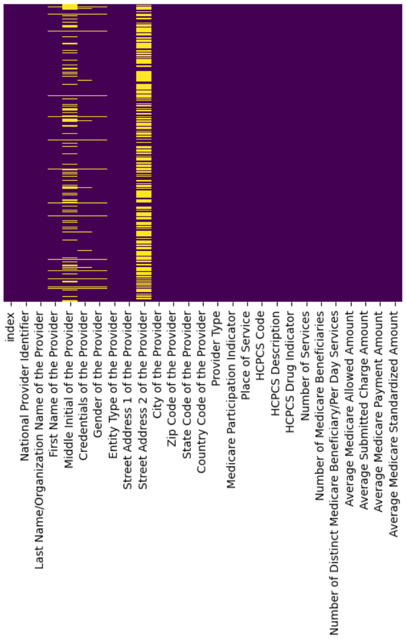
PRML - MINOR PROJECT
ANOMALY DETECTION

We have been given health care data and to detect the anomalies in the dataset.

PREPROCESSING AND EXPLORATORY DATA ANALYSIS

Data Cleaning:

Some of the columns contained ‘,’ , ‘.’ , etc. in the numeric data, and contained nan values in the dataset for some columns. The column ‘Credentials of the Provider’ contained the same value ‘MD’ in different formats like ‘M.D.’, ‘M,D’ etc. which are considered as noise in the dataset and have been removed. The heatmap plot for nan values in the dataset is shown below:

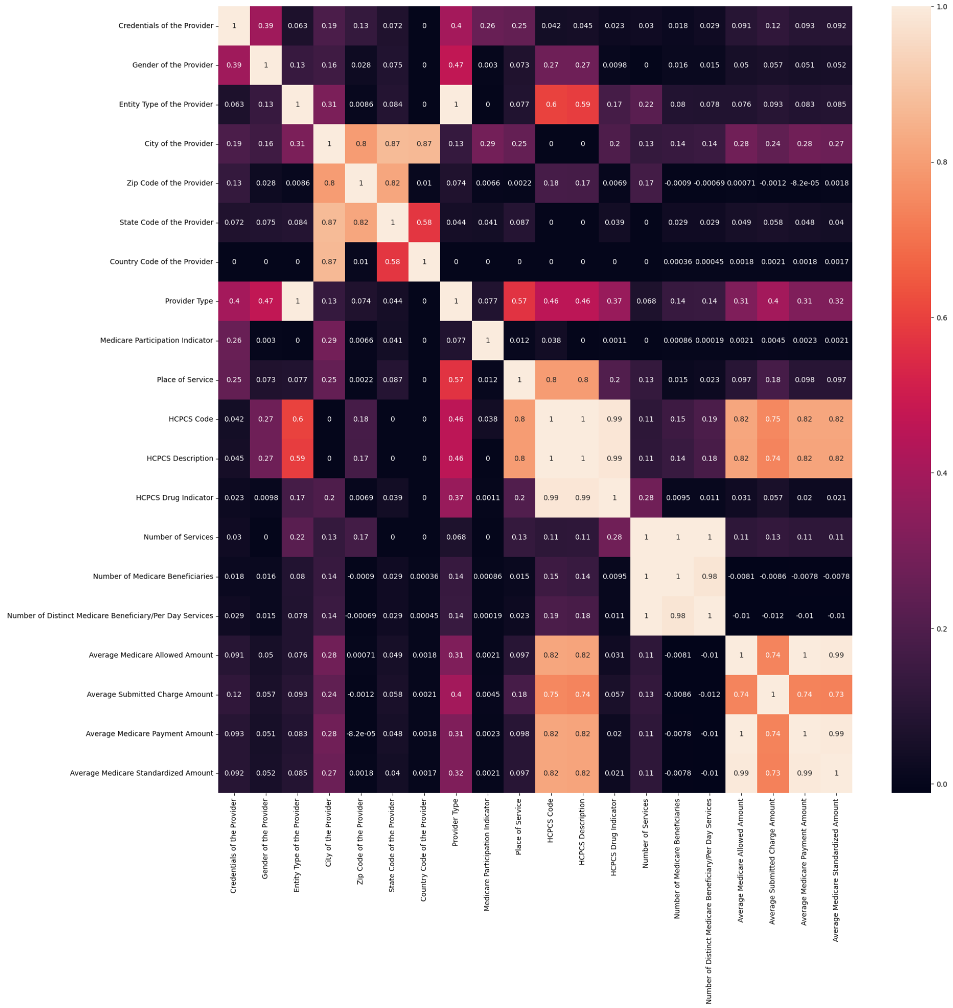


As from the plot we can clearly see that the column ‘Street Address 2 of the Provider’ and ‘Middle Initial of the Provider’ has many nan values so they are dropped. And for the remaining columns containing nan values the nan values are removed by mode of that column as they are categorical features.

Some features ('index','National Provider Identifier','Last Name/Organization Name of the Provider','First Name of the Provider','Street Address 1 of the Provider') are dropped from the data as they seem to do not contribute to the prediction of anomalies as the name,address, National Provider Identifier (id of providers) etc. these do not much affect the provider to be fraud and they also have a high number of unique values in column which is difficult to handle for model and may give wrong predictions.

FEATURE SELECTION (USING CORRELATION MAP) :

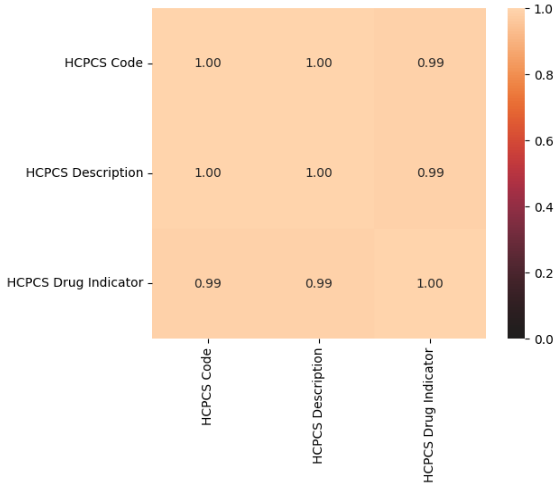
Here we have plotted a heatmap of correlation using seaborn using dython library (plots the association(correlation) between features) and if some of the features are highly associated (correlated) to each other so only one of them is kept because that one single column will define those columns as well.

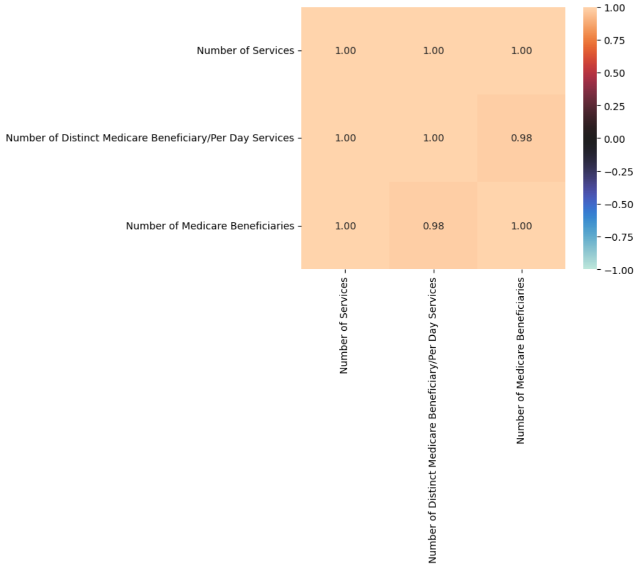


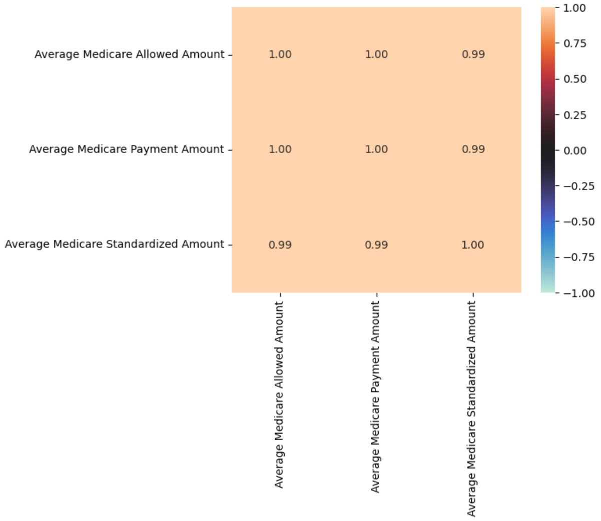
Following is the correlation of some of the highly correlated features –
Entity Type of the Provider -> Provider Type and correlation = 0.9978872254160679
HCPCS Code -> HCPCS Description and correlation = 0.99909744497O5192
HCPCS Code -> HCPCS Drug Indicator and correlation = 0.9867671833768509
HCPCS Description -> HCPCS Drug Indicator and correlation = 0.9874849263612132
Number of Services -> Number of Medicare Beneficiaries and correlation = 0.9982778502073709
Number of Services -> Number of Distinct Medicare Beneficiary/Per Day Services and correlation = 0.9991333301674521
Number of Medicare Beneficiaries -> Number of Distinct Medicare Beneficiary/Per Day Services and correlation = 0.9810718190199162

Average Medicare Allowed Amount -> Average Medicare Payment Amount and correlation = 0.99870397489615393
Average Medicare Allowed Amount -> Average Medicare Standardized Amount and correlation = 0.9948313068806548
Average Medicare Payment Amount -> Average Medicare Standardized Amount and correlation = 0.994648750972444

So we will drop the features – Entity Type of the Provider,HCPCS Code','HCPCS Description,'Number of Services' , 'Number of Distinct Medicare Beneficiary/Per Day Services','Average Medicare Allowed Amount', 'Average Medicare Payment Amount' , "Zip Code of the Provider",'City of the Provider'.
For better visualization of these correlated features –



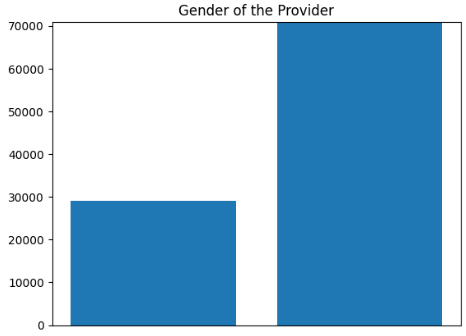


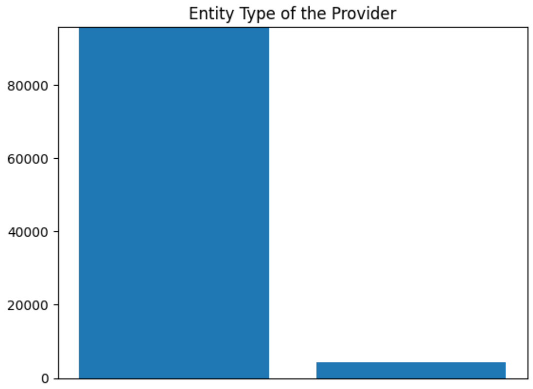


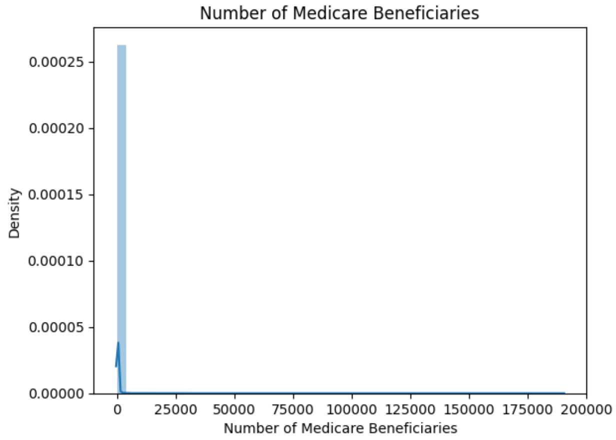
Also the features "Zip Code of the Provided" has good correlation with "State Code of the Provider" and "City of the Provider" but not good with "Country Code of the Provider" but "State

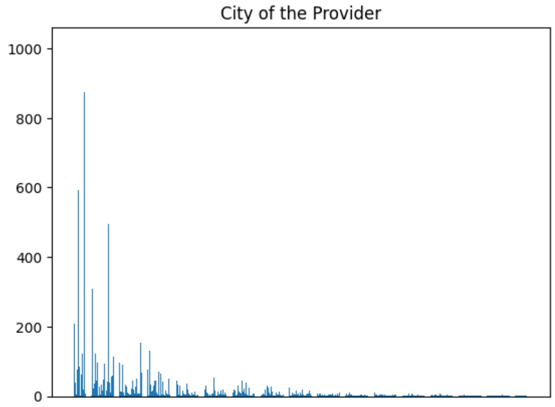
Code of the Provider" and "City of the Provider" have good with Country Code of the Provider from these we can drop "Zip Code of the Provider" and "City of the Provider"

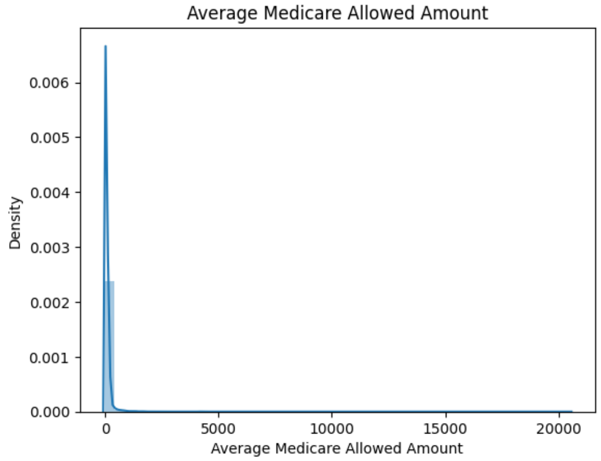
VISUALISING THE DATA:

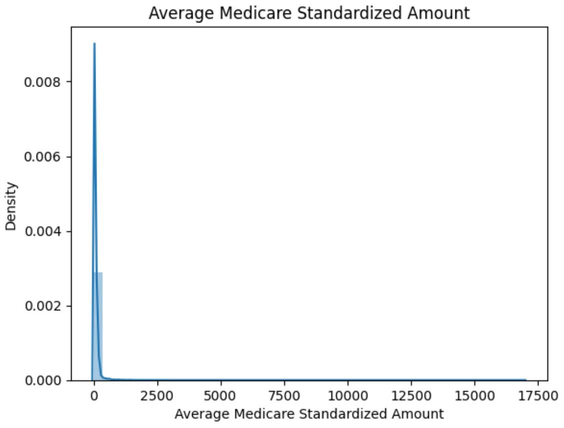












ENCODING THE FEATURES AND STANDARDIZING THE CONTINUOUS FEATURES

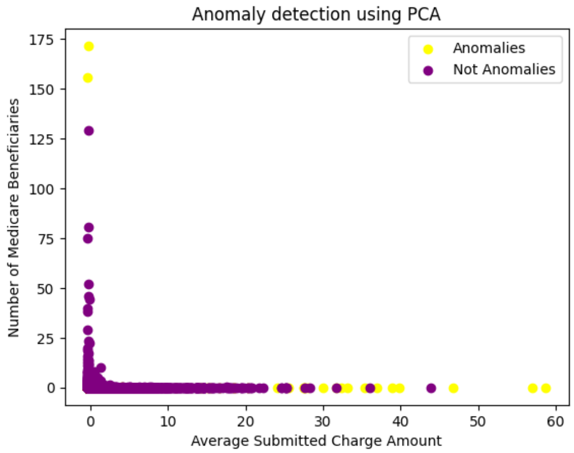
Now we have standardized the continuous features and encoded the categorical and nominal columns and finally performed one hot encoding on those features. For normal encoding for some of the columns we have a large amount of categorical data so we have taken a threshold (by observing the data) and then if count of any value in that column is less than threshold then encoded as 0 and named as 'others' and encoding the remaining features as we do. Finally after all this we have performed one hot encoding and we have 60 columns in our data.

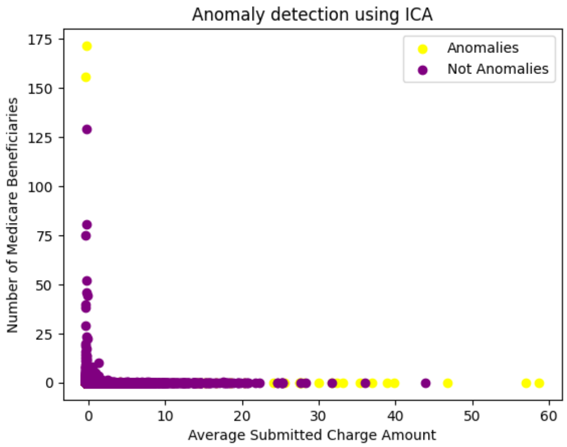
ANOMALY DETECTION USING DIFFERENT METHODS:

We have used different methods to identify the anomalies in the data, basically the outliers in data and finally did voting for the anomalies found for different methods and if most of the methods told it an anomaly then it is considered an anomaly (fraud provider).

PCA (Principal Component Analysis) and ICA (Independent Component Analysis)

In both methods pca and ica we follow the same technique to find the anomalies.
To detect anomalies, the reduced data is reconstructed from its lower-dimensional representation using the inverse transformation of PCA or ICA. Then, the error between the original data and its reconstruction is calculated, and any data points with an error above a predefined threshold are flagged as anomalies. The threshold here we have taken is 95 percentile (considering that there may be 5% of the data with anomalies) .
Following is the plot we obtained for two features of dataset (5th and 6th i.e. Number of Medicare Beneficiaries and Average Submitted Charge Amount).

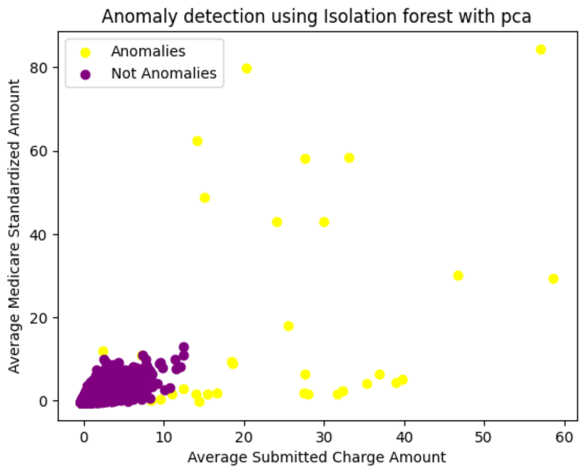


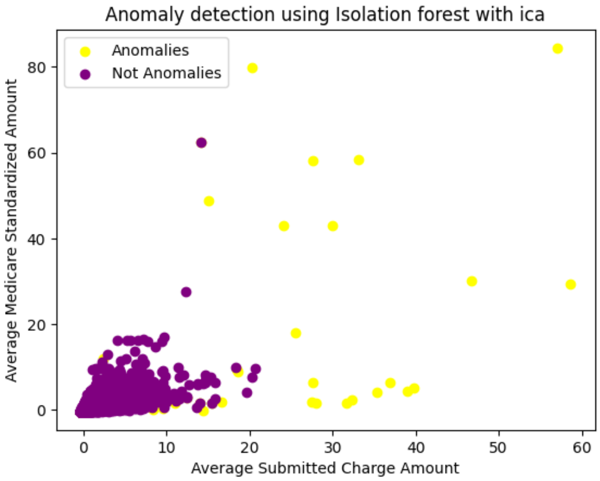


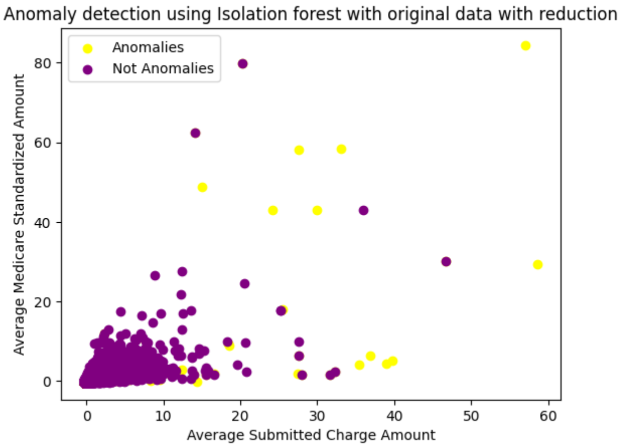
ANOMALY DETECTION USING ISOLATION FOREST

Here, an isolation forest has been employed to find anomalies in the data set. Its foundation is the idea of employing binary tree topologies to isolate abnormalities. To divide the data into two

sub-samples, the algorithm randomly chooses a feature and a random split value for that feature. Until the samples are divided into manageable groups of anomalies that are segregated in the tree structure, this process is repeated recursively. The Isolation Forest algorithm is based on the idea that because anomalies are uncommon, they can be isolated with fewer splits than regular data points. Each data point receives an anomaly score from the algorithm based on how many splits are necessary to isolate it in the tree structure. The anomaly score will be lower for anomalies than for regular data items.
The same method is applied on the dataset three times first reducing the data by pca , then reducing by ica using the same number of components as obtained in previous and third with the original data.

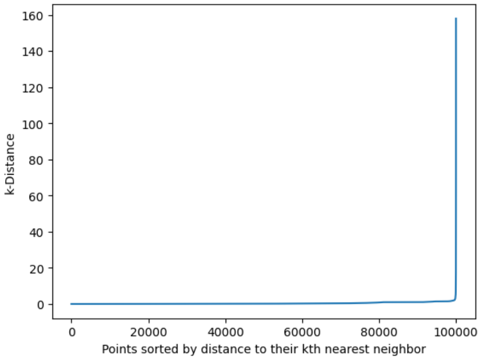


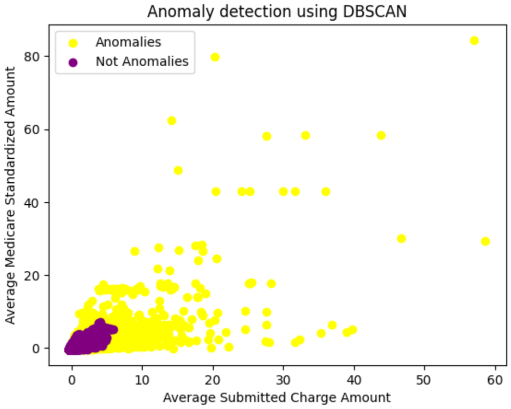




ANOMALY DETECTION USING DBSCAN

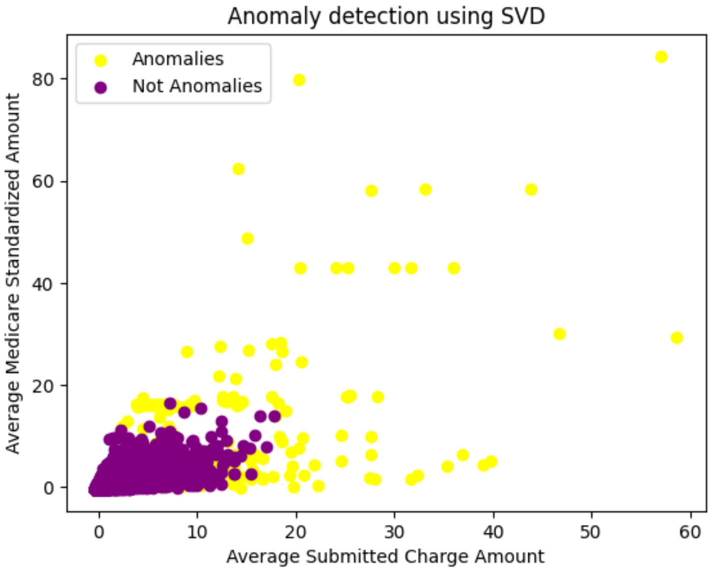
Here we have used dbscan to find the anomalies in the dataset. It works by grouping data points that are closely packed together in high-density regions and flagging data points that are in low-density regions as anomalies. We have plotted the k distance graph using nearest neighbors by taking the value of k = 20 (n_neighbors) and from there it looks like the value where the distance value changes is very close to zero so we have taken epsilon = 0.5 and then predicted the anomalies in the dataset.





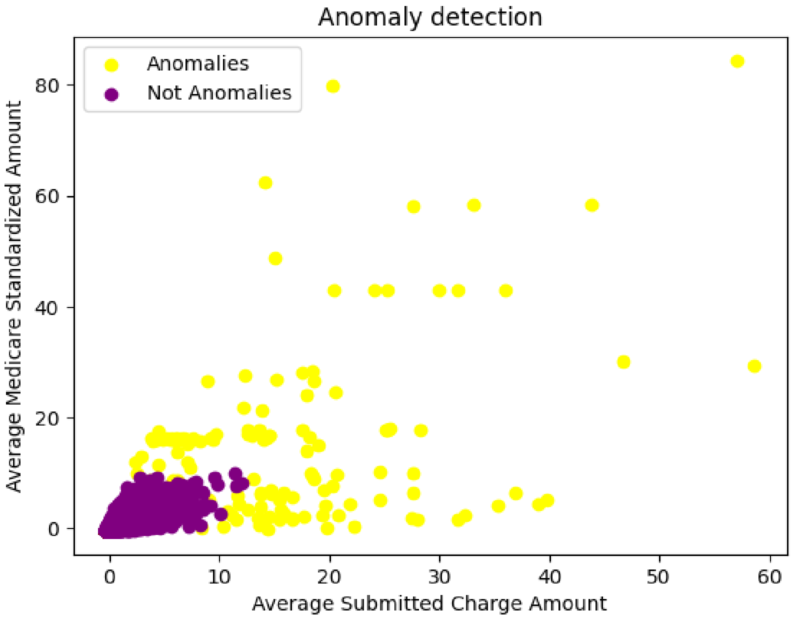
ANOMALY DETECTION USING SVD (Singular Value Decomposition)

SVD is a matrix factorization method that decomposes a matrix into three matrices: a left singular matrix, a diagonal matrix, and a right singular matrix. SVD is commonly used for dimensionality reduction and data compression, but it can also be used for anomaly detection by reconstructing the original matrix and comparing it to the original data points. Reconstruct the original matrix by multiplying the reduced left singular matrix, the reduced diagonal matrix, and the reduced right singular matrix.Then calculate the error between the original data points and the reconstructed matrix. The error can be calculated using metrics such as the mean absolute error or the mean squared error. Then take the threshold of 95 percentile and then for the error greater than that threshold are considered as anomalies.



ANOMALIES IN THE DATA

From the above methods finally we chose a data point as anomaly if 2 or more methods declared it as anomaly .
Anomalies ::



Prediction for new testing data:

Now we have converted our data to supervised and training on voting classifier with random forest , gaussian naive bayes , svm classifier and logistic regression to get the output whether the testing data is an anomaly (fraud or not) which has been deployed in our website.