Machine Learning Engineer Nanodegree
Capstone Project Proposal

By,
Aakar Mutha

Domain Addressed

Suicide is one of the major causes of deaths in the world. It is estimated that around 20 Million people attempt to commit suicide every year. People commit suicide because of reasons like Mental disorders, including depression, bipolar disorder and many more. The people who are more prone to commit suicide are in the age groups of 15 – 30 years and 70 + years. The dataset I am using is posted in Kaggle by the World Health Organization.

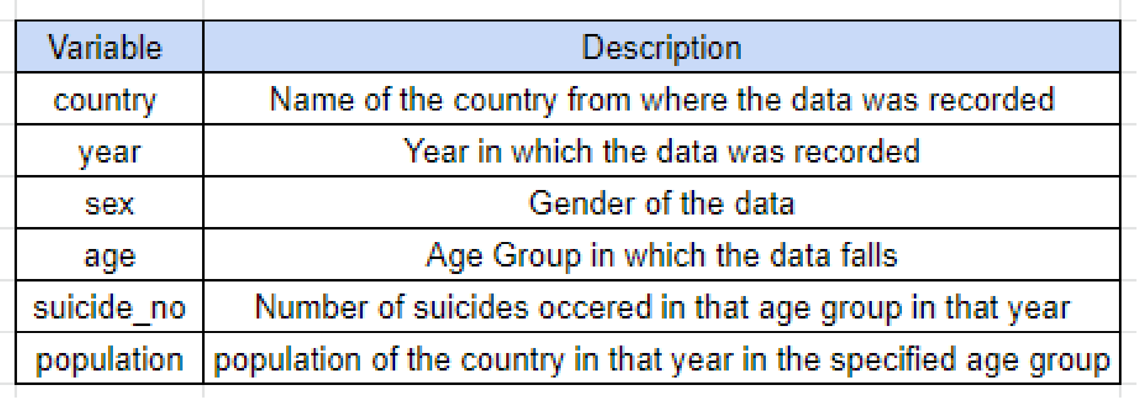
The motivation for this project was to combat this problem and to predict approximately the number of suicides that can take place in a given country in a given year according to the historical data. This prediction will help the government as well as the NGOs to spread awareness.

Problem Statement

The goal is to create a machine learning model which will give us the number of suicides that occur in a particular age group in a country. This model will help the government and NGOs to better organise campaigns for suicide prevention. This will also help to organize campaigns that are targeted to a particular age group.

Datasets and Inputs

The dataset is provided by Szamil and is named as WHO Suicide Statistics. It has 43,776 entries in each of its 6 columns.
https://www.kaggle.com/szamil/who-suicide-statistics



Variable	Description
country	Name of the country from where the data was recorded
year	Year in which the data was recorded
sex	Gender of the data
age	Age Group in which the data falls
suicide_no	Number of suicides occered in that age group in that year
population	population of the country in that year in the specified age group

https://docs.google.com/spreadsheets/d/1px2CKqXjjgkzWolwIgsf-QKvG0IfSAI4H6TqLf3NTbU/edit?usp=sharing

The input to the model will be country, year, sex, age and population. This will all be passed through Label encoder so that it is converted to categorical variables.

The dataset consists 141 different countries.

The dataset is for 38 years (1979 – 2016) of historical data.

The dataset is divided in 6 age groups.

Solution Statement

The solution will utilize the fact that the historical data can be used correlate and predict the number of suicides that may occur. This will be helpful for us to use supervised learning models and predict the suicidal behaviour of people in the coming years. I will be using the 5 columns namely country, year, sex, age and population as the input to the model. I will be testing various regression techniques like SVR, Linear Regression, Polynomial Regression, Decision Trees, etc.

I will also evaluate other bagging and boosting models to determine the best model. I will also compare all the results obtained from all the models.

Benchmark Model

My benchmark model is Decision Tree Regression. This model got `MSE : 187271.9595 , RMSE : 432.7493033 and r2_score : 0.7534.`

My goal is to improve the r2 score along with decreasing the MSE and RMSE.

Evaluation Matrix

The goal is to maximize the r2 score and minimize the MSE and RMSE values. This will allow us to account the variability in the data. Minimizing the RMSE and MSE will allow us to know how well our model has performed.

Product Design

1) The first step of any machine learning project is to define the problem. This proposal reflects the first part of this task and after acceptance the first step is to gather the data and access it.
2) Second step is to analyse and prepare the data.

   It includes the following steps:

   i. Sorting the data in ascending order based on the year.
   ii. Dropping the rows which have any null values.
   iii. Changing data types if necessary.
   iv. Encoding the features of categorical data type using the label encoder.
   v. Normalizing the data
   vi. Etc.
3) Step three is choosing the model. I will experiment with many types of models and will choose the one that best suits my data.
4) Step four is to calculate the accuracy of all the models and then comparing them by representing them with a bar plot so that the best one can be chosen as the final model
5) Step five is hyperparameter tuning. In this step I will further try to improve the r2 score of the model and get the maximum value possible.

I will be using a jupyter notebook with python 3.8 to document all the steps taken.