MSBD5001 Group Project Proposal
IEEE-CIS Fraud Detection

BASUCHAUDHURI, Priyanko
pbab@connect.ust.hk

FORK, Wing Yin
wyfork@connect.ust.hk

HO, Kwan Yu
kyhoat@connect.ust.hk

LEUNG, Cheuk Hei
chleungbc@connect.ust.hk

WONG, Ka Wing
kwwongbv@connect.ust.hk

WONG, Chin Hang
chwongar@connect.ust.hk

WONG, Tsz Fung
tfwongar@connect.ust.hk

Abstract

This project will explore a dataset of credit card transactions, provided by a leading e-payment merchant, applying various data analytic processes and identify fraudulent transaction through supervised learning on CART Decision Tree, LGBM, XGBoost.

1 Introduction

With the continual rise in usage of e-payment methods, it has become imperative for e-payment system providers to use real time fraud detection system to safeguard their customers security away from fraud and losing billions dollar in a year. [1] In 2018, there is a losses of $27.85 billion US Dollar globally due to the card fraud.

Payment fraud can occur in a variety of ways, such as identity theft, synthetic identity fraud, card skimming, or account takeovers. Most of these case, the identities are stolen by email or fake website phishing. Friendly fraud is another form of payment fraud when the actual cardholder is not admitting the transaction [2]. Card and payment fraud are well-known problems in society while machine learning may become a powerful assistant to this. [3]

2 Dataset

IEEE-fraud-detection dataset [4] contains over 600,000 credit card transactions, which have transaction details, payment method and identity of cardholder. Dataset is provided by Vesta Corporation, a e-commerce payment solution provider.

The data is labelled and indicate the fraudulent class in each transaction. There are around 400 features in this dataset and they are divided into transaction and identity parts.

Transaction part contains features which describe the transaction, such as product category, payment card features, issue bank, card type, country, address, purchaser, receiver email domain. Time Delta features present time distance from last transactions. Part of the data columns have been integrated as count features, such as the count of address which associated to the card. Also the match features represent the count which matches the name and address of that card's information.

Identity part contains features related to identity of the entity who made the transaction, such as device type, device information, network connection information (IP, ISP, Proxy), digital signature (browser, OS).

Challenges

The dataset contains over 400 columns, the abundance of features need data cleansing, exploratory analysis to find relevant features for model fitment and apply dimension reduction to reduce model complexity. [5] [6] Otherwise, dataset has an imbalanced label distribution on prediction classes, there are only 21K of 600K transactions are classified as fraud in dataset. This will be an issue for learning and evaluation as the model can always predict non-fraudulent and achieve a high accuracy. [7]

There are several methods to handle imbalanced dataset in practice. [8] The first method is to duplicate the fraudulent records to match the number of non-fraudulent records. Duplicating records will mimic a balanced dataset, hence avoiding the accuracy exploit. The second method is to tune the hyper-parameter "class_weight" in classification models provided by the sci-kit learn library. Tuning this hyper-parameter allows the model to adjust the penalties to each class when the model makes a mistake. There are also other approaches, such as using a Encoder-Decoder Neural Network [6] to learn a dataset, and apply a classifier on top of the latent space.

3 Expected outcome

- Card Fraud Classification models with using multiple algorithms.
- According to the grid search metrics, choose the best model from models with different parameters.
- Model has to achieve low loss, high F1-score and accuracy.
- Model is good fitted, model accuracy on validation and testing dataset should be close to training data.

4 Solution approach

This project intend to follow a system of classical waterfall software engineering approach where outcome of each of the constituent phase is fed into the next phase and periodic feedback is to considered to modulate parameters of the model.

The phases of development as following:

1. Data Analysis
2. Data Cleansing
3. Feature Extraction
    - 3.1 One-Hot Encoding
    - 3.2 Dimension Reduction (PCA, t-SNE)
4. Models
    - 4.1. SVM
    - 4.2. CART Decision Tree
    - 4.3. LightGBM
    - 4.4. XGBoost
5. Model Training
    - 5.1. k-fold Cross Validation
    - 5.2. Grid Search
6. Evaluation on Benchmarks
    - 6.1. Accuracy
    - 6.2. Loss
    - 6.3. F1-Score
7. Fine-tuning
    - 7.1. Model Parameters Tuning

5 Testing platform

A record of model training and testing, the hardware and spec as following:

OS: Ubuntu 18.04
RAM: 32GB
GPU: 2080ti
CUDA Driver: 450.80
Python: 3.7
Docker & Nvidia docker
Library:
Sci-kit Learn
pandas
matplotlib
seaborn

References

[1] European Central Bank. Ecb report shows a fall in card fraud in 2016, Sep 2018.

[2] SamanehSorournejad, Zahra Zojaji, Reza Ebrahimi Atani, and Amir Hassan Monadjemi. A survey of credit card fraud detection techniques: Data and technique oriented perspective, 2016.

[3] Yvan Lucas and Johannes Jurgovsky. Credit card fraud detection using machine learning: A survey, 2020.

[4] Vesta Corporation. Ieee-fraud-detection, Jun 2019. (https://www.kaggle.com/c/ieee-fraud-detection/data).

[5] Kai Shen, Anya Mcguirk, Yuwei Liao, Arin Chaudhuri, and Deovrat Kakde. Fault detection using nonlinear low-dimensional representation of sensor data, 2019.

[6] Yasuhiro Ikeda, Kengo Tajiri, Yuusuke Nakano, Keishiro Watanabe, and Keisuke Ishibashi. Estimation of dimensions contributing to detected anomalies with variational autoencoders, 2018.

[7] Jan Brabec, Tomáš Komárek, Vojtěch Franc, and Lukáš Machlica. On model evaluation under non-constant class imbalance, 2020.

[8] Kathleen Kerwin and Nathaniel D. Bastian. Stacked generalizations in imbalanced fraud data sets using resampling methods, 2020.