Regression model for Happiness score

Sai Vamsi Ogety*,* Olesch Ansel Dsouza, Sri Charan Reddy Padidam

# INTRODUCTION

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating the International Day of Happiness on March 20th. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and shows how the new science of happiness explains personal and national variations in happiness.

The happiness scores and rankings use data from the Gallup World Poll. The scores are based on answers to the main life evaluation question asked in the poll. This question, known as the Cantrell ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The scores are from nationally representative samples for the years 2013-2016 and use the Gallup weights to make the estimates representative. The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world’s lowest national averages for each of the six factors. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[[4]](https://en.wikipedia.org/wiki/Linear_regression#cite_note-4) This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine. Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other (for example, higher SAT scores do not *cause* higher college grades), but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables. A linear regression line has an equation of the form ***Y = a + bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

A multilayer perceptron (MLP) is a class of feedforward artificial neural network. An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training] Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

Multilayer perceptron’s are often applied to supervised learning problems: they train on a set of input-output pairs and learn to model the correlation (or dependencies) between those inputs and outputs. Training involves adjusting the parameters, or the weights and biases, of the model in order to minimize error. Backpropagation is used to make those weigh and bias adjustments relative to the error, and the error itself can be measured in a variety of ways, including by root mean squared error (RMSE).

# Method

*A. Data Pre-processing*

In order to apply the model on the training data, the data should not have any missing values. We checked for missing values in all the datasets but fortunately there is no data that is missing. The data should contain only numerical columns in order to do linear regression, so we dropped the columns with text data. One of the things we noticed is that the features are not same across the three datasets. We have decided to keep the columns which are common to all three datasets because we must append three datasets into a single pandas dataframe and having same columns is one the constraint.

*B. Linear regression*

We are using sklearn library to create a linear regression model. First, we create an instance of the Linear regression

Model then we fit our training data to the model. Now that we have our model trained, we use prediction method to predict happiness score which is our target variable. In order to see how well our model fitted to our training data we have used Mean-Square-Error as the metric.

C*. Experiments*

In order to see which features are contributing more in predicting the happiness score we have divided our seven features’Economy','Family','Health','Freedom','Trust','Generosity','Dystopia' into three sets.

Our First set consists of features ’Economy', 'Family' and 'Health'. Second set consists of ‘Freedom’ and ‘Trust’ and last set consists of ‘Generosity’ and ‘Dystopia’.

Before we started our experiments, we selected all features and ran our model. These are the distribution plots for this experiment.

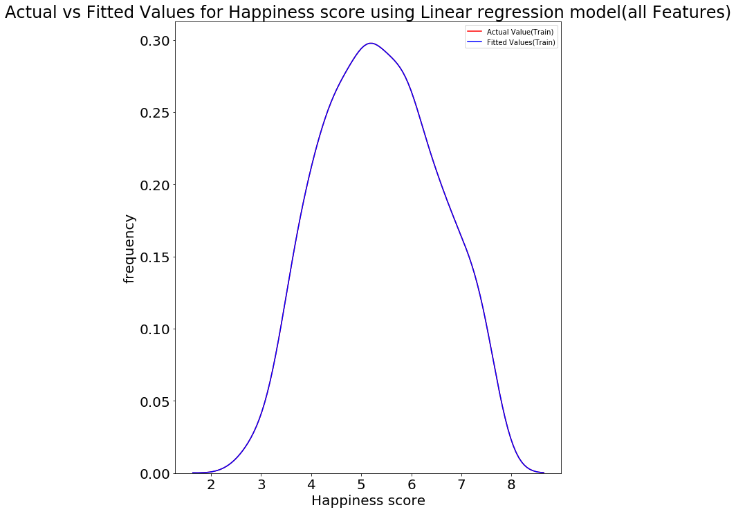


Fig1. Distribution plot between Actual and fitted value Linear Regression with all the features.

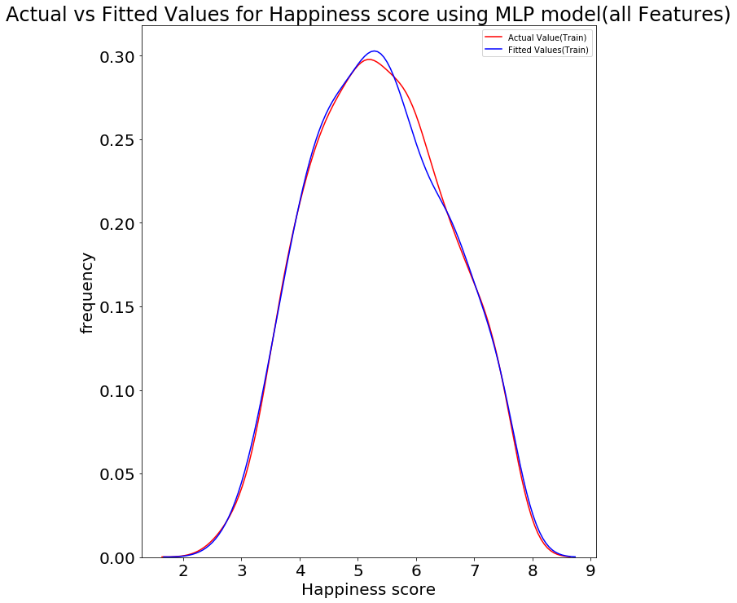


Fig2: Distribution plot between Actual and fitted value MLP with all the features.

The same plots with only two features gave these results.

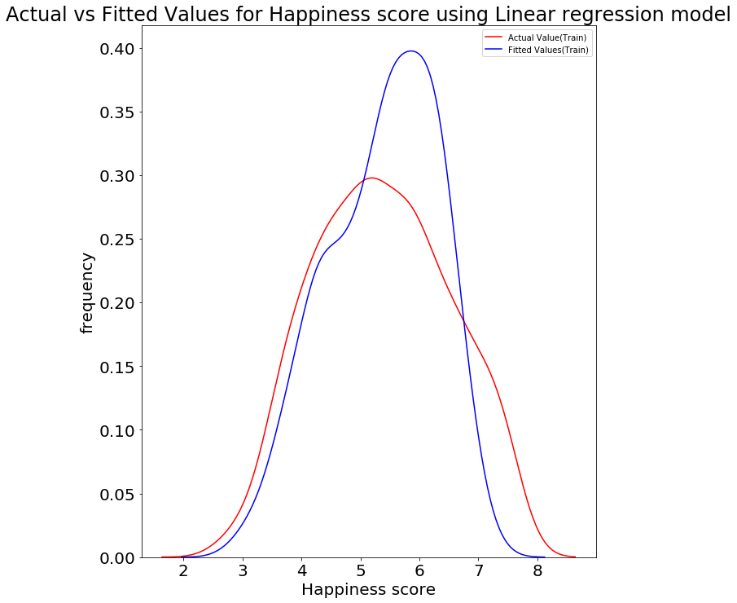


Fig3. Distribution plot between Actual and fitted value Linear Regression with only two features.

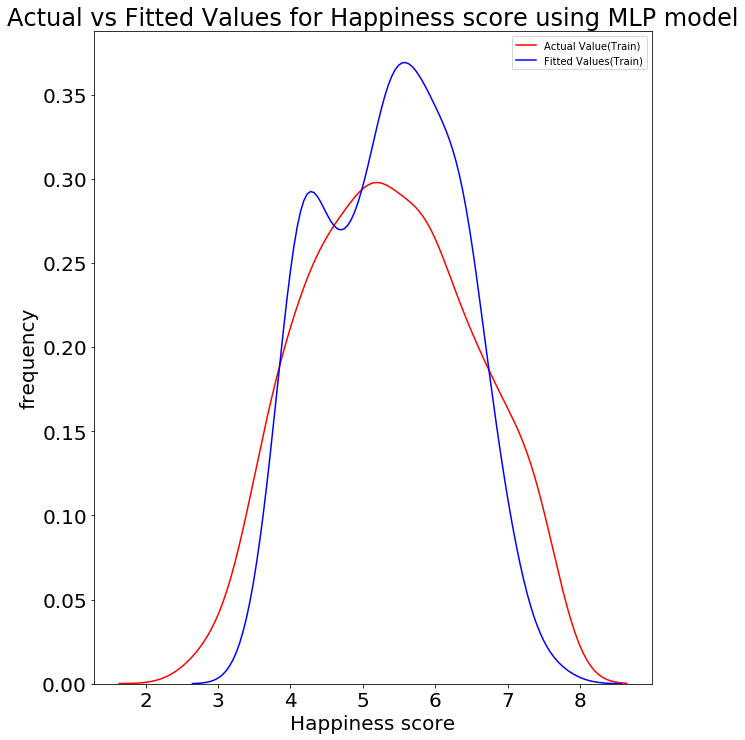


Fig4. Distribution plot between Actual and fitted value MLP with only two features.

We have also made 3-D plots to compare between Target value and the predicted value. The following are the plots.

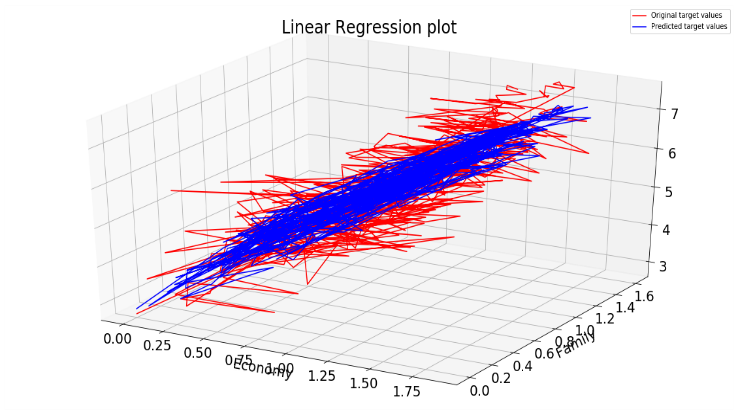


Fig5. 3-D plot for Linear Regression with two features.

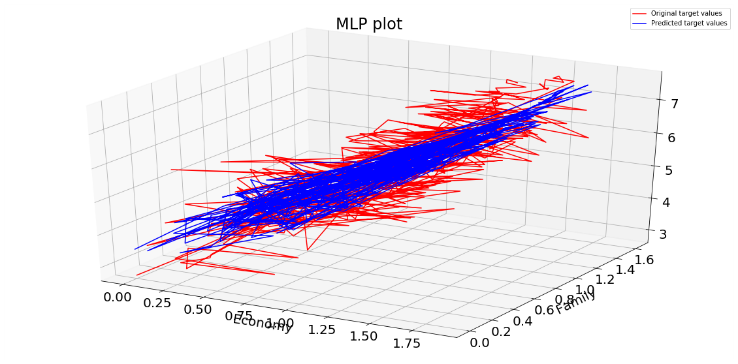


Fig6. 3-D plot for MLP with two features.

# Conclusions

Linear regression model overfitted the training data with an error of 8.1869445091806e-08 when we predicted values using all the values. The MLP model fitted well for the training data with an error of 0.017458290440188224. We can say that linear regression model performed better here when we used all the features to predict the target variable.

The table below shows the Means-Square error for different set of features.

|  |  |  |
| --- | --- | --- |
| Features | MSE(linReg) | MSE(MLP) |
| First Set | 0.38313 | 0.31557 |
| Second Set | 0.43288 | 0.41214 |
| Third Set | 0.92976 | 0.81252 |

We can see from the results above that when only few features were selected MLP model performed better than the Linear Regression model.