* 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  
P*emaildomain R*emaildomain  
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.  
(The field names are masked and pairwise dictionary will not be provided for privacy protection and contract agreement)

Categorical Features:  
DeviceType  
DeviceInfo  
id*12 - id*38

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

both are for purchaser  
addr1 as billing region  
addr2 as billing country

Hi Lynn! Thank you for taking your time to make the competition more interesting :)   
I have one question.   
In this dataset, are all cards/users unique?   
(in other words, can there be multiple transactions of the same user or multiple rows of the same card?)

It can be, some transactions are from the same card or from the same account, or both the same.

Is this all real data, and if so, how confident are you that all fraud instances are correctly labeled? I ask because I see some suspicious rows that look like undetected fraud, or synthetic rows added to confuse one of the more obvious fraud detection algorithms.

It's a good question.   
Yes, they're all real data, no synthetic data. 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.

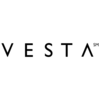
Hi Lynn! thank you for this explanation:)  
you say:D1-D15: timedelta, such as days between previous transaction, etc.   
so one question is, does D\* mean the value for the same user/card ? or somthing else?

[Lynn@Vesta](https://www.kaggle.com/linwangatvesta)Competition Host•10 days ago•Options•Reply

0

It could be on the same card/account/email/IPaddress, etc.

I guess, it's two line in the accounts holder address form.  
So addr1 + " " + addr2 = entire card holders address.

[[novice tier](https://www.kaggle.com/linwangatvesta)](https://www.kaggle.com/linwangatvesta)

[Lynn@Vesta](https://www.kaggle.com/linwangatvesta)Competition Host•10 days ago•Options•Reply

5

You may understand as billing zipcode and country.   
(It's not street address - which would cause too many levels)