Overview

Data No

Notebooks Discussion

Leaderboard

rd Rules

s Team

My Submissions

A Research Prediction Competition

IEEE-CIS Fraud Detection

\$20,000

Prize Money

Can you detect fraud from customer transactions?



IEEE Computational Intelligence Society \cdot 5,531 teams \cdot 11 days to go (3 days to go until merger deadline)

V E S T A^{**}

Data Description (Details and Discussion)

111

posted in IEEE-CIS Fraud Detection 2 months ago

Lynn@Vesta

Hi All,

I see many questions regarding data description, so it maybe a better idea to open a thread for discussion. The following is a bit more details about it:

- Transaction Table *
- TransactionDT: timedelta from a given reference datetime (not an actual timestamp)
- TransactionAMT: transaction payment amount in USD
- ProductCD: product code, the product for each transaction
- card1 card6: payment card information, such as card type, card category, issue bank, country, etc.
- · addr: address
- dist: distance
- P_ and (R_) emaildomain: purchaser and recipient email domain
- C1-C14: counting, such as how many addresses are found to be associated with the payment card, etc. The actual meaning is masked.
- D1-D15: timedelta, such as days between previous transaction, etc.
- M1-M9: match, such as names on card and address, etc.
- Vxxx: Vesta engineered rich features, including ranking, counting, and other entity relations.

Categorical Features:

ProductCD

card1 - card6

addr1, addr2

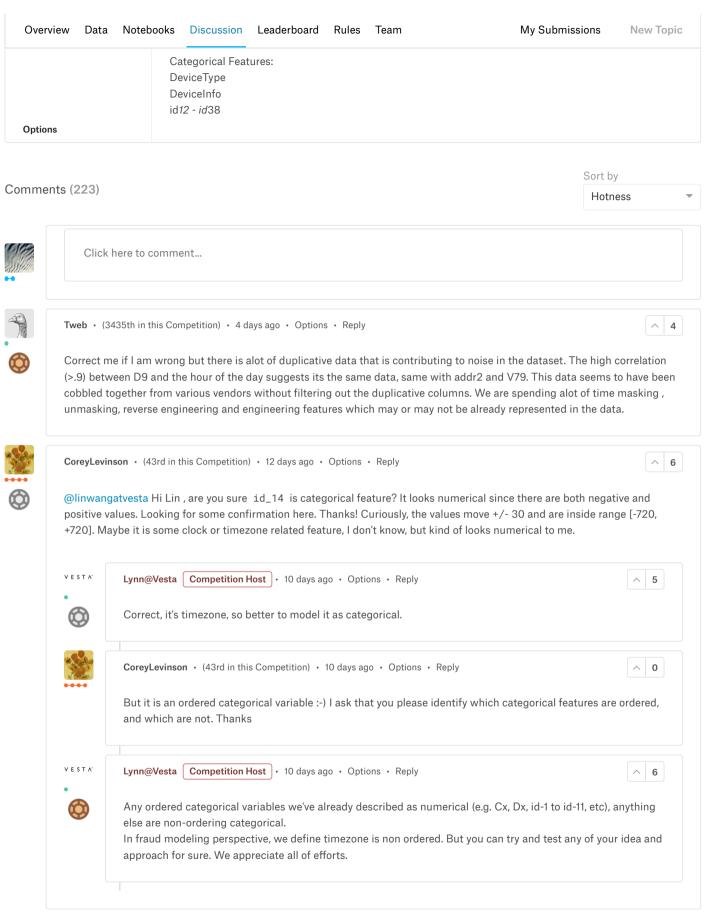
Pemaildomain Remaildomain

M1 - M9

Identity Table *

Variables in this table are identity information – network connection information (IP, ISP, Proxy, etc) and digital signature (UA/browser/os/version, etc) associated with transactions.

They're collected by Vesta's fraud protection system and digital security partners.





Overview Data Notebooks Discussion Leaderboard Rules Team My Submissions New Topic

นสเรองเบเที , มว เธ responsible ror mow many days nave passed ท**om the previous ตลกรองเบเท**



Tudor Lapusan • (4120th in this Competition) • 24 days ago • Options • Reply

^ 188



This discussion has a lot of comments and it takes time to read all of them:)

I created a summary with the most important information I found. I hope it will help. Please put a comment if I lost some important info.

Transaction table

"It contains money transfer and also other gifting goods and service, like you booked a ticket for others, etc."

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

"TransactionDT first value is 86400, which corresponds to the number of seconds in a day (60 * 60 * 24 = 86400) so I think the unit is seconds. Using this, we know the data spans 6 months, as the maximum value is 15811131, which would correspond to day 183."

TransactionAMT: transaction payment amount in USD

"Some of the transaction amounts have three decimal places to the right of the decimal point. There seems to be a link to three decimal places and a blank addr1 and addr2 field. Is it possible that these are foreign transactions and that, for example, the 75.887 in row 12 is the result of multiplying a foreign currency amount by an exchange rate?"

ProductCD: product code, the product for each transaction

"Product isn't necessary to be a real 'product' (like one item to be added to the shopping cart). It could be any kind of service."

card1 - card6: payment card information, such as card type, card category, issue bank, country, etc.

addr: address

"both addresses are for purchaser

addr1 as billing region

addr2 as billing country"

dist: distance

"distances between (not limited) billing address, mailing address, zip code, IP address, phone area, etc."

P_ and (R_) emaildomain: purchaser and recipient email domain "certain transactions don't need recipient, so Remaildomain is null."

C1-C14: counting, such as how many addresses are found to be associated with the payment card, etc. The actual meaning is masked.

"Can you please give more examples of counts in the variables C1-15? Would these be like counts of phone numbers, email addresses, names associated with the user? I can't think of 15.

Your guess is good, plus like device, ipaddr, billingaddr, etc. Also these are for both purchaser and recipient, which doubles the number."

D1-D15: timedelta, such as days between previous transaction, etc.

M1-M9: match, such as names on card and address, etc.

Vxxx: Vesta engineered rich features, including ranking, counting, and other entity relations.

"For example, how many times the payment card associated with a IP and email or address appeared in 24 hours time range, etc."

"All Vesta features were derived as numerical. some of them are count of orders within a clustering, a time-period or condition, so the value is finite and has ordering (or ranking). I wouldn't recommend to treat any of them as categorical. If any of them resulted in binary by chance, it maybe worth trying."

Identity Table

Variables in this table are identity information – network connection information (IP, ISP, Proxy, etc) and digital signature (UA/browser/os/version, etc) associated with transactions.

They're collected by Vesta's fraud protection system and digital security partners.

(The field names are masked and pairwise dictionary will not be provided for privacy protection and contract agreement)

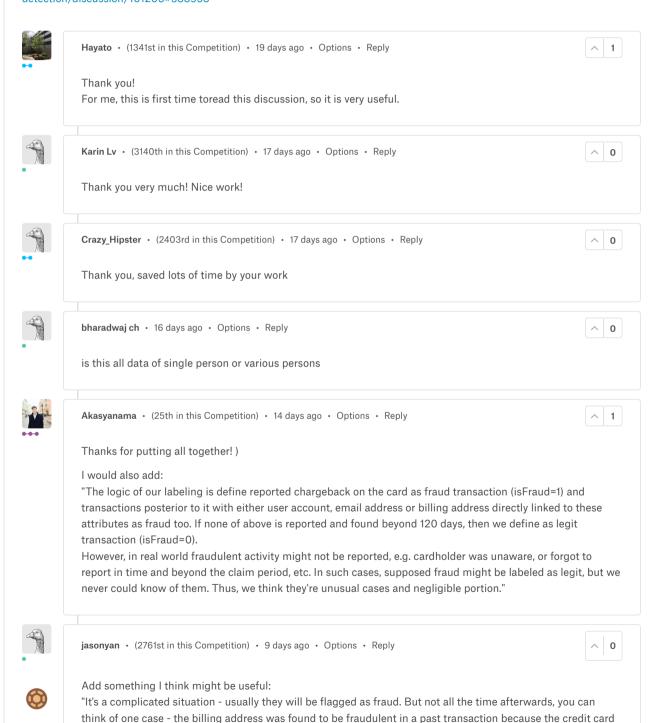
DeviceInfo: https://www.kaggle.com/c/ieee-fraud-detection/discussion/101203#583227

long an account stayed on the page, etc. All of these are not able to elaborate due to security partner T&C. I hope you could get basic meaning of these features, and by mentioning them as numerical/categorical, you won't deal with them inappropriately."

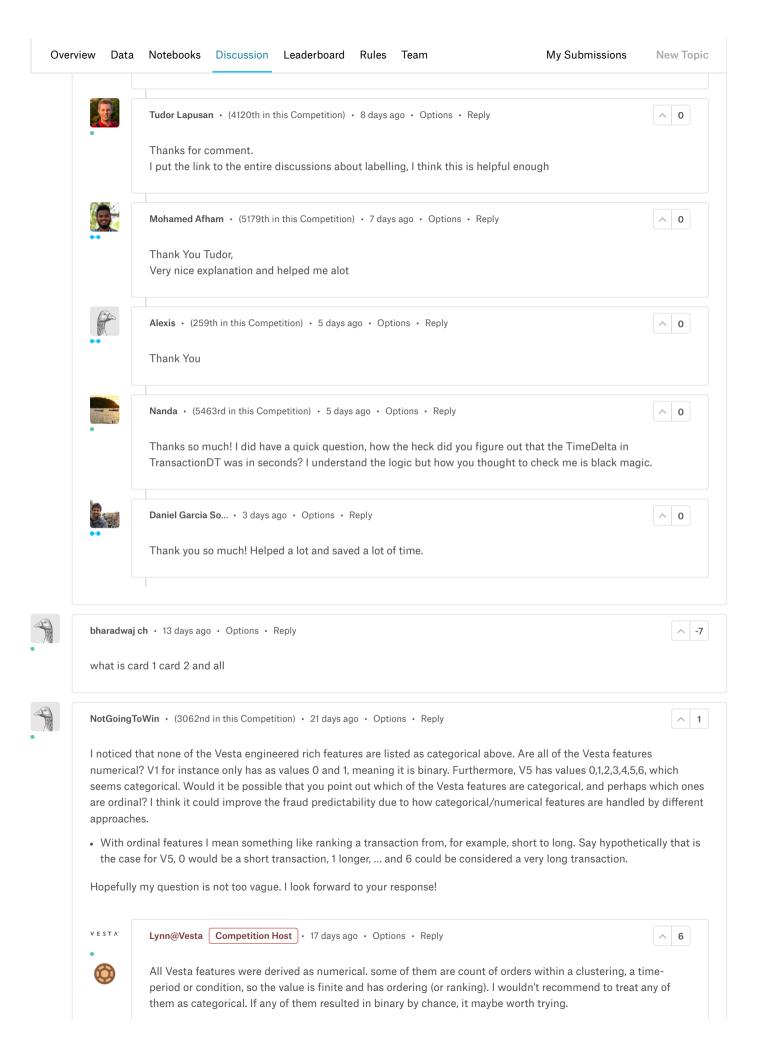
Labeling logic

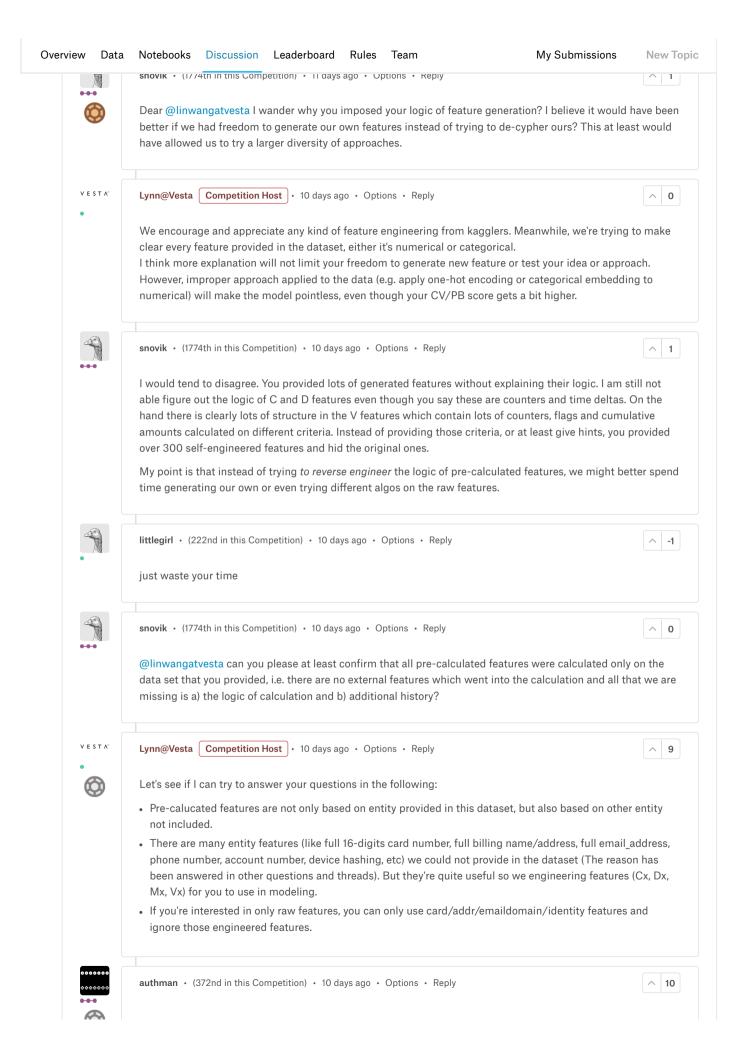
"The logic of our labeling is define reported chargeback on the card as fraud transaction (isFraud=1) and transactions posterior to it with either user account, email address or billing address directly linked to these attributes as fraud too. If none of above is reported and found beyond 120 days, then we define as legit transaction (isFraud=0).

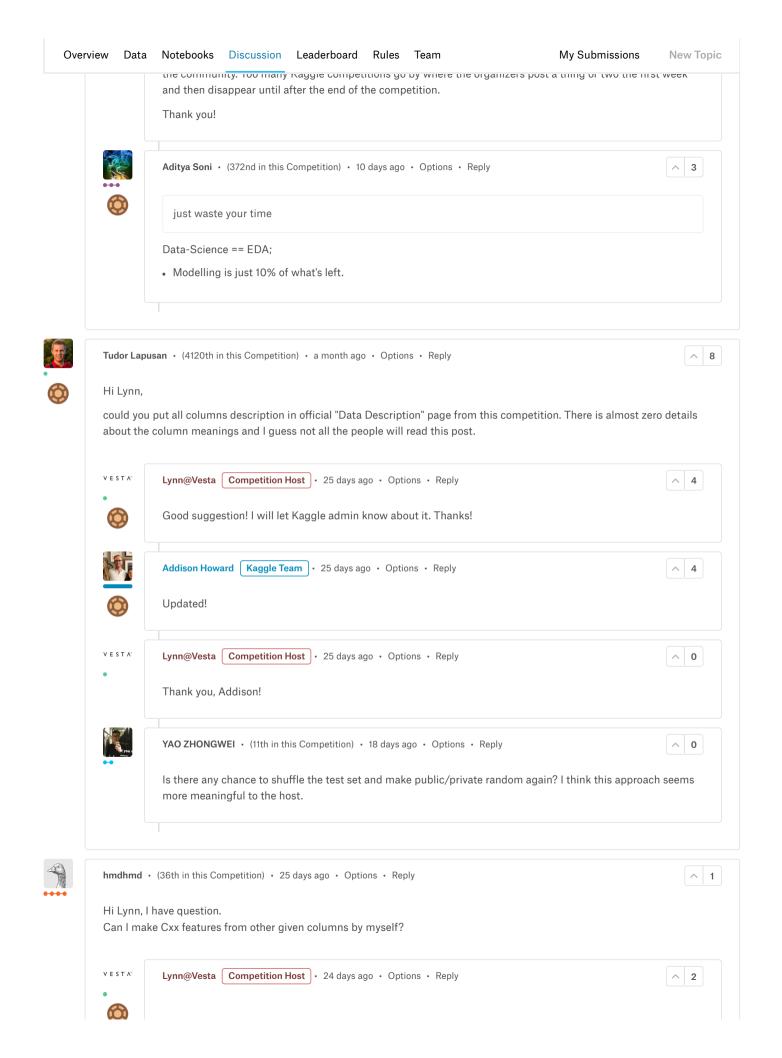
However, in real world fraudulent activity might not be reported, e.g. cardholder was unaware, or forgot to report in time and beyond the claim period, etc. In such cases, supposed fraud might be labeled as legit, but we never could know of them. Thus, we think they're unusual cases and negligible portion." Read more: https://www.kaggle.com/c/ieee-fraud-detection/discussion/101203#588953

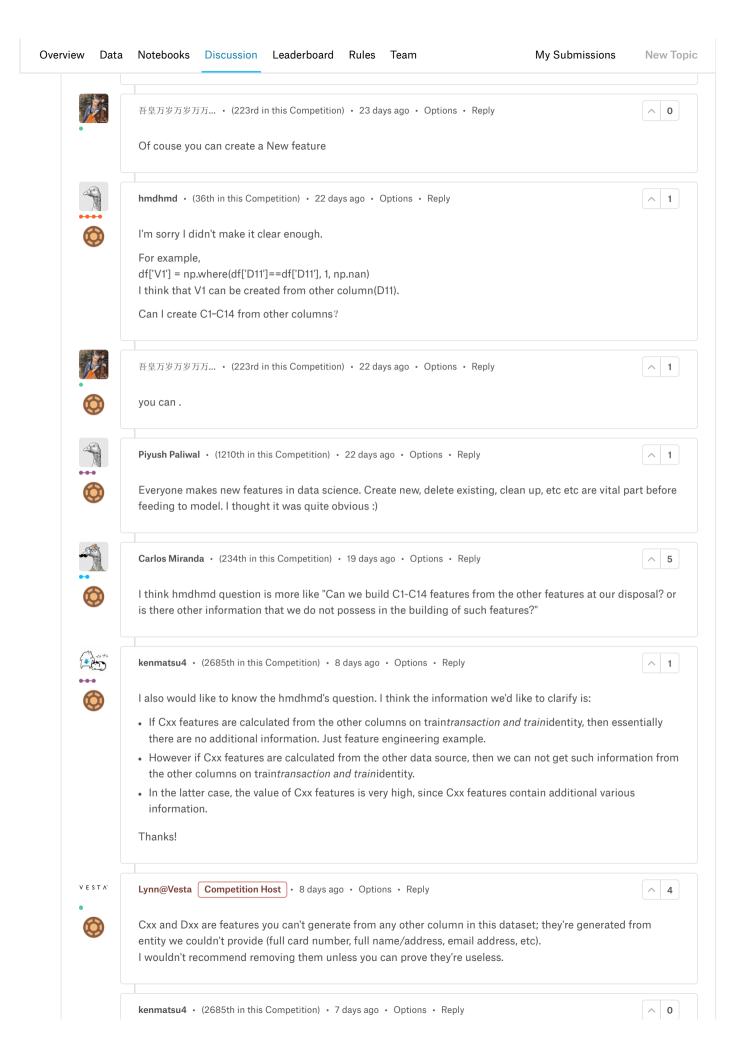


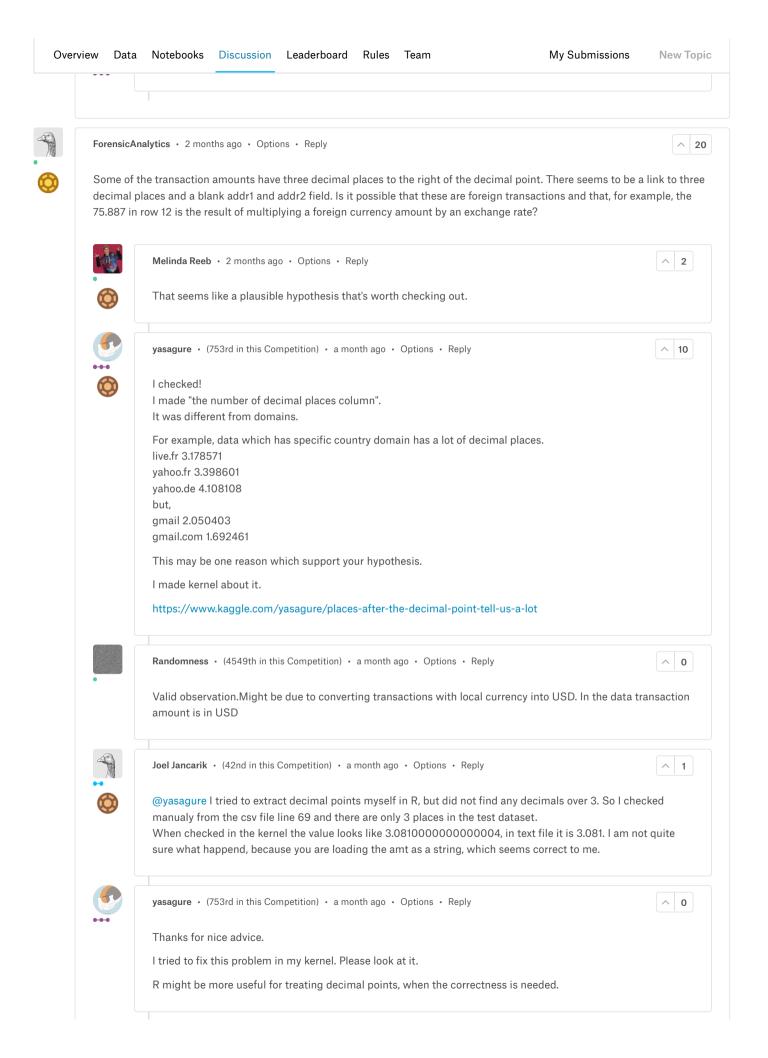
associated with it was stolen. But the cardholder is actually the victim, we're not going to blacklist him forever if

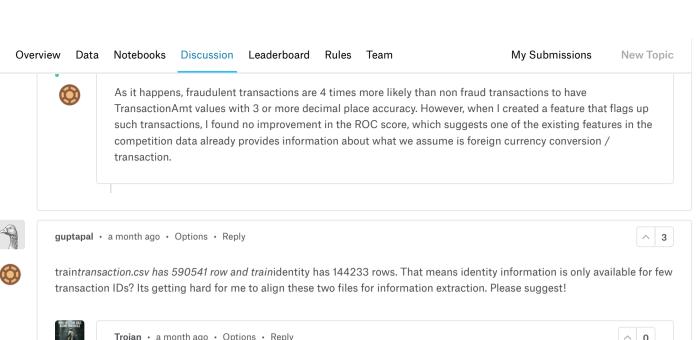


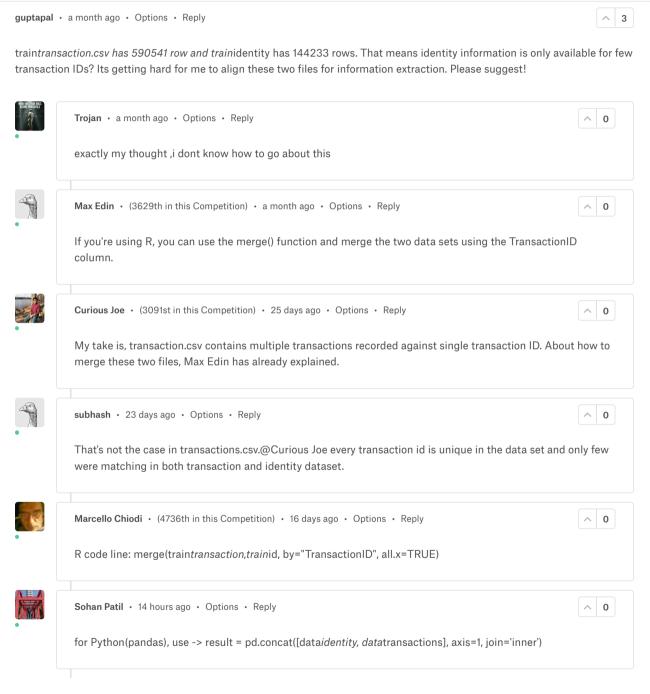












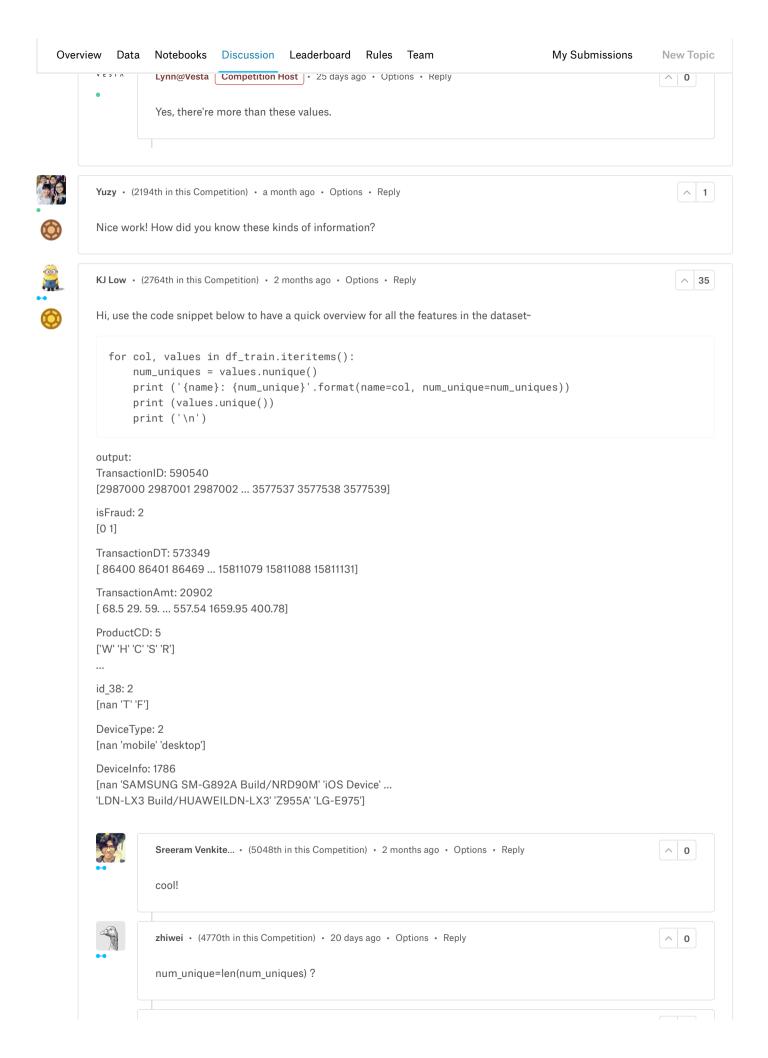


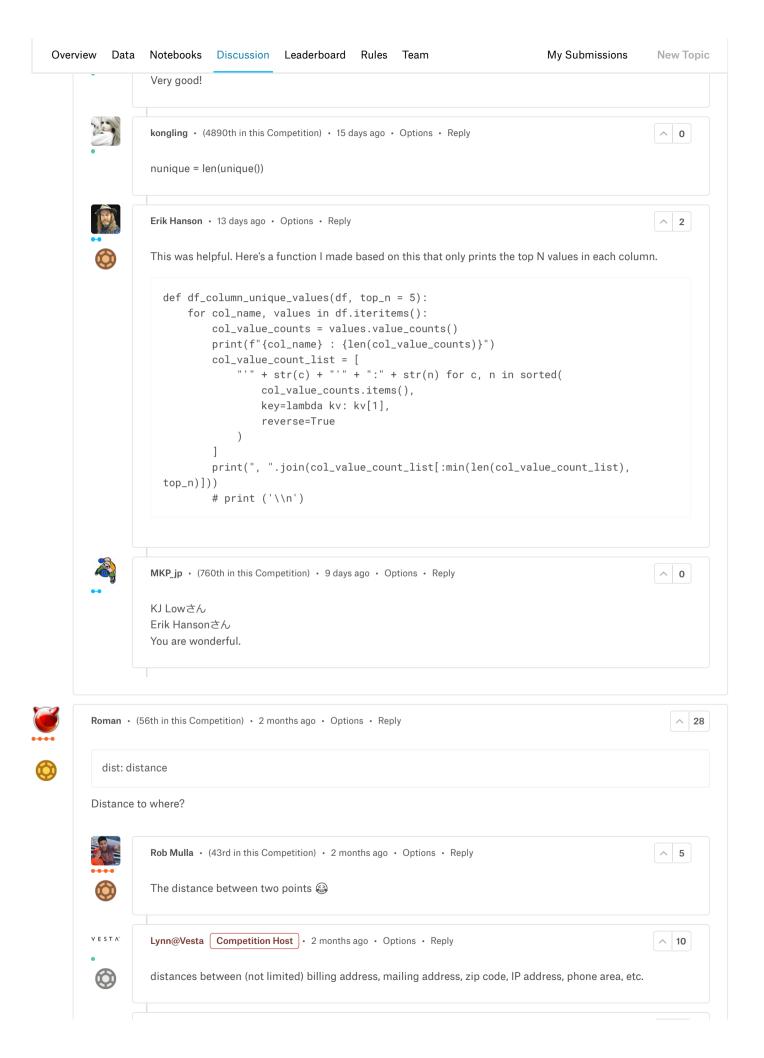
estelle1728 • (3856th in this Competition) • a month ago • Options • Reply

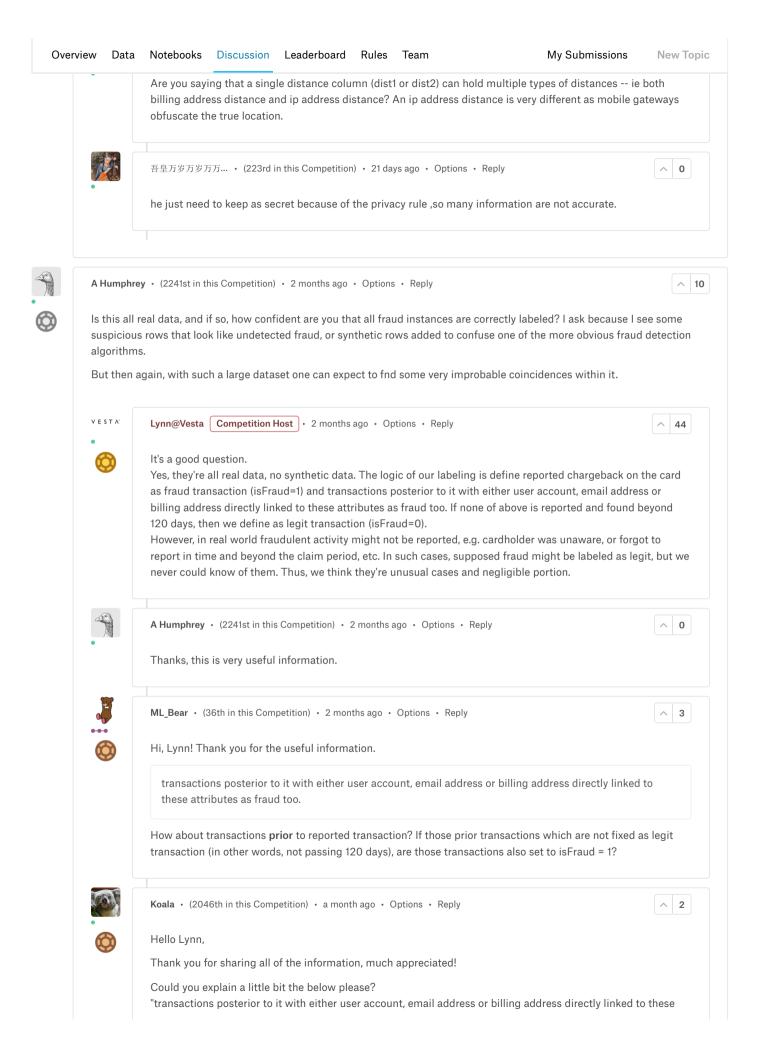


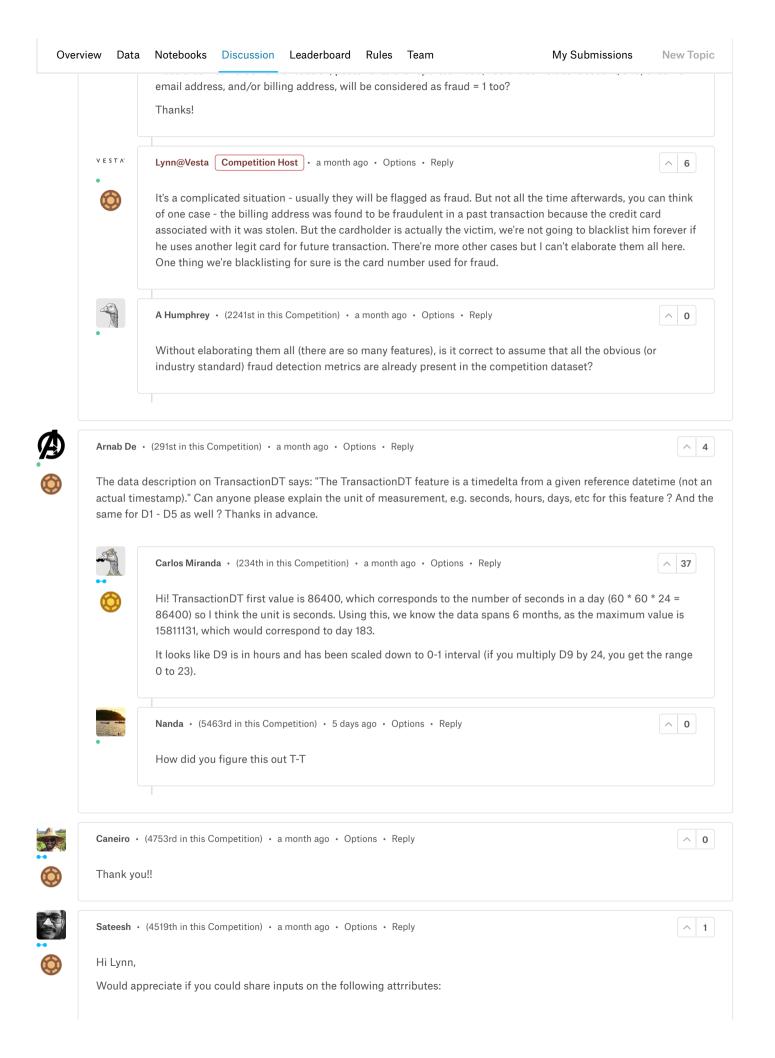


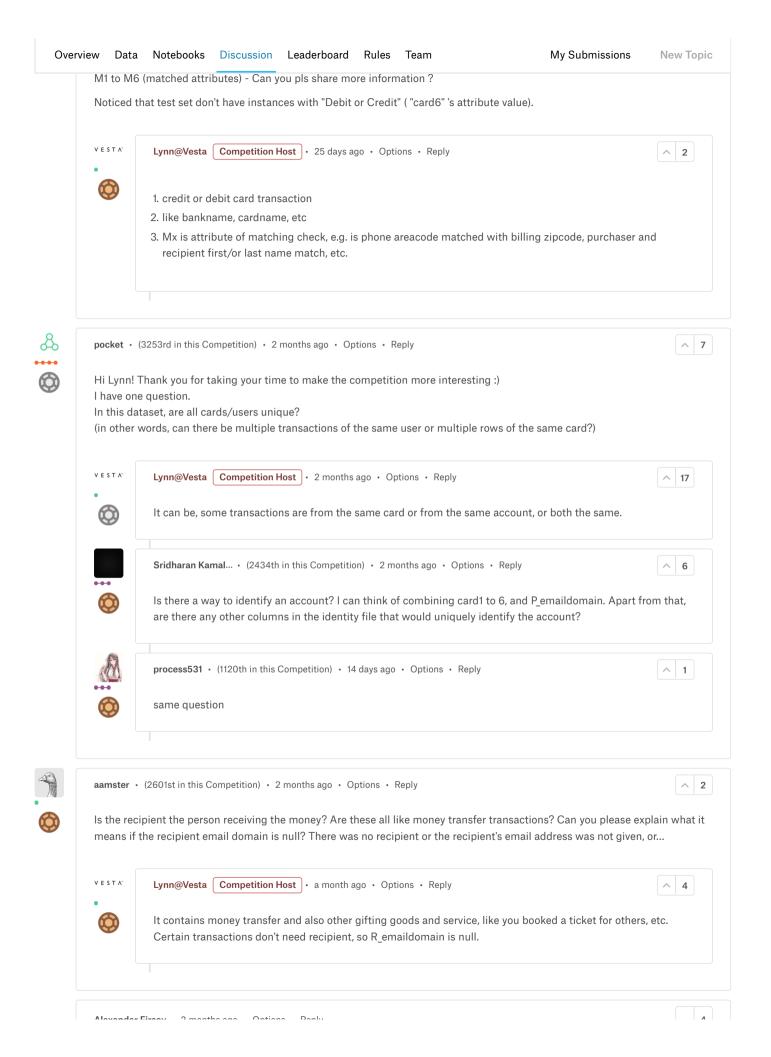
Thank you for this great work. I have a question about the feature card1 - card6. As is described, card1 - card6 denote payment card information. However, there are various value for specfied feature. For example, there are over 10, 000, 500, 150 different values in feature card1, card2, card3 respectively. So does so many different kinds of payment card exist? Thanks in advance.











S

mean that identity data is heterogeneous? So part of rows comes from Vesta and part from partners?

It would explain why "DeviceInfo" column sometimes contains device code like "LG-H840 Build/NRD90U" and sometimes OS version like "Android 7.1.2".

If identity data is a mixture of several system outputs, is it possible to find out what was the source of identity data? Is there a column indicating partners system/Vesta system?

VESTA

Lynn@Vesta Competition Host • 2 months ago • Options • Reply





If I understand your question well, all rows are homogeneous but columns are heterogeneous. In other words, you can think e.g. id-01 to id-20 are collected from Vesta, id-21 to id-30 are from sourceA and the rest are from sourceB.

However, I didn't quite understand your example - with DeviceInfo='LG-H840 Build/NRD90U' and OS='Android 7.1.2'. They're collected from the same source.



Alexander Firsov • 2 months ago • Options • Reply





Lynn, thanks! You have answered my question.

As for my example, I had assumed that the same column is combined from different sources as values have various nature. This is not the case as per your answer.

Original example was about "DeviceInfo" column, which sometimes contains device (LG-H840 Build/NRD90U) and sometimes OS (Android 7.1.2). So my guess that either parsing is broken or there are several sources for this column.



Gunes Evitan • (372nd in this Competition) • 2 months ago • Options • Reply

^ 2



I also thought the parsing is broken. DeviceInfo feature is probably parsed from user agent string. That's why the value can be different things like OS version, device model or other user agent fragments.



AmirH • (17th in this Competition) • 2 months ago • Options • Reply

^ 4



Thanks!

Why would a transaction not have a row of identity?

Would an offline transaction have info of the business?

VESTA



For various technical reasons, it's challenging to collect all identity information at the real-time transactions.



AmirH • (17th in this Competition) • 2 months ago • Options • Reply



Thanks Lynn!

Could you just share if all transactions are offline, online or both, and if offline, would the data show the business information.

Also, address information is of the payer or the Recepient?

Thanks!

