Suicide predictions on an international level

Mathias Løkkebø Øvreseth
Computer engineering
Norwegian University of Science and Technology
Ålesund, Norway
mathilov@stud.ntnu.no

Michal Åsebø Berg
Computer engineering
Norwegian University of Science and Technology
Ålesund, Norway
michalb@stud.ntnu.no

Contents

ABSTRACT

More than 700 000 people die by suicide worldwide each year **(World Health Organization 2021)**. A better understanding of what factors that affect the number of suicides can help prevent groups of people that are prone to suicide. In this article we look at suicide trends worldwide and then develop a machine learning suicide prediction model, using publicly available data from the period 1987-2016. An eXtreme Gradient Boosting (XGBoost) based machine learning model **(XGBoost 2021)** and RandomForestRegressor **(Sciki-learn 2022)** was used with six features to predict suicides rates for the scandinavian countries (Norway, Sweden and Denmark. XGBoostRegressor provided a R^2 value of 0.96, meanwhile RandomForestRegressor scored an R^2 score of 0.973. Shapley Additive exPlanations were used to acquire the impact of the six features. These features were as follows: population, age group, gender, year, country, and Gross Domestic Product. These six features were then used to create a Suicide Vulnerability Index for the 101 countries.

INTRODUCTION

Suicides were responsible for 1.3% of all deaths worldwide. It has become one of the leading causes of death worldwide, ranking it the 17th leading cause of death in 2019 (World Health Organization 2021). Cause of death is labeled as a suicide if the person were intentionally causing self-death. The model developed can predict suicide rates on an international level. This could be proven useful in effectively allocating resources and providing necessary help where needed. The SVI developed in this paper could be extended further to create an SVI for each country instead of internationally to get a better understanding of which countries are more prone to higher suicide rates.

METHODS

Data description

This work used publicly available data and compares socio-economic information with suicides by country in order to predict suicide rates on an international level. The data set contains suicide data and country characteristics for 101 different countries from year 1987 to 2016. This section will provide a detailed description of the data.

The country level suicide data used in this study is a compiled dataset pulled from four other datasets linked by time and place and was built to find signals correlated to increased suicide rates among different cohorts globally.

The data set describe country level features such as country, year, gender, age group, population and GDP per capita, with the target feature being number of suicides (United Nations 2022). Some input features provided in the data set were removed, such as "country year", "generation" due to it being duplicate of data already present in the data set, as well as human development index, due to it containing null values. The input feature "suicides per 100 000 population" were also dropped from the data set. The figure below displays the features used for our machine learning model.

	country	year	gender	age_group	suicide_number	population	gdp_per_capita
0	Sweden	1987	male	35-54 years	394	1124000	22813
1	Sweden	1988	male	35-54 years	394	1140400	25731
2	Sweden	1993	male	35-54 years	385	1214000	25880
3	Sweden	1989	male	35-54 years	381	1157700	26978
4	Sweden	1991	male	35-54 years	375	1189400	33623

Figure 1 The input features used in our model

The age groups were divided into six different groups; 15 – 24 years, 35-54 years, 75 years and above, 25 – 34 years, 55 – 74 years and 5 – 14 years. Each group is represented with 164 unique instances.

Prediction model description

This study was performed by finding a correlation between the number of suicides and the six input features provided in the dataset. Due to the fact our models would predict a numerical value, being number of suicides, it was modeled as a Regression based Predictive Model problem. Thus, by using pipelines, the chosen models were the XGBRegressor and RandomForestRegressor for our predictions. Each model was implemented through the use of pipelines. SHAP, which is used to reverse-engineer the output of any predictive algorithm, were integrated into the solution and seamlessly computed SHAP values used to evaluate the two models (Lundernberg 2018).

The dataset used in this study were split into a training set and a testing set with a ratio 80:20 by the train_test_split function provided by sklearn.

Feature importance

The various input features in the dataset would naturally not be contributing an equal amount to determine suicide numbers. Thus, it is a necessity to identify the determining factors in our model. To accomplish this and evaluate our model, we used SHAP and its calculated values, to determine which factors made the most impact. The figure below displays the weight contribution of each feature.

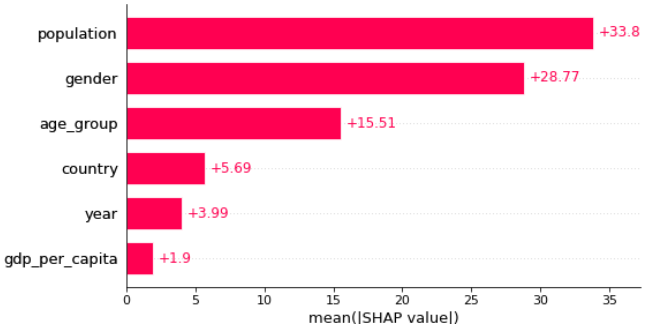

Figure 2 Plot of the contribution of each input feature

Python, alongside its libraries, were used for both the development of the model and the analysis of the model.

RESULTS

Suicide prediction

XGBoost Regressor and Random Forest Regressor was used to implement the machine learning Suicide prediction in python and the model performance was measured using the R^2 scores. An R^2 value of 0.96 for the XGBoost regressor and 0.97 for the Random Forest Regressor was calculated using the six unique features.

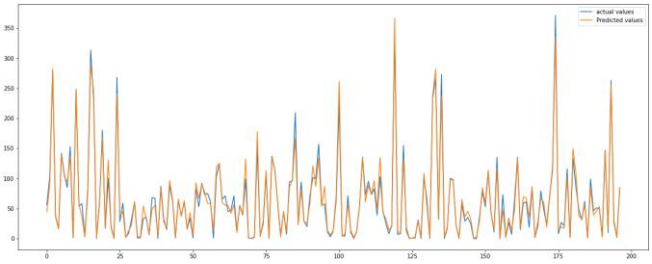

Figure 3 plot of actual suicide values versus predicted values generated by the machine leaning model

From Figure 3 we can see the predicted suicide numbers (marked in orange) and the actual suicide numbers (marked in blue). The Y-axis represents the number of suicides. The X-axis represents a row in the dataset.

Feature importance

To predict suicides, six features were used. To identify the impact of each feature, SHAP feature importance values were used. The values generated by SHAP represents the importance by each feature on a log-odds scale (logarithm of the ratio of high suicides to low suicide). Figure 4 shows the SHAP value plot for all the features used within the training dataset.

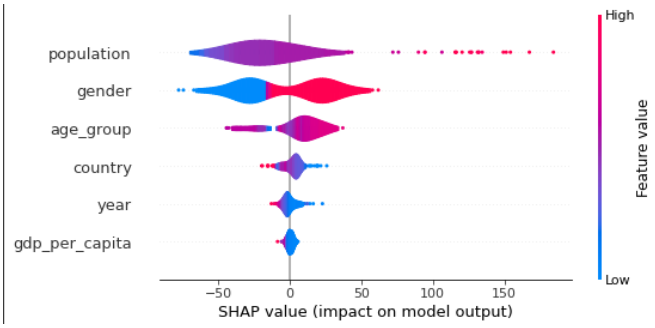

Figure 4 SHAP summary plot reveals the relative impact of each feature on the prediction results.

In figure 4 every single feature is listed and arranged based on their significance. From the figure 4 we can see population had the biggest impact on predicting suicides, followed by gender, age group, country, Gross Domestic Product and year. The X axis shows the SHAP values corresponding to each feature. The color of the dots is red if the corresponding feature value was high, and blue if it was low.

SVI

The six unique features were used creating the SVI. The SVI calculated measures the resilience towards suicides. Therefore, a high SVI score interprets a tendency to higher suicide rates.

DISCUSSION

Suicide is a substantial concern and one of the leading sources of death worldwide. This makes it essential to employ a targeted suicide control and prevention efforts. This work predicted suicide rates between 1987 to 2016, and the model can be used to predict suicides in upcoming years. Furthermore, an SVI can be generated for each country to get a more accurate representation of which country are more prone to suicide, and which features are the leading factors. The six features used to predict suicides were proven to have a high level of correlation according to the R^2 score generated.

The generated SHAP values provided the computed importance of each feature. Population and gender had the biggest impact on the suicide prediction. A country with a smaller population will naturally have fewer cases of suicide compared to a country with a high population. Thus, population will be a driving factor in this model. The other driving factor were proved to be gender, showing that more men tend to commit suicide than women. The figure below displays the correlation between some of the features used in this study.

	year	suicide_number	population	gdp_per_capita
year	1.000000	-0.088172	0.081589	0.787501
suicide_number	-0.088172	1.000000	0.706751	-0.156210
population	0.081589	0.706751	1.000000	-0.100961
gdp_per_capita	0.787501	-0.156210	-0.100961	1.000000

Figure 5 Correlation between features

REFERENCES

Lundernberg, Scott. 2018. *https://shap.readthedocs.io.* january 15. Accessed september 23, 2022. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html.

Sciki-learn. 2022. *scikit-learn.org.* september 25. Accessed october 5, 2022. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

United Nations. 2022. *hdr.undp.org.* september 23. Accessed september 2023, 2022. https://hdr.undp.org/data-center/country-insights#/ranks.

World Health Organization. 2021. *who.int.* june 16. Accessed September 29, 2022. https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data.

XGBoost. 2021. *https://xgboost.readthedocs.io.* january 15. Accessed September 23, 2022. https://xgboost.readthedocs.io/en/stable/.