**\* 진우 머리속 지식**

대포통장을 활용한 사기거래는 대포통장이 잘 되는지 확인하려고 짧은 시간동안 계좌에 돈을 넣고 빼는 경향이 있음. 여기서 TransactionAmt 데이터를 중점으로 봤는데 사기 거래일 경우에 같은 카드에서 같은 돈이 왔다갔다 한 경우가 엄청 많다. 기본카드는 없고 없는것은아니지만 내가 볼때는 못봤어 한번 그것도 코드로 짜서 확인해봐야지

사기거래인 경우에는 $1 단위로 많이 거래가 되었고

**\* Characteristics of CNP fraud**

CNP fraud is a low risk/high reward crime, and will continue to grow until major changes occur.”

<https://www.digitaltransactions.net/facing-a-big-jump-in-u-s-cnp-fraud-retailers-turn-to-automated-tools/>

**\* About CNP**

It is most commonly used for payments made over [Internet](https://en.wikipedia.org/wiki/Internet), but also mail-order transactions by mail or [fax](https://en.wikipedia.org/wiki/Fax), or over the [telephone](https://en.wikipedia.org/wiki/Telephone).

**\* About 'chargeback'**

If a fraudulent CNP transaction is reported, the [acquiring bank](https://en.wikipedia.org/wiki/Acquiring_bank) hosting the [merchant account](https://en.wikipedia.org/wiki/Merchant_account) that received the money from the fraudulent transaction must make restitution to the cardholder - this is called a [chargeback](https://en.wikipedia.org/wiki/Chargeback).

**\* To prevent fraud**

The [card security code](https://en.wikipedia.org/wiki/Card_security_code) (commonly CVV2) system has been set up to reduce the incidence of credit card fraud arising from CNP

**\* Mail-order fraud**

A carder would obtain the credit card information for a local resident and then intercept delivery of the illegitimately purchased merchandise at the shipping address, often by staking out the porch of the residence.

**\* Characteristics of CNP fraud**

Fraudsters bet on the fact that many fraud prevention features are not used for small transactions.

**\* Investigation limit**

Credit card companies only investigate if the charge is more than $10 because it costs about that much to run an investigation.

**\* Criminal sites**

Then the money was moved to bank accounts in Lithuania, Estonia, Latvia, Bulgaria, Cyprus, and Kyrgyzstan where the money could not be traced or recovered.

**\* Etc**

Only about 10 percent of the fraudulent charges were ever reported or contested by the card owner that was billed.

<https://en.wikipedia.org/wiki/Card_not_present_transaction>

**\* Description of CNP fraud**

Except, for merchants, it’s not so simple. CNP fraud happens when real customers’ card details are stolen and used to make purchases, most frequently online.

**\* Chargeback fraud**

if customers notice fraudulent transactions in their bank statements, then they speak with their banks and organize the money to be returned to their accounts, which is called a chargeback. But chargeback fraud is rife too.

<https://bdaily.co.uk/articles/2019/07/10/cnp-the-fraud-which-cost-the-uk-506-million-over-one-year>

**\*Data involved in CNP fraud**

Chargebacks911 says this is the data that's generally involved in card-not-present fraud:

Card number

Card expiration date

Card security code

Personal billing information (such as the cardholder’s address)

**\* Methods that a cyberthief employ to collect card data**

**Hacking**. A hacker invades a computer system to grab, alter or destroy information, often by installing malware, short for malicious software. Often, a hacker gains this data by breaking into computer systems operated by places such as retailers, restaurants, hotels, banks, schools and government agencies. With this stolen data, cybercrooks can produce cloned credit cards or commit other kinds of fraud, such as card-not-present fraud. They also can sell the data via underground marketplaces online.

**Skimming**. A crook will install a device called a skimmer at a gas pump, ATM or other location to copy data from credit or debit cards that have magnetic strips. Also, a waiter or store clerk with a portable skimmer might steal card data when you're not watching. What happens to the data? A crook typically uses it for fraudulent activity or sells it to other criminals.

**Phishing.** As the name suggests, a phisher goes fishing for sensitive data by posing as a legitimate organization, such as a credit card company or bank, via email, phone or text to reel in information like card data or passwords. This data can be gathered, for instance, by urging an email recipient to open a malicious attachment or click on a malicious link.

<https://creditcards.usnews.com/articles/what-is-card-not-present-fraud>

**\* 각 column들에 대한 reference**

<https://www.kaggle.com/alijs1/ieee-transaction-columns-reference>

**\* 각 column에 대해서**

\* TransactionDT: timedelta from a given reference datetime (not an actual timestamp)

\* TransactionAMT: 거래 금액

\* productCD: w c r h s가 있는데 각각 뭐를 의미할까

\* card: payment card info. such as card type, card category, issue bank, country, etc.

card1: other values가 85% 가까이 차지. 13,555개 card info일 확률이 높음

card2: 나름 고르게 분포되어있음. 501개 issue bank일 확률이 높음

card3: 150과 185 값에 97%가 몰려있음. 115개

card4: 카드 종류

card5: 226값에 50%가 몰려있고 224값에 13%가 몰려있음. 120개 -> isFraud랑 같이 groupby

card6: 신용카드, debit card

\* addr: address

\* dist: distance

dist1: 59%가 null값

dist2: 93%가 null값

\* P\_ and R\_ emaildomain: purchaser and recipient email domain

\* C1 ~ C14: counting, how many addresses are found to be associated with the payment card

\* D1 ~ D15: timedelta, such as days btw previous transaction

\* M1 ~ M9: match, such as names on card and address

Vxxx:: Vesta engineered rich features

**\* id\_ker column 알아보기**

id\_01: -5가 57%

id\_12: Not found가 85%, Found가 15%

id\_15: Found, New, Unkonw이 97%. 나머지 null

id\_19: 굉장히 고르게 value가 퍼져있음. Not skewed

id\_20: 굉장히 고르게 value가 퍼져있음. Not skewed

id\_28: Found, New이 97%. 나머지 null

id\_29: Found, NotFound가 97%. 나머지 null

id\_31: 인터넷 브라우저를 알려주는 듯함. 골고루 퍼져있음.

Device type: Desktop, mobile이 끝임.

Device info: Window, iOS, MacOS가 55% 차지

**\* Definition**

Trans\_train: 전체 데이터

id\_train: 전체 데이터

tns\_train: Trans\_train에서 Vxxx column 전체 제거

id\_ker: id column중에서 중요 & null값 적은 column들만

**\* 해볼 것들**

1) 전체 거래금액 vs 사기 당한 금액의 비교

(작은 거래의 기준점을 정하고 그에 따라서 기준점 이상과 이하의 거래금액 비교)

-> 사기 안 당한 사람은 $134, 사기 당한 사람은 $149

2) 사기 당한 데이터와 사기 당하지 않은 데이터를 구별해서 상관관계 차이 확인하기

->

3) id\_12: Categorical. 각 범주별로 fraud 여부와 상관이 있는지 확인 요

4) isFraud == 1로 했을 때 profiling 해보기

5) fraud와 null값 유무의 상관관계를 조사해야 한다.

6) card5 국가일 확률이 제일 높기에 확인해보기

226값에 50%가 몰려있고 224값에 13%가 몰려있음. 120개 -> isFraud랑 같이 groupby해보기. VESTA 거래국가 및 현황과 비교해서 매칭시켜보기

7) 이메일 도메인이랑 isFraud 비교해서 유의미한 거 찾아보기

8) card4 column이랑 isFraud와의 상관관계

9) transactionAmt의 isFraud 0과 1의 median을 비교해보기!

10) 모델은 어떤거? RanFo? Boost?

구체적인 값을 정해야 하는 이유?

-> 사기자가 사기금액이 너무 낮다. 그걸 반영하기 위한 것

-> 최적값은 Tree 써서 구하면 되니까 그렇게 하쥬(by.조조)

11) 시간이 관계가 있는가?

-> DT가 짧으면서 어느정도 Card ID가 같으면서 Amt가 같은거는 Labeling해야 될 것이다.

-> DT 차이를 바탕으로 Column을 더 만들자

-> Index 차이가 나는 것만.

-> 말한 증상이 나오는 것의 DT max, min을 구해서

-> 실제 시간을 DT를 구해서 알 수 없지는 않다.

-> 굳이 시간을 찾아야 되나?

-> 조조: DT를 그냥 날리고 Amt만 비슷한 것만 묶어서 봐도 충분히 볼 수 있을것이다.

-> 대포통장 확인한 과정 뒤에 같은 카드로 거래가 있는지\

-> 이상한 게 왜 그럼 싹다 fraud가 찍히지?

12) Product CD가 'c'일 때는 $0.001 단위까지 내려가는데 Fraud일 경우에는 모두 $0.001 단위까지 내려간다. 이거 확인해보자!

13) 같은 금액이 계속 왔다갔다 한 애들이 fraud인지 아닌지 확인할 필요가 있다.

14) C와 D columns 다 합쳐보고 fraud랑 correlation. nonnull과 null을 비교한다는 것이다. value 차이를 무시하고 하는 것.

가장 좋은 방법은 binarary로 해서 보고, 값으로 해서 보고 C1과 C2 correlation 높으면 묶고 이 방법이 제일 좋다.

15) AutoML을 해보는 건 어떤가요

16) V column 전처리 한 번 하기