fraud detection is a branch of anomaly detection, which is a branch of statistics.

fraud detection algorithms can be used for detecting insurance fraud, credit card fraud, identity theft, money laundering and more.

fraud detection algorithms usually deal with imbalanced classes because typically fraud occurrences are an extreme minority in such transactions.

resampling methods

to combat class imbalance, you can undersample the majority (non-fraud) class or oversample the minority (fraud) class. however with random oversampling you train your model on a lot of duplicates and with random undersampling you throw away a lot of useful data. to avoid duplicating fraud observations, you can use the SMOTE method, which uses characteristics of nearest neighbors of fraud cases to create new synthetic fraud cases.

Use resampling methods on the training set, not on the test set, and test your model only on real data.

rule based systems or ML based systems

you can either used rule based systems or ML based systems as your fraud detection algorithm, however ML systems are usually preferred because they yield a probability while rule based systems only give a yes or no answer, and ML systems can adapt to data over time while rule based systems are static

With an ML based system, you can also combine resampling methods and models together through pipelining, such as combining a logistic regression with a SMOTE method.

fraud detection using labeled data

First you have to distinguish between normal and abnormal behavior, as abnormal data points might potentially be fraudulent. To understand normal behavior, you describe the data by plotting histograms, checking for outliers, and exploring correlations.

clustering models are used to detect fraud by analyzing patterns in the data.

K-means clustering groups the data into clusters that contain data points similar to each other but different from the data in other clusters. There's also a cluster centroid in the middle of each cluster, marked with x or a star.

The objective of k-means clustering is to make each cluster as tightly packed and far away from each other as possible.

With k-means clustering, you have to set the number of clusters beforehand, and to estimate the right number of clusters you can either use the elbow method or the silhouette method. Because k-means is a distance-based method, you also have to scale your data prior to clustering.

finding outliers

after the clusters are formed, you define a cut-off point and take the distance of each data point to its cluster centroid. You use the euclidean distance metric to find the distances. if the distance is greater than the cut-off point, you flag that data point as an outlier and therefore abnormal. here, all the data points outside the circle are abnormal. abnormal data points can be fraudulent or just rare cases of legit data, so you have to investigate them in more detail.

DBSCAN

Another clustering method is DBSCAN, which works better on weirdly shaped data and clusters of similar density. With DBSCAN, you don’t need to predefine the number of clusters, but you have to predefine the maximum allowed distance between points in a cluster and the minimal number of data points in each cluster.

This approach can be used when fraudulent behavior has commonalities, which cause clustering. The fraudulent data would cluster in tiny groups, rather than be the outliers of larger clusters. So for this method, you label the smallest clusters as abnormal and potentially fraudulent clusters.

other clustering algorithms

here are some other clustering algorithms you would use in different cases

Mean-shift: Many clusters, uneven cluster size, non-flat geometry

Spectral clustering: Few clusters, even cluster size, non-flat geometry

Ward hierarchical clustering: Many clusters, possibly connectivity constraints

OPTICS: Non-flat geometry, uneven cluster sizes, variable cluster density

only points, lines, and hyper-planes are flat