CS6890 - Fraud Analytics Using Predictive and Social Network Techniques

P. Ram Anudeep - EE16BTECH11027
V. Revanth Babu - ES16BTECH11027
. C.Jeevan Chandra - CS16BTECH11042.

Dataset:

Given dataset has 98,310 entries of the form :

id	month	Total sales	Exempt sales	sgst liability	cgst liability	igst liability	total liability
sgst_cash setoff	cgst_cash setoff	igst_cash setoff	total_cash setoff	Sgst_itc claimed	cgst_itc claimed	igst_itc claimed	Total_itc claimed

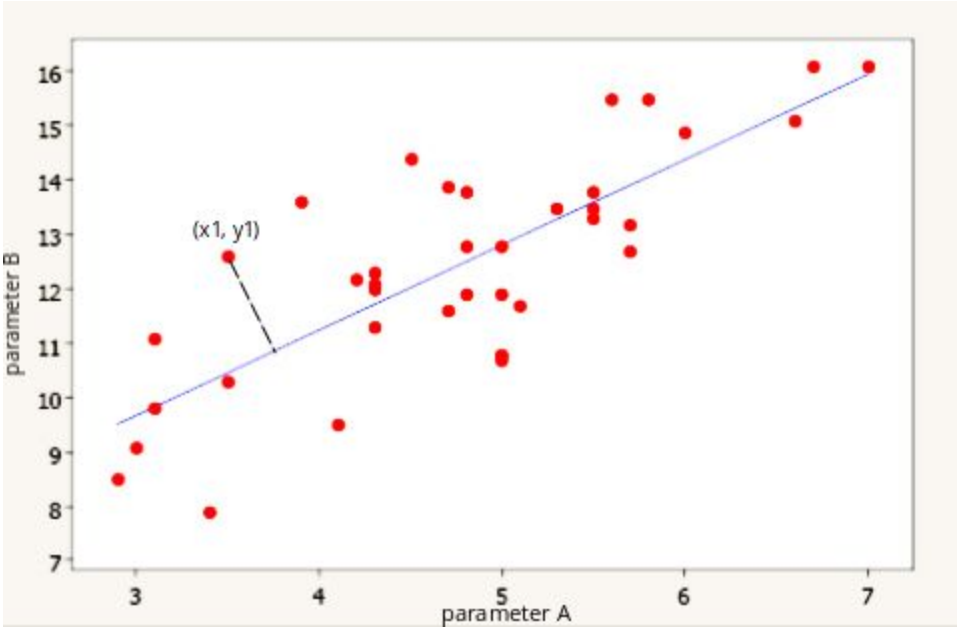
There are 11201 number of unique ids and the months given are ranging from 7-2017 to 6-2018. Also we can find a large number of entries.

Approach:

We first found the 9 parameters for an id in every month. The 9 parameters are:

1. Correlation of Total Sales Amount and Total GST Liability.
2. Correlation of Total GST Liability and Total SGST Liability.
3. Correlation of Total SGST Liability and Total SGST paid in cash.
4. Correlation of Total Sales amount and Total SGST paid in cash.
5. Correlation of Total Tax Liability and Total ITC.
6. Correlation of Total ITC and IGST ITC.
7. Ratio of Total Sales vs Total Purchases.
8. Ratio of IGST ITC vs Total ITC.
9. Ratio of Total Tax Liability vs IGST ITC

Once we have identified two variables that are correlated, we would like to model this relationship. We want to use one variable as a **predictor** or **explanatory** variable to explain the other variable, the **response** or **dependent** variable.
In the below graph for the Correlation of Total Sales Amount and Total GST Liability the parameter A will be Total Sales Amount and the parameter B will be total GST Liability.
The point (x1, y1) represents the total sales amount and total gst liability for an id in a month.
Then we find the distance between the point and the best fit line.



We do this process for get the correlation values for all months. Also we find the 3 ratio parameters directly using values for an id for all months. If the denominator is 0 for any of these values we are making that a high value.

Now we have 9 parameter values for all the transactions.
We can now find the euclidean distance between two transactions using these parameters and create an adjacency matrix for a month.
Now that we have a graph we applied spectral clustering algorithm to find the outliers for each month. We can then give a threshold value of months for which if the id is in the outliers for months greater than the threshold then the id is a true outlier.

Algorithm : (Parameters : Number_Of_Clusters, Cluster_Threshold_Size)

1. Split the Dataset into month wise data.
2. Removed all the entries which has zero sales.
3. Converted month-wise data with actual attributes to new data matrix with attributes as those 9 parameters mentioned above.
4. For each month :
    a. Convert the above data_Matrix to adjacency matrix for clustering.
    b. Applied Spectral Clustering with (Number_Of_Clusters) as a parameter for each of the month-specific data.
    c. Marked all the traders as suspicious whose cluster size is less than the threshold mentioned above.
4. Returned all the suspicious traders into a text file.

Parameters given :
(Number_Of_Clusters = 10, Cluster_Threshold_Size = 30)

Results :
We got a total of 256 suspicious traders, which are mentioned in the file **Suspicious Tax Payers.txt.**

Validation :
We did not have a proper validation set for realising what we have done is right. So below I have plotted t-SNE embeddings of the dataset for each of these months. Whoever turned out to be suspicious in a given month from the above algorithm is marked red in the below graph. We can clearly see that the shady traders are on the boundary on all of these t-SNE plots. No where I can see the plots of these traders well-inside. This is a sort of validation for us that the traders we got are shady.

T-SNE Plots :