Autoencoders for Strategic Decision Support

Sam Verboven[a], Jeroen Berrevoets[a], Chris Wuytens[b], Bart Baesens[c], Wouter Verbeke[a]

[a] *Vrije Universiteit Brussel, Belgium*
[b] *Antwerp Management School, Belgium*
[c] *Katholieke Universiteit Leuven, Belgium*

Abstract

In the majority of executive domains, a notion of normality is involved in most strategic decisions. However, few data-driven tools that support strategic decision-making are available. We introduce and extend the use of autoencoders to provide strategically relevant granular feedback. A first experiment indicates that experts are inconsistent in their decision making, highlighting the need for strategic decision support. Furthermore, using two large industry-provided human resources datasets, the proposed solution is evaluated in terms of ranking accuracy, synergy with human experts, and dimension-level feedback. This three-point scheme is validated using (a) synthetic data, (b) the perspective of data quality, (c) blind expert validation, and (d) transparent expert evaluation. Our study confirms several principal weaknesses of human decision-making and stresses the importance of synergy between a model and humans. Moreover, unsupervised learning and in particular the autoencoder are shown to be valuable tools for strategic decision-making.

Keywords: Unsupervised learning, Strategic Decision Support, Outlier Detection

1. Introduction

1.1. Problem Description

Data-driven approaches, such as machine learning and artificial intelligence methods are being adopted across industries to support, optimize and automate operational decisions. Examples include the adoption of machine learning in credit scoring to optimize decisions to extend credit [1], and in customer churn prediction to optimize customer relationship management [2].

Data-driven methods perform best at well-defined tasks that are repetitive and have tractable short-term effects. These strengths stand in stark contrast with what constitutes strategic decision making. Strategic decisions are often described as infrequent decisions, typically taken by management, that are not well defined and have high impact, long-term effects [3, 4]. For the remainder of this paper, we follow this definition. Examples of strategic decisions include; deciding on a remuneration policy, the composition of the board of directors, launching a new product, or investing in new machinery.

So in spite of technological progress, which has made operational task support such as churn prediction accessible to the average company, strategic decisions are still predominantly made without any learning-based grounding.

As such, the most important long-term decision-making in an organization is arguably the least supported by learning systems. Hence, strategic decision-making has to date been guided by expert knowledge even though humans are known to be prone to various biases [5], and managers are known to have preconceptions that lack objective grounding [6]. A lack of data-driven strategic decision support thus represents a large-impact problem across industries.

A data-driven solution of this problem would entail a system capable of learning from data that provides management with actionable information, as conceptually displayed in Figure 1.

1.2. Peer Influence in Strategic Decisions

Organisations are heavily influenced by others when making strategic decisions. On the one hand peer information is used to imitate, and on the other hand as a baseline to differentiate from through innovation. For both use cases, the key question amounts to 'what is normal' in a given peer group. This question underlies many strategic actions and their respective evaluation e.g., the definition of a correct remuneration policy is dependent on the market, and a

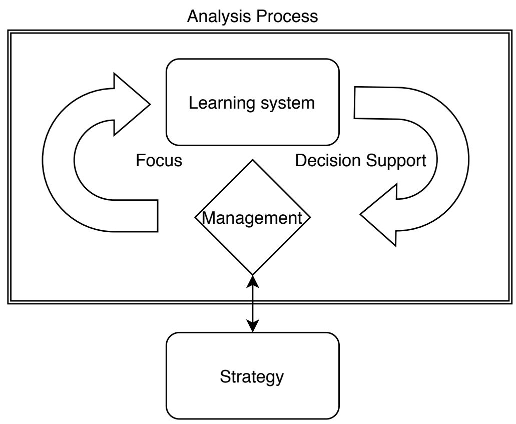

Figure 1: Conceptual diagram of the role of a learning system in decision support

policy leading to a revenue increase of one percent would be less lauded if all competitors grow by ten percent. Currently, lacking a learning-based grounding, the manager is restricted to peer information gathered through heuristics and readily available descriptive statistics. As such, a notion of normality that accurately represents complex multi-variate relationships is necessary for management to function and intelligently outline strategy [7].

The field concerned with this notion of normality is called outlier or anomaly detection, the identification of unexpected or abnormal behavior [8]. However, discrete classification of outliers does not suffice to enable provision of detailed and actionable information tuned to strategic decision making.

1.3. From outlier detection to strategic decision support

In outlier detection the decision to make usually applies at the observation-level, i.e. take a single action or not depending on the discrete classification outlier/no outlier of a single observation. Correct classification is thus central to the outlier detection task, which is reflected by the dominant evaluation strategies in this field.

In strategic decision support the (in case of the autoencoder same) unsupervised learning method is used to characterise the whole strategic peer environment, i.e. strategically relevant information is to be extracted. Not deviating from others is usually inconsequential in outlier detection decision tasks. For strategic decisions on the other hand, the implications of the aggregate result depend on the specific strategy under review. For example, not deviating in certain areas may require policy adjustment when one wants to innovate, but could also imply successful policy that brought the organization in line with industry leaders. Disentanglement of these deviations at the dimension-level, indicating whether you are below the norm or exceeding it and to what extent, also represents actionable quantitative information for managers. For example, in which dimensions (how) is the organization different, and how much do we need to change to get in line with industry leaders?

In other words, to provide actionable strategic information, managers require **granular feedback** on the outlyingness of each entity in the population. In line with above paragraph, this implies (1) whether, (2) how, and (3) to what extent an organization is different from relevant peers.

However, the output of the solution is required to be actionable, interpretable, justifiable [9], and ultimately accepted by the decision-makers. We group these qualitative characteristics under the umbrella term (4) **synergy**. This fourth requirement is often ignored but is key to ensuring the eventual adoption of the proposed decision support system. Past research has shown strong distrust and even dislike towards algorithmic decision support in the managerial domain [10] and, ultimately, the management is responsible for the executive decisions.

For strategic decision support, labeled data is not available and annotation is either largely incorrect, expensive, or both. As such, active learning or label noise strategies to enable supervised learning cannot be applied. Hence an unsupervised approach is adopted. Figure 3 visually motivates the need for unsupervised learning.

1.4. Solution

In this study, we introduce a framework for providing strategic data-driven decision support by utilizing an autoencoder (AE) neural network. The recon-

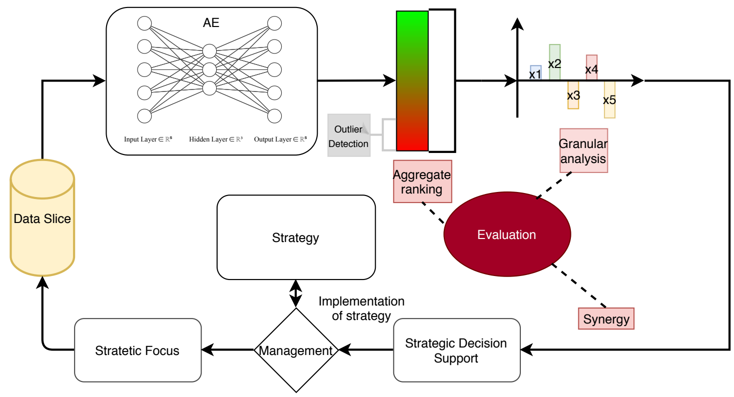

Figure 2: Comprehensive Diagram of the Extension of Data-driven Decision Support

struction error of the AE facilitates provision of granular feedback based on data by explicitly scoring in terms of how, to what extent and in what sense an observation deviates from a learned normal state, i.e., from what is expected, given the particular context when comparing to a set of relevant peers. Such a diagnosis is relevant in the strategic decision process which is dependent on peer information. Simply put, by means of comparison with a relevant benchmark, one may more accurately take position in the strategic landscape, and learn how to make more precise adjustments to either mimic or diverge from others. Other traditional outlier detection methods such as Isolation Forest and Local Outlier Factor techniques focus solely on classification performance, and do not yield this additional feedback necessary for the strategic support task. The structure of the solution is visualized in Figure 2.

To evaluate our approach, we introduce an extensive experimental setup specifically designed to gauge the capacity to fulfill every single requirement for effective strategic decision support defined in Section 1.2. As Figure 2 shows, this still involves an evaluation of not only the aggregate ranking but also elements of the granular analysis as well as synergy with management. The evaluation framework is discussed in full detail in Section 5.

For the purpose of the presented study, two proprietary datasets were obtained from a European HR services provider; these are sets of observations representing employees (D1) and employers (D2), including a selection of five and eleven dimensions of employees and employers, respectively. These datasets allow us to evaluate the use of the proposed approach to leverage unlabeled datasets for providing relevant input to the strategic decision-making process.

1.5. Contributions

In this article, we introduce the use of autoencoders for providing strategic decision support. We introduce and apply an assessment procedure to validate the proposed methodology using two HR datasets. We leverage data quality issues, expert opinion, expert validation and synthetic observations to demonstrate that the AE-based method does the following:

- Outperforms humans and other benchmark models;
- Offers granular dimension-level feedback, yielding extensive insights beyond the aggregate outlier scores; and
- Outputs information considered relevant and interpretable and is thus highly synergetic with human experts.

We present experimental results that validate (a) the business need for a data-driven diagnosis and (b) the adequacy of the proposed methodology in providing such decision support. The presented application of the proposed methodology is in the field of human resources management. However, the methodology is versatile and can be applied across strategic domains by selecting an appropriate dataset relevant for the envisioned analysis e.g. for financial strategy select relevant financial features, for a relative mapping of company culture one would need different features.

The remainder of this paper is structured as follows. In the next section, the related literature is reviewed. Subsequently, in Section 3 the proposed methodology is discussed. In Section 4, the need for data-driven strategic decision support is experimentally demonstrated. Next, Section 5 describes a series of experiments evaluating the effectiveness of the autoencoder as a solution. The

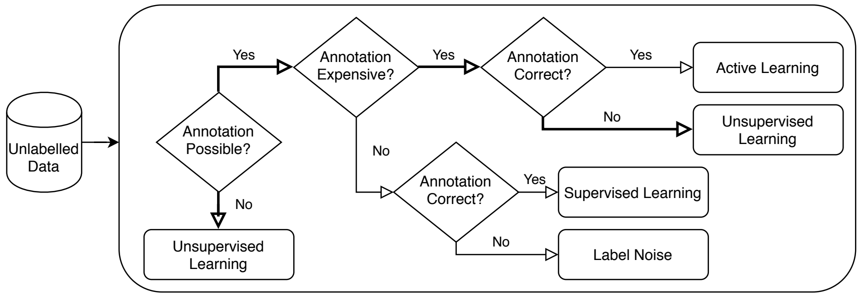

Figure 3: Label Imperfection

implications of these experiments are reported and discussed in Section 5.5. Finally, conclusions and future research opportunities are presented in Section 6.

2. Related Work

In this section, the existing decision support literature is revisited. Next, we review the literature relevant to characterize human decision making, and the influence of peer organizations thereon in a strategic context. Furthermore, the outlier detection literature is reviewed, as it methodologically and conceptually relates closely to our vision of strategic decision support. Finally, unsupervised learning for outlier detection is specifically examined to accommodate the absence of labels in the setting of this paper.

2.1. Analytics for Decision Support

Applications of supervised learning to model well-defined, repeatedly occurring events and/or corresponding decisions are rife in the literature, e.g. [1, 2, 11, 12]. In the strategic decision support literature specifically, applications include a multi-agent system for strategic bidding in electricity markets [13] and fire extinguishing method effectiveness prediction. However, different than the current work, these tasks are well defined and repeated. When the tasks are not well defined, no labels are available and unsupervised algorithms must be applied.

From a technical perspective, unsupervised algorithms have been applied in other decision support settings. Applications of unsupervised outlier detection aim to support decisions through flagging of outliers, e.g. in fraud detection [14].

From a conceptual perspective, the literature most in line with ours, i.e. concerning those strategic decisions that are one-off and ill-defined, is largely focused on non-parametric [15] and qualitative studies [16, 17]. [15] is especially similar to our work as it also focuses on quantitative feedback about peers in strategic decision making, although the information is limited to identification of best practice organization and the peer-groups they affect.

2.2. Peer Influences and Strategic Decision Making

Strategic decisions are not taken on a purely rational basis. Initial emotional or intuitive responses affect judgment [18] and when faced with complex information, humans fall back to simple heuristics [19, 20, 21]. As such, decision support should provide simple information that is relevant to the strategic decision making process. Several studies have established that organizations that gather more information about their environment achieve a higher performance through improved and more rational decision making [22, 23]. This environmental or peer information is used to make conscious choices to be similar to (i.e. mimicry) or different (i.e. innovation) than peers [24, 25, 7]. In the strategic management literature the optimal trade-off between differentiation and conformity is referred to as 'optimal distinctiveness' [25]. Empirical examples of mimicking decisions without rational basis include the appointment of CMOs [26]. Even when not outright mimicking, managers draw ideas from the practices of others [4, 3]. A tool able to comprehensively present similarity or quantifies best practice profiles thus provides information relevant to strategic decision making.

In summary, the impact on strategic decision support systems is twofold:

(a) Human decision making suffers from several flaws and biases and needs objective grounding through relevant information.

(b) Strategic decision making is strongly influenced by peer information

To assess the impact of these limitations on learning and human handling of complex strategic information and thus the need for a system yielding interpretable condensed information, it is paramount to study human expertise.

2.3. Outlier Detection

Outlier detection has been successfully applied in a plethora of fields, including fraud detection [27], computer vision [28], network intrusion detection [29], and medicine [30]. The interest in outlier detection stems from the assumption that identification of outliers and their characteristics translates into actionable information [31] towards these outlying observations. We extend this assumption and argue that common unsupervised methods can uncover information relevant to general strategic decision-making, beyond actions taken towards individual observations. Note that in the absence of labeled observations, unsupervised methods allow the ranking of observations based on the level of outlyingness indicated by outlier scores.

2.4. Unsupervised Outlier Detection

Approaches to unsupervised outlier detection are mostly based on statistical reasoning, distances, or densities [32]. The capacity of methods to accurately identify outliers varies across applications and depends on the dimensionality of the dataset, although some methods appear to be robust and generalize better than others [33].

Typically, a score is produced that can subsequently be used to rank and classify observations. The nature of such rankings produced by unsupervised outlier detection techniques is not yet well understood [33, 34]. This implies that every method inherently adopts its own implicit definition of what constitutes normality. Moreover, the optimal definition varies across application domains [33]. The autoencoder is a method that combines strong performance with a possibility of granular feedback. Deep autoencoding architectures have achieved outstanding results in traditional outlier detection [35, 12, 36].

To evaluate unsupervised models, expert input, e.g., a set of observations labeled by an expert, can be used; alternatively, if labels are available (though unused by the unsupervised learning method), then a holdout test set can be used as in the evaluation of supervised models [32]. An evaluation based on expert input hinges on two critical assumptions:

(i) the expert's labeling is correct, and

(ii) the expert's semantic understanding is relevant or desirable.

Note here that labeling observations becomes exceedingly difficult if the dimensionality of the observations, i.e., the number of available dimensions, increases [37]. This implies that the relevance of expert input is limited.

3. Methodology

In this section, we will discuss the autoencoder as well as two other state-of-the-art outlier detection methods, namely, the local outlier factor (LOF)[38] and isolation forest (Iforest)[39].

3.1. Autoencoders for decision support

The autoencoder facilitates an extension to decision support by offering granular and actionable information in addition to an overall outlier score and ranking. This additional information makes the output highly interpretable, as the overall causes of abnormality are readily quantified in terms of the original feature space [40]. LOF and Iforest do not offer such granular feedback but will be used in the experiments to benchmark the outlier ranking obtained from the autoencoder.

Autoencoders are symmetric artificial neural networks trained with the objective of reconstructing their inputs, i.e., observations. A basic autoencoder (Figure 4) maps an input vector $\mathbf{x} \in \mathbb{R}^n$, where $n \in \mathbb{N}^+$ is the dimension of $\mathbf{x}$, to an output vector of an equal dimension, i.e., the reconstructed observation $\mathbf{r} \in \mathbb{R}^n$. An autoencoder essentially consists of two main components: (1) an

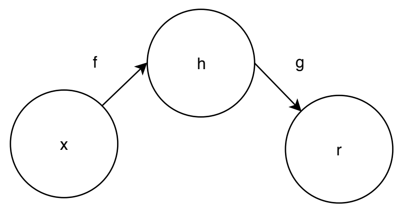

Figure 4: Autoencoder (adapted from [41])

encoder f that maps $\mathbf{x}$ to an internal representation $\mathbf{h} \in \mathbb{R}^m$, where $m \in \mathbb{N}^+$, and (2) a decoder g that maps $\mathbf{h}$ to $\mathbf{r}$.

Most applications of autoencoders aim to extract useful properties of the dataset through the internal representation $\mathbf{h}$. Such applications include pre-training [42], dimensionality reduction [43], and vectorizing word representations [44]. However, in outlier detection and by extension in decision support, we are primarily interested in the output $\mathbf{r}$. More specifically, here we are interested in the similarity between $\mathbf{r}$ and $\mathbf{x}$ expressed by a loss function $\mathcal{L}(\mathbf{x}, g(f(\mathbf{x})))$. Generally, a loss function that penalizes the distance from $\mathbf{r}$ to $\mathbf{x}$ is selected, thereby defining the reconstruction error. By restricting the capacity of $\mathbf{h}$, useful properties of the data may be learned [41]. In an undercomplete autoencoder, the internal representation acts as a bottleneck since $\mathbf{h}$ is of a lower dimension than $\mathbf{x}$, i.e., $n > m$. Through this bottleneck, an incomplete reconstruction is forced since model capacity no longer suffices for an exact reconstruction. When training the autoencoder with the objective of minimizing the reconstruction loss, we implicitly favor the reconstruction of inputs that *are closest to the data*. Hence, inputs that are the farthest from the learned reconstruction exhibit the largest errors. If more hidden layers are used in the autoencoder architecture, the capacity of the network increases, enabling it to construct a more complex hidden encoding of the data.

An undercomplete autoencoder combines multiple characteristics of an attractive solution to our problem:

- It can handle a mix of continuous and discrete data [45].
- The reconstruction errors can be interpreted as deviations for each indi-

vidual dimension from the *normal* or *expected* state, and

- The errors offer information about both the size and the direction of the deviation.

3.2. Outlier ranking methods

3.2.1. Local Outlier Factor.

The local outlier factor method (LOF) [38] is a state-of-the-art unsupervised outlier detection algorithm [46]. LOF is a density-based scheme in which an outlier score $LOF_k(p)$ is computed for each observation.

The k nearest neighbors $N_k(p)$ are determined for each observation p, where $k \in \mathbb{N}^+$. Afterwards, the local reachability density $lrd_k(p)$ for one observation p is computed:

$$lrd_k(p) = \left(\frac{\sum_{o \in N_k(p)} d_k(p, o)}{|N_k(p)|} \right)^{-1}, \tag{1}$$

where d_k is the reachability distance. In (1), the local reachability density is thus inversely proportional to the average reachability distance from p to its k neighbors. The reachability distance is almost always computed as the Euclidean distance [46]. Intuitively, a larger distance between observations implies a lower density.

Given $lrd_k(p)$, $LOF_k(p)$ can be computed:

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|}. \tag{2}$$

$LOF_k(p)$ is the average ratio of the lrd of p to the lrds of its k neighbors.

The number of nearest neighbors being considered (k), and the distance measure for the reachability distance, i.e., Euclidean, are hyperparameters of the model (cf. Table .10 in the Appendix). Observations with a density that is substantially lower than those of their neighbors are considered outliers or anomalies.

As the average ratio between the densities of the observation and the neighborhood increases, so does $LOF_k(p)$. Hence, $LOF_k(p)$ being equal to one implies that lrd_k of observation p is on average equal to lrd_k of its neighbors. A higher $LOF_k(p)$ indicates that p lies, on average, in a lower-density area than those of its neighbors and can thus be considered to be more outlying. LOF outputs a score that can subsequently be used to rank observations from high to low level of outlyingness.

3.2.2. Isolation Forest (Iforest).

Isolation forest (Iforest) [39] is a powerful outlier detection algorithm that extends decision tree and ensemble methods, such as random forests. Isolation implies "*the separating of an instance from the rest of the instances*" [39]. The key assumption behind Iforest is that anomalies are fewer and different and are thus more susceptible to *isolation* when the input space is randomly segmented.

Compared to inlying observations, an outlying observation will on average require fewer splits of a decision tree that randomly partitions the input space, for the observation to be isolated from other observations. If a forest of such random trees collectively produces shorter *path lengths* for some observations to be isolated, the latter are likely outliers.

The number of edges an observation x traverses in an isolation tree from the root node to termination at an external node is denoted by $h(x)$. Moreover, a normalization factor $c(n)$ enables comparisons across different subsampling sizes. The Iforest method then calculates a score $s(x, n)$,

$$s(x, n) = 2^{-\frac{\mathbb{E}[h(x)]}{c(n)}}, \tag{3}$$

where $\mathbb{E}[h(x)]$ is the expectation of $h(x)$ from a collection of trees. The resulting anomaly score $s(x, n)$, for which $0 < s(x, n) \leq 1$, can be utilized as follows:

- The closer $s(x, n)$ is to 1 for observation p, the more likely p is to be anomalous.
- Conversely, if $s(x, n)$ is significantly lower than 0.5, the observation is

almost certainly non-anomalous.

An existing study of explainability of Iforest identifies the dimensions that contribute the most to the final score [47]. In contrast to an autoencoder, an isolation forest does not offer insight as to the size and sign of the deviation from normality. A more detailed explanation of Iforest is available in [39].

4. Experimental Validation of the Problem

In the literature section, it was established that managing optimal distinctiveness is key in strategic management, and that gathering of peer information is associated with enhanced performance. As such, if managers can swiftly and consistently process large amounts of complex peer information, there is no need for a support system. In other words, the assumption that this assessment of relative normality of complex strategic data is difficult ultimately determines the added value of this study, the type of algorithm we should use, and the evaluation strategy to be applied.

In this section, we report the setup and results of an experiment designed to test this assumption.

4.1. Set-up

Ten study subjects were selected by an HR services company as experts based on their expertise. All subjects were from the consulting division, and had either consulting, business intelligence, or director roles in the organization.

The data used in this study belongs to an HR services provider, and includes data on both employees and employers. Two datasets were composed: the first dataset (D1) consisted of $128,820$ observations of employees and included five dimensions (see Table .7); the second dataset (D2) consisted of $1,864$ observations of employers and included eleven dimensions (see Table .7).

For both datasets the subjects were asked to label a subset of observations. First, the three methods were run on both D1 and D2. Second, using these results, subsets were selected to (i) span the full range of normality, including observations with high, medium and low outlier rankings across methods, and (ii) ensure discrimination between methods by including a mix of observations the three methods disagreed on, i.e., ranked in very different deciles. Third, the ten subjects were given as much time as needed to review and label the observations as normal ($Y = 0$), outlier ($Y = 1$), or undecided if a subject could not decide on a label ($Y = na$). Furthermore, two additional indicators of aptitude of subjects were collected for both D1 and D2. Each subject was asked to score the following:

- The relevance of the subject's professional experience to the labeling task on a scale of one to ten, with a score of ten meaning very relevant; and
- The difficulty of the labeling task on a scale of one to ten, with a score of ten meaning very difficult.

To assess whether humans indeed rely on certain heuristics when faced with complex, i.e., high-dimensional, strategically relevant data, the subjects were asked to identify the main dimensions that contributed to deciding on the label for an observation.

A key requirement for logical decision-making, either by humans or systems, is consistency [18, 5]. To assess the consistency of subjects, in both series of observations that were to be labeled, a number of duplicates, i.e., copies of observations, were included. The consistency of a subject is then evaluated as the proportion of the copied observations that were assigned the same label, or, for a number of subjects $s = 1, 2, ..., N$ and $\mathcal{D}$ duplicates,

$$Consistency = \sum_{i=1}^{\mathcal{D}} \frac{c_{s,i}}{\mathcal{D}}, \tag{4}$$

$$\text{where } c_{s,i} = \begin{cases} 1 & \text{if } s \text{ assigned } i \text{ the same label.} \\ 0 & \text{if } s \text{ assigned } i \text{ a different label.} \end{cases} \tag{5}$$

This measure of consistency is interpreted as a proxy for proficiency at the task at hand. A higher level of inconsistency in making decisions points to irrational

Table 1: Expert Results

		Average	StDev	Min	Max	Correlation: Consistency	Correlation: Difficulty	Correlation: Job Relevance
Employee	Consistency	71.11%	1.26	4.00	8.00	1.00	-0.64	0.67
n = 49	Difficulty	6.90	3.11	2.00	10.00	-0.64	1.00	-0.85
$\mathcal{D} = 9$	Job relevance	5.20	3.19	1.00	10.00	0.67	-0.85	1.00
Employer	Consistency	60.00%	1.05	1.00	5.00	1.00	0.04	0.38
n = 40	Difficulty	6.00	2.45	2.00	9.00	0.04	1.00	-0.49
$\mathcal{D} = 5$	Job relevance	5.60	2.50	1.00	9.00	0.38	-0.49	1.00

and non-systematic judgment.

4.2. Results

For both datasets, consistency scores, indicators of aptitude, and the correlation matrix between consistency and aptitude indicators are listed in Table 1. Four key observations can be made with respect to the results.

1. First, the experts are often in disagreement with *each other*, as shown in Figure 5. Note that the experts do not unanimously agree for even a single observation on the appropriate label. Spearman rank correlation results for the judgments of individual experts are shown in Table .8 and Table .9 for D1 and D2, respectively.

2. Second, the experts do not agree with *themselves*. With average consistency rates of the duplicate labels of 71.11% and 60.00% for datasets D1 and D2, respectively, human experts are remarkably inconsistent. They appear to barely surpass random performance, characterized by a consistency rate of 50%. In agreement with the literature, the consistency of experts is observed to decline as complexity increases. Inconsistencies are not related to a specific subset of observations that are difficult to assess, as all fourteen duplicate observations were inconsistently labeled at least once.

3. Third, experts focus on a relatively small number of dimensions, indicating heuristic decision making. This can be inferred from Figure 6. Moreover, experts take into account different (combinations of) dimensions in deciding on the appropriate labels. The tenth dimension is the only characteris-

tic reported as having been used at least once by every expert. Conversely, only four experts indicated using the eighth dimension.

4. Fourth, for the employee dataset (D1), consistency is positively correlated with self-reported professional relevance, and negatively with perceived difficulty. For the employer set (D2), which includes more dimensions, the experts' self-assessment of perceived difficulty did not correlate significantly with consistency.

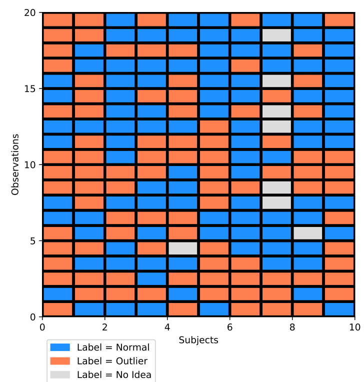

Figure 5: Expert labels for the first twenty observations of the employer dataset (D2) described in Section 4. The colors represent the labels, each column contains the labels assigned by a given expert, and each row visualizes the labels assigned to a given observation.

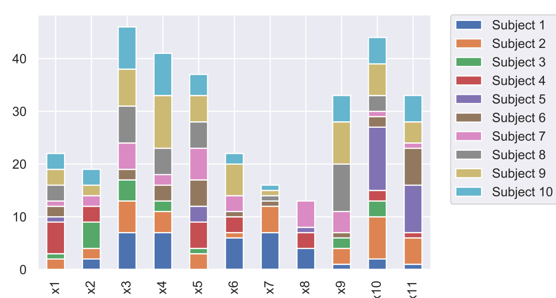

Figure 6: Distribution of the use of eleven dimensions (x1–x11) in the employer dataset (D2) by the ten experts

These results indicate that human experts rely on heuristics and that in a strategic setting, they are not able to process complex, strategically relevant peer information. These results therefore highlight the need to support strategic decision making.

5. Experimental Validation of the Solution

Section 4 presented experimental evidence of the limited ability of human experts to consistently analyze complex data within their field of expertise, which is the problem we aim to address in this study. In Section 5, use of the autoencoder as a strategic decision support system is validated as a potential solution to this problem.

We identify three dimensions in validating the proposed approach:

(i) **Outlier detection performance**: We assess the correctness of the obtained outlier score ranking, with outlier scores being the aggregated amount of deviation across all dimensions.

(ii) **Dimension-level feedback**: To ensure the added value of providing granular feedback, i.e., feedback regarding size and sign of a deviation provided by the system at the level of individual dimensions, we assess the reliability and accuracy of the provided feedback.

(iii) **Synergy between the model and human assessment**: A seamless integration within the management decision-making process is vital for a successful adoption of the proposed system; here, we ensure that users correctly understand the output of the system, can use the output for practical decision-making, and do not find their personal beliefs to be in persistent conflict with the output.

To validate the autoencoder-based support system across these three dimensions, we perform four experiments involving blind expert validation (Section 5.1), transparent expert validation (Section 5.2), an observed case of corrupted data (Section 5.3), and synthetic observations (Section 5.4).

Table 2 summarizes the contributions of these four experiments to the validation of the system across the three dimensions identified above. The following sections will provide full details on the setup of these experiments and discuss the results. Hyperparameters and correlations are consistent with previous studies and reported in the Appendix in Table .10 and Table .11.

Table 2: Validation of Methodology

	(i) Outlier detection performance	(ii) Synergy	(iii) Dimension-level feedback
5.1. Blind expert validation	(x)	x	(x)
5.2. Transparent expert validation		x	x
5.3. Data quality	x		x
5.4. Synthetic observations	x	x	x

[a] (x) indicates a moderate contribution.
[b] x indicates a sizable contribution.

5.1. Blind Expert Validation

This first experiment aims at evaluating the accuracy of outlier scores produced by the autoencoder. Since the observations in the data are unlabeled, there is no objective ground truth that can be used for assessing the accuracy of the ranking. As argued in Section 4, the alternative of using labels assigned by an individual human expert cannot be assumed to yield a trustworthy assessment. As an improved alternative to using the labels of a single expert for validation, we may instead compare the assessment of the autoencoder system with that of a group of experts, which can be considered to be an ensemble classification system. An ensemble classifier benefits from accurate and diverse members [48, 49]. Hence, we use ensemble theory to construct a weighted aggregate classifier from individual expert opinions. Every subject is considered to be a weak classifier, and it is hypothesized that their joint performance may be better, leveraging the wisdom of crowds [50].

5.1.1. Set-up.

To combine individual estimates, two variants of majority voting are implemented:

Unweighted Majority Voting. Denote the decision of the s^{th} subject (i.e., expert) by $d_{s,j} \in \{0,1\}$ for $s = 1, ..., S$ and $j = 1, ..., C$, where S is the number of subjects, and C is the number of classes, such that $d_{s,j} = 1$ for the class the subject selected, and zero otherwise. For an observation, J_{uv} is the voted label, and the summation tabulates the number of votes for class j:

$$J_{uv} = argmax_{j \in \{0,1,2\}} \sum_{s=1}^{S} d_{s,j}. \tag{6}$$

Weighted Majority Voting. Here, w acts as a weighting factor for the vote. The weighted majority vote is J_{wv}, and the summation in this case tabulates the weighted vote for class j. Hence, the votes of individuals who perceive their expertise to be more relevant to the task will have larger weights in the vote.

$$J_{wv} = argmax_{j \in \{0,1,2\}} \sum_{s=1}^{S} w_s d_{s,j}, \tag{7}$$

where $w = 1, ..., 10$, and w_s is either the self-perceived job relevance of subject s or the inverse of the self-perceived difficulty of the task of subject s.

The labels of individual experts, obtained in the experiment discussed in Section 4 and combined using the two majority voting schemes described above, are used to assess the outlier scores of the autoencoder, LOF, and iForest by subsequently labeling five, ten, and fifteen percent of observations with the highest outlier scores as outliers. Afterwards, we measure the accuracy of the weighted and unweighted majority expert ensemble against this labeling. Under the assumptions that (i) the models are valid tools for outlier detection in this setting, and (ii) humans make different mistakes that can average out when combined, convergence between the labels of outlier detection methods and those of the expert ensemble is to be expected.

5.1.2. Results.

Table 3 shows the results for the unweighted and weighted expert ensembles, both when weighting with the self-reported job relevance and difficulty scores. A higher accuracy means there is a stronger match between the expert ensemble and the outlier detection method. This table demonstrates that, generally, the autoencoder attains the highest accuracy, at least in comparison with the weighted ensembles. This indicates that, among the three models, the autoencoder best matches with the weighted aggregate judgment of human experts. The absolute and percentage accuracy increases achieved by weighing the expert labels are also the highest for the autoencoder. The experts were relatively correct in their self-assessments, and the models are accurate, as evidenced by

Table 3: Majority Voting Results

	AE			Iforest			LOF		
	5 %	10%	15%	5%	10%	15%	5%	10%	15%
Unweighted	0.54	0.56	0.62	0.56	0.54	0.59	**0.64**	0.51	0.49
JobRel_Weight	0.69	0.72	**0.77**	0.69	0.67	0.69	0.72	0.64	0.62
Difference	**0.15**	**0.15**	**0.15**	0.13	0.13	0.10	0.08	0.13	0.13
% Increase	**28.57%**	27.27%	25.00%	22.73%	23.81%	17.39%	12.00%	25.00%	26.32%
Difficulty_Weight	0.77	**0.79**	**0.79**	0.72	0.69	0.72	**0.79**	0.67	0.64
Difference	**0.23**	**0.23**	0.18	0.15	0.15	0.13	0.15	0.15	0.15
% Increase	**42.86%**	40.91%	29.17%	27.27%	28.57%	21.74%	24.00%	30.00%	31.58%

the accuracy increase after weighting.

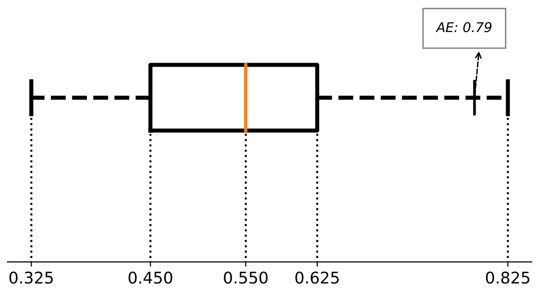

Figure 7: Boxplot of individual accuracy distribution between the experts and all models and cutoffs (n = 90), with an indicator of the ensemble AE result from Table 3.

The boxplot in Figure 7 represents the distribution of accuracy values of ten individual experts across the three methods (AE, LOF, Iforest) and the three cutoff values for turning outlier scores into labels (5%, 10%, and 15%), thus yielding 90 data points (3x3x10). The median accuracy is barely higher than the performance of a random model. Out of these ninety combinations of cutoff values, experts and models, only one has a higher accuracy than that consistently reached by the ensemble-weighted AE. In this respect, it is remarkable that the expert-weighted ensemble stabilizes at an accuracy of just under 80%. We can conclude that there is high variance in accuracy between individual experts, but the AE can consistently represent majority expert opinion.

5.2. Transparent Expert Validation

To evaluate synergy, we assess whether experts understand and agree with the output provided by the autoencoder system.

Table 4: Outlier Detection Performance – Detection Results

	A Data Quality			B Synthetic Observations			C Average Performance		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
AE	**61.80%**	82.05%	85.39%	50.00%	70.00%	80.00%	**55.90%**	**76.01%**	**82.70%**
Iforest	60.67%	**95.51%**	**100.00%**	10.00%	20.00%	40.00%	35.34%	57.76%	70.00%
LOF	0.00%	4.49%	11.24%	**60.00%**	**80.00%**	**100.00%**	30.00%	42.25%	55.62%

Table 5: Dimension-level Feedback – Data Quality Set

Dimension	No. Obs.	Dimension rank	Direction % correct
x1	3		100.00%
x2	5		100.00%
x3	8		100.00%
x4	22		100.00%
x5	69		100.00%
All	89	85.39%	100.00%

5.2.1. Set-up.

During a two-hour panel session, the group of experts was presented with the output of the autoencoder for the observations in the two datasets that the experts labeled in the previous experiment, as reported in Section 4. The aim of the session was to gauge whether and how each expert could extract insights useful for decision-making from the output of the system. Specifically, the experts discussed the outlier score ranking as well as the granular feedback, i.e., deviations at the dimension level, provided by the autoencoder. To facilitate analysis, observations were presented using interactive visualizations that were implemented in a business intelligence software.

5.2.2. Results.

The panel was able to interpret the results provided by the system. The panel did not object to a single assessment of the autoencoder (either at the aggregate outlier score level or at the granular dimension level). The interpretability and justifiability of the system, as confirmed by the experts, indicates synergy between experts and the model. While the autoencoder output was being studied, a data quality issue was noticed in the employee dataset (D1), highlighting synergy and yielding concrete actionable benefits of the model. Moreover, the experts proposed new applications of the system beyond the employees-and-employer dataset. Alternative employee- or employer-level datasets could be analyzed to provide specific insights on themes such as work fatigue, hiring, onboarding, etc. The versatility of the autoencoder-based approach, as recognized by the experts, indicates that the system is effective and useful in decision-making. Such versatility is a valuable property, allowing the proposed approach to be adopted as a comprehensive decision-support instrument for performing ad hoc analysis in support of any decision-making process, by merely compiling a dataset including a set of relevant dimensions.

5.3. Data Quality

Data quality issues are closely related to outlyingness. As reported in the previous section, a large-impact data quality issue was discovered in the employee dataset during the transparent expert validation of the autoencoder output. The discovered data quality issue (Section 5.2) was fixed by in-house experts. By comparing the pre- and post-fix versions of D1, the affected points could be reliably identified. Moreover, one could discern the involved dimensions as well as the direction of the effect. For the affected points, the logical relations the variables abide by were violated. Consequently, the affected observations are sufficiently distinct to have a close affinity with the concept of an outlier.

5.3.1. Setup.

Using this data quality event to our advantage, two experiments were devised. First, labeling the affected observations as one, and the others as zero allowed an evaluation of the detection performance of the algorithms. Second, utilizing knowledge about the affected dimensions and the direction of the effect permitted testing of the granular feedback capabilities of AE.

To validate the dimension-level feedback of AE, we define two measures of accuracy: dimension rank accuracy, and direction accuracy:

- **Dimension rank accuracy** equals 1 for an observation if the AE error is the highest in the actual affected dimension(s) and equals 0 otherwise.

Table 6: Dimension-level Feedback – Synthetic Dataset

Obs.	Perturbations	Dimension rank acc.	Direction
1	1	1	correct
2	1	0	correct
3	1	1	correct
4	1	0	correct
5	1	1	correct
6	2	1	correct
7	2	1	correct
8	2	0	correct
9	3	1	correct
10	3	1	correct
Average		70.00%	100.00%

- **Direction accuracy** of an observation is equal to 1 if for all affected dimensions, the direction is correctly represented by the sign of the difference between the observed value and the output value and is 0 otherwise.

Since the data quality issue was identified, we were able to assign ground truth labels to the affected dimensions and the direction. Using these labels, we could calculate the dimension rank and direction accuracy.

5.3.2. Results.

Table 4A displays the data quality detection performance for the three algorithms. Iforest and AE perform well, as both have a high proportion of affected observations in the top percentiles of their respective rankings. In contrast, LOF performs poorly.

Considering the granular feedback, as shown in Table 5, AE consistently recognizes the direction of the deviation (100%) and ranks the perturbed dimension(s) the highest for 85.39% of the observations. Interestingly, this performance does not change significantly if observations that AE did not correctly classify as affected (84.21%) are omitted. This is particularly relevant to the extension from the top x percentile analysis to full-population decision support; even without high outlier scores, the granular feedback is accurate and valuable.

5.4. Synthetic observations

Due to the instability of outlier detection algorithms reported in the literature across domains [33], examining performance on the data quality dataset is insufficient for evaluating performance in general. Therefore, we adopt the approach proposed and applied in [51, 52] and inject synthetic outliers into the dataset.

5.4.1. Setup.

Observations in the employer dataset that were evaluated as *non-outlying* by three outlier detection methods were selected. Next, perturbations to these observations were devised by a panel of four experts to achieve impossibility, illogicality, or implausibility beyond a reasonable doubt with a minimum amount of perturbation. As such, variance between the methods' rankings is ensured, making it possible to discern the best-performing method.

The synthetic observations were varied across the data plane with five unidimensional, three two-dimensional, and two three-dimensional perturbations. After their inception, these perturbed observations were added to the full dataset, and outlier detection models were retrained. To evaluate the dimension-level feedback, we use the same measures as reported in Section 5.3. In this experiment, we can observe the ground truth, label accordingly, and evaluate the accuracy.

5.4.2. Results.

Table 4B shows that AE performs well. Additionally, and in contrast with Table 4A, LOF reports great results, with perfect discrimination at 15% cutoff. Iforest, however, performs poorly. The results for the dimension rank accuracy and the directional feedback are displayed in Table 6. For 70.00% of the perturbed observations, AE correctly ranks all perturbed dimension(s). Furthermore, the autoencoder obtains the correct direction of the perturbation in all dimensions for every observation.

5.5. Discussion

We validated an autoencoder-based approach to support strategic decisions on three levels (cfr. Table 2:

- **Outlier detection performance.** Validation results using experts, data quality and synthetic observations reported in Tables 4A, 4B and 3 indicate a strong performance of the autoencoder in detecting outliers in various experiments. Tables 4C and 3 illustrate performance for various settings and show that the autoencoder significantly outperforms two other state-of-the-art algorithms assessed in the experiments. The instability of results due to data quality and synthetic observations' settings confirms earlier results reported in [33], who reported instability of methods when comparing performance for different outlier detection settings. We observe this phenomenon for two datasets in the same setting. A desirable solution should therefore generalize well across semantic definitions of outliers without requiring significant hyperparameter tuning. Moreover, excessive tuning to a specific semantic definition may prevent the model from identifying interesting semantically varying patterns. Tuning on already discovered data quality issues seems especially inappropriate. Based on the results of the conducted experiments, we conclude that the autoencoder generalizes well across settings.

- **Synergy.** The autoencoder is shown to be highly synergistic with human decision-making processes due to (i) strongly correlating with joint weighted human decision-making, (ii) being unanimously accepted during a two-hour panel discussion that explored the insights provided by the approach, and (iii) matching the semantic definition of outliers on synthetic observations. The experts in the panel were able to interpret and explain the results, placing them in a richer context than that the model had direct access to through the input data.

- **Granular feedback at the dimension level.** The autoencoder is a powerful tool for discerning the rank and deviation direction of the main dimensions contributing to abnormality. Traditional unsupervised methods do not offer such granular feedback. Moreover, the autoencoder achieves perfect accuracy in assessing the direction of the deviation in our experiments (Tables 5 and 6). An interesting implication is that the dimension-level feedback seems remarkably stable even for low-ranked observations. This supports the idea of adopting unsupervised outlier detection methods for obtaining actionable information beyond a small set of top-ranked observations with high aggregate outlier scores.

6. Conclusions

In this paper, we propose an unsupervised learning approach to support strategic decision-making by adopting the autoencoder, a powerful artificial neural network-based method that can provide detailed insights in regard to large and small deviations from what is expected. Such deviations relative to relevant peers support decision-making, providing feedback on the "as-is" situation and the direction towards an improved "to-be" situation.

To validate the proposed approach, a unique dataset was obtained from a European HR services provider, including information on a large set of employees and employers. Using a panel of ten experts, we observe that, as a first contribution to this domain, human experts are inconsistent and non-comparable in their judgments. This finding strongly motivates the need for support in the first stage of the business decision-making process, i.e., the analysis of business problems.

To this end, we investigate the detection performance, synergy, and granular feedback of our autoencoder-based solution. We acknowledge that in this setting, there is no single guaranteed evaluation method for assessing the performance and use of the proposed method. In the absence of a generally accepted evaluation procedure, we devise and perform four experiments for validation using (i) transparent expert validation, (ii) blind expert validation, (iii) data quality classification, and (iv) generation of synthetic observations.

The results of these experiments indicate that the proposed autoencoder method meets business users' requirements in terms of outlier detection performance, synergy, and dimension-level feedback. Moreover, the method is versatile and can be adopted to support decision-making across various management areas by compiling appropriate datasets.

Unsupervised learning for decision support is an underexplored research area. Decision support systems that interconnect humans and machines are urgently needed to unlock the potential of big data for optimizing strategic decision-making. Several challenges remain:

(i) A framework for objective and trustworthy validation of analytical models in an unsupervised setting is missing;

(ii) Models need to be more robust, reducing the risk of failure modes;

(iii) Developing a system to provide strategic decision support is a challenge to data scientists since the development of a system that aligns with high-level strategy requires a higher level of business understanding than development of traditional decision support systems, e.g., a customer churn prediction model, and

(iv) A lack of familiarity with unsupervised learning methods may hamper swift industry adoption.

Along with challenges, unsupervised decision support offers exciting possibilities for future research. Possible areas for further development include the following:

(i) The incorporation of a temporal dimension to capture and describe the time-varying nature of the data distribution;

(ii) The demonstration and prediction of causal effects of actions with regard to their abnormality profile;

(iii) The extension of other unsupervised algorithms to deliver granular population-wide decision support;

(iv) The investigation of the generalization capacity of various algorithms across different semantic definitions of normality; and

(v) A pragmatic alternative offered by our approach to the bandit model literature proposing fully autonomous decision systems [11] that may offer opportunities for extending the proposed approach that are yet to be explored.

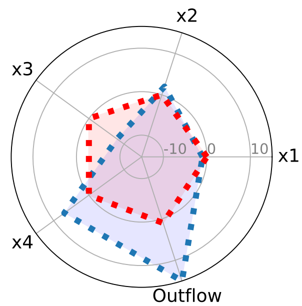

Figure .8: Label Imperfection: the red area shows a non-deviating profile, while the blue area shows a significant deviation with respect to the expected outflow.

Table .7: D1 and D2 Features

Features	Type
Fixed Wage	Continuous
Variable Wage	Continuous
Vacation bonus	Continuous
End-of-Year Bonus	Continuous
Benefits in Kind	Continuous

Features	Type
Age	Continuous
Company Cars	Continuous
Inflow	Continuous
Outflow	Continuous
Average Wage	Continuous
Average Variable Wage	Continuous
Wage Range (Max-Min)	Continuous
Gender	Continuous
Hours of Education Leave	Continuous
Education level	Continuous
Number of Employees	Continuous

Table .8: Expert Spearman Rank Correlation D1

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7	Expert 8	Expert 9	Expert 10
Expert 1	1.000	0.605	0.675	0.381	0.397	0.418	0.468	0.321	0.200	0.376
Expert 2	0.605	1.000	0.494	0.245	0.779	0.731	0.677	0.704	0.261	0.602
Expert 3	0.675	0.494	1.000	0.322	0.387	0.281	0.260	0.327	0.055	0.245
Expert 4	0.381	0.245	0.322	1.000	0.205	-0.001	0.074	0.162	0.187	0.404
Expert 5	0.397	0.779	0.387	0.205	1.000	0.684	0.691	0.698	0.333	0.443
Expert 6	0.418	0.731	0.281	-0.001	0.684	1.000	0.779	0.534	0.246	0.333
Expert 7	0.468	0.677	0.260	0.074	0.691	0.779	1.000	0.643	0.263	0.446
Expert 8	0.321	0.704	0.327	0.162	0.698	0.534	0.643	1.000	0.432	0.641
Expert 9	0.200	0.261	0.055	0.187	0.333	0.246	0.263	0.432	1.000	0.344
Expert 10	0.376	0.602	0.245	0.404	0.443	0.333	0.446	0.641	0.344	1.000

Table .9: Expert Spearman Rank Correlation D2

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7	Expert 8	Expert 9	Expert 10
Expert 1	1.000	0.053	0.221	-0.234	-0.082	-0.016	0.305	0.004	0.430	0.219
Expert 2	0.053	1.000	0.218	0.258	0.261	0.402	0.100	0.519	0.027	0.277
Expert 3	0.221	0.218	1.000	0.169	0.130	0.175	0.065	0.150	0.224	0.381
Expert 4	-0.234	0.258	0.169	1.000	0.380	0.129	-0.181	0.061	-0.109	0.424
Expert 5	-0.082	0.261	0.130	0.380	1.000	0.032	-0.177	0.205	0.251	0.023
Expert 6	-0.016	0.402	0.175	0.129	0.032	1.000	0.090	0.471	-0.067	0.373
Expert 7	0.305	0.100	0.065	-0.181	-0.177	0.090	1.000	0.168	0.034	0.428
Expert 8	0.004	0.519	0.150	0.061	0.205	0.471	0.168	1.000	0.177	0.201
Expert 9	0.430	0.027	0.224	-0.109	0.251	-0.067	0.034	0.177	1.000	-0.042
Expert 10	0.219	0.277	0.381	0.424	0.023	0.373	0.428	0.201	-0.042	1.000

Table .10: Implementations and parameter selection

Model	**Parameters** {employee, employer}
AE	Hidden layers: 3,8 Encoding dimensions: 4,7 Activation function: ‘SELU’ Loss = ‘MSE’ Optimizer: Adam Learning rate: 9.5e-3
Iforest	*Contamination* = 0.5
LOF	Distance: Minkowski with $p = 2$ k=max(n*0.1,50)

Table .11: Correlation Methods

Correlations		AE			Iforest			LOF		
		5%	10%	15%	5%	10%	15%	5%	10%	15%
	5%	1.00	0.83	0.67	0.57	0.47	0.39	0.55	0.64	0.61
AE	10%	0.83	1.00	0.81	0.46	0.46	0.45	0.77	0.77	0.73
	15%	0.67	0.81	1.00	0.49	0.44	0.59	0.60	0.75	0.70
	5%	0.57	0.46	0.49	1.00	0.73	0.54	0.53	0.46	0.44
Iforest	10%	0.47	0.46	0.44	0,73	1.00	0.75	0.42	0.41	0.38
	15%	0.39	0.45	0.59	0.54	0.75	1.00	0.45	0.34	0.30
	5%	0.55	0.77	0.60	0.53	0.42	0.45	1.00	0.68	0.65
LOF	10%	0.64	0.77	0.75	0.46	0.41	0.34	0.68	1.00	0.95
	15%	0.61	0.73	0.70	0.44	0.38	0.30	0.65	0.95	1.00

References

[1] B. Baesens, R. Setiono, C. Mues, J. Vanthienen, Using neural network rule extraction and decision tables for credit-risk evaluation, Management science 49 (2003) 312–329.

[2] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens, New insights into churn prediction in the telecommunication sector: A profit driven data mining approach, European Journal of Operational Research 218 (2012) 211–229.

[3] C. R. Schwenk, Cognitive simplification processes in strategic decision-making, Strategic management journal 5 (1984) 111–128.

[4] K. M. Eisenhardt, M. J. Zbaracki, Strategic decision making, Strategic management journal 13 (1992) 17–37.

[5] D. Kahneman, A. Tversky, Choices, values, and frames, in: Handbook of the Fundamentals of Financial Decision Making: Part I, World Scientific, 2013, pp. 269–278.

[6] J. A. Doukas, D. Petmezas, Acquisitions, overconfident managers and self-attribution bias, European Financial Management 13 (2007) 531–577.

[7] F. Liebl, J. O. Schwarz, Normality of the future: Trend diagnosis for strategic foresight, Futures 42 (2010) 313–327.

[8] C. C. Aggarwal, P. S. Yu, Outlier detection for high dimensional data, in: Proceedings of the 2001 ACM SIGMOD international conference on Management of data, 2001, pp. 37–46.

[9] D. Martens, J. Vanthienen, W. Verbeke, B. Baesens, Performance of classification models from a user perspective, Decision Support Systems 51 (2011) 782–793.

[10] M. K. Lee, Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management, Big Data & Society 5 (2018) 2053951718756684.

[11] J. Berrevoets, S. Verboven, W. Verbeke, Optimising individual-treatment-effect using bandits, in: Neural Information Processing Systems (NeurIPS) 2019, Curran Associates, Inc., 2019.

[12] C. Zhou, R. C. Paffenroth, Anomaly detection with robust deep autoencoders, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 665–674.

[13] T. Pinto, T. M. Sousa, I. Praça, Z. Vale, H. Morais, Support vector machines for decision support in electricity markets strategic bidding, Neurocomputing 172 (2016) 438–445.

[14] E. Stripling, B. Baesens, B. Chizi, S. vanden Broucke, Isolation-based conditional anomaly detection on mixed-attribute data to uncover workers' compensation fraud, Decision Support Systems 111 (2018) 13–26.

[15] D. L. Day, A. Y. Lewin, H. Li, Strategic leaders or strategic groups: A longitudinal data envelopment analysis of the us brewing industry, European Journal of Operational Research 80 (1995) 619–638.

[16] V. L. Sauter, Competitive intelligence systems: qualitative dss for strategic decision making, ACM SIGMIS Database: the DATABASE for Advances in Information Systems 36 (2005) 43–57.

[17] L. A. Leskinen, P. Leskinen, M. Kurttila, J. Kangas, M. Kajanus, Adapting modern strategic decision support tools in the participatory strategy process—a case study of a forest research station, Forest Policy and Economics 8 (2006) 267–278.

[18] B. De Martino, D. Kumaran, B. Seymour, R. J. Dolan, Frames, biases, and rational decision-making in the human brain, Science 313 (2006) 684–687.

[19] D. Kahneman, S. Frederick, Representativeness revisited: Attribute substitution in intuitive judgment, Heuristics and biases: The psychology of intuitive judgment 49 (2002) 81.

[20] G. Gigerenzer, W. Gaissmaier, Heuristic decision making, Annual review of psychology 62 (2011) 451–482.

[21] G. H. Van Bruggen, A. Smidts, B. Wierenga, Improving decision making by means of a marketing decision support system, Management Science 44 (1998) 645–658.

[22] K. D. Brouthers, F. Andriessen, I. Nicolaes, Driving blind: Strategic decisionmaking in small companies, Long range planning 31 (1998) 130–138.

[23] R. L. Daft, J. Sormunen, D. Parks, Chief executive scanning, environmental characteristics, and company performance: An empirical study, Strategic management journal 9 (1988) 123–139.

[24] M. Yang, M. Hyland, Who do firms imitate? a multilevel approach to examining sources of imitation in the choice of mergers and acquisitions, Journal of Management 32 (2006) 381–399.

[25] E. Y. Zhao, G. Fisher, M. Lounsbury, D. Miller, Optimal distinctiveness: Broadening the interface between institutional theory and strategic management, Strategic Management Journal 38 (2017) 93–113.

[26] C. Wiedeck, A. Engelen, The copycat cmo: firms' imitative behavior as an explanation for cmo presence, Journal of the Academy of Marketing Science 46 (2018) 632–651.

[27] B. Baesens, V. Van Vlasselaer, W. Verbeke, Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection, John Wiley & Sons, 2015.

[28] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 1975–1981.

[29] M. H. Bhuyan, D. K. Bhattacharyya, J. K. Kalita, Network anomaly detection: methods, systems and tools, Ieee communications surveys & tutorials 16 (2013) 303–336.

[30] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, P. Suetens, et al., Automated segmentation of multiple sclerosis lesions by model outlier detection, IEEE transactions on medical imaging 20 (2001) 677–688.

[31] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM computing surveys (CSUR) 41 (2009) 15.

[32] E. Schubert, R. Wojdanowski, A. Zimek, H.-P. Kriegel, On evaluation of outlier rankings and outlier scores, in: Proceedings of the 2012 SIAM International Conference on Data Mining, SIAM, 2012, pp. 1047–1058.

[33] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, M. E. Houle, On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study, Data Mining and Knowledge Discovery 30 (2016) 891–927.

[34] A. Zimek, R. J. Campello, J. Sander, Ensembles for unsupervised outlier detection: challenges and research questions a position paper, Acm Sigkdd Explorations Newsletter 15 (2014) 11–22.

[35] W. Lu, Y. Cheng, C. Xiao, S. Chang, S. Huang, B. Liang, T. Huang, Unsupervised sequential outlier detection with deep architectures, IEEE transactions on image processing 26 (2017) 4321–4330.

[36] C. Fan, F. Xiao, Y. Zhao, J. Wang, Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data, Applied energy 211 (2018) 1123–1135.

[37] M. Sakurada, T. Yairi, Anomaly detection using autoencoders with non-linear dimensionality reduction, in: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, ACM, 2014, p. 4.

[38] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, Lof: identifying density-based local outliers, in: ACM sigmod record, volume 29, ACM, 2000, pp. 93–104.

[39] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE, 2008, pp. 413–422.

[40] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[41] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[42] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, Journal of machine learning research 11 (2010) 3371–3408.

[43] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, science 313 (2006) 504–507.

[44] S. C. AP, S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V. C. Raykar, A. Saha, An autoencoder approach to learning bilingual word representations, in: Advances in Neural Information Processing Systems, 2014, pp. 1853–1861.

[45] J. Chen, S. Sathe, C. Aggarwal, D. Turaga, Outlier detection with autoencoder ensembles, in: Proceedings of the 2017 SIAM International Conference on Data Mining, SIAM, 2017, pp. 90–98.

[46] M. Goldstein, S. Uchida, A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, PloS one 11 (2016) e0152173.

[47] M. A. Siddiqui, J. W. Stokes, C. Seifert, E. Argyle, R. McCann, J. Neil, J. Carroll, Detecting cyber attacks using anomaly detection with explanations and expert feedback, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2872–2876.

[48] T. G. Dietterich, Ensemble methods in machine learning, in: International workshop on multiple classifier systems, Springer, 2000, pp. 1–15.

[49] L. K. Hansen, P. Salamon, Neural network ensembles, IEEE Transactions on Pattern Analysis & Machine Intelligence (1990) 993–1001.

[50] F. Galton, Vox populi (the wisdom of crowds), Nature 75 (1907) 450–451.

[51] K. Das, J. Schneider, Detecting anomalous records in categorical datasets, in: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2007, pp. 220–229.

[52] G. Nychis, V. Sekar, D. G. Andersen, H. Kim, H. Zhang, An empirical evaluation of entropy-based traffic anomaly detection, in: Proceedings of the 8th ACM SIGCOMM conference on Internet measurement, ACM, 2008, pp. 151–156.