Machine Learning Report of Human Freedom Index

Yuzhe Yang*, Siyuan Liu, Yuqian Dai, Laiyuan Zhang, and Lei Guo

***Abstract*—The Human Freedom Index (HFI) presents the state of human freedom in this world based on a broad measurement, which encompasses personal, civil, and economic freedom. Based on the conclusions from our data exploration report, we rethink the work related to the missing value processing and try different approaches to fill the missing cells for improving the data quality and model performance. In addition, we keep the regression and cluster analysis for digging out the correlation between HFI, GDP, and immigration rate. The result indicates that the immigration rate has a limited impact on the HFI score, but it could still be regarded as a positive correlation. In the meantime, the GDP has a clearer positive correlation with the HFI score. Our code and related experiment results can be obtained from https://github.com/YuzheYang/AdvancedML2019-2020.**

***Index Terms*—Human Freedom Index, data visualization, data cleaning, correlation analysis**

I. Introduction

The research of HFI works on exploring a sufficient number of objective indicators that can comprehensively represent the social freedom of every country, in the aspects of the individual, society, religion, culture, regime, and economics. Apart from the strong subjective comment about some countries in front of every yearly report, it is neutral and academic. With the data from this report, it is able to dig implicit conclusions under their surface values by machine learning technologies about freedom.

II. Recap the data structure.

There are 76 basic indicators in two top categories, the personal freedom sub-index, and the economic freedom sub index, for each country of 162 around the world, and all of them are the root nodes in that chart which belong to a higher category. The value of each more top category, also called a non-basic indicator, is the mean of all the marks from its sub-categories. The value of each basic indicator is acquired from the statistical data of the source like an influential and authoritative global organization or a reliable ongoing research program. And all the basic indicators are normalized discretely or continuously in the range of 0 to 10 in reasonable methodologies which can be referred in their full report. Apart from these basic and non-basic indicators, 112 in total, which denote the conditions of a country in a variety of areas; there are also 8 columns provide the information of the rankings by different scores and the quarterly interval. Moreover, it is noteworthy that 10 basic indicators of 76 are empty, with no data inside the columns, but the scores of their more top categories can demonstrate that there should be some values but still missing for some reasons.

Yang et al. are with the School of Electronics and Computer Science, University of Southampton, Southampton, e-mail: (yy1a19, sl18n19, yd6u19, lz4y19 and lg1y19) @soton.ac.uk, and this report is for our group project in COMP6208: Advanced Machine Learning (2019-2020).

III. Rethinking of the Data Cleaning

In practice, a very common problem is that data-sets often have missing values. And the primary purpose of this work is to explore the effect of different data completion methods on model performance. By applying mean data filling, machine learning-based, and deep learning-based data filling methods, the performances of them on 10-classes classification problems are discussed. In this section, we will implement different filling algorithms to fix the missing value in our data-set.

A. Mean value filling

The mean value filling step is to select a column with missing values, remove all missing values of the column, and calculate the mean value using the remaining values. At last, fill all the missing values in the column with the mean value.

B. KNN filling

The step of KNN filling is to select the columns other than the selected columns, and at the same time remove the rows with missing values in the columns that need to predict the missing values, and then calculate the distance. Take the following figure as an example. A batch of data has 6 samples and three features. The first feature of sample 5 and sample 6 needs to be filled. Then first need to select the second and third features of the sample and calculate the nearest sample. The sample closest to sample 5 is sample 2, and the sample closest to sample 6 is sample 1; thus, 8 and 10 are filled into the first features of samples 5 and 6, respectively. Technically, if the missing values are discrete, use the K-nearest neighbour classifier to vote for the largest category of K neighbours for filling; if it is a continuous variable, use the K-nearest neighbour regressor to fill in the average of the K neighbours.

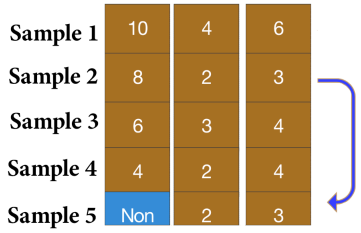

Figure 1: Example of the KNN filling.

C. Deep learning-based filling

In particular, a 3-layer neural network is used for missing value filling based on deep learning knowledge. Since the missing value columns are all continuous values, it can be regarded as a regression problem. More specifically, the input dimension is 119 dimensions (excluding the selected label column), the number of nodes in the first hidden layer is 50,

the number of nodes in the second hidden layer is 20, and the number of nodes in the output layer is 1. The activation function between layers is RELU, the optimiser is Adam, and the MSE is used as the loss function.

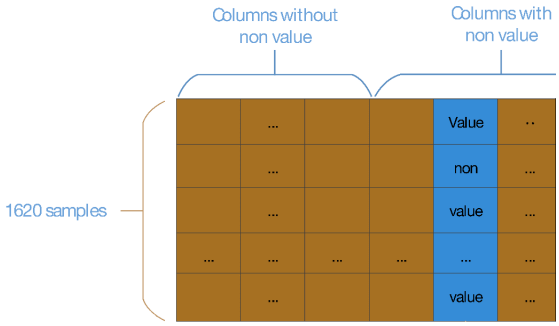

Figure 2: Example of the deep learning-based filling.

The filling step is to sequentially select columns with missing values as labels, where samples with values in the column are used as the training set, and samples without values are used as the test set. The features of the training set and test set are all columns except this column.

In Figure. 2 the brown column is the label column selected by the current model, and the blue column is the feature column of the training set and test set. The samples with corresponding values in the label column are regarded as the training set, and the samples with the corresponding labels in the missing value are regarded as the test set. Besides, considering that 96.67 percent of the data columns have missing values, the feature columns will inevitably have missing values, and the neural network cannot handle this situation. Therefore, the feature columns are filled with mean values. There are 116 columns with missing values in the data, then a total of 116 models are required to predict the missing values of each column, and all the final samples are the samples filled with deep learning-based knowledge.

The experimental results are shown in Figure 3 and Figure 4. Similarly, the classification model is a 3-layer fully connected neural network with an input dimension of 118 dimensions (excluding the year column and label column) the first one hidden layer has 50 hidden nodes, the second hidden layer has 30 hidden nodes, and the output layer has 10 nodes, corresponding to 10 categories. The activation function between layers is ReLU, and the optimizer is Adam; however, the loss function we applied is cross-entropy. Also, after 20 iterations, the indicator will no longer improve and stop to prevent overfitting. The measurement indicators are the correct rate and F1-score. For the populated data set, the data with the year 2017 is selected as the test set, and the other samples are used as the training set.

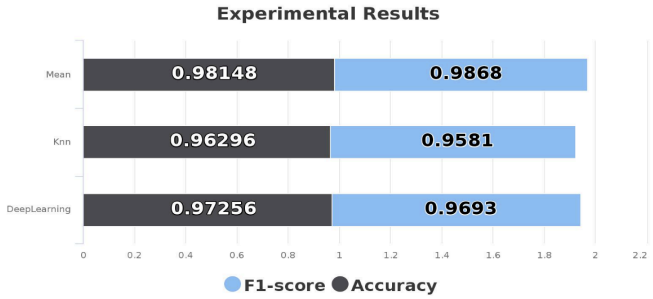

Figure 3: The F1-scores and accuracy of three approaches.

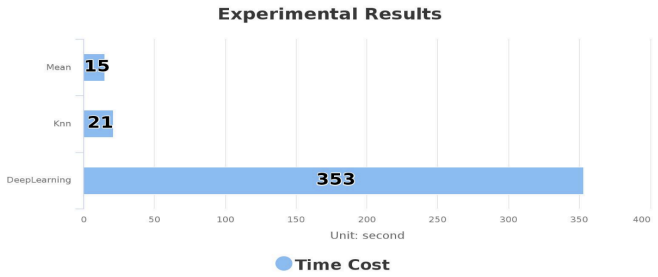

Figure 4: The time costs of three approaches.

In conclusion, the speed of the mean filling is the fastest, due to the simple calculation. The second is the KNN filling algorithm based on machine learning knowledge because KNN filling is basically to determine the object according to the relevant distance. Conversely, the filling method based on deep learning takes the longest time. The total time often exceeds the sum of the first two. The reason is that it requires training up to 116 models to make predictions for missing values, so it takes the most time.

From the perspective of F1-score and accuracy, we normally think that methods based on machine learning or deep learning will bring better performance, but actually not. Regardless of F1-score or accuracy, the method of mean filling shows the best performance, and the methods of deep learning and machine learning are second and third. The potential reasons might be that there are too many missing values while the total amount of data is relatively small, so filling methods based on machine learning or deep learning cannot learn enough features. In fact, the performance of the three filling methods is highly related to the data set, and each filling method has its own applicable scenarios. It is impossible to judge the best method based on an experiment or a data-set.

IV. Regression Analysis

In this section, we will give a regression analysis that can map fewer features into the HFI score. This work can help the survey designer to reduce their designing features and remove some unnecessary items to save both time and money. This section consists of two subsections. First, feature selection and correlation analysis will be illustrated for reducing the dimensionality. Then, we will make a comparison among different algorithms for the regression task, which include random forest, support vector regression (SVR), multi-layer perceptron (MLP) regression, AdaBoost, and XGBoost. As mentioned before, the data-set contains 120-dimensional features, which are all the accordance for calculating the HFI

score. However, some features are highly correlated with the others, which may cause the regression model sensitive to parameters, and also it will take up more weight parameters. Moreover, reducing the input features can make the relation between labels and features be highlighted. Hence, selecting features is vital to machine learning-based approaches. In this work, we first compare the hf score to all those qualified features. Then, we use a settled threshold to divided the features into high correlated and low correlated to hf score. After this classification, we use a scatter figure to observe the correlation between hf score and each feature. The result shows that there actually exist some features which are highly correlated with hf score and their cross-correlation is also high. Those features are labeled as redundant data, which should be re-sampled for reducing the dimensionality. Then, we use sklearn built-in feature selection approaches, such as SelectKBest, and Tree-based selections to select features. In addition, we also select one hand-crafted group feature, which is set as a comparison group.

A. Regression Analysis

For analyzing the regression results, we choose to divide the data-set into two parts, 130 samples as our training set and 32 samples as the testing set. In this work, we choose five regression algorithms, which are random forest, support vector regression (SVR), multi-layer perceptron (MLP) regression, AdaBoost, and XGBoost for comparison. The result is shown in TABLE IV-A.

Model	SelectKBest	Tree-based	Ours
Random Forest	0.87	**0.93**	0.78
SVR	**0.98**	0.97	0.94
MLPRegressor	0.97	**0.99**	0.86
AdaBoost	0.86	**0.94**	0.76
XGBoost	0.95	**0.96**	0.88

It can be observed that the Tree-based selection can outperform the others and select effective features for the regression task and the average accuracy on test data is 0.96 which is quite high, and it proves that our feature selection is successful and one good thing is the tree-based selection reduces the features into 5 dimensions which is quite low.

V. Cluster Analysis

To better understand and cluster the data, we first analyse the feature structure and clean the data based on the structure graph. Then, we use k-means to cluster the data for analysing the country HFI score, and we direct the clustering result on the world map for observing the HFI cluster distribution.

A. K Selection

As always, K-selection is a noteworthy question for K-required clustering algorithms. In cluster algorithms, the elbow method could be a reasonable heuristic to determine. Hence, we select both the silhouette coefficient and the average distance of each point to cluster centers to measure the performance while consistently changing the K value. A related result is shown in Figure. 5. Through abandon overly small and overly significant K values, we can tell that there is an elbow point where K is 8. Then we use eight as the number of clusters of K-means.

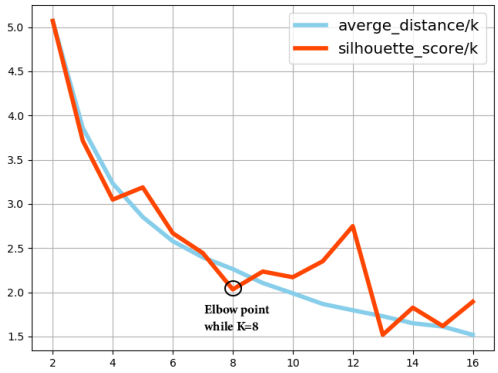

Figure 5: K value selection via elbow method.

B. Data Decomposition

To decomposite the full-scale features into two-dimension coordinate-like data for K-means, we select three different decomposition algorithms, which are Non-negative matrix factorization (NMF), t-distributed stochastic neighbor embedding (t-SNE) and the principal component analysis (PCA) as a comparison experiment. The result indicates that t-SNE can outperform the others as shown in Figure 6.

Figure 6: Decomposition results via t-SNE, PCA and NMF.

We project that clustering result of K-means on the composited data from t-SNE which are the best result that acquired by running it over 100,000 epochs while considering the silhouette coefficient, and it is shown in the world map.

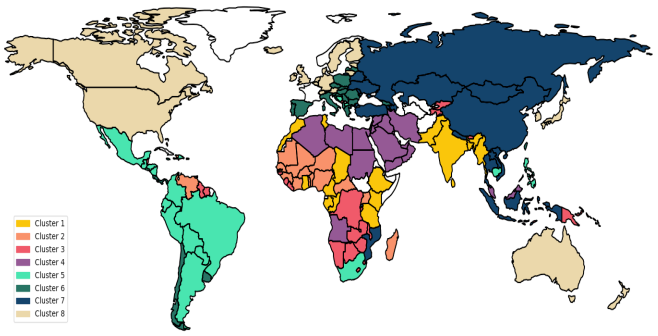

Figure 7: Clustering result reflected on the world map.

Here, each color represents a cluster class, and some countries may have different colors for different regions, such as Taiwan and Hongkong of China. The white area with black boundary means no statistical data on these areas, and it also means that country has very few residents or rather small. And it is worth pointing out that some places and Antarctica are removed from this world map due to limited resolution. The interesting thing is, the adjacent countries are very apt to be in the same cluster while there is no geographic data that has been used in the selected 67 essential features. It might indicate that adjacent countries can share a similar sense of freedom to some extent. Moreover, beyond the aforementioned phenomenon, some countries are surrounded by variant clusters which are contrary to the common laws.

VI. HFI EXPANSIONS

To discover other factors that can be mirrored by HFI score, we choose two indexes to do further discoveries, which are the immigration rate and GDP.

A. Immigration rate

In the first experiment, we use a scatter plot to show the relationship between HFI and immigration rate and the result is shown in Figure 8. Here, the x axis is the 'International migrant stocks', which is the number of immigrants in the country divided by the number of immigrants in the world and y axis is the related HFI score.

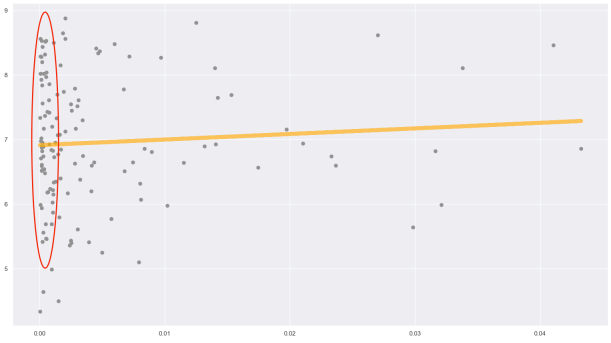

Figure 8: Correlation between the HFI and immigration rate.

It can be observed that despite the fact that the immigration rate of some countries is missing (135/165), they are still sharing a positive correlation in general. However, some countries which are labeled by the red circle are not involved in this phenomenon, which is strange and we may need further discussion on this point in the future. Therefore, the general conclusion is the higher HFI, the higher the immigration rate. However, the correlation between them is relatively low under some circumstances

B. Gross domestic product (GDP)

Gross domestic product (GDP) is a monetary measure of the market value of all the final goods and services produced in a specific period. GDP reflects differences in the cost of living and the inflation rates of the countries; therefore, using a basis of GDP is arguably more useful when comparing living standards between nations, while nominal GDP is more useful comparing national economies on the international market. It is natural thinking that the HFI will correlate with living standards, which can be reflected by GDP. Therefore, we try to discover the correlation between GDP and HFI score, and their related correlation is shown in Figure 9. Here, the x axis is GDP per-person and y axis is its related HFI score.

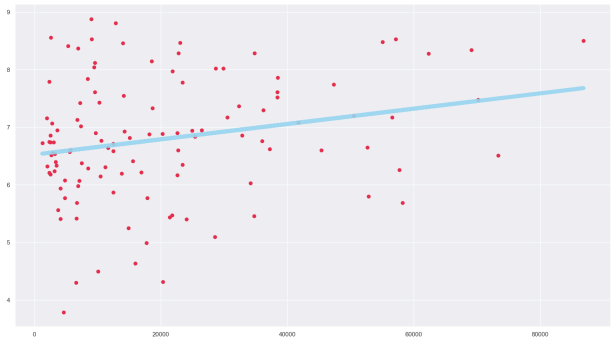

Figure 9: Correlation between the HFI and GDP.

From Figure 9, it can be observed that they are sharing a positive correlation overall. The fitting result indicates that the higher GDP, the higher HFI score, which is similar to the correlation between HFI and immigration rate. Moreover, this correlation is also in line with our intuitive feeling.

VII. CONCLUSION

In conclusion, in the data exploration report, we first dig out some general features for intuitive observation. Then, we do data cleaning and feature selection through correlation analysis, and we proposed our approach for feature selection. Moreover, regression analysis and clustering analysis are applied for understanding the country's freedom status. The regression analysis indicates that we can use Tree-based feature selection for predicting hf score and the number of selected features is 5, which is quite low and effective. The cluster analysis shows that adjacent countries are usually sharing a similar sense of freedom.

In this report, we rethink the filling approaches for missing cells and the conclusion is that the mean value filling can outperform the other approaches which are machine learning-based and deep learning-based filling. Moreover, we keep the regression and cluster analysis in this report since it is our objective. In addition, we do some expansions exploration related to the HFI score. The conclusion is that the HFI score has a positive correlation with GDP and immigration rate, which respects the initial hypothesis we made at the beginning.

The HFI score has shown its massive vitality and reference-able value. However, we still have a remaining problem in the correlation between HFI and immigration rate, a bunch of countries are not in line with the general phenomenon. Hence, we need to do more experiments to discover the potential reasons behind this in the future. Moreover, we could still continue discovering more indexes correlated with HFI to see the importance of freedom in different aspects of our life.