Project Report

February 4, 2021

CONTENTS

[1]Word Count: 1258

1 ABSTRACT

*F*rom the beginning of time, governments have tried to elevate the quality of life of their citizen. Many surveys and systems have been developed to recognize what affects the happiness level. These results have been widely used across the globe to ensure better policy-making. In this project, we will use machine learning with the world happiness dataset taken from Kaggle to conclude what is the root of happiness.

2 INTRODUCTION

There is a consensus that productivity and general performence of a country has a direct relationship with the phychological status of its people. To keep prosperity level intact in a country, we must know what indicators affect the happiness level of the masses. Psychologically, every individual may have their triggers for happiness, but there has to be a bigger picture of it, therefore, we must find a way to detect that particular indicator that affects the prosperity level of a country or a region.
The main objective of this study is to develop a system that can clearly show us what indicators of a country's system increases the happiness level if maintained positively.
In this project, we will be using World Happiness Dataset from the public domain of Kaggle. The dataset has 156 data points. Each data point has seven features, and every feature represents a particular measure. Such as GDP, Social Support, Life Expectancy, Generosity, and many more. The data is in CSV format with a total size of 78.96 KB.

3 DESCRIPTION OF DATA STATS AND THEIR INTERPRETATION

We have analyzed the dataset and, we found out that the dataset contains 156 rows (data points) and seven features. We have extracted more useful information from each feature column, including Minimum, Median, Mean, standard deviation, and many more. The tabular form of statistics of each column is given below.
Column Names: Overall Rank, Happiness Score, GDP Per Capita, Social Support, Healthy Life Expectancy, Freedom to Make Life Choices, Generosity, and Perception of Corruption.

3.1 Data Stats

Column Name	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Rank	1	39.75	78.50	78.50	117.25	156.00
Score	2.853	4.545	5.380	5.407	6.184	7.769
GDP	0.0	0.6028	0.9600	0.9051	1.2325	1.6840
S.Support	0.0	1.056	1.272	1.209	1.452	1.624
Life Exp	0.0	0.5477	0.7890	0.7252	0.8818	1.1410
Freedom	0.0	0.3080	0.4170	0.3926	0.5072	0.6310
Generosity	0.0	0.1087	0.1775	0.1848	0.2482	0.5660
Corruption	0.0	0.0470	0.0855	0.1106	0.1412	0.4530

3.2 Data Interdependence

As we know that we are trying to find out which factor affects the happiness score the most. For this, We will be using a correlation technique used in data statistics to find out the positive or negative correlation (relationship) between entities. The tabular form of correlation between Happiness Score and all other features is given below:

S.No	Feature	Correlation With Happiness Score
1	GDP Per Capita	0.7938829
2	Social Support	0.770578
3	Healthy Life Expectancy	0.7798831
4	Generosity	0.07582369
5	Freedom to make choices	0.5667418
6	Perceptions of Corruption	0.3856131

The above table displays correlation values of happiness score with all other features. It can be concluded that the greater the GDP Per Capita the more happy the country will be. This can also be proved through the dataset as in 2015, Switzerland was the most prosperous country in the world with a GDP Per Capita of 1.39651 and a happiness score of 7.587, and now Finland with a happiness score of 7.769.

4 Machine Learning Model

Machine learning is a subset of artificial intelligence (AI) in which the system can automatically learn and improve from experience without being explicitly programmed. It has applications in a wide variety of domains where the system or model is provided with some data. The data has input features and an output label (result). The model then tries to find out the mathematical relation between the input and output, so when there is a new input,

the model uses its experience based on previously given data to predict the new output label.

Below are some widely used machine learning algorithms or models:
1) Linear Regression.
2)Logistic Regression.
3) Decision Tree.
4) Support Vector Machine.
5) Naive Bayes.
6) K-NN.
7) K-Means 8) Random Forest and many more.

In this project, we have used Linear Regression to train our model on World Happiness Dataset. The main reason for choosing linear regression is because it has always been used to identify relationships in big datasets. The dataset is split into train data as 70% and test data as 30%, where train data is used to train the model and test data is used to predict and test how accurate the model is.
As mentioned previously, we wanted to predict the happiness score of a country, but as we already know from the data analysis that happiness score is tightly correlated with GDP Per Capita. We used loops to train multiple linear regression models one by one with different features while having the happiness score column as a constant target. We created a separate model for GDP and happiness score and compared it with models of every other column attribute. This gave us a further clear picture of which feature understood patterns of happiness score more accurately.

5 Accuracy and Its Methods

We have used three different methods to gauge the accuracy score of our models. The names are given below:
1) Co-efficient of Determination or R-Squared
2) Mean Squared Error
3) Root Mean Squared Error

Note: The R-Squared represents the degree to which the feature has understood the variance of the target feature. The higher the R squared value, the better it is. On the other hand, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) represent errors with corresponding feature column.

5.1 Accuracy Table with Happiness Score

Methods	Rank	Score	GDP	Social Support
R-Squared	0.97	1	0.63	0.60
MSE	0.026	6.7	0.45	0.48
RMSE	0.163	8.1	0.67	0.69

Methods	Life Expectancy	Freedom of Choice	Generosity	Perception of Corruption
R-Squared	0.60	0.32	0.005	0.14
MSE	0.48	0.83	1.22	1.04
RMSE	0.69	0.91	1.10	1.02

6 Tests

We have used three different kinds of test to evaluate the trained machine learning models, description about tests is given below:
1) Ramsey RESET Test
2) Breusch–Pagan Test
3) J-Test

In machine learning, when using Linear Regression, one can specify Linear Regression Model that's linear on data or, one can specify a Linear Regression Model that is non-linear on data. We have observed GDP is leading the way from the correlation and accuracy tests. But we need some confirmation that our model accurately fits on the data, and for that, we have used multiple tests as mentioned above, and in case the true relationship between input feature as X and output label Y turns out to be non-linear then this model will not fit the data perfectly, and this is what these tests are designed for.
In statistics, the Ramsey Regression Equation Specification Error Test (RESET) test is a test for the linear regression model. It tests whether non-linear combinations of the fitted values help explain the response variable. p value if greater then 5% means this model is linear. The Breusch–Pagan test is a chi-squared test, the test statistic is distributed n2 with k degrees of freedom. If the test statistic has a p-value below an appropriate threshold (e.g. p is less than 0.05) then the null hypothesis of homoskedasticity is rejected which means the variable which is predicting is not giving perfect insight.
The idea of the J test is the following:
if the first model contains the correct set of regressors, then including the fitted values of the second model into the set of regressors should provide no significant improvement. But if it does, it can be concluded that model 1 does not contain the correct set of regressors.

6.1 Test Table Between Score and Rank

Test	P-Value	Estimated Dev.	Error
Ramsey Reset=202.05	p-value < 2.2e-16	Nill	Nill
Breusch-Pagan=0.7382	p-value = 0.3902	Nill	Nill
J-Test: M1 fitted on M2	Nill	0.00240	0.025104
J-Test: M2 fitted on M1	Nill	0.99845	0.020149

6.2 Test Table Between Score and Score

Test	P-Value	Estimated Dev.	Error
Ramsey Reset=21.065	p-value = 8.418e-09	Nill	Nill
Breusch-Pagan=4.0927	p-value = 0.04307	Nill	Nill
J-Test: M1 fitted on M2	Nill	0	1.2204e-16
J-Test: M2 fitted on M1	Nill	1	9.4399e-17

6.3 Test Table Between Score and GDP

Test	P-Value	Estimated Dev.	Error
Ramsey Reset=2.8701	p-value = 0.05977	Nill	Nill
Breusch-Pagan=0.0549	p-value = 0.8147	Nill	Nill
J-Test: M1 fitted on M2	Nill	Nill	Nill
J-Test: M2 fitted on M1	Nill	1	Nill

Here the Ramsey Reset Test and Breusch-Pagan Test returns p-value greater than 0.05 which is 5% hence we can conclude that model works perfectly fine and show linearity on feature column GDP Per Capita.
The J Test here shows no significant improvement in accuracy, therefore the model is perfectly in harmony with the linearity.

6.4 Test Table Between Score and Life Expectancy

Test	P-Value	Estimated Dev.	Error
Ramsey Reset=9.8625	p-value = 9.392e-05	Nill	Nill
Breusch-Pagan=0.4188	p-value = 9.392e-05	Nill	Nill
J-Test: M1 fitted on M2	Nill	0.59361	0.10559
J-Test: M2 fitted on M1	Nill	0.49516	0.10749

6.5 Test Table Between Score and Freedom

Test	P-Value	Estimated Dev.	Error
Ramsey Reset=3.4478	p-value = 0.03433	Nill	Nill
Breusch-Pagan=3.9933	p-value = 0.04568	Nill	Nill
J-Test: M1 fitted on M2	Nill	0.85178	0.058981
J-Test: M2 fitted on M1	Nill	0.54770	0.082620

6.6 Test Table Between Score and Generosity

Test	P-Value	Estimated Dev.	Error
Ramsey Reset=1.6884	p-value = 0.1883	Nill	Nill
Breusch-Pagan=14.303	p-value = 0.0001556	Nill	Nill
J-Test: M1 fitted on M2	Nill	1.0140	0.06046
J-Test: M2 fitted on M1	Nill	1.8458	0.63306

6.7 Test Table Between Score and Perception of Corruption

Test	P-Value	Estimated Dev.	Error
Ramsey Reset=8.3387	p-value = 0.0003663	Nill	Nill
Breusch-Pagan=5.7408	p-value = 0.01658	Nill	Nill
J-Test: M1 fitted on M2	Nill	0.062735	0.06046
J-Test: M2 fitted on M1	Nill	0.42233	0.129155

The test satisfies and fails to reject null hypothesis when tested on featuer column GDP.

7 Conclusion

We used a linear regression model to test and figure out which factors of a country play a vital role in its prosperity. We used World Happiness Dataset from the public domain of Kaggle and divided it into train and test accordingly. After training and testing on an adequate number of data points, we found out the happiness score of a country has a strong relationship with the GDP Per Capita. Further, to confirm linearity between happiness score and GDP Per Capita, we used multiple tests on each feature column.
From the above results of accuracy score and test results, we can conclude that countries with higher GDP per capita have happier population.

8 REFERENCES

J.B. Ramsey (1969), Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis. Journal of the Royal Statistical Society, Series B 31, 350–371

W. Kr<e4>mer H. Sonnberger (1986), The Linear Regression Model Under Test. Heidelberg: Physica

T.S. Breusch A.R. Pagan (1979), A Simple Test for Heteroscedasticity and Random Coefficient Variation. Econometrica 47, 1287–1294

R. Koenker (1981), A Note on Studentizing a Test for Heteroscedasticity. Journal of Econometrics 17, 107–112.

W. Kr<e4>mer H. Sonnberger (1986), The Linear Regression Model under Test. Heidelberg: Physica

R. Davidson J. MacKinnon (1981). Several Tests for Model Specification in the Presence of Alternative Hypotheses. Econometrica, 49, 781-793.

W. H. Greene (1993), Econometric Analysis, 2nd ed. Macmillan Publishing Company, New York.|

W. H. Greene (2003). Econometric Analysis, 5th ed. New Jersey, Prentice-Hall.

9 APPENDIX

```
data <- read.csv("2019.csv") #import dataset
library(dplyr) #import all libraries
library(pROC)
library(dplyr)
library(caTools)
library(Metrics)

library(lmtest) # all ml model tests are in this library
options(warn=-1)

#data summary task
nrow(data)
summary(data)
data <- data[,-2] #dropping names of countries becuase they're insignificant
columns<-colnames(data, do.NULL = TRUE, prefix = "col")
columns

#correlation between happiness score and everyother attribute
cor(data$Score, data$GDP.per.capita)
cor(data$Score, data$Social.support)
cor(data$Score, data$Healthy.life.expectancy)
cor(data$Score, data$Generosity)
cor(data$Score, data$Freedom.to.make.life.choices)
cor(data$Score, data$Perceptions.of.corruption)



#Train test split
set.seed(2)
split<-sample.split(data, SplitRatio=0.7)
split
train<-subset(data, split="TRUE")
test<-subset(data, split="FALSE")


Model2 <-lm(formula = data$Score ~data$GDP.per.capita,data=data)

#a loop for training model of score attribute with everyother attribute one by one
counter=1
for (val in columns) {
  #filterdData=data[val]
```

```
  print(counter)
  print(val)

  #Linear regression model is being trained with score and  each attribute from loc
  Model <-lm(formula = train$Score ~train[,counter],data=train)
  #print(summary(Model))

  pred<-predict(Model,test) #trained model is being predicted
  #pred

  #accuracy 3 techniques
  #coefficent of determination or r squared
  Y_test<- test$Score
  error <- Y_test - pred
  R2=1-sum(error^2)/sum((Y_test- mean(Y_test))^2)
  print(paste("R2",R2))

  #mean squared error
  mean_squared_error <- mse(test$Score,pred)
  print(paste("mean_squared_error",mean_squared_error))

  #root mean squared error
  root_mean_squared_error <- rmse(test$Score,pred)
  print(paste("root␣mean_squared_error",root_mean_squared_error))

  # linear regression validation/testing
  RamseyRESET<-resettest(Model,type="regressor",data=data)
  print(RamseyRESET)

  BPTEST<-bptest(Model, varformula = NULL, studentize = TRUE, data = data)
  print(BPTEST)
  #This is the basis of the BreuschPagan test. It is a chi-squared test: the tes

  #coxtest(Model, Model2)
  JTEST<-jtest(Model, Model2)

  #encomptest(Model, Model2)
  print(JTEST) #compares two fitted models GDP with score and score with some other

  #rmse
  counter<-counter+1
}
```