data and your approach (You can use PCA here or Feature Selection / Engineering / Scaling Normalizing)
    c. Results
        Plots/Graphs/Accuracy Score under different parameters etc
    d. Conclusions
        How it helps solve that problem and how you can improve this in the future.