Identifying Fraudulent Activities

Goal

E- commerce websites often transact huge amounts of money. And whenever a huge amount of money is moved, there is a high risk of users performing fraudulent activities, e.g. using stolen credit cards, doing money laundry, etc.

Machine Learning really excels at identifying fraudulent activities. Any website where you put your credit card information has a risk team in charge of avoiding frauds via machine learning.

The goal of this challenge is to build a machine learning model that predicts the probability that the first transaction of a new user is fraudulent.

Challenge Description

Company XYZ is an e-commerce site that sells hand-made clothes.

You have to build a model that predicts whether a user has a high probability of using the site to perform some illegal activity or not. This is a super common task for data scientists.

You only have information about the user first transaction on the site and based on that you have to make your classification ("fraud/no fraud").

These are the tasks you are asked to do:

- For each user, determine her country based on the numeric IP address.
- Build a model to predict whether an activity is fraudulent or not. Explain how different assumptions about the cost of false positives vs false negatives would impact the model.
- Your boss is a bit worried about using a model she doesn't understand for something as important as fraud detection. How would you explain her how the model is making the predictions? Not from a mathematical perspective (she couldn't care less about that), but from a user perspective. What kinds of users are more likely to be classified as at risk? What are their characteristics?
- Let's say you now have this model which can be used live to predict in real time if an activity is fraudulent or not. From a product perspective, how would you use it? That is,

what kind of different user experiences would you build based on the model output?

Data

We have 2 table downloadable by clicking

here. The 2 tables are:

```
"Fraud_Data" - information about each user first transaction
```

Columns:

- **user_id** : Id of the user. Unique by user
- **signup_time** : the time when the user created her account (GMT time)
- **purchase_time** : the time when the user bought the item (GMT time)
- **purchase_value** : the cost of the item purchased (USD)
- **device_id** : the device id. You can assume that it is unique by device. I.e., 2 transactions with the same device ID means that the same physical device was used to buy
- **source** : user marketing channel: ads, SEO, Direct (i.e. came to the site by directly typing the site address on the browser).
- **browser** : the browser used by the user.
- **sex** : user sex: Male/Female
- **age** : user age
- **ip_address** : user numeric ip address
- **class** : this is what we are trying to predict: whether the activity was fraudulent (1) or not (0).

```
    "IpAddress_to_Country" - mapping each numeric ip address to its country.
For each country, it gives a range. If the numeric ip address falls within
the range, then the ip address belongs to the corresponding country.
```

Columns:

- **lower_bound_ip_address** : the lower bound of the numeric ip address for that country
- **upper_bound_ip_address** : the upper bound of the numeric ip address for that country
- **country** : the corresponding country. If a user has an ip address whose value is within the upper and lower bound, then she is based in this country.

Example

```
    Let's check the first user activity
```

head(Fraud_Data,1)

Column Name	Value	Description
user_id	22058	this is the user id
signup_time	2015-02-24 22:55:49	this user signed up on the site on Feb, 24, at 10:55 and 49sec PM GMT
purchase_time	2015-04-18 02:47:11	his first transaction happened on April, 18 at 2:47 AM GMT
purchase_value	34	the item he bought cost 34$
device_id	QVPSPJUOCKZAR	this is his device id
source	SEO	came to the site via SEO
browser	Chrome	was using Google Chrome
sex	M	he is a Male
age	39	he is 39 y/o
ip_address	732758369	this is his numeric ip address
class	0	his activity was not fraudulent

```
    Let's check where that user is based. Its IP address should fall within
 some range in the IpAddress_to_Country table.
```

subset (IpAddress_to_Country, lower_bound_ip_address <= 732758369 & upper_bound_ip_address>= 732758369)

Column Name	Value	Description
lower_bound_ip_address	729808896	this the lower bound of the range that included that user numeric ip address
upper_bound_ip_address	734003199	this the upper bound of the range that included that user numeric ip address
country	Japan	the user is based in Japan since we found out that his ip address falls within the Japanese range