Regression Models for Happiness Score

Prasanth Sukhapalli
Department of Computer Science
Arizona State University
Tempe, Arizona, USA
psukhapa@asu.edu

Manish Vishnoi
Department of Computer Science
Arizona State University
Tempe, Arizona, USA
mvishnoi@asu.edu

Ratan Ravindra Shenoy
School of Electrical Engineering
Arizona State University
Tempe, Arizona, USA
rshenoy2@asu.edu

***Abstract*—The World Happiness Report uses survey data, demographic information and statistical analysis to rank 155 countries based on the happiness of their people. A country's overall happiness score is a composite of different factors including GDP per capita, healthy life expectancy and perceptions of corruption. This visualization charts happiness scores with independent datasets to explore possible trends that might also contribute to a countrys happiness. Using these variables, we constructed a linear regression model as well as a multilayer perceptron (MLP) which would help us predict the happiness score in these 155 countries. A comparison between the multi-layer perceptron model and linear regression models was conducted based on the RMSE of both the models to check the accuracy of both the models. The results concluded that the RMSE of the linear regression model is slightly better to the RMSE found by the multi-layer perceptron model for the given small dataset.**

***Index Terms*—Perceptron, RMSE, Regression.**

I. Introduction [1], [2], [3]

The World Happiness Report is a landmark survey of the state of global happiness which ranks approximately 157 countries by their happiness levels. The report continues to gain global recognition as governments, organizations and civil societies 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 columns following the happiness score estimate the extent to which each of six factors Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, Dystopia Residual contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the worlds lowest national averages for each of the six factors. Using these seven variables, we can then attempt to construct a linear regression model which may help us predict the happiness score in the 155 countries. We can then compare the predicted score to the actual score to observe how accurate how model is. Moving on, while were using these variables to just build an understanding of the ranking we can still see which variable(s) are highly correlated with happiness score and if there are any differences in these variables between 2015 to 2017. Patterns in the data can often be summarized by means of mathematical models. One of the simplest patterns (or models) for data on a pair of variables is a techniques for fitting lines to data and checking how well the line describes the data are called linear-regression methods. Using these methods, we can examine the relation between a change in the value of straight line. The change in the main variable of interest in the study The two variables have somewhat different roles in the interpretation of a linear regression. One of the variables, called the predictor variable or the independent variable.

II. Method and Approach [7]

In our paper, we will compare Perceptron and Linear regression results on the same database. The perceptron is a linear classifier an algorithm that classifies input by separating two categories with a straight line. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

This same concept can be extended to cases where there are more than two variables. This is called multiple linear regression. For instance, consider a scenario where you have to predict the price of the house based upon its area, number of bedrooms, the average income of the people in the area, the age of the house, and so on. In this case, the dependent variable(target variable) is dependent upon several independent variables. A regression model involving multiple variables can be represented as:

$$y = b_0 + m_1 b_1 + m_2 b_2 + m_3 b_3 + ... m_n b_n$$

The perceptron is a linear classifier an algorithm that classifies input by separating two categories with a straight line. Input is typically a feature vector x multiplied by weights w and added to a bias b:

$$y = wx + b$$

Perceptrons produce a single output based on several real-valued inputs by forming a linear combination using input weights (and sometimes passing the output through a non-linear activation function). In math terms:

$$y = \varphi(\sum_{i=1}^{n}) w_i x_i + b = \varphi(w^T x + b)$$

In this article, we will briefly study what linear regression is and how it can be implemented using Scikit-Learn, which is one of the most popular machine learning libraries for Python. Similarly, we will implement perceptron with only one hidden layer using scikit-learn. In the end, we will compare and the results.

A. Data Analysis [5]

Before moving on applying linear regression and perceptron, we will do some basic data analysis to look at our data distribution and select relevant columns as features. Data analysis will help us removing unwanted data and resolving data discrepancies.

- Happiness Score database provides us with 3 years of data, 2015, 2016 and 2017 of approximately 157 countries. Happiness score is directly proportional to Happiness rank, i.e higher the score better the rank.
- The basic features of happiness score for all three years are :

TABLE I: Common description of Happiness Score

Iterations	2015	2016	2017
Mean	5.38	5.38	5.35
std	1.15	1.14	1.13
min	2.84	2.90	2.69
max	7.59	7.53	7.53

- Every year has multiple factors to reach at that score, but to make our results more universal, we analysed only common features across all three years. Which are :
    - Economy (GDP per Capita)
    - Family
    - Health (Life Expectancy)
    - Freedom
    - Trust (Govt. Corruption)
    - generosity

Once data is homogeneous across all data sources, we will now find most relevant features and then feed them to our Linear Regression and perceptron models. We will compare results among all features and most relevant features to figure out that whether it helped us to feed only relevant features or more features the better.

B. Feature Selection [4], [6]

As we have selected some common features, we will use different techniques to find more relevant features than others. If a feature is effecting more to happiness score than that feature is more important.

- First technique that we are using is heat map. Heat maps will show us how relevant a feature is to our class. Generally speaking a feature is directly effecting our class if score is above 0.66 or below -0.66.
  We can clearly observe that below three features are highly co related to the Happiness Score,
    - Economy (GDP per Capita) (0.77)
    - Family (0.72)
    - Health (Life Expectancy) (0.71)

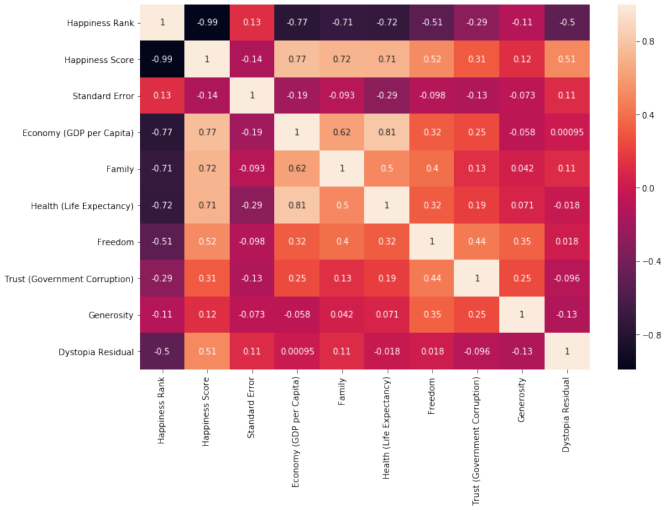

Fig. 1: Heat Map for different features with Happiness Score

- Another way to check which features are ore relevant is pairplot from seaborn library. It will create a pairwise plot of all features/columns in a data base. And more linear and dense a feature is to happiness score, more relevant it is.

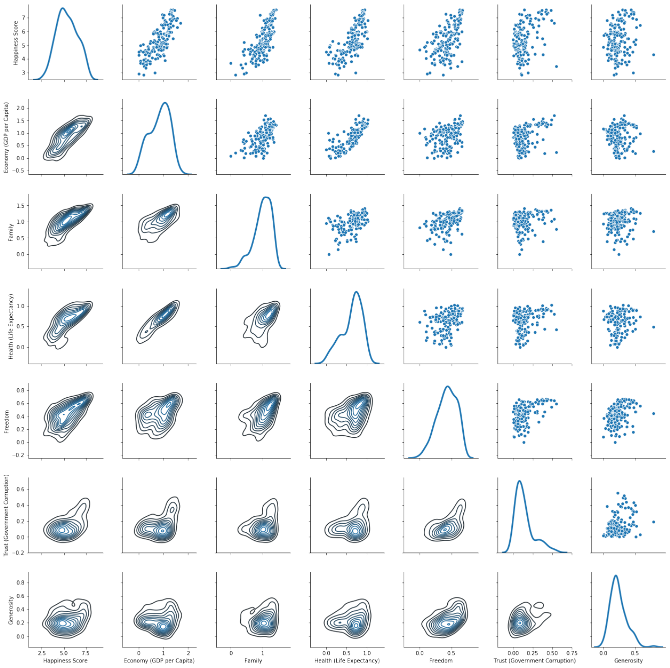

Fig. 2: Pairwise relevancy plots

  We can clearly observe from image that again only those three features are linearly aligned with Happiness Score and all other three are just randomly scattered.

Thus, we will create two kinds of model. One, which has all features selected and another one with only relevant three features selected. We will compare the results on Linear Regression and Perceptron for both.

III. Implementation and Results

We trained using all three years of data together with all selected features. After doing the necessary preprocessing steps, the models are trained using the training data and is calculated for Root Mean Square Error with the test data. It is found that for the RMSE for both the models are comparable under the context of both the subset of features selected.

For perceptron, we took one hidden layer with 48 neurons. We tried different numbers and reached the conclusion based on RMSE results.

Both algorithms behave better if data is standardized (Specially Perceptron). Standardization involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one. For the purpose of standardization, we used sklearn's preprocessing library and standardscaler function on both the models.

To show that ignored feature are in fact impacting our analysis negatively, we trained our models on them as well.

Results found are :

TABLE II: Results based on RMSE comparison

Model	All Six	Most relevant	Less relevant
Linear Regression	0.58	0.65	1.02
Perceptron	0.55	0.67	0.98

Though these results are very close, we tried to implement only one feature at a time implementation to observe how closely linear regression and perceptron are related.

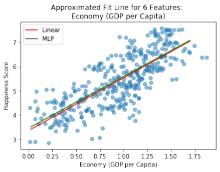

Fig. 3: Economy

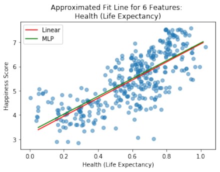

Fig. 4: Health

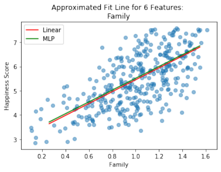

Fig. 5: Family

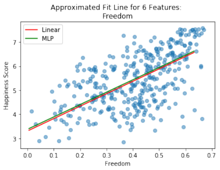

Fig. 6: Freedom

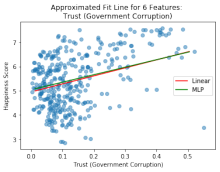

Fig. 7: Trust

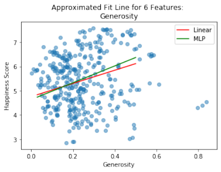

Fig. 8: Generosity

IV. Conclusions and Discussions

From the above experiment, we can conclude that three most relevant features are :

1) Economy
2) Health
3) Family

Both Linear Regression model and perceptron model provide almost same accuracy when most relevant features were selected. When working only on less relevant features, we observed a spike in RMS error, which was expected.

Interesting observation from this experiment was that even if we select all features, we are getting similar results as of most relevant features. This shows that even if we select features which are not relevant, models were not effected drastically. Linear regression and Perceptron both decreases the weights associated with less important features and thus not effecting the accuracy.

Although, as this is not much of concern to select only most relevant features in this small dataset, it may effect performance drastically if dataset is huge and features are in very large numbers.

References

[1] Sustainable Development Solutions Network ."World Happiness Report": https://www.kaggle.com/unsdsn/world-happiness/version/2?.

[2] Seber, G. A., Lee, A. J. (2012). Linear regression analysis (Vol. 329). John Wiley Sons.

[3] Montgomery, D. C., Peck, E. A., Vining, G. G. (2012). Introduction to linear regression analysis (Vol.821). John Wiley Sons.

[4] Ruck, D. W., Rogers, S. K., Kabrisky, M. (1990). Feature selection using a multilayer perceptron. Journal of Neural Network Computing, 2(2), 40-48.

[5] Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W. (2005). Applied linear statistical models (Vol. 5). Boston: McGraw-Hill Irwin.

[6] Belue, L. M., Bauer Jr, K. W. (1995). Determining input features for multilayer perceptrons. Neurocomputing, 7(2), 111-121.

[7] Myers, R. H., Myers, R. H. (1990). Classical and modern regression with applications (Vol. 2). Belmont, CA: Duxbury press.