Local Outlier Factor Model

Now that we have visualized the High Dimensional Data, let us see how a machine learning model would perform analysis on the materials and segregate them into Outliers and Inliers.

Our aim is to find the Local Outliers. From T-SNE plots we have observed that the similar material come closer and form dense clusters. Whereas, different materials go away from the cluster and are likely to be the Outliers.

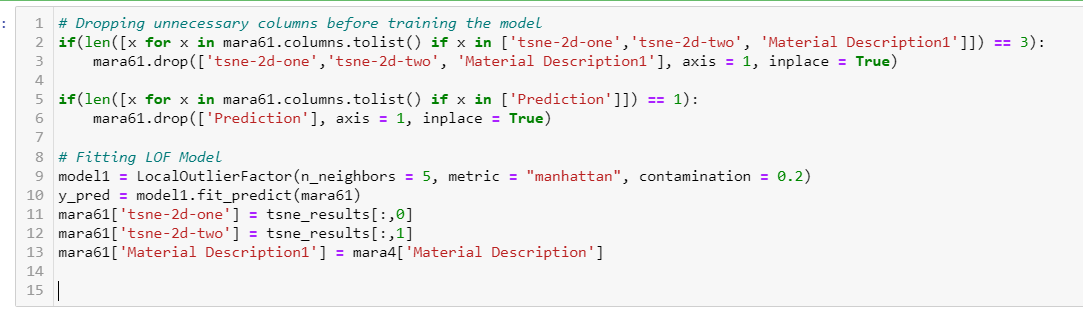
**LOF Model Functioning:**

The Local Outlier Factor (LOF) model is a density - based approach. It measures the local deviation of a data point from its neighbors, and it assumes that outliers will have a significantly lower density of neighboring points compared to the points in the dense regions.

Here is how the LOF model works:

* For each data point in the dataset, LOF calculates its distance to its nearest neighbors. LOF then compares the density of the point to the densities of its nearest neighbors.
* If the density of the point is significantly lower than the densities of its neighbors, it is considered an outlier. The degree of outlier status is determined by the ratio of the density of the point to the densities of its neighbors.
* Finally, the LOF algorithm assigns an LOF score to each data point, representing its degree of outlier status. The materials which have a high LOF score are termed as Outliers, and the materials with low LOF score are termed as inliers.

The code snippets shown here are for Material Group A880. It contains Detachable Components.

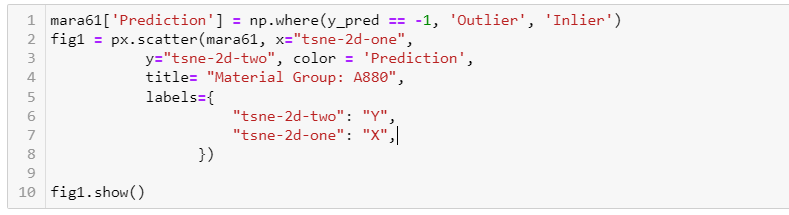


We have started by dropping out 'tsne-2d-one', 'tsne-2d-two', 'Material Description1' and 'Prediction' before fitting the LOF model. The first if statement checks if these three columns are present in the data frame and drops them if they are present. Similarly, the second if statement checks if the 'Prediction' column is present in the data frame and drops it if it is present.

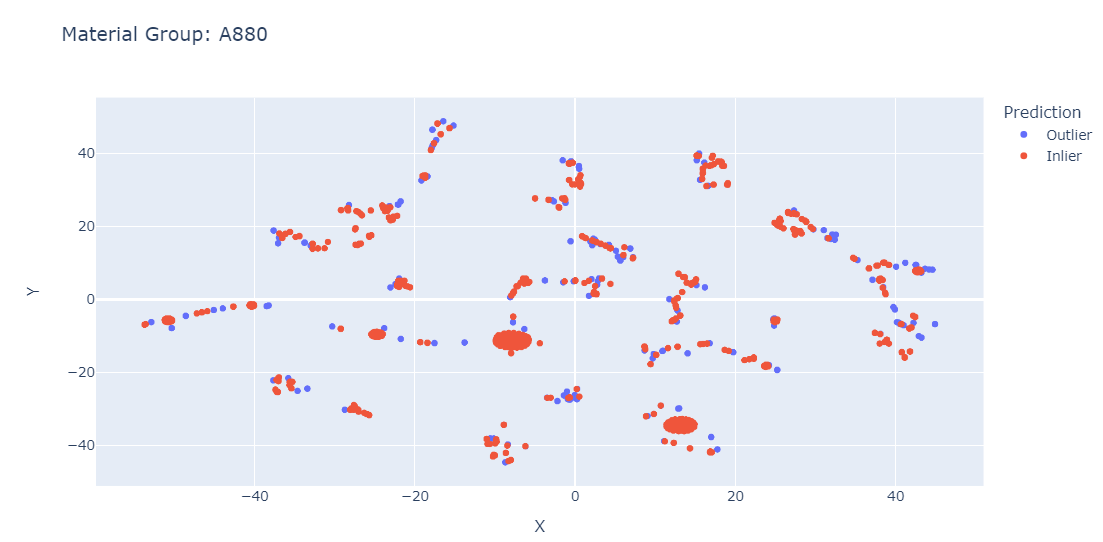
After that, the LOF model is initialized with the specified hyperparameters and fit on the remaining columns in the 'mara61' data frame. The predictions made by the LOF model are stored in the 'y\_pred' variable.

We have used a trial and error approach to assign the values in the Contamination, n\_neighbors parameters. We keep changing the values until the model clearly defines the outlier point as outliers For example, here we have assignment contamination as 0.2 and number of neighbors as 3.

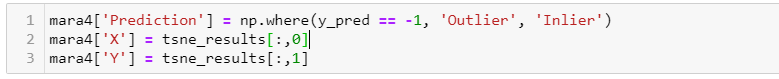
Finally, the 'tsne-2d-one', 'tsne-2d-two', and 'Material Description1' columns are added back to the 'mara61' data frame.



Here, we have created a scatter plot using the “matplotlib” library. The code creates a new column in the "mara61" data frame called "Prediction". If the value in the "y\_pred" column is -1, the value in the "Prediction" column is set to "Outlier". Otherwise, the value is set to "Inlier".



The above graph is the output of our model. Here, the points in blue color are outliers and the points in red color are inliers. With this graph, we can determine if the points are being classified into inliers and outliers properly or not. If not, then we can change the contamination and n\_neighbor values. Else, we can continue doing analysis with this.



In the python code, if we hover the cursor on the data points, its co-ordinates are displayed. These

co-ordianates are attached to mara4 data frame. Hence with these columns we can identify the outliers and inliers with their co-ordinate values.

The update mara4 Data Frame is converted into csv.

Upon analysis following are the reason for a point to be an outlier:

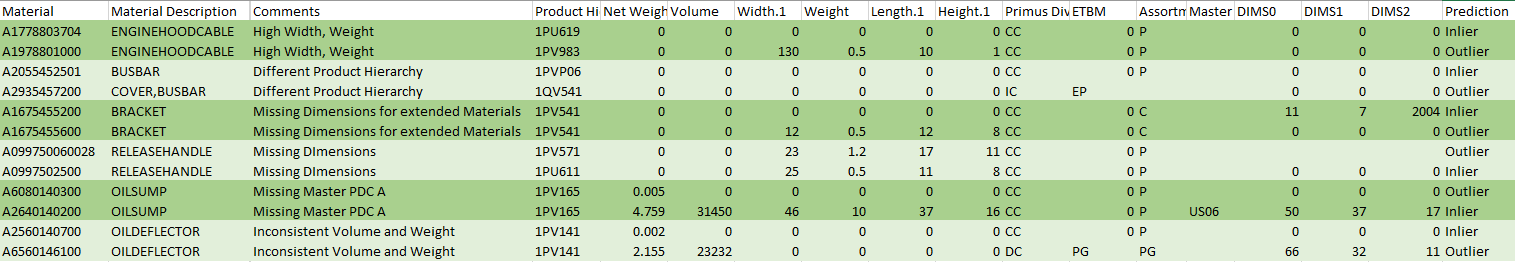
|  |
| --- |
| **Categories** |
| Large values for Volumes and Dimensions |
| Missing Dimension values for Extended Materials |
| Different Product Hierarchy, ETBM, Assortment class, Primus Division Code, Basket ID |
| Missing Master PDC A |
| Inconsistent Weights and Volume |
| Missing Values for Net Weight |

**Understanding flat files:**



Finally, the updated mara4 Data Frame (with “Prediction”, “X”, and “Y” columns is converted into a CSV Flat File).

Here is a list of Materials that have been classified as Outlier and Inliers by LOF Model for different Material Groups.



Outliers/ Inliers

Data Columns

Material Identifiers

Reason for being an Outliers