Anomaly detection

Anomaly detection

Anomaly detection is a process of finding those rare items, data points, events, or observations that make suspicions by being different from the rest data points or observations. Anomaly detection is also known as outlier detection.

If you think of this in the context of time-series continuous datasets, the normal or expected value is going to be the baseline, and the limits around it represent the tolerance associated with the variance. If a new value deviates above or below these limits, then that data point can be considered anomalous.

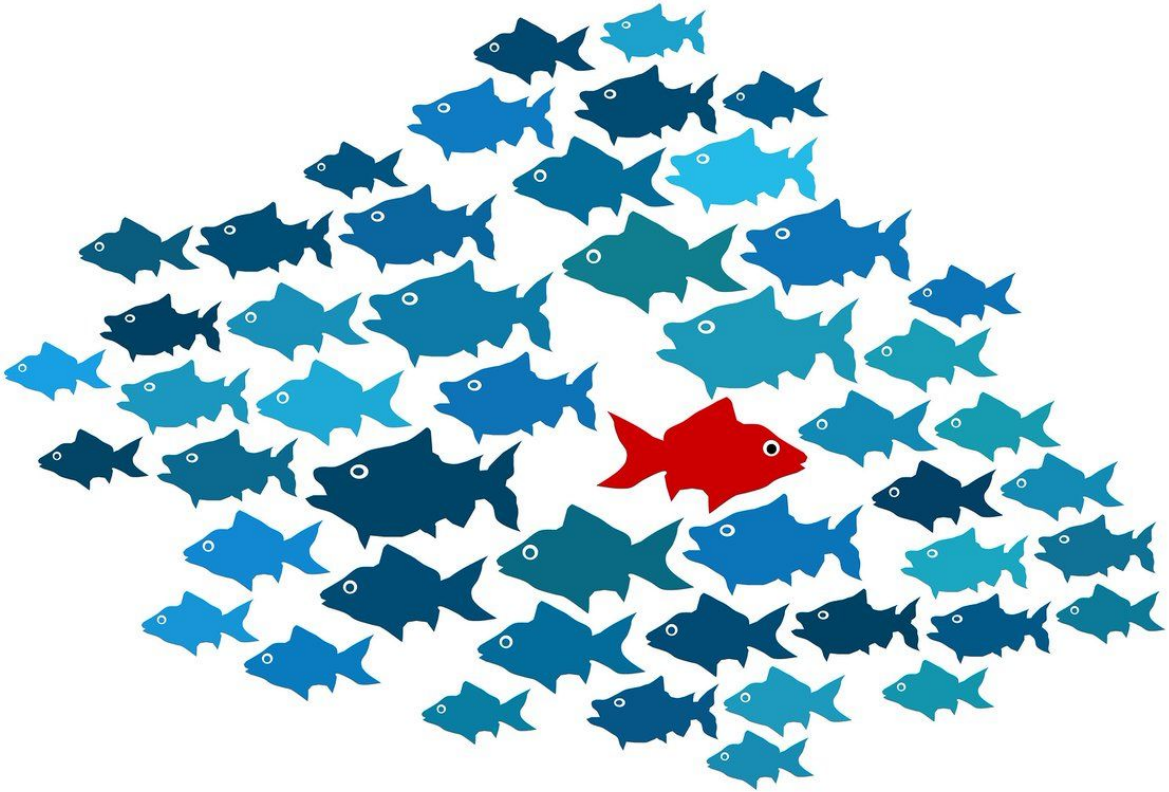
Anomaly detection is a key use case for machine learning algorithms, and one that might seem like magic. We know, of course, that accurate anomaly detection relies on a combination of historical data and ongoing statistical analysis. Importantly, these models are highly dependent on the data quality and sample sizes used that affect the overall alerting.

Anomaly detection

Common reasons for outliers are:

- data preprocessing errors;
- noise;
- fraud;
- attacks.

Anomaly detection



Real Life Anomaly detection

Success in today's high-velocity business environments means having the correct information to make the right decisions at the right time. As marketplaces grow more competitive and customer expectations continually rise, the "right time" is often real-time. Every transaction generates a plethora of data.

Anomalies within your company's data set can represent opportunities and threats to the business. Real-time detection of anomalies empowers enterprises to make the right decisions to seize revenue opportunities and avoid potential losses.

Three Types of Anomaly Detection

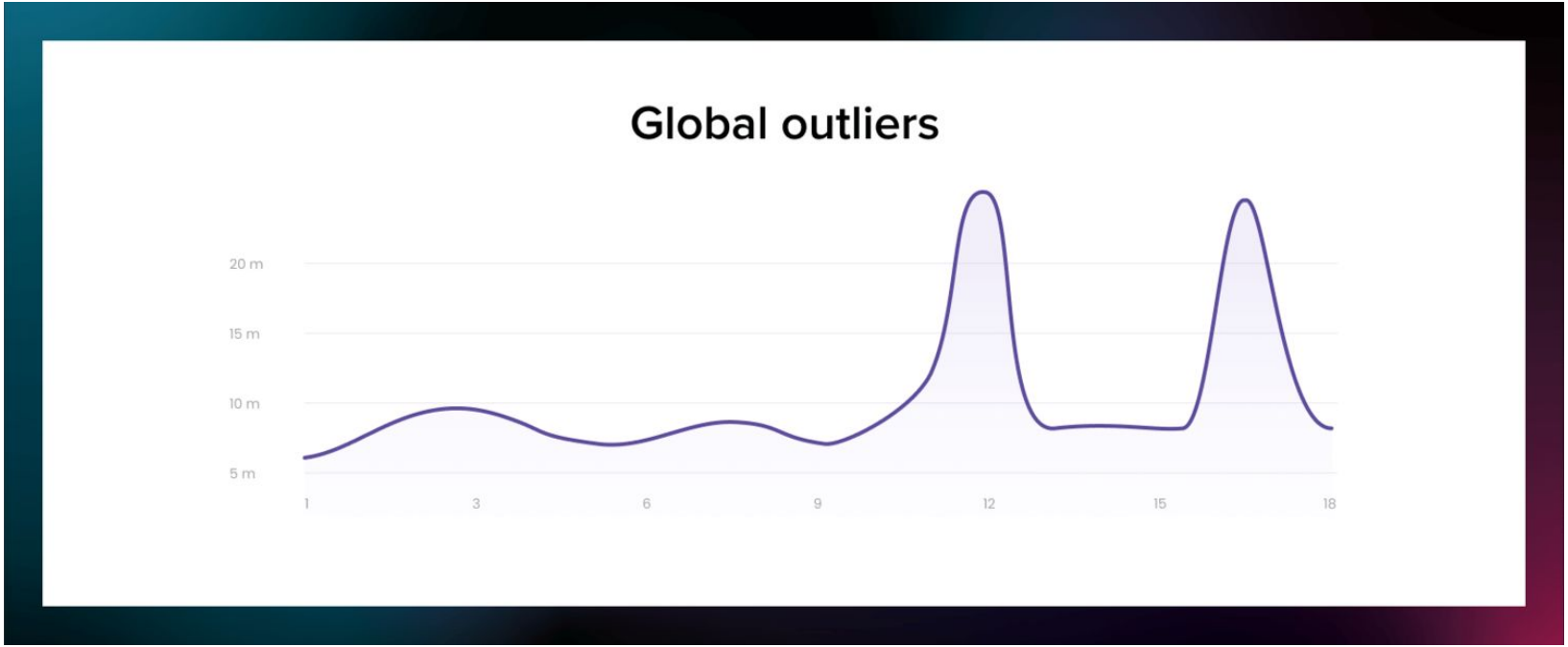
There are three commonly accepted types of anomalies in statistics and data science: **Global outliers, contextual outliers, and collective outliers**.

Global outliers

When a data point assumes a value that is far outside all the other data point value ranges in the dataset, it can be considered a global anomaly. In other words, it's a rare event.

For example, if you receive an average American salary to your bank accounts each month but one day get a million dollars, that would look like a global anomaly to the bank's analytics team.

Global outliers



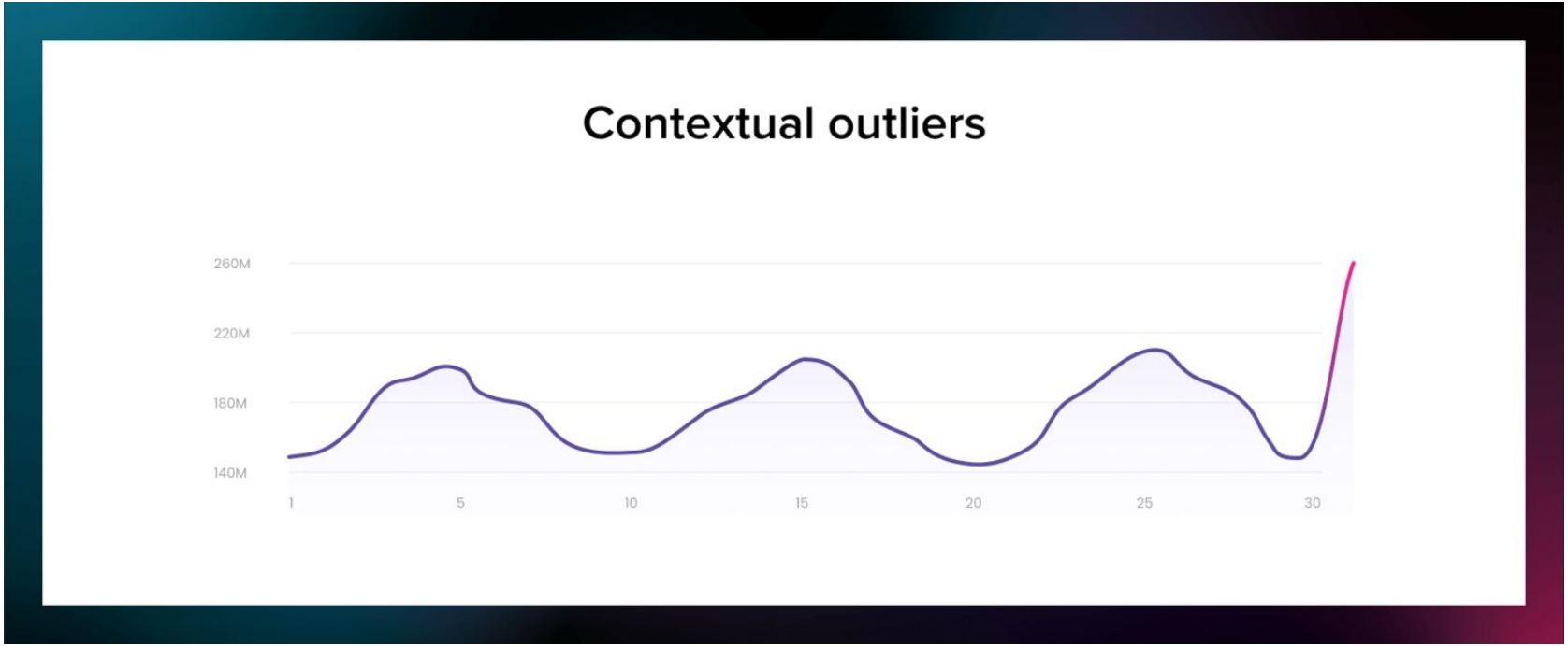
Contextual outliers

Contextual outliers

When an outlier is called contextual it means that its value doesn’t correspond with what we expect to observe for a similar data point in the same context. Contexts are usually temporal, and the same situation observed at different times can be not an outlier.

For example, for stores it’s quite normal to experience an increase in customers during the holiday season. However, if a sudden boost happens outside of holidays or sales, it can be considered a contextual outlier.

Contextual outliers



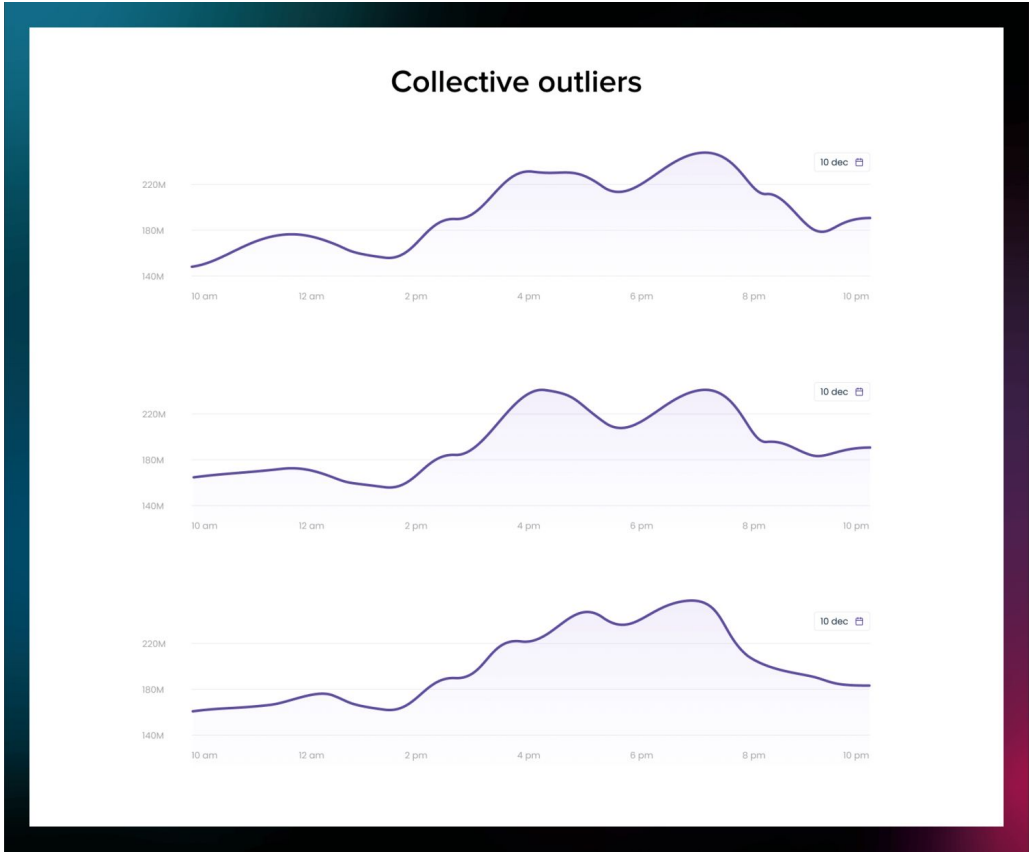
Collective outliers

Collective outliers

Collective outliers are represented by a subset of data points that deviate from the normal behavior.

In general, tech companies tend to grow bigger and bigger. Some companies may decay but it’s not a general trend. However, if many companies at once show a decrease in revenue in the same period of time, we can identify a collective outlier.

Collective outliers



Why do you need machine learning for anomaly detection?

This is a process that is usually conducted with the help of statistics and machine learning tools.

The reason is that the majority of companies today that require outlier detection work with huge amounts of data: transactions, text, image, and video content, etc. You would have to spend days going through all the transitions that happen inside a bank every hour, and more and more are generated every second. It is simply impossible to drive any meaningful insights from this amount of data manually.

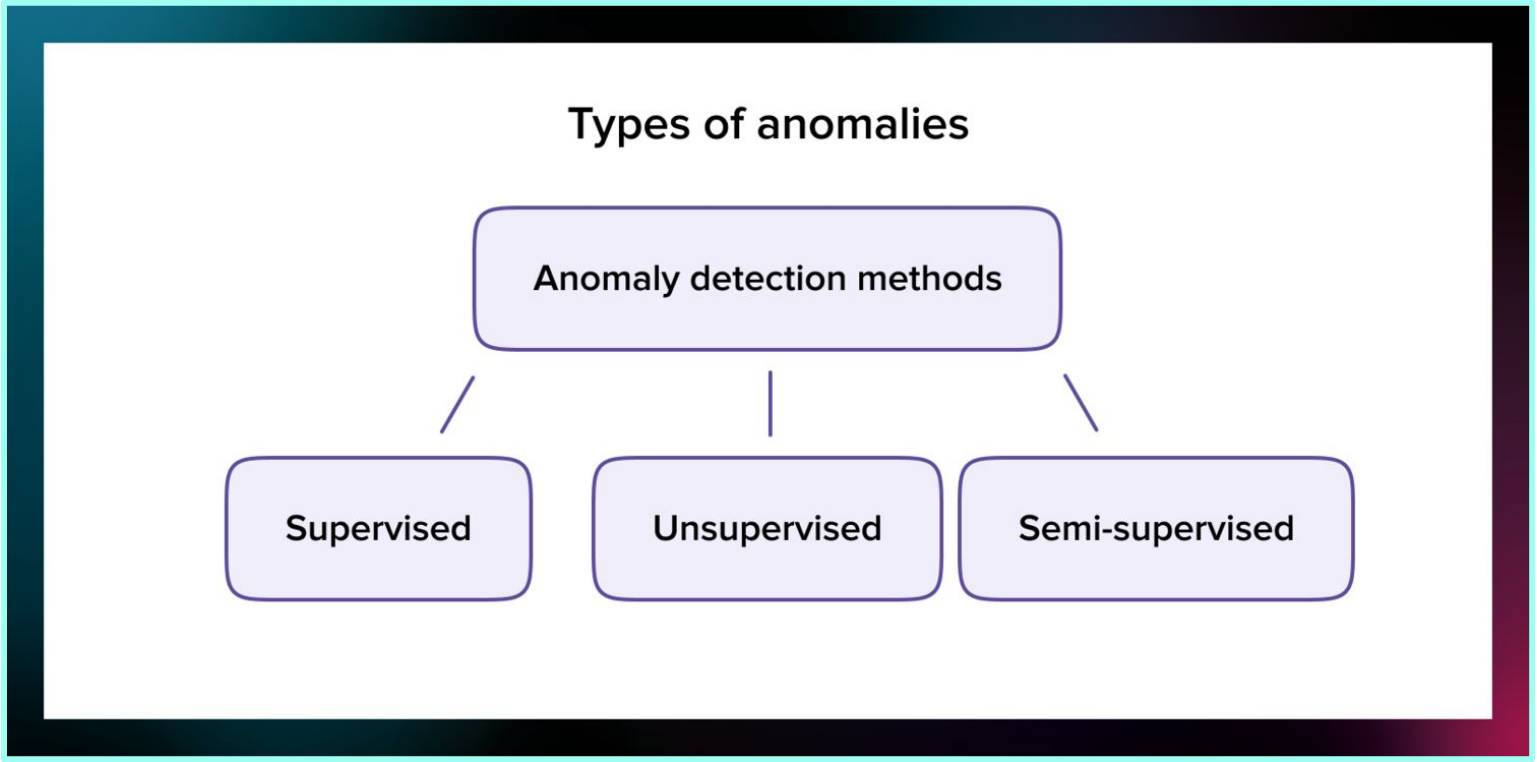
Moreover, another difficulty is that the data is often unstructured, which means that the information wasn't arranged in any specific way for the data analysis. For example, business documents, emails, or images are examples of unstructured data.

Why do you need machine learning for anomaly detection?

To be able to collect, clean, structure, analyze, and store data, you need to use tools that aren’t scared of big volumes of data. Machine learning techniques, in fact, show the best results when large data sets are involved. Machine learning algorithms are able to process most types of data. Moreover, you can choose the algorithm based on your problem and even combine different techniques for the best results.

Machine learning used for real-world applications helps to streamline the process of anomaly detection and save the resources. It can happen not only post-factum but also in real time. Real-time anomaly detection is applied to improve security and robustness, for instance, in fraud discovery and cybersecurity

What are anomaly detection methods?



Supervised

In supervised anomaly detection, an ML engineer needs a training dataset. Items in the dataset are labeled into two categories: normal and abnormal. The model will use these examples to extract patterns and be able to detect abnormal patterns in the previously unseen data.

In supervised learning, the quality of the training dataset is very important. There is a lot of manual work involved since somebody needs to collect and label examples.

Note: While you can label some anomalies and try to classify them (hence it's a classification task), the underlying goal of anomaly detection is defining "normal data points" rather than "abnormal data points". So in real world applications with very few anomaly samples labelled, it's almost never regarded as a supervised task

Unsupervised

This type of anomaly detection is the most common type, and the most well-known representative of unsupervised algorithms are neural networks.

Artificial neural networks allow to decrease the amount of manual work needed to pre-process examples: no manual labeling is needed. Neural networks can even be applied to unstructured data. NNs can detect anomalies in unlabeled data and use what they have learned when working with new data.

The advantage of this method is that it allows you to decrease the manual work in anomaly detection. Moreover, quite often it’s impossible to predict all the anomalies that can occur in the dataset. Think of self-driving cars, for example. They can face a situation on the road that has never happened before. Putting all road situations into a finite number of classes would be impossible. That is why neural networks are priceless when working with real-life data in real-time.

Semi-supervised

The architecture of neural networks is a black box. We often don't know what kinds of events neural networks will label as anomalies, moreover, it can easily learn wrong rules that are not so easy to fix. That is why unsupervised anomaly detection techniques are often less trustworthy than supervised ones.

Semi-supervised

Semi-supervised anomaly detection methods combine the benefits of the previous two methods. Engineers can apply unsupervised learning methods to automate feature learning and work with unstructured data. However, by combining it with human supervision, they have an opportunity to monitor and control what kind of patterns the model learns. This usually helps to make the model's predictions more accurate

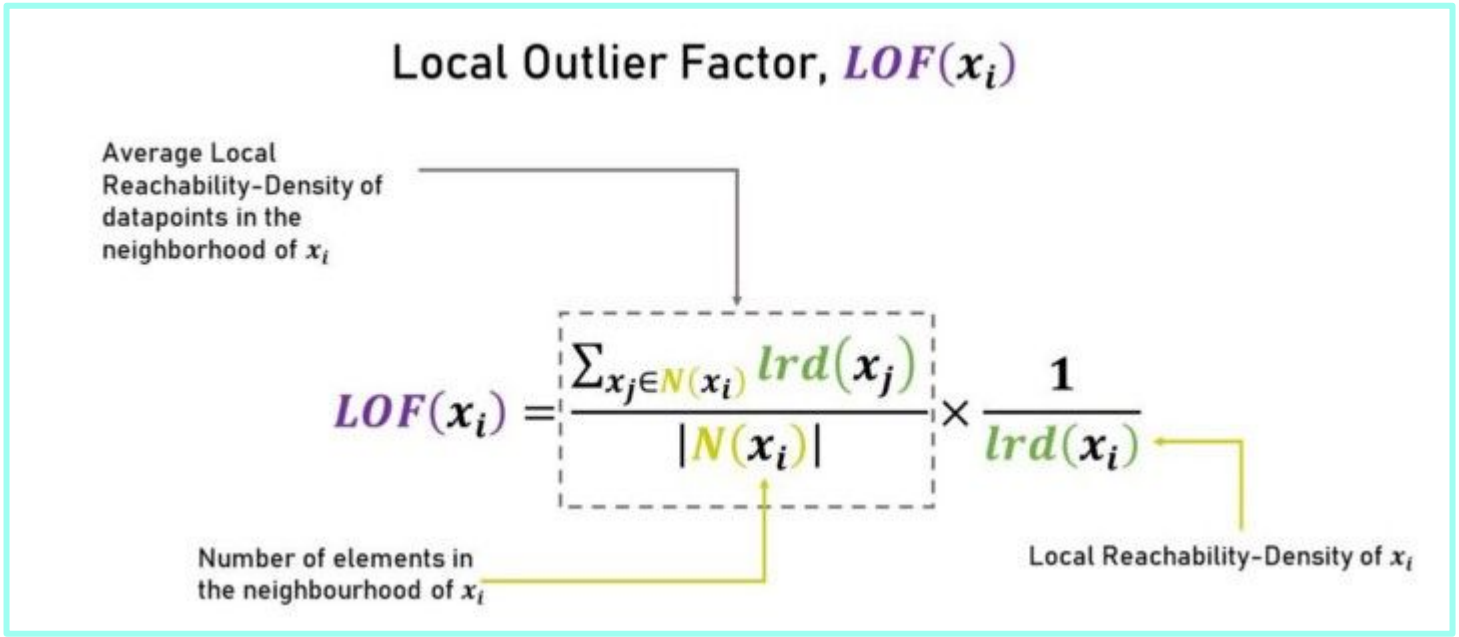
Machine learning algorithms for anomaly detection

Multiple machine learning algorithms can be used for anomaly detection depending on the dataset size and the type of the problem.

Local outlier factor (LOF)

Local outlier factor is probably the most common technique for anomaly detection. This algorithm is based on the concept of the local density. It compares the local density of an object with that of its neighbouring data points. If a data point has a lower density than its neighbours, then it is considered an outlier.

Machine learning algorithms for anomaly detection



Machine learning algorithms for anomaly detection

K-nearest neighbors

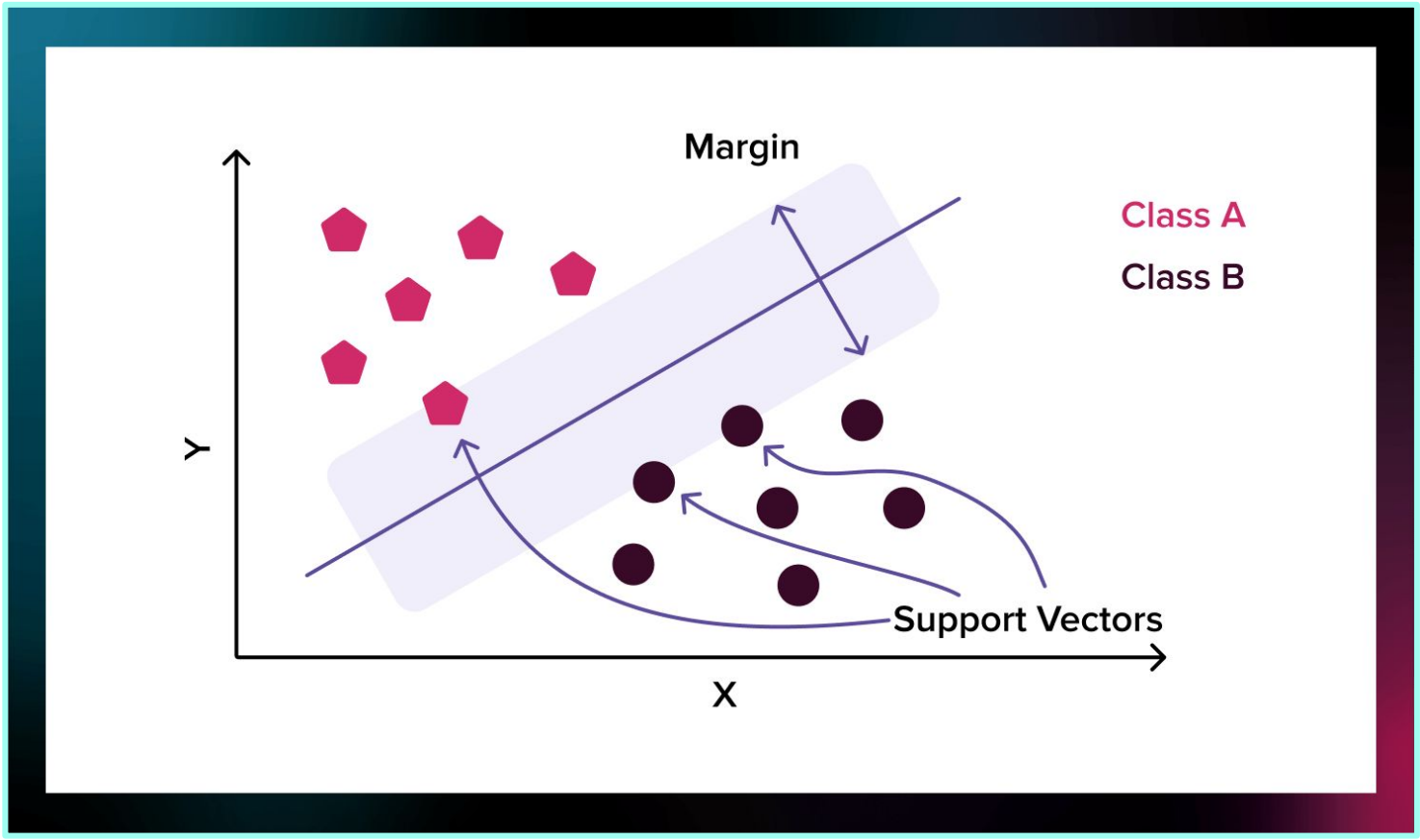
kNN is a supervised ML algorithm often used for classification. When applied to anomaly detection problems, kNN is a useful tool because it allows to easily visualize the data points on the scatterplot and make anomaly detection much more intuitive. Another benefit of kNN is that it works well on both small and large datasets.
Instead of learning ‘normal’ and ‘abnormal’ values to solve the classification problem, kNN doesn’t perform any actual learning. So when it comes to anomaly detection, kNN works as an unsupervised learning algorithm. A machine learning expert defines a range of normal and abnormal values manually, and the algorithm breaks this representation into classes by itself.

Machine learning algorithms for anomaly detection

Support vector machines

Support vector machine (SVM) is also a supervised machine learning algorithm often used for classification. SVMs use hyperplanes in multi-dimensional space to divide data points into classes. The hyperparameter nu is the threshold (percentage) for outliers which you have to choose manually.
SVM is usually applied when there are more than one classes involved in the problem. However, in anomaly detection it is also used for single class problems. The model is trained to learn the ‘norm’ and can identify whether unfamiliar data belongs to this class or represents an anomaly.

Machine learning algorithms for anomaly detection



Machine learning algorithms for anomaly detection

DBSCAN

This is an unsupervised ML algorithm based on the principle of density. DBSCAN is able to uncover clusters in large spatial datasets by looking at the local density of the data points and generally shows good results when used for anomaly detection. The points that do not belong to any cluster get their own class: -1 so they are easy to identify. This algorithm handles outliers well when the data is represented by non-discrete data points.

Machine learning algorithms for anomaly detection

Autoencoders

This algorithm is based on the use of artificial neural networks that encode the data by compressing it into the lower dimensions. Then, ANNs decode the data to reconstruct the original input. When we reduce the dimensionality, we don't lose the necessary information because the rules have already been identified in the compressed data. Now we can already discover outliers.

Bayesian networks

Bayesian networks enable ML engineers to discover anomalies even in high-dimensional data. This method is used when the anomalies that we're looking for are more subtle and harder to discover and visualizing them on the plot might not produce the desired results.

Need of Anomaly Detection

1. Anomaly detection for application performance

Application performance of any company can either generate or reduce workforce productivity and revenue. General or traditional approaches for monitoring the application performance allow to react to issues, but still business used to suffer, and hence it affects the user. But with the help of anomaly detection using machine learning, it is easy to identify and resolve the application performance issues before they affect the business as well as users.

Anomaly detection using machine learning algorithms can simply correlate data with corresponding application performance metrics and find out the complete knowledge of the issue. There are different industries that also employ anomaly detection techniques for their businesses, such as **Telco, Adtech**, etc.

Need of Anomaly Detection

2. Anomaly detection for product quality

It is not enough for product managers to trust another department for taking care of required monitoring and alerts. It is always required for product managers to be able to trust that product will work smoothly. It is because the product always needs changes, from each version release to new feature upgradation, and generates anomalies. If you don't properly monitor these anomalies, it may cause millions of revenues lost and can also affect the brand reputation.

Need of Anomaly Detection

3. Anomaly detection for user experience

If you release a faulty version, you may experience a DDoS attack, risk of usage lapses across customer experiences. So, it is required to react to such issues before they impact user experience to reduce the chances of revenue loss.

Proactively streamlining and improving user experiences will help improve customer satisfaction in a variety of industries, including Gaming, online business, etc.

Need of Anomaly Detection

4. Fraud detection

Fraud detection with machine learning helps to prevent activities aimed at obtaining money or property unlawfully. Fraud detection software is used by banks, credit organizations, and insurance companies. For example, banks check loan applications before making a decision. If the system detects that some of the documents are fraudulent, for example, that your tax number doesn't exist in the system, it will notify the bank employer.

5. Health monitoring

Anomaly detection systems are incredibly helpful in healthcare. They help doctors with diagnosis detecting unusual patterns in MRI and test results. Usually, neural networks that have been trained on thousands of examples are applied here, and sometimes they give a more accurate diagnosis than doctors with 20 years of experience.

Need of Anomaly Detection

6. Defect detection

Manufactures can lose millions in lawsuits supplying their clients with mechanisms or mechanism details that have defects. One detail that doesn’t correspond to the production standards can cause a plane to crash, thus, killing hundreds of people. Anomaly detection systems that use computer vision can detect if the detail has a defect even among thousands of other similar details on the beltline. Moreover, anomaly detection systems can be connected to the mechanisms to monitor internal systems such as engine temperature, fuel levels, and other parameters.

Thank You