IT1244 Team 21 Project

Introduction

Welcome to the IT1244 Team 21 Project. In this project, we performed machine learning on the Fraud Electricity Gas Consumption dataset. The notebooks can be found in the "Code" folder, and the models that we have selected are K-Nearest Neighbors (KNN), Linear Regression, Multi-Layer Perceptron (MLP) and XGBoost.

File structure

```
IT1244_Team21_Project
├── Code
│   ├── KNN.ipynb
│   ├── Linear.ipnyb
│   ├── XGBoost.ipynb
│   ├── MLP.ipynb
│   ├── client.csv (given dataset to be placed here)
│   ├── invoice.csv (given dataset to be placed here)
│   └── Outputs (Contains all our outputs [more information below])
├── Team21 Project Report
└── Readme.pdf
```

Required packages

Before running any of the notebooks, please ensure that you have the necessary Python packages installed.

1. `numpy` - A powerful library for numerical operations and mathematical functions.
2. `pandas` - Used for data manipulation and analysis.
3. `matplotlib` - A data visualization library for creating charts and plots.
4. `scikit-learn` - Provides various machine learning functionalities for classification, regression, and more.
5. `imbalanced-learn` - Used to handle imbalanced datasets in machine learning tasks.
6. `xgboost` - A popular gradient boosting library for building robust machine learning models.
7. `torch` - PyTorch, as used in the Multi-Layer Perceptron (MLP) and Linear Regression notebook.

You can install these packages using the following pip command:

```
pip install numpy pandas matplotlib scikit-learn imbalanced-learn xgboost torch
```

or

```
pip3 install numpy pandas matplotlib scikit-learn imbalanced-learn xgboost torch
```

Instructions for Running the Notebooks

1. **First and foremost, ensure that the dataset is located in the same folder as the Jupyter notebooks (ipynb files) within the "Code" folder.**
2. Select a python kernel (preferably python 3.9 - 3.10).
3. Ensure that all required packages are installed.
4. For each notebook, press `Run All` and wait. It should take a few minutes per file, with the pre-trained hyperparameters already selected. Do note that the code for the GridSearch is commented out, and for your reference.
5. There is no need delete the new files that have been generated.

Outputs

It contains our outputs in the form of graphs and txt files for our statistics from the the models paired with the different sampling methods [Random Oversampling, SMOTE, Vanilla]. For the machine learning models such as Linear and Multi-Layer Perceptron, the folder name refers to the hyperparameters. i.e

```
(learning_rate, batch_size, num_epochs)
```

The folder name with (0.001,50,100) means that `learning_rate = 0.001, batch_size = 50, num_epochs = 100`.

The data consists of our Confusion Matrix, Metrics.txt and the ROC Curve.

We have a PyTorch file in the folder `(0.001,50,100)/Vanilla` for both `Linear Outputs` and `MLP Outputs` to run if required. It needs to be placed in the same directory as the python notebooks (in `Code`).