Explorations on the Levels of Happiness across Countries

02433236

1 Introduction

The World Happiness Report, annually published by the WHR editorial board, offers a concrete method to evaluate the extent of happiness across various countries. From "the World Happiness Report" (n.d.), the happiness score is based on a metric called Cantril Ladder, which asks the survey respondents to evaluate their current life satisfaction and happiness from 0 to 10 - where 0 represents the worst life evaluation and 10 represents the best possible life. In particular, we focus on the World Happiness Report from 2022 to 2024 to figure out the association between the ladder score and social factors within a country that are expected to contribute to the satisfaction of individual's happiness evaluation. By predicting the happiness score from different factors, we can gain insights into global life satisfaction. This gives government feasible directions to make policy decisions to enhance well-being and enables international organizations to assess their global work.

2 Data

2.1 Data collection

In this data science report, the world happiness datasets from 2022 to 2024 were collected directly from the survey results of the Gallup World Poll. The datasets also include six social and economic factors with continuous values for our further investigation. The values of each factor are calculated by comparing to the states of Dystopia (represents the worst possible conditions of six variables):

- Social Support: The extent of social support available to an individual.
- Perceptions of Corruption: Public responses to the corruption, which reflects people's confidence to their government.
- Healthy Life Expectancy: The expectance of living in a healthy life.
- Generosity: The generosity among the population, higher values indicate more common altruistic behaviours among population.
- GDP: Measure of the economic health of a country.
- Freedom to Make Life Choices: The extent of personal freedom in daily lives.

2.2 Data processing

The datasets of 2022 and 2023 share an identical data structure then the dataset of 2024 was preprocessed to align the same column names with those in other two datasets. There are different number of rows in three datasets (145 in 2022 dataset, 137 in 2023 dataset, 143 in 2024 dataset) indicating that different countries were researched across three years. Then after removing rows with missing values presenting in one row in the 2023 dataset and three rows in the 2024 dataset, the datasets retained the rows common to all three years, resulting in 131 rows per dataset. To provide an initial glimpse of the world happiness scores globally, we can then add a column specifying the region of each country to 2024 dataset following the country region in 2022 and 2023 datasets. By calculating the average scores of 10 regions each year, we can notice that North America, Australia, New Zealand and Western Europe reach the highest

average happiness scores each year while South Asia shows the lowest average happiness scores but also an increasing trend across three years.

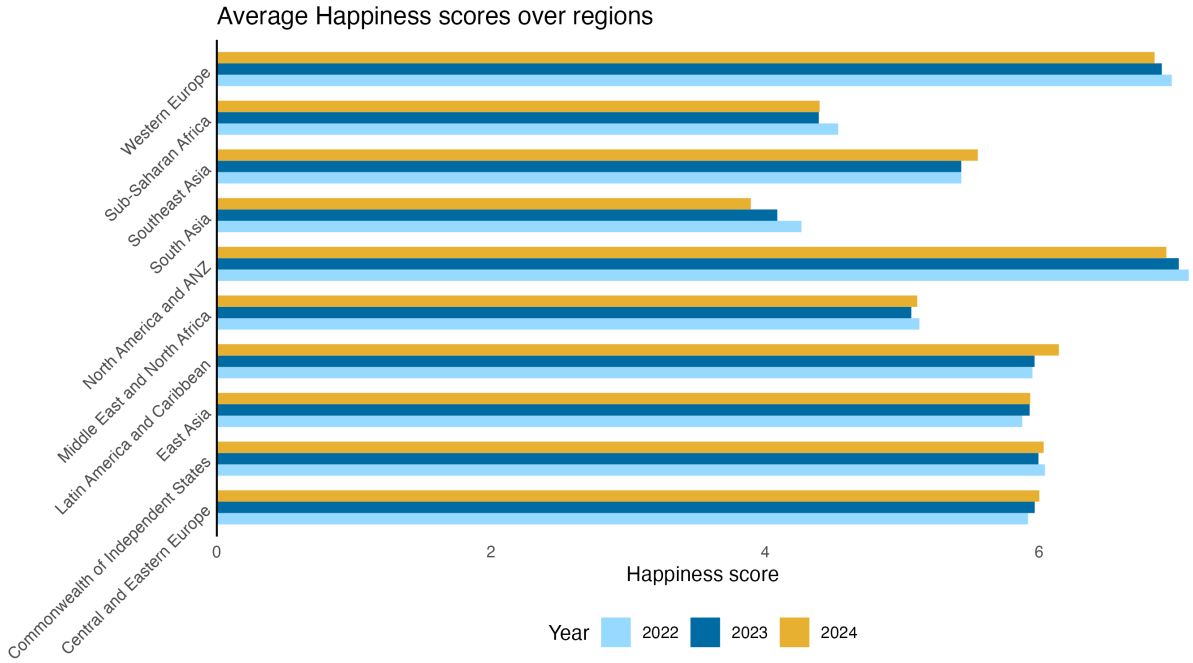

Figure 1: Histograms of Happiness Scores

To select the most significant features for happiness score model, we can first compute the correlations between variables and plot the correlation matrix in 2. The variable year has almost no correlation to the happiness score as expected, then we can use the merged dataset from 2022 to 2024 for predictions. Additionally, the correlation coefficient between generosity and the happiness score is also very small (0.05). Also, it can be noticed from the correlation plot that there are moderate strong positive correlations between healthy life expectancy and GDP (0.71) and social support and GDP (0.69). For further verification, we will perform Lasso regression to find the most predictive features and make reliable estimates even in the presence of multicollinearity by penalizing the coefficients. The social support and GDP of a country shows a strong correlation with the happiness scores.

3 Methods

The merged dataset is split into training sets (80%) and test sets (20%). After constructing the model and finding the optimal hyperparameters, the predictions are made on test set to check the model performance.

3.1 Lasso Regression

We first construct a Lasso Regression model, which adds a penalty term to the ordinary least squares model. Since the happiness score appears to have little correlation to the population generosity, the Lasso Regression model can penalizes the less important features. Lasso Regression model is implemented by adding regularisation to the loss function and we compute the estimated vector of coefficients $\hat{\boldsymbol{\beta}}$ by:

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{X}_i \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

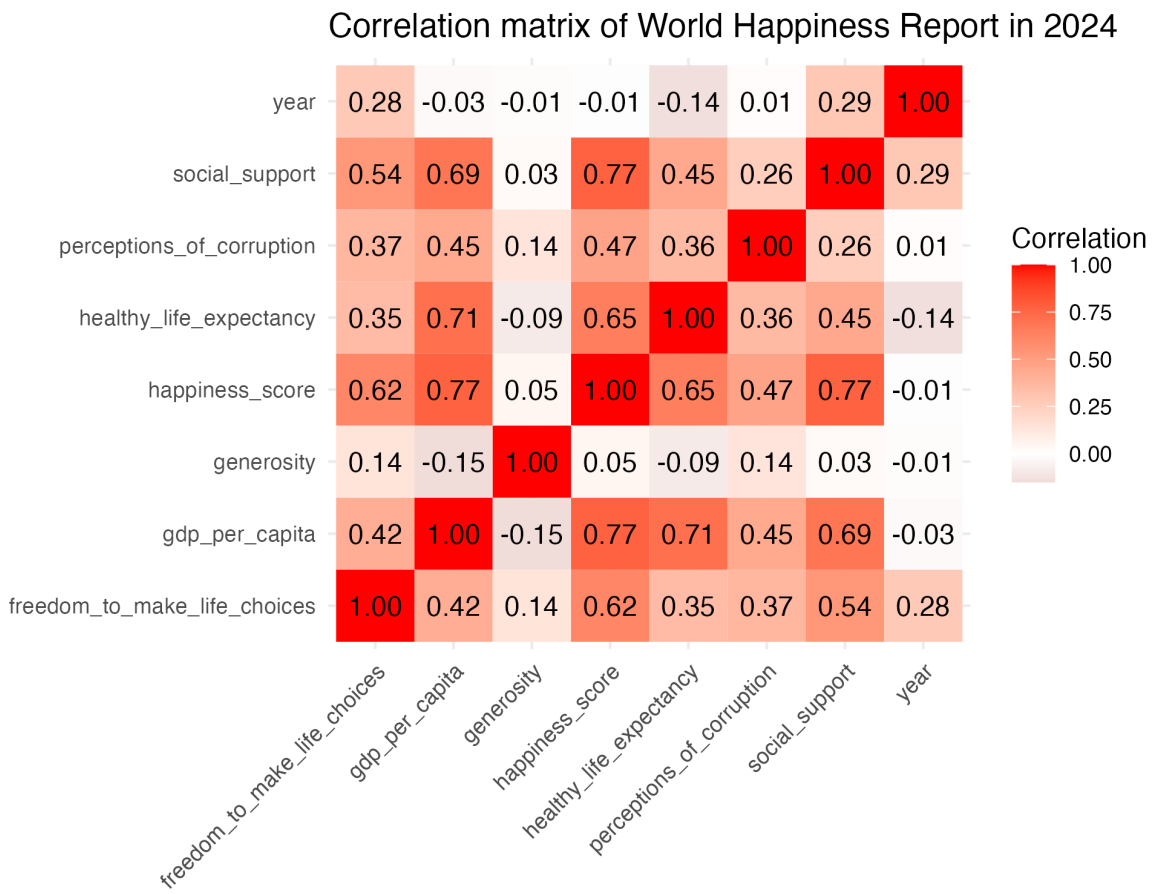

Figure 2: Correlation matrix

where $\boldsymbol{X}_i$ is the feature vector for i_{th} country, y_i is the observed happiness score for i_{th} observation and λ is the regularization parameter and determines the strength of the penalty. The hyperparameter λ is tuned by a 10-fold cross validation and the optimal λ is the one that mininmizes the cross-validation error.

3.2 Random Forest

Secondly, the Random Forest model is implemented as it can also model the nonlinearity presenting in the data. A Random Forest is composed of B decision trees $\{T_1, T_2, ..., T_B\}$ and each decision tree is built from a bootstrap sample from the training data. For each node in each decision tree, the algorithm chooses the optimal split value s by:

$$s = \underset{s}{\operatorname{argmin}} \left[\min_{c_1, c_2} \left(\sum_{i \in \mathcal{I}_1} (y_i - c_1)^2 + \sum_{i \in \mathcal{I}_2} (y_i - c_2)^2 \right) \right],$$

where I_1 and I_2 are the child nodes split by s, c_1 and c_2 are the mean responses for the left child nodes I_1 and right nodes I_2 respectively.

In this model, we set the number of trees to 200. After tuning the hyperparameters and model

construction, the predictions of the happiness score is made by:

$$\hat{y} = \frac{1}{200}\sum_{b=1}^{200}\hat{y}_b, \quad \text{for the } b^{th} \text{ tree.}$$

3.3 K-Nearest Neighbours Regression

The third model we applied is the K-Nearest Neighbours regression, which makes predictions by taking the average of the happiness score outputs of the nearest points. Firstly, the hyperparameter k is tuned by the 10-folds cross validation method within a range of k values and $k = 9$ was selected with the smallest root mean squared error. The distance between feature vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ is Euclidean distance:

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sqrt{\sum_{k=1}^{6}(x_{ik} - x_{jk})^2},$$

where k is the k^{th} feature in the predictors. The predictions $\hat{y}$ of the happiness scores based six social and economic factors are made by:

$$\hat{y}(\boldsymbol{x}) = \frac{1}{9}\sum_{x_i \in N(\boldsymbol{x})} y_i,$$

where $N(\boldsymbol{x})$ is the set of 9 nearest neighbours of $\boldsymbol{x}$.

4 Results

Lasso Regression The Lasso coefficients table 1 shows that the generosity among population is not correlated to the happiness score and the coefficient is set to zero by the Lasso model. In particular, the social support offered to individuals indicates a strong positive effect on the happiness score of a country increasing the happiness score by around 1.5199 with one unit increase in this variable. All other variables also show positive correlations with the happiness score. The optimal λ is 0.03191781.

Table 1: Regression Coefficients

Variable	**Coefficient**
Intercept	2.2116
GDP per Capita	0.2477
Social Support	1.5199
Healthy Life Expectancy	0.7941
Freedom to Make Life Choices	1.2379
Generosity	0.0000
Perceptions of Corruption	0.5727

Random Forest The Random Forest model explains 81.19% variance in the data suggesting a good fit and the mean squared error is 0.252. The histogram 3 shows how much the mean squared error is increased if the value of the variable is rearranged while keeping other variables unchanged. It demonstrates the importance of each feature and the generosity appears to contribute the less to the happiness score predictions. Additionally, people's freedom to make life choices and the social support are the two predictors that the most likely to contribute to the accuracy of the model.
K-Nearest Neighbours The K-Nearest Neighbours model show a much smaller mean squared error of 0.0657 compared to Lasso Regression and Random Forest (0.309, 0.252) indicating a very good fit to the world happiness dataset from 2022 to 2024.

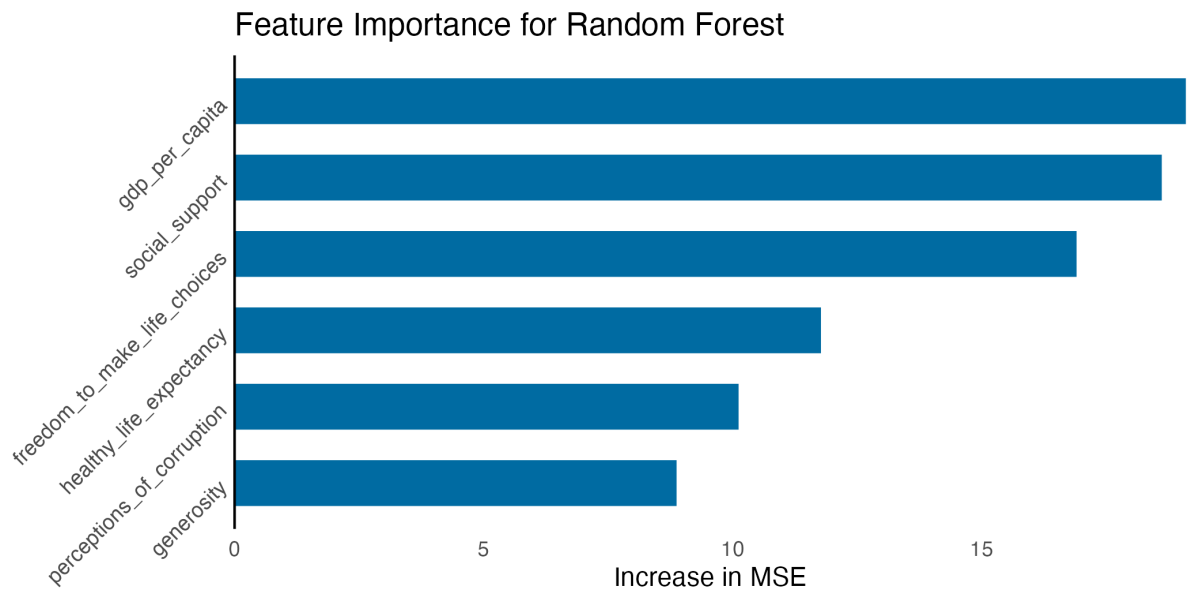

Figure 3: Feature Importance

Performance of Predictions Figure 4 shows the predictions of the happiness scores from the three models versus the true happiness scores in the test set. The dashed line $y = x$ is used to identify the fit of each model. All three models show a good fit to the data while the KNN model with $k = 9\%$ indicates an outstanding predictions on the test set.

Figure 4: Predictions made by three models

5 Conclusion

Although there appears to have some small fluctuations in the happiness scores from year to year from our analysis of the World Happiness Report from 2022 to 2024, years does not mainly relate to the happiness score of each country. It can also be concluded that the K-Nearest Neighbours model performs clearly better in predicting the happiness scores based on six main features than the Lasso Regression model and the Random Forest model. From the feature importance demonstrated from the Lasso Regression and the Random Forest model, both show that the freedom that people make life choices and the social support contribute the most to the accuracy of the predictions in happiness scores. However, the generosity of people tends to be the personal characteristics rather than the features related to the extent of happiness in a country.

6 Future work

In the process of exploratory data analysis, it can be noticed that different regions show quite different happiness scores. In addition to the six social factors we investigated in the report, more features like geographical features, weather, infrastructure of a country can also be collected to model the happiness scores. In terms of the model construction, there are various hyperparameters in the Random Forest such as the number of forests, the minimum leaves at each node can be further tuned to reach better prediction results.

7 Reference

World Happiness Report. (n.d.). Retrieved [10/05/2024], from https://worldhappiness.report/about/.