ITM885 Case Study 2: Identifying Fraudulent Activities

In this case study, we want to implement feature engineering and create features with high predictability.

Company XXX 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:
1. For each user, determine her country based on the numeric IP address.
2. 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.
3. 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?
4. Let's say you now have this model which can be used to predict in real time if an activity is fraudulent or not. From a product perspective, how would you use it?

We have two data sets. “Fraud_data.csv” contains information about each user’s first transaction. Concretely,
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.csv” is 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. Concretely,

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.