Report: Bank Note Forgery Clustering Analysis: Clustering Prediction Model
To: Chief Data Officer
Date: August 23rd 2022

Main objective of the analysis

The main objective of this analysis is to see if a clustering model can automatically differentiate between forged and genuine banknotes. Several clustering models were developed, and dimension reduction transformation is discussed as a recommendation for further work:

- K-means clustering algorithm
- Hierarchical agglomerate clustering algorithm (Best performing algorithm)
- Mean shift algorithm
- DBSCAN algorithm
- Principal Component Analysis (PCA): discussed in further work and next steps section

These models were developed and evaluated and the most appropriate one chosen in terms of interpretability and accuracy. The best model is then recommended, based on its predictive performance when presented with unseen banknotes; i.e. are the unseen notes genuine or a forgery. This model could then be put into production. By automating this model large numbers of banknotes can be quickly scanned and examined, and forged banknotes detected.

Section describing data

The original dataset used is 1,372 instances of banknote images comprising of;

- 762 genuine banknotes
- and 610 forged banknotes

Although the dataset has a column indicating whether a banknote is genuine or a forgery, this data was only used to test the accuracy of the cluster models. Cluster models were developed to determine if bank notes could be clustered into forgery and genuine banknote clusters.

The following features were extracted from the 1,372 bank note images:

- V1 (variance of wavelet)
- V2 (skewness of wavelet)
- V3 (kurtosis of wavelet)
- V4 (image entropy)

There were no missing values.

During the exploratory data analysis phase of the project, it was found that the data set had outliers; data points that differ significantly from other observations. Thirty-four outliers were identified and removed from the dataset, representing less than 2.5% of the raw dataset. For the cluster models developed 2 features were used, V1 and V2.

As the data ranges for V1 and V2 were different, the data for these features was normalized (range 0 to 1), prior to developing a K-Means model.

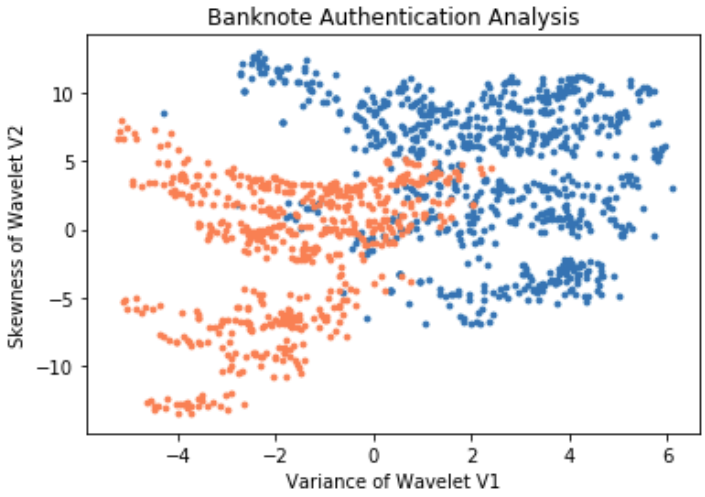
Exploratory Data Analysis (EDA)

The exploratory data analysis phase of the project included generating summary statistics for the two features, V1 (variance of wavelet) and V2 (skewness of wavelet); counts, mean, standard deviation, max, and min.

Out[11]:

	V1	V2
count	1372.000000	1372.000000
mean	0.433735	1.922353
std	2.842763	5.869047
min	-7.042100	-13.773100
25%	-1.773000	-1.708200
50%	0.496180	2.319650
75%	2.821475	6.814625
max	6.824800	12.951600

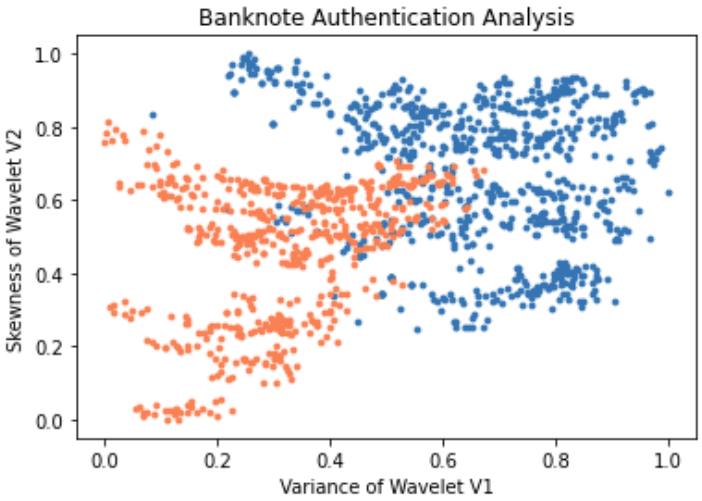
The original data was visualized by graphing V1 and V2 of the raw dataset for the 1,372 banknote instances on a graph with the x axis labelled V1, and the y axis labelled V2.



The coral-coloured dots indicate forged banknote data instances, and the blue dots indicate genuine banknote instances. Although there are two distinct clusters in the above graph, there is some overlapping in the centre.

Outliers were removed from the original data set i.e. points lying more than plus or minus 2 standard deviations from the mean were identified and removed. As the max and min values for the features V1 and V2 were significantly different, the V1 and V2 data points were normalized i.e. transformed in the range between 0 and 1. A scatter plot of the normalized features is shown below.

Scatter plot of normalized V1 and V2 features



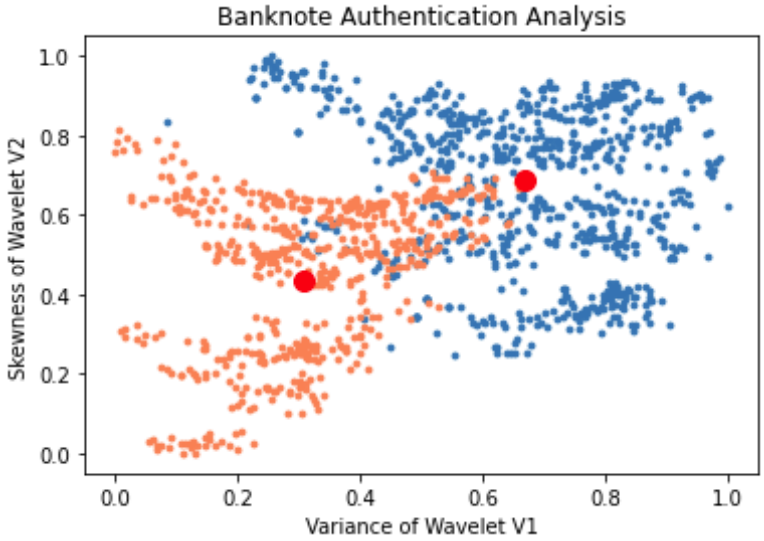
Unsupervised learning models developed

1. K-means Clustering Algorithm

K-means clustering was the first clustering algorithm developed. This algorithm was developed with 2 features V1 (Variance of wavelet) and V2 (Skewness of wavelet).

A value of K=2 (2 clusters) was chosen for the K-means clustering algorithm. The distance metric chosen was “Euclidean” as this is a good metric for coordinate based measurement’s, in lower dimensional space. For higher dimensional space the “Manhattan” metric would have been the preferred distance metric.

The K-means clustering algorithm appears to have done a reasonable job of separating the two clusters, as indicated by the centroid locations. However an examination of the K-means cluster algorithm accuracy will need to be done; see next section “key insights and findings”.



The value of the centroids are shown below.

Value of centroid clusters

```
In [49]:  1  kmean_result.cluster_centers_

Out[49]: array([[0.67049091, 0.68974512],
                [0.30842708, 0.4353383 ]])
```

The instance of the class ‘kmeans’ (KMeans) was fit on the data and new data clusters predicted.

Examination of the KMeans algorithm accuracy was undertaken; see next section “key insights and findings”.

2. Hierarchical Agglomerate clustering algorithm

Hierarchical agglomerative clustering algorithm was the second clustering algorithm developed. The distance metric chosen was ‘Euclidean’, and the number of clusters chosen ‘n_clusters’ was 2.

The algorithm was run 4 times using the following hierarchical linkage types.

(i). Single linkage; minimum pairwise distance between clusters; produced the best results
(ii) Complete linkage; maximum pairwise distance between clusters
(iii) Average linkage; average pairwise distance between clusters
(iv). Ward linkage; merge based on inertia

The instance of the class ‘agg’ (Agglomerative Clustering) was fit on the data and new data clusters predicted.

Examination of the Agglomerative cluster algorithm accuracy was undertaken; see next section “key insights and findings”.

3. Mean shift clustering algorithm

Mean shift algorithm was the third algorithm developed to cluster the banknotes into genuine and forged banknotes. Initially a bandwidth window size of 0.9 was used. This produced two distinct clusters. The bandwidth window size was then increased slightly to 0.93. This bandwidth produced the best results.

The instance of the class ‘ms’ (MeanShift) was fit on the data and new data clusters predicted.

Examination of the Mean Shift cluster algorithm accuracy was undertaken; see next section “key insights and findings”.

4. DBSCAN

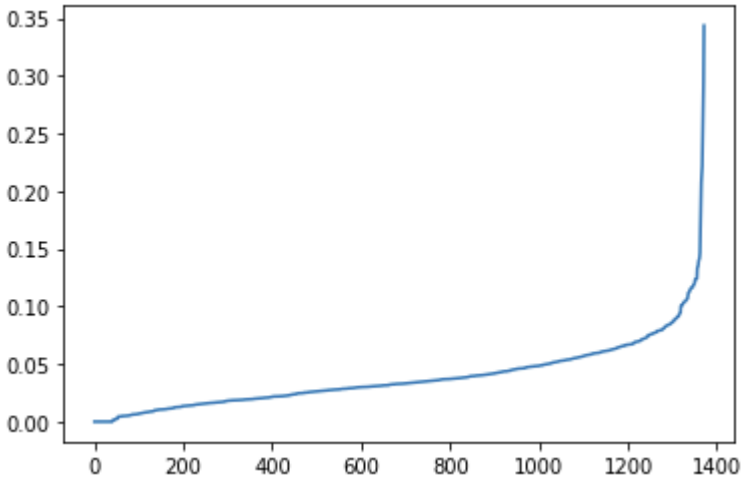
DBSCAN was the fourth algorithm developed to cluster the banknotes into genuine and forged banknotes.

The model's behaviour is dictated by parameters below:

An Epsilon (eps) value of 0.13 was used in the model; eps is where two points are considered neighbors if the distance between the two points is below the threshold epsilon.

A min sample value of 5 was used in the model; min_samples is the minimum number of neighbors a given point should have in order to be classified as a core point.

A K_Nearest Neighbors algorithm was utilized and the results graphed to determine optimal Epsilon (eps) value. The optimal Epsilon value can be found at the point of maximum curvature: see graph below.



The instance of the class 'db' (DBSCAN) was fit on the data and new data clusters predicted.

Examination of the DBSCAN cluster algorithm accuracy was undertaken; see next section "key insights and findings".

Best Performing Model

Of the four clustering models developed the model with the best performance was the hierarchical agglomerative clustering model. This algorithm identified all the genuine banknotes correctly, and more than 98% of the forged banknotes were correctly identified.

Key Insights and Findings

Summary

Of the four clustering models developed the model with the best performance was the hierarchical agglomerative clustering model.

1. K-Means Algorithm Clustering Model

The first clustering model developed was the K-Means (K=2) clustering model.

K-means cluster model accuracy

Of the banknote population analyzed, 78 forged banknotes and 90 genuine banknotes were incorrectly identified. This represents approximately 13% of the population of the two banknote categories. The overlapping of clusters of forgery and genuine banknotes may indicate a K-Means model with more than two clusters, may be a better model, especially if some forgery banknotes approximate well to genuine banknotes.

In the light of this, the silhouette algorithm was run to determine the optimal number of clusters.

The optimal number of clusters k is 6; the silhouette coefficient analysis is shown in the appendix. The K-means accuracy also improves when K=6 clusters.

2. Hierarchical Agglomerative Clustering Model

The second clustering model developed was the hierarchical agglomerative clustering model.

Hierarchical agglomerate clustering model accuracy

The ‘single’ hierarchical linkage type produced the best results. The algorithm identified all the genuine banknotes correctly. For the forged banknote population, the algorithm incorrectly identified 10 forged banknotes as genuine banknotes; this is less than 2% of the forged banknote population.

The Hierarchical Agglomerative clustering algorithm was found to be the best performing algorithm of the four.

3. Mean Shift Clustering Model

The third clustering model developed was the hierarchical agglomerative clustering model.

Mean shift clustering model accuracy

The increase in bandwidth window size from 0.9 to 0.93 improved model accuracy, where the number of forged banknotes identified as genuine dropped to 27, which is less than 5% of the forged banknote population incorrectly identified.

However, more than 25% of the genuine banknote population was incorrectly identified as being forgeries.

4. DBSCAN Clustering Model

The fourth clustering model developed was the DBSCAN clustering model.

DBSCAN clustering model accuracy

The optimal number for Epsilon was found to be approximately 0.13. An eps of 0.13 and a min_samples value of 5 was used in the model.

The number of outliers as identified by '-1' label were as follows:

- Genuine banknote category 3 outliers
- Forged banknote category 13 outliers.

The number of forged banknotes incorrectly identified as genuine was approximately 7% of the forged banknote population with 93% correctly identified.

However, more than 15% of the genuine banknote population was incorrectly identified as being forgeries.

Further work (next steps): Model Improvement

A larger dataset is available with 4 features (V1, V2, V3, V4). The performance of the model may be improved by using all four features in the model.

These four features or dimensions may be reduced to two dimensions using a dimension reduction transformation such as Principal Component Analysis (PCA). The K-Means algorithm would then be run again on the reduced dimensions.

By utilizing all four features a more accurate clustering model may be developed.

Conclusion and Recommendations

The K-Means clustering machine learning model performs well using the two features V1 and V2, and is a good benchmark model.

However the best performing model was the Hierarchical Agglomerative Clustering Model, which correctly identified all genuine notes and more than 98% of the forged notes correctly.

A larger dataset is available with 4 features (V1, V2, V3, V4). The performance of the model may be improved by using all four features in the model. These four features or dimensions may be reduced to two dimensions using Principal Component Analysis (PCA), and the K-Means algorithm or another algorithm run again.

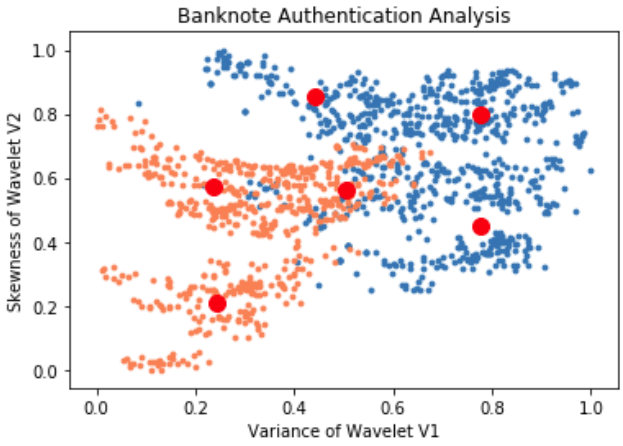
It is recommended that with further work using the larger dataset the bank could use a clustering algorithm model to predict/detect whether unseen banknotes are genuine or forgery.

Appendix

K-Means with Optimal number of clusters (6 clusters):

The optimal number of clusters ‘k’ is 6; silhouette coefficient analysis.

Graph of genuine and forged banknotes with centroid locations (red dots)



Summary of results

Centroid positions

```
([[0.2357243 , 0.57359129],
  [0.50473609, 0.56284568],
  [0.24351117, 0.20774766],
  [0.44098184, 0.85494262],
  [0.77617328, 0.45008065],
  [0.77593176, 0.80009398]])
```

The coral coloured dots indicate forged banknote data instances, and the blue dots indicate genuine banknote instances. There are 6 centroids in the above scatter plot; these are indicated by red dots.

For unseen banknotes to be tested using the k-Means model whose feature V1 (variance of wavelet) and V2 (skewness of wavelet) values are close to the following centroid coefficients, the table below would indicate with reasonable probability whether the banknote was genuine, forgery or indeterminate.

The model does a good job at predicting forgery bank notes that are cluster around the bottom left centroid 0.242 and 0.207 and the mid left centroid 0.235 and 0.573. Where the model does not perform well is in the centre at centroid position 0.504 and 0.562. For the remaining 3 centroids the model performs well predicting genuine bank notes.

Centroid position	V1	V2	Result
Top middle centroid	0.4420	0.8548	Indicates genuine banknotes
Top right centroid	0.7768	0.7938	Indicates genuine banknotes
Bottom right centroid	0.7778	0.4400	Indicates genuine banknotes
Bottom left centroid	0.2435	0.2077	Indicates forged banknotes
Left center centroid	0.2363	0.5738	Indicates forged banknotes
center centroid	0.5073	0.5627	Indeterminate result