Predicting Suicide rates of Countries: Verification of country scale factor contribution

Jonguk Jeon
ukj0806@kaist.ac.kr
KAIST
Daejeon, South Korea

Seungjun Chung
s.j.chung@kaist.ac.kr
KAIST
Daejeon, South Korea

Abstract

We propose a new, socially-impactful model that predicts suicide rate of countries from a given set of features. We present wide domain of 54 features highly contributing to suicide rate, and present the best predicting model to address ignorance of such. Our work is novel in that the features considered were country scale; not individual scale. Our model successfully predicts suicide rates of countries with few or no error. We built additional models with far less number of features for practical usage in the future. This work can be later used to reduce suicide rate in a country scale.

Software and data are available online at: https://github.com/33577/suicideRatePrediction

[This paper is for a CS492D course project.]

1 INTRODUCTION

As the world develops, many people are suffering isolation from the society. This results in serious mental, physical health symptoms including major depression disorder or social anxiety disorder. Many interdisciplinary researches try to address such problem and come up with plausible solutions. However, suicide rate has been increasing all over the world and doesn't seem to stop its walk yet no good approach has come up.

Thanks to development of machine learning technique, world are now able to select appropriate technique to build a model with desirable functionality. Lots of studies have been taking advantage of the technology which provided a great skeleton of this research. We propose to make a prediction on suicide rates of countries and verify its feature importance. This can be formulated as four-fold tasks:

1. Find the most contributing features.
2. Build a regression model for prediction of suicide rate.
3. Build an additional and practical regression model with the reasonable and optimal number of features.
4. For each of the most contributing features verify relationship between suicide rate and each single feature.

2 RELATED WORK

Recently, several studies on suicide combined with machine learning technique has come up. Most of them mainly focused on that of individual's suicide. For instance, studies on self-injurious thoughts[2] or predicting individuals with suicide ideation[3] are such. However, there has been no study on suicide rate of country scale. We applied our novel approach of focusing on a country scale factors upon the previous studies.
Main idea of methodology was borrowed from *Nowcasting: Towards Real-time GDP Prediction*[4] and *Step by Step House Prices Prediction*[9], which left a trace of building regression models with several widely used regression algorithms such as XGBoost, LGBM, etc. Upon their previous research, we conducted with a larger number of regression algorithm technique.

3 DATA

We used three datasets from *Kaggle* and *Economist Intelligence Unit* from *Economist.* These datasets are 'suicide rates overview 1985 to 2016', 'countries of the world' and 'environmental variables for world countries'. Description for each dataset is as follows:

1. 'Suicide Rates Overview 1985 to 2016' is about suicide rates over countries.
2. 'Countries of the World' have 19 information include population, area size, infant mortality over countries.
3. 'Environmental variables for world countries' have 27 information include travel time to city, precipitation, percentage of country covered by cropland or trees.
4. EIU is about index of politics, democracy, economy over countries.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$100.00
https://doi.org/10.1145/1122445.1122456

Since the data of 'Countries of the World' was from 2010, we use 2010 year data of 'suicide rates overview 1985 to 2016' and EIU of 2010. The resulting number of features is 54. For those with datatype of non-int or non-float, we pre-processed it to ensure every data is arithmetically computable. Note that for the feature 'sex', male is represented as 1, female is represented as 2 in our data.
The results of correlation between features and suicide rate are as follows. Table 1 shows the five most positively correlated features. For instance, higher your age is, higher possibility of committing suicide. Table 2 shows the four most negatively correlated features. For instance, men tend to commit suicide more than female.

	Age	Lit.	D.r	C.c	T.c.c
Correlation	0.34	0.24	0.23	0.16	0.15

Table 1: Positive correlation between features and suicide rate. Shown only top eleven. 'Lit.' denotes 'Literacy', 'D.r' denotes 'Death rate', 'C.c' denotes 'Cropland cover', 'T.c.c' denotes 'Tree canopy cover.'

	Sex	T.d.q	Rain seasonality	Birth rate
Correlation	-0.42	-0.27	-0.26	-0.23

Table 2: Negative correlation between features and suicide rate. Shown only top fourteen. 'T.d.q' denotes 'Temperature of driest quarter.'

Besides age and sex, the magnitudes of correlation is less than 0.2. On the other hand for age and sex, they shows those of near 0.4 which means there is few or no linear relationship except age and sex.

4 METHOD

We made eight models built upon Lasso, ElasticNet, KernelRidge, GradientBoosting, XGBoost, Light GBM, SVR, Random Forest respectively. Then computed mean square error in prediction to get the best working model. The best performing four algorithms were XGB, LGB, GB, RF which denotes XGBoost, Light GBM, GBoost, Random Forest each. We discarded the other four.

	XGB	LGB	GB	RF
MSE	47.37	44.97	28.23	20.84
Prediction Score	0.66	0.68	0.80	0.86

Table 3: Number of features, prediction score for four models of the best performing scenario

This research investigates predicting suicide rates of countries. The following points describe the main research questions of research:

(1) **RQ1**: Which feature is the most significant?
(2) **RQ2**: Prediction score by number of features
(3) **RQ3**: Relationship between each feature and suicide rate

For each research question, we utilized methodology which can capture each concept well.

RQ1. Get feature importance of four algorithms each. Then visualize our result in descending order.

RQ2. Make 200 models with 1, 2, 3,..., 49, 50 number of features for 4 algorithms. Select subsets of features by its order of importance. Calculate prediction score of each model and find the best model with highest score.

RQ3. Select 6 features that are highly important in model and vicissitudinous in country. We picked GDP, population, literacy, phone user rate, death rate and EIU10VA which denotes voice and accountability including democratic index and human rights. Create new dataset from data of Korea by changing values of 6 features each. We made 576(=6*8*6*2) rows with 6 features, 8 changes, 6 age classes and 2 sex class. Then Try to figure out how suicide rate will change if a value of a feature of Korea increases or decreases while maintaining the other values.

5 RESULTS

For each research question, we yielded meaningful results as follows.

RQ1.
Figure 1 shows relative importance of each feature of XGBoost model in descending order. This is one of four models; XGBoost, Light GBM, GBoost, Random Forest. Sex, isothermality, environmental factors, age, literacy are some of the highest ranked. Figure 3 and 4 shows that of GBoost and Random Forest models. The result was highly leaned only too age and sex features, meaning the two factors are dominant. Figure 2 shows that of Light GBM model. Age feature ranked the highest, followed by sex and environmental features.

RQ2.
Figure 6 shows prediction score over number of features. The best score was 0.84 from GBoost model which training with 4 features; sex, age, isothermality and temperature of driest quarter.

	XGB	LGB	GB	RF
Number of Features	8	15	4	17
Prediction Score	0.74	0.75	0.84	0.83

Table 4: Number of features and prediction score for four models of the best performing scenario

RQ3. Figure 7 shows predicted suicide rates of 15 to 25 years old Korean male by Light GBM model according changes of each value of feature of Korea increase or decrease. All other values of features are same except one feature denoted on x axis of graphs. While country's GDP is increase, suicide rate is decrease. While EIU10VA increase, suicide rate is decreased. High death rate is related with high suicide rates.

However Figure 8 shows different results. Figure 10 is about suicide rate of 25 to 35 years old Korean male predicted by same

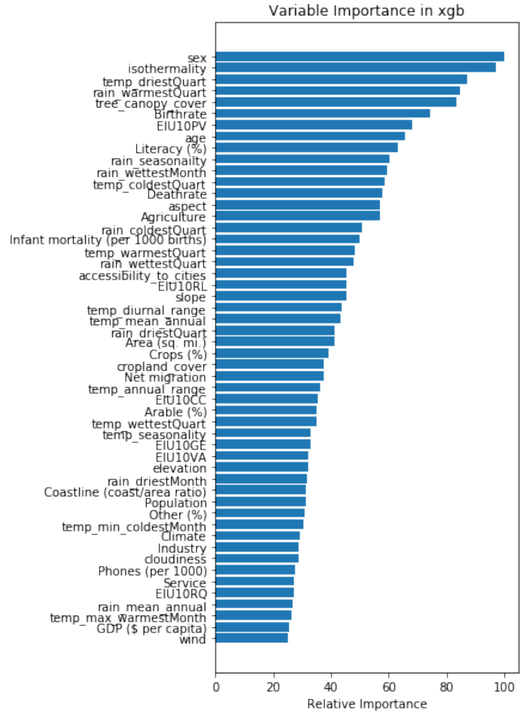

Figure 1: Feature importance of XGBoost model

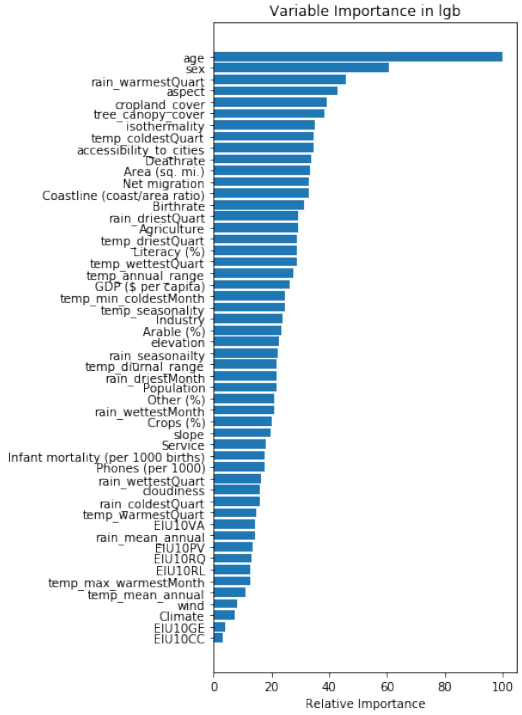

Figure 2: Feature importance of Light GBM model

model, Light GBM. While GDP is increase, suicide rate is increase. Graphs of population and death rate show almost opposite shapes.

6 DISCUSSION / CONCLUSION

We built a regression model that predicts suicide rates of countries. We considered wide domain of features ranging from literacy rate to democratic index. Among these, we proposed features that have high correlation with suicide rate which can be later used to reduce suicide rate in country scale. Moreover, in that the models resulted in more or less minor error, sufficiently successful predictions of suicide rates of each country are made.

For the limitation, hyperparameter optimization is needed. To better fit our models, it would have been better to build upon various combinations of hyperparameters. Naturally there are some space left for future work. First, try ensemble model of the best performing ones in order to yield better prediction with less error. Second, now that a model with less number of features is built, gather corresponding values for each feature to try predicting suicide rate of Korea if possible. Finally, since the reason for each feature contribution to suicide rate is interesting yet exponible, addressing the reasons may be requested as an interdisciplinary work.

7 REFERENCES

[1] Kathryn P. Linthicum. 2019. Machine Learning in Suicide Science: Applications and Etics

[2] Taylor A Burke et al. 2018. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review

[3] Seunghyong Ryu et al. 2018. Use of Machine Learning Algorithm to Predict Individuals with Suicide Ideation in the General Population

[4] Teo Susnjak, Christoph Schumacher. 2018. Nowcasting: Towards Real-time GDP Prediction

[5] Rusty 2019, Suicide Rates Overview 1985 to 2016, viewed 29 April 2019, <https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>.

[6] Fernando, L 2018, Countries of the World, viewed 29 April 2019, <https://www.kaggle.com/fernandol/countries-of-the-world>.

[7] Zander, V 2018, Environmental variables for world countries, viewed 30 April 2019, <https://www.kaggle.com/zanderventer/environmental-variables-for-world-countries>.

[8] The Economist, 2018, Democracy Index 2018, viewed 29 April 2019,<https://www.eiu.com/topic/democracy-index>.

[9] Arman, U 2017, Step by Step House prices prediction, viewed 29 April 2019, <https://www.kaggle.com/auygur/step-by-step-house-price-prediction-r-2-0-77>.

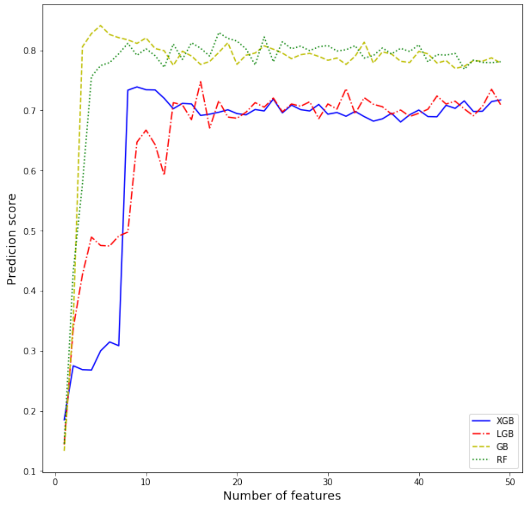

Figure 5: Prediction score over number of features

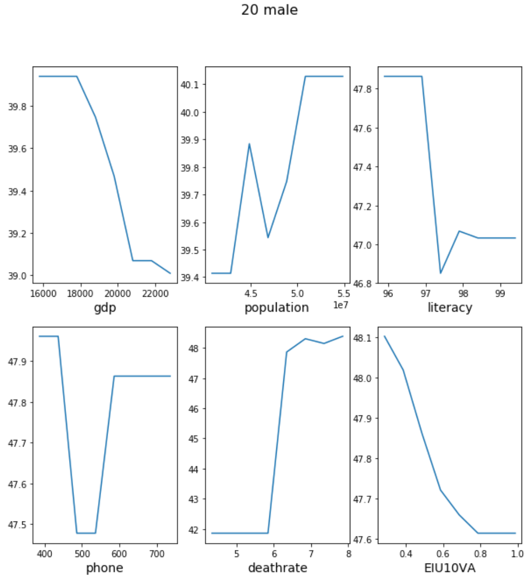

Figure 6: Suicide rate prediction of age 15 - 25 Korean male with LGBM model

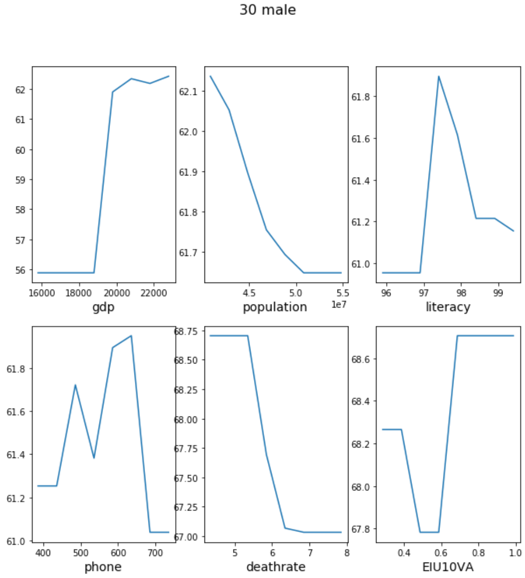

Figure 7: Suicide rate prediction of age 25 - 35 Korean male with LGBM model