3.Architecture Description

3.1 Data Description

The dataset contains transactions made by credit cards in September 2013 by European cardholders.

The dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions . Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset.

The feature 'Amount' is the transaction Amount.

3.2 Data **Pre-processing**

* **Performing EDA to get insight of data like identifying distribution, statistics , outlier among data etc.**
* **Check for null values in the columns. If present impute the null values.**
* **Visualizing scatter plot of independent variables with each other and output variables.**
* **Perform Standard Scalar to scale down the values.**
* **Perform PCA**
* Data cleaning: this step involves identifying and removing missing, inconsistent, or irrelevant data. This can include removing duplicate records, filling in missing values, and handling outliers.
* Data integration: this step involves combining data from multiple sources, such as databases, spreadsheets, and text files. The goal of integration is to create a single, consistent view of the data.
* Data transformation: this step involves converting the data into a format that is more suitable for the data mining task. This can include normalizing numerical data, creating dummy variables, and encoding categorical data.
* Data reduction: this step is used to select a subset of the data that is relevant to the data mining task. This can include feature selection (selecting a subset of the variables) or feature extraction (extracting new variables from the data).
* Steps Involved in Data Pre-processing:

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a). Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

Ignore the tuples:

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

Fill the Missing values:

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b). Noisy Data:

Noisy data is a meaningless data that can’t be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

* Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

* Regression:

Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

* Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

3.3 Data Modelling And Training:-

Created pipeline:- Created pipeline for every model (Logistic regression, Gradient boosting classifier, Random forest classifier, etc.) We splitted our datset in training and testing set.

Before diving the data in training and validation set we performed preprocessing and featuring scaling.The scaling was performed over training and validation data

We splitted our datset in training and testing set.

Algorithms like K-mean, XGBoost, Random Forest etc were used and select our final model and we saved that as Model.pkl.

Model selection :- We find the best model for each pipeline. For each pipeline, we used hyper tuned algorithms. After calculating the AUC scores for all pipeline models

3.4 Best Fit Model

Algorithms like K-mean, XGBoost, Random Forest etc were used and select our final model. We have found “Random forest” best fit for our data.

3.5 Saving Model:-

we saved our model as Model.pkl.

3.6 Front-end Website For project :-

* Created “index.html”
* Created Form for our features (independent variables) and used “Post” method.
* We feed all that data to our saved best fit model ( Random forest model).
* Then if our output is “0” then our transaction is not a fraud transaction otherwise it is a fraud transaction.
* Display this result on screen.

3.7 Form Data from the website:-

3.8 Fed Data To ML model