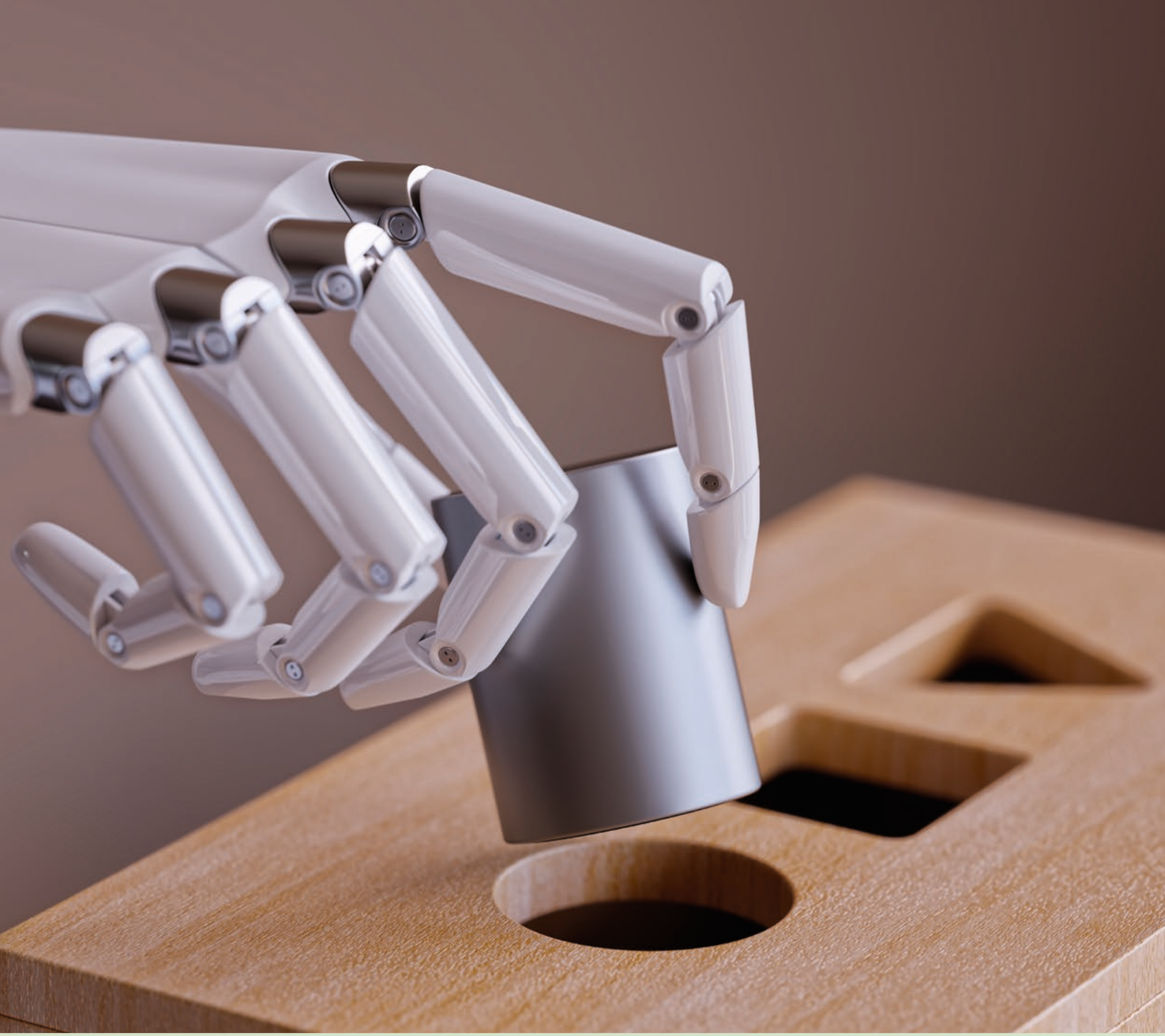
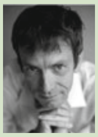


Using machine learning in a financial statement audit

Machine learning is a powerful technique that uses artificial intelligence to learn from data. It has uses ranging from virtual personal assistants to consumer preference prediction. However, these techniques are not commonly used in a financial statement audit. In this article we will review examples of unsupervised and supervised methods. We will also have a look at what keeps auditors from applying these techniques.



Lucas A. Hoogduin RA
is an audit director at the KPMG Global Solutions Group.
lhoogduin1@kpmg.com

Despite its use in other disciplines, machine learning is not widely used in audits

INTRODUCTION

Digitizing the audit is an ongoing effort. It started with the development of a digital filing system of audit work. This provides access to digital sources of information that are subject to the audit or that can assist in the audit work. In addition, audit procedures are now executed on entire data sets rather than on samples taken from it. The next logical step in this process is to replace conventional audit procedures by more efficient data-driven procedures.

There are many different data-driven techniques that are available to the auditor. A particular subset is machine learning, a technique that relies on the development of algorithms and statistical models that the computer can then use to carry out a specific task without further instructions. What's more, the learning aspect refers to the fact that these techniques improve over time as they get exposed to more data.

Take for example the email Inbox. The email application comes straight out of the box with some instructions (an algorithm) that are used to identify potential junk or spam messages. As the user starts scanning the Inbox, they may have encountered some messages that were missed, and a regular scan of the Junk mail folder may reveal some messages that were inadvertently tagged as spam. Most email applications now use the fact that the user moved messages from Inbox to Junk or vice versa to update the algorithm. This happens behind the scenes; the user doesn't have to do anything to set this in motion. And as time goes by, the application gets better and better at recognizing spam.

Machine learning in audit is not a widely used application, despite the fact that machine learning has been used in other disciplines. New applications include the personal virtual assistant, product recommendations, search engine result refining, online customer support, and online fraud detection. A good introduction on the various modeling techniques is given by [Hair14]. The authors not only explain how the techniques work; they also provide examples from literature where they were applied. For example, Multiple Regression Analysis is explained by [Hise83], who modeled the performance of retail outlets using 18 independent variables. [Desh82] used Factor Analysis to look at why some consumer product companies make more use of marketing research than others. And [Dant90] used Multiple Discriminant Analysis in their 1990 paper to determine whether there is a difference between the patients of private physicians and those of walk-in clinics. The methods and examples in [Hair14] assist in clarifying the structure of any data problem (are there similarities, is the variable of interest a continuous or a binary/categorical variable). They provide inspiration when solving audit-related problems.

In this article, we first introduce some recent innovations, and then the question why it is so hard to implement these innovations in the auditor's toolbox will be answered.

EXAMPLES OF MACHINE LEARNING IN THE AUDIT

Machine-learning techniques can be split into two classes, unsupervised learning and supervised learning. In unsupervised learning the data scientist 'lets the data talk'. The computer gets only general instructions, for example in the case of a clustering algorithm the number of desired clusters, which variables to use for clustering, and how the degree to which data points are similar (their distance) is defined. Supervised methods provide the algorithm with specific information on the research question. For example, when the user moves mail messages from or to the Junk mail folder, the algorithm receives information about what constitutes spam and what not, according to the user.

Example 1: Unsupervised learning – ratio analysis

Unsupervised learning can be used for the clustering of financial statement ratios. Assume that the auditor would like a dashboard with no more than six different ratios that collectively summarize the financial performance of the client. The problem is that there are close to 80 different ratios available. Is it possible to choose six of these, so that the information provided is still reasonably complete, and the information shown avoids overlap? In machine-learning terms, groups or clusters of ratios are needed, where the ratios within a cluster are closely correlated, and the correlation between the clusters is as low as possible. The factorial analysis method can help achieve this.

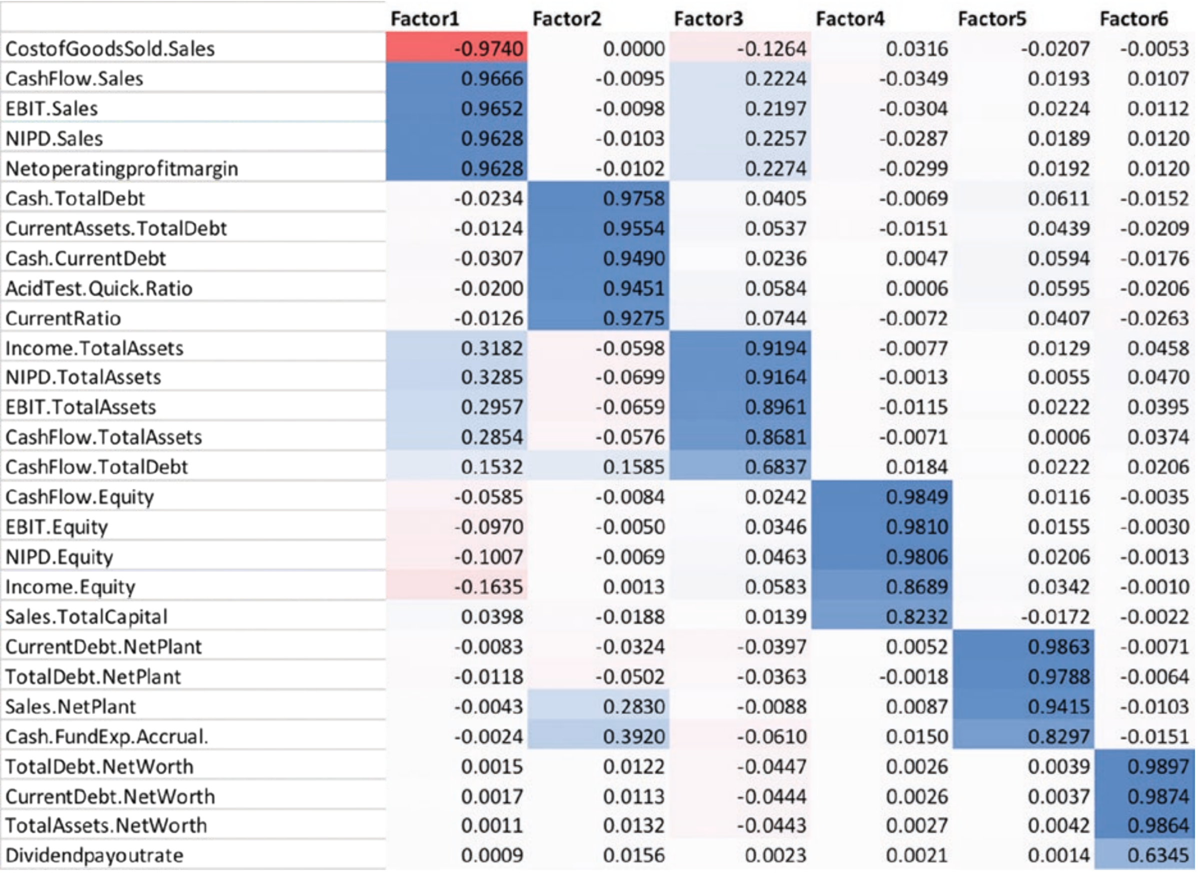
The starting point is to take financial statement information for a large number of companies in a specific industry. The hardest part of this exercise, and something that occurs very often when dealing with voluminous data, is to clean up the data. For example, most of the financial statement data may include numbers for Total Current Assets, and its constituent parts Cash and Cash Equivalents, Total Receivables, and Total Inventories. In some cases, the constituent parts may be missing however, and the face of the balance sheet only provided Total Current Assets, whereas in other cases one or more details are present, but the field of the total may be missing. One solution would be to discard all these cases, but the resulting database may then be dramatically reduced. Instead, going through the painful exercise of reconciling the information available may reward the researcher with a much larger database for consecutive use.

Once the base data are cleaned, the factorial analysis is carried out. The researcher makes a choice about the number of clusters desired; the computer does the hard work. If the user has specified six factors, the computer tries to find six vectors in a six-dimensional space in such a way that the total distance between each of the 80 original ratios and the six resulting vectors is as small as possible. The result of the analysis reveals the six requested factors representing clusters of financial statement ratios, as shown in Table 1. These clusters generally coincide with measurements that are of interest to financial statement users, like liquidity, solvency, profitability, asset utilization, return on invested capital, and financial market. For example, the first factor is related to profitability, the second to liquidity, the third to return on assets, etc. The next step is to choose for each of these factors which particular financial statement ratio correlated with it closest. For example, the Profitability factor (Factor 1) appears to correlate the strongest with the ratio of Cost of Goods Sold to Sales. The resulting six ratios are as dissimilar as possible, and the information contained in ratios that are not displayed resembles that of at least one of the displayed ratios. For example, once the ratio of Cost of Goods Sold to Sales is provided in the dashboard, the ratio of Cash Flow to Sales doesn't provide much additional information, since both ratios correlate strongly with Factor 1 derived.

Example 2: Unsupervised learning – journal entries

A similar approach has been taken to classify journal entries. Using the general ledger accounts and the

Table 1. Partial results of a factorial analysis on financial statement ratios.



	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
CostofGoodsSold.Sales	-0.9740	0.0000	-0.1264	0.0316	-0.0207	-0.0053
CashFlow.Sales	0.9666	-0.0095	0.2224	-0.0349	0.0193	0.0107
EBIT.Sales	0.9652	-0.0098	0.2197	-0.0304	0.0224	0.0112
NIPD.Sales	0.9628	-0.0103	0.2257	-0.0287	0.0189	0.0120
Netoperatingprofitmargin	0.9628	-0.0102	0.2274	-0.0299	0.0192	0.0120
Cash.TotalDebt	-0.0234	0.9758	0.0405	-0.0069	0.0611	-0.0152
CurrentAssets.TotalDebt	-0.0124	0.9554	0.0537	-0.0151	0.0439	-0.0209
Cash.CurrentDebt	-0.0307	0.9490	0.0236	0.0047	0.0594	-0.0176
AcidTest.Quick.Ratio	-0.0200	0.9451	0.0584	0.0006	0.0595	-0.0206
CurrentRatio	-0.0126	0.9275	0.0744	-0.0072	0.0407	-0.0263
Income.TotalAssets	0.3182	-0.0598	0.9194	-0.0077	0.0129	0.0458
NIPD.TotalAssets	0.3285	-0.0699	0.9164	-0.0013	0.0055	0.0470
EBIT.TotalAssets	0.2957	-0.0659	0.8961	-0.0115	0.0222	0.0395
CashFlow.TotalAssets	0.2854	-0.0576	0.8681	-0.0071	0.0006	0.0374
CashFlow.TotalDebt	0.1532	0.1585	0.6837	0.0184	0.0222	0.0206
CashFlow.Equity	-0.0585	-0.0084	0.0242	0.9849	0.0116	-0.0035
EBIT.Equity	-0.0970	-0.0050	0.0346	0.9810	0.0155	-0.0030
NIPD.Equity	-0.1007	-0.0069	0.0463	0.9806	0.0206	-0.0013
Income.Equity	-0.1635	0.0013	0.0583	0.8689	0.0342	-0.0010
Sales.TotalCapital	0.0398	-0.0188	0.0139	0.8232	-0.0172	-0.0022
CurrentDebt.NetPlant	-0.0083	-0.0324	-0.0397	0.0052	0.9863	-0.0071
TotalDebt.NetPlant	-0.0118	-0.0502	-0.0363	-0.0018	0.9788	-0.0064
Sales.NetPlant	-0.0043	0.2830	-0.0088	0.0087	0.9415	-0.0103
Cash.FundExp.Accrual.	-0.0024	0.3920	-0.0610	0.0150	0.8297	-0.0151
TotalDebt.NetWorth	0.0015	0.0122	-0.0447	0.0026	0.0039	0.9897
CurrentDebt.NetWorth	0.0017	0.0113	-0.0444	0.0026	0.0037	0.9874
TotalAssets.NetWorth	0.0011	0.0132	-0.0443	0.0027	0.0042	0.9864
Dividendpayoutrate	0.0009	0.0156	0.0023	0.0021	0.0014	0.6345

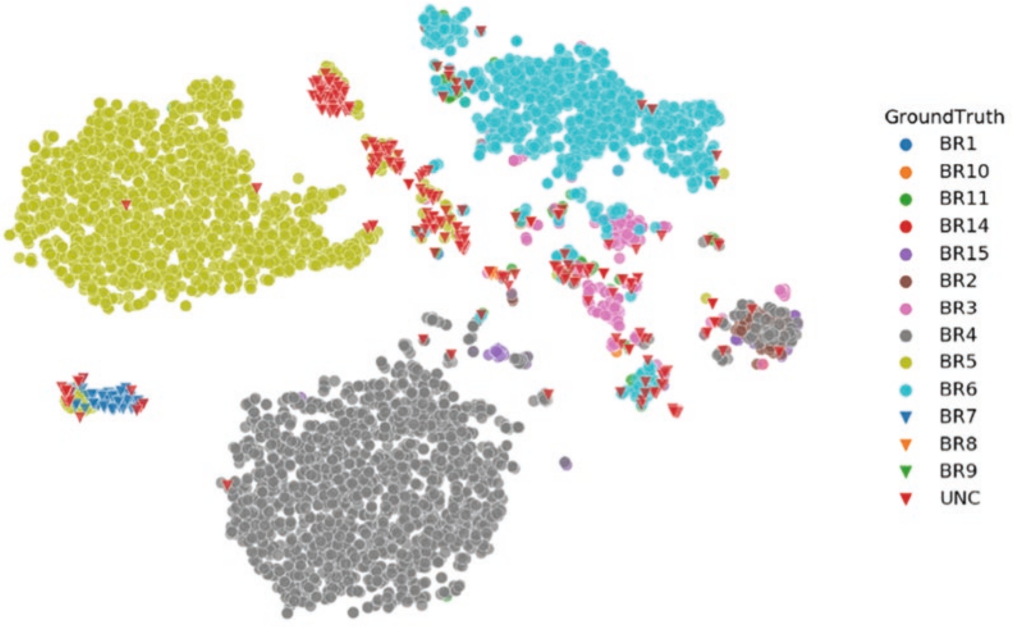


Figure 1. Results of an unsupervised Hierarchical Agglomerative Clustering of journal entries.

amounts of the debits and credits, Hierarchical Agglomerative Clustering provides the desired number of clusters of similar entries. A graphical representation is provided in Figure 1. The clusters found are displayed in a two-dimensional scatterplot, with a rotation that is optimized to see as many different clusters as possible. Transactions are color-coded according to auditor knowledge about the business process to which they belong.

This helps in different ways. It identifies the main transaction streams within a company, like purchases, sales, payments, receipts, payroll, fixed asset additions, etc. In the representation in Figure 1, these show up as distinctly separate groups or clusters of transactions. It visualizes the complexity of the bookkeeping process: were control accounts used or not; how often are amounts transferred to another account until they settle in a final destination. Investigating the clusters may reveal unusual entries, for example manual entries, unexpected users or ledgers. And finally, this technique reveals common structures in companies, assisting the auditor in finding relationships between processes and financial statement accounts, as a basis for other (supervised) machine-learning techniques.

Example 3: Supervised learning – regression analysis

The most widely known and least complicated of machine-learning techniques is regression analysis. It uses the presumed relationship between some variable of interest (the 'dependent'), typically a financial statement account that the auditor wants to examine, and a set of predictors, financial or non-financial data that the auditor believes has a plausible relationship with the dependent variable. The relationship is 'trained' on data available. This typically encompasses historical data for time-series applications, or similar data (historical data or from other entities) for a cross-sectional analysis.

Using regression analysis can be a very efficient technique to identify outliers, observations that are so unexpected that they merit further investigation. The absence of outliers allows to assess the probability that the dependent is free of material misstatement. The lower this probability, the less additional audit work is required.

Regression analysis applications may be widely deployed. Obvious examples are: analysis of depreciation charges against the historical cost of fixed assets, interest expense against the balance of long-term debt, or a margin analysis between revenue and cost of sales. Particularly the availability of external data sources like economic indicators, price indices, and industry trends could have a strong effect on the effectiveness of the regression model to identify outliers and obtain audit evidence as to whether the account is materially misstated. For example, Figure 2 shows the relationship between the Revenue of an airline company and passenger seat miles (the total number of miles passengers have travelled).

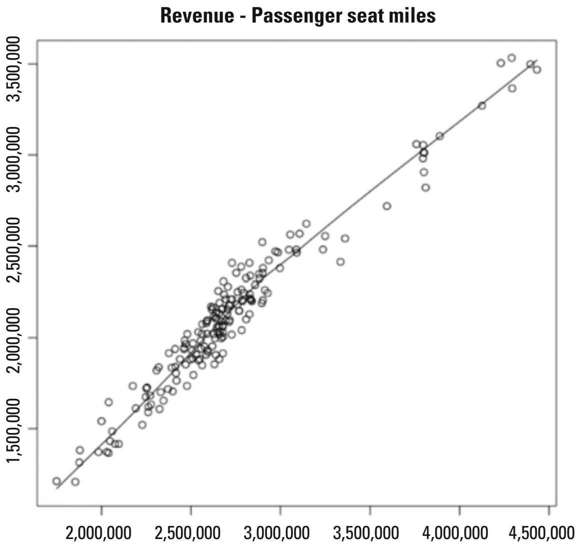


Figure 2. Scatter plot as part of a regression analysis.

Example 4: Supervised learning – loan ratings

As explained in Example 3, regression analysis is used on a dependent variable that is continuous, i.e. it can take any value in a certain range. But what if the dependent is a nominal variable (like gender) or ordinal (a sequence of categories in a particular order)? Similar to regression analysis there are machine-learning techniques such as logistic regression to construct models using a set of predictor variables, but rather than using the category representation (which could be completely arbitrary), the statistical technique is employed on the probability of class membership, which is a continuous variable between zero and one.

A useful application is the review of loan ratings. A loan rating is an ordinal variable ranging from AAA (the best rating) to F (the worst). The algorithm trains itself with information on the debtor, the loan, and the collateral to fit a predictive model that helps estimating the most probable class membership for each loan. The auditor then compares them to the ratings established by the client and investigates those loans that show the largest classification differences.

Example 5: Supervised learning – journal entry testing

It is very common, nowadays, that banks and credit card companies send their customers a message when they have identified a suspect transaction. This is a great example of a supervised learning technique on nominal data. The algorithm is trained on large volumes of transactional data, some of which are fraudulent, and others are not. There could be very many predictors, each of which have predictive power in the model. If a customer rarely buys shoes on the internet and suddenly does so, this raises a flag, but it does not for users who buy shoes online on a regular basis.

These techniques can also be used as part of an audit and have been used for a long time in forensic audits. The hardest part in this case is to obtain sufficient examples of fraudulent or otherwise incorrect transactions. This is the reason why these techniques are not very common yet as part of a normal financial statement audit.

BARRIERS AND ROADBLOCKS

Despite the potential of these new techniques, their application is not as common as could be. Even though the main barriers, like the availability of good data, statistical software and fast computers that can work with huge amounts of data, have now been overcome, there are a number of remaining issues that need resolution.

1. Data pooling. There has been much debate as to whether information obtained from one audit client could be used to assist in the audit of another. Does training the algorithms on sets of data from different audit clients therefore constitute a problem?
2. Audit evidence. Auditors clearly see the advantage of machine-learning tools where it comes to identifying outliers. But the question as to whether the absence of outliers can be regarded as substantive audit evidence has not been solved until recently ([Boer19]).
3. Data accuracy. Using supervised learning to assess whether or not an account is materially misstated using independent data shifts the burden of auditing the data of the dependent data to auditing the independent data.
4. Familiarity with statistical models. Auditors have not been sufficiently trained in the use of machine learning and statistics. Field auditors are therefore reluctant to use techniques they don't understand, and internal and external regulators cannot approve the use of techniques that cannot be explained to them.
5. Innovation cost. Development of machine-learning techniques for general use in the audit is costly, and the development cycle can stretch over multiple years. It is tempting to use innovation budgets on short-term wins however short-lived they may be.

THE WAY FORWARD

Each of the barriers and roadblocks mentioned before can be overcome. It will require some fundamental discussion, but the outcome will be an audit approach that allows the auditor to reduce spending time on areas that do not merit such a lot of manual work, to provide audit evidence on a much timelier basis, and to focus on high-risk areas. The use of machine-learning techniques is an essential first step that enables continuous monitoring and continuous assurance. There are solutions for the barriers and roadblocks identified earlier:

1. Data pooling. Rather than pooling data and run an algorithm on the data mass, the algorithm continues to train itself on each new data set. Its effectiveness therefore increases after every single use. This requires that the algorithm does not store any of the data it used for training, but only new coefficients for each of the predictors used.
2. Audit evidence. [Boer19] explains how the risk of material misstatement should be calculated if a predictive regression model reveals no outliers. Similar techniques can be developed for classification techniques and other predictive analytical procedures.
3. Data accuracy. There are many data sets publicly available that can be used as independent variables for a predictive model. Data accuracy is verified once by those maintaining the repository, allowing the auditor to rely on its accuracy.
4. Familiarity with statistical models. The author experiences an increase in auditor willingness to expand their personal skillset to include statistical modelling. Also, many universities now offer data science courses for auditors. As soon as field auditors experience the advantages that machine-learning techniques provide, their skepticism and fear will be reduced.
5. Innovation cost. Not all innovation needs a multi-million investment. It could be as easy as a single application of a new technique on a single engagement, then share the success and try to use the same approach on similar engagements. Procedures are now in place to enable the sharing of intellectual property between engagement teams, as well as safeguards to ensure that the application works as advertised. Once approved, these applications can then be made available within a country, or even globally.

CONCLUSION

Auditors have access to vast amounts of data. They can use these to more effectively gather evidence to support their opinion. Machine-learning techniques are very powerful tools that the auditor can employ to reach his audit objectives. Even though these techniques have so far not been widely used as part of a financial statement audit, in other areas they have proven their added value. The barriers and roadblocks that keep auditors from their use can be overcome if needed.

References

[Boer19] Boersma, M., Hoogduin, L., Sourabh, S., & Kandhai, D. (2019). *Audit Evidence from Substantive Analytical Procedures.* Proceedings of the 2019 American Accounting Association Annual Meeting.

[Dant90] Dant, R.P., Lumpkin, J.R., & Bush, R.P. (1990). Private Physicians or Walk-in Clinics: Do the Patients Differ? *Journal of Health Care Marketing, 10*(2), 25-35.

[Desh82] Deshpande, R. (1982). The Organizational Context of Market Research Use. *Journal of Marketing, 46*(Fall), 91-101.

[Hair14] Hair Jr., J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2014). *Multivariate Data Analysis, Seventh Edition.* Harlow.

[Hise83] Hise, R.T., Gable, M., Kelly, J.P., & McDonald, J.B. (1983). Factors Affecting the Performance of Individual Chain Store Units: An Empirical Analysis. *Journal of Retailing, 59*(2), 22-39.

[Hoog11] Hoogduin, L. & Touw, P. (2011). *Statistiek voor Audit en Controlling.* Amsterdam.

About the author

Lucas A. Hoogduin RA is an audit director at the KPMG Global Solutions Group in Berlin, Germany. He is primarily concerned with the development of machine-learning techniques for financial statement audits, including a methodology that is consistent with the applicable auditing standards. He also provides data science training for auditors, post-graduate students and other interested parties. He is one of the authors of the book *Statistiek voor Audit en Controlling* (Statistics for Audit and Control) ([Hoog11]).