The ultimate guide to anomaly detection

Key use cases, techniques, and autoencoder machine learning models

Why is anomaly detection important?

Large volumes of data from business operations are generated daily. If used correctly, this data can help businesses make better decisions—and one way to gain a competitive advantage with this data is through anomaly detection. Detecting anomalies, using Spotfire® analytics, can stop a minor issue from becoming a widespread, time-consuming problem. In this white paper, we cover the basics of anomaly detection, its main use cases, and a few key techniques to keep in mind.

What are anomalies?

Before we dive in, let's take a step back. What exactly are anomalies?

An anomaly is an unexpected change or deviation from the expected pattern in a dataset. Therefore, anomaly detection is a way of detecting abnormal behavior. It's important to note that anomalies aren't necessarily good or bad, but companies should be alerted to any break in pattern to assess whether actions need to be taken.

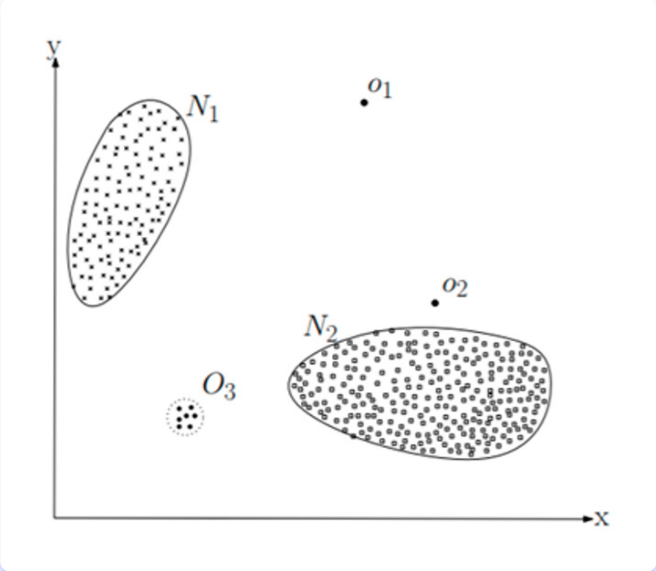


Figure 01: Example of Anomalies (O1, O2, O3) in a 2D Dataset

The difference between anomalies and outliers

There is much debate on this topic, and many people use the terms interchangeably—although they are similar, anomalies are not identical to outliers. Synonyms for outliers may include ”discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different application domains.”[1]

Assuming that all data is generated by a set of processes, outliers are points with a low probability of occurrence within a given dataset generated by those processes. They are observation points that are distant from other observations within the normal population; however, they don’t necessarily represent abnormal behavior or behavior that occurred because of a different process. Outliers warrant attention because they can distort predictions and affect model accuracy if not detected and handled appropriately. They are generated by the same process but occur with lower probability whereas anomalies are patterns that are generated by different processes.

Use cases for anomaly detection

Manufacturing defects

Autoencoders are used in manufacturing to find defects. Manual inspection to find anomalies is a laborious and offline process, and building machine-learning models for each part of the system is difficult. Therefore, some companies have implemented a process where sensor data on manufactured components is continuously monitored and any defects (anomalies) are detected using an autoencoder model that scores the new data.

Monitoring equipment sensors

Many types of equipment, vehicles, and machines now have sensors. For example, your “simple” smartphone has many that include ambient light and back-illuminated sensors, accelerometers, digital compasses, gyroscopes, proximity, near-field communication, GPS, and fingerprint sensors. Monitoring these outputs can be crucial to detecting and preventing breakdowns and disruptions. Unsupervised learning algorithms like autoencoders are widely used to detect anomalous data patterns that may indicate impending problems.

Fighting financial crime

In the financial world, transactions worth trillions of dollars are executed every minute. Identifying suspicious ones in real time can provide organizations with a necessary competitive edge. Over the last few years, leading financial companies have increasingly adopted big data analytics to identify abnormal transactions, clients, suppliers, or bad actors. Machine learning models are used extensively to detect anomalies in streaming data. By one estimate, “Mastercard’s network is estimated to process up to 5,000 transactions per second.”[2]

Healthcare claims fraud

Insurance fraud is a common occurrence in the healthcare industry. It is vital for insurance companies to identify fraudulent claims and ensure that no payout is made for them. In the past few years, many companies have invested heavily in big data analytics to build supervised, unsupervised, and semi-supervised models to predict the likelihood of insurance fraud for each claim submitted.

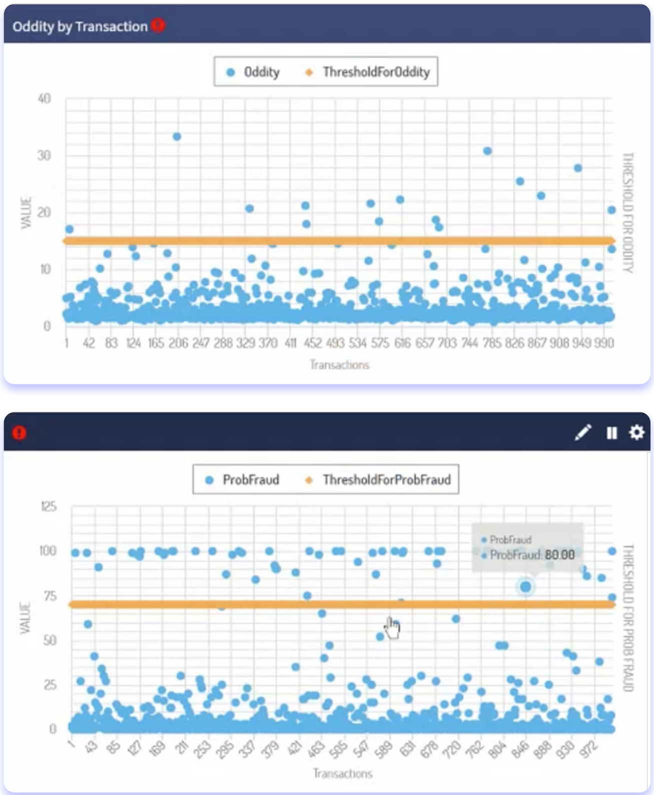


Figure 02: Real-time Fraud Detection Accelerator

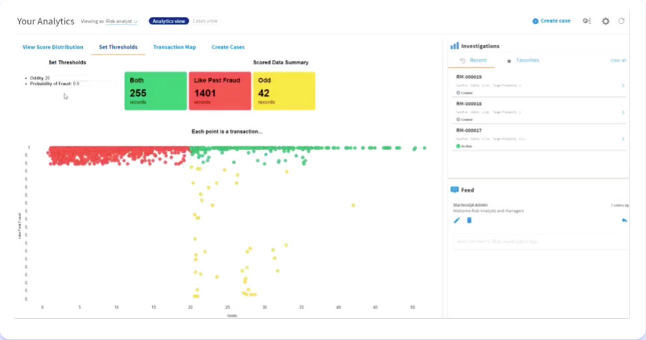
[1] Chandola, Varun; Banerjee, Arindam; Kumar, Vipin. Anomaly Detection: A Survey, ACM Computing Surveys, https://dl.acm.org/doi/10.1145/1541880.1541882

[2] Rodrigues, Francisco. Bitcoin Lightning Network vs Visa and Mastercard: How Do They Stack Up?, Cointelegraph, August 24, 2022. https://cointelegraph.com/news/bitcoin-lightning-network-vs-visa-and-mastercard-how-do-they-stack-up

Further examples

Beyond the most common use cases already described, there are many other examples of how anomaly detection can be applied across a wide variety of industries:

- Military surveillance: Image recognition
- Cybersecurity: Intrusion detection
- Safety systems: Fault detection



Set Thresholds

- Oddity 20
- Probability of Fraud: 0.9

Both 255 records

Like Past Fraud 1401 records

Odd 42 records

StartersQA Admin
Welcome Risk Analysts and Managers

Figure 03: Spotfire Cloud Risk Investigation App

- Hack protection: Anomalous network traffic detection
- Weather variables: Heat wave or cold snap detection
- MRI imaging: Alzheimer's or malignant tumor detection
- Spacecraft sensors: Faulty component detection

Techniques for anomaly detection

Companies around the world have used many techniques to detect data abnormalities in their markets. While the list below is not comprehensive, three anomaly detection techniques have been popular:

1. **Visual discovery:** Anomaly detection can be accomplished through visual discovery. In this process, a team of data or business analysts visually monitors dashboards containing bar charts, scatter plots, statistical process control (SPC) graphs, and other visualizations to find unexpected behavior. This technique often requires prior business knowledge in the industry of operation and creative thinking to ensure the right visualizations are used to find targets. Because humans are inherently visual, we can quickly spot patterns in data. However, the downside is that we may not be able to visually inspect all of the data available in a reasonable timeframe.
2. **Supervised learning:** Supervised learning is an improvement over visual discovery. In this technique, persons with business knowledge in a particular industry label a set of data points as normal or abnormal. An analyst then uses this labeled data to build machine learning models that will be able to predict anomalies in unlabeled new data. Supervised learning is a great technique to use when you have known patterns in data that you would like to model. However, in a case where new patterns may emerge (fraud, manufacturing), you may need to implement an unsupervised learning method.
3. **Unsupervised learning:** Another technique that is very effective but not as popular as others is unsupervised learning. In this technique, unlabeled data is used to build unsupervised machine learning models. These models are then used to predict new data. Because the model is tailored to fit normal data, anomalous data points will stand out.

Examples of unsupervised learning algorithms

Autoencoders

Unsupervised neural networks, or autoencoders, are used to replicate the input dataset but only approximately; if the input is replicated exactly, the model cannot usefully be applied to new data. The approximation is made by restricting the number of hidden layers in a neural network. A reconstruction error is generated upon prediction. The reconstruction error is defined as the difference between the model output and the new input data. The higher the reconstruction error, the higher the possibility that the data point is an anomaly.

Clustering

In this technique, the analyst attempts to classify each data point into one of many pre-defined clusters by minimizing the within-cluster variance. Models such as K-means clustering, K-nearest neighbors, and others, are used for this purpose. A K-means or a K-NN model serves the purpose effectively because it assigns a separate cluster for all data points that do not look similar to normal data.
In general, this technique is most useful when the data are well separated into natural clusters—a requirement that should also be checked.

One-class support vector machine

A support vector machine defines a hyperplane that best divides a set of labeled data into two classes. For this purpose, the distance between the two nearest data points that lie on either side of the hyperplane is maximized. For anomaly detection, a one-class support vector machine is used to classify as anomalies those data points that lie much farther away than the rest of the data points. Similar to clustering, this is most effective when the points are well separated into natural groupings.

Time series techniques

Anomalies can also be detected through time series analytics by building models that capture trends, seasonality, and levels and slope changes. These models are then used along with new data to find anomalies. When the new data diverges too much from the model prediction, either an anomaly or a model failure is indicated. Recent developments include the MASS techniques (Mueen's Algorithm for Similarity Search) for fast scanning of time series for unusual subsequences (discords) and related methods for change detection and for comparisons of multiple time series.

Autoencoders explained

Autoencoders use unsupervised neural networks that are both similar to and different from a traditional feed-forward neural network. They are similar in that they use the same principles (for example, backpropagation) to build a model. They are different in that they do not use a labeled dataset containing a target variable for building the model. They use a training dataset and attempt to replicate the output dataset by restricting hidden layers/nodes.

The focus of this model is to learn and identify a function or an approximation of it that would allow it to predict an output that is similar to the input. The function achieves this by placing restrictions on the number of hidden units in the data. For example in the figure shown above, if we have seven columns in a dataset (L1) and only four hidden units (L2), the neural network is forced to learn a more restricted representation of the input. By limiting the hidden units, we can force the model to learn a pattern in the data, if one exists.

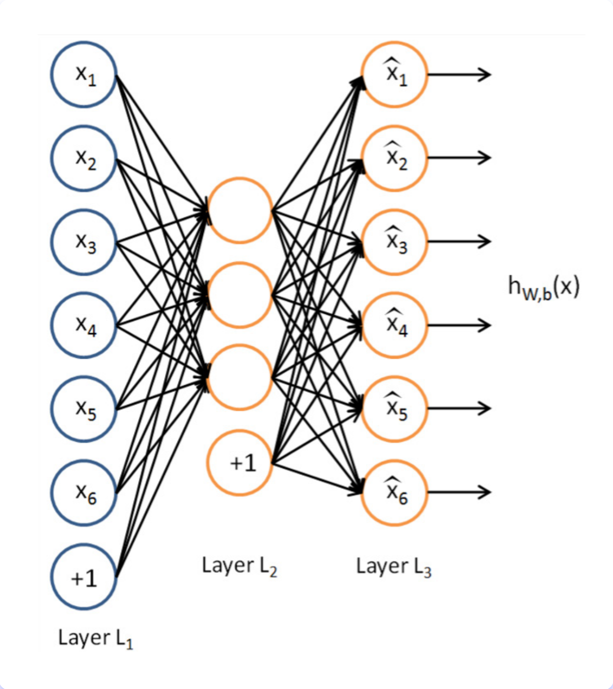


Figure 04: Example Schematic of a Single-layer Sparse Autoencoder

Each of the hidden units can be either active or inactive, and an activation function, such as tanh or rectifier, can be applied to the input of these hidden units to change their state.

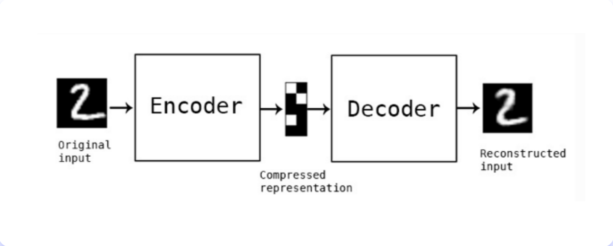


Figure 05: Example of Auto Encoding and Decoding Process for Digit Recognition

Some forms of autoencoders:

- **Under-complete autoencoders:** These are normally used for anomaly detection, for example in manufacturing or fraud detection. The autoencoder is built with constraints so that the input cannot be exactly reproduced, and the reconstruction error can be used for detecting anomalies.
- **Regularized autoencoders:** In contrast, these are more able to reproduce the input but are constrained differently to provide other capabilities. A good example is denoising autoencoders that can be used to remove extraneous data points and recognize patterns of interest.
- **Representational power, layer size, and depth:** The architecture of the network along these dimensions (layer size and depth) determines whether the result is incomplete and how complex the features represented can be.
- **Stochastic encoders and decoders:** These use probability distributions instead of point estimates of connection weights. For example, variational autoencoders are used to generate new images or texts that are extrapolations of the patterns found in the original data.
- **Denoising autoencoders:** Random changes are made to the input, and the network is trained to recognize/reconstruct the original data.

Learn more about anomaly detection

Both anomaly detection and autoencoder machine learning models present rich opportunities for companies able to successfully implement them. Hopefully, this guide gave you a better understanding of the value such tactics provide and the possible applications across different industries.

Now that you've got a handle on the basics, get started with anomaly detection using Spotfire® Visual Analytics software.

Cloud Software Group
Headquarters
851 W Cypress Creek Rd.
Fort Lauderdale, FL 33309
www.spotfire.com

Spotfire® goes beyond basic rearview dashboards to offer a single visual analytics platform for data exploration and real-time decisions. Backed by point-and-click, no-code data science, Spotfire allows even the non-developer to analyze both data-at-rest and data-in-motion, together, for faster time-to-insight and better business outcomes.

15Feb2024