Bank Note Data Science Report

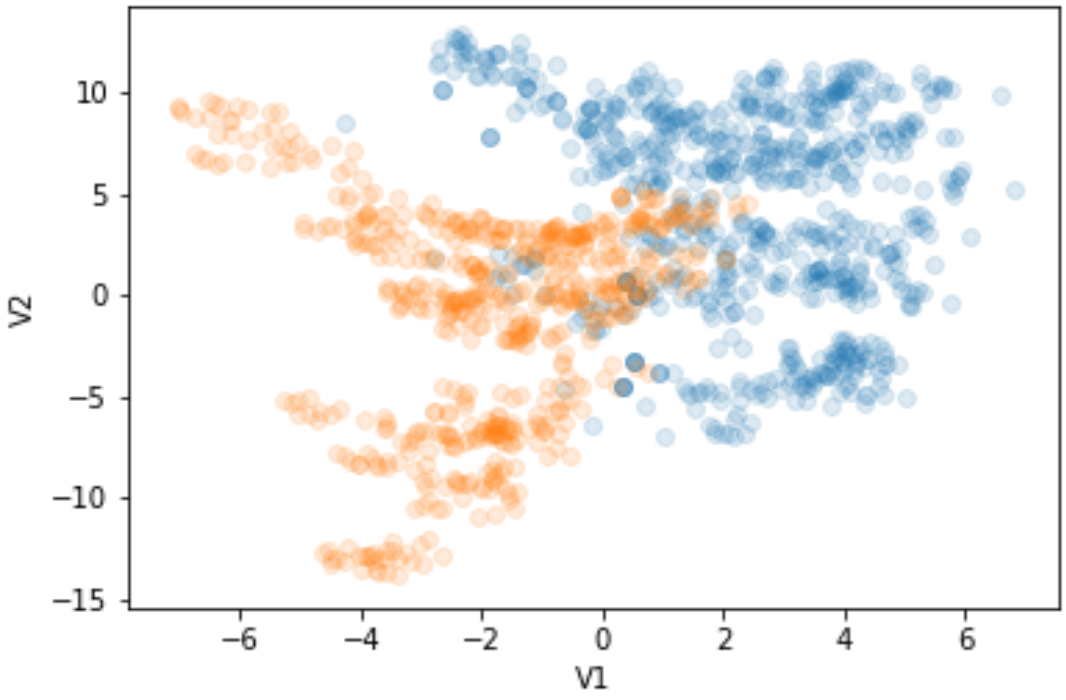
Introduction:

In this report I will describe the project on bank note data that I have done. I will discuss the data that I used, the methods and algorithms that I used, and finally the conclusions that I can draw from. The purpose of this project is to see whether we can use data science techniques to evaluate whether we can distinguish authentic banknotes from fake banknotes in this dataset.

Data:

The data I used was "banknote-authentication", which was from OpenML. The data contains 1372 entries and has two features: V1 and V2. The data also contains a Class column where one of the classes is authentic and the other is fake. A plot of the data is shown in Figure 1, with Class 1 displayed in orange, and Class 2 displayed in blue.

Figure 1



Methods:

I used a unsupervised learning algorithm called K-means clustering. The algorithm takes the data, which is each data point's two features as input. Note that I do not give the class to the algorithm as input. The output of the algorithm is K centroids which are cluster centers. The data is partitioned such that each data point is belongs to the cluster with the closest centroid. For this project, I used the K-means clustering algorithm with K=4 clusters.

Results:

By eyeballing the data in Figure 1, it looks like there is a separation between the classes of banknotes. Next, I will use a more formal method to perform analysis.

I ran the K-means clustering algorithm with K=4 and found the following centroids (cluster centers). In Figure 2, the centroids are shown in blue and the clusters are distinguished by color. Figure 3 shows the centroids in red, and again shows the true class designation from the data. By comparing Figures 2 and 3, we can see that K-means clustering algorithm does a good job of finding nodes that are in the same class. In particular, the banknotes in leftmost clusters appear to be similar to the ones that are in Class 1 (orange in Figure 3).

Figure 2

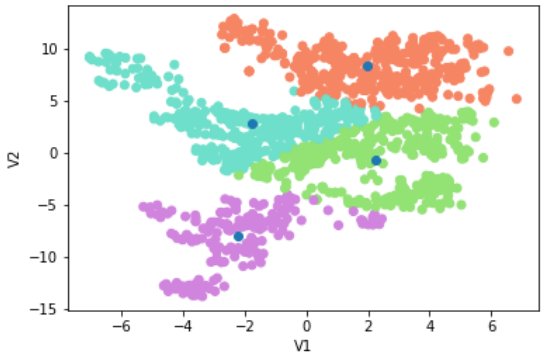
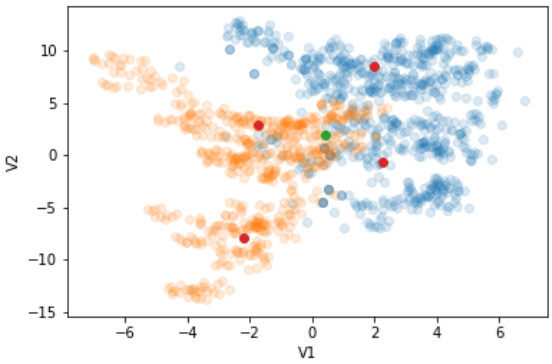


Figure 3



Recommendations to the client:

Based on our results, I recommend using our method to distinguish authentic banknotes from fake ones. In particular, I recommend using the K-means clustering algorithm with K set to 4, in order to partition the data into clusters.