Bank Notes Validation

Shreyank Buddhadev

October 2021

1 Problem

Given a dataset of extracted features of bank notes images task is to classify whether a bank note is authentic or not.

1.1 Dataset Overview

Feature	Type	Summary
variance	numerical	variance of Wavelet Transformed image
skewness	numerical	skewness of Wavelet Transformed image
curtosis	numerical	curtosis of Wavelet Transformed image
entropy	numerical	entropy of image
target	categorical	**1** is for authentic note and **0** for not.

1.2 Data Visualization

- Imbalanced dataset pose a challenge of poor modeling. To better visualize imbalance in dataset fig 1 depicts count plot for feature *target*.

  As depicted in figure 1 data has slight imbalance.

- Figure 2a and 2b depicts boxplot and distribution plot for feature *variance*

- Figure 3a and 3b depicts boxplot and distribution plot for feature *skewness*

- Figure 4a and 4b depicts boxplot and distribution plot for feature *curtosis*

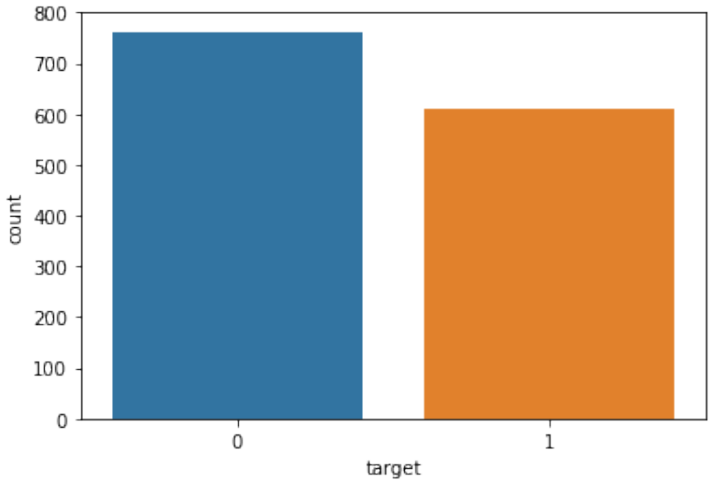
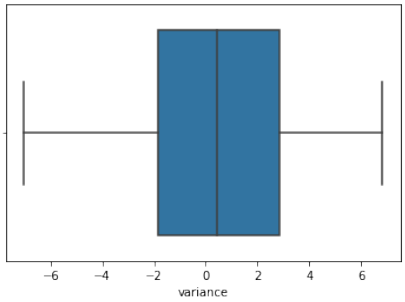
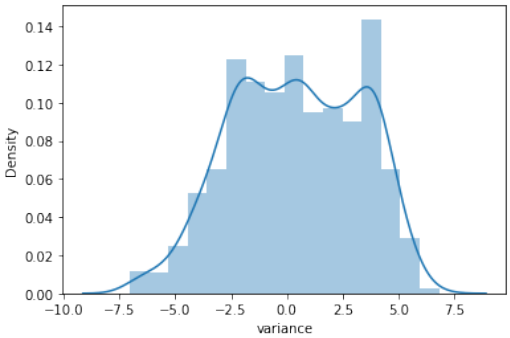


Figure 1: count plot for target features



(a) box plot for variance

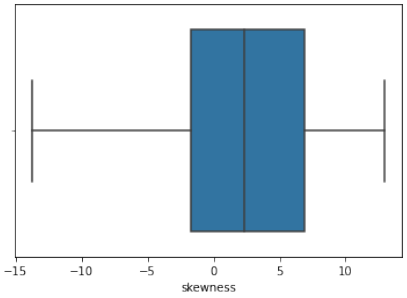


(b) dist plot for variance

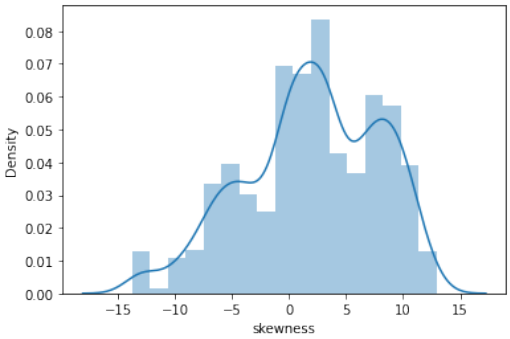
Feature *curtosis* has right skewed data. Also, from the plots we can infer that there is possibility of outlier in the feature *curtosis*.

- Figure 5a and 5b depicts boxplot and distribution plot for feature *entropy*

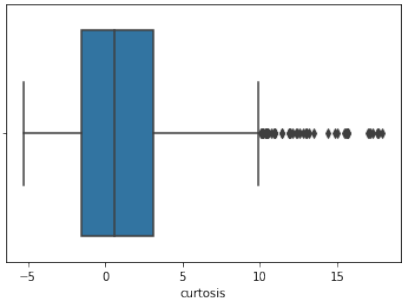
  Feature *entropy* has left skewed data. Also, from the plots we can infer that there is possibility of outlier in the feature *entropy*.



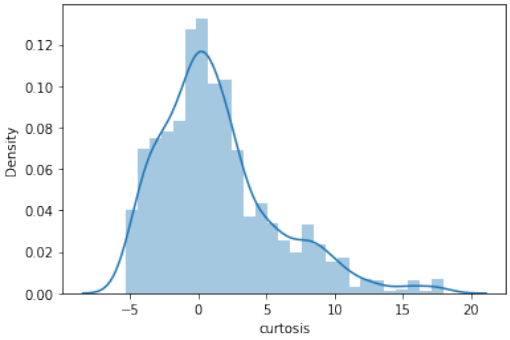
(a) box plot for skewness



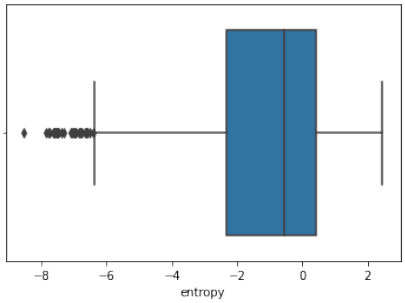
(b) dist plot for skewness



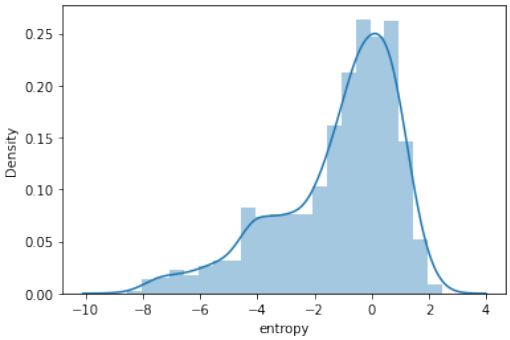
(a) box plot for curtosis



(b) dist plot for curtosis



(a) box plot for entropy



(b) dist plot for entropy

1.3 Data Pre processing

Provided data had several inconsistencies to overcome those below mentioned steps are applied on data set. To handle inconsistencies there were several possible alternatives, this section contains approach which were best suited to achieve high accuracy.

1.3.1 Filling NaN values

Features *curtosis* and *entropy* had some outliers which contributed to higher standard deviation. Hence, to replace data with central values NaN values for all features were replaced by corresponding feature's median value.

1.3.2 Outlier removal

Feature *curtosis* and *entropy* have few data points as outliers. But, since logistic regression does sigmoid squshing it doesn't had much impact. Analysis for effect of outliers on accuracy is added in analysis section.

1.3.3 Normalizing or Standardizing data

The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. Similarly, Standardizing the features around the center mean 0 and a standard deviation of 1 is important when we compare measurements that have different units.

- Normalization process $x_i = (x_i - min)/(max - min)$
- Standardization process $x_i = (x_i - \mu)/\sigma$

  here x_i = data point, μ = mean, σ = standard deviation

Here, feature data is around similar scale. Yet, to be computationally efficient normalization was applied.

1.3.4 Linearity and Multi collinearity

Below heatmap shows linear independence between columns. Here, skewness and curtosis data is highly corelated.

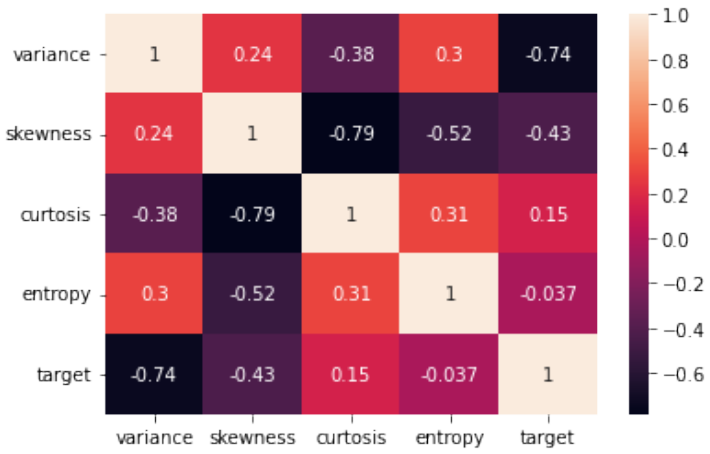


Figure 6: heat map for features

1.4 Multivariate Gaussian Naive Bayes

From fig 6 we can infer that since curtosis and skewness column has high corelation we can choose *variance, skewness* and *target* column for multivariate gaussian.

1.5 Logistic Regression loss function plots

- Loss function graph for logistic regression using gradient descent

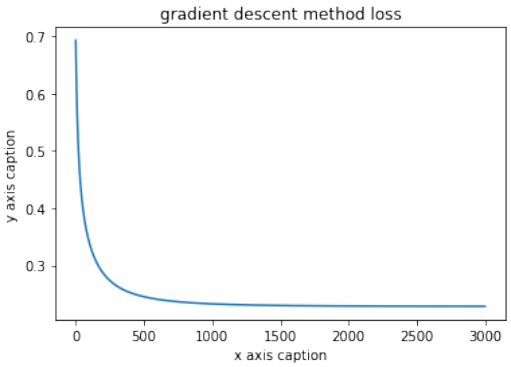


Figure 7: loss graph per iteration for gradient descent

- Loss function graph for logistic regression using newton's method

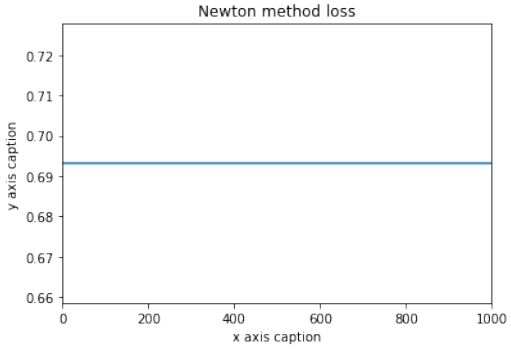


Figure 8: loss graph per iteration for newton's method

1.6 Accuracy Analysis

For logistic regression

type	accuracy
newton's method - training	89.79033728350045
newton's method - testing	90.54545454545455
gradient descent - training	89.88149498632635
gradient descent - testing	89.45454545454545

For gaussian

features taken	accuracy
all-training	81.03919781221514
all - testing	77.818181818181815
variance - training	81.13035551504102
variance - testing	79.63636363636364
skewness - training	61.16681859617138
skewness - testing	62.909090909090914
curtosis - training	49.68094804010939
curtosis - testing	56.36363636363636
entropy - training	55.51504102096627
entropy - testing	47.63636363636364