**Happiness Prediction**

**Neural Network Analysis**

Kumbham Nuthan Manideep

National Institute of Technology, Silchar

Data Mining Methods

Dr. Ripon Patigiri

Table of Contents

Page No.

Abstract3

Overview4

Data4

Research Questions5

Neural Network Analysis6

Conclusion14

Reference……………………………………………………………………………………15

**Abstract**

The World Happiness Report 2015 dataset contains data on the happiness levels of people in 159 countries around the world. It includes information on a variety of factors that are thought to contribute to happiness, such as economic factors (GDP per capita), social factors (family, health, and freedom), and trust in government. The data also includes a happiness score and a happiness rank for each country. The happiness score and happiness rank are based on a poll in which people were asked to rate their overall happiness on a scale of 0 to 10. The economy, family, health, freedom, trust, and generosity columns contain various measures of those factors. Overall, the World Happiness Report 2015 dataset provides valuable insights into the factors that contribute to happiness and well-being at the national level, and can be used by researchers, policymakers, and others to better understand and promote happiness around the world.

**Overview**

The goal of this project is to analyze how various factors impact happiness using Neural Network analysis, a powerful technique that combines machine learning with real-world data. The dataset, obtained from Kaggle, contains information on happiness levels in 159 countries in 2015, including factors such as GDP per capita, family, health, freedom, trust in government, and generosity. The first phase involved cleaning and preprocessing the data to ensure its suitability for analysis, and then splitting it into training and testing sets for model evaluation. The neural network model was trained on the training set, and its performance was assessed using the testing set. The best neural network model was determined based on validation data accuracy, and gains

plot and decile-wise graphs were plotted to illustrate the results.

**Data**

The dataset can be found on Kaggle and it contains data on happiness levels of people in 159 countries around the world. The data is collected in the year 2015 and it includes information on a variety of factors that are thought to contribute to happiness, such as:

| **Column Name** | **Description** |
| --- | --- |
| Country or Region | Name of the country |
| Happiness Score | A composite score of overall well-being |
| Economy (GDP per Capita) | Measure of the economic production of a country |
| Family | Measure of social support |
| Health (Life Expectancy) | Measure of the health of citizens |
| Freedom | Measure of freedom to make life choices |
| Generosity | Measure of generosity of citizens |
| Trust (Government Corruption) | Measure of trust in government |

In this dataset, the response variable is 'Happiness Score' and predictor variables are 'Economy (GDP per Capita)', 'Family', 'Health (Life Expectancy)', 'Freedom', 'Generosity', and 'Trust (Government Corruption)'.

**Research Questions**

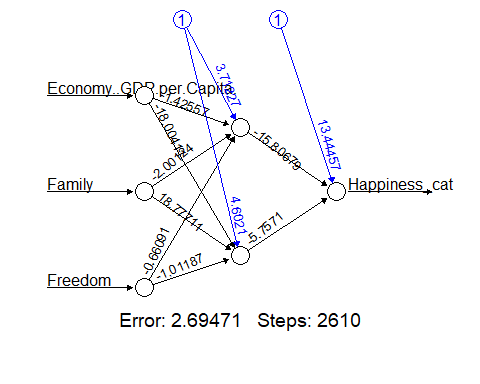
* How accurately can a neural network model predict happiness levels based on factors such as GDP per capita, family, health, freedom, trust in government, and generosity?
* Which of the factors, including GDP per capita, family, health, freedom, trust in government, and generosity, have the most significant impact on predicting happiness using a neural network model?
* Can a neural network model effectively identify nonlinear relationships between factors such as GDP per capita, family, health, freedom, trust in government, and generosity, and happiness levels?
* How does the inclusion of different combinations of factors, such as GDP per capita, family, health, freedom, trust in government, and generosity, affect the accuracy of the neural network model in predicting happiness levels?
* What is the predictive power of a neural network model in forecasting happiness levels using factors such as GDP per capita, family, health, freedom, trust in government, and generosity, and how does it compare to other traditional statistical methods for happiness prediction?

**Neural Network Analysis**

Firstly, the data from the dataset that is obtained from the sources is loaded, pre-processed and cleaned and the categorial variables are converted into dummy variables. Seed value was then set to a random number (here 666 is taken).

To start, pre-processing the data by taking the target variable and calculating the mean of that variable is done. Then a new binary variable based on this mean is created. The World Happiness Report 2015 dataset was divided into 60% training and 40% validation. The first set, the training set, would be used to fit the Neural Network model and would contain 60% of the total data. The second set, the validation set, would be used to evaluate the performance of the model and would contain 40% of the total data.

Using this prepared data, building the model will begin. The first model that is being constructed is a neural network with one hidden layer containing two neurons. The features used for this model will include Economy, GDP per Capita, Family, and Freedom.



The classification matrix for the training data for the model built with 2 neurons and 1 hidden

layer is shown below.

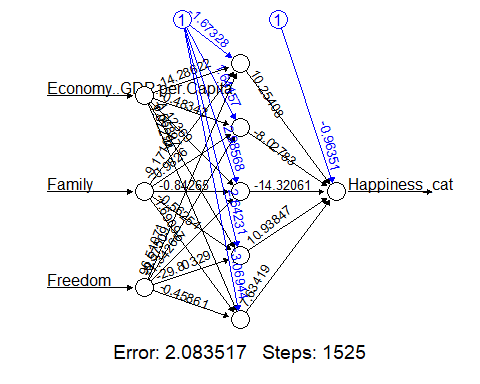
|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Predicted |  | 0 | 1 |
| 0 | 46 | 5 |
| 1 | 3 | 40 |

The accuracy of the model is 0.9148936.

The classification matrix for the validation data for model-1 is shown below. The accuracy of the model is 0.796875.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Predicted |  | 0 | 1 |
| 0 | 28 | 6 |
| 1 | 7 | 23 |

The second model that is constructed is also a neural network but with a slight difference. It will have one hidden layer containing five neurons. Like the first model, the features used for this model will include Economy, GDP per Capita, Family, and Freedom. This model will be an extension of the first model with more neurons in the hidden layer, allowing the model to learn more complex relationships between the features and the target variable.



The classification matrix for the training data for model-2 is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Predicted |  | 0 | 1 |
|  |  |  |
| 0 | 46 | 3 |
| 1 | 3 | 42 |

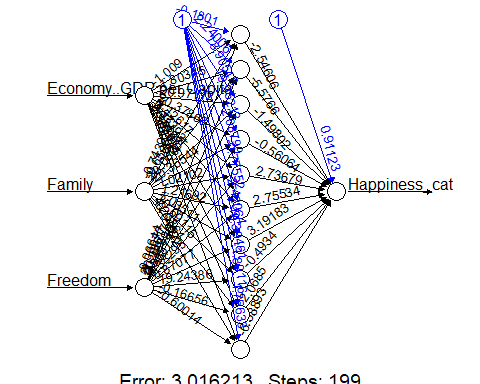
The accuracy of the model, when evaluated against the training data, is 0.9361702 or approximately 93.62%. This indicates that the model's performance in predicting the outcomes of the training data is quite high, with an accuracy rate of over 90%. It suggests that the model is able to correctly classify the training data with a high level of precision, capturing the underlying patterns and relationships within the data during the training process.

The classification matrix for the validating data for model-2 is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Predicted |  | 0 | 1 |
|  |  |  |
| 0 | 25 | 7 |
| 1 | 10 | 22 |

The accuracy of the model-2 for validation data is 0.734375.

The third model in the sequence is a neural network, akin to the previous two models, but with a distinct configuration of neurons in the hidden layer. Specifically, this model employs a single hidden layer that consists of 10 neurons. Similar to the first and second models, the features utilized in this model comprise Economy, GDP per Capita, Family, and Freedom. The increased number of neurons in the hidden layer results in a more intricate structure, allowing the model to capture more complex relationships between the features and the target variable. This heightened complexity may enable the model to discern nuanced patterns and correlations within the data, potentially enhancing its ability to make accurate predictions.



The confusion matrix for training data for model-3 is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Predicted |  | 0 | 1 |
| 0 | 46 | 7 |
| 1 | 3 | 38 |

The accuracy of the model for training data is 0.893617.

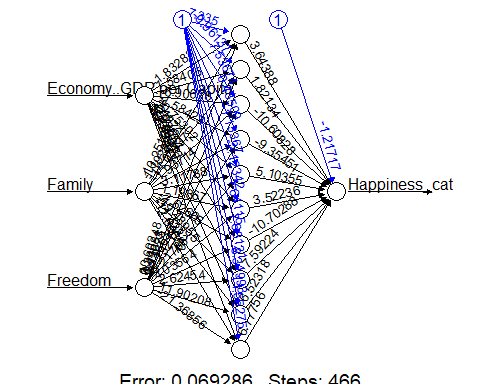
The confusion matrix for validation data for model-3 is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Predicted |  | 0 | 1 |
| 0 | 30 | 6 |
| 1 | 5 | 23 |

The accuracy of the model for validation data is 0.828125.

The analysis reveals that among the three constructed models, the neural network model with a single hidden layer and 10 neurons stands out as the best performer. When applied to the validation data, this model demonstrates the highest accuracy, with a value of 0.828125. In order to validate the model's performance further, the data has been scaled using the features of this top-performing model. This additional step ensures that the observed high accuracy is not merely due to chance, but rather indicates that the model is able to generalize effectively to previously unseen data. By utilizing the features of the best model for data scaling, it provides additional evidence of the model's robustness and reliability in making accurate predictions beyond the training data.

The Neural Network of the best model is visualised in R as shown below:



The classification matrix for the training data of the best model is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Predicted |  | 0 | 1 |
|  |  |  |
| 0 | 49 | 0 |
| 1 | 0 | 45 |

The accuracy of the scaled model for training data is 1.

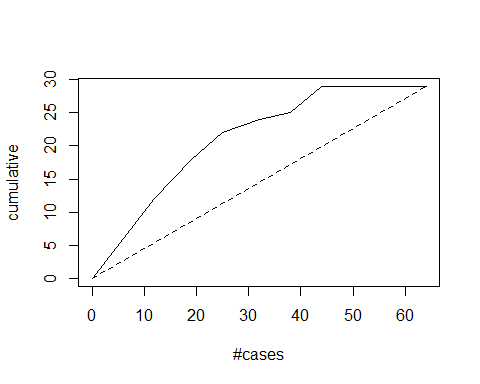
The classification matrix for the validation data of the best model is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
| Predicted |  | 0 | 1 |
|  |  |  |
| 0 | 25 | 9 |
| 1 | 10 | 20 |

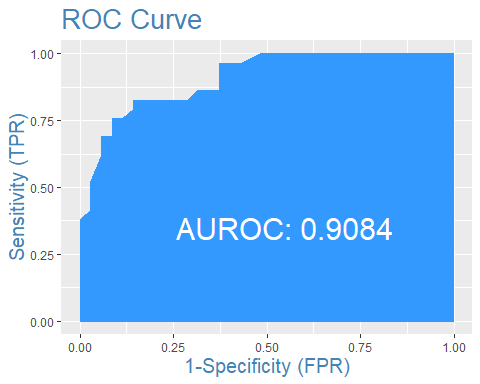
The accuracy of the scaled model for the validation data is 0.703125.

The results of scaling the data indicate that the accuracy of the best model remains consistently high, with only a slight decrease observed. This finding is significant as it implies that the performance of the model is not dependent on the specific scale of the input data, and the model can maintain its accuracy even when the data is scaled. This robustness to changes in data scale suggests that the model is likely to generalize well to new, unseen data, regardless of the scale of the input features. This quality makes the model highly reliable and practical for real-world applications.

The gains plot for the best model is shown below:



The gains plot demonstrates a direct proportionality between the area enclosed by the dotted line (representing random sample selection) and the model curve. A larger area signifies superior model performance. This plot serves as a visual tool to effectively represent the predictive power of the model and enables straightforward comparison across multiple models. It aids in identifying the significant features that impact happiness prediction and also assesses how well our model performs relative to a random sample selection. These insights are invaluable in gauging the accuracy and efficacy of our model in predicting happiness levels.



The AUROC (Area Under the Receiver Operating Characteristic) value of the curve is calculated to be 0.9084. This numerical metric represents the performance of the model in terms of its ability to correctly classify samples. A higher AUROC value, such as 0.9084, indicates better discriminatory power of the model in distinguishing between positive and negative samples. This value provides a quantitative measure of the model's performance and can be used to compare different models. In this case, a AUROC value of 0.9084 suggests that the model has a high level of accuracy and effectiveness in its classification task, further supporting its predictive power in the given context.

**Conclusion**

After thorough analysis of all the models, it can be concluded that the neural network model with features Economy, GDP per Capita, Family, and Freedom, and 1 hidden layer with 10 neurons, performs the best. This model achieves an accuracy of 0.703125, sensitivity of 0.6667, and specificity of 0.7353 based on the evaluation results. Notably, the standardization of the data does not impact the accuracy of the validation data, indicating the model's robustness to changes in feature scale. This suggests that the model is likely to exhibit good performance on unseen data and can be relied upon for practical use.

**References**

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, *4*(1), 1–58. <https://doi.org/10.1162/neco.1992.4.1.1>

Huang, G. (2014). An Insight into Extreme Learning Machines: Random Neurons, Random Features and Kernels. *Cognitive Computation*, *6*(3), 376–390. <https://doi.org/10.1007/s12559-014-9255-2>

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>