Regression Analysis of World Happiness Score

Aalap Doshi
Electrical and Computer Engineering
Arizona State University

Poojan Patel
Electrical and Computer Engineering
Arizona State University

Varadaraya Ganesh Shenoy
Electrical and Computer Engineering
Arizona State University

***Abstract*—In this report, we create a regression model for the happiness score using the provided features. This is done via two ways (ie) Linear Regression and Multi-Layer Perceptron.First,a linear regression model for the given dataset is created, then the RMSE for the corresponding model is recorded.The same is repeated using Multi-Layer Perceptron which performs as well as the linear regression model.[Si20]**

***Index Terms*—Linear Regression, MultiLayer Perceptron, RMSE**

I. Introduction

Regression analysis is a set of statistical processes to estimate the relationships between a dependent variable and one or more independent variables. One of the most popular implementation of regression analysis is linear regression, in which a researcher attempts to find a line (or a complicated linear combination) that fits the data most closely according to a specific mathematical criterion. This method of analysis is used for two conceptually distinct purposes. First, regression analysis is mainly used for prediction and forecasting, substantially overlapping with the field of machine learning. Second, it helps to infer causal relationships between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset.[Wikc]

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.[Wika]

Multi-Layer Perceptron is a class of feed-forward artificial neural network. It typically refers to networks composed of multiple layers of perceptrons with threshold. Multilayer perceptrons with a single hidden layer are ccolloquially called "vanilla " neural networks. It usually contains at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node has a neuron that uses a nonlinear activation function.This model helps to distinguish data that is not linearly separable.[Wikb]

Root Mean Square Error (RMSE) is the standard deviation of the prediction errors. Prediction errors also referred to as,residuals are a measure of how far the data points are from the regression line.RMSE signifies the concentration of data around the line of the best fit. [SH]

In this project we have developed a linear regression model for the happiness score of different countries base on different parameters such as GDP per capita, Economy, Generosity, Freedom, Health (Life expectancy) etc. The training of the model was done by using a package called Linear regression in python and then predicting the value of our data based on the actual data given. The accuracy of our predication based on actual data was done by measuring the amount of error. We measured different types of data such as root mean square error (RMSE), Mean square error (MSE) and Mean absolute error (MAE). By doing this error calculation we were able to determine how good our prediction is based on trained data. In the second part of our project we trained a multilayer perceptron with one hidden layer. All the analysis such asLinear regression was done on multilayer perceptron as well. The analysis of multilayer perceptron in python by using the package MLPRegressor[Mlp]. After training both the models, we compared the RMSE for both the models and saw how closely related both are.

II. Dataset and Model Parameters

The World Happiness Score dataset for the year 2015 is used for the regression analysis. [Net]

In this paper, the features used for linear regression are: Economy (GDP per Capita), Family, Health(Life Expectancy), Freedom, Trust (Government Corruption). It is implemented on the linear regression model using [Sci]. For the MLP model created using [Mlp] , all the features are used. Finally, RMSE is recorded for both models using Eq(1).

$$RMSE = \sqrt{\overline{(f-o)^2}} \tag{1}$$

where,

f = expected values,
o = observed values

III. Methods

The Linear Regression Model used for the analysis takes the form of Eq(2).

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon \tag{2}$$

where ,

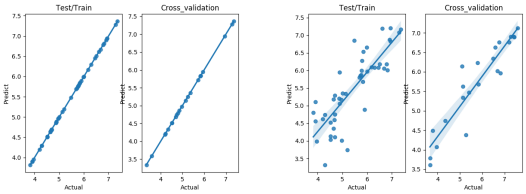
$$i = 1,...,n$$

n = number of statistical units

The MLP model used here has 15 neurons in its hidden layer to extract the features of the dataset.

IV. Results

In this section,the results of the model implementation is inspected and corresponding plots are depicted.



(a) Linear Regression using all features (b) Linear Regression using first 5 features

Fig. 1: Linear Regression Analysis

Fig. 2: 3D plot for Economy vs Family vs Happiness

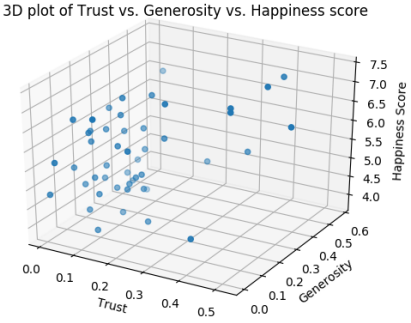
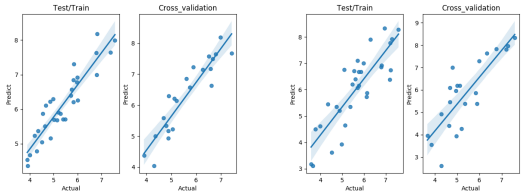


Fig. 3: 3D plot for Trust vs Generosity vs Happiness



(a) MLP Regression using all features (b) MLP Regression using first 5 features

Fig. 4: MLP Regression Analysis

Fig. 5: 3D plot for Economy vs Family vs Happiness

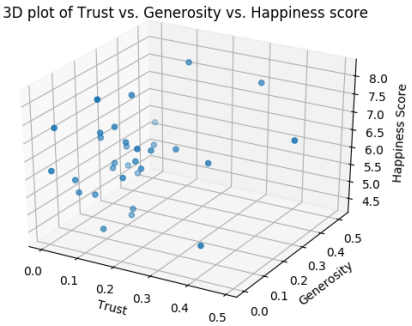


Fig. 6: 3D plot for Trust vs Generosity vs Happiness

Features	Train-Test	Cross-Validation
7	0.0003	0.0002
5	0.5866	0.5602

(a) Linear Regression Model

Features	Train-Test	Cross-Validation
7	1.0114	0.8131
5	1.0325	0.9890

(b) MLP Regression Model

TABLE I: RMSE for the regression models

From the plots and metrics obtained, it's quite obvious that the regression using MLP is not recommended as RMSE is almost 1. On the other hand, Linear Regression achieves an

RMSE of 0.003 when all features are utilized. Thus,for this dataset the linear regression model gives a lower RMSE and reveals better relation between dependent and independent variables.

V. Conclusions

In this paper, the 2015 dataset of the World Happiness Report is analyzed using regression models. The linear regression model results in a train-test RMSE of 0.0003 and 0.5866 while using all and 5 features respectively. The primary objective of this work is bound by implementation of these models on dataset of a single year while the datasets for the next few years is available. In future, the above models can be implemented for datasets for next few years , while also using the regression analysis of one dataset to reveal the relation of an existing feature in the upcoming year to the happines score. Although, this would be affected by the political and demographic issues for that particular year the relation can help realize the significance of each feature to the happiness score.Another regression model, Logistic Regression can also be applied. It is said that Logistic Regression is highly recommended when the response variable is categorical in nature, it can still be applied to investigate the relevance of these type of models to examine the correlation of the dependent variable to independent variable.

References

[Si20] Jennie Si. "Team Assignment – A regression model for happiness score". In: *EEE 511- Spring 2020* (2020).

[Net] Sustainable Development Solutions Network. *World Happiness Report*. URL: https://www.kaggle.com/unsdsn/world-happiness.

[Sci] *sklearn.linear*$_{m}odel.LinearRegression$. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[Mlp] *sklearn.neural*$_{n}etwork.MLPRegressor$. URL: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html.

[SH] Statistics-HowTo. *RMSE-Root Mean Square Error*. URL: https://www.statisticshowto.datasciencecentral.com/rmse/.

[Wika] Wikipedia. *Linear Regression*. URL: https://en.wikipedia.org/wiki/Linear_regression.

[Wikb] Wikipedia. *Multilayer Perceptron*. URL: https://en.wikipedia.org/wiki/Multilayer_perceptron.

[Wikc] Wikipedia. *Regression Analysis*. URL: https://en.wikipedia.org/wiki/Regression_analysis.