Fraud Transaction Detection

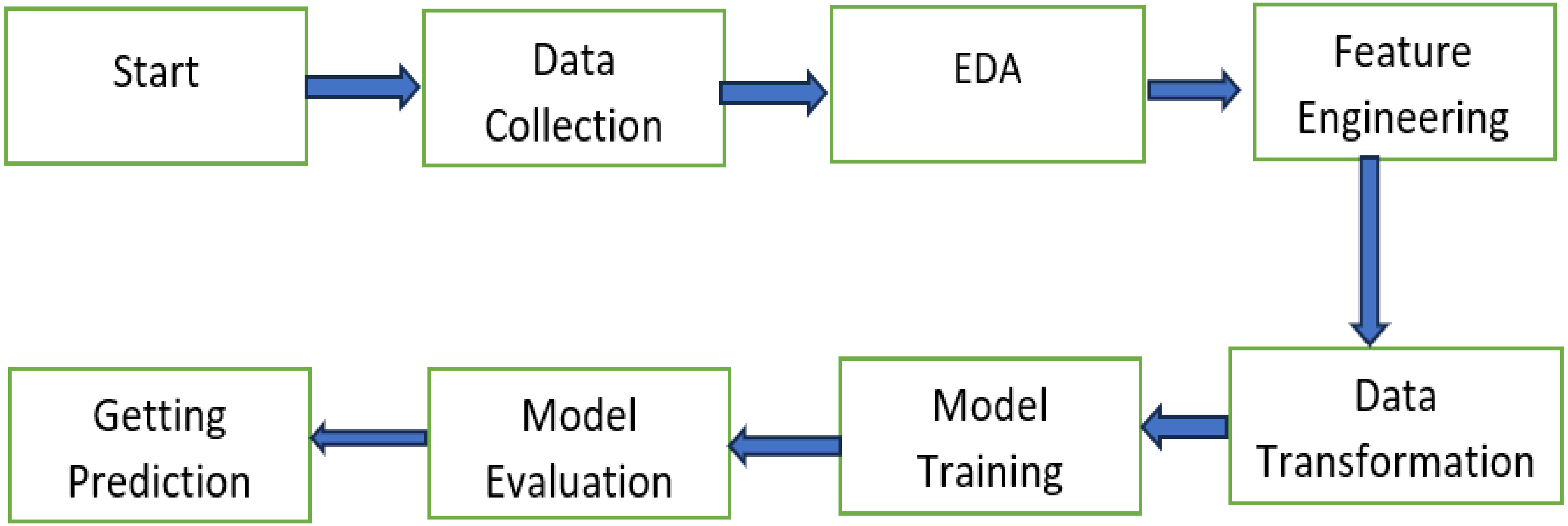
Objective:

Development of a predictive model for predicting fraud transaction. The model will determine whether a customer’s transaction is placing a fraudulent or not.

Benefits:

- Detection of upcoming frauds.
- Gives better insight of customers base.
- Helps in easy flow for managing resources.
- Manual inspection if fraud is identified .

Architecture



Start

Data Collection

EDA

Feature Engineering

Data Transformation

Model Training

Model Evaluation

Getting Prediction

Data Transformation :

- Prepare the data for training the fraud detection model.
- Load training and testing datasets from CSV files.
- Examine the structure and format of the input data.
- Set up pipelines for numerical and categorical feature transformations.
- Use imputation strategies for missing data in both numerical and categorical features.
- Apply StandardScaler and RobustScaler to normalize numerical features.
- Convert categorical variables into numerical format using one-hot encoding.
- Implement a ColumnTransformer to manage multiple feature transformations.
- Divide data into dependent and independent features.
- Save the preprocessing object for future use in the artifacts folder.

Model Training:

- Train machine learning models for fraud transaction detection.
- Split the input data into training and testing sets.
- Choose from multiple classifiers, including Logistic Regression and Decision Tree Classifier
- Utilize hyperparameter tuning with predefined parameter distributions for each classifier.
- Evaluate each model's performance using cross-validation scores.
- Identify the best-performing model based on accuracy scores.
- Display a comprehensive report containing accuracy scores for each model
- Save the best model (highest accuracy) in the artifacts folder for future predictions.
- Log information about the best model, including its name and accuracy score.
- Implement error handling to manage exceptions during the model training process.

- Model Selection –

After the model training are completed, we find the best model and save the pickle file as best_model.pkl file. For training 3 algorithms “Logistic Regression” ,”Decision Tree” and “Random Forest“ algorithms are used. For each algorithm both the hyper tunned algorithms are used. We calculate the accuracy_score for all models and select the model with the best score.

Prediction:

- The accumulated data from artifacts is exported in csv format for prediction
- We perform data pre-processing techniques on it.
- Decision Tree model created during training is loaded for the preprocessed data for prediction
- Based on the cluster number respective model is loaded and is used to predict the data for that cluster.
- Once the prediction is done then the predictions are saved in csv format and shared.

Q & A:

Q1) What’s the source of data?

Data was available in csv file which was provided by iNeuron.

Q 2) What was the type of data?

The data was the combination of numerical and Categorical values.

Q 3) How logs are managed?

We are using logs as per the steps that we follow the flow of model training and prediction. For that purpose , we used logging.error and logging.info to get more relevant and detailed information regarding the flow of execution.

Q 4) What techniques were you using for data pre-processing?

- Removing unwanted attributes
- Visualizing relation of independent variables with each other and output variables
- Checking and changing Distribution of continuous values
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.
- Scaling the data

Q 5) How training was done or what models were used?

- Before diving the data in training and validation set we performed clustering over fit to divide the data into clusters.
- As per cluster the training and validation data were divided.
- The scaling was performed over training and validation data
- Algorithms like logistic regression , decision tree and random forest were used we saved the best performing model

Q 6) How Prediction was done?

We have testing file in artifacts folder. We Perform the same life cycle till the data is transformation .Then using the best_model.pkl file prediction performed. In the end we get the accumulated data of predictions as csv file.