automating the detection of forged banknotes

Part 1 – The purpose of this project

The detection of forged bank notes is an incredibly important part of the daily operation of any bank. In 2019, around 427,000 counterfeit Bank of England banknotes with a face value of £9.8 million were taken out of circulation.

The process of detecting forged banknotes manually may be time consuming and possibly inaccurate, and for these reasons it may be worth considering automating the process of detecting forgeries.

In this report we will consider a model which has been trained to test features of various banknotes and attempt to detect forgeries. We will then discuss the results of this model and whether this solution could be of benefit to your bank.

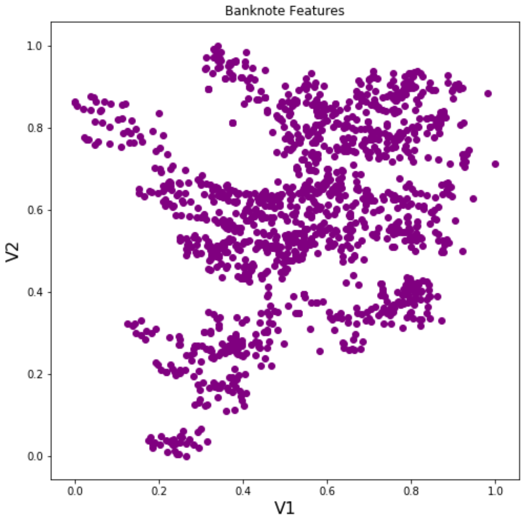
The model used for this demonstration is an algorithm called *K-means Clustering*. This algorithm looks at a given set of data and sorts the data into a user-defined number of categories, or *clusters*. Since we want to test if a bank note is real or not, we set the algorithm to provide us with just two of these clusters: **Genuine** and **Forged**.

Part 2 – Describing the data

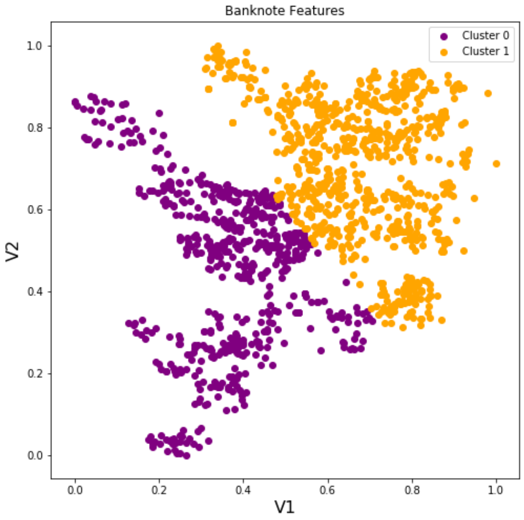
The data we used to train our model is the **Banknote authentication Data Set** from University of California, Irvine. This dataset contains observations of 1372 individual bank notes, each with four attributes we can measure with our algorithm. The attributes themselves are based on *image wavelet transformations*, which are essentially methods used to convert images in to numbers the algorithm can make use of. For the purpose of this model we are only considering two of these transformations which are labelled **V1** and **V2**.

Part 3 – Analyzing the data

To visualise the banknote data, we first produced a scatter plot to observe the relationship between our two parameters **V1** and **V2.** Each point in the below graph represents a single bank note in the data. Note that at this point the data is *unlabelled* – that is we do not know which notes are genuine or counterfeit.

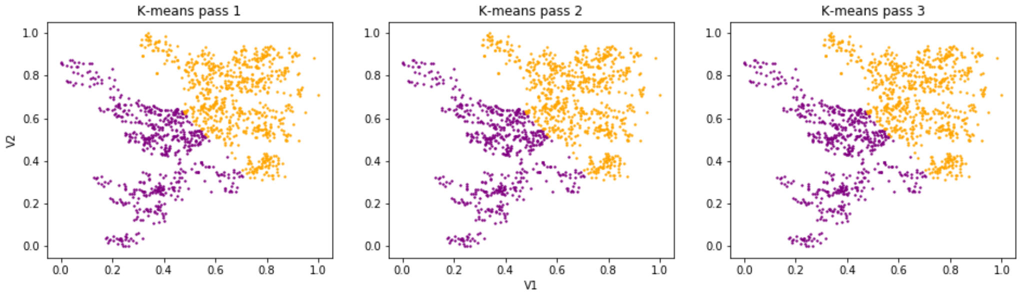


We can see from the data that there is a significant degree of variance for both **V1** and **V2** in the bank notes in our sample, however by looking at this data it is difficult to suggest where a line could be drawn to divide these notes in to two categories. This is where our algorithm comes in to use.

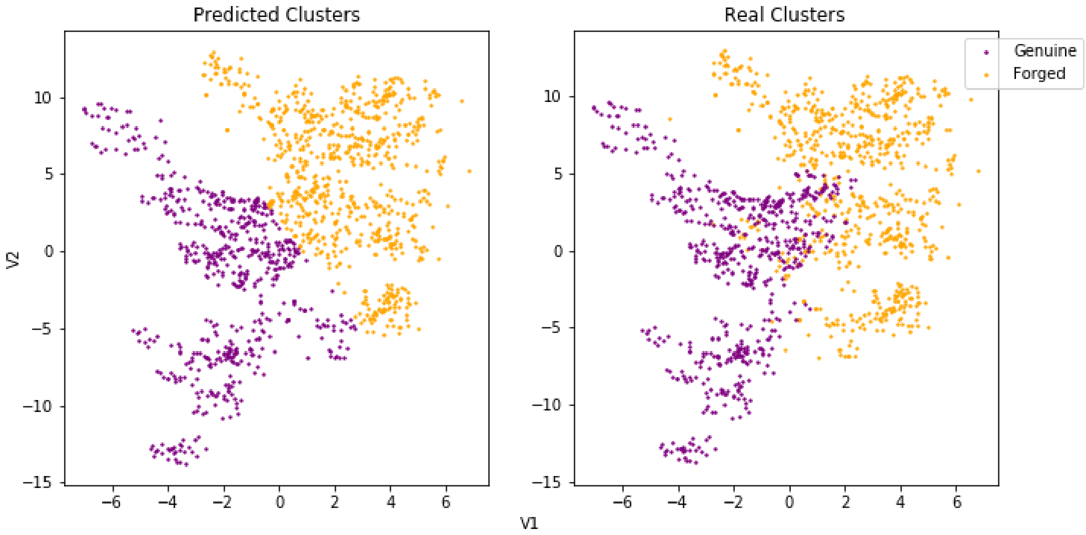


Here we have used the *K-means Clustering* algorithm to split the data in to two *clusters*. Though we haven't labelled which of these clusters represents real or forged bank notes at this time, what is important is that the algorithm has analysed the data and defined the two most prominent clusters.

Part 3 cont.



Each time we run K-means Clustering, the algorithm begins by creating starting points for each cluster at random. This means we can run the algorithm multiple times and test if we obtain the same results. Since in this case we have consistently identified the same clusters in the data we can be confident the model is working as intended and the clusters it has identified are a reliable means by which to group our data points.



Here we have plotted our predicted results (left) against that of the real data (right). We can see from these visualisations that the clusters identified by our model are somewhat accurate. In fact, out of 1372 bank notes, our model was able to correctly classify 1197, or **87.2%.**

Part 4 – Recommendations

The K-means Clustering algorithm was able to successfully identify genuine and forged bank notes 87.2% of the time. While 87.2% is a reasonable rate of success, it is important to consider several factors when considering whether to implement this model in your business.

It is worth noting that 87.2% is an *overall* accuracy score for the model. Of the banknotes labelled incorrectly, these are a combination of *false positives* (i.e. genuine notes labelled as forged) and *false negatives* (forged notes labelled as genuine). This could be investigated further, and it would be worth considering which of these errors is best avoided. Is it more damaging to dispose of genuine notes, or to unknowingly recirculate fraudulent notes?

A solution to be considered may be using multiple methods of identification. This model could be used alongside a secondary source of identification in order to 'double check' the results of the other system and create a more accurate system overall.

Finally, it is left to the business to consider the cost of any given system. While using a model to predict the authenticity of banknotes may be more accurate than what is possible by a human or alternative system, steps need to be taken to measure the various qualities of any given note and input these measurements in to the model. These measurements could be made by machines or people, but either way would incur costs which would need to be considered.