Graduation Project – 2022

PCS – Department of Computer Engineering and Digital Systems

Computer Engineering

Subject:

Machine Learning Applied to Card Payment Fraud Detection

Introduction

Despite of being a global trend because of its many facilities and being leveraged by e-commerce, payments made by bank card are not completely secure. A 2019 study by NilsonReport [1] estimates that in the next 10 years, $408.5 billion will be due to bank card payment frauds.

Historically, anti-fraud systems were based on a pre-programmed set of rules that highlights a payment as fradulent, but with online shopping, fraudsters have much more frexibility, making these single ruleset systems weak to detect frauds.

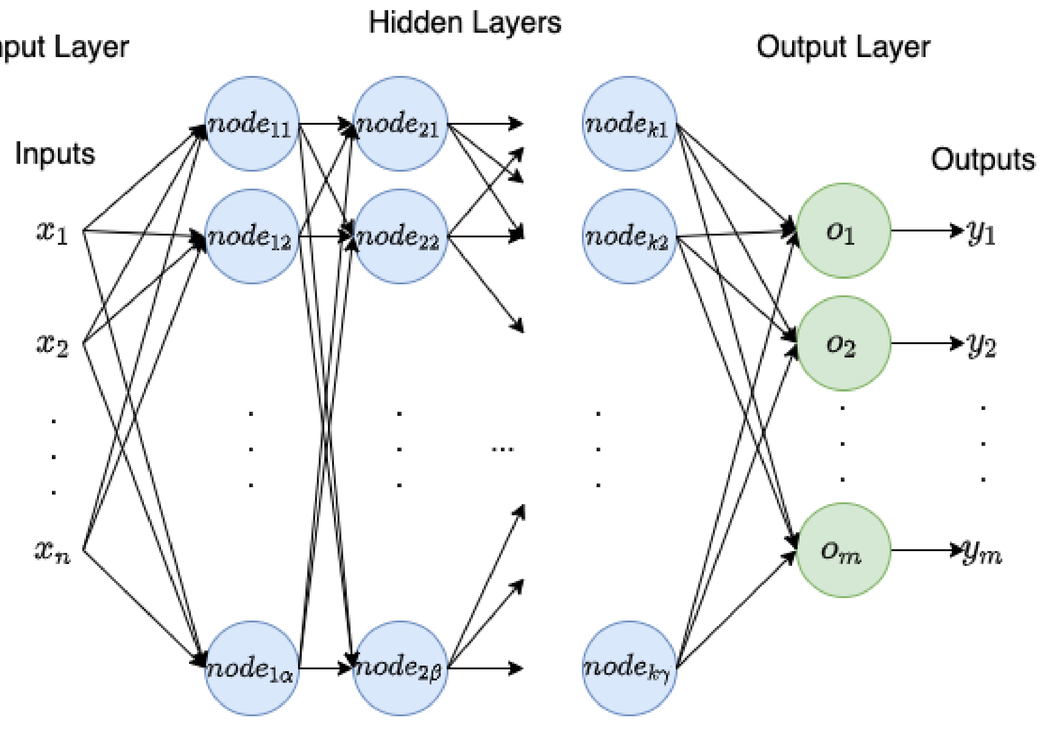
In contrast, the advancement in computational processing in recente decades allowed technologies such as Machine Learning to enter the domain of bank card fraud detection. In this type of system, a ML model analyzes historical data and learns the main fraud patterns from it.

Objective

Based on [2], [3] and [4], compare whether recent findings in Deep Learning (DL) for tabular data can outperform Gradient Boosted Decision Trees (GBDT), the state-of-art in this domain, considering a highly imbalanced tabular dataset.

Moreover, it will propose an optimized model with significant efficiency in detecting fradulent card payments.

Figure 1: The Multilayer Perceptron General Architecture



Metodology

In a GPU-enabled environment and using a highly imbalanced tabular dataset with millions of card payment transactions labeled as fradulent or not, a whole data pipeline was developed from scratch to train, validate, optimize, test and compare two GBDT and four DL models.

Furthermore, several techniques, such as over sampling and adapted loss function, were discussed and used to compensate the data imbalancing.

Results

The results showed that the GBDT models outperformed so far the DL ones in all three metrics: performance by F1 score, training time and ease of code implementation.

And after several optimization steps and techniques, the XGBoost model ended up performing greatly on the considered dataset, keeping the number of false positives low and increasing significantly the number of true positives.

Table 1: The results

Model	TN	FP	FN	TP	F1 Score	Train Time
XGBoost	297987	643	846	524	41.31	10min 3s
LightGBM	297905	725	1074	296	24.76	6.6s
MLP	297250	1380	994	376	24.06	25min
ResNet	298535	95	1197	173	21.12	57min 13s
FTT	298625	5	1277	93	12.67	3h 13s
XBNet	298421	209	1287	83	9.90	8h

References

[1] Card Fraud Report. Available in: https://nilsonreport.com/upload/content_promo/NilsonReport_Issue1209.pdf.

[2] KADRA, A. et al. *Well-tuned Simple Nets Excel on Tabular Datasets.* 2021.

[3] GORISHNIY, Y. et al. *Revisiting Deep Learning Models for Tabular Data.* 2021.

[4] SARKAR, T. *XBNet: An Extremely Boosted Neural Network.* 2021.

Student: Pedro Henrique Carvalho dos Reis

Advisor Professor: Prof. Dr. Reginaldo Arakaki