IA Assignment N° 2
Supervised Learning

Bank Account Fraud Dataset Suite

3LEIC – G33
António Ferreira – up202004735
Hugo Gomes – up202004343
João Moreira – up202003550

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Problem Specification

In the sector of banking, the use of machine learning models and algorithms linked to supervised learning is becoming increasingly significant.

Banking is now simpler than ever thanks to technological advancements, but the number of fraudulent activity cases has risen as well.

For this project, our main goal is to investigate several strategies for identifying trends and forecasting future fraudulent activity. To accomplish this task, we’ll use a data set, made up of 32 attributes, with information on bank account opening applications.

Related Work

Documentation:

- https://pandas.pydata.org/docs/reference/
- https://scikit-learn.org/stable/
- https://seaborn.pydata.org/generated/seaborn.pairplot.html

Dataset:

- https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022?select=Variant+II.csv
- https://github.com/feedzai/bank-account-fraud/tree/main

Other:

- https://ruslanmv.com/blog/The-best-binary-Machine-Learning-Model
- https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/#Impute_Replace_Missing_Values_with_Mean

Tools & Algorithms Developed

Regarding the tools/libraries that are going to be used, we opted to highlight the following ones:

- **pandas** – essential to perform the main data analytical tasks for this project
- **NumPy** – provides support for a wide range of mathematical functions
- **Scikit-learn** - offers a variety of supervised and unsupervised learning algorithms
- **Seaborn** - provides a higher-level interface for creating informative and visually appealing statistical graphics

Support Vector Machines

SVMs are commonly used for binary classification problems and can be useful when the dataset has a **clear boundary** between the two classes. SVMs can also be effective in handling **high-dimensional data**.

K-Nearest Neighbours

K-NN is a **non-parametric** algorithm that can be used for both regression and classification problems. It works by finding the K closest data points (based on some distance metric) to a new data point, and then using the **majority vote** of the **K-nearest neighbours** to classify the new data point.

Decision Trees

Decision Trees are also a commonly used supervised learning algorithm for fraud detection in banking applications. Decision Trees are a type of **tree-based** algorithm that can handle both **categorical** and **numerical** features, making them well-suited for datasets with **mixed data types**.

Neural Networks

Neural networks are a versatile and powerful class of models that can be used for a wide range of supervised learning problems. They can handle **non-linear relationships** between features and the target variable, and can work well with **large, complex datasets**.

Logistic Regression

This is a simple and widely used algorithm that can be used for **binary classification** problems, such as identifying fraudulent vs non-fraudulent transactions.

Implemented Work

Data Pre-processing

After analyzing the given data set, we've come to realize that, taking into consideration the project's context, there were attributes with faulty data. Consequently, we removed the rows with errors and replaced the missing values with the mean of the respective column. Finally, we opted to convert the categorical data into numerical data in order to implement the previous mentioned algorithms.

Implemented algorithms

In the first place, we divided the data set into two parts: training set consisted by 75% of the data and the testing set with the remaining 25%. On top of that, the decision tree, Logistic Regression and a neural network (MLPClassifier) are already implemented.