Fraud Detection & Predication on Bank Payments

Fraudulent behavior can be seen across many different fields such as e-commerce, healthcare, payment and banking systems. Fraud is a billion-dollar business and it is increasing every year. The PwC global economic crime survey of 2018 [1] found that half (49 percent) of the 7,200 companies they surveyed had experienced fraud of some kind.

Even if fraud seems to be scary for businesses it can be detected using intelligent systems such as rules engines or machine learning. Most people here in Kaggle are familier with machine learning but for rule engines here is a quick information. A rules engine is a software system that executes one or more business rules in a runtime production environment. These rules are generally written by domain experts for transferring the knowledge of the problem to the rules engine and from there to production. Two rules examples for fraud detection would be limiting the number of transactions in a time period (velocity rules), denying the transactions which come from previously known fraudulent IP's and/or domains.

Rules are great for detecting some type of frauds but they can fire a lot of false positives or false negatives in some cases because they have predefined threshold values. For example let's think of a rule for denying a transaction which has an amount that is bigger than 10000 dollars for a specific user. If this user is an experienced fraudster, he/she may be aware of the fact that the system would have a threshold and he/she can just make a transaction just below the threshold value (9999 dollars).

For these type of problems ML comes for help and reduce the risk of frauds and the risk of business to lose money. With the combination of rules and machine learning, detection of the fraud would be more precise and confident.

Banksim dataset

We detect the fraudulent transactions from the Banksim dataset. This synthetically generated dataset consists of payments from various customers made in different time periods and with different amounts. For more information on the dataset you can check the Kaggle page for this dataset which also has the link to the original paper

Data As we can see in the first rows below the dataset has 9 feature columns and a target column. The feature columms are :

- **Step**: This feature represents the day from the start of simulation. It has 180 steps so simulation ran for virtually 6 months.
- **Customer**: This feature represents the customer id
- **zipCodeOrigin**: The zip code of origin/source.
- **Merchant**: The merchant's id
- **zipMerchant**: The merchant's zip code

- **Age**: Categorized age
    - 0: <= 18,
    - 1: 19-25,
    - 2: 26-35,
    - 3: 36-45,
    - 4: 46:55,
    - 5: 56:65,
    - 6: > 65
    - U: Unknown
- **Gender**: Gender for customer
    - E : Enterprise,
    - F: Female,
    - M: Male,
    - U: Unknown
- **Category**: Category of the purchase. I won't write all categories here, we'll see them later in the analysis.
- **Amount**: Amount of the purchase
- **Fraud**: Target variable which shows if the transaction fraudulent(1) or benign(0)