BANKNOTE AUTENTICATION USING MULTIPLE LINEAR REGRESSIONS AND SIMPLE LINEAR REGRESSION

[1]Sofia Firdaus, [2]M.AfsharAlam,
[1](Department of Computer Science & Engineering, Jamia Hamdard, New Delhi, India.
Email: sofiafirdaus786@gmail.com), Corresponding author
[2](Professor, Department of Computer Science & Engineering, Jamia Hamdard, New Delhi, India)

ABSTRACT
Bank note forgery has become important economic issue in India. People from urban to rural civilization are depending on bank note for their livelihood. These bank notes are used to buy things, which are important in day to day life or invest in other sectors or to live luxurious life. But sometime these notes are forged by some unauthenticated entity for their own gain. This act can disrupt the economy of the country and disrupt life of innocent people. It can result in a huge crisis. To avoid this some techniques are used to determine the authentication of the bank. These techniques will ensure the integrity of the bank note. In this paper we are discussing linear regression technique. This technique includes multiple linear regression and simple linear regression, which comes under Classification, using data mining. The outcomes of these techniques will be then analyzed and result will be derived from it.

Keywords: *Linear regression, Simple linear regression, Data mining, Authentication, Classification*

1. INTRODUCTION
Bank note fraud is described as the use of illegal means by which original bank notes are forged or duplicated and then they are supplied in the markets, so that citizens of that state or country can be fooled and exploited in a terrible way. It is a criminal offence. In bank note forgery, unauthenticated organization or people will make a bank note which is look alike of the original bank note. Then they will release these fake notes in market so that these fake notes can be easily used as original notes. This will cause a huge loss to the economy of the state or country. It will affect everyone including businessmen, paid workers, self-employed, government, etc.

It is roughly calculated that two hundred and fifty out of ten lakhs are always fake since, it is also said that forged bank notes with the face value of seventy rupees enter the economy every year but only a third is intercepted and it was also said by Prime Minister that terrorism was funded by fake Indian currency so this was an important reason for removing the notes of five hundred and one thousand, that made up of major portion of market [1]. Government and authentic organization like Reserve Bank of India takes drastic measures to ensure that such situation does not occur frequently.Many techniques are used to determine the authentication of the bank note. Here data mining techniques will be used. Data mining refers to the process of extraction and processing useful data. Now, these data can range from text data and multimedia data to xml data. Useful patters are retrieved after processing these data. Data mining has key five steps. First it has to identify the source from where information or data will be collected for further use. Then next step is to select the data that is going to be analyzed. After that relevant and useful information is extracted from the analyzed data. Next step will be to identify the useful pattern from the extracted dataset. Last step will be to interpret the outcome and draw some results then report them to party. Data mining techniques like Linear Regression will be used here to determine the authentication of bank note in real world scenario by processing the attributes which are extracted from the images. Images of the real and fake banknotes are taken and then attributes are extracted from them. These attributes are: Variance, Skewness, Kurtosis, and Entropy.

Variance is defined as how each pixel, in the image, varies from its neighboring pixel. Normally, variance attribute gives us an idea that how those pixels are spread in the image. Skewness is defined as a measure of lack of symmetry in an image. A dataset is said to be symmetric if its left-hand side is similar to the right-hand side. Kurtosis is describing the shape of probability. High kurtosis datasets will have outliers or heavy tails and light kurtosis will have no outliers or light tails. Entropy is defined as measure that classifies the texture of the image.

2. LITERATURE REVIEW

Use of Data Mining in Banking - Kazi Imran Moin, Dr. Qazi Baseer Ahmed[2]
Availability of large amounts of data and need for transforming such data into knowledge can encourage the Information Technology industry to use the phenomena of data mining. The banking organization in today's world has come under a big change in the way business operations is conducted. The banking organization has started knowing the need of the processes like data mining phenomena. Author discussed about techniques of this phenomena and its uses in banking organization.
Study of Data Mining on Banking Database In Fraud Detection Techniques- Sayali Kishor Rodge[3]
Bank organization have a very big database to handle for different purposes. It is also used to secure data elements and stop the frauds data Mining processes is used. Also includes the whole Data Mining technique. It is to avoid the conflicts of bank database processes.
Principles of Data Mining- David J. Hand[4]
Discusses that how due to data mining some interesting features have been found in huge datasets. Paper gives a general view about data mining and its relationship with statistic field by using techniques like regression. This paper also discusses about the detection of unfavorable drugs using the techniques of data mining.
Realtime Fraud Detection In The Banking Sector Using Data Mining Techniques/Algorithm - S.N. John, C.Anele, Okokpujie Kennedy O., F.Olajide [5]
Banking organization has become very important; therefore, fraud has also become common. To detect it they are using techniques of data mining which also include linear regression. Customer data is analyzed using these techniques to detect fraud by adding some advance level of security.

3. MATERIALS AND METHODS

Linear Regression analysis is a type of predictive analysis. It is a technique of data mining which comes under Classification. It shows the relationship between the dependent variable and independent variable. It is done by using linear function. Linear Regression shows that how independent variables can be used to predict the outcome of dependent variable. The equation of regression is normally written as, y=mx+c. Here, 'c' is constant, 'm' is coefficient of regression and 'x' is value of independent variable. [6]. Now, linear regression can be used in applications in huge scale as it is not difficult to accommodate the model which is dependent on their unspecified parameters linearly then the models who are dependent on their unspecified parameter non-linearly [7]. If there is one independent or explanatory variable, then this approach is called Simple Linear regression. Here, independent variable can be denoted by 'x' and dependent variable can be denoted by 'y'. If there are multiple independent values then it is called as Multiple Linear Regression.
To authenticate the integrity of bank note, Weka will be used. Weka (Waikato Environment for Knowledge Analysis) is an open source software which is used to apply the data mining techniques on dataset. For that dataset is used which is downloaded from UCI website [8]. This dataset contains data which were extracted from images of the banknote. These images are taken from real and fake banknotes. For this purpose, an industrial camera was used and the resultant image have 400x400 pixels. Wavelet transform tool was the name of the tool which was used to extract features from the image. There are 1378 entries in dataset. Variance, Skewness, Kurtosis, Entropy and Class are the attributes who are extracted from the image. Then import these attributes in Weka and save it as '*.arff' extension. Attributes must be in the from of numeric or binary. Dataset is divided into total two parts. Here, sixty five per cent of them is used as training data and rest of the thirty five of them is used as training dataset. Since in regression, a line is formed which is called regression line. Here, two classes are taken which we call 0 and 1 and then they will be used as numeric values. So, for regression line in case of 0 instances the values are pretty low and for 1 instances, the value is high. Then a threshold is decided according to which it is decided that whether value comes under class 0 or class 1. In figure-*fig.1,* the data which was imported in Weka is shown. Here, four attributes are present which is discussed above. If these attributes are in the form of nominal then they are changed to numeric
.

No.	1: VarianceT	2: SkewnessT	3: Kurtosis	4: Entropy
	Numeric	Numeric	Numeric	Numeric
1	3.6216	8.6661	-2.8073	-0.44699
2	4.5459	8.1674	-2.4586	-1.4621
3	3.866	-2.6383	1.9242	0.10645
4	3.4566	9.5228	-4.0112	-3.5944
5	0.32924	-4.4552	4.5718	-0.9888
6	4.3684	9.6718	-3.9606	-3.1625
7	3.5912	3.0129	0.72888	0.56421
8	2.0922	-6.81	8.4636	-0.60216
9	3.2032	5.7588	-0.75345	-0.61251
10	1.5356	9.1772	-2.2718	-0.73535
11	1.2247	8.7779	-2.2135	-0.80647
12	3.9899	-2.7066	2.3946	0.86291
13	1.8993	7.6625	0.15394	-3.1108
14	-1.5768	10.843	2.5462	-2.9362
15	3.404	8.7261	-2.9915	-0.57242
16	4.6765	-3.3895	3.4896	1.4771
17	2.6719	3.0646	0.37158	0.58619
18	0.80355	2.8473	4.3439	0.6017
19	1.4479	-4.8794	8.3428	-2.1086
20	5.2423	11.0272	-4.353	-4.1013

Fig.1 Representation of dataset

3.1. Multiple Linear Regression

After that apply the filter called Linear Regression which is present in Classification. Now, Classification is the function, or it can be also called technique, of data mining where it will assign data elements to their intended categories or classes. During the application of linear regression Variance, Skewness and Kurtosis are treated as independent attributes on which Class is dependent. Using linear function, dependent and independent attributes will be computed.

```
Linear Regression Model

Class =

     -0.1428 * VarianceT +
     -0.0781 * SkewnessT +
     -0.1016 * Kurtosis +
      0.7987

Time taken to build model: 0.01 seconds

=== Predictions on test split ===

    inst#     actual  predicted      error
        1       0          -0.036    -0.036
        2       0           0.168     0.168
        3       1           0.906    -0.094
        4       0           0.027     0.027
        5       0          -0.101    -0.101
        6       0          -0.028    -0.028
        7       1           0.92     -0.08
        8       1           0.864    -0.136
        9       1           1.171     0.171
       10       0           0.083     0.083
```

Fig.2 Multiple linear regression execution

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.11 seconds

=== Summary ===

Correlation coefficient                  0.9315
Mean absolute error                      0.1322
Root mean squared error                  0.1807
Relative absolute error                 26.7951 %
Root relative squared error             36.432  %
Total Number of Instances              480
```

Fig.3 Multiple linear regression result

In Multiple Linear Regression, there are multiple independent variables. Here, in figure- *fig.2* it is shown that model body is built in 0.01 seconds. It is also showing that class of instance has different values in actual and predicted column and in figure- *fig.3* it is shown that time taken to test model on test data is 0.11 second.

3.2. Simple Linear Regression

After the result is produced then apply the filter called 'Simple Linear Regression' which comes under Classification.

```
Linear regression on VarianceT

-0.13 * VarianceT + 0.5

Predicting 0 if attribute value is missing.


Time taken to build model: 0 seconds

=== Predictions on test split ===

    inst#     actual  predicted      error
        1       0          0.346     0.346
        2       0          0.847     0.847
        3       1          1.042     0.042
        4       0          0.106     0.106
        5       0         -0.187    -0.187
        6       0          0.05      0.05
        7       1          0.794    -0.206
        8       1          0.885    -0.115
        9       1          0.661    -0.339
       10       0          0.28      0.28
```

Fig.4 Simple linear regression execution

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.1 seconds

=== Summary ===

Correlation coefficient                  0.7877
Mean absolute error                      3.0171
Root mean squared error                  3.6498
Relative absolute error                 63.0215 %
Root relative squared error             61.5794 %
Total Number of Instances              480
```

Fig.5 Simple linear regression result

In Simple Linear Regression, there is single independent variable. Here, in figure- *fig.4* it is shown that model body is built in 0 second. It is also showing that class of instance has different values in actual and predicted column and in figure- *fig.5* andit is shown that time taken to test model on test data is 0.1 second.

4. RESULT AND DISCUSSION

Total number of instances which are processed here are 480. There is a difference in time taken by multiple and simple linear regression to build their model bodies. There is also a difference between original and predicted values of an instance. Therefore, there will be some error due to difference between original and predicted value.

Table 1. Comparison between Multiple linear regression and Simple linear regression

Technique	Mean Absolute Error	Root Mean Square Error
Multiple Linear Regression	0.1322	0.1807
Simple Linear Regression	3.0171	3.6498

5. CONCLUSION

Authentication of bank note is done by using Multiple Linear regression and Simple Linear Regression where Linear regression took 0.01 second to build the model and Simple Linear regression took 0 second to build the model. Time taken to test on test model is 0.11 seconds in Multiple linear regression and in Simple linear regression is 0.1 second. Error in between original and predicted value is high in Simple linear regression in comparison to Multiple Linear regression. Multiple Linear Regression has less mean absolute error (0.1322) and root mean square error (0.1807) in comparison to Simple Linear regression's mean absolute error (3.0171) and root mean square error (3.6498). Therefore, multiple linear regressions are much better than simple linear regression.

REFERENCES

[1] Fake notes or counterfeit notes or forged notes in India. (April 16, 2017). Retrieved from https://www.bemoneyaware.com/blog/fake-notes-counterfeit-forged-india/

[2] Kazi Imran Moin, Dr. Qazi Baseer Ahmed, "Use of Data Mining in Banking" International Journal of Engineering Research and Applications Vol. 2, Issue 2, Mar-Apr 2012, pp.738-742

[3]Sayali Kishor Rodge, "Study Of Data Mining On Banking Database In Fraud Detection Techniques", YMT College of Management, Maharashtra, India, International Research Journal of Engineering and Technology (IRJET), May (2016).

[4] David J. Hand, "Principles of Data Mining", Department of Mathematics, South Kensington Campus, Imperial College London, London, UK, Drug Safety, July (2017)

[5] S.N. John, C.Anele, Okokpujie Kennedy O., F.Olajide, "Realtime Fraud Detection In The Banking Sector Using Data Mining Techniques/Algorithm", International Conference on Computational Science and Computational Intelligence, 2016.

[6] Hilary L. Seal (1967). "The historical development of the Gauss linear model". Biometrika. 54 (1/2): 1–24. doi:10.1093/biomet/54.1-2.1. JSTOR 2333849

[7] Linear regression-Wikipedia. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Linear_regression

[8] Dua, D. and KarraTaniskidou, E., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2017.