3.2 Descriptive Analytics for Fraud Detection

3.2.1 Descriptive vs Predictive #Descriptive #Predictive

Descriptive Analytics	*Predictive Analytics*
Aims at finding anomalous behavior deviating from the norm	Assumes the availability of historical data with known frauds
● Behavior of the *average customer* across a time period.	● The analytical models built *can only detect known fraud patterns*.
● The *average behavior* of a given customer across a time period.	● Useful to *explain the anomalies* found by descriptive analytics.
Can identify emerging unknown cyber threats.	**Are based on known fraudulent patterns.**
Unsupervised Learning	***Supervised Learning***

3.2.2 Descriptive Analytics = Unsupervised learning = Anomaly Detection

It aims at finding anomalies, suspicious observations, outliers or exceptions.

> **Grubbs (1969)**
>
> An **outlying observation**, or outlier, is one that appears to *deviate markedly from other members of the sample in which it occurs*.

Relevant in environments where:

- Organizations are *starting* doing fraud detection.
- There is *no labeled historical data* set available.
- *Fraudsters are* continuously *adapting* their strategies.

3.2.2.1 Unsupervised Learning Challenge

Define the average behavior or norm:

- *Depend* on the application field.
- *Boundary* between norm and outliers *is not clear-cut*.
  - Fraudsters try to blend into norm
- The *norm may change* over time

    - Analytical models built need to be continuously monitored and updated.
- *Anomalies do not necessarily represent frauds*.

> ***Unsupervised learning for fraud detection*** *require extensive validation of the identified suspicious observations.*

3.2.3 Graphical outlier detection procedures

Ideal tools to explore the data and get preliminary insights:

- **One-dimensional outliers** (*histogram or box plot*).
- **Two-(three) dimensional outliers** (*scatter plot*).

👎🏽 Disadvantages:

- *Less formal* and only limited to a few dimensions.
- *Require active involvement* of the end-user.
- For a large dimensional data set, is cumbersome.

Next chapter: Statistical Outlier Detection