# Machine Learning in Fintech : Financial Fraud Detection

**Anass Akrim**
Ecole des Mines de Saint-Etienne, France
anass.akrim@etu.emse.fr

## Introduction :

Fraud detection is an active field of research due to increasing fraud risk. The fight against frauds became a major issue with the development of organizations : insurance frauds, credit card frauds...
Therefore, it becomes necessary to equip itself with the tools to face this threat and find a solution to detect, manage frauds.

## Aim :

➢Semi-supervised approach : apply and implement a semi-supervised method using firstly labelled data.

➢Real-time fraud detection : detect outliers or anomalies (semi) automatically

## Objectives :

Put in place a **robust** method capable of identifying financial fraud in real-time, or anomaly in a data set collected, through semi-automatic learning :

✓ Set a **robust real-time** methodology (in **Deep Learning**), while calibrating and optimizing the predictive model built.

✓ Comparison of the method used with **unsupervised** statistical learning algorithms.

## Materials and methods :

- **Dataset :**

Dataset from a European cardholder from September 2013 (source : kaggle.com/datasets), transactions that occurred in two days : **284, 807 transactions**, including **492 frauds**. **31 features**.

We consider two subsets, so we can compare the **robustness** of the methods:
-The **'small' dataset** : 15.000 transactions and 492 frauds.
-The **'big' dataset** : the whole dataset.

- **Deep Learning Semi-supervised technique :**

Restricted Boltzmann Machines (RBM), able to detect outliers in **real-time**.
Trained only on normal (genuine) transactions set, the model will **learn the pattern** of normal transactions, can **reconstruct what non fraudulent transactions looks like** and learns how to **discriminate** whether or not new transactions belong to that same class.

- **Unsupervised Statistical Learning algorithms:**

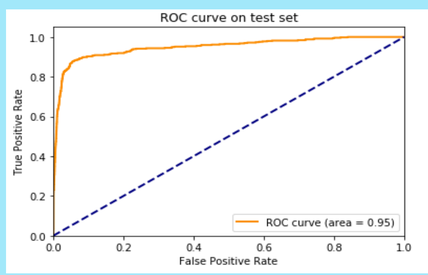
They use **distances** or **densities** to estimate what is normal and what is aberrant.
- Clustering by density (DBSCAN, OCSVM)
- Clustering by 'isolation' (Isolation Forest).

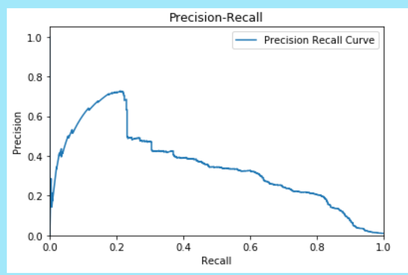
## Results :

Comparison between the proposed model (**RBM**) and three studied models (**DBSCAN, OCSVM and Isolation Forest**) :

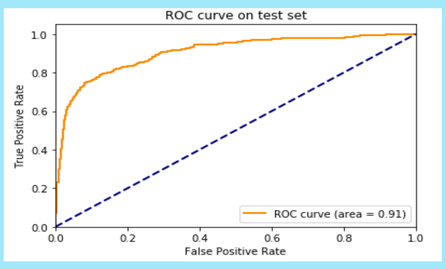
| Dataset | Evaluation metrics | DBSCAN | OCSVM | Isolation Forest | RBM |
|---|---|---|---|---|---|
| 'Small' Dataset | Precision | **0.55** | 0.29 | 0.41 | 0.41 |
| | Recall | **0.84** | 0.59 | 0.83 | 0.83 |
| 'Big' Dataset | Precision | 0.10 | 0.10 | 0.12 | **0.15** |
| | Recall | 0.83 | 0.60 | 0.71 | **0.84** |

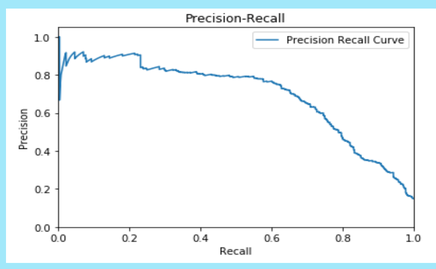
Comparison based on the **Precision** (the percentage of truly known fraudulent transactions to transactions classified as fraudulent by the model) and the **Recall** (the percentage of transactions correctly classified as fraudulent by the model).




RBM model's performance for 'Big' Dataset




RBM model's performance for 'Small' Dataset

## Conclusion and perspectives :

RBM is a **semi-supervised** technique which is a very good tool for **real-time fraud detection** (possibility to train the model with new genuine data and improve the accuracy of the model), while the statistical learning algorithms used are built for an unsupervised outlier detection.

RBM seems to perform very well on large and small data set (**detects almost 84% of the frauds, ie. 413 over 492 frauds**), which is not the case for clustering algorithms used : very good with small data sets but not that good with larger ones (poor precision mostly) : **lack of robustness**.

**Perspectives** :
❑ Tune the hyper parameters of the RBM : make it **more robust**.
❑ Interesting to conduct future work on **other datasets** : verify the **generalization** of the results presented by this study.