Data Mining Report By Ameya Sinha

The data for the credit card transactions was given to us in a csv file. After loading the data into a pandas DataFrame an initial **data exploration** lead to the following discoveries -

- The Data consisted of 284909 rows and 31 columns. The columns represented the Time, Amount, components from the PCA and the categorical variable.[1]
- There were no NULL values in the original dataset.[2]
- There were twenty-one columns which had a dtype of float64, two columns which had a dtype of int64 and 8 columns (V1 - V6 and V24 - V25) which had a dtype of object.[3]
- On further exploration of the columns which were of object type it was seen that they have data such as '?' and ','. **This leads us to the conclusion that yes data preprocessing is needed as the given data is corrupted.** The corrupted data was then converted to nan and removed from the DataFrame.[4]
- There were also duplicates present in the data. They were also removed. [5]
- The class value denoted where the particular transaction was a fraud or not. The data in the columns that were not corrupted consisted of 283253 non fraud transactions and 473 transactions that were of fraud type. [6]
- On generating a heatmap from the seaborn library it was seen that columns - 'V1', 'V3', 'V5', 'V6', 'V7', 'V9' ,'V10', 'V12', 'V14', 'V16', 'V17', 'V18' have a relatively high correlation with the class column. [7]
- It was also seen that columns 'V28', 'V24' ,'V23', 'V22', 'V26', 'V13', 'V15', 'V25', 'Amount' had a correlation of less than 0.01 with the class column. They were further dropped. [8]
- The correlation of the data after this preprocessing step appeared to be uniformly correlated.
- This histograms of the columns that were most correlated to the class appeared to have less variation then the column which had the least variation. [9]
- The column V19 had a range of about 13 whereas other columns such as V5 and time had a larger range. As the range between the columns are not twice or thrice other columns and as clustering algorithms use the measure of distance in their internal workings we conclude that yes **Normalization of the needed.** The data was then normalized using the MinMaxScaler. [10]

Insights / Inferences / Results of running the K-Means Algorithm - When we used the Elbow Method [11] to see what should be the ideal number of clusters in the dataset the number turned out to be 2. There

---

[1] Out[3] in Data Mining Assignment.pdf
[2] Out[4] in Data Mining Assignment.pdf
[3] Out[4] in Data Mining Assignment.pdf
[4] Out[5] and Out[6] in Data Mining Assignment.pdf
[5] Out[14] in Data Mining Assignment.pdf
[6] Out[15] in Data Mining Assignment.pdf
[7] Out[17] in Data Mining Assignment.pdf and correlation_matrix_initial.jpg in Figures
[8] Out[18] in Data Mining Assignment.pdf
[9] Out[19] and Out[20] and Out[21] in Data Mining Assignment.pdf and histogram_V17.jpg, histogram_V14.jpg, histogram_V25.jpg in Figures
[10] Out[31] in Data Mining Assignment.pdf
[11] Out[32] in Data Mining Assignment.pdf and Elbow_Method.jpg in Figures

was a huge drop that was seen when we increased K from 1 to 2 but on increasing the value of K further we found out that the values decrease by a little amount from here.
K-Means was then run on the dataset with the values of n_clusters as 2, init as k-means++ and random_state as 0.[12] We then assumed that the cluster with the larger number of points is the one which should have the label as non-fraud (0) and the cluster with the smaller number of points the label fraud (1).[13]
We thus got 152714 tuples predicted as non fraud and 131012 tuples predicted as fraud.[14] The **mean square error**[15] of the clustering was found to be 0.4623932949394839 and the **root mean square error** [16]was 0.6799950697905712. Accuracy[17] of the clustering was found to be 53.76067050605161% and the Recall[18] and Precision[19] values were found out to be 30.866807610993657 and 0.11144017341922878 respectively. The **correlation** between the predicted value and the original values were found out to be -0.01254835.[20]
From the above insights we infer that K-Means is not a suitable algorithm to cluster the data as it not only yields a poor accuracy but also the recall and the precision values are not very promising. The precision value in fact is too low. This means that of all the people that we notify to have a fraudulent transaction very few of them actually have so. K-Means , therefore, is not a suitable algorithm for this dataset.

Insights / Inferences / Results from computing the Two Main Principle Components -
When we compute the two main Principle Components and plot them in a scatter plot we see that the data appears to be quite Density based.[21]
There seems to be a huge cluster and all the points far away from these clusters seem to be anomalies.
Here we infer that Density based clustering should be the most appropriate for this dataset to mimic the real world dataset. However the dataset appears to be huge and DBSCAN does not run on the dataset as the kernel dies quickly.

[12] In[33] in Data Mining Assignment.pdf
[13] In[36] in Data Mining Assignment.pdf
[14] Out[36] in Data Mining Assignment.pdf
[15] Out[37] in Data Mining Assignment.pdf
[16] Out[38] in Data Mining Assignment.pdf
[17] Out[42] in Data Mining Assignment.pdf
[18] Out[44] in Data Mining Assignment.pdf
[19] Out[46] in Data Mining Assignment.pdf
[20] Out[48] in Data Mining Assignment.pdf
[21] Out[50] in Data Mining Assignment.pdf and Most_Significant_Principle_Components.jpg in Figures