CS F415: Data Mining Assignment-1
REPORT

Name: Arkil Patel
ID: 2016A7PS0665G

I. Answers to Questions

Ans 1. Yes, we need data pre-processing. The data given consists of non-numeric data types (or missing values) in certain columns as well as duplicate values. These need to be removed by data cleaning techniques.

Ans 2. Yes, we need normalization. This is because the attributes have differing ranges of values. Hence it is imperative to normalize all values to a common scale, else some attributes may incorrectly dominate the clustering process.

Ans 3. a) Root Mean Squared Error was found to be 0.03 as calculated in code snippet at In[33].
b) Correlation between Class and Predicted Column was found to be 0.69 as calculated in code snippet at In[32].

Ans 4. Theoretically, DBSCAN should get me closest to the actual result. This is because DBSCAN is a density-based algorithm and since our task is to find outliers in the data (i.e. fraudulent data) which will be characterized by low density region, DBSCAN would be the optimal choice. However, running DBSCAN on such a large dataset requires higher computational power and hence I have developed my model based on K-Means Algorithm.

II. Clustering Model

I have developed a clustering model using the K-Means Algorithm. Initially, I ran the algorithm on all attributes, but the results were disappointing (see accuracy at Out[15]). Hence, I decided to drop all those columns which do not have a considerable correlation with Class (see In[15]). From the initial 30 columns, I now had 14 columns which were making a considerable impact on Class. On deriving the Elbow curve for this (see Fig2.), I found the K value to be 8. After running K-means on the data, I found that the Fraudulent data was best represented by the 8th cluster (index 7) and the Non-Fraudulent data by the other 7 clusters.

Columns Dropped: V25, V15, V13, V26, V22, Amount, V23, V24, V28, Time, V20, V27, V21, V8, V19, V6
Number of Clusters: 8
Predicted Classification: All points lying in Cluster Index 7 are fraudulent. All other points are not fraudulent.

III. Result

Accuracy: 99.91%
Precision: 85.81%
Recall: 54.97%
Root Mean Square Error: 0.03

IV. Observations and Inferences

- The data given is very skewed in the favour of Non-Fraudulent Transactions. Hence, the Fraudulent Transactions may be considered as Outliers.
- We cannot derive any domain knowledge from the attribute names since they are marked as V1, V2, and so on. Hence, we will have to rely on the Correlation heatmap to find domain knowledge.

- Almost half the columns in the data have negligible correlation with Class. These columns contain data which does not contribute significantly towards predicting the Class and hence may cause the Clustering algorithm to go wrong. Hence, the accuracy of the clustering algorithm may be improved by dropping these columns. This was found to be true.
- The elbow curve is smooth and doesn't give a clear value of k. However, it is easy to see that K will belong in the range around 8. By means of trial and error, 8 was found to give the maximum accuracy.
- We have no basis for initialising the initial points required for K-Means ourselves, since we cannot consider the Class column for clustering.
- Even though the accuracy and precision of the model are high, the recall is low. This means that the model is very picky. It doesn't classify many points as Fraudulent. Those that it classifies to be fraudulent are very likely to be actually fraudulent (high precision). The drawback is that, in being so selective, the model misses out on a considerable number of actually fraudulent points (low recall).
- An attempt to apply Agglomerative Clustering failed, giving Memory Error. For this assignment, Agglomerative Clustering runs on a sample of about 30000 tuples, but gives Memory Error on passing a larger sample. This defeats the purpose since we hope to cluster about 275000 tuples.
- To visualize the clustering, we can plot a 2D graph between any attribute and Class. The points should be colour separated based on class value as 0 or 1 (see Fig4.).

V. Conclusion

In this Assignment, we have hoped to Classify a transaction as either fraudulent or not by means of Unsupervised Learning, i.e., Clustering. After dropping the unrelated columns, I have developed a model by running the K-Means algorithm to divide the data into 8 clusters. All points lying in one particular cluster (index 7) are considered fraudulent while all others are not. If given an unknown tuple, we can add it to the dataset and run clustering on the dataset to find the predicted Class of the tuple. This model classifies the points with an accuracy of 99.91% and a Root Mean Square Error of 0.03.