DATAVISOR

The Dummy Handbook on MACHINE LEARNING FOR FRAUD DETECTION

Machine learning (ML) is fascinating. In fact, it might be one of the most interesting fields of knowledge out there today. If you want to see what we mean, check out these incredible examples. From music to financial services, ML is here to stay.

But fascinating concepts can sometimes become enshrouded as they become "buzzwords". At DataVisor, we are very passionate about machine learning and decided to publish this handbook to settle the debate of what **"supervised machine learning"** and **"unsupervised machine learning"** are and which one is better to protect your business from fraud.

First things first. Let's define machine learning as the branch of artificial intelligence and computer science that focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving accuracy. (IBM) In other words, machine learning refers to computers that can analyze large volumes of data and learn how to make decisions regarding that data, similar to what a human would do but faster and at a much larger scale.

Let's dive in!

WHAT IS SUPERVISED MACHINE LEARNING?

Supervised machine learning (SML) uses labeled datasets to train or "supervise" algorithms in order to classify new data or predict the outcome of similar situations in the future. It can be divided into "classification algorithms" that accurately assign data to specific categories and "regression algorithms" that understand the relationship between dependent and independent variables in given datasets.

Ok, but what are some everyday use cases of SML?

A good example of SML classification algorithms is the system in your inbox that classifies spam as such and moves it into a specific folder.

Have you wondered how Google Maps "predicts" how long your commute will take? It uses an SML regression model that learns from all the trips people have taken and then takes into account the specific variables from your route to yield an estimated time.

How Can SML Benefit My Company's Fraud Strategy?

To use SML in fraud detection, a process similar to this one must be implemented:

- First, large fraud databases are analyzed by people who create labels to indicate when events and user profiles are fraudulent or not.
- Then, features are associated with each entry in the database. Features are data attributes and can represent different traits or details like identity information, payment methods, locations, and shopping history.
- Finally, an algorithm, which is essentially a set of rules to solve problems (think about a mathematical equation as an example), is run on the feature-enriched dataset with the objective of learning how to make predictions for future entries that look similar to past ones.

Once this training is complete, the algorithm can check new instances and, if they share features with past ones, determine with high probability if they are fraudulent or not.

Adding SML to fraud detection efforts offers improvements over rules-based systems because of the ability to generalize patterns from previous instances of fraud. SML models can leverage many more features than a manually created rule and simultaneously weight features more accurately.

Limitations of SML for Fraud Detection and Prevention

What makes fraud detection a unique challenge for SML is that the former is a "moving target" where digital criminals constantly innovate in their tactics. Given that SML generalizes patterns from known instances of fraud, it remains bound to the data defined in those cases; so ML-only approaches ultimately offer no protection against new or unknown fraud attacks. In other words, SML offers a reactive protection against fraud.

Another limitation of SML-only approaches is that their models decay fast and require frequent re-tuning. This is especially relevant to companies that build SML models in-house because they need to service those models themselves without relying on the economies of scale that benefit well-established fraud detection software providers who service many clients at the same time.

WHAT IS UNSUPERVISED MACHINE LEARNING?

Unlike SML, unsupervised machine learning (UML) analyzes data sets that have not been labeled yet. It is called "unsupervised" because a computer processes large amounts of data without the need for humans to tell it what categories each datum belongs to (AKA "labeling").

UML algorithms are used for three main tasks:

Clustering:
AI that finds patterns in large datasets and "clusters" data based on their similarities or differences.

Association:
AI that finds relationships between variables in a given dataset.

Dimensionality Reduction:
A technique used when a data set has too many features to reduce the number of inputs to a more usable size without losing data integrity.

Examples of UML applications:

Marketing teams sometimes leverage UML association algorithms to group together customer segments based on their shared characteristics (demographics, interests, ideologies, etc.) and then build marketing campaigns that are tailored to the different groups of customers.

When you shop at your favorite online retailer and get a banner telling you that 'customers who bought this also bought', you are seeing the result of an association UML algorithm.

Key Insight:

In layman's terms, SML learns from labeled datasets by making predictions and adjusting for correct answers, and UML works on its own to discover the inherent structures of unlabeled data sets.

UML for Fraud Detection and Prevention

The signature advantage of unsupervised machine learning is its ability to operate without the need for labels; it can analyze data in real time, with no prior knowledge required. UML models are fundamentally self-tuning, and UML-powered solutions are free of the delays associated with supervised machine learning and rules-based approaches.

The use of UML-based fraud management tactics can yield stellar results when a given organization is wrestling with many different fraud types—and especially when the attacks are new or previously unknown. There are many fraud use cases for which UML can be applied, including:

APPLICATION FRAUD
Using UML, banks and financial institutions can analyze whole networks of applications to detect hidden connections that may appear legitimate when viewed in isolation.

TRANSACTION FRAUD
UML algorithms can be used to detect fraudulent accounts before those accounts can be used to conduct transactions that result in financial loss.

BOT ATTACKS
Using a UML-driven holistic data analysis approach, it is possible to analyze user histories, behavior changes, and suspicious patterns across millions of accounts. This enables the capture of significantly more bot-powered attacks.

PROMOTION ABUSE
UML solutions enable the capture of all members of a given fraud ring by identifying hidden linkages between fake account registrations and discovering unknown attacks without labels or training data.

MONEY LAUNDERING
UML algorithms can look at complex networks of transactions instead of individual ones, and can detect and eliminate launderers who deposit small denominations of funds to avoid CTR reporting.

WHICH ONE IS BETTER FOR MY BUSINESS' FRAUD PREVENTION NEEDS?

Drumroll, please…

It depends!

This might not be the quick answer you were looking for, so apologies in advance. But really, trust us: The choice between UML and SML depends on the goals you seek and other practical limitations that need to be considered on a case-by-case basis.

Do you want to predict outcomes for new data? Do you expect the new data to look very similar to your existing data? Then you probably only need an SML algorithm.

On the other hand, if you answer yes to any of following questions, then you should probably look into UML algorithms:

- Do you need to uncover patterns and anomalies in large data sets that are not apparent to the naked eye?
- Are you being attacked using new fraud patterns?
- Is your data unlabelled and would it take a lot of time/resources to label it?
- Are you launching new products or markets and need to be protected from the start, without waiting to build and label data?

WHAT KIND OF ML DOES DATAVISOR USE WITH ITS CUSTOMERS?

This question does have a simple answer:

The right one, depending on what our customer needs.

We work with companies that operate in dozens of different businesses and we know that they are all unique, so we always make sure to understand their use cases, pain points, and needs before we recommend any solution.

We told you that we are very passionate about what we do, and we meant it. We like to listen to our customers and always provide clear recommendations that are based on evidence and backed by our expertise in the field.

Key Insight:

When choosing a fraud detection provider, the most important aspect you need to consider is whether or not vendors' solutions really align with your business goals and KPIs and if they are a right match for the resources you have available. When it comes to fighting fraud and its effect on your bottom line, it's not enough to implement the shiniest technologies if they are not the right solution to your problems.

Are you still curious? Do you want to know how machine learning can help your business fight fraud?

About DataVisor

DataVisor is the world's leading AI-powered Fraud and Risk Platform that delivers the best overall detection coverage in industry. With an open SaaS platform that supports easy consolidation and enrichment of any data, DataVisor's solution scales infinitely, enabling organizations to act on fast-evolving fraud and money laundering activities as they happen in real time. Its patented unsupervised machine learning technology, combined with its advanced device intelligence, powerful decision engine and investigation tools, provides guaranteed performance lift from day one.

For more information on DataVisor:

info@datavisor.com

www.datavisor.com

967 N. Shoreline Blvd. | Mountain View | CA 94043

DATAVISOR