

TECH ALPHARETTA

2972 Webb Bridge Rd, Alpharetta, GA 30009

**Anomaly Detection**

This challenge is quite tough, and it might push you out of your comfort zone.

You'll need to train an unsupervised model to detect potentially fraudulent credit card transactions.

This technique, also known as anomaly detection, is frequently used in finance, meteorology, and even cyber security.

**Background**

Detecting fraud is like playing cat and mouse. Criminals are always looking for new ways to commit fraud, while banks must try to stay one step ahead to protect their customers.

However, if you train supervised classification model on past incidents, your model might not be able to catch new forms of fraud.

This is where anomaly detection comes in.

**Anomaly detection** reframes of problem. Instead of trying to classify transactions as fraudulent or not, you'll train an unsupervised model that learns what's "normal."

Transactions that stray too far from the norm, or outliers, will then raise alerts.



You live in California, and someone swipes your card in Siberia... Hmm...

Just to be clear, anomaly detection doesn't *need* to be unsupervised. However, the context of many anomaly detection exercises will require you to use unsupervised techniques.

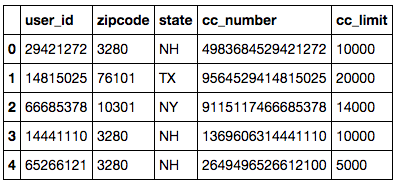
For example, your boss may simply give you the latest month of credit card transaction data and ask you to detect if something looks weird.

In this case, you wouldn't have any labels for past fraud, which would prevent you from performing supervised classification.

**Data**

The first table is called **cc\_info.csv**.

It account details for 843 credit cards from a major bank.

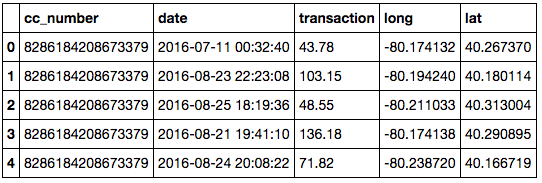


*Data Dictionary:*

* **user\_id -**Unique ID of customer.
* **zipcode** - Home zipcode of customer.
* **state** - Home state of customer.
* **cc\_number** - Anonymized credit card number.
* **cc\_limit** - Spending limit on the credit card per calendar month.

The second table is called **tx\_data.csv**.

It contains 230,000 transactions for those 843 credit cards.



*Data Dictionary:*

* **cc\_number -**Anonymized credit card number.
* **date** - Datetime of transaction.
* **transaction** - Transaction amount in USD.
* **long** - Longitude of transaction location.
* **lat** - Latitude of transaction location.

**Objectives**

You've been asked to build a few tools to protect the bank's customers from potential fraud and from overspending.

* Use an unsupervised algorithm to return a list of all anomalous transactions that may be worth investigating further.
  + Tip: If you go with a clustering algorithm, experiment with different sets of input features.
* Write a function that can be run every day to detect if any customer exceeds his/her monthly limit.
  + Tip: One way to implement the function would be to use 4 arguments:
    - user\_id
    - date (default would be today)
    - DataFrame of credit card info
    - DataFrame of transactions
* Return a list of customers who went over their monthly limit in the most recent month.