



Fraud Transaction

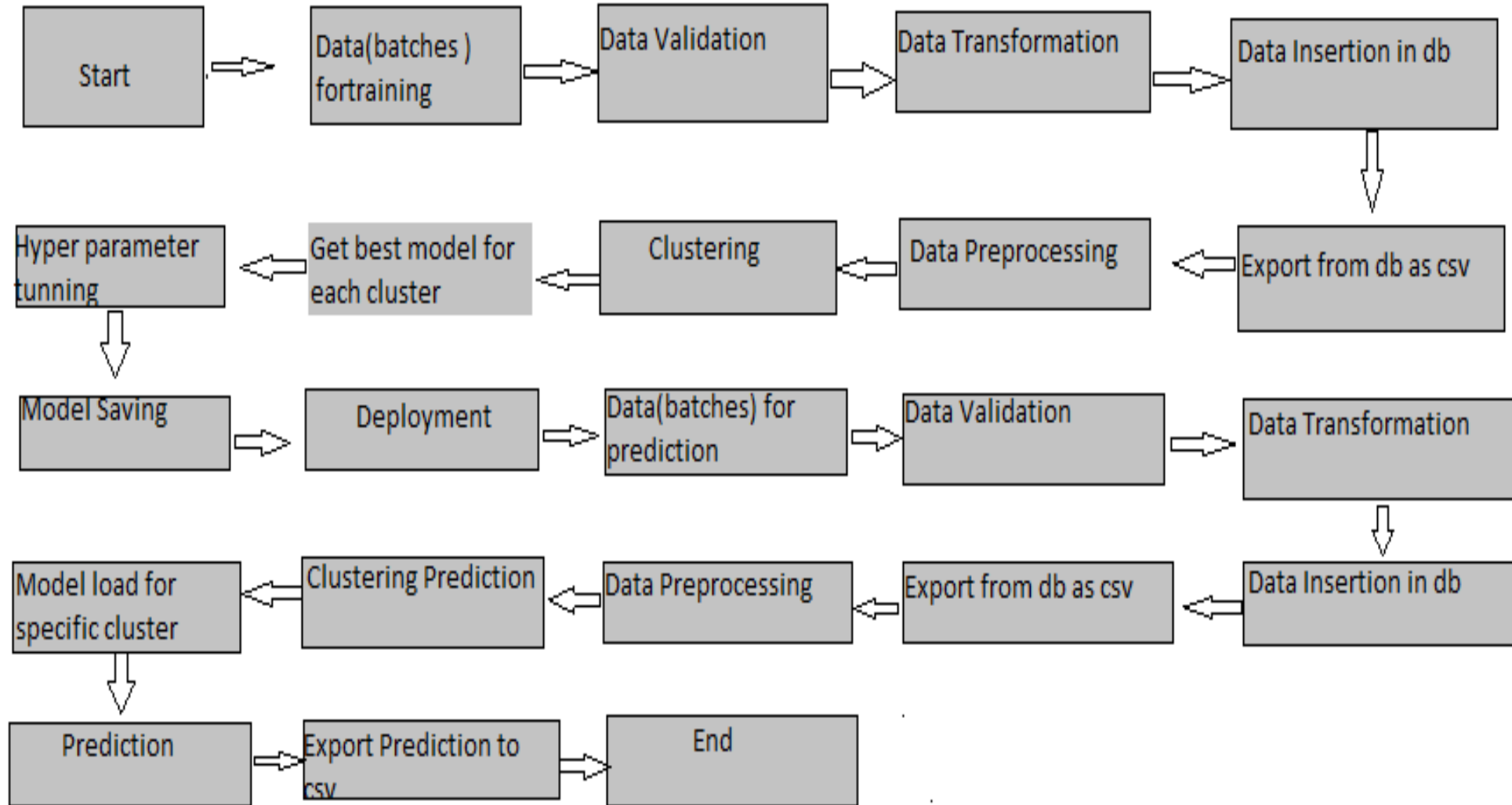


Objective:

Development of a predictive model for monitoring fraud transaction. The model will determine whether a customer is placing a fraudulent transaction 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 .



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 “colab_file_Folder”
- ❑ Number of Columns – Validation of number of columns present in the files
- ❑ Name of Columns - The name of the columns is validated and should be the same as given in the schema file
- ❑ 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.

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 among data etc.
- Check for null values in the columns. If present impute the null values.
- Perform Standard Scalar to scale down the values.
- Perform PCA .

□ **Pipeline—**

- Created pipeline for every model

□ **Model Selection –**

After the pipelines are created, we find the best model for each cluster. By using pipeline (algorithms “Random Forest” ,”Xgboost” etc.). For each pipeline the hyper tuned algorithms are used. We calculate the AUC scores for all pipeline models and select the model with the best score. Similarly, the model is selected for each pipeline. We select random forest with the best accuracy and did 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.
- ❑ Random Forest model created during training is loaded and hyper tuned for the preprocessed data predicted
- ❑ Based on the respective model was loaded and was used to predict the data for that model.
- ❑ Once the prediction is done for all the Random Forest. The predictions are saved in csv format and shared.

Q & A:

Q1) What's the source of data?

The data for training is provided by the client in multiple batches and each batch contain multiple files

Q 2) What was the type of data?

The data was the contain only of numerical values.

Q 3) What's the complete flow you followed in this Project?

Refer slide 5th for better Understanding

Q 4) After the File validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client.

Stage

Tasks

Requirements Analysis

Business Analysis

Business Requirements
Business Constraints

Technical Requirements

Use Cases
Usage Analysis
Quality of Service Requirements

Logical Design

Logical Architecture
Deployment Scenario

Deployment

Deployment Design

Sizing
Deployment Architecture
Implementation Design

Implementation

Hardware Build-out
Installation
System Configuration
Customization and Development
Testing
Production Rollout

Operations

System Maintenance

Monitoring
Performance Tuning
System Enhancements/Upgrades
Data Replication and Backup