***Customer Personality Analysis***

**Objective:**

Customer personality analysis is essential for personalizing marketing, developing products that meet diverse needs, enhancing customer experiences, and creating accurate segments. It helps in identifying churn risks and developing retention strategies, building brand loyalty through strong emotional connections, and gaining a competitive advantage with relevant offerings. For Nasher Miles, this analysis can improve marketing efforts, product development, and customer service, resulting in higher customer satisfaction and loyalty.

**Benefits**

Customer personality analysis enhances personalization by tailoring marketing messages and product recommendations, improving product development to meet unique needs, and providing personalized customer experiences. It aids in accurate customer segmentation, increases retention by identifying churn risks, and builds stronger brand loyalty through aligned values and messaging. This analysis offers a competitive advantage with relevant offerings, optimizes resource allocation, and provides deep insights into customer behaviors, guiding strategic decisions for better overall performance.

**Data Sharing Agreement :**

Sample file name

Length of date stamp(8 digits)

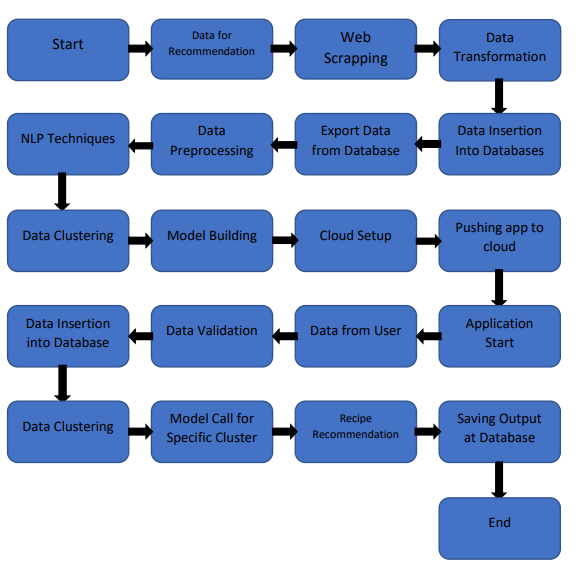
Length of time stamp(6 digits)

Number of Columns

Column names

Column data type

**Architecture**



**Data Validation and Data Transformation :**

Name Validation - Validation of files name as per the DSA. We have created a

regex pattern for validation. After it checks for date format and time format if

these requirements are satisfied, we move such files to "Good Data\_Folder" else

"Bad Data\_Folder “

Number of Columns – Validation of number of columns present in the files, and if

it doesn't match then the file is moved to "Bad Data Folder “

Name of Columns - The name of the columns is validated and should be the same

as given in the schema file. If not, then the file is moved to "Bad Data\_Folder".

Data type of columns - The data type of columns is given in the schema file. It is

validated when we insert the files into Database. If the datatype is wrong, then the

file is moved to "Bad Data\_Folder".

Null values in columns - If any of the columns in a file have all the values as

NULL or missing, we discard such a file and move it to "Bad Data\_Folder".

**Data Insertion in Database:**

Table creation :- Table name “t motorpv fraud" is created in the database for

inserting the files. If the table is already present then new files are inserted in the

same table.

Insertion of files in the table - All the files in the "Good Data\_Folder" are inserted

in the above-created table. If any file has invalid data type in any of the columns,

the file is not loaded in the table

**Model Training:**

Data Export from Db :

The accumulated data from db is exported in csv format for model training

Data Preprocessing

Performing EDA to get insight of data like identifying distribution , outliers ,trend

among data etc.

Check for null values in the columns. If present impute the null values.

Encode the categorical values with numeric values.

Perform Standard Scalar to scale down the values.

**Clustering –**

KMeans algorithm is used to create clusters in the preprocessed data. The

optimum number of clusters is selected by plotting the elbow plot, and for the

dynamic selection of the number of clusters, we are using KneeLocator

function. The idea behind clustering is to implement different algorithms on

the structured data

The Kmeans model is trained over preprocessed data, and the model is saved

for further use in prediction

**Model Selection** –

After the clusters are created, we find the best model for each cluster. By using 2

algorithms “SVM” and "XGBoost". For each cluster both the hyper tunned

algorithms are used. We calculate the AUC scores for both models and select the

model with the best score. Similarly, the model is selected for each cluster. All

the models for every cluster are saved for use in prediction

**Prediction:**

The testing files are shared in the batches and we perform the same Validation

operations ,data transformation and data insertion on them.

The accumulated data from db is exported in csv format for prediction

We perform data pre-processing techniques on it.

KMeans model created during training is loaded and clusters for the preprocessed

data is predicted

Based on the cluster number respective model is loaded and is used to predict the

data for that cluster.

Once the prediction is done for all the clusters. The predictions are saved in csv

format and shared.