Objective

The objective of our project is to predict which clients of Vesta Corporation, the world’s leading payment service company, are likely to be fraud, based on their transaction histories and personally identity information.

It will help us improve the efficacy of fraudulent transaction alerts for millions of people around the world, helping hundreds of thousands of businesses reduce their fraud loss and increase their revenue. And of course, can help save people from false positives.

Data sets:

The datasets from Vesta Corporation have two parts: ‘Identify’ and ‘Transaction’.

Since the target variable is in Transaction data, the identity data is left join to it.

After joined by ID, there are 600K observations and 433 features, including 393 features of transaction and 41 features of identity.

Before rush to the analysis, we check basic situation of datasets:

First, there are no duplicated observations in original datasets.

Second, most features, including transaction time, are anonymous.

Third, not all transactions observations have corresponding identity information.

Finally, the primary key, ‘TransactionID’, is unique.

EDA:

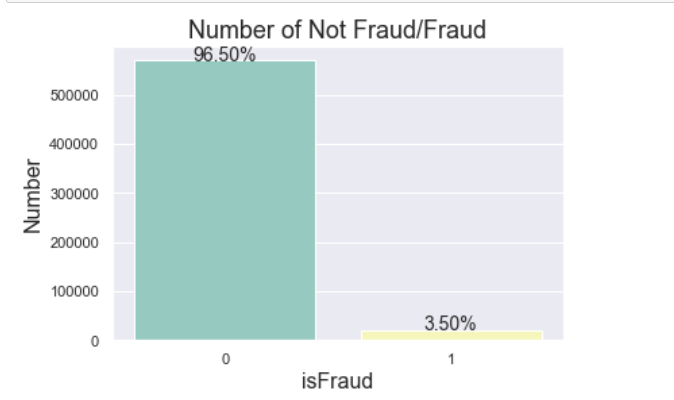
The quality of data is not good.

(show plot first)

Firstly, most features have missing values and there are large amount of feature with high percentage of missing.

Among 433 features, we find that only 18 features are without missing values. Plenty features have more than 70~80% missing values. There are even 74 features with more than 80% missing values and 12 columns with more than 90% missing data.

Secondly, there exist severe imbalance of target variables



Thirdly, we explore data type and meanings of some anonymous variables and find some interesting factors. For example, most transaction is of small amount less than 400, the visa takes up 65% of the transactions and Mastercard takes up 35% of the transactions. You can find out more interesting facts in our report.