Fraud Data Analytics Techniques Catalogue

Fraud Data Analytics Maturity

Technique Sophistication

Descriptive Statistics

Data Matching

Trend Analysis

Bayesian Statistics

Fuzzy Matching

Decision Trees

Regression Analysis

Neural Language Processing

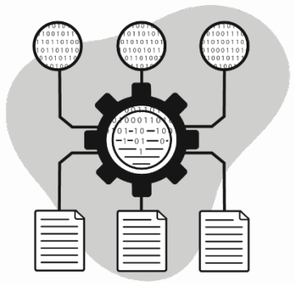
Network Analytics

Support Vector Machines

Clustering Algorithms

Neural Networks

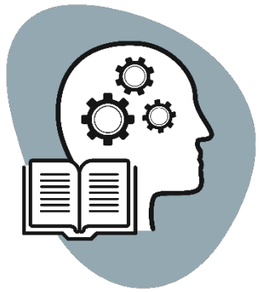
Ctrl click



Rules Based Analytics:

Human knowledge of fraud indicators is converted into rules that are run against data sets. When an indicator of fraud is detected by the rules, the transaction is flagged or a response triggered.

Analysis techniques are deductive - insights are generated directly from data, without inference.



Supervised Learning:

The prediction of results based on relationships between starting values and end results. Predictive models are reviewed (and corrected if required) by a human based on knowledge of fraud indicators.

Supervised learning techniques are used for three main purposes:

1) Predicting outcomes;
2) Determine causal relationships;
3) Categorisation.

Unsupervised learning:

Algorithms are used to identify patterns in the data, using only starting values. As there are no 'correct' answers, there is no human intervention in the algorithm's pattern identification.

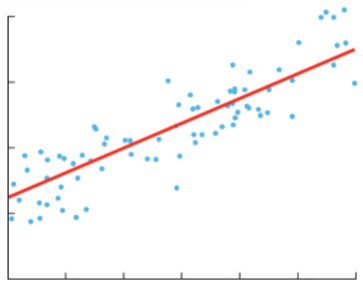
Analysis techniques are used for two main purposes:

1) Clustering values into meaningful cohorts;
2) Solving association problems, (e.g. people who have X will generally do Y).

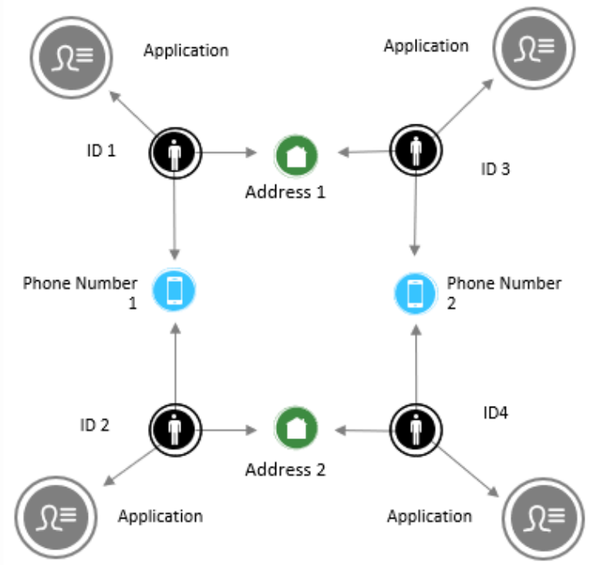
Category	Technique	Examples	Purpose of this technique	Maturity	Technology requirements	Focus areas
Rules based analysis 	**Descriptive Analytics** The use of basic statistical calculations (such as mean, median, standard deviation) as well as visualisations (such as histograms) to provide insights around datasets and their distribution. ↥ Back to Top	Statistics can highlight unusual spikes or dips in transaction activity which may indicate fraudulent activity. For example, if the difference in total value or volume of payments made to an individual is statistically significant (higher) compared to those made to similar individuals, there may be an indication of fraudulent claims.	Many fraud indicators used by counter fraud teams are based on descriptive statistics. For example, calculations of acceptable thresholds are often informed using statistical data. 	Working knowledge of statistics, data engineering, and coding	No specialist software or technology required	Detection  Identification  Decision Support 
Rules based analysis 	**Data Matching** Comparison of two or more data points to identify inconsistencies or confirm compliance. ↥ Back to Top	An example of data matching that is already used in many organisations is comparing of names to “watchlists” of known criminals or sanctioned individuals. Data matching is also used to identify abnormalities in transactions. For example, comparing payments made to a master invoice table – if a payment has been made for an invoice that does not exist in the master invoice table, this may indicate the invoice is fraudulent or has been inserted.	There are many applications for data matching. In general, this technique involves matching a dataset of unknown risk to a set of data with already identified risk factors. 	Working knowledge of statistics, data engineering, and coding	No specialist software or technology required	Detection  Identification  Decision Support 

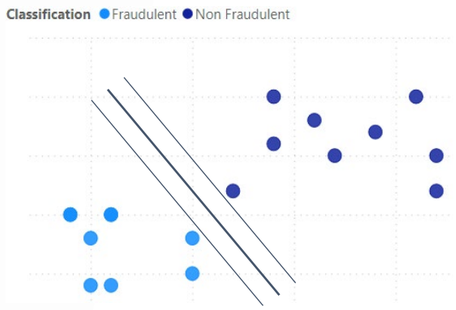
Category	Technique	Examples	Purpose of this technique	Maturity	Technology requirements	Focus areas
Rules based analysis 	**Trend Analysis** Analysis of a series of data to identify patterns and abnormalities. In most cases, trend analysis is conducted over time-series data. ↥ Back to Top	Trend analysis can be useful to analyse a single series of data, or be used to compare one series against the trend in a population. When combined with visualisation it can also be used to automatically highlight trend outliers. For example: • a gradual increase over time in the amount of data being uploaded to the internet from an employee is trend analysis of a single series of data. This could indicate an insider fraud event (leaking of confidential information). • an accounts payable manager approves invoices from a certain supplier twice as quickly as invoices from other suppliers over a period of 6 months. The trend analysis between suppliers and the accounts manager could indicate invoice fraud.	Trend analysis uses statistical techniques to identify patterns in data, and flag where any data element within a series fails to fit within the parameters of the trend. 	Working knowledge of statistics, data engineering, and coding	No specialist software or technology required	Detection  Identification  Decision Support 
Rules based analysis 	**Bayesian statistics** Bayesian statistics are calculated using Bayes theory of probability, and can also be used for supervised and unsupervised learning. ↥ Back to Top	Bayesian statistic techniques can be used for risk scoring of populations, based on known attributes. For example, it is known that 10 out of 100 claims are fraudulent. Based on this, statistically 10% of a population should be risk scored as high risk of fraud. Analysis showed that of fraud events, 50% are committed on a Sunday. This added information, using Bayes inference, allows an analyst to update the probability of any individuals flagged as high risk, increasing the risk score if they have submitted a claim on a Sunday.	These methods utilise Bayes' theorem to update probabilities once new data is obtained to express a degree of belief that an event will occur. Any new data elements that fail to fit within the probability distribution are flagged as abnormalities. 	Working knowledge of statistics, data engineering, and coding	No specialist software or technology required	Detection  Identification  Decision Support 

Category	Technique	Examples	Purpose of this technique	Maturity	Technology requirements	Focus areas
Rules based analysis 	**Fuzzy Matching** Also known as approximate string matching, this technique attempts to connect text elements that are similar but not exactly the same. ↥ Back to Top	Fuzzy matching algorithms are built into most analytical packages, to allow comparison of text elements (strings) for better data matching. For example, fuzzy matching would allow the matching of two similarly but differently spelled names (e.g. Jonathan and Johnathan) or short form of one name (John and Johnathan). Different algorithms are available to measure the "distance" between two strings of text. Examples include Levenshtein, Soundex, and Jaro-Winkler distance algorithms. Analysts should experiment with different algorithms to find the most effective for the data they are analysing.	The application of fuzzy matching techniques is important when combined with data matching to increase the percentage of matches (noting that matches on their own are not an indicator of fraud).	Working knowledge of statistics, data engineering, and coding	No specialist software or technology required	Detection  Identification  Decision Support 
Supervised Learning 	**Decision Trees** Decision tree algorithms map out logical decisions and their possible consequences/ outcomes, creating a branching structure. Training a decision tree on an existing data set provides a structure to allow new data to be classified. ↥ Back to Top	Decision tree algorithms and other predictive techniques can be used to classify populations into risk cohorts. For example, if several instances of fraud have been identified following a certain pattern, the system can "learn" which attributes are characteristic of that activity and look for those across the entire customer, vendor, or employee population. For example, an application process for a government grant is performed online, and an applicant must answer a series of questions to submit their application. Of all the applications submitted, approved, and paid, fraud has been detected in 1%. Each of these applicants is flagged in the data. Assuming that there is a correlation to how questions have been answered and what turns out to be a fraudulent application, a decision tree algorithm will learn from the data, and classify all applicants into categories, based on which applicants were flagged as fraudulent and the questions that have been answered. This classification structure can then be applied to all new applicants, with a risk score and classification applied based on their answer to questions. This allows higher risk applications to be diverted for additional review steps.	From a fraud analytics perspective, decision tree algorithms and other predictive techniques can be used to classify populations into risk cohorts. While decision trees are easy to visualise and implement due to their simplicity, they usually do not perform as well as other predictive algorithms.	Intermediate knowledge of statistics and data engineering, advanced coding	Specialist analysis software packages Specialist hardware with sufficient RAM Large volumes of data storage	Population Profiling  Identification  Decision Support 

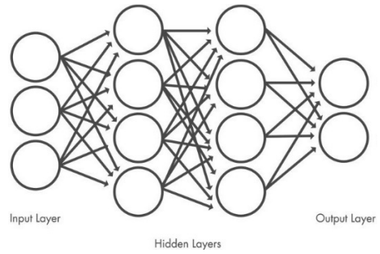
Category	Technique	Examples	Purpose of this technique	Maturity	Technology requirements	Focus areas
Supervised Learning 	**Regression Analysis** Regression algorithms attempt to understand if there is a relationship between a dependent variable (the data point that is being analysed) to independent variables (data points that are of unknown impact). ↥ Back to Top	In its simplest form, regression compares two different factors and determines if there is a correlation. For example, in the case of application fraud, there is a hypothesis that the amount of fraud increases when an applicant has a poor credit rating. A simple linear regression would map these two factors on a graph (X axis the amount of fraud, Y axis a worsening credit rating). In this instance, the amount of fraud is the dependent variable, and the credit rating the independent variable.  The regression line (in red) describes the relationship between the two factors, confirming the relationship. Now suppose in the example of application fraud, there are hundreds of independent variables (questions answered by applicants). Using multivariate regression allows the analyst to identify which of these variables (or combination of variables) are significant. The resultant output is a prediction of the probability that the application is fraudulent, for any given individual, based on the attributes of that individual's application. This differs from classification, as the result is not a category, but a value.	Where data is complex with a large number of attributes, use of regression techniques allows fraud data analysts to identify the data points that are likely to indicate fraud and flag these for further investigation. There are many regression algorithms and techniques that have been developed. Common regression techniques include: • Linear regression: Attempts to use a linear equation to model the relationship between variables, minimising the error for each provided datapoint. • Lasso regression (Least Absolute Shrinkage and Selection Operator): a type of linear regression that reduces parameters and "shrinks" the model by penalising overfit coefficients. • Logistic Regression: A statistical model that is an extension of linear regression, using the logistic function, to classify data based on a dichotomous outcome. • Multivariate regression: Regression analysis that models 2 or more dependent/outcome variables • K-Nearest Neighbour: A classification/regression algorithm that classifies new datapoints based on that of its k-nearest neighbours.	Intermediate knowledge of statistics and data engineering, advanced coding	Specialist analysis software packages Specialist hardware with sufficient RAM Large volumes of data storage	Population Profiling  Identification  Decision Support 

Category	Technique	Examples	Purpose of this technique	Maturity	Technology requirements	Focus areas
Supervised Learning 	**Natural Language Processing** Algorithms that have been developed to examine and convert text data (unstructured) to structured data, to facilitate further analysis. ↥ Back to Top	For example, a government agency is providing hardship payments to individuals who have been affected by a natural disaster. In order to apply for a payment, an individual must call a hotline, which is recorded, and speak to an operator who determines eligibility for payment. To detect fraud in this situation, an agency manually audits a sample of recordings by listening to the call and to determine if the applicant is indeed eligible. Out of the sampled calls, 1% are determined to be fraudulent through collusion between the applicant and the operator. Using natural language processing, instead of relying on manual auditing, the entire population of recordings can be screened for fraud. Recordings are converted into text through speech recognition. The text is then processed through algorithms to detect known patterns of words and phrases that are of high risk for fraud, which are flagged for review.	Natural language processing algorithms can be used to process text data sets that may contain indicators of fraud, for example, complaints data, social media feeds, or call centre recordings. Natural language processing is often used to “mine” text data to identify themes, keywords, or emotions within the text. For example, in call centre transcripts it can be used to identify particular calls which left the customer unsatisfied/angry or call operators who seem to perform poorly in servicing callers.	Intermediate knowledge of statistics and data engineering, advanced coding	Specialist analysis software packages Specialist hardware with sufficient RAM Large volumes of data storage	Population Profiling  Identification  Decision Support 

Category	Technique	Examples	Purpose of this technique	Maturity	Technology requirements	Focus areas
Supervised Learning 	**Network Analytics** Establishes relationships between entities based on common data attributes to create a network of activity. ↥ Back to Top	For example, in an application for a government grant, 4 individuals apply and submit their details, including address and phone number (which is required for contact purposes). These individuals are from a criminal gang, looking to defraud the government. In order to not raise suspicions, the individuals use different combinations of two addresses and two phone numbers, so that no applicants have the same details. Using this technique, the relationship between the individuals is identified as a network of relationships is built per the diagram below:  On a larger scale with many more attributes, networks can span beyond simple relationships and identify fraudulent activity.	Network analytics are enabled by graph database technology, where each data element is stored based on its matching of attributes to other data elements. Network analytics have been used in fraud analytics to identify fraud networks, based on matching attributes between different customers that identify undeclared and unknown relationships. This is particularly useful in situations where fraudulent activity is executed by a network of actors or where relationships are masked through a series of complex company ownership structures.	Intermediate knowledge of statistics and data engineering, advanced coding	Specialist analysis software packages Specialist hardware with sufficient RAM Large volumes of data storage	Population Profiling  Identification  Decision Support 

Category	Technique	Examples	Purpose of this technique	Maturity	Technology requirements	Focus areas
Supervised Learning 	**Support vector machines** The algorithmic equivalent of drawing a line (or plane) through data to separate it into different categories by identifying the plane which is the furthest from the datapoints in the training dataset. ↥ Back to Top	The use case for support vector machines in fraud is similar to decision trees, differing in the approach taken to solving classification. As with decision trees, a training data set is required where known instances of fraud have been labelled. In the example of application fraud, when training a support vector machines model, each application in the data set is mapped into *n* dimensional space, where *n* is the number of features (questions in the application). Each application is also labelled as being either fraudulent or non fraudulent. The algorithm attempts to find an optimal line to separate every application into categories of fraudulent/non fraudulent. The line itself is the furthest distance (margin) from the nearest data point.  In the example above, the middle line is the optimal ‘plane’, furthest from the data points. By applying this theory across the entire training data set, a classification model is produced, which can then be used against new applications to class them as potentially fraudulent or non fraudulent.	Support vector machines can be used for either classification or regression, but is more widely used for classification problems. Support vector machines are commonly used in text classification, linked to natural language processing algorithms, to identify topics, sentiment, or intent within the text. For example, identifying complaints within large populations of feedback.	Intermediate knowledge of statistics and data engineering, advanced coding	Specialist analysis software packages Specialist hardware with sufficient RAM Large volumes of data storage	Population Profiling  Identification  Decision Support 

Category	Technique	Examples	Purpose of this technique	Maturity	Technology requirements	Focus areas
Unsupervised Learning 	**Clustering algorithms** Clustering refers to the process of dividing data points into groups (clusters), so that each data point in a cluster has more in common with their cluster than with other clusters. ↥ Back to Top	Clustering has a number of applications within fraud management. Most commonly it is used to stratify populations by their behaviour to allow more accurate monitoring of activity. For example, a clustering algorithm might identify 5 natural groups of customers within a population with one of those populations representing a cohort of large corporations. If one of those large corporations suddenly began behaving more like an individual, systems would be able to detect that change and flag the customer for review. Clustering groups data points together based on similarities in their attributes. This allows identification of outliers within specific clusters that warrant further investigation. For example, suppose a group of 100 healthcare providers submit claims to a government agency for treatment of a disease. The attributes available in this data set may include if the provider is public, private, or an individual, the treatment type and cost, age of patients treated, etc. Using clustering, each provider is clustered with providers that have similar characteristics, using one of the clustering techniques listed:  By clustering providers into cohorts, further analysis can be performed to identify outliers within each cluster, rather than across the entire population, with outliers scrutinised and investigated for potential fraud.	It is one of the most widely used unsupervised learning techniques. From a fraud analytics perspective, clustering can be used to classify entity populations into cohorts. There are many types of clustering algorithms, including: • K-means clustering - Algorithm that assumes similar things exist in close proximity in the data. It creates *k* "clusters" in order to minimise the distance between datapoints within each. • Mean-shift clustering - An algorithm that creates and uses a probability density function describing known data to measure distance. Can be used alone or to improve upon *k*-means outputs. • Hierarchical clustering - Creates a hierarchical, branching approach to finding clusters, similar to decision trees.	Advanced knowledge of statistics, data engineering, and coding	Specialist analysis software packages Specialist hardware with sufficient RAM Large volumes of data storage	Population Profiling  Identification  Decision Support 

Category	Technique	Examples	Purpose of this technique	Maturity	Technology requirements	Focus areas
Unsupervised Learning 	**Neural Networks** Modelled loosely on the way the human brain works, neural networks create a series of decision layers (nodes) that "weight" their inputs and "fire" if the result is above a threshold. ↥ Back to Top	Neural networks work by repeatedly building predictive models (hidden layers) which operate in sequence to give a likelihood of one or more outcomes. As such, they are able to use a complex set of inputs to determine key pieces of information about a situation. For example, there is a requirement to not only classify applications as either potentially fraudulent or non-fraudulent, but also to apply a probability to the applications classified as potentially fraudulent. This could be achieved by performing the two tasks separately; using separate techniques to classify and calculate probabilities. However, this is problematic if the results produced by each model are inconsistent or diverge. A neural network is able to take a large set of inputs (features) such as the key data in the application together with the known customer information and use training data, in this case past instances of fraud, to build a model capable of both identifying likely application fraud and presenting that likelihood.	In a neural network, nodes are defined in three layers; input nodes contain the data that is accepted into the model, hidden nodes take the input and perform calculations or decisions, and output nodes contain the results of the problem.  Neural networks can be either supervised or unsupervised, and perform regression, classification, or a combination of both (multi-output). Neural networks (also known as Artificial Neural Networks) can be applied to even the most complex problems; they are used on the forefront of machine learning research. In fraud detection, they can learn from large quantities of historical data including known fraud cases to build a model that can be applied going forward for detection. The system can be retrained periodically to continue to refine to model and increase its accuracy.	Advanced knowledge of statistics, data engineering, and coding	Specialist analysis software packages Specialist hardware with sufficient RAM Large volumes of data storage	Population Profiling  Identification  Decision Support