Credit Card Fraud

Project II Checkpoint I - Artificial Inteligence

Grupo 04_1A

up201906086@up.pt Marcelo Henriques Couto

up201907361@up.pt Francisco Pinto de Oliveira

Faculdade de Engenharia da Universidade do Porto

May 30, 2022

Problem Specification

Credit card fraud is a crime easily stopable if detected. However, this detection must be quick to occur, as transactions are supposed to be fast. The **objective** of this project is to **develop a machine learning model** based on supervised learning classification algorithms able to distinguish, based on some input data, if a given transaction related to a bank account is fraudulent or not.

In order to come up with the ideal model, we must experiment with at least three different classification algorithms, as well as carry out an **Exploratory Data Analasys**.

Problem Specification

The dataset contains the following data:

- **Distance from home** - continuous
- **Distance from last transaction** - continuous
- **Ratio to median purchase price:** Ratio of the value of the transaction to the median transaction value - continuous
- **Repeat retailer:** Whether the retailer where the transaction happened had other transactions registered for the same person
- **Used chip:** Transcation via credit card - discrete, binary
- **Used pin number** - discrete, binary
- **Online order** - discrete, binary
- **Fraud** - discrete, binary - **label**

We considered that all distances are expressed in kilometers.

Related Work

Code

- Imbalanced Learn - over sampling and undersampling tools
- Scikit Learn - machine learning algorithms

Websites

- https://machinelearningmastery.com/what-is-imbalanced-classification/
- https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

Articles

- Bekkar, M., Djemaa, H. K., Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. J Inf Eng Appl, 3(10).
- Tharwat, A. (2021), "Classification assessment methods", Applied Computing and Informatics, Vol. 17 No. 1, pp. 168-192.

Tools and Algorithms

Tools

We will use **Python** as programming language, programming in a **Jupyter Notebook** environment. For machine learning algorithms, we will resource to **Scikit-Learn** and **Imbalanced Learn** libraries, as well as **Pandas** to read and handle the data and **Seaborn and Matplotlib** to visualize it.

Algorithms

Because our dataset is imbalanced, we plan to explore **Synthetic Minority Oversampling** and **Undersampling** techniques to handle this issue.

For the classification we plan on using **Logistic Regression**, **Random Forest**, **Decision Trees**, **Support Vector Machine** and possibly others.

Implementation Already Carried Out

Language: Python - Anaconda

Environment: Visual Studio Code using JupyterNotebooks

Data Structures: The dataset is represented as a pandas.DataFrame **Data**

Preprocessing and **Exploratory Data Analysis**:

- Analysis of dataset's validity and integrity
- Analysis of dataset balance
- Analysis of outliers, missing values and other errors
- Analysis of correlations between the multiple variables

Model Training

- Logistic Regression with cross validation using Area Under ROC Curve as evaluation measure