A project proposal for ITI-105

Credit Card Fraud Detection

by
Liew Chooi Chin

Introduction

Background

Credit Card (CC) frauds are happening at an ever increasing rate. The amount lost in 2022 is estimated to be 6.77 cents per $100 by Moneytransfers.com (n.d.). Frauds come in all *sizes*. Usually, CC frauds involve large amounts of money. The card owners can be immediately alerted if they have set the transaction warning threshold value. However, frauds also occur with small amounts of money transacted over a period of time. The illegal transactions are not discovered by the card owners until sometime later (Ang, 2023).

CC frauds are basically divided into whether the card is present or not at the point of transaction.

First, we will talk about CP (Card Present) fraud. Before EMV enabled cards became the default, magnetic stripe cards were the standard. The magnetic stripe stores all the information about the card and transmits the information at the POS (Point of Sale). Because the information is static, the magnetic stripe information can be easily copied illegally and loaded onto other cards. This card skimming technique was rampant in the past. However, with the standard EMV card nowadays, skimming becomes difficult (Credit Card Numbers: What Do They Mean? - Forbes Advisor, 2022).

Second, there is CNP (Card Not Present) fraud. To use a CC without the presence of a card, the number and CVV (Security Code) is required. But, Bank Identification Number attacks are common (Ang, 2023). This attack can be carried out by generating the CC numbers which are based on a standard format (Payment Card Number, 2023). A search in Google using keywords "credit card number generator" yields about 258,000,000 results. With the card number format being easily generated, and given the proliferation of online shopping and business transactions, CNP frauds are becoming the norm.

Besides that, online data leaks that contain users' CC numbers and personal details are common. This data is often traded on the dark web. Thus, CNP fraud becomes the norm for fraudulent transactions.

Fraud detection

The Straits Times (6 Jul 2023) reported that Mastercard is selling its AI tools to help banks in Britain to help detect fraudulent transactions. One of the main selling points of the tool is the ability to detect APP (Authorized Push Payment). This APP is a trick where scammers trick people into transferring money into what seems to be legitimate payees like families and companies. For example, people were tricked into paying small amounts of unauthorised transactions to ChatGPT and Apple (Ang, 2023).

We are intrigued by how well classical ML (machine learning) classification methods can be used to detect CC fraud. However, being a public member, we do not have access to any real world credit card transaction data. We will use a dataset found on Kaggle for our purpose.

Organisation of proposal

This proposal comes in the following sections:

- Section one, which is this section, introduces the problem.
- Section two discusses the various research related to fraud detection.
- Section three presents the EDA and preliminary findings.
- Section four will discuss the proposed methodology like methods and evaluation metrics to be considered for model learnings.
- Section five will discuss the deployment architecture, possible problems and limitations of the project.

```
"Evil never sleeps, and Virtue is ever-vigilant."
  --Good Omens, Terry Pratchett and Neil Gaiman
```

Literature survey

The surveys show that most research uses classification algorithms and outliers detection algorithms. It is interesting to note that outliers detection algorithms generally perform worse than other classifiers.

Using more sophisticated GA for feature selection (Ileberi et al., 2022) does not improve the results any better than using traditional classifiers without feature selections.

Resampling techniques are commonly employed by the authors in the survey. This technique is used to counter the highly imbalanced dataset that is used. Financial fraud detection is always highly imbalanced in terms of legitimate and fraudulent transactions.

Also using sophisticated ANN (Artificial Neural Networks) in Khan et al. (2022) does not yield better results than using LR (Linear Regression) and SVM (Support Vector Machine).

Use of threshold point in ROC curve in (Khan et al., 2022) is a novel idea. ROC curve plots true positive rate (recall or sensitivity) against false positive rate (1 - specificity). The diagonal line connecting (0, 0) and (1, 1) represents random prediction. The ideal classifier with zero false negative rate and a perfect one true positive rate will score a point (0, 1) on the ROC curve. Khan et al. (2022) uses a threshold point along with ROC Curve as a performance measure. The nearer the threshold point to the perfect (0, 1), the better the trade-off between true positive rate and false positive rate. The threshold points give more nuance than just AUC (area under curve) alone.

The table below summarises the methods and results from the literature surveyed.

Authors	Methods	Results
Alfaiz et al. (2022)	All K-Nearest Neighbors (AllKNN) undersampling technique along with CatBoost (AllKNN-CatBoost), with stratified K-fold cross-validation. Notes: Initially first best three selected algorithms are LR, KNN and DT.	AUC value of 97.94%, Recall value of 95.91%, F1-Score value of 87.40%

Dornadula et al. (2019)	Methods: Local outlier factor (LOF) Isolation forest (IF) Logistic regression (LR) Decision tree (DT) Random forest (RF)	Accuracy, Precision, MCC 0.4582, 0.2941, 0.1376 0.5883, 0.9447, 0.2961 0.9718, 0.9831, 0.9438 0.9708, 0.9814, 0.9420 0.9998, 0.9996, 0.9996
Ileberi et al. (2022)	DT, RF, LR, NB, ANN with min-max scaling, genetic algorithm feature selection. Results for full feature vector is shown here.	Model, Accuracy, Recall, Precision, F1-Score RF, 87.95%, 77.87%, 92.63%, 84.61% DT, 96.91%, 76.10%, 71.07%, 73.50% ANN, 97.80%, 74.33%, 42.85%, 54.36% NB, 80.31%, 64.60%, 13.95%, 22.95% LR, 93.88%, 60.17%, 62.96%, 61.53%
Maniraj et al. (2019)	Methods: LOF IF	Class, Precision, Recall, F1-Score, Support 0 Non-fraud, 1.00, 1.00, 1.00, 284315 1 Fraud, 0.05 0.05, 0.05, 492 0 Non-fraud, 1.00, 1.00, 1.00, 284315 1 Fraud, 0.33, 0.33, 0.33, 492
Plakandaras et al. (2022)	AutoML called JAD Bio (JustAddData Bio), with KFoldCrossValidation	Best model Precision, Recall, F-measure 0.796, 0.848, 0.821
Khan et al. (2022)	Methods: LR ANN (Artificial Neural Network) SVM	MCC, ROC Curve (Threshold point) 73.3%, 0.97 (0.38, 1) 76.32%, 0.90 (0.24, 1) 82.69% , 0.94 (0.22, 1)

Table 1: Comparisons of methods and results from various literatures

Exploratory Data Analysis

The dataset used is a made-up dataset found on Kaggle. The url is https://www.kaggle.com/datasets/kartik2112/fraud-detection?select=fraudTest.csv

Columns in the dataset:

- index - Unique Identifier for each row
- trans_date_trans_time - Transaction DateTime
- cc_num - Credit Card Number of Customer
- merchant - Merchant Name
- category - Category of Merchant
- amt - Amount of Transaction
- first - First Name of Credit Card Holder
- last - Last Name of Credit Card Holder
- gender - Gender of Credit Card Holder
- street - Street Address of Credit Card Holder
- city - City of Credit Card Holder
- state - State of Credit Card Holder
- zip - Zip of Credit Card Holder
- lat - Latitude Location of Credit Card Holder
- long - Longitude Location of Credit Card Holder
- city_pop - Credit Card Holder's City Population
- job - Job of Credit Card Holder
- dob - Date of Birth of Credit Card Holder
- trans_num - Transaction Number
- unix_time - UNIX Time of transaction
- merch_lat - Latitude Location of Merchant
- merch_long - Longitude Location of Merchant
- is_fraud - Fraud Flag <--- Target Class

We have conducted an EDA to understand the features of the dataset. For codes on EDA, please refer to the attached Notebook. Selected features for model learning are discussed below.

Findings

unix_time and trans_date_trans_time are the same information. To make use of the various time functionality in pandas, we convert the trans_date_trans_time to datetime64[ns] type.

We found that each of the unique cc_num corresponds to its own unique address. Dob (date of birth), gender and job is not useful is prediction, so we will not be using these features.

We also look at the merchant names and their categories of shops. The merchants' names are not useful for fraud prediction. But for the categories, we found that some categories of shops are more prone to fraudulent transactions, so we keep the categories for features to be used in learning.

Distance between a customer and a point of sale is useful in predicting frauds. When a distance is unusually far for a transaction, the transaction is suspicious. We use the latitude and longitude of the customer and the merchant to calculate the distance between them.

We use the pandas rolling window to calculate the average number of transactions, the average amount, and the average distance between a customer and a point of sale. Average amount is useful as a guide to find an unusually high amount of fraudulent transactions. While the average number of transactions is useful as a guide to find high volume but low amount frauds.

When we grouped the frauds into hours in a day, we found that a high number of frauds occur between 22 to 03 hours at midnight. So, we include this feature for learning.

Summary of selected features

These features are explained here.

- `cat_code`: Categories of shopping or merchants. We found that some categories have high chances to be used for fraudulent transactions.
- `is_midnight`: We group the transactions by hours. From the hourly groups, we found that at 22 to 03 hours, the number of frauds is unusually high. Therefore, we keep this feature for model learning.

- `count_3d`, `count_7d`, `count_30d`: Rolling window periods of 3, 7 and 30 days. This is the number of average transactions of a customer. If the number of transactions is unusually high, then we can suspect that the transactions are suspicious.
- `amt_3d`, `amt_7d`, `amt_30d`: Rolling window periods of 3, 7 and 30 days. This is the average of transactions of a customer. If the amount of transactions is unusually high, then we can suspect that the transactions are suspicious.
- `amt`: This is the actual transaction amount occuring at a point of sale.
- `distance_3d`, `distance_7d`, `distance_30d`: Rolling window periods of 3, 7 and 30 days. This is the average distance between a customer and a point of sale. If the distance between a customer and a point of sale is unusually high from the average distance, we can suspect that the transactions are suspicious.
- `distance`: This is the actual distance between a customer and a point of sale when a transaction is carried out.
- `is_fraud`: The label of whether a transaction is fraud or legitimate.

Proposed Methods

We consider the following methods in order to develop useful models. We will develop about 3 models and compare the performance of the 3 models. For model learning, we are likely to choose supervised learning. With the availability of labels, supervised learning comes in handy because the evaluation between the predictions and the actual labels can be compared.

Scaling of features

Scaling of data may be necessary for some algorithms to perform optimally. Generally when regularisation is used in linear regression algorithms, features need to be scaled or normalised. This is because regularisation algorithms are sensitive to a large range of extreme values. Whereas algorithms like decision trees, scaling is not required. The decision at every node can be made based on actual values.

So, we will scale features according to the algorithms that we choose.

Resampling techniques

There are three types of resampling techniques. They are over sampling, under sampling and mixed sampling that combines the over and under sampling methods.

Alfaiz et al. (2022) uses all three types of sampling techniques. Dornadula et al. (2019), Ileberi et al. (2022) uses SMOTE. In their results, resampling techniques have proven to improve the performance of prediction.

Our labels for fraud are far fewer than legitimate transactions. Thus, we will need to either synthetically generate more fraud samples, synthetically reduce the samples of legitimate samples, or do both at the same time.

Stratified K-fold Cross Validation

Stratified K-fold cross validation is a method to split the dataset into training and test sets. Alfaiz and Fati (2022) and Plakandaras et al. (2022) use this technique in their models' learning process.

Because of the large number of non-fraud labels, stratified K-fold CV is used to perform cross validation. With ordinary splitting, the algorithm might totally miss out fraud labels in some folds because the number of frauds is extremely few. Therefore, using stratified K-fold ensures that the percentages of the two classes are balanced according to their original ratio in the dataset.

Proposed Evaluation Metrics

Evaluation metrics

There are two classes, fraud and non-fraud, to be classified, hence a binary classification problem. We have ratios of TP, TN, FP and FN to consider. Classification evaluation metrics consists of the following: Accuracy, Precision, Recall, Specificity, False Positive Rate, F1-Score, ROC-AUC, PR-AUC, and MCC. To interpret the classification evaluation metrics meaningfully, we will look into the general desirable characteristics of a fraud detection system.

Accuracy

In this fraud detection scenario, accuracy is not a good performance measure. It is because the number of non-fraud cases is over 99.42% percent and any prediction will have a 99.42% chance of being correct!

Precision, Recall and PR-Curve

In fraud detection, it is better to have a false alarm (i.e. false positive) that warns the users wrongly than to let fraud happen. However, we do not want to have overly high FP alarms so that customers will start ignoring the warnings. At the same time, we want to be able to catch as many frauds as possible. Thus, the ability to identify fraud when it is actually a fraud is highly valuable in this scenario. Thus, we will want recall to be high.

Recall and precision is a trade-off of each other, i.e. high recall will result in lower precision and vice versa. Since, we have decided that catching as many frauds as possible is important, then we will accept a trade-off of slightly lower precision.

Having determined the priority of recall and precision, we can thus use a PR-curve that will give an idea of the various thresholds and performance of precision and recall.

MCC (Matthews Correlation Coefficient)

In Dornadula (2019) and Khan (2022), MCC is used as a performance measure. MCC is widely used in Bioinformatics as an evaluation metric (Boughorbel et al., 2017). In Bioinformatics, imbalanced data like rare diseases is similar to fraud detection. The positive rare disease or fraud classes are usually very tiny compared to the general negative non-disease or non-fraud classes.

MCC uses all the values provided in the confusion matrix; they are precision, recall, specificity and negative predictive value (i.e. 1- specificity). Thus, MCC takes into account all TP, TN, FP and FN of the confusion matrix where none of the precision, recall, F1-score or ROC-AUC have been able to do. In essence, MCC computes the correlation between the actual value and the predicted values (Khan et al., 2022).

In addition, MCC is unaffected by imbalanced data (Chicco & Jurman, 2020, Chicco and Jurman, 2023). Thus, it is highly valuable evaluation metrics for our fraud detection project.

Potential Problems and Limitations

In the synthetic dataset, we assume that the customers remain at one place during the whole period of the dataset. The distance between the customer and the point of sale always refers to the origin, which is the customers' residential addresses. If a customer travels to another place, the transactions that are performed at that place would have a different origin.

Finally, the limitation of our time and technical skills. We are limited to carry out a small scale experimentation with the various ML available at our disposal.

References

Alfaiz, N. S., & Fati, S. M. (2022). Enhanced Credit Card Fraud Detection Model Using Machine Learning. *Electronics*, *11*(4), 662. https://doi.org/10.3390/electronics11040662

Ang, Q. (2023, June 10). Singapore credit and debit cardholders report multiple unauthorised charges from ChatGPT, Apple. *The Straits Times*. https://www.straitstimes.com/singapore/courts-crime/singapore-credit-and-debit-cardholders-report-multiple-unauthorised-charges-from-chatgpt-apple

Appendix B - Credit Card Number Formats. (n.d.). Support.worldpay.com. Retrieved July 9, 2023, from http://support.worldpay.com/support/CNP-API/content/appendbcredit.htm

Aurélien Géron. (2019). *Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media, Inc.

Borgne, Y.-A. L., Siblini, W., Lebichot, B., & Bontempi, G. (2020). *Reproducible Machine Learning for Credit Card Fraud detection - Practical handbook*. Fraud-Detection-Handbook.github.io. https://fraud-detection-handbook.github.io/fraud-detection-handbook

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, *12*(6), e0177678. https://doi.org/10.1371/journal.pone.0177678

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-019-6413-7

Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1). https://doi.org/10.1186/s13040-023-00322-4

Credit Card Numbers: What Do They Mean? – Forbes Advisor. (2022, December 20). Www.forbes.com.

https://www.forbes.com/advisor/credit-cards/what-does-your-credit-card-number-mean/#:~:text=Credit%20card%20numbers%20are%20usually

Dornadula, V. N., & Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 631–641. https://doi.org/10.1016/j.procs.2020.01.057

Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00573-8

Khan, S., Alourani, A., & Mishra, B. (2022). Developing a Credit Card Fraud Detection Model using Machine Learning Approaches. *International Journal of Advanced Computer Science and Applications*, *13*(3).

Machine Learning Group of ULB (Université Libre de Bruxelles), & Worldline. (2020). *Credit Card Fraud Detection*. Www.kaggle.com. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Maniraj, S. P., Saini, A., Ahmed, A., & Sarkar, S. (2019). Credit card fraud detection using machine learning and data science. *International Journal of Engineering Research*, *8*(9), 110–115.

Moneytransfers.com. (n.d.). *Shocking Credit Card Fraud Statistics & Facts for 2023*. Moneytransfers.com. https://moneytransfers.com/news/2022/09/21/credit-card-fraud-statistics

Payment card number. (2023, June 14). Wikipedia. https://en.wikipedia.org/wiki/Payment_card_number

Plakandaras, V., Gogas, P., Papadimitriou, T., & Tsamardinos, I. (2022). Credit Card Fraud Detection with Automated Machine Learning Systems. *Applied Artificial Intelligence*, *36*(1). https://doi.org/10.1080/08839514.2022.2086354

The Straits Times. (2023, July 6). *Mastercard's new AI tool helps British banks tackle scams*. The Straits Times. https://www.straitstimes.com/business/mastercard-s-new-ai-tool-helps-british-banks-tackle-scams

---

Appendix

A. eda_v1_notebook.pdf