knowledgehut upGrad

CREDIT CARD FRAUD PREDICTION PROJECT DOCUMENT (BY HARIDOSS ANNAVI)

TABLE OF CONTENTS

knowledgehut upGrad

Overview and Requirement

This project aims to build a machine learning model to predict credit card faults, such as defaults or fraudulent transactions. The goal is to assist financial institutions in identifying high-risk transactions or customers, thereby reducing losses and enhancing customer security.

Please refer to the below links for detailed requirements:

https://github.com/haridoss-annavi/HA-Capstone-project/blob/Credit-Card-Fraud-Prediction/Requirement

Key Components

1. DATA SOURCE: The dataset (creditcard.csv) containing transaction data, including fraudulent and non-fraudulent transactions.
2. DATA INGESTION & LOADING: The CSV file is read into a Pandas DataFrame. Data quality checks, including missing values and data types, are performed.
3. EXPLORATORY DATA ANALYSIS (EDA): Visual and statistical analysis to understand patterns, correlations, and data distribution. Key outputs include class distribution, outlier detection, and correlation matrices.
4. DATA PREPROCESSING:

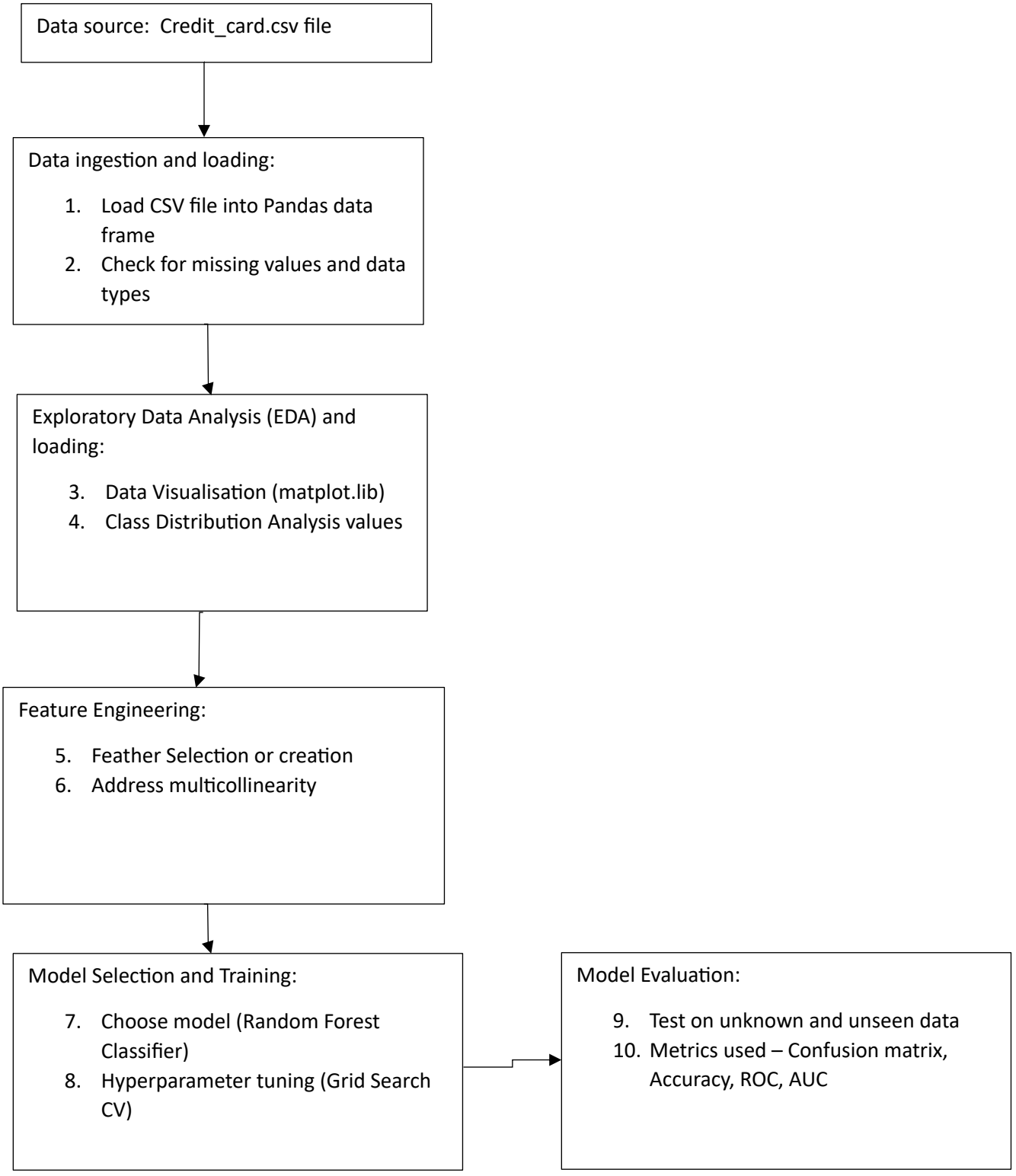
   o FEATURE SCALING: Standardize features using Standard Scaler.

   o IMBALANCED DATA HANDLING: Use SMOTE to generate synthetic samples for the minority class.

   o TRAIN/TEST SPLIT: Split the data into training and testing sets.
5. FEATURE ENGINEERING: Select or create new features and address issues like multicollinearity to improve model performance.
6. MODEL SELECTION & TRAINING: o Select a machine learning model (Random Forest). o Perform hyperparameter tuning using GridSearchCV. o Train the model using the balanced dataset.
7. MODEL EVALUATION: Evaluate the model using metrics like accuracy, confusion matrix, and ROC AUC to ensure the model's robustness.
8. MODEL DEPLOYMENT PLAN: Outline steps to deploy the model, including containerization, serving via an API, and integration into the production environment.

knowledgehut upGrad

Solution Design



knowledgehut upGrad

Source code and Data Visualisation

Please refer to the GitHub link for Project source code and Visuals (charts indicating the ML Analytics)

https://github.com/haridoss-annavi/HA-Capstone-project/blob/Credit-Card-Fraud-Prediction/ha-credit-card-fraud-prediction-rf-smote.ipynb

Installation and Deployment plan

Please refer to the GitHub link for README.md file

https://github.com/haridoss-annavi/HA-Capstone-project/blob/Credit-Card-Fraud-Prediction/README.md

Model Training and Results

The project uses the following machine learning models to predict credit card faults:

• K-Nearest Neighbours (KNN)

• Logistic Regression

• Random Forest Classifier

The best-performing model is chosen based on the accuracy, precision, recall, and F1 score.

Results • The final model achieved an accuracy of the highest values of Normal transactions are 284315, while of Fraudulent transactions are just 492.

The average value of normal transactions is small(USD 88.29) than fraudulent transactions that is USD 122.21

Best score:

SMOTE (OverSampling) = RandomForest =

Accuracy: 0.9995611109160493

Precision: 0.9041095890410958

Recall: 0.7857142857142857

F2: 0.806845965770171

This is a considerably difference by the second-best model that is 0.8252 that uses just RandomForests with some Hyper Parameters.

Worst Score:

Logistic Regression with GridSearchCV to get the Best params to fit and predict where the recall = 66.67% and f2 = 70%.