Machine Learning Nanodegree

Capstone Proposal

Amr Eid Abd Elgawad

January 25, 2019

Definition

project overview

suicide is a crime that individual do to him self and is increasing through years , number of suicide people is affected by many circumstances like wars or starvations or economic circumstances etc ,Death by suicide is an extremely complex issue that causes pain to hundreds of thousands of people every year around the world. The objective of this project is to contribute to put ways to prevent suicide ,there is alot of different researches in this field such as this research from association for psychologial science (APS) with title :Predicting Risk of Suicide Attempts Over Time Through Machine Learning [5] ,and this article from The American Journal of Psychiatry with title : Suicide Prediction With Machine Learning [6] , the dataset from kaggle [named :WHO SUICIDE STATISTICS [1] and is collected from the World Health Organization (WHO) [2] ,dataset has 4 features and one target the number of suicides in each country through years ,and the input to the model is age :of people that the user needs to predict number of suicides ,sex : if people male or female,country : which country that wanted to predict suicide , population :population in countries that wanted to predict suicide so the inputs are (country , sex ,age ,population) the Dataset shape is (43776, 6) it contains 43776 rows ,and after removing the rows that contain nan values in any column so it become 36060 rows .

Problem Statement

the main problem that this project focuses on is to help in reducing and preventing suicide ,as Death by suicide is an extremely complex issue that causes pain to hundreds of thousands of people every year around the world , this is a real life problem and is using real data set from (WHO) that is collected through years for each country in the world ,

by applying supervised learning (regression),and use regressors like: linear regression ,decision tree regressor , bagging regressor , random forest regressor and choose the regressor that gives higher r2score. then using this model to predict the number of suicides for a specific country and specific age ,and sex ,and population in that country ,this solution is helping in reducing the suicides by predicting number of suicides and see where the highest number of suicides and see which ages they are and which gender for a specific country to apply methods and make media programs to speak with a specific life stage.

Evaluation Metrics

R2SCORE is a good estimate to evaluate the model

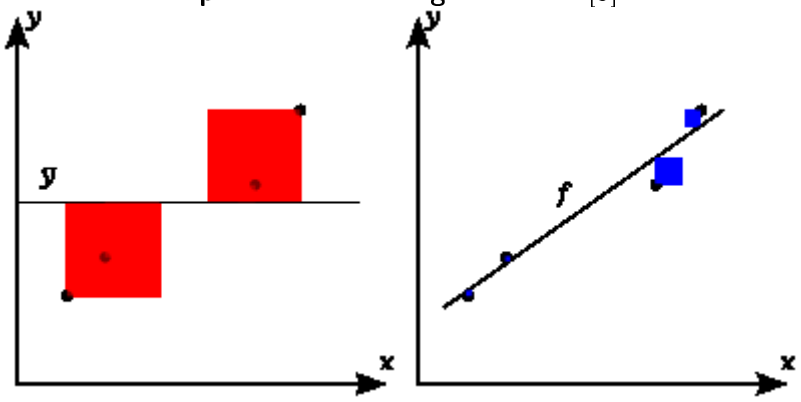
the model that results the hieghst R2SCORE on training and testing is bagging regressor

Training score 0.9895040189816754

Testing score 0.9534959132816964

r2 score[7]—varies between 0 and 100%, It is closely related to the MSE , but not the same. Wikipedia defines r2 like this, " . . . is the proportion of the variance in the dependent variable that is predictable from the independent variable(s)." Another definition is "(total variance explained by model) / total variance." So if it is 100%, the the two variables are perfectly correlated, i.e., with no variance at all. A low value would show a low level of correlation, meaning a regression model that is not valid, but not in all cases.

The better the linear regression (on the right) fits the data in comparison to the simple average (on the left graph), the closer the value of R 2 {\displaystyle R^{2}} R^{2} is to 1. The areas of the blue squares represent the squared residuals with respect to the linear regression. The areas of the red squares represent the squared residuals with respect to the average value . [8]



Analysis

Data Exploration

the features in this data set is (age ,sex,country ,population) the target is (number of suicides) the dataset contains records for each country in the world from 1985 to 2015 each year have records based on ages as age columns is categorized in 5 categories (15-24 years ,25-34 years ,35-54 years ,55-74 years, +74 years) in sex column each gender (male ,female) have records based on age records suicides columns contains the value of suicides number in each age type and each sex tyoe for a specific year and country in the population column it contains values of number of population for each sex and age type for a specific year on specific country .

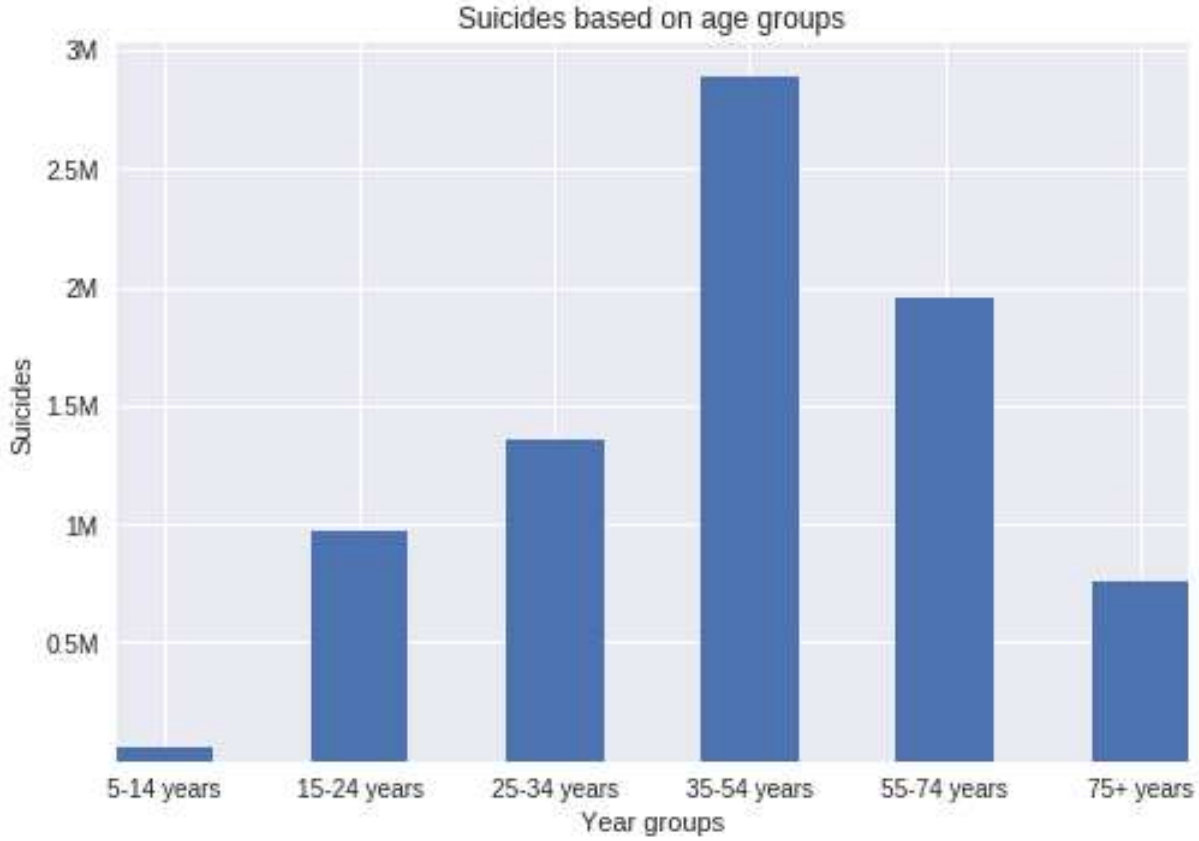
	country	year	sex	age	suicides_no	popula
0	Albania	1985	female	15-24 years	NaN	277
1	Albania	1985	female	25-34 years	NaN	246
2	Albania	1985	female	35-54 years	NaN	267
3	Albania	1985	female	5-14 years	NaN	298
4	Albania	1985	female	55-74 years	NaN	138
5	Albania	1985	female	75+ years	NaN	34
6	Albania	1985	male	15-24 years	NaN	301
7	Albania	1985	male	25-34 years	NaN	264
8	Albania	1985	male	35-54 years	NaN	296
9	Albania	1985	male	5-14 years	NaN	325

here is some statistics about the data set gathered using describe command

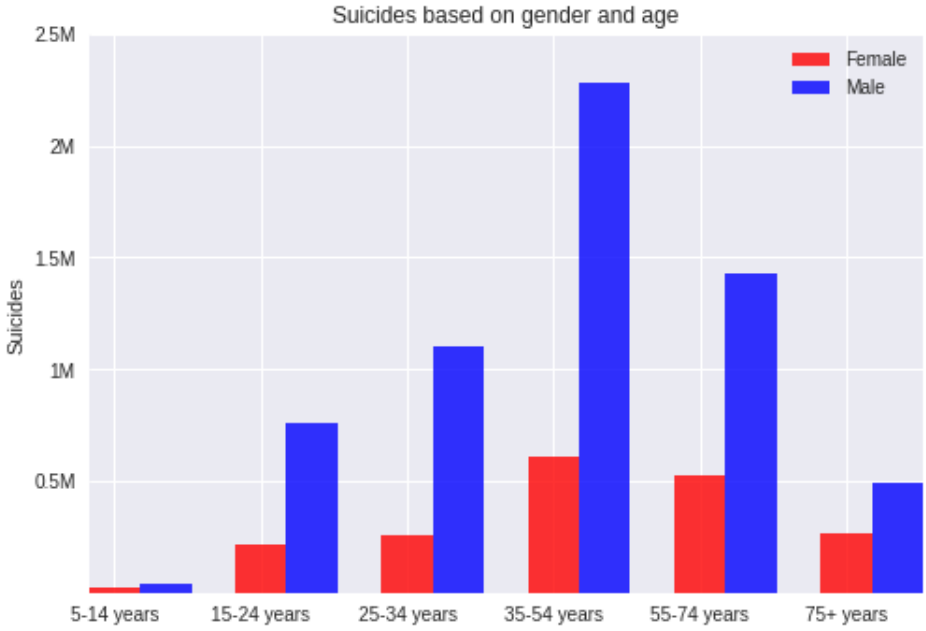
	year	suicides_no	population
count	43776.000000	41520.000000	3.831600e+04
mean	1998.502467	193.315390	1.664091e+06
std	10.338711	800.589926	3.647231e+06
min	1979.000000	0.000000	2.590000e+02
25%	1990.000000	1.000000	8.511275e+04
50%	1999.000000	14.000000	3.806550e+05
75%	2007.000000	91.000000	1.305698e+06
max	2016.000000	22338.000000	4.380521e+07

Exploratory Visualization

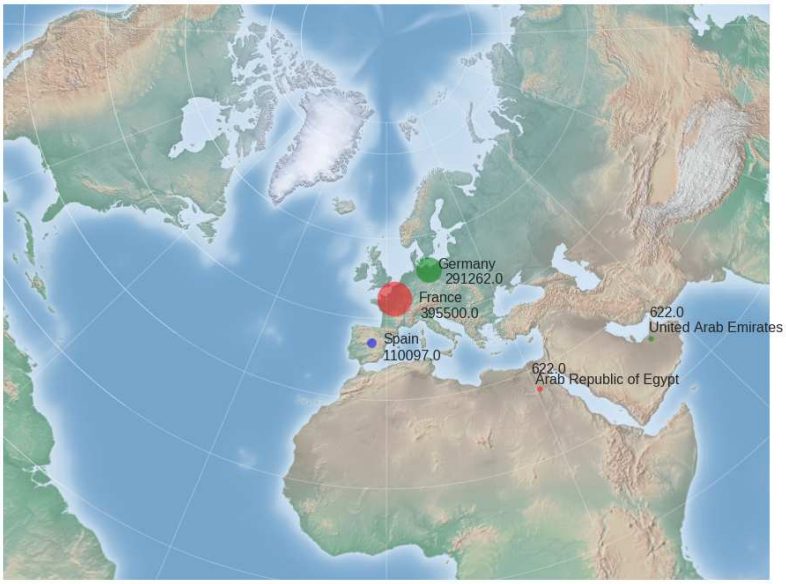
SUICIDES BASED ON AGE GROUPS: in this graph show the number of suicides in the whole world classified by the 5 types of ages as illustatrted in previous sections : in x-axis represents the age categories on y-axis the number of suicides in millions concluded from this graph that the most number of suicides is between (35-54 years).



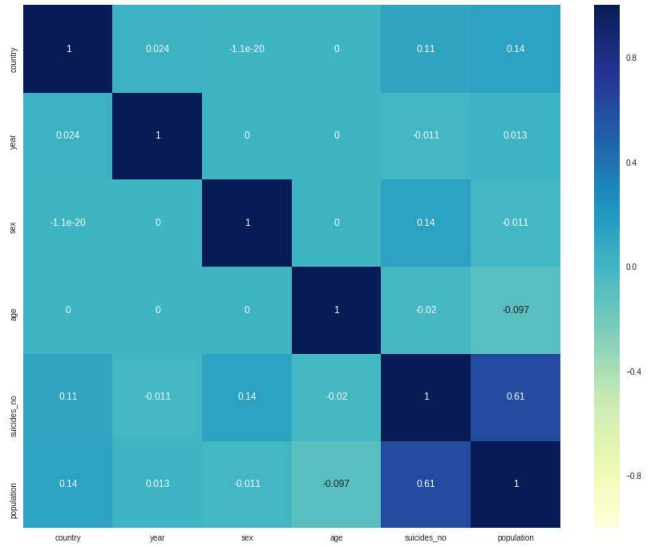
SUICIDES BASED ON GENDER AND AGE: this graph shown the same information as the previous one but in this the number of suicides in each age category is shown in two categories based on gender if male or female and conclusion from this graph that the most number of suicides is between (35-54 years) (and males).



draw world map for each country with suicide number :by using mpl_toolkits.basemap draw a world map with some countries and it's total records of suicides in all ages and genders from 1985 to 2015 the countries are(Egypt ,uae,spain,france,germany)



SNS heat map to visualize the correlations between variables in the dataset as shown in the figure drawn by seaborn lib



Algorithms and Techniques

* first algorithm used is linear regression[9] which is the simplest regression technique that algorithm tries to draw a line that matches or closer to the most points of data and this also the bench mark model of this problem

* Decision tree regressor [10] and it works as follows :Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

* Bagging regressor[11] : also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach

* Random forest regressor [12]: or random decision forests is an ensemble learning method for classification or regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

* Gradient boosting [13]is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function

Benchmark Model :

in this project will create model to predict the number of suicides by using features : age,sex,country ,population ,and target in the dataset is the suicide number, my solution is to build a regression model and making analyasis on data using matplot lib ,seaborn lib there is many kernels in kaggle[3] that doing visualization and statistics and also make regression to predict suicides but only for one country (sweden) [14] ,but i will use the country as a feature to predict for any country ,tried the linear regression model on the data set and the results on testing and training as follows, will do optimization techniques and try another regressors and ensemble methods to get higher R2SCORE.

```
[28] from sklearn.model_selection import train_test_split
     x_train,x_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=434)

[29] from sklearn.linear_model import LinearRegression
     reg=LinearRegression()
     reg.fit(x_train,y_train)

     LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
              normalize=False)

[30] y_pred = reg.predict(x_test)

     from sklearn.metrics import r2_score
     y_pred2=reg.predict(x_train)
     print(r2_score(y_test, y_pred))
     print(r2_score(y_train, y_pred2))

     0.39504839095750277
     0.3969447878461171
```

The score of linear regression on testing data

The score of linear regression on training data

Methodology

data preprocessing

by using the preprocessing module that is in sklearn ,made label encoder on the countrry ,sex,age columns to make it numbers so that can pass it to the model for training and testing ,for the the countries coulmn it took numbers from 0 which represnts (albania) to 117 which represents Virgin Islands (USA)

	country	year	sex	age	suicides_no	population
24	0	1987	0	0	14.0	289700.0
25	0	1987	0	1	4.0	257200.0
26	0	1987	0	2	6.0	278800.0
27	0	1987	0	3	0.0	311000.0
28	0	1987	0	4	0.0	144600.0

Implementation

Programming Language and Libraries

- Python 3.
- jupyter notebook.
- scikit-learn
- mpl_toolkits.basemap
- matplot lib .
- numpy .
- pandas .

implemented draw_map function from scratch which is used to draw the number of suicides on the map , this accepts basemap object as input and this function uses global variables frequencies of suicidess of each country an is calculated outside the function .

implementing regression models begins with the implementation of the bench mark model (linear regression) the results on testing and training is very low .

then choosed different types of regression models which is decision tree regressor the results on training and testing is improved from the bench mark model.

then tried to improve the decision tree regressor to reduce overfitting and improve performance by using ensemble methods like randomforest regressor and bagging regressor and GradientBoostingRegressor , ExtraTreesRegressor.

Refinement

No refinement is done as the baggingregressor model is doing perfectly on the data using the default hyper parameters so no need for parameters tuning .

the model that results the hieghst R2SCORE on training and testing is bagging regressor

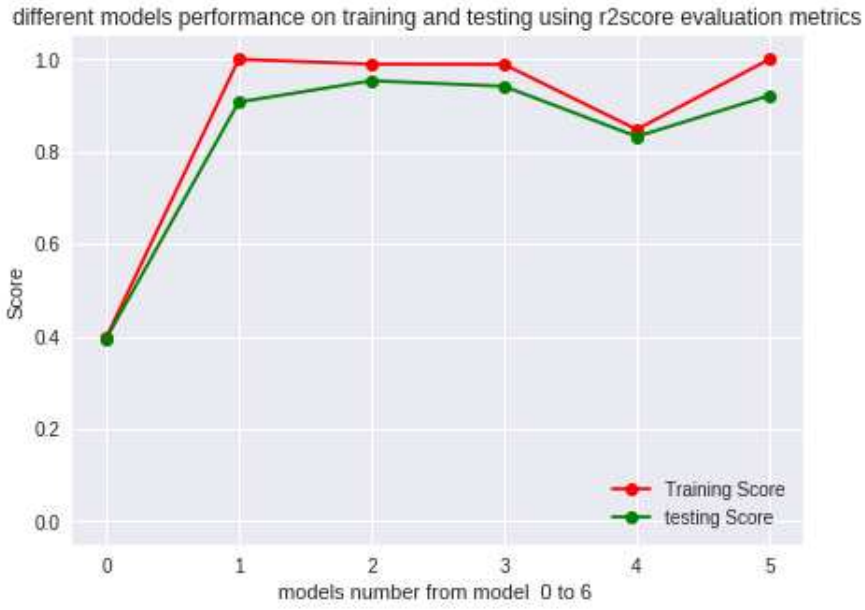
Training score 0.9895040189816754

Testing score 0.9534959132816964

Results

Model Evaluation and Validation

models in this project is evaluated using r2score evaluation matrix ,i made evaluatio of each model on trainig and testing and put the records in two array (training, testing) then by using matplotlib draw a graph the show the performances of each model .



Justification

in the bench mark model (linear regression) the r2score on testing and training is very low less than 50% then decision tree regressor results better performance and any model that is used in this project it's results is higher than the bench mark model .

but the best regressoin model for this problem is by using bagging regressor (endemble method) which r2score results is :

Training score 0.9895040189816754

Testing score 0.9534959132816964

Conclusion

Free-Form Visualization

creating a histograms to see the number of suicides in each age category ,the created a histogram to recognize the hieghest number of suicides if it in male or females then creating histogram to visualize each country suicides in all years and age and sex,

then draw basemap to visualize some countries on the world map with it's number of suicides the same as previous histogram but this is for more clarification, then by using seaborn draw heatmap to draw the correlations between features in the data set ,at last section draw a plot to visualize the results(training, testing) of the different models that is implemented .

Reflection

end-to-end problem

in this project i wanted to help in reducing and preventing suicides using machine learning and to apply what i learned in the last few months to solve a real life problem to solve this problem first i read dataset and check for null values i droped the nans and make data cleaned then preprocess the data by doing label encoding then do some analysis and visualization on the data to understand the data set and relations between it's variable the split the data set for training and testing sets then apply regression algorithms on the data and i choosed the best model which is :bagging regressor .

Chalenges

the difficulties that i faced is to analyze the data and get counts of suicides based on gender and sex as the data set not well organized (in my humble opinion).

improvement

i think this model will be more efficient and accurate if the data set first i well organized like making one hot encoding of age column , and also by adding more features to the data set like marital status and job describtion and other important features which will make the model more realistic and confident.

Bibliography

[1] https://www.kaggle.com/szamil/who-suicide-statistics

[2] https://www.who.int/about/copyright/en/

[3] https://www.kaggle.com/szamil/who-suicide-statistics/kernels

[4] https://www.kaggle.com/cengizeralp/practice-2-suicide-analysis-with-regression

[5] https://static1.squarespace.com/static/54de6056e4b0409b0654ceb4/t/59809fc2c534a5a9d7879cb0/1501601732516/Walsh%2C+Ribeiro%2C+%26+Franklin%2C+proof+version+%28ML+and+sui+attempt+prediction%29.pdf

[6] https://ajp.psychiatryonline.org/doi/10.1176/appi.ajp-rj.2017.120105

[7] https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/

[8] https://en.wikipedia.org/wiki/Coefficient_of_determination

[9] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[10] https://www.saedsayad.com/decision_tree_reg.htm

[11] https://en.wikipedia.org/wiki/Bootstrap_aggregating

[12] https://en.wikipedia.org/wiki/Random_forest

[13] https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

[14]