WHITE PAPER

INTERNAL AUDIT APPLICATIONS OF MACHINE LEARNING SAMPLE SELECTION

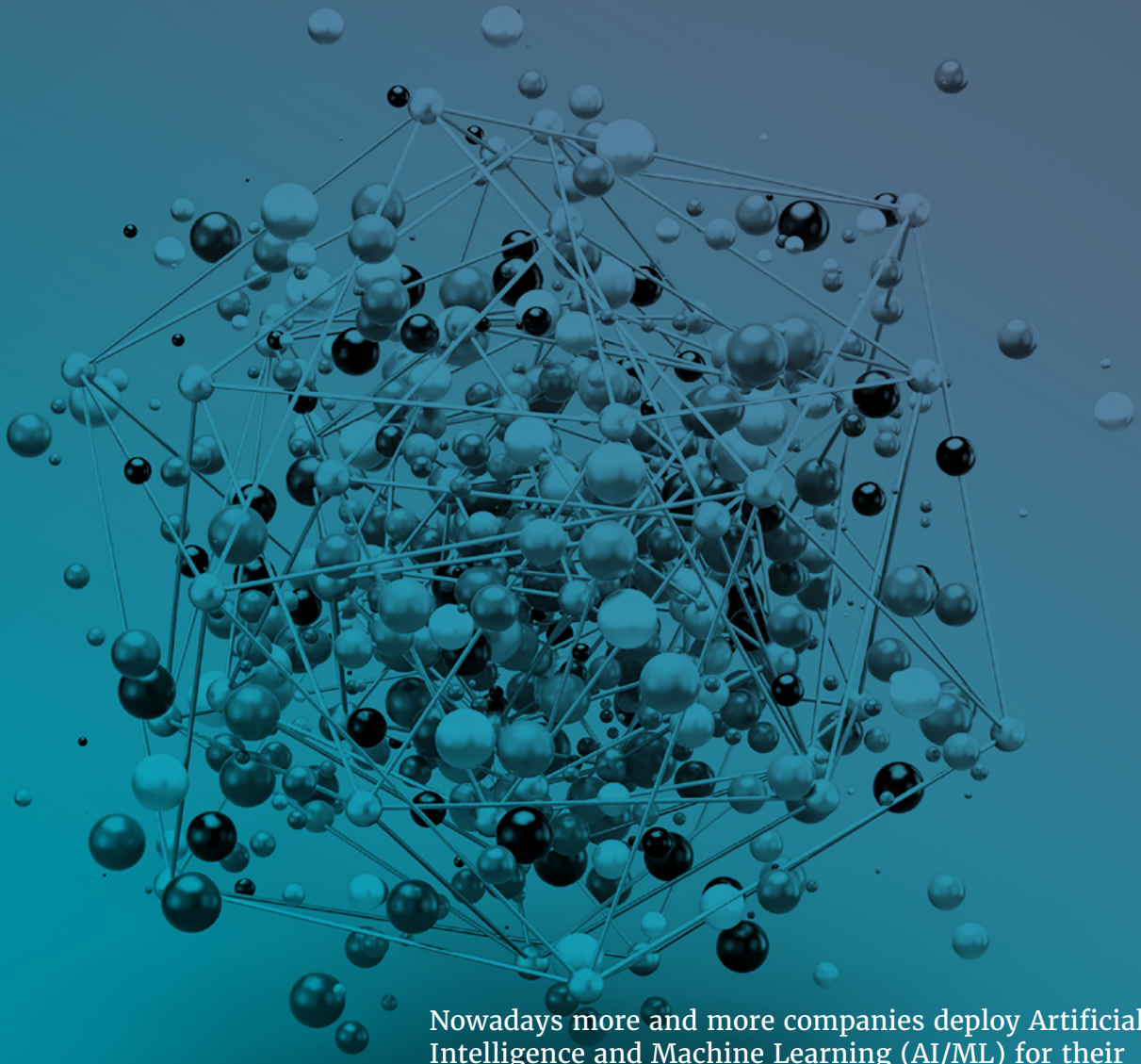


Foto: getty images

Nowadays more and more companies deploy Artificial Intelligence and Machine Learning (AI/ML) for their everyday operations. The immense capabilities of AI/ML allow for large-scale analysis, greatly increase efficiency, and generate new insights. In many operations and processes AI/ML algorithms can very effectively support specialists, particularly when it comes to data analysis.

AUTHORS: DENIS LIPPOLT, DR. XENIYA KOZINA

protiviti®
Global Business Consulting

WHITE PAPER

INTERNAL AUDIT APPLICATIONS OF MACHINE LEARNING SAMPLE SELECTION

Foto: iStock

CONTENT

protiviti®
Global Business Consulting

When analysing and addressing risks, internal audit functions deal with the analysis of large amounts of data. This is exactly where ML algorithms can be extremely helpful to identify hidden patterns and anomalies in the data. Isolating these most risky elements in a given audit area increases efficiency, effectiveness and overall risk coverage in the internal audit process.

Modern developments in AI/ML enable going beyond the capabilities of heuristic approaches for identifying outliers. This paper outlines one of these methods, which can be successfully applied to enhance the effectiveness of sampling: Isolation Forest, an unsupervised learning algorithm, in the sample selection step of internal audit projects.

UNDERSTANDING & MITIGATING RISK

The day-to-day business of any organisation inevitably involves dealing with uncertainties, including unknown, unpredictable, or unexpected events as well as cases when information is lacking. This can impact 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. [11]

As this applies to almost every single process in an organisation, understanding risks helps to reduce and mitigate them to ensure the achievement of business goals and operational effectiveness. Therefore, modern organisations have to operate at all levels with risk and control functions that employ various risk management frameworks.

"THREE LINES" APPROACH

These functions become increasingly diverse and separated and a systematic approach is needed for their effective functioning. One such approach is provided by the "three lines" model, where the first line is assigned to the operational management while various risk and compliance functions are within the second line. Internal auditing forms the third line and is key for any company.

The broad range of internal audit responsibilities includes assessing and evaluating how effective and appropriate the entire risk management and internal control system of the respective organisation is, how effective the first two lines are as well as assessing compliance with internal policies and procedures, laws, and external regulative requirements. [7,8,9,11,14]

The audit function, in comparison with the two other lines, is performed very independently in an organisation. At the same time there is an active reporting line to senior management. This enables auditors to provide the management board and senior management with an objective and independent assessment and offer consulting and insights aiming for improvement of all business processes and company corporate governance. Employing a systematic, risk-oriented, and process-independent approach helps internal auditors to achieve their objectives in their daily work.

TECHNOLOGY-ENABLED INTERNAL AUDITS

The results of internal audits have a direct impact on vital business decisions. Tools improving the analysis capabilities of audits are therefore of high importance.

»Modern developments in AI/ML enable going beyond the capabilities of heuristic approaches for identifying outliers.«

DENIS LIPPOLT, SERVICE LINE LEAD
ADVANCED ANALYTICS, PROTIVITI GERMANY

»The results of internal audits have a direct impact on vital business decisions. Tools that improve the analysis capabilities of audits are therefore of high importance.«

DENIS LIPPOLT, SERVICE LINE LEAD ADVANCED ANALYTICS, PROTIVITI GERMANY

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. Whether certain controls exist or not may also be part of audit testing activities.[10] Testing should provide reasonable evidence on how effective the perceived risk is mitigated by controls.

CHALLENGES OF RISK DETECTION

Any type of risk might be in the focus of audit activities, depending on the current business situation, organisational structure, external regulatory activities, or new developments in an organisation. However, investigations mainly revolve around operational risks. In general, the loss distribution function of operational risks is a heavy tailed one, meaning that among many events with relatively noticeable probability, beyond specified Value-at-Risk (VaR) extreme so-called "black-swan" (high impact, low probability) events can appear that are not in all the cases covered by economic capital. [12,13]

Addressing these risks often comes close to identifying a needle in a haystack. At this stage of the audit process, the population size (data set from which a sample is selected) becomes very important. For a small population it is possible to check all the items, while for a large population it becomes more complicated. If it is possible to sufficiently specify what constitutes an exception and identify such instances in the population using data analytics techniques, the use of full population testing techniques is still feasible. Usually, however, a sample from the entire data set is being selected and the auditor draws conclusions about the entire population based on the analysis of items in the sample.

DESIGNING SAMPLES

Depending on the objectives of the particular audit, different methods can be applied. Quite often a judgmental approach to sampling is deployed for auditors to come to an opinion, especially in the case of testing residual risks. In this case, an auditor, based on experience, information available from other sources and good knowledge of processes designs the sample containing "risky" items without applying statistical methods. [10] Items containing "non-risky" elements may also be included in the sample, especially 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 already be addressed by well-functioning controls of the existing internal controlling system (ICS). Otherwise, it would have led to certain difficulties or problems and would have been known before the audit 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". Thus, audit activities mainly focus on these items and samples therefore should mainly consist of these 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 manually may impose further or, in some cases, enhance sampling risks, i.e., the risk that the conclusion based on the sample analysis might differ from the one drawn from the entire population analysis (if it had been performed). [10]

Moreover, outliers 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 only applying one's own judgment.

ISOLATION FOREST WITH SUPERIOR CAPABILITIES

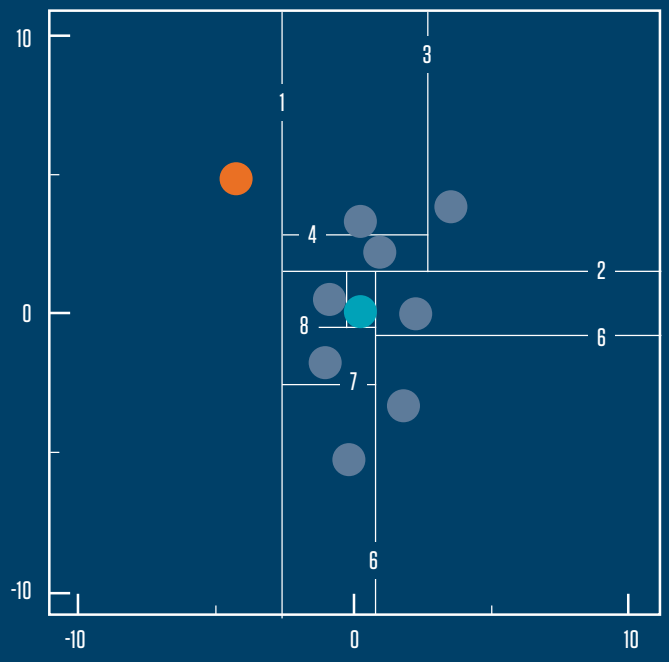
Apart from other established methods, the Isolation Forest (IF) algorithm has become a highly effective way of identifying anomalies. First introduced in 2008 by Liu et. al. [1], it is now being widely used and applied in many areas where identification of anomalies is of key importance, such as finance, IT technologies, engineering, astronomy and seismology. Spotting anomalies in streaming data can serve for resolution of cyber security problems, where abnormal patterns may be indicative of a system intrusion events. [3,4] Finance applications include, e.g., the detection of fraudulent credit card activities and unusual behaviour patterns of third-party agents. [5,6]

The IF algorithm makes uses of the characteristics inherent to anomalies in a given dataset, i.e., being very minor and having very different attribute-values (being very divergent) from genuine data points. **It uses an absolutely different concept for searching anomalies:** Instead of profiling based on 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, the IF algorithm isolates the instances that are abnormal [1]. Moreover, the superior performance of this method in comparison with others, based on density and distance measures, such as

- → the ability to process high dimensional datasets
- → low memory requirement
- → short runtime and
- → detection accuracy

have also been proven [2].
Common datasets in finance – and consequently those used during audit procedures – are very often multidimensional. Usually, outliers in such datasets constitute a very small fraction of the entire population and can follow very unpredictable observation patterns or observation patterns that diverge from normal ones. Hence, pooling a sample containing anomalies just "by eye" from a given dataset can get very complicated.

FIGURE 1:
ILLUSTRATION HOW THE ALGORITHM ISOLATES THE POINTS STEP BY STEP IN A 2D DATASET.



SIMPLIFYING SAMPLING

The IF algorithm significantly simplifies the common sampling techniques applied in internal audits. In fact, in the cases of retrieving anomalies only, it enables the whole process. 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.

To build a relevant dataset, which corresponds to the specific audit project objectives and scope and reflects relevant business considerations as well as enriches data with external data per discretion inevitably requires the expertise and experience of an auditor. We therefore recommend using this method mainly as a complementary and extremely practical tool and not as a self-sustained mechanism that can enable audits from beginning to end.

In order to identify the risky items in a dataset supervised learning algorithms can be used, which

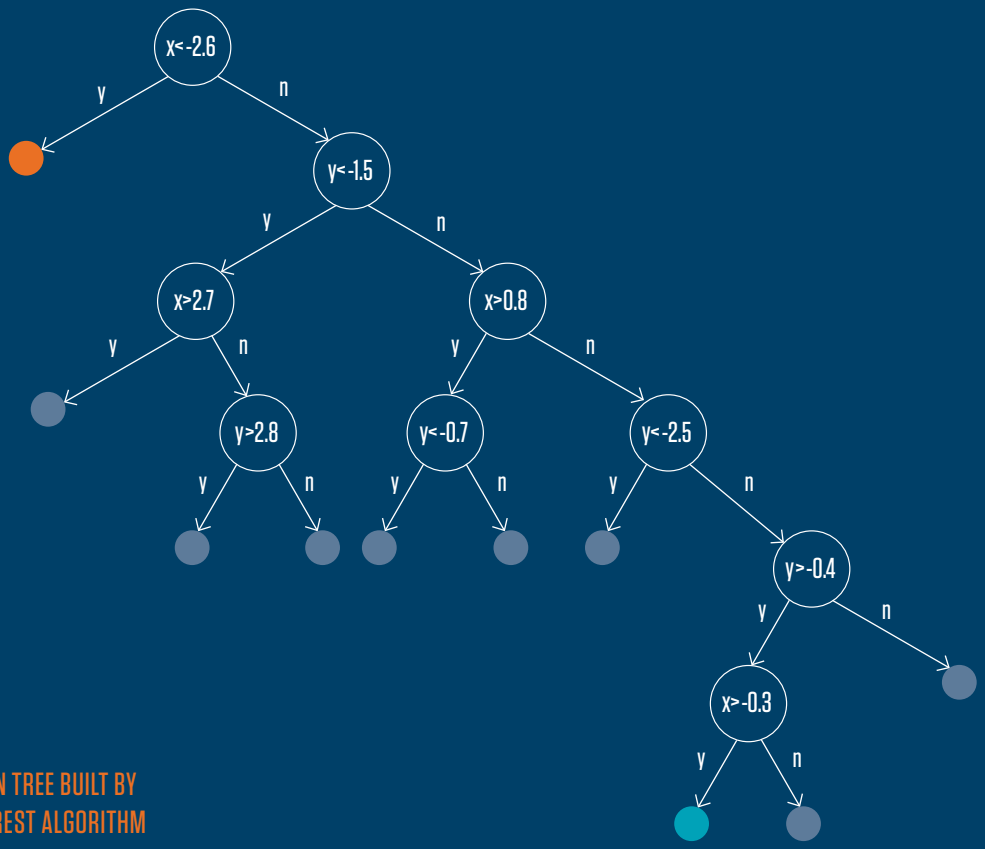


FIGURE 2: DECISION TREE BUILT BY THE ISOLATION FOREST ALGORITHM

involves an additional labelling process. In many cases it becomes almost impossible or at least very time-consuming and expensive to label the data set. In contrast to this, the IF algorithm can be used in an unsupervised mode, meaning that no prior labelling of data is required, which is very beneficial for audit applications.

STEP-BY-STEP: HOW ISOLATION FOREST WORKS

The IF algorithm separates outliers from the entire dataset, based on random partitioning:

1. When the data are fed, the algorithm first randomly chooses a feature from the entire multivariate population.
2. Next, a random value is chosen between its minimal and maximal values.
3. The original dataset is divided into two sub-sets.

In 2D, these steps can be illustrated by 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)*

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.

In the following, the three steps described are repeated recursively on each sub-set until all the observations are isolated to the external node. As every separation is based on an "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)*. The path length or 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.

A decision tree, built by the IF algorithm (presenting one possible tree in an IF ensemble) in order to separate all the points in the dataset (2D dataset, depicted in Fig.1). 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 shows that to isolate outliers 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.

DEFINING THE ANOMALY SCORE

When executing the IF algorithm an entire ensemble – or "forest" – of binary decision trees is being built, each having a random set of partitions. This can be defined with the number of trees in the forest as a parameter. Since for each tree in an ensemble the path length to every observation is defined, an anomaly score, which classifies instances, can be calculated. For every point the path length is then 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 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 and 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 consist of entirely normal observations. [1] Ultimately, the algorithm builds the trees in a training stage and in 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]

PROGRAMMING LANGUAGES FOR ISOLATION FOREST

Using open-source software has become widespread in data science. From the range of programming languages that are currently being used, R and Python have emerged as dominating, offering a variety of tools and standard solutions for data analytics purposes, and, in particular, for the detection of outliers.

A standard implementation of the IF 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 utilised in many BI platforms for further analysis. Combining all these software resources with visualisation capabilities can considerably simplify analysis, make understanding easier and speed up the recognition of the items of interest in comparison with common methods.

CONCLUSION AND OUTLOOK

When dealing with high dimensional data, i.e., with many features, the use of different ML techniques can be very beneficial. In finance, the items or transactions associated with risks are rare and very often extremely difficult to detect. **Therefore, innovative technologies supporting data analytics become highly relevant and they can easily be used in day-to-day auditing processes.** The IF method directly and swiftly provides auditors with the necessary information on the "riskiest" items, capturing information from the entire population.

When applied in addition to common audit procedures, such methods can serve as an additional source of information providing suggestions what to investigate in more detail or pay especial attention to. Supported by IF, **Protiviti can deliver new levels of efficiency, further improve common routines and as a result provide new information and a deeper understanding of business processes.** This opens new ways for companies to implement changes for improvement. The information gained offers key insights for auditors and, ultimately, highly valuable input for the senior management to enable timely decision-making.

Sources

[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] *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*

[4] *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*

[5] *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*

[6] *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*

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

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

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

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

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

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

[13] *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*

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

CONTACT

PETER GRASEGGER
Managing Director
+49 173 653 8922
peter.grasegger@protiviti.de

DENIS LIPPOLT
Director
+49 172 698 30 48
denis.lippolt@protiviti.de

www.protiviti.de