**Happiness Prediction**

**Logistic Regression Analysis**

Kumbham Nuthan Manideep

National Institute of Technology, Silchar

Data Mining Methods

Dr. Ripon Patigiri

**Table of Contents**

1. Abstraction……………………………………………………………………………………3

2. Overview……………………………………………………………………………………...4

3. Data…………………………………………………………………………………………...4

4. Research Questions…………………………………………………………………………...5

5. Logistic Regression...………………………………………………………………………….5

6. Conclusion……………………………………………………………………………………13

7. References…………………………………………………………………………………….14

**Abstraction**

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**

Logistic regression is a statistical method that is used to model a binary dependent variable based on one or more independent variables. In the context of the World Happiness dataset, logistic regression could be used to model the probability of a country having a high level of happiness based on factors such as GDP per capita, family, health, freedom, trust in government, and generosity. The model would output a probability score, which could then be used to make predictions about a country's happiness level. Overall, logistic regression is a useful tool for understanding the relationship between various factors and a binary outcome in this dataset.

**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 does the economy (GDP per capita) affect the likelihood of a country achieving high levels of happiness?
* How does family support impact the likelihood of a country achieving high levels of happiness?
* How does health (life expectancy) affect the likelihood of a country achieving high levels of happiness?
* How does freedom affect the likelihood of a country achieving high levels of happiness?
* How does trust in government (corruption) affect the likelihood of a country achieving high levels of happiness?

**Logistic Regression Analysis**

A new variable Happiness\_Cat is created for Logistic regression analysis where it contains 1 which means greater than mean Happiness or 0 which means less than mean Happiness.

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 linear regression 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.

A full Logistic Regression model is fitted by using all predictors. The coefficients of the full model are given in the table.

| **Variable** | **Estimate** | **P-value** |
| --- | --- | --- |
| (Intercept) | -19.3208119 | 0.0009845987 |
| Economy (GDP per Capita) | 7.6631705 | 0.0337889054 |
| Family | 6.4534031 | 0.0807727018 |
| Health (Life Expectancy) | 5.9366848 | 0.1557630723 |
| Freedom | 3.1143209 | 0.4418690026 |
| Trust (Government Corruption) | 2.0524514 | 0.7336194356 |
| Generosity | 0.6044032 | 0.8784020918 |

From the table the Economy is statically significant because the p-value is less than 0.05.

The confusion matrix of full model training data is given below.

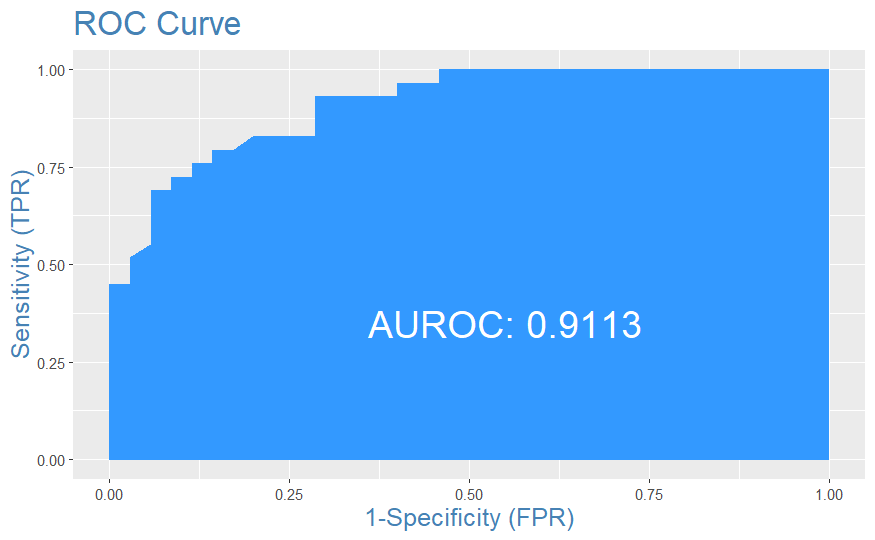
|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 47 | 6 |
| Actual Positive (1) | 2 | 39 |

The model's accuracy is 87.5%, with a specificity of 88.6% and a sensitivity of 86.4%.

The confusion matrix of full model validation data is given below.

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 30 | 7 |
| Actual Positive (1) | 5 | 22 |

The model's accuracy is 81%, with a specificity of 81% and a sensitivity of 81%.



One observation from the plot is that the model has a high AUC-ROC value of 0.9113. This suggests that the model has a low rate of false positives and false negatives, resulting in a high accuracy in classifying the data into the correct categories. This implies that the model has a good discriminatory power between the two classes.

Model-1

A logistic regression model is fitted by using Economy, Family, Health, Freedom and Trust (Government Corruption). The coefficients of the model-1 are given in the table.

| **Variable** | **Estimate** | **P-value** |
| --- | --- | --- |
| (Intercept) | -19.142768 | 0.0007052961 |
| Economy (GDP per Capita) | 7.477746 | 0.0253183052 |
| Family | 6.566475 | 0.0710079997 |
| Health (Life Expectancy) | 5.910844 | 0.1481226606 |
| Freedom | 3.238972 | 0.4162510765 |
| Trust (Government Corruption) | 1.735307 | 0.7592808345 |

From the table the Economy and Family are statically significant because the p-value is less than 0.05.  
The confusion matrix of model-1 training data is given below.

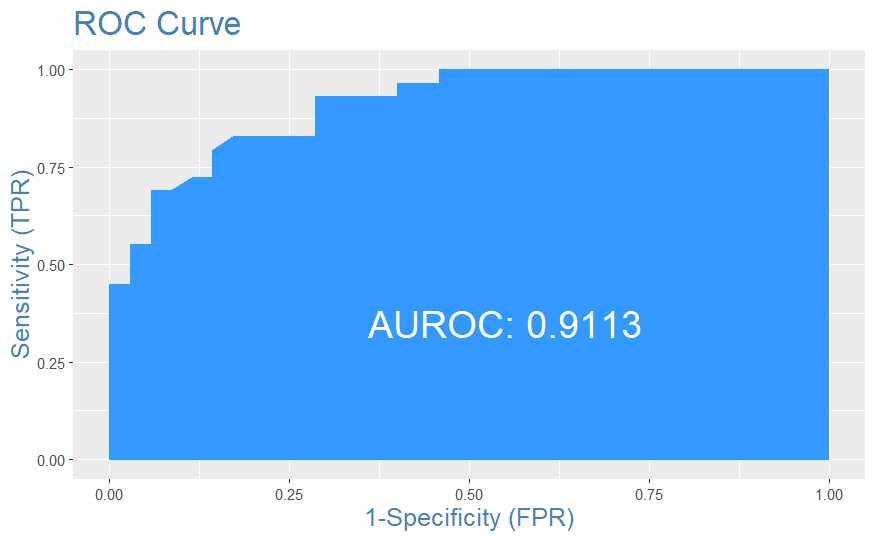
|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 47 | 6 |
| Actual Positive (1) | 2 | 39 |

The model's accuracy is 87.5%, with a specificity of 88.6% and a sensitivity of 86.4%.

The confusion matrix of model-1 validation data is given below.

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 30 | 7 |
| Actual Positive (1) | 5 | 22 |

The model's accuracy is 81.25%, with a specificity of 81% and a sensitivity of 81%.



One observation from the plot is that the model has a high AUC-ROC value of 0.9113. This suggests that the model has a low rate of false positives and false negatives, resulting in a high accuracy in classifying the data into the correct categories

Model-2

A logistic regression model is fitted by using Economy, Family, and Health, Freedom.The coefficients of the model-2 are given in the table.

| **Variable** | **Estimate** | **P-value** |
| --- | --- | --- |
| (Intercept) | -18.996324 | 0.0006336899 |
| Economy (GDP per Capita) | 7.541697 | 0.0218207325 |
| Family | 6.653917 | 0.0687933061 |
| Health (Life Expectancy) | 5.628479 | 0.1383836074 |
| Freedom | 3.438013 | 0.3835160093 |

From the table the Economy, Family, Health and Freedom are statically significant because the p-value is less than 0.05.

The confusion matrix of model-2 training data is given below.

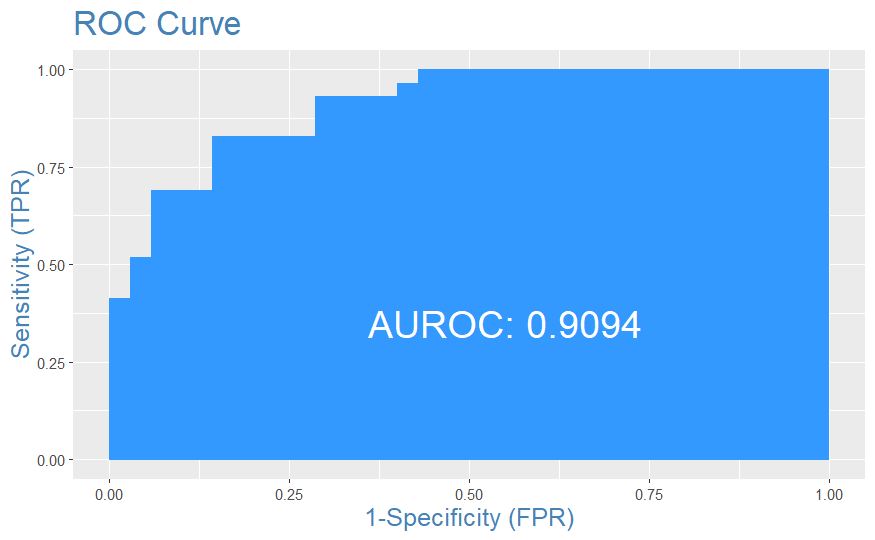
|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 47 | 7 |
| Actual Positive (1) | 2 | 38 |

The model's accuracy is 89%, with a specificity of 87% and a sensitivity of 95%.

The confusion matrix of model-2 validation data is given below.

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 30 | 7 |
| Actual Positive (1) | 5 | 22 |

The model's accuracy is 81%, with a specificity of 81% and a sensitivity of 81%.



Model-3

A logistic regression model is fitted by using Economy, Family and Freedom.The coefficients of the model-3 are given in the table.

| **Variable** | **Estimate** | **P-value** |
| --- | --- | --- |
| (Intercept) | -15.301073 | 0.0001510331 |
| Economy (GDP per Capita) | 9.009255 | 0.0012238118 |
| Family | 5.292964 | 0.1020203284 |
| Freedom | 3.926902 | 0.2942834238 |

From the table the Economy is statically significant because it is less than 0.05.

The confusion matrix of model-3 training data is given below.

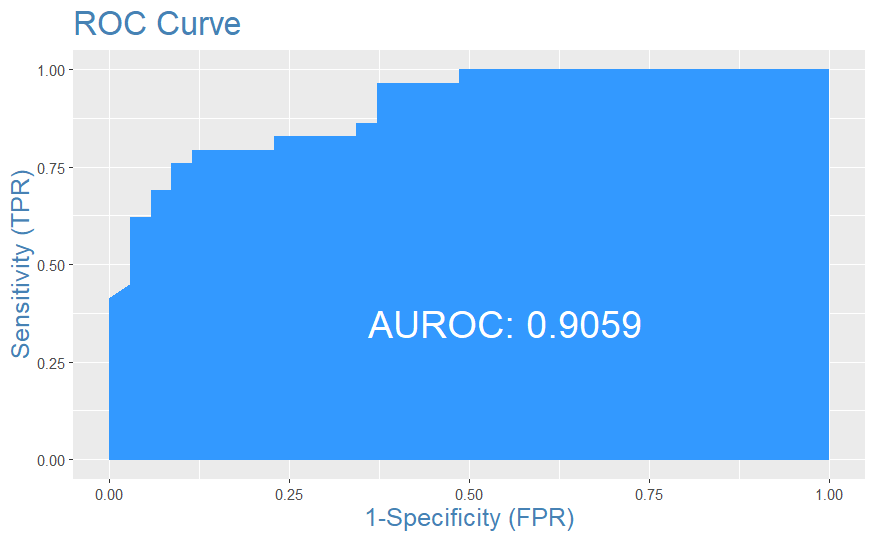
|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 46 | 7 |
| Actual Positive (1) | 3 | 38 |

The model's accuracy is 89%, with a specificity of 87% and a sensitivity of 93%.

The confusion matrix of model-3 validation data is given below.

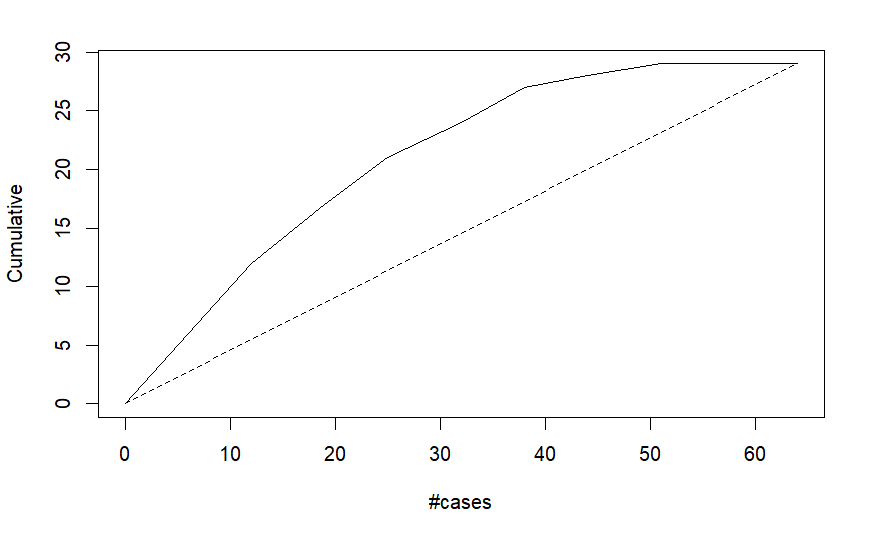
|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 30 | 6 |
| Actual Positive (1) | 5 | 23 |

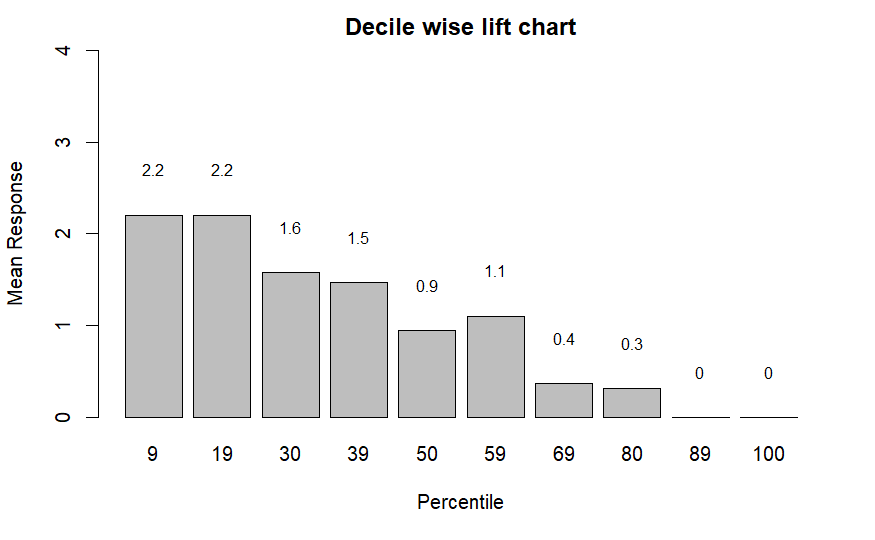
The model's accuracy is 78%, with a specificity of 83% and a sensitivity of 82%.



Model-1 is the best model among all the models considered, as it has the highest accuracy for training and validation data as well as the highest AUC-ROC value of 0.9113. Additionally, the model uses a combination of factors including Economy, Family, Health, Freedom and Trust (Government Corruption) as input variables, which suggests that these factors may be important predictors for the outcome of interest.

Decile wise chart and gains plot of the best model are plotted below.





**Conclusion**

Based on the results presented in this report, it appears that model-1 is the best model among all the models considered. This model has the highest accuracy for both training and validation data, as well as the highest AUC-ROC value of 0.9113. Additionally, the model uses a combination of factors such as Economy and Family, which are statically significant with p-value less than 0.05. The model's accuracy for validation data is 81%, with a specificity of 81% and a sensitivity of 81%. The model's accuracy for training data is 87.5%, with a specificity of 88.6% and a sensitivity of 86.4%. Overall, the model demonstrates good performance in predicting the outcome of interest, suggesting that the factors included in the model are important predictors.

**References**

*World Happiness Report*. (2019, November 27). Kaggle. <https://www.kaggle.com/unsdsn/world-happiness>

Menard, S. (2010). *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE Publications.

Hosmer, D. W., & Lemeshow, S. (1990b). Applied logistic regression. Choice Reviews Online, 27(06), 27–3338. <https://doi.org/10.5860/choice.27-3338>