Estimating Suicide Rates Using Data Science

Setup

The goal of this project is to build a model to estimate the amount of suicides for a given group provided its socio-economic conditions. This will be of great use for support organizations in enabling them to focus on the right groups. We will focus our research to answer the below.

- For a given group can we classify the amount of suicides into high, medium or low. We will use macro f1-score for evaluation in order to give importance to every class.
- Are suicides more in a particular age / gender? We will take help of visual plots for evaluation.
- Does any socio-economic condition i.e. GDP, HDI has a linear relation on suicides? We will use R-squared to estimate the variability explained.

The null hypothesis here would be that the classifiers have the same performance over the given data while the alternate hypothesis would be that the classifiers have different performance over the given data. We will use one-sided paired t-test to reject null hypothesis for $p < 0.05$.

Approach

Each data point in the dataset contains age group, gender, population, GDP, GDP per capta, suicide count, suicide count per 100k population. **GDP per capta** and **suicide count per 100k population** are really good information as they normalize the values so that they can be compared against countries with different population sizes with ease. We then performed the below.

Visualize: We have created bar plots and line plots for suicide count over the years 1985 to 2016 against other features to understand their distribution and trend. Some important plots and distributions are Male vs Female suicides from 1985 to 2016, suicides of various age groups from 1985 to 1986, distribution of suicides in various age groups further divided by gender. We have also plotted choropleth maps for suicide count, GDP, HDI, population.

Identify relationships: We created a new data frame with GDP, GDP per capta, HDI, suicide count per 100k population for every country as the original data set did not has these details and then tried to identify positive/negative relationship, correlation among these features. The analysis was still not very helpful in deriving conclusions.

Linear and local regression: We have removed columns which do not provide enough information and which are redundant i.e. country (as we already have the socio economic conditions of the country), country-year (as we already have country and year in different columns), generation (age column has similar information), suicides_no (has related count to suicide_100k_pop), gdp_for_year (as gdp per capta has related information). We used these modifications for classification as well.

We have then converted all categorical variables like age, sex, year into dummy variables and performed linear regression and loess regression on different combinations of features with dependent variable as suicides.

Classification: We then converted **suicide count per 100k population** from numerical to categorical i.e. classes - Z0 (for 0-19 suicide per 100k), Z1 (for 20-39), Z2 (for 40-79), Z3 (for 80+). The idea is that a Z3 class can be categorized as “Attention Zone” as it has high number of suicides compared to other groups.

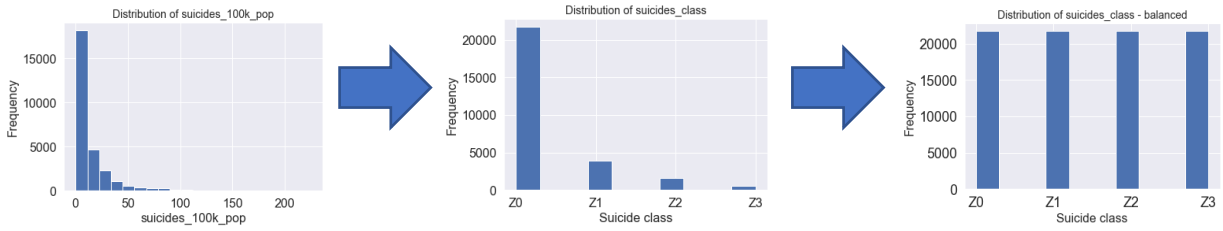

Fig 1: Conversion of suicide count from numerical to categorical and balancing

We then realized that the classes are imbalanced and hence performed oversampling to balance the classes. We were now able to build and evaluate classifiers which would estimate the suicide count for a given group with its socio-economic conditions. We have tried various classifiers as below with different combinations of features i.e. with and without dropping rows with NAs for HDI column, with and without year column. We got good f1-score after dropping year column, rows with NA in HDI column. We then leveraged GridSearchCV of sklearn to find optimal hyper parameters for the below classifiers.

Classifier	**Hyper Parameters**	**Optimal Parameters**
RandomForestClassifier	n_estimators = 1,2,4,8,16,32,64 max_depth = 2,3,4,5,6,8,16,32,None criterion = entropy, gini max_features = 6, 11, 16, 21, 26, 31	n_estimators = 32 max_depth = None criterion = gini max_features = 6
DecisionTreeClassifier	max_depth = 2,3,4,5,6,8,16,32,None criterion = entropy, gini splitter = best, random	max_depth=None criterion=entropy splitter= best
KNeighborsClassifier	n_neighbors = 1 to 30 weights = uniform,distance, metric = euclidean,manhattan	n_neighbors= 1 weights= uniform metric = manhattan
LinearSVC	C = 0,10..90	C = 0.01
LogisticRegression	penalty = l1, l2 C = np.logspace(-4, 4, 20) solver = liblinear	penalty = l1 C = 206.9 solver = liblinear
LinearDiscriminantAnalysis	solver = svd, lsqr,eigen, shrinkage = None,auto	solver = svd shrinkage = None

Table 1: Classifiers and Hyper Parameters experimented

We have then measured macro f1-score of validation data set for all 6 classifiers (with optimal hyperparameters) using 10-fold cross validation approach to determine the best performing classifier. We then also used one sided paired t-test to determine if the classifiers have similar performance.

Results

Our visualization techniques revealed the below:

- Suicides are more in male population than in female population.
- Rate of suicide increases with age and hence most of the suicides are in 75+ age group.
- Rate of suicide is also decreasing globally.

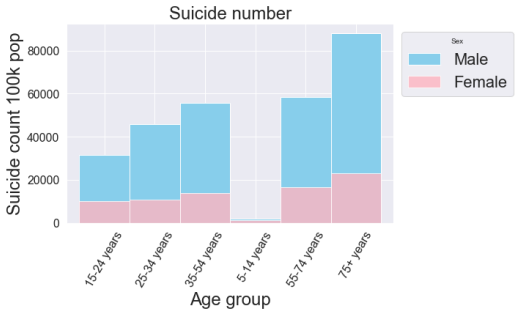

Fig 2(a): Suicide count in various gender

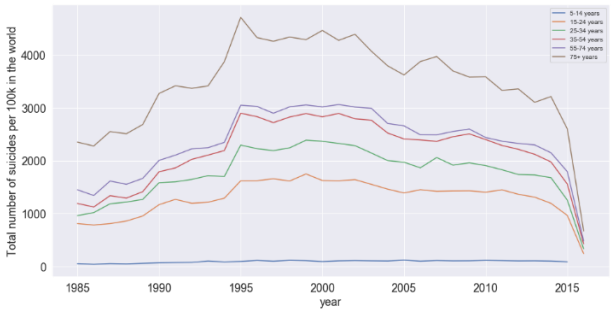

Fig2(b): Suicide count across years in various age groups

Relationship analysis revealed that there is a weak positive relationship between a countries GDP (per capita) and suicide rate with an R-squared of 0.54. There is also a relationship between HDI and suicide_100k_pop which can be ignored as its R-squared is 0.13.

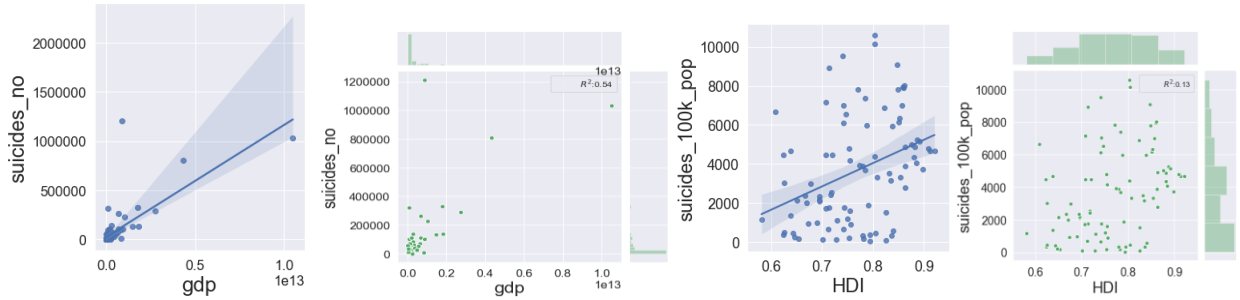

Fig 3: Relationship between gdp and number of suicides; HDI and number of suicides.

Results of linear regression analysis are below, as R-squared values are not close to 1, it is clear that the variability of suicides cannot be explained by the input variables, this is because number of suicides does not have a linear relationship with any input variables as understood from our relationship analysis.

Input variable	Dependent variable	R-squared
gdp,gdp_per_capita,HDI	Suicide_100k_pop	0.173
HDI	Suicide_100k_pop	0.132
gdp,gdp_per_capita,HDI, year, sex, age, population	Suicide_100k_pop	0.301
gdp,gdp_per_capita, sex, age, population	Suicide_100k_pop	0.295
gdp,gdp_per_capita,HDI, sex, age, population	Suicide_100k_pop	0.296

Table 2: R-squared values of linear regression

We also tried local regression to predict number of suicides using different combination of input variables. As per Fig 6 we could not find any valuable insights.

Results of various classifiers on training and validation data are as below. We have better f1-score with Random Forest Classifier and with Decision Tree performing second and Linear SVC doing the worst. The performance of Decision Tree and Random Forest over different folds seems to be same, there does not seem to be any kind of variation in data that is affecting its performance.

Fig 4(a): Evaluation of various classifiers on training and validation data.

Fig 4(b): Evaluation of Random Forest and Decision Tree over 10 folds.

Random Forest classifier also has an accuracy of 0.98 indicating that most of the predictions are correct. The f1-scores of each individual class is above 0.95 as in the below figure indicating we could achieve the goal of the project.

	precision	recall	f1-score	support
Z0	1.00	0.93	0.96	1676
Z1	0.94	1.00	0.97	1626
Z2	0.99	1.00	0.99	1440
Z3	1.00	1.00	1.00	1234
accuracy			0.98	5976
macro avg	0.98	0.98	0.98	5976
weighted avg	0.98	0.98	0.98	5976

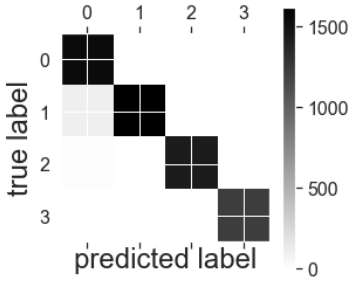

Fig 5: Classification Report of Random Forest Classifier

The results from one-sided student's T test as below indicate that we can reject our null hypothesis that these classifiers have similar performance.

Classifier 1	**Classifier 2**	**p-value**
Random Forest	Decision Tree	0.0000032
Random Forest	kNN	0.00000031
Random Forest	Linear SVM	0.000000000000028
Random Forest	Logistic Regression	0.00000000000011
Random Forest	Linear Discriminant Analysis	0.000000000000015

Underlying function:
The amount of suicides attribute to a sequence of outcomes of different factors like old age, lower HDI of the country etc. Hence Random Forest and Decision Tree is able to perform better as they are able to identify the right sequence of outcomes in determining the amount of suicides. As Random Forests construct a multitude of decision tress, they have slightly better performance. Linear SVM performed the worst as the groups cannot be divided by a maximum-margin hyperplane as they have a tree-based relationship. kNN did a decent job as it tried to identify the groups based on distance and the groups with same outcomes has similar distance.

Limitation:
We noticed below with regards to our dataset:

- All countries are not included in the dataset.
- HDI for some countries and years are not present which makes the analysis incomplete.

We can achieve better performance and infer more patterns if we can overcome these limitations. It would be a good idea to get data related to other social, economic, geographical factors for these countries across the years 1985-2016 and use it to draw more inferences.

Conclusion

Random Forest classifier on the current data set is a good starting point but is not a complete solution as the dataset lacks certain things like all countries data, some more socio-economic factors like life expectancy, pollution, rate of diseases which can be strong indicators to suicides. We can also reject out null hypothesis that all classifiers have similar performance.

As a Data Scientist, this study reiterated the fact that visualizations are a good way to identify critical patterns in data and also understanding categorical data is important in creating better plots. It is also important to take a step back and reformat some features of the data to bring out the required results.

References

1. Kaggle 2021, Suicide Rates Overview 1985 to 2016 Dataset. https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016/download

Appendix

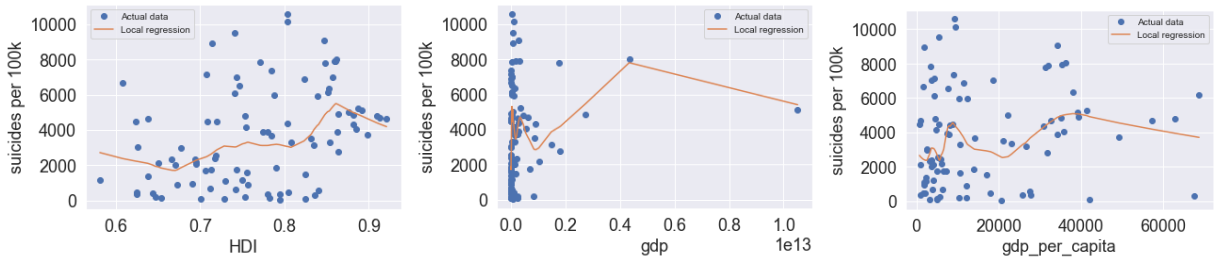

Fig 6: Local regression to predict suicides via different input variables

Classifier	**Training f1-score**	**Test f1-score**
RandomForestClassifier	0.98	0.98
DecisionTreeClassifier	0.97	0.98
KNeighborsClassifier	0.92	0.93
LinearSVC	0.16	0.17
LogisticRegression	0.57	0.57
LinearDiscriminantAnalysis	0.57	0.57

Table 3: Macro f1-score on training and validation data