Ensemble Outlier Detection using Clustering for Identifying Tax Evaders.

Nithin

Surya Pramod

Krishna Chaitanya

Preprocessing and Data Analysis:

First we remove some records where the following parameters are zero.

- total_sales : When there are no sales, there is no need to pay GST so such return need not be considered.
- total_liability : The ratio of IGST ITC and total_liability should not be NaN as we are going to use this as an attribute while clustering.
- total_itc_claimed : The ratio of IGST ITC and total_itc_claimed should not be NaN as we are going to use this as an attribute while clustering.

Next we drop the month attribute and group the data by the id of the taxpayer and then average the payments of each taxpayer to a single record. This helps to decrease the size of the data so the clustering algorithms converge faster.

Benford Analysis:
Performing Benford Analysis on values in different attributes shows us considerable deviation on attributes like exempt_sales, igst_liability, igst_cashsetoff and igst_itc_claimed. This is in accordance with our intuition that records involving inter-state sales and exempt sales are more likely to be altered.

The results of Benford analysis can be found at the end of the report.

Correlation Parametres:

We obtain the correlation parametres by first fitting a linear regression model to the parameters that we are correlating. Then we find the distance of each record from this fitted line. This distance is used as an attribute while clustering.

Then using the correlation parameters and ratio parameters mentioned in the paper we create a new set of attributes for use in clustering.

Note: The ratio parameter of total_sales vs total_purchases is not used as the purchase information is not contained within the data set (this would have been possible if gst rate was fixed, but this is not the case).

Clustering:

We are using an ensemble of multiple clustering algorithms to predict the outliers in the data.

K-means Clustering:

K-means clustering first chooses n points at random as clusters centres and then clusters the other points based on their distance from these centres. It then recalculates the centres and continues this process until there is no change in the clusters.

K-medoids Clustering:

In k-medoids clustering we first choose n points at random as medoids and then cluster the other points based on their distance from these medoids. Then we swap each medoid with a point and check if we get better clusters (with lower cost) we keep the swap else reverse it. We continue this as long as the cost decreases.

DBSCAN:

DBSCAN starts from a point and classifies all points closer than a distance **r** as being in the same cluster and continues this process from each of these points. Points that don't have any other points in their radius are classified as noise.

Hierarchical Clustering:

In hierarchical clustering we start with **n** clusters and merge the closest pair of clusters. We keep doing this until we have the required number of clusters.

Spectral Clustering:
In spectral clustering we first find the graph laplacian and use the eigenvectors of that to split the graph till we have the required number of clusters.

Most of these clustering methods do not work on the complete dataset. They either take a lot of memory or a lot of time or both. Some of them like kmeans, DBSCAN, can be made to run by adjusting the parameters. But these clusters are not very useful.

Ensemble:

Each clustering method produces a bunch of outliers. These count as votes for each person(id) in the dataset. We then sort the people in the descending order of the votes to get people who are likely to have committed tax fraud. We give more weight to the outliers in DBSCAN because it has intrinsic outlier detection.

The parameters used for clustering were arrived at based on the silhouette-index of the clusters which is used to evaluate clusters when their true labels are not known.

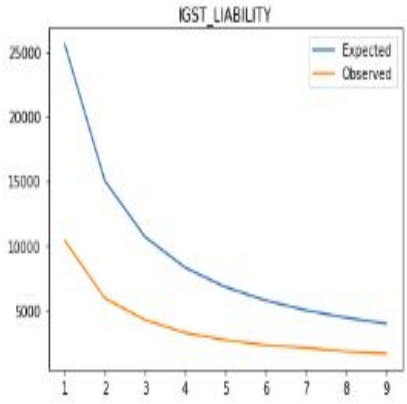
Analysis:

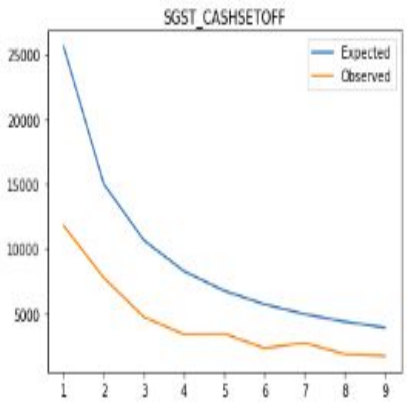
We can see a strong correlation between the predicted outliers and igst returns. The percentage of people in outliers who file igst is considerably more (65%) than that in the total data (35%).

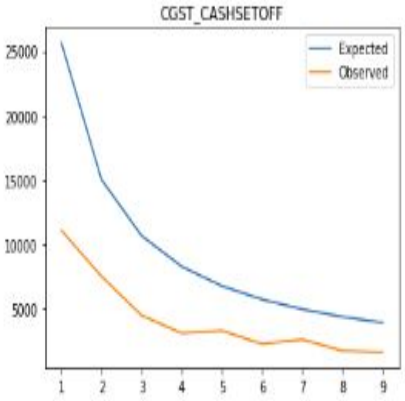
Similarly we can see a strong correlation between the outliers and exempt sales, which shows an increase in percentage from 4% to 16% from total data to outliers.

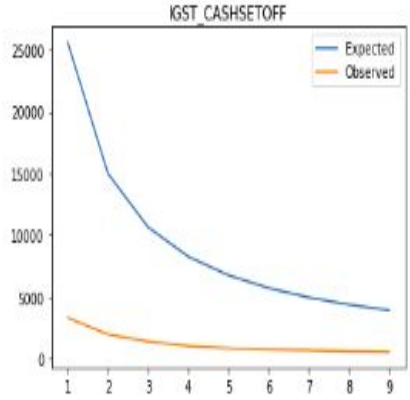
This correlation agrees with the Benford analysis of the attributes which shows igst return and exempt sales to be heavily manipulated.

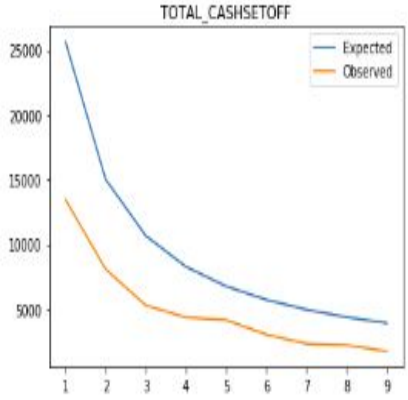
Benford Analysis Plots

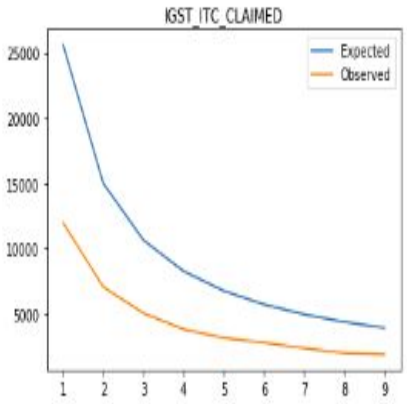


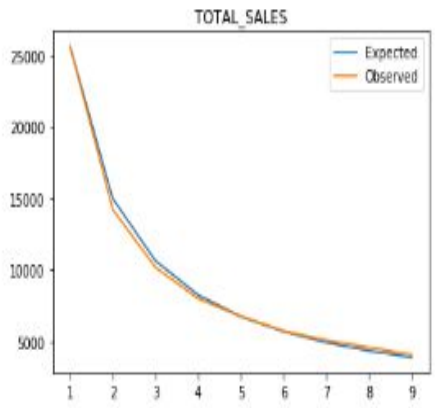


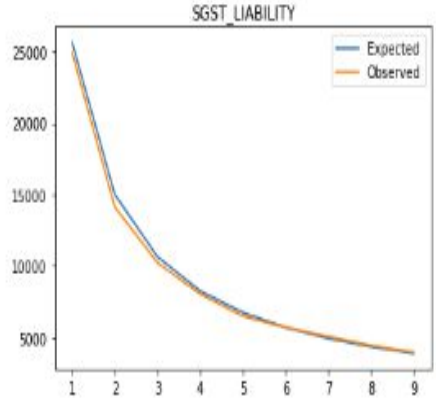


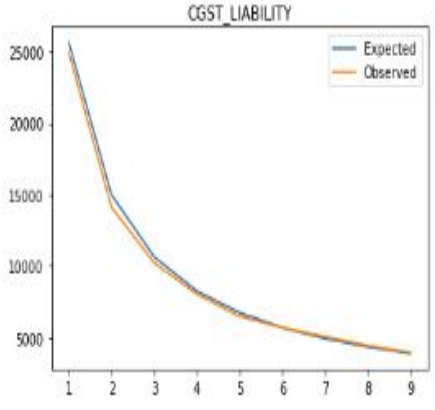


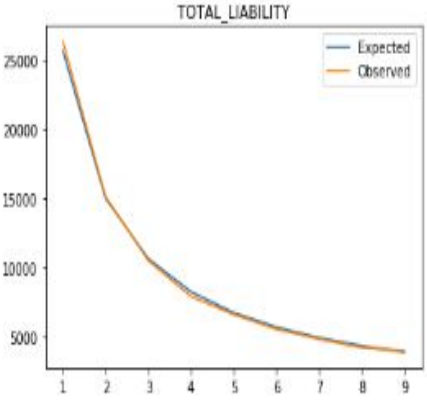


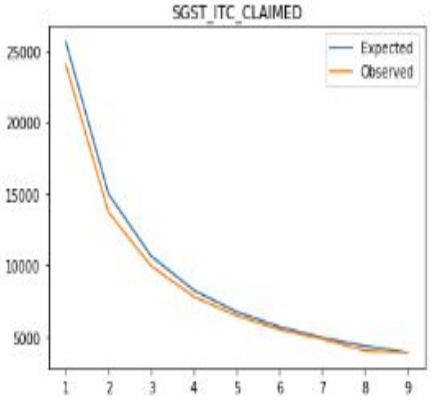


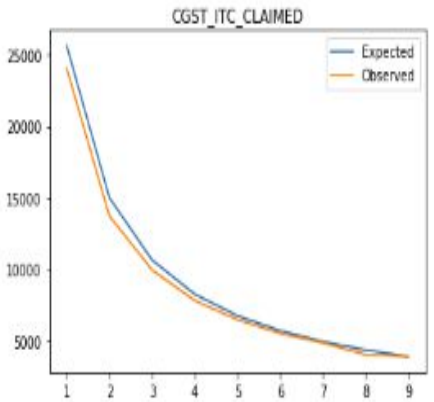

TOTAL_SALES
Expected
Observed
25000
20000
15000
10000
5000
1 2 3 4 5 6 7 8 9

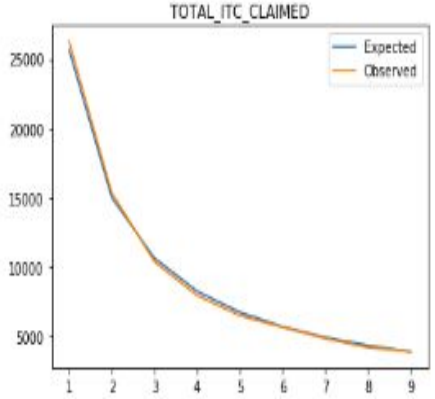

SGST_LIABILITY
Expected
Observed
25000
20000
15000
10000
5000
1 2 3 4 5 6 7 8 9


CGST_LIABILITY
Expected
Observed
25000
20000
15000
10000
5000
1 2 3 4 5 6 7 8 9


TOTAL_LIABILITY
Expected
Observed
25000
20000
15000
10000
5000
1 2 3 4 5 6 7 8 9


SGST_ITC_CLAIMED
Expected
Observed
25000
20000
15000
10000
5000
1 2 3 4 5 6 7 8 9


CGST_ITC_CLAIMED
Expected
Observed
25000
20000
15000
10000
5000
1 2 3 4 5 6 7 8 9


TOTAL_ITC_CLAIMED
Expected
Observed
25000
20000
15000
10000
5000
1 2 3 4 5 6 7 8 9