protiviti®

Face the Future with Confidence

Application of Machine Learning in Internal Audit for Sample Selection

By Denis Lippolt and Xeniya Kozina

Nowadays more and more companies deploy Artificial Intelligence and Machine Learning (AI/ML) into their every-day operations. Vast capabilities of AI/ML provide more comprehensiveness and are of a great help for increasing efficiency and generating new insights. In many operations and processes AI/ML algorithms can effectively serve specialists a helping hand, especially when it comes to analysis of data.

When analysing and addressing risks, Internal Audit functions deal with analysis of large amount of data. And it is exactly here where ML algorithms can be very helpful to identify hidden patterns and abnormal observations in the data and therefore the most risky elements in a given audit area, increasing efficiency, effectiveness and overall risk coverage in the Internal Audit process.

In the following we give insights into applying an unsupervised learning algorithm named Isolation Forests in the sample selection step of Internal Audit projects.

Sample Selection

Every-day business of any organisation inevitably involves dealing with uncertainties, that may include unknown, unpredictable, or unexpected events as well as the cases when the information is lacking. As a result, this can influence the expected or planned outcome of a process or, in general, any objective of a company. According to the International Organisation for Standardisation such "effect of uncertainty on objectives" is defined as risks. [14] From here it becomes clear that almost every single process in an organisation is related to risks. Understanding and learning from risks helps to reduce and mitigate them and, thereby, to assure effective achievement of business goals and operational effectiveness. Therefore, in a modern organisation through all the levels risk and control functions perform their responsibilities employing various risk management frameworks. Nowadays these functions often become more and more diverse and separated, that, of course, requires a systematic approach of how to assure their more effective functioning. Such an approach is provided by the "three lines (of defense)" model, where the "first" line is assigned to the operational management and various risk and compliance functions are within the "second" line. Internal audit function forms the "third line" and plays one of the key roles in an undertaking. The broad range of internal audit responsibilities includes i.a. performing assessment and evaluation of effectiveness and appropriateness of the entire risk management and internal control system of the respective organisation, how effective the "first" and the "second" lines are as well as of compliance to internal policies and procedures, laws and external regulative requirement. [10,11,12,14,17]

The Audit function, in comparison with the two other lines, has a high level of independence in an organisation, and at the same time, an active reporting line to senior management. This enables it with the capacities to provide the management board and senior management with an objective and independent assessment, consulting and

insights aiming for improvement of all business processes and company corporate governance. Employment of systematic, risk-oriented and process-independent approach helps internal audit in its day-by-day work to achieve its objectives.

The results of internal audit work have a direct impact on vital business decisions of a company. Therefore, further techniques that make it possible to extend and improve the capabilities of audit analysis, are of high importance.

During the Audit process, internal auditors investigate and test the design and effectiveness of controls and groups of controls, which address certain identified inherent risks and whether certain controls exist or not may also be in scope of audit testing activities. [13] Testing should provide reasonable evidence on how effective the perceived risk is mitigated by controls.

Which types of risks might be then mostly in focus of audit activities? Although there may be any of them (depending on the current business situation, organisational structure, external regulatory activities, new development activities in an organisation etc..), the audit investigations are mainly targeted on operational risks. In general the loss distribution function of operational risks is heavy tailed, meaning that among many events with relatively noticeable probability, beyond specified Value-at-Risk (VaR), extremal "black-swan" (high impact, low probability) events can appear that are not in all the cases covered by economic capital. [15,16] Addressing these risks is often "identifying a needle in a haystack". And it is where the population size starts to play an important role in the audit process. For a small population it is possible to check all the items, while for a large population it becomes more complicated. In the case one can sufficiently specify what constitutes an exception and identify such instances in the population using data analytics techniques, employment of full population testing techniques is still feasible. Usually, however, a certain part (sample) from the entire population of data is being selected and the auditor draws the conclusion about the entire population based on the analysis of items in the sample.

Depending on the objectives of the particular audit engagement, different methods can be applied. Quite often to form an audit opinion a judgmental approach to sampling is being used, especially where it is about testing residual risks. In this case, an auditor, based on experience, good knowledge of processes or on available information from other sources designs the sample containing "risky" items without application of statistical methods. [13] The items, containing "non-risky" elements may also be included in the sample, especially by the necessity to test a control based on the knowledge of how a standard process is functioning and how the risks are mitigated by this control. However, such "normal" instances should be already addressed by well-functioning controls of the existing internal control system (ICS), since otherwise it would have already led to certain difficulties or problems and would have been known before the audit is taking place. Therefore, the "normal" transactions or items in the entire population, reflecting business-as-usual processes, do not contain too much risk, and those items that impose risk can be considered "abnormal". Such items are then mainly in focus and covered by audit activities. Hence it is very desirable that a sample consists mainly of those items from the entire population.

In practice, a heuristic or judgmental approach to sampling is very common and widespread, although, designing a sample by selecting "risky" items just "by hand" may impose further or, in some cases, enhance sampling risks – the risk that the conclusion based on the sample analysis might differ from that one drawn from the entire population analysis (if it had been performed). [13] Moreover, "abnormal" items may follow quite complicated patterns that can contain not only one deviating feature, but also comprise combinations of several parameters. Hence, it is not always possible to identify them by application only of a judgmental approach.

Modern developments in AI/ML now may provide the opportunity to go beyond the capabilities of heuristic approaches for identifying anomalies. In this paper we are going to present one of these methods, that can be successfully applied and enhance the effectiveness of sampling.

Isolation Forest

Apart from other established methods,[1] a high-performance identification of anomalies has recently become feasible due to development and implementation of the Isolation Forest (IF) algorithm. First introduced in 2008 by Liu et. al. [1] the method is currently widespread and being applied in many areas where identification of anomalies is of a significant relevance, such as finance, IT technologies, engineering, astronomy and seismology. In astronomy it is being used for visualisation, detection and complex analysis of measured datasets from different astronomical objects. [3]. Application of IF in geophysics enables detection and prediction of events [4]. Spotting anomalies in streaming data can serve for resolution of cyber security problems, where abnormal patterns may be indicative of system intrusion events. [5,6] Detection of fraudulent credit card activities and unusual behaviour patterns of third-party agents has recently become feasible due to application of such ML methods in the finance domain. [7,8]

The Isolation Forest algorithm draws advantage from the characteristics inherent to anomalies in a given dataset: (i) being very minor and (ii) having very different attribute-values (being very divergent) from genuine data points using an absolutely different concept for searching anomalies. With its name speaking for itself, instead of profiling through the normal values, where the algorithm learns what the "normal" observations in the dataset are and then assigns all out-of-model ones to anomalies, Isolation Forest method isolates the instances that are abnormal [1]. Moreover, superior characteristics of Isolation Forest in comparison with other methods, based on density and distance measures, such as ability to process high dimensional datasets, low memory requirement, short runtime and detection accuracy were later proven and presented by the same authors in 2012 [2].

Common datasets in finance (and consequently those used during audit procedures) are very often multidimensional. Usually the abnormal observations in such datasets constitutes a very small fraction of the entire population and can follow very unpredictable or divergent from normal observations patterns. Hence, to pool a sample, containing anomalies, just "by eye" from a given dataset very often turns out to be complicated. Application of Isolation Forest algorithm here is beneficial as it significantly simplifies (or in the cases, when one speaks about retrieving only anomalies, makes it even possible) the common sampling techniques employed in Audit. This method, of course, only leads to substantial results under the assumption, that the information contained in the anomalies is quite different from that in the rest of the dataset; or in other words, when the target entries with high underlying risks are anomalies for a given dataset. Construction of the relevant dataset, corresponding to the specific audit project objectives and scope, reflecting relevant business considerations as well as enriching data with external data per discretion/input inevitably requires the expertise and experience of an auditor.[2] Therefore here we propose to consider the method foremost as a complementary and extremely practical tool and not a solely self-sustained mechanism entirely enabling Audit procedures from the beginning till the end.

In order to identify the risky items in a dataset the supervised learning algorithms can be used, which implies additional labelling process. In many cases it becomes almost impossible or at least very time-consuming and expensive to label the set of data. In contrast to this, the isolation forest can be effectively used in the unsupervised mode, meaning that no prior labelling of data is required, and thus be very beneficial for applications in Audit.

So how does it work? Isolation Forest algorithm

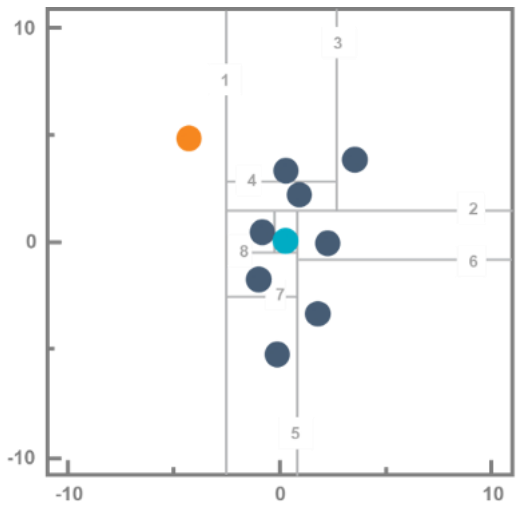
[1] In contrast to the approach used in Isolation Forest algorithm, most of the methods used so far for anomaly detection (e.g. classification-based methods, Replicator Neural Network, one-class SVM and clustering-based ones) profile normal instances and identifying anomalies as those, that do not conform a "normal" pattern [2]

[2] As examples, the information on weekdays obtained from dates can help to monitor occurrence of transaction outside the business days; location information can be retrieved from IBAN or from ZIP codes; for more exact spatial information, addresses can be enriched with latitude/longitude information etc.

separates anomalies away from the entire dataset, based on the random partitioning. When the data are fed, from the entire multivariate population the algorithm first choses randomly a feature, then, between its minimal and maximal values it chooses a random value, followed by a separation of the original dataset into two sub-sets.

In 2D space this can be illustrated/imagined as drawing a line perpendicular to an axis, while in higher dimensional space the separation into the sub-sets is correspondingly being made by a separation (hyper-)plane. (See Figure 1)

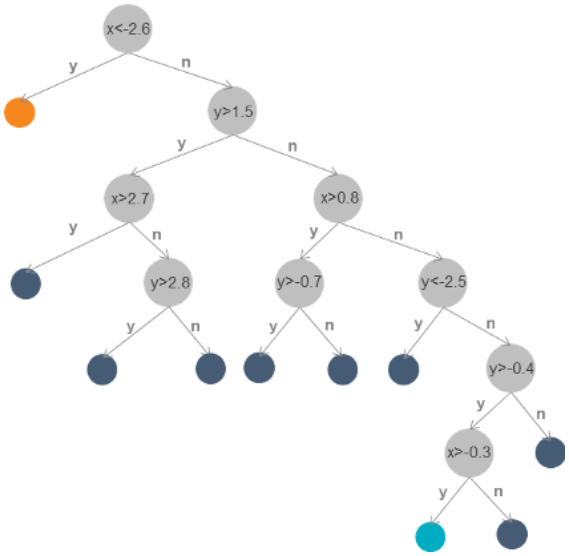
Figure 1 Illustrates how the algorithm step by step isolates the points in a 2D dataset.



The lines that at every step separate the current set into the two sub-sets are marked with numbers. The anomaly (marked in orange) is separated earlier than all other points (here at the first step). The algorithm termination point (marked in blue) is separated at the latest step (here it required 9 partitions) and all the points, isolated at the intermediate stages are marked with dark blue.

After that, the previously described steps are being repeated recursively on each sub-set until all the observations are isolated to the external node. As every separation is based on "exclusive-OR" operation, the algorithm can be represented by a decision tree structure, where every node is depicted by a random partitioning step (see Figure 2). Path length or the number of nodes, required for an instance to be separated from the very root of the tree to the external node, is shortest for anomalies.

Figure 2 A decision tree, built by the isolation forest algorithm



A decision tree, built by the isolation forest algorithm (presents one possible tree in isolation forest ensemble) in order to separate all the points in the dataset (2D dataset, depicted in Fig.1). As one can see the length from the upper node to the anomaly (orange) is the shortest. The path length to the termination point (blue) is much longer.

This also makes clear the fact that in order to isolate anomalies in a dataset it is not necessary to use the entire path length to the termination point at the very last datapoint. In many cases it is quite sufficient to define a shorter depth (path length) that, consequently, shortens execution time as well.

During execution of the Isolation Forest algorithm an entire ensemble (or Forest) of binary decision trees is being built, each having a random set of partitions; with the number of trees in the forest as a parameter, that can be defined.[3] Since for each tree in an ensemble the path length to every observation is defined, anomaly score, classifying instances, can be calculated. For every point the path length then is being averaged after all recursive operations over the entire number of trees. [1] Thus, the anomaly score is defined as $S(x) = 2^{-\frac{E(h(x))}{c(n)}}$, where E(h(x)) is the average of path lengths h(x) for a given observation and c(n) is an average path length of all terminations to external node (normalisation parameter). [1, 2] As it follows from this formula, the anomaly score is inversely proportional to the average path length, thus, giving the possibility to classify observations with S(x) diverging to 1 as anomalies; those having S(x) significantly smaller than 0.5 as normal observations, and the dataset where all the instances got the anomaly score close to 0.5 to consisting from entirely normal observations. [1] Thus, on a training stage the algorithm builds the trees and on the testing stage, going through the entire set of data points, calculates the number of nodes for each tree and hence evaluates the anomaly score. [2]

Employment of open source software nowadays becomes very widespread in Data Science. From the variety of programming languages that currently are being used, R and Python emerged as dominating and offering variety of instruments and standard solutions for Data Analytics purposes, and, in particular, for detection of outliers.

Standard implementation of isolation Forest algorithm can be realised by means of "solitude" and "scikit-learn" packages in R and Python respectively and the solutions established in R and Python can be further utilised in many BI Platforms for further analysis. Further combination with visualisation capabilities of all these software resources can make analysis much easier, more understandable and recognise the items of interest faster in comparison with common methods.

Conclusion and outlook

When dealing with high dimensional data application of different ML techniques is very beneficial. In finance the items or transactions associated with risks are rare and very often hardly detectable and this is where modern developments in the field of data analytics can help. Their employment in every-day processes of audit function is already feasible. The suggested here isolation forest method directly provides audit with the necessary information on the most "risk-containing" items in the shortest time and capturing the information from the entire population. At the same time, being applied additionally to the common audit procedures such methods can serve an additional source of information playing suggestive role in what to investigate in more detail or pay especial attention to. This assures the new level of efficiency, provides further improvement of common routines and as a result delivers new previously unattainable information and deeper understanding of all the processes and herewith opening new ways and providing capacities for further improvements in an organisation. The information gained can provide direct insights for the audit and therefore necessary input for timely decision-making to the senior management.

[3] As usual in „forest" algorithms, this allows for multi-threading and therefore efficient performance of the algorithm.

References

[1] Fei Tony Liu, Kai Ming Ting, Zhi-Hua Zhou, Isolation forest, 2008 Eighth IEEE International Conference on Data Mining; DOI: 10.1109/ICDM.2008.17

[2] Fei Tony Liu, Kai Ming Ting, Zhi-Hua Zhou, Isolation-based anomaly detection, ACM Transactions on Knowledge Discovery from Data (TKDD), 2012; DOI: 10.1145/2133360.2133363

[3] D. Baron et al, ML in Astronomy: a Practical Overview, 2019: http://research.iac.es/winterschool/2018/media/summaries/ml_summary_dbaron.pdf

[4] C. Hulbert et al, A silent build-up in seismic energy precedes slow slip failure in the Cascadia subduction zone, 2019, https://arxiv.org/abs/1909.06787

[5] N. S. Arunraj et al, Comparison of supervised, semi-supervised and unsupervised learning methods in network intrusion detection systems (NIDS) application, 2017, ISSN: 2296 – 4592: https://ojs-hslu.ch/ojs302/index.php/AKWI/article/view/89

[6] Zh. Ding et al, An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window, 3rd IFAC International conference on Intelligent Control and Automation Science, 2013, Chengdu, China

[7] M. R. Miller et al, Sleuthing for adverse outcomes: Using anomaly detection to identify unusual behaviors of third-party agents, Proceedings of Machine Learning Research 71:121-125, 2017 KDD 2017: Workshop on Anomaly Detection in Finance

[8] H. A. Shukur et al, Credit card fraud detection using machine learning methodology, International Journal of Computer Science and Mobile Computing, Vol. 8, Issue. 3, March 2019, pg. 257 – 260

[9] International Auditing Standard (IAS) 530

[10] Anlage 1: Erläuterungen zu den MaRisk in der Fassung vom 27.10.2017, BaFin, 2017

[11] Rundschreiben 2/2017 (VA) - Mindestanforderungen an die Geschäftsorganisation von Versicherungsunternehmen (MaGo)

[12] Internationale Grundlagen für die berufliche Praxis der Internen Revision 2017 Mission, Grundprinzipien, Definition, Ethikkodex, Standards, Implementierungsleitlinien, DIIR, 2017

[13] White Paper – Internal Audit Sampling, The Institute of Internal Auditors–Australia, 2017

[14] International standard ISO 31 000:2009(E)

[15] S. Strzelczak, Operational risk management, 2007, https://www.researchgate.net/publication/312491702

[16] P.Teplý et al, The theoretical background of operational risk management, Published in International Conference on Education and Management Technology, 2010, DOI:10.1109/icemt.2010.5657656

[17] IIA Position Paper: The Three Lines of Defense in Effective Risk Management and Control, IIA, 2013

Contacts

Peter Grasegger
Managing Director
+49.173.653.8922
peter.grasegger@protiviti.de

Denis Lippolt
Associate Director
+49.172.698.3048
denis.lippolt@protiviti.de