COMP5310 Principles of Data Science

Numbers of terrorist attacks prediction based on national multiple indicators

Student Name: Kai Hu

Student ID: 490597347

Unikey: kahu5036

Abstract:

Today, frequent terrorist attacks have become a major concern for people around the world. What are the main factors that will affect the terrorist attacks? And what is the correlation between numbers of terrorist attacks and national multiple indicators. In order to find the answer for this question, some major living condition indicators would be analyzed in this report, mainly focus on how these factors affect the numbers of terrorist attacks and which one of them is the main factor for this question. In this assignment, below method will be used to address upper issues:

1) The most important factors that affects the numbers of terrorist attacks will be analyzed using heatmap.
2) The data clean and the custom matrix build based on top 10 numbers of terrorist attacks prediction.
3) The prediction of terrorist attacks numbers based on national multiple indicators using keras sequential pipeline.
4) The results analyze based on the prediction.

Approach

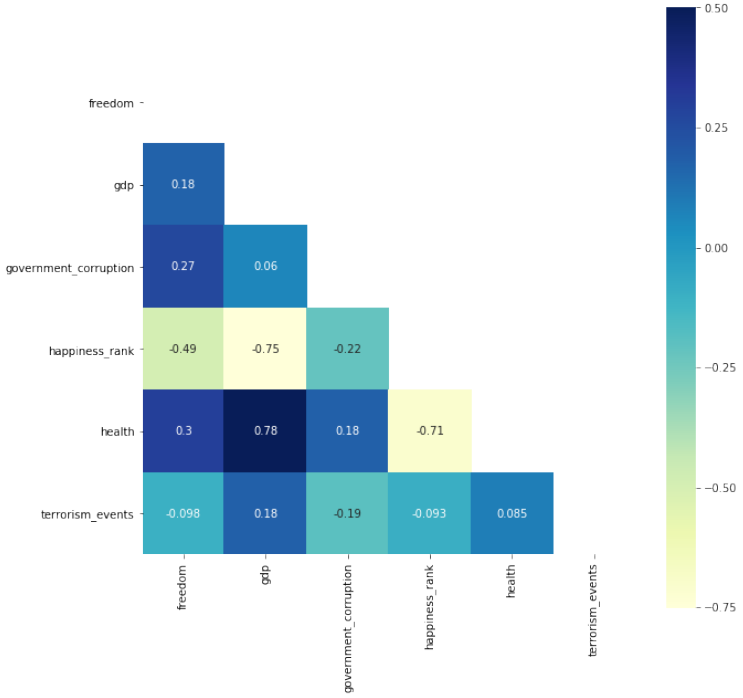
1) Data clean and reform

   In assignment 1 all "null" data has already been removed, so assignment 2 won't repeat that step. However, in order to train the model properly and correctly, the correct data type need to be aligned. Firstly, the column "country" is an object instead of float. To address this, as an example, add the country of TOP10 as a new column back to the original database and remove the countries in the original database. Please notice here, in the new database, countries that are not top 10 simply don't have labels, but the data is still used to build the model. Secondly, there are a large number of random terrorist attacks, and their proportion is not very large, but it will affect the accuracy of the training model throughout the database. Therefore, by directly using the SQL statement to filter, the number of terrorist attacks below 40 was removed.

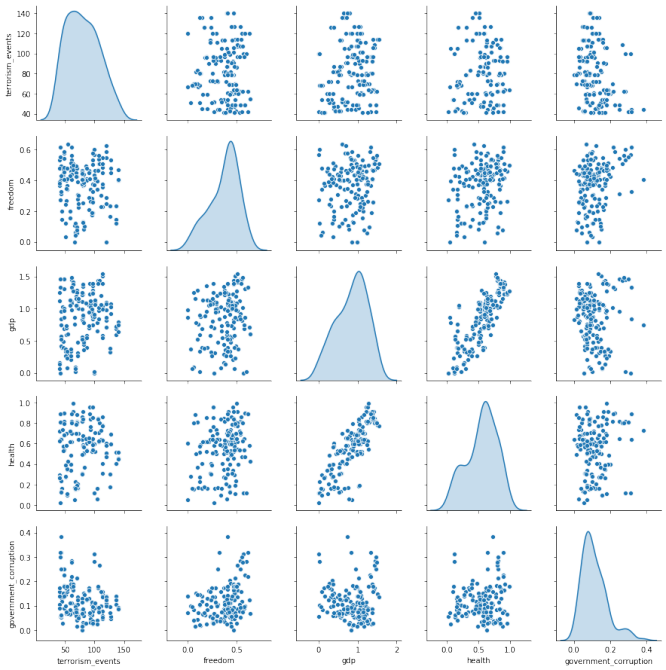
2) Correlation heatmap and joint distribution of pairs of columns from training set

   A heatmap was generated to show correlation between terrorism attack numbers, freedom, GDP, government corruption, happiness rank and health rate. (Graph 1. Correlation heatmap between national multiple indicators and numbers of terrorist attacks), it is very obvious that GDP has the largest impacts on terrorist attack numbers.

A joint graph of pairs of columns from training set was generated to show the data distribution, for most data pairs, no special patterns were identified except GDP and health index has linear relation. (Graph 2. Joint distribution of pairs of columns from training set)



Graph 1. Correlation heatmap between national multiple indicators and numbers of terrorist attacks)

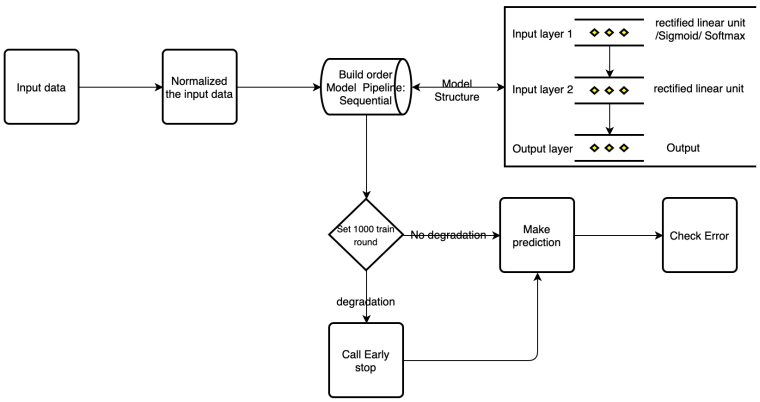


Graph 2. Joint distribution of pairs of columns from training set

3) Model Training

The first operation on data set is to normalize the input data as the input data set have different data range and scale, the normalization limits multiple inputs to same scale date. It is useful for later model training. Then a sequential pipeline uses 3 layers pipeline was introduced to train the model. Layer 1 is initial layer that contains the input shape and beginning method to be used for the pipeline. For example, “Relu”, “Sigmoid”, “Softmax” and etc. and be used to initial the process. After comparison, “Relu” was selected as initial layer method for this pipeline as it has the lowest error rate and best data fit curve.

The 2nd layer is mainly responsible for more accurate and detailed training on the incoming data. When the training was finished, we can find layer trains the 8 times parameters than layer 1. The role of the last layer is to output the data one by one. (Graph 4. Model Summary)



Graph 3. Model training logic diagram

Layer (type)	Output Shape	Param #
dense_33 (Dense)	(None, 128)	2048
dense_34 (Dense)	(None, 128)	16512
dense_35 (Dense)	(None, 1)	129

Total params: 18,689
Trainable params: 18,689
Non-trainable params: 0

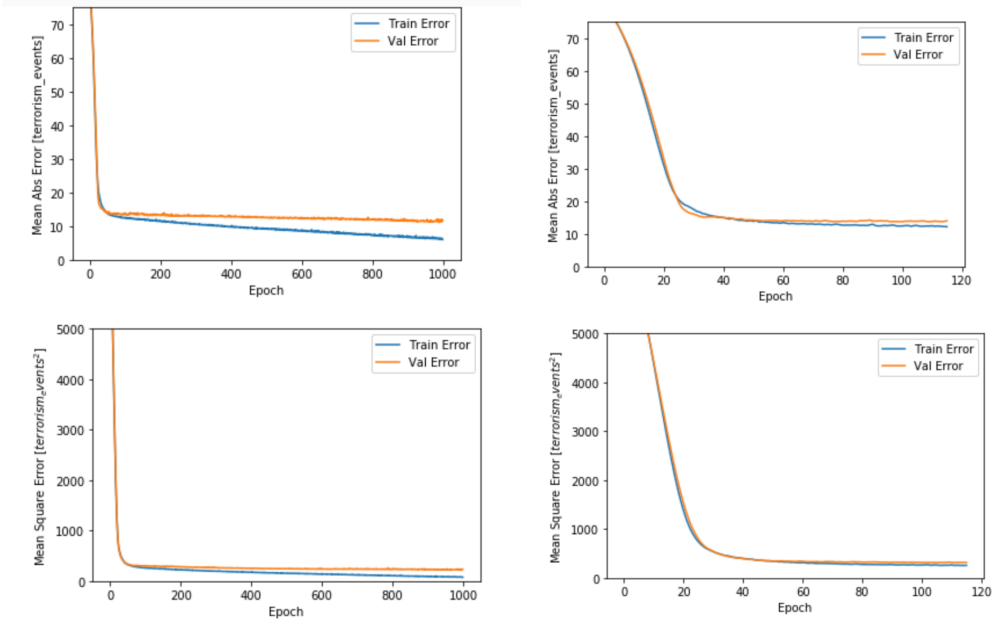
Graph 4. Model summary

Result and Evaluation

1) Model error comparison

After 1000 rounds training, mean absolute error and mean square error were used to show the difference between the two models. (Graph 5a, Error comparison of train data and actual data of 1000 rounds)

Graph 5a shows, the degradation happens after 100 rounds, which means it is unncessary to train the model for more than 100 rounds. In this case, early stop function is called to avoid further training if no improvement appeared. (Graph 5b, Error comparison of train data and actual data of early stop)

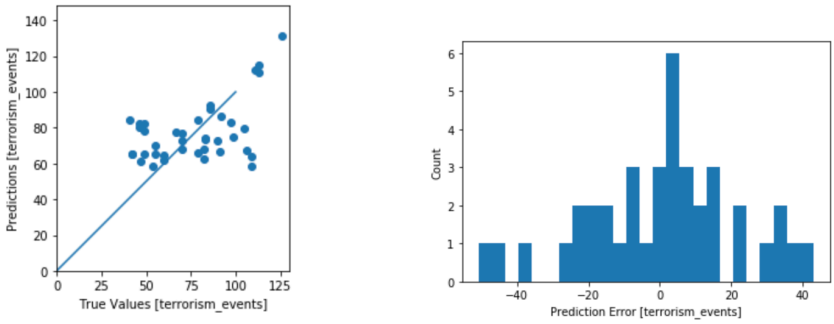


Graph 5a, Error comparison of train data and actual data of 1000 rounds

Graph 5b, Error comparison of train data and actual data of early stop

2) Prediction based on trained model

Finally, use the previously trained model to make predictions and compare the predicted results with real results.



Graph 6a, Comparison between predicted results with real results

Graph 6b, Prediction minus Test

It can be found that the predicted results and the real results are distributed on both sides of the midline, which is caused by the lack of correlation of the data itself. The mean abs error between predicted results and the real results is 16.6570 events. This error can be further reduced by feeding more data. As current national multiple indicators only include 3 years data. Therefore, this error range is acceptable for this assignment.

Reference:

[1] "Global Terrorism Database",

https://www.kaggle.com/START-UMD/gtd

[2] "World Happiness Report",

https://www.kaggle.com/unsdsn/world-happiness

[3] "Google Tensorflow tutorials",

https://www.tensorflow.org/tutorials