+) Variables no missing value :

Transaction characteristics : ID, time, amount, productCD -> obliviously

isFraud = Y -> obliviously

card 1 -> maybe present the obligatory card number

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

+) In general, the proportion of missing values of a feature on the training set and the one on the test seta are pretty equals, except:

-) C14 has no missing on training but not on test set

-) D and C variables

+) Fraud transactions present very small proportion in total data -> imbalanced problem

+) Transaction Amount

Same distribution on train and test

has significantly impact on Fraud : Distribution of log TA for Fraud observation has a heavier tail.

+) ProductCD has impact on Fraud.

+) For each card variables, the number of categories on train set is higher than on test set

+) Card4 missing value can be replaced by the value of the group generated by card 1 as the key

+) 88% of card 3 == ‘150’ -> used to replace missing

+) Addr1 and Addr2 are in pair

+) Addr1 has impact but no special pattern.

+) 88% Addr2 == ‘87’ -> missing replacement. For Addr2 = ‘65’ the Fraud rate is very high = 60%

+) dist1 has ~60% of NaN, dist2 >90% of NaN in both train and test.

-) Discard dist2

-) No clearly pattern for the distribution of dist1 by Fraud. However, the percentage of Fraud is 2 times higher while dist1 is missing. Transforming dist1 to categorical variable (1 if missing, 0 if not)

+) We merge P(R)\_emaildomain by the company (ex: hotmail.es = hotmail.fr) only if the frequency of one group is too low.

+) Logically, if Receipt email is missing, that maybe means that purchaser and receiver are the same person -> replace missing R\_email by P\_email, create a new dummy variable to record this replacement.

+) There are a lot of email domain which contain abbreviation ‘mx’ (or Mexico) -> create new 0-1 variable to specify that -> there are a slightly impact by this new variable. Probability of Fraud: 2.5% vs 3.5%. However, the proportion of that new variable is very imbalanced (1 : <1%), so we delete it.

+) ‘Gmail’ is the most common class. Interesting thing is that the class "mail.com" has high probability of Fraud.

+) M1, M2, M3 are triplet, if one of each is missing, the other are also missing.

+) M8, M9 are in pair

+) For M feature, the probability of Fraud in the missing value is highest among all class (expect M4) -> Create Miss group for missing value

+) After transforming timeDelta Feature, we observe that the frequence

Card1

**Transaction variables**

* 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 - Transaction**

* ProductCD
* emaildomain
* card1 - card6
* addr1, addr2
* P\_emaildomain
* R\_emaildomain
* M1 - M9

**Categorical Features - Identity**

* DeviceType
* DeviceInfo
* id\_12 - id\_38