What is an Isolation Forest?

Isolation Forest is an algorithm similar to Random Forest in that both use decision trees. However instead of focusing on identifying common data patterns, Isolation Forest is designed specifically to detect anomalies . It does this by randomly selecting a split value within the range of a particular feature. This process is repeated recursively to create isolation trees. Due to this mechanism or using random partitions, the algorithm for isolation forest runs very fast with a linear time complexity. In comparison, algorithms like DBSCAN  (which we discussed in our [previous article](https://www.knime.com/blog/fraud-detection-using-dbscan)) have log-linear time complexity. This makes isolation forests more efficient, improving runtime performance significantly.

Although Isolation Forest runs quickly, and it can handle large datasets, it is not without limitations.. The algorithm often needs parameter tuning depending on the dataset given and is sensitive to the sample size it is trained on as larger sample sizes have a higher risk of “swamping” and “masking” effects on the model, where the anomalies get buried within normal data, reducing the model's effectiveness

Despite the advantages of Isolation Forest, its limitations need to be carefully managed, especially when applied to real-world datasets where anomalies are sparse. Parameter tuning and understanding the data distribution are still necessary for the algorithm to perform optimally.

With this understanding, let's explore how we can apply Isolation Forest to a practical use case: detecting fraudulent transactions with real credit card data.

Identify fraudulent transactions with isolation forest

Credit card transactions can generally fall into two categories: normal and suspicious. The goal is to accurately identify fraudulent transactions while minimizing the number of false positives—cases where legitimate transactions are incorrectly flagged as fraudulent. Ideally, only a small percentage of flagged transactions should turn out to be false positives.

In our use case, we will  focus on automating fraud detection by training a model on a labeled dataset and applying it to a new transaction to simulate incoming data from an outside data source.

## Training a machine learning model to identify fraudulent transactions

Isolation Forest is a type of unsupervised learning model that is good at identifying outliers by using random partitioning to create trees. This method improves efficiency, particularly for finding outliers.