Identifying Return Defaulters

Nithin
CS16BTECH11005

Surya Pramod
ES16BTECH11015

Krishna Chaitanya
CS16BTECH11011

Preprocessing

First we generate new parameters from the attributes given in the dataset. This is done because previous work in the literature shows that directly using the raw data of tax returns rarely works well for identifying return defaulters. So instead we derive new parameters.

Multi-stage Benford Analysis

We first run Benford analysis on the complete data set to see if it contains manipulated data. We also run Benford analysis on the final clusters obtained to see how manipulated they are.

Generating Correlation Parameters

We fit linear regression lines between the following pairs of attributes given in the raw data.

- total_sales and total_liability
- total_liability and sgst_liability
- sgst_liability and sgst_cashsetoff
- total_sales and sgst_cashsetoff
- total_liability and total_itc_claimed
- total_itc_claimed and igst_itc_claimed

Then we find the distance of each data point from this fitted line. This distance is used as a new attribute for clustering.

Generating Ratio Parameters

We generate the following ratio parameters to use in clustering:

- Ratio of total_sales and total_purchases [1]
- Ratio of igst_itc_claimed and total_itc_claimed
- Ratio of total_liability and igst_itc_claimed

Note that this list of parameters is by no means exhaustive. Other parameters that might turn out to be useful could be added later.

Finding Clusters

We plan on implementing and comparing the following approaches towards clustering:

- k-means, k-medoids

  We expect this method to not work very well. As these clustering approaches assume that the underlying clusters are somewhat spherical in nature. But this might not be the case for the data we have.

- Hierarchical Clustering

  The problem with this approach is that it is very sensitive to noise and also quite expensive to compute, but as our underlying goal is to find outliers among millions of records, this might not be a suitable approach.

- Spectral Clustering
- DBSCAN

  Spectral and DBSCAN do not suffer from the above mentioned disadvantages (computational complexity and structure of data), so we expect them to give useful clusters.

Detecting Outliers

Once we have the clusters we can get two types of outliers. Those that do not get classified into any cluster and those that are at the boundaries of clusters. These are reported as suspicious. If any of the obtained clusters are reasonably big, we can use Benford analysis on them to see how legitimate these clusters are.

[1] The parameter total_purchase is not present in the data set, but it can be derived using the given data.