AUGUST 8, 2019

FAULTY STEEL PLATES

EXECUTIVE SUMMARY

MOHAN RAJENDRAN

Purpose

The aim of this analysis was to correctly classify the type of surface defects in stainless steel plates, with six types of possible defects (plus "other"). Multi Class Classification will be used to predict the fault type based on image parameters.

Faulty Steel Plates Dataset

Link: https://www.kaggle.com/uciml/faulty-steel-plates

This dataset contains 1941 observations of 34 variables that describe the picture of the fault in the steel plates. The first 27 fields describe the kind of steel plate faults seen in images and serve as predictor variables. There appears to be a lot of correlation and collinearity between the variables. Though we can make some legitimate guess around how these variables are corelated, the absence of variable description in the original dataset prevents us from removing these variables.

Response Variable

The last 7 variables are the response variables each denoting one of the fault class. These response variables can be combined into one with each fault as the observation value.

Audience

Direct audience for this analysis will be any industry which employs steel in its machinery like conveyor belts, food processing industry using heavy machinery and basically most of the automated production industry which uses some form of conveyer belts in its production unit. They can use this model to predict the fault types based on the image parameters.

Methodology

The columns in the dataset define various parameters of the faulty steel image. Hence high correlation between the variables is expected. Also since the descriptions of the columns are not given it is decided to go ahead without removing any of the columns based on colleniarity.

Artificial Neural Networks

Almost all the variables are numeric and can be nonlinear considering that they are variables that explain the image of the faulty steel. Most of them are x, y coordinates. Hence Neural Network approach is best which has no data assumptions. We will perform 10-fold cross validation to come up with the best 'size' and 'decay rate' parameters for the ANN model. Upon completion of the 10-fold cross validation, the best model based on accuracy has a size = 6 and Decay rate = 0.4

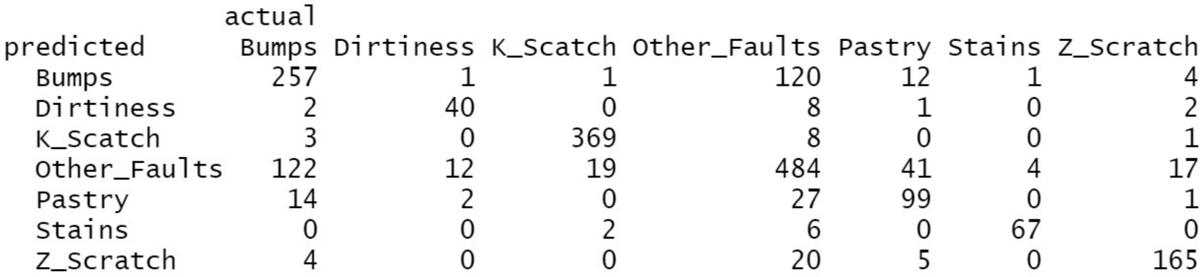
Decision Trees

Another approach that can be very useful in scenarios like image data processing is Decision Trees. Even decision trees have no assumptions and hence useful in data like this. We performed RandomForest and used 10-fold cross validation to arrive at the best value for 'mtry' argument. Upon completion we arrived at a mtry parameter value of 24.

Double Cross Validation
Finally, a double cross validation is performed between the selected two models. Artificial Neural Network model performed well between these two models.

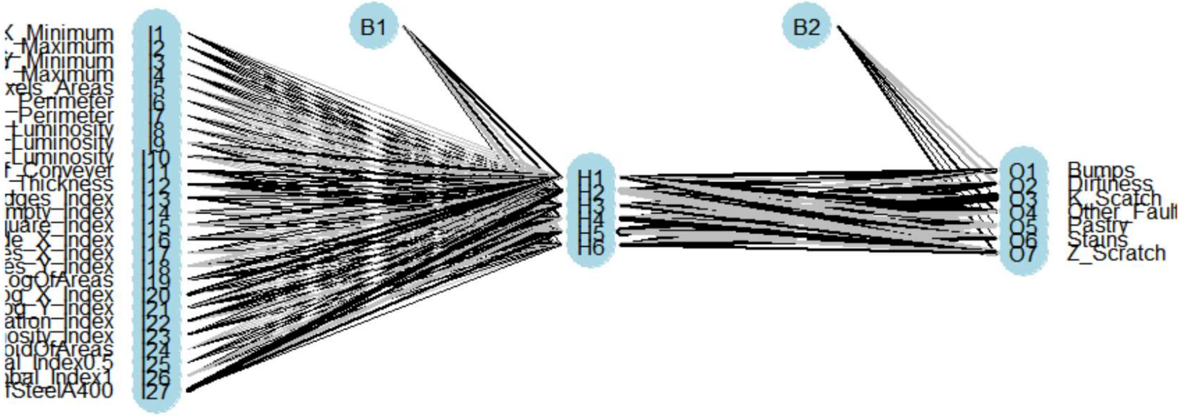
Final Model

ANN model with size = 6 and decay rate = 0.5 performs better overall and the same is fitted against the entire dataset. The accuracy rate of the model is 76.30% with the following confusion matrix.



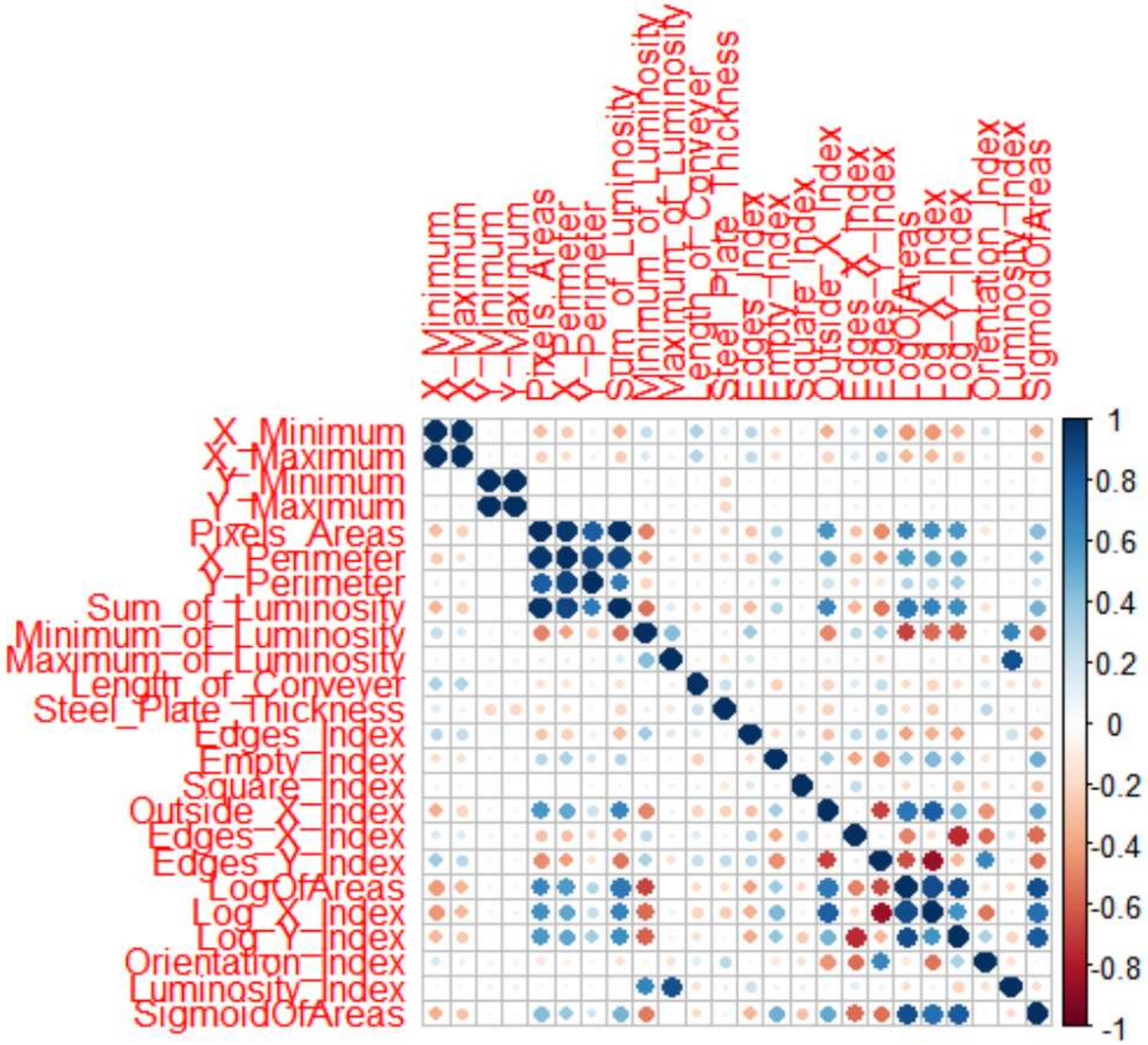
```
              actual
predicted      Bumps Dirtiness K_Scatch Other_Faults Pastry Stains Z_Scratch
  Bumps          257         1        1          120     12      1         4
  Dirtiness        2        40        0            8      1      0         2
  K_Scatch         3         0      369            8      0      0         1
  Other_Faults   122        12       19          484     41      4        17
  Pastry          14         2        0           27     99      0         1
  Stains           0         0        2            6      0     67         0
  Z_Scratch        4         0        0           20      5      0       165
```

The Neural Network is plotted as follows



Limitations

Since the description for the variables are not available, we are not able to reduce the number of variables though we identified collinearity plot given below.



Next Steps

AS next steps, we can work around these limitations by doing variable selection based on principal component analysis. From the PCA analysis, 10 components explain more than 90% of the variability in the data. So, its fair to use the top 10 contributors. We can repeat the entire analysis based on the top 10 contributors to see if the accuracy improves.