DATA MINING

REPORT

KIRAN T S-B160244CS

CHRISTY SABU-B160143CS

C P ROSHAN-B160258CS

HASSAN JAMIL-B160510CS

DATA PREPROCESSING

The world happiness dataset downloaded from kaggle.com contained separate datasets for 3 years(2015,2016,2017).the datasets were merged together using open refine and the unwanted attributes were removed using remove columns option in open refine.finally the happiness dataset was brought into the format (Country Name, Year, Happiness Score).the world development indicator dataset had the 35 indicators as tuples which had to be transposed in open refine to change it into attributes, then the previous world happiness dataset was merged with it on country name and year.

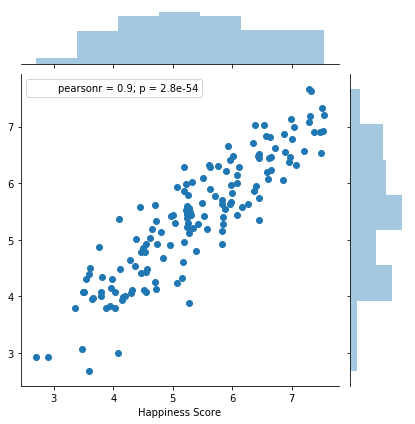
The final dataset contained several abnormalities as many of the (about 98%)tuples for certain attributes(indicators) were found to be NULL .we used open refine to filter out the unwanted attributes using the above mentioned procedure.About 5 of the attributes were dropped for the same.the tuples relating to regions rather than a country was removed using the text facet option available in open refine,the same method was used to calculate the no of null values for a particular attribute.

The next hurdle in data preprocessing was NULL values in happiness score for some countries for certain years.for this a seperate attribute (region) was introduced in the dataset which represented the region to which each country belonged. At first the columns were grouped based on the “country name” and the mean of available values were used to fill the NULL values.further the indicators were filled using the value from the same regions ,assuming the values doesn’t show much variance from the indivitual countries indicator values.the categorical values(Country Name,Region,Year) were dropped to apply linear regression to create a new model to predict the happiness score.

MODEL CREATION

The linear regression model was used to predict the happiness score . we used datasets of 2015 and 2016 as training data and 2017 dataset as test data as instructed in the questions.the train data was used succesfully to fit the regression model with happiness score along the y axis and rest of the indicators along x axis ,excluding country name and year (categorical values).The test data was loaded and the predictions for the test data were made and compared with the original Happiness Score.

The jointplot for the y\_test and predictions were found



The happiness score can be predicted with the model with errors.

Mean Absolute Error(MAE): 0.41186492617895337

Mean Squared Error(MSE): 0.25496690392933397

Square root of MSE (RMSE): 0.5049424758616906

BEST FEATURE

Recursive feature elimination (RFE) was used to find the best feature to predict the happiness score.RFE is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached

***model = LinearRegression()***

***rfe = RFE(model, 1)***

***rfe = rfe.fit(x\_train,y\_train)***

***print(rfe.ranking\_)***

[14 22 2 3 1 19 7 4 5 15 17 18 23 28 12 21 16 20 8 9 27 11 10 24

26 13 25 6]

Which showed the 5th indicator was the best feature to predict the happiness score

***print(x\_train.columns[4])***

NY.ADJ.AEDU.GN.ZS

Which was found out to be Adjusted savings: education expenditure (% of GNI) from the dataset