Research Prediction Competition

IEEE-CIS Fraud Detection

Can you detect fraud from customer transactions?

$20,000
Prize Money

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

VESTA

Lynn@Vesta

Data Description (Details and Discussion)

posted in IEEE-CIS Fraud Detection 2 months ago

411

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
P*emaildomain* 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.

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

Options

Comments (223)

Sort by
Hotness

Click here to comment...

Tweb • (3435th in this Competition) • 4 days ago • Options • Reply 4

Correct me if I am wrong but there is alot of duplicative data that is contributing to noise in the dataset. The high correlation (>.9) between D9 and the hour of the day suggests its the same data, same with addr2 and V79. This data seems to have been cobbled together from various vendors without filtering out the duplicative columns. We are spending alot of time masking , unmasking, reverse engineering and engineering features which may or may not be already represented in the data.

CoreyLevinson • (43rd in this Competition) • 12 days ago • Options • Reply 6

@linwangatvesta Hi Lin , are you sure `id_14` is categorical feature? It looks numerical since there are both negative and positive values. Looking for some confirmation here. Thanks! Curiously, the values move +/- 30 and are inside range [-720, +720]. Maybe it is some clock or timezone related feature, I don't know, but kind of looks numerical to me.

Lynn@Vesta Competition Host • 10 days ago • Options • Reply 5

Correct, it's timezone, so better to model it as categorical.

CoreyLevinson • (43rd in this Competition) • 10 days ago • Options • Reply 0

But it is an ordered categorical variable :-) I ask that you please identify which categorical features are ordered, and which are not. Thanks

Lynn@Vesta Competition Host • 10 days ago • Options • Reply 6

Any ordered categorical variables we've already described as numerical (e.g. Cx, Dx, id-1 to id-11, etc), anything else are non-ordering categorical.
In fraud modeling perspective, we define timezone is non ordered. But you can try and test any of your idea and approach for sure. We appreciate all of efforts.

Akasyanama • (25th in this Competition) • 13 days ago • Options • Reply 5

transaction , D3 is responsible for "how many days have passed **from the previous transaction**

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 R*emaildomain 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

Hayato • (1341st in this Competition) • 19 days ago • Options • Reply 1

Thank you!
For me, this is first time toread this discussion, so it is very useful.

Karin Lv • (3140th in this Competition) • 17 days ago • Options • Reply 0

Thank you very much! Nice work!

Crazy_Hipster • (2403rd in this Competition) • 17 days ago • Options • Reply 0

Thank you, saved lots of time by your work

bharadwaj ch • 16 days ago • Options • Reply 0

is this all data of single person or various persons

Akasyanama • (25th in this Competition) • 14 days ago • Options • Reply 1

Thanks for putting all together!)

I would also add:
"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."

jasonyan • (2761st in this Competition) • 9 days ago • Options • Reply 0

Add something I think might be useful:
"It's a complicated situation - usually they will be flagged as fraud. But not all the time afterwards, you can think of one case - the billing address was found to be fraudulent in a past transaction because the credit card associated with it was stolen. But the cardholder is actually the victim, we're not going to blacklist him forever if

Tudor Lapusan • (4120th in this Competition) • 8 days ago • Options • Reply 0

Thanks for comment.
I put the link to the entire discussions about labelling, I think this is helpful enough

Mohamed Afham • (5179th in this Competition) • 7 days ago • Options • Reply 0

Thank You Tudor,
Very nice explanation and helped me alot

Alexis • (259th in this Competition) • 5 days ago • Options • Reply 0

Thank You

Nanda • (5463rd in this Competition) • 5 days ago • Options • Reply 0

Thanks so much! I did have a quick question, how the heck did you figure out that the TimeDelta in TransactionDT was in seconds? I understand the logic but how you thought to check me is black magic.

Daniel Garcia So... • 3 days ago • Options • Reply 0

Thank you so much! Helped a lot and saved a lot of time.

bharadwaj ch • 13 days ago • Options • Reply -7

what is card 1 card 2 and all

NotGoingToWin • (3062nd in this Competition) • 21 days ago • Options • Reply 1

I noticed that none of the Vesta engineered rich features are listed as categorical above. Are all of the Vesta features numerical? V1 for instance only has as values 0 and 1, meaning it is binary. Furthermore, V5 has values 0,1,2,3,4,5,6, which seems categorical. Would it be possible that you point out which of the Vesta features are categorical, and perhaps which ones are ordinal? I think it could improve the fraud predictability due to how categorical/numerical features are handled by different approaches.

- With ordinal features I mean something like ranking a transaction from, for example, short to long. Say hypothetically that is the case for V5, 0 would be a short transaction, 1 longer, ... and 6 could be considered a very long transaction.

Hopefully my question is not too vague. I look forward to your response!

VESTA

Lynn@Vesta Competition Host • 17 days ago • Options • Reply 6

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.

snovik • (1774th in this Competition) • 11 days ago • Options • Reply 1

Dear @linwangatvesta I wander why you imposed your logic of feature generation? I believe it would have been better if we had freedom to generate our own features instead of trying to de-cypher ours? This at least would have allowed us to try a larger diversity of approaches.

Lynn@Vesta Competition Host • 10 days ago • Options • Reply 0

We encourage and appreciate any kind of feature engineering from kagglers. Meanwhile, we're trying to make clear every feature provided in the dataset, either it's numerical or categorical.
I think more explanation will not limit your freedom to generate new feature or test your idea or approach. However, improper approach applied to the data (e.g. apply one-hot encoding or categorical embedding to numerical) will make the model pointless, even though your CV/PB score gets a bit higher.

snovik • (1774th in this Competition) • 10 days ago • Options • Reply 1

I would tend to disagree. You provided lots of generated features without explaining their logic. I am still not able figure out the logic of C and D features even though you say these are counters and time deltas. On the hand there is clearly lots of structure in the V features which contain lots of counters, flags and cumulative amounts calculated on different criteria. Instead of providing those criteria, or at least give hints, you provided over 300 self-engineered features and hid the original ones.

My point is that instead of trying *to reverse engineer* the logic of pre-calculated features, we might better spend time generating our own or even trying different algos on the raw features.

littlegirl • (222nd in this Competition) • 10 days ago • Options • Reply -1

just waste your time

snovik • (1774th in this Competition) • 10 days ago • Options • Reply 0

@linwangatvesta can you please at least confirm that all pre-calculated features were calculated only on the data set that you provided, i.e. there are no external features which went into the calculation and all that we are missing is a) the logic of calculation and b) additional history?

Lynn@Vesta Competition Host • 10 days ago • Options • Reply 9

Let's see if I can try to answer your questions in the following:

- Pre-calucated features are not only based on entity provided in this dataset, but also based on other entity not included.
- There are many entity features (like full 16-digits card number, full billing name/address, full email_address, phone number, account number, device hashing, etc) we could not provide in the dataset (The reason has been answered in other questions and threads). But they're quite useful so we engineering features (Cx, Dx, Mx, Vx) for you to use in modeling.
- If you're interested in only raw features, you can only use card/addr/emaildomain/identity features and ignore those engineered features.

authman • (372nd in this Competition) • 10 days ago • Options • Reply 10

and then disappear until after the end of the competition.

Thank you!

Aditya Soni • (372nd in this Competition) • 10 days ago • Options • Reply 3

> just waste your time

Data-Science == EDA;

- Modelling is just 10% of what's left.

Tudor Lapusan • (4120th in this Competition) • a month ago • Options • Reply 8

Hi Lynn,

could you put all columns description in official "Data Description" page from this competition. There is almost zero details about the column meanings and I guess not all the people will read this post.

Lynn@Vesta Competition Host • 25 days ago • Options • Reply 4

Good suggestion! I will let Kaggle admin know about it. Thanks!

Addison Howard Kaggle Team • 25 days ago • Options • Reply 4

Updated!

Lynn@Vesta Competition Host • 25 days ago • Options • Reply 0

Thank you, Addison!

YAO ZHONGWEI • (11th in this Competition) • 18 days ago • Options • Reply 0

Is there any chance to shuffle the test set and make public/private random again? I think this approach seems more meaningful to the host.

hmdhmd • (36th in this Competition) • 25 days ago • Options • Reply 1

Hi Lynn, I have question.
Can I make Cxx features from other given columns by myself?

Lynn@Vesta Competition Host • 24 days ago • Options • Reply 2

吾皇万岁万岁万万... • (223rd in this Competition) • 23 days ago • Options • Reply 0

Of couse you can create a New feature

hmdhmd • (36th in this Competition) • 22 days ago • Options • Reply 1

I'm sorry I didn't make it clear enough.

For example,
df['V1'] = np.where(df['D11']==df['D11'], 1, np.nan)
I think that V1 can be created from other column(D11).

Can I create C1~C14 from other columns?

吾皇万岁万岁万万... • (223rd in this Competition) • 22 days ago • Options • Reply 1

you can .

Piyush Paliwal • (1210th in this Competition) • 22 days ago • Options • Reply 1

Everyone makes new features in data science. Create new, delete existing, clean up, etc etc are vital part before feeding to model. I thought it was quite obvious :)

Carlos Miranda • (234th in this Competition) • 19 days ago • Options • Reply 5

I think hmdhmd question is more like "Can we build C1-C14 features from the other features at our disposal? or is there other information that we do not possess in the building of such features?"

kenmatsu4 • (2685th in this Competition) • 8 days ago • Options • Reply 1

I also would like to know the hmdhmd's question. I think the information we'd like to clarify is:

- If Cxx features are calculated from the other columns on train*transaction and train*identity, then essentially there are no additional information. Just feature engineering example.
- However if Cxx features are calculated from the other data source, then we can not get such information from the other columns on train*transaction and train*identity.
- In the latter case, the value of Cxx features is very high, since Cxx features contain additional various information.

Thanks!

VESTA

Lynn@Vesta Competition Host • 8 days ago • Options • Reply 4

Cxx and Dxx are features you can't generate from any other column in this dataset; they're generated from entity we couldn't provide (full card number, full name/address, email address, etc).
I wouldn't recommend removing them unless you can prove they're useless.

kenmatsu4 • (2685th in this Competition) • 7 days ago • Options • Reply 0

ForensicAnalytics • 2 months ago • Options • Reply 20

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?

Melinda Reeb • 2 months ago • Options • Reply 2

That seems like a plausible hypothesis that's worth checking out.

yasagure • (753rd in this Competition) • a month ago • Options • Reply 10

I checked!
I made "the number of decimal places column".
It was different from domains.

For example, data which has specific country domain has a lot of decimal places.
live.fr 3.178571
yahoo.fr 3.398601
yahoo.de 4.108108
but,
gmail 2.050403
gmail.com 1.692461

This may be one reason which support your hypothesis.

I made kernel about it.

https://www.kaggle.com/yasagure/places-after-the-decimal-point-tell-us-a-lot

Randomness • (4549th in this Competition) • a month ago • Options • Reply 0

Valid observation.Might be due to converting transactions with local currency into USD. In the data transaction amount is in USD

Joel Jancarik • (42nd in this Competition) • a month ago • Options • Reply 1

@yasagure I tried to extract decimal points myself in R, but did not find any decimals over 3. So I checked manualy from the csv file line 69 and there are only 3 places in the test dataset.
When checked in the kernel the value looks like 3.0810000000000004, in text file it is 3.081. I am not quite sure what happend, because you are loading the amt as a string, which seems correct to me.

yasagure • (753rd in this Competition) • a month ago • Options • Reply 0

Thanks for nice advice.

I tried to fix this problem in my kernel. Please look at it.

R might be more useful for treating decimal points, when the correctness is needed.

As it happens, fraudulent transactions are 4 times more likely than non fraud transactions to have TransactionAmt values with 3 or more decimal place accuracy. However, when I created a feature that flags up such transactions, I found no improvement in the ROC score, which suggests one of the existing features in the competition data already provides information about what we assume is foreign currency conversion / transaction.

guptapal • a month ago • Options • Reply 3

train*transaction.csv has 590541 row and train*identity has 144233 rows. That means identity information is only available for few transaction IDs? Its getting hard for me to align these two files for information extraction. Please suggest!

Trojan • a month ago • Options • Reply 0

exactly my thought ,i dont know how to go about this

Max Edin • (3629th in this Competition) • a month ago • Options • Reply 0

If you're using R, you can use the merge() function and merge the two data sets using the TransactionID column.

Curious Joe • (3091st in this Competition) • 25 days ago • Options • Reply 0

My take is, transaction.csv contains multiple transactions recorded against single transaction ID. About how to merge these two files, Max Edin has already explained.

subhash • 23 days ago • Options • Reply 0

That's not the case in transactions.csv.@Curious Joe every transaction id is unique in the data set and only few were matching in both transaction and identity dataset.

Marcello Chiodi • (4736th in this Competition) • 16 days ago • Options • Reply 0

R code line: merge(train*transaction,train*id, by="TransactionID", all.x=TRUE)

Sohan Patil • 14 hours ago • Options • Reply 0

for Python(pandas), use -> result = pd.concat([data*identity, data*transactions], axis=1, join='inner')

estelle1728 • (3856th in this Competition) • a month ago • Options • Reply 8

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.

Lynn@Vesta Competition Host • 25 days ago • Options • Reply 0

Yes, there're more than these values.

Yuzy • (2194th in this Competition) • a month ago • Options • Reply 1

Nice work! How did you know these kinds of information?

KJ Low • (2764th in this Competition) • 2 months ago • Options • Reply 35

Hi, use the code snippet below to have a quick overview for all the features in the dataset~

```
for col, values in df_train.iteritems():
    num_uniques = values.nunique()
    print ('{name}: {num_unique}'.format(name=col, num_unique=num_uniques))
    print (values.unique())
    print ('\n')
```

output:
TransactionID: 590540
[2987000 2987001 2987002 ... 3577537 3577538 3577539]

isFraud: 2
[0 1]

TransactionDT: 573349
[86400 86401 86469 ... 15811079 15811088 15811131]

TransactionAmt: 20902
[68.5 29. 59. ... 557.54 1659.95 400.78]

ProductCD: 5
['W' 'H' 'C' 'S' 'R']
...

id_38: 2
[nan 'T' 'F']

DeviceType: 2
[nan 'mobile' 'desktop']

DeviceInfo: 1786
[nan 'SAMSUNG SM-G892A Build/NRD90M' 'iOS Device' ...
'LDN-LX3 Build/HUAWEILDN-LX3' 'Z955A' 'LG-E975']

Sreeram Venkite... • (5048th in this Competition) • 2 months ago • Options • Reply 0

cool!

zhiwei • (4770th in this Competition) • 20 days ago • Options • Reply 0

num_unique=len(num_uniques) ?

Very good!

kongling • (4890th in this Competition) • 15 days ago • Options • Reply

0

nunique = len(unique())

Erik Hanson • 13 days ago • Options • Reply

2

This was helpful. Here's a function I made based on this that only prints the top N values in each column.

```
def df_column_unique_values(df, top_n = 5):
    for col_name, values in df.iteritems():
        col_value_counts = values.value_counts()
        print(f"{col_name} : {len(col_value_counts)}")
        col_value_count_list = [
            "'" + str(c) + "'" + ":" + str(n) for c, n in sorted(
                col_value_counts.items(),
                key=lambda kv: kv[1],
                reverse=True
            )
        ]
        print(", ".join(col_value_count_list[:min(len(col_value_count_list),
top_n)]))
        # print ('\\n')
```

MKP_jp • (760th in this Competition) • 9 days ago • Options • Reply

0

KJ Lowさん
Erik Hansonさん
You are wonderful.

Roman • (56th in this Competition) • 2 months ago • Options • Reply

28

dist: distance

Distance to where?

Rob Mulla • (43rd in this Competition) • 2 months ago • Options • Reply

5

The distance between two points 😂

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

10

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

Are you saying that a single distance column (dist1 or dist2) can hold multiple types of distances -- ie both billing address distance and ip address distance? An ip address distance is very different as mobile gateways obfuscate the true location.

吾皇万岁万岁万万... • (223rd in this Competition) • 21 days ago • Options • Reply 0

he just need to keep as secret because of the privacy rule ,so many information are not accurate.

A Humphrey • (2241st in this Competition) • 2 months ago • Options • Reply 10

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.

But then again, with such a large dataset one can expect to fnd some very improbable coincidences within it.

Lynn@Vesta Competition Host • 2 months ago • Options • Reply 44

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.

A Humphrey • (2241st in this Competition) • 2 months ago • Options • Reply 0

Thanks, this is very useful information.

ML_Bear • (36th in this Competition) • 2 months ago • Options • Reply 3

Hi, Lynn! Thank you for the useful information.

> transactions posterior to it with either user account, email address or billing address directly linked to these attributes as fraud too.

How about transactions **prior** to reported transaction? If those prior transactions which are not fixed as legit transaction (in other words, not passing 120 days), are those transactions also set to isFraud = 1?

Koala • (2046th in this Competition) • a month ago • Options • Reply 2

Hello Lynn,

Thank you for sharing all of the information, much appreciated!

Could you explain a little bit the below please?
"transactions posterior to it with either user account, email address or billing address directly linked to these

email address, and/or billing address, will be considered as fraud = 1 too?

Thanks!

Lynn@Vesta Competition Host • a month ago • Options • Reply 6

It's a complicated situation - usually they will be flagged as fraud. But not all the time afterwards, you can think of one case - the billing address was found to be fraudulent in a past transaction because the credit card associated with it was stolen. But the cardholder is actually the victim, we're not going to blacklist him forever if he uses another legit card for future transaction. There're more other cases but I can't elaborate them all here. One thing we're blacklisting for sure is the card number used for fraud.

A Humphrey • (2241st in this Competition) • a month ago • Options • Reply 0

Without elaborating them all (there are so many features), is it correct to assume that all the obvious (or industry standard) fraud detection metrics are already present in the competition dataset?

Arnab De • (291st in this Competition) • a month ago • Options • Reply 4

The data description on TransactionDT says: "The TransactionDT feature is a timedelta from a given reference datetime (not an actual timestamp)." Can anyone please explain the unit of measurement, e.g. seconds, hours, days, etc for this feature ? And the same for D1 - D5 as well ? Thanks in advance.

Carlos Miranda • (234th in this Competition) • a month ago • Options • Reply 37

Hi! 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.

It looks like D9 is in hours and has been scaled down to 0-1 interval (if you multiply D9 by 24, you get the range 0 to 23).

Nanda • (5463rd in this Competition) • 5 days ago • Options • Reply 0

How did you figure this out T-T

Caneiro • (4753rd in this Competition) • a month ago • Options • Reply 0

Thank you!!

Sateesh • (4519th in this Competition) • a month ago • Options • Reply 1

Hi Lynn,

Would appreciate if you could share inputs on the following attrributes:

M1 to M6 (matched attributes) - Can you pls share more information ?

Noticed that test set don't have instances with "Debit or Credit" ("card6" 's attribute value).

VESTA

Lynn@Vesta Competition Host • 25 days ago • Options • Reply 2

1. credit or debit card transaction
2. like bankname, cardname, etc
3. Mx is attribute of matching check, e.g. is phone areacode matched with billing zipcode, purchaser and recipient first/or last name match, etc.

pocket • (3253rd in this Competition) • 2 months ago • Options • Reply 7

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?)

VESTA

Lynn@Vesta Competition Host • 2 months ago • Options • Reply 17

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

Sridharan Kamal... • (2434th in this Competition) • 2 months ago • Options • Reply 6

Is there a way to identify an account? I can think of combining card1 to 6, and P_emaildomain. Apart from that, are there any other columns in the identity file that would uniquely identify the account?

process531 • (1120th in this Competition) • 14 days ago • Options • Reply 1

same question

aamster • (2601st in this Competition) • 2 months ago • Options • Reply 2

Is the recipient the person receiving the money? Are these all like money transfer transactions? Can you please explain what it means if the recipient email domain is null? There was no recipient or the recipient's email address was not given, or...

VESTA

Lynn@Vesta Competition Host • a month ago • Options • Reply 4

It contains money transfer and also other gifting goods and service, like you booked a ticket for others, etc.
Certain transactions don't need recipient, so R_emaildomain is null.

Alexander Firsov • 2 months ago • Options • Reply 4

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 6

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 1

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

Lynn@Vesta Competition Host • 2 months ago • Options • Reply 0

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 0

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!

Lynn@Vesta Competition Host • 2 months ago • Options • Reply 4

address is of purchaser.

aamster • (2601st in this Competition) • 2 months ago • Options • Reply 1

How can there be only ~300 unique addresses (addr1) but over half a million transactions in the train set?

Alvaro • (895th in this Competition) • 2 months ago • Options • Reply 6

addr1 is the billing zip code (https://www.kaggle.com/c/ieee-fraud-detection/discussion/101203583749). I don't see any reason why the number of addr1 values would be unreasonably low, considering how many countries (addr2) these zip codes belong to.

Use this to see how many instances of each country and zip code combination there are in the train dataset:

```
train.groupby(['addr2', 'addr1']).size()
```

Viraj Bagal • (1305th in this Competition) • 2 months ago • Options • Reply 1

Thanks for sharing this valuable info. I have doubt about addr feature. Does addr1 refer to address of purchaser and addr2 refer to address of recepient ?

Lynn@Vesta Competition Host • 2 months ago • Options • Reply 13

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

Gunes Evitan • (372nd in this Competition) • 2 months ago • Options • Reply 1

Region is city/state/province, am I correct?

Aleksandr Kosol... • (66th in this Competition) • 2 months ago • Options • Reply 2

zipcode

João Nogueira • (426th in this Competition) • 2 months ago • Options • Reply 2

Nice, thank you very much!

Chizuchizu • (119th in this Competition) • 2 months ago • Options • Reply 1

Thank you!

Nice

Shengquan Wang • (2267th in this Competition) • 2 months ago • Options • Reply 2

prefect

Ang Li • 2 months ago • Options • Reply 3

whats about the work?

cristinasitiki • 2 months ago • Options • Reply 0

yes,it s very great

Ants • (1351st in this Competition) • 2 months ago • Options • Reply 2

Thanks for the insight! That helps knowing how to treat the columns.

Is there a way to reliably group transactions for the same purchaser or recipient?

Rahul • (1415th in this Competition) • 2 months ago • Options • Reply 2

Thanks for the description, I was just dropping most of the columns before, I'm keen to see what the Vesta engineered features are and how to use them.

Nirjhar Roy • (2400th in this Competition) • 2 months ago • Options • Reply 1

exact same thought ..

yohei • (2373rd in this Competition) • 2 months ago • Options • Reply 0

Thanks!

Elliot • (1660th in this Competition) • 3 days ago • Options • Reply 0

Great Job!

Sunny gupta • 6 days ago • Options • Reply 0

Hi Lynn, I want to reduce number of variable by combining according to their features, so can you just tell me about Vxx variable , like some Vxx variable are ranking, so i will choose '1 'with the highest value but some Vxx variable has rating, so i will

Mohit Singh • (511th in this Competition) • 6 days ago • Options • Reply

0

Thank you

psoni • (2065th in this Competition) • 9 days ago • Options • Reply

0

Any reason, why ROC score is different than score which is getting calculated while submitting the prediction results?
for example, for me ROC is 98% but after submitting score is ~92%.

sl1me - OedaLab • 13 days ago • Options • Reply

0

I've noticed that identity features like id*06, id*08 have a spike of -100 value in histogram.
Those columns already has NaN, so -100 can have different meaning.
So can someone tell me what -100 exactly mean?
Or, can we treat -100 as a part of NaN?
Thanks in advance.

VESTA

Lynn@Vesta Competition Host • 10 days ago • Options • Reply

0

I would not suggest making -100 as NaN

Shiu-Tang Li • (1162nd in this Competition) • 13 days ago • Options • Reply

0

Thanks for sharing.

Hi Im Ryo • (1516th in this Competition) • 15 days ago • Options • Reply

0

This is very special!

xiechunhui • 15 days ago • Options • Reply

0

thank you very much, nice work!

Nanashi • (616th in this Competition) • 15 days ago • Options • Reply

0

@linwangatvesta thanks for your help :)
my question is: kaggle only splitted the dataset (train and test), they didn't encoded nothing or changed the data... right?
thanks in advance.

VESTA

Lynn@Vesta Competition Host • 15 days ago • Options • Reply

0

yes, encoding is on our side, kaggle didn't do anything on that

wa007 • (4649th in this Competition) • 16 days ago • Options • Reply 0

What is Vesta...
Who can get me some information about Vesta ?

Ayush Goel • (3538th in this Competition) • 15 days ago • Options • Reply 1

I think this is the link to their website: https://trustvesta.com/ . I think they provide solutions for fraud detection and provides secure payment solutions.

VESTA

Lynn@Vesta Competition Host • 14 days ago • Options • Reply 2

Welcome to visit our website

wa007 • (4649th in this Competition) • 14 days ago • Options • Reply 0

thanks !

haoran.yang • (4011th in this Competition) • 19 days ago • Options • Reply 0

Valuable information!

yuto16 • (2772nd in this Competition) • 19 days ago • Options • Reply 0

Thank you for your explanations!

Hxr • (675th in this Competition) • 20 days ago • Options • Reply 0

cool

Shalin • (2717th in this Competition) • 21 days ago • Options • Reply 0

Lynn,
Are some of the Vesta engineered features categorical? V1, V14, V41, V65, V305 etc have just two unique values. Also it would help if they were classified as either ordinal or nominal. Thanks

VESTA

Lynn@Vesta Competition Host • 17 days ago • Options • Reply 0

They were derived as numerical

igman • (701st in this Competition) • 22 days ago • Options • Reply 0

imagination • (3191st in this Competition) • 24 days ago • Options • Reply 0

I see some discussion about foreign transction, is any column represent transaction in which foreign country?
If it is online shoppoing, then any column represent it happen in which website ?

Lynn@Vesta Competition Host • 24 days ago • Options • Reply 0

card feature contains which country of the card is issued; billing address contains which country of the cardholder.
There's no website information included.

Keisuke Daimon • (3869th in this Competition) • a month ago • Options • Reply 0

- DeviceType
- DeviceInfo
- id*12* - *id*38

Could you explain the difference between id*01* - *id*11 and id*12 and id*38? I am wondering why id*01* - *id*11 are not mentioned.

Lynn@Vesta Competition Host • 25 days ago • Options • Reply 3

id01 to id11 are numerical features for identity, which is collected by Vesta and security partners such as device rating, ip_domain rating, proxy rating, etc. Also it recorded behavioral fingerprint like account login times/failed to login times, how 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.

Joel Jancarik • (42nd in this Competition) • a month ago • Options • Reply 0

Is a 3.5% real Fraud rate? It seems to be too high.

Max Edin • (3629th in this Competition) • a month ago • Options • Reply 0

2-3% is what the fraud rate generally is with credit card transactions, so it's not unrealistic. Looking at how the fraud is distributed over e.g. card1 looks realistic too. I don't think they've added extra fraudulent transactions.

Lynn@Vesta Competition Host • 25 days ago • Options • Reply 2

In real it's slightly lower, because to make labeling more accurate, we removed a portion of Vesta denied transactions already. But we haven't added any extra fraudulent transactions.

Debarpan Dhar • a month ago • Options • Reply 0

Arman Kuzembayev • a month ago • Options • Reply 0

Thanks

Mark Sandal • (4290th in this Competition) • a month ago • Options • Reply 0

Oh, thanks. Was tired of guessing what these mysterious features might actually be.

Fadi Badine • a month ago • Options • Reply 0

Thanks for the description

Owen Xu • (1853rd in this Competition) • a month ago • Options • Reply 0

Thanks a bunch for such great detailing post. I have one question regarding address. Typically when you use your card at some store, your transaction should include the store address as I have experienced. Is this information included in any address feature as well? Cheers.

Lynn@Vesta Competition Host • 25 days ago • Options • Reply 1

These are all online payment transactions (CNP). store address was not included in the dataset.

Bryan • (2375th in this Competition) • a month ago • Options • Reply 0

This was extremely helpful, thank you!

Kirill • (2443rd in this Competition) • a month ago • Options • Reply 0

Thank, I am joining the competition.!!

Hxr • (675th in this Competition) • a month ago • Options • Reply 0

thanks

季风 • (734th in this Competition) • a month ago • Options • Reply 0

nice topic ,thanks

Randomness • (4549th in this Competition) • a month ago • Options • Reply 0

Hi Lynn, can we assume purchaser email domain will be same as card holder email domain.

Yes, you could.

CinKate • (1703rd in this Competition) • a month ago • Options • Reply 0

Thank you very much!

Joel Jancarik • (42nd in this Competition) • a month ago • Options • Reply 0

Hi, how was the split between train and test data performed? Is it just a random sample or is there any other difference worth mentioning?

aamster • (2601st in this Competition) • a month ago • Options • Reply 0

Max Edin • (3629th in this Competition) • a month ago • Options • Reply 2

If you merge the train and test data sets and plot the data over transactionDT you can see that the data is split in time. I.e. that the test data is a continuation from the train data.

Lynn@Vesta Competition Host • 25 days ago • Options • Reply 1

Test dataset contains payment transactions posterior to Train dataset. Public/private set was split by random though.

Arturo Garcia • (29th in this Competition) • 24 days ago • Options • Reply 0

Hi Lynn. Thank you for your insights. But there are a couple of discussions where is probed that the Test set is split by time. Are they wrong?

Lynn@Vesta Competition Host • 24 days ago • Options • Reply 1

I read them discussing train/test split, and yes it was split by time; but public/private was split random.

Arturo Garcia • (29th in this Competition) • 24 days ago • Options • Reply 3

Thank you again for answering me so quick.

I was talking about this two ones:

https://www.kaggle.com/c/ieee-fraud-detection/discussion/105788#latest-608190
https://www.kaggle.com/c/ieee-fraud-detection/discussion/101040

They create a fake submission with all zeros in the last 80% percent of the test data and the public LB doesn't change.

Exactly, what Lynn@Vesta claims here about "random" public/private split doesn't sound to be true. Guys have tested by submissions. I wonder everything else mentioned is true.

Lynn@Vesta Competition Host • 24 days ago • Options • Reply 2

Hmm, that's odd. I'm told by Kaggle team that public/private was split by random, and that's kind of our intention. But Vesta only provides the dataset and Kaggle team takes care of splitting. I wish I hadn't spread incorrect information, maybe Kaggle team could make the clarification (or maybe competition policy won't allow them to reveal that, I don't know).

authman • (372nd in this Competition) • 23 days ago • Options • Reply 3

> I wish I hadn't spread incorrect information

It's all good; I think more people probably saw and are aware of the existing public/private split discussion topics and kernels than those who have saw your note on the split being random. Besides, your mention was only up for a day before you corrected.

David J. Slate • (1518th in this Competition) • 22 days ago • Options • Reply 1

I just joined the competition and discovered the news that the public/private split was not random but was based entirely on transaction time. It seems quite strange that Kaggle would do that not only without consulting Vesta but apparently, at least according to Lynn, without even informing them. Sounds like there was some serious miscommunication between host and sponsor on this point.

I'm a veteran of many Kaggle competitions, and a large fraction of them have suffered from one or more goofy issues, including data leaks, unintentional "magic" features, etc. Some are annoying; some are just entertaining. It's a good thing that Kagglers are generally alert and clever enough to uncover most of them.

CPMP • (2389th in this Competition) • 17 days ago • Options • Reply 11

@linwangatvesta

> I read them discussing train/test split, and yes it was split by time;
>
> Hmm, that's odd. I'm told by Kaggle team that public/private was split by random, and that's kind of our intention. But Vesta only provides the dataset and Kaggle team takes care of splitting. I wish I hadn't spread incorrect information, maybe Kaggle team could make the clarification (or maybe competition policy won't allow them to reveal that, I don't know).

It is great that public/private is split by time and not randomly. otherwise you would see kagglers do LB probing and the models you'd get would not be worth a dime.

inversion Kaggle Team • 17 days ago • Options • Reply 11

@cpmpml is very much correct. We always split time-based data by time for Train/Test, and Public/Private. If we didn't do that, the Public leaderboard score would provide feedback (leakage) for future time points.

Lynn@Vesta Competition Host • 17 days ago • Options • Reply 1

CPMP • (2389th in this Competition) • 15 days ago • Options • Reply 1

@inversion thanks.

Mohamad Falah • (1639th in this Competition) • a month ago • Options • Reply 0

Thanks, this is very useful information.

YoonSoo • (14th in this Competition) • a month ago • Options • Reply 0

Thanks! It will help us engineer features.

Sri Yogesh • (439th in this Competition) • a month ago • Options • Reply 0

Thanks a ton!

ddemo • (5435th in this Competition) • 2 months ago • Options • Reply 0

how to deal with missing TransactionID(s) in test dataset?

Filip Noworolnik • (2600th in this Competition) • 2 months ago • Options • Reply 1

I don't think it's possible for TransactionID to be missing... The submission score is based on those values and isFraud column, that is your predictions... Try to download test csv again...

ddemo • (5435th in this Competition) • a month ago • Options • Reply 0

Hi Filip, thank you for your reply.i downloaded it again but, there are missing transactionid values (the one which we use in submission that too, more than 4L transaction ids are missing in test*identity.csv when compared to test*transactions.csv
thanks in advance.

Filip Noworolnik • (2600th in this Competition) • a month ago • Options • Reply 0

If I'm not wrong, your question is already answered in AmirH comment. There is just no such an information :)

dashinggirl • (4401st in this Competition) • 2 months ago • Options • Reply 0

thanks

Thanks ! this description will be very helpful in feature engineering

aamster • (2601st in this Competition) • 2 months ago • Options • Reply

0

Can you please be more specific on the Vxx features? How would these counts be different from C1-15 for example? What kind of ranking?

aamster • (2601st in this Competition) • 2 months ago • Options • Reply

0

Can you please be more specific on D1-15? Are these timedelta using different attributes of the transaction, such as time since email address was used, time since address was used, etc?

Lynn@Vesta Competition Host • a month ago • Options • Reply

1

Yes, you've got their meaning, time since xxx last time used or first time used, etc.

aamster • (2601st in this Competition) • 2 months ago • Options • Reply

0

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.

Lynn@Vesta Competition Host • a month ago • Options • Reply

2

Your guess is good, plus like device, ip*addr, billing*addr, etc. Also these are for both purchaser and recipient, which doubles the number.

Sogna • (31st in this Competition) • 2 months ago • Options • Reply

0

Thanks

sssjunjun • (1247th in this Competition) • 2 months ago • Options • Reply

0

@linwangatvesta Thank you for providing this description. I have one question.
I noticed for each line on ProductCD column there is only one value. Does that mean purchaser only buy one category of goods for one order? That sounds impossible

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

2

Well, actually 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.

Thanks a lot for these precious informations !!

MuzafferSenkal • 2 months ago • Options • Reply 0

Thanks

kce330 • (4743rd in this Competition) • 2 months ago • Options • Reply 0

thanks!

EmiY • (1656th in this Competition) • 2 months ago • Options • Reply 0

Thank you so much!Really helpful!

Maksim Drobchak • 2 months ago • Options • Reply 0

cool!

ummar • 2 months ago • Options • Reply 0

really wonderful

Bill Miao • (170th in this Competition) • 2 months ago • Options • Reply 0

Cool, thanks!

Richard Frost • 2 months ago • Options • Reply 0

Thanks

Sur • (3754th in this Competition) • 2 months ago • Options • Reply 0

@linwangatvesta
Hi Lynn could you eloberate little more about Vxxx features ?

Lynn@Vesta Competition Host • 2 months ago • Options • Reply 4

Like what description says, it includes 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.

aamster • (2601st in this Competition) • 2 months ago • Options • Reply 0

How are these different from C1-15? Please be more specific?

Ezekizel • 2 months ago • Options • Reply 0

nice

Mohit Sharma • 2 months ago • Options • Reply 0

Great!!

Big Si • 2 months ago • Options • Reply 0

Thank you!

HONG_YP • 2 months ago • Options • Reply 0

great work thanks.

T.U.V • (10th in this Competition) • 2 months ago • Options • Reply 0

Thank you very much, information you shared is very helpful👍

Jolayemi olubunmi • 2 months ago • Options • Reply 0

Nice. Thank you

Abhishek Sankar • 2 months ago • Options • Reply 0

How to get started?

Ilham Firdausi P... • (1697th in this Competition) • 2 months ago • Options • Reply 3

Some public kernel on the Kernels tab provides a good starting point

Siddharaj Junnarkar • 2 months ago • Options • Reply 0

Thanks

batu bayraktar • 2 months ago • Options • Reply 0

thanks !

Jianfei Wang • (40th in this Competition) • 2 months ago • Options • Reply 0

so one question is, does D* mean the value for the same user/card ? or somthing else?

Lynn@Vesta Competition Host • 2 months ago • Options • Reply 0

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

Randomness • (4549th in this Competition) • a month ago • Options • Reply 0

Hi Lynn,

I want to understand D1-D15 calculations bit more. Please help me.

Measurement Unit:
`Unit` of D1-D15 is in days.

Definition of D1-D15:
Let's say current transaction is C ; Most recent historical transactions are MR1,MR2,MR3.....

D1 - Difference between latest & most recent transaction in history (MR1)
D2 - Difference between Most recent (MR1) & immediate previous transaction in history (MR2)
....
If so, why we have scenario as D1(entry), D2 (NA),D3(entry),D4(NA),.....Please refer first transaction in image.

Thanks,
Siva Kumar

D1-D15.PNG (41.63 KB)

Carlos Miranda • (234th in this Competition) • a month ago • Options • Reply 1

Hi Siva Kumar,
Using your notations and your example, one can assume missing values can come from unkown values for MR2, MR3, etc, e.g. if both C and MR1 are known, the value is computed, but if MR2 is not present in the dataset, then NaN is used, thus C - MR2 would give NaN.

Thinking of broader definitions of D# features, one can come up with "time difference between current and last transaction **issued in that country/region**", or more specific ones like "time difference between current and last transaction **issued in that country/region for the given card holder**" and "time difference between current transaction and last recorded transaction **with higher amount for the given card holder**".

Features meanings can be very specific and not related to each other, which leads to sparsity and the cases you have pointed out.

Eric Vos • (350th in this Competition) • 2 months ago • Options • Reply 0

Many thanks Lynn,
These information are useful.

Ilham Firdausi Putra • (1697th in this Competition) • 2 months ago • Options • Reply 0

So what's the difference between addr1 and addr2?

May be one gives country and other gives the street.

Roman S. Shche... • 2 months ago • Options • Reply 0

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

VESTA

Lynn@Vesta Competition Host • 2 months ago • Options • Reply 8

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

23 days ago

This Comment was deleted.

a month ago

This Comment was deleted.

a month ago

This Comment was deleted.

2 months ago

This Comment was deleted.