South African Journal of Business Management
ISSN: (Online) 2078-5976, (Print) 2078-5585

# Data-related risks for the use of machine learning in retail customer demand forecasting

Check for updates

**Authors:**
Lee-Ann Pietersen[1]
Riaan J. Rudman[1]

**Affiliations:**
[1]School of Accountancy, Faculty of Economic and Management Sciences, Stellenbosch University, Stellenbosch, South Africa

**Corresponding author:**
Riaan Rudman,
rjrudman@sun.ac.za

**Dates:**
Received: 24 June 2024
Accepted: 19 Mar. 2025
Published: 09 May 2025

**How to cite this article:**
Pietersen, L., & Rudman, R.J. (2025). Data-related risks for the use of machine learning in retail customer demand forecasting. *South African Journal of Business Management, 56*(1), a4766. https://doi.org/10.4102/sajbm.v56i1.4766

**Copyright:**
© 2025. The Authors. Licensee: AOSIS. This work is licensed under the Creative Commons Attribution License.

**Read online:**
Scan this QR code with your smart phone or mobile device to read online.

**Purpose:** The use of machine learning in customer demand forecasting is reliant on quality data sources. Data should be governed and managed appropriately to ensure that customer demand forecasting is accurate. Most retailers, however, do not understand the technology and are unable to identify all the risks. The purpose of this study is to identify significant data-related risks which arise from the use of machine learning for customer demand forecasting.

**Design/methodology/approach:** A structured literature review was conducted to obtain an understanding of machine learning used for customer demand forecasting and data governance mechanisms required to appropriately manage data assets. Using this understanding, the data governance principles and objectives of the Data Management Body of Knowledge developed by The Global Data Management Community (DAMA DMBOK) and Control Objectives for Information and Related Technologies 2019 (COBIT-2019) governance frameworks were used to identify the data-related risks in a comprehensive manner.

**Findings/results:** Several significant data-related risks arising from the implementation of machine learning for retail customer demand forecasting were identified. These risks link to each stage and component of the machine learning system development life cycle.

**Practical implications:** The risks can be used by internal and external auditors, as well as those charged with governance and other management functions within an organisation, to identify and evaluate risks arising from the use of machine learning within their organisation.

**Originality/value:** While previous studies identify risks on an *ad hoc* basis, this study used the COBIT-2019 and DAMA DMBOKv2 governance frameworks as the foundation for the identification of risks to ensure completeness and rigour of the risks identified.

**Keywords:** machine learning; retail customer demand forecasting; significant data risks; risk assessment; risk management; IT governance; data governance; COBIT-2019.

## Introduction

Traditional customer demand forecasting techniques, which use statistical and mathematical processes, rely on historical sales information of retailers (Vandeput, 2021). However, historical sales information may not be sufficient to predict the customer demand of fast-moving consumer goods such as perishable food items. Furthermore, traditional techniques are not timely and do not take contextual information, which provide further insights into customer behaviour (such as weather patterns, regional and competitor data), into account (Kaleva & Småros, 2023). Machine learning, which is capable of analysing different types of data from various sources, may address this concern. Using a machine learning system for customer demand forecasting does, however, come with a range of significant data-related risks that retailers should consider at the different stages of the machine learning life cycle. Those responsible for designing and implementing machine learning systems in retail often lack the necessary data governance skills and experience to prevent these project failures. Implementing machine learning-based customer demand forecasting without properly understanding the data requirements of the project or risks related to the use of the data, could lead to project failures (Sankaran, 2021). Limited guidance is, however, available to the Chief Information Officers of retailers to identify these significant data-related risks associated with the use of machine learning for customer demand forecasting (Van Wyk, 2018). Data-related risks are the risks that the objectives for data assets used in machine learning for retail customer demand forecasting of perishable foods are not met. These objectives consist of the quality, integrity, privacy and security of data assets.

**Note:** Special collection: The manuscript is a contribution to the themed collection titled 'Corporate Governance and Sustainable Business Practices in the Fourth Industrial Revolution', under the expert guidance of guest editors Prof. Nicolene Wesson and Dr. George Frederick Nel.

Furthermore, the current literature on the data-related risks associated with the use of machine learning (such as Angarita-Zapata et al., 2021; Raguseo, 2018; Sharma et al., 2020; Van Wyk, 2018) is not comprehensive and does not focus on the specific context of using the technology for customer demand forecasting in the retail industry. This study aims to identify the significant data-related risks introduced by the use of machine learning systems for customer demand forecasting of perishable foods by retailers. The research focuses on the significant data-related risks introduced by a machine learning system and the data used and does not attempt to provide an exhaustive list of all risks associated with the use of machine learning. The use of two well-known governance frameworks ensures that the risks identified are comprehensive and complete.

## Methodology

A qualitative, non-empirical study was conducted with a review of existing literature to obtain an understanding of the technology of machine learning and the context in which it is utilised to make predictions which form the basis of the research study. Once an understanding was obtained about the use of machine learning systems for customer demand forecasting of perishable foods by retailers, the control objectives and underlying guidance of two governance frameworks were used to identify the significant data-related risks when machine learning is utilised. These frameworks were selected and utilised as the foundation of the research to ensure the academic rigour and the completeness of the risks identified. This addresses the weaknesses in other studies (by *inter alia* Angarita-Zapata et al., 2021; Raguseo, 2018; Sharma et al., 2020; Van Wyk, 2018) that investigate risks in technologies that identify risks on an *ad hoc* basis and that do not use an established theory or framework. The following structured approach, similar to that used by *inter alia* Sahd (2016) and Van Wyk (2018), was followed:

- **Structured literature review:** The structured literature review was performed to obtain a better understanding of the different elements needed to achieve the research objective, namely: an understanding of the industry, customer demand forecasting, the data, the technology and data governance. The sources were limited to focus on the identification of risks to data used by the machine learning system in retail customer demand forecasting of perishable foods. In order to add scientific rigour to a literature review and obtain a strong theoretical basis for the research, a four-stage approach is suggested by Sylvester et al. (2013). The four stages include the searching, mapping, appraisal and synthesis of literature. A wide selection of articles and readings was selected in the beginning stages to enable a comprehensive understanding of the underlying literature and the selection was narrowed down to more specific areas in the latter stages. This process resulted in the original search result of 395 sources being narrowed down to 95 relevant sources, which were read in depth and used as the data sources for this study. While not all of these data sources were used in the final identification of significant data-related risks, they served as a source to broaden the authors' understanding of the different topics discussed above. The final selection of 49 data sources which were used as a base for the identification of data-related risks include a wide range of literature types, which are set out in Table 1.
- **Understand data governance:** A review of COBIT-2019 was performed to identify the relevant governance and management objectives useful for data governance. In addition, the principles and guidelines on data governance and management detailed in the 10 knowledge areas of the DAMA DMBOKv2 were investigated to provide more detail for the objectives of corporate governance listed in COBIT-2019. The knowledge areas of the DAMA DMBOKv2 were then linked to the COBIT-2019 objectives for which they provide a clearer guidance with a specific focus on data governance to form a governance matrix that will be used to ensure that risks identified are comprehensive.
- **Identify significant data-related risks:** The data-related risks identified from the prior literature as part of the structured literature review were mapped to the objectives and knowledge areas of the governance matrix developed in step two. The objectives and knowledge areas were then used to perform a completeness check and identify additional risks that were not found as part of the structured literature review. The risks identified using the governance frameworks were then grouped into strategic, operational and technical risks.

The Literature Review details a brief literature review to provide an understanding of the industry and customer demand forecasting, the technology, the data and data governance. The subsequent section, titled Data-related risks, discusses the identified significant data-related risks at each stage of the machine learning life cycle. Finally, the Conclusion summarises the key findings of the study and identifies potential future areas of research.

### Ethical considerations

Ethical clearance to conduct this study was obtained from the Stellenbosch University, Research Ethics Committee: Social, Behavioural and Education Research (REC: SBE). (No. ACC-2023-27695).

**TABLE 1:** Types of data sources used.

| Literature type | Number of sources |
|---|---|
| Articles in accredited journals | 30 |
| Popular articles and web pages | 7 |
| Books | 4 |
| Conference papers | 3 |
| White papers, laws and regulations | 4 |
| Dissertations | 1 |
| **Total data sources** | **49** |

# Literature review

## The perishable food industry and customer demand forecasting

Retailers are often the ones in the food supply chain that bear the cost of carrying the stock until it is sold to consumers. When a retailer overstocks, the excess inventory needs to be stored and managed, leading to additional costs. Moreover, when perishable foods are past their sell-by dates or start to spoil, it is the retailer's responsibility to properly dispose of these foods, which contributes towards food waste (Domingo, 2023). Food waste has gathered increasing attention globally in recent years, which is emphasised by the United Nations Sustainable Development Goals calling for a 50% reduction in food waste created at retail and consumer level by 2030 (United Nations, 2022). It is frequently argued that inadequate customer demand forecasting is seen as the cause of food waste by retailers (De Moraes et al., 2020). De Moraes et al. (2020) add that the lack of forecasting accuracy and demand variation are the main reasons for this. The demand for food products by retail customers constantly fluctuates because of various factors which are mostly out of the retailer's control (Tsoumakas, 2019), such as the weather, economic conditions and trends in customer preference. Perishable foods typically have short shelf-lives which leave retailers with a limited timeframe in which these foods need to be sold. This is a balancing act – if a retailer chooses to be more conservative in the quantity of food stocked and there is a shortage, the retailer loses out on potential income that could have been gained had the demand of customers been met. Accurate and timeous, real-time customer demand forecasting is therefore imperative for retailers who stock perishable foods.

### The data used in retail customer demand forecasting

Perishable food items have a limited timeframe in which they need to reach the consumer to prevent spoilage. Traditional forecasting methods typically use historical sales figures per product (Guo et al., 2013; Kumar et al., 2020). While historical sales data is still valuable in providing insights into the purchasing behaviours of customers (Guo et al., 2013; Kumar et al., 2020), constantly changing market conditions are reducing the value of this historical data and trends may change significantly over time (Zhu et al., 2021). Traditional methods are also unable to process large volumes of complex data and are too slow for the dynamic perishable food retail industry, making them ineffective (Tarallo et al., 2019). Including other non-demand data, otherwise known as contextual data, can vastly increase the accuracy of customer demand predictions (Kumar et al., 2020). Contextual data include information on the economic environment, weather conditions, regional and competitor data, seasonality, social media data, availability of and promotions on alternative products and the attributes of a specific product (Guo et al., 2013). Contextual data are often also presented as unstructured data. Unstructured data are more difficult to sort and analyse than structured data are as these do not have a predefined data model and are stored in its unprocessed, native format until it is used (Yang et al., 2019). While this makes it more adaptable to the needs of the organisation and increases the speed at which it can be collected, analysis of unstructured data requires specialised tools and more expertise. Furthermore, it requires more storage space and processing power than structured datasets (Yang et al., 2019).

### Laws and regulations governing the data used by retailers

The use of customer data is critical to understand and predict customer demand (Oosthuizen, 2021), but using a customer's data comes with various privacy concerns, which can lead to legal challenges as there are several laws which regulate how organisations can use and store personal data (Kaplan & Haenlein, 2019; Oosthuizen, 2021). One such law is the Protection of Personal Information (POPI) Act No. 4 of 2013 (POPI Act) which is applicable in South Africa. Although there are some differences between the POPI Act and other laws governing the protection of personal data around the world, such as the European Union's General Data Protection Regulation (Regulation [EU] 2016/679) (GDPR), the principles of these acts are all aligned (De Bruyn, 2014). The focus point of most of these laws and regulations is the protection of the privacy of information that can be used to identify an individual person, otherwise known as Personally Identifiable Information (PII). As the customer data used by retailers for customer demand forecasting often contains PII, the protection of the privacy of retail customers is of the utmost importance.

## Machine learning

With the large volumes and variety of data being generated and collected by retailers, machine learning, which is heavily dependent on data, can be used to enable more accurate customer demand predictions. Machine learning can be defined as a computational system that acts intelligently by accurately interpreting data, learning from that data and then using what it has learned to achieve specific goals without being explicitly programmed to do so (Alpaydin, 2020; Kaplan & Haenlein, 2019; Mohri et al., 2018).

### The building blocks of a machine learning system

Machine learning works by feeding input data to a machine; selecting an algorithm; configuring and adjusting parameters and settings and then instructing the machine to analyse the data. The machine then continues to look for patterns within the input data through a process of trial and error and forms a data model. This model can then be used to predict future values, gain knowledge about a dataset or both (Alpaydin, 2020; Theobald, 2017). The three main building blocks of a machine learning system are (Oosthuizen, 2021):

- **Input data:** Data are fundamental to the effective functioning of a machine learning system as it needs input data to learn from its past experiences. Each set of data used by a machine learning system contains data items with different features (Theobald, 2017). Each data item would also have different labels. Labels are values or categories assigned to data items that are used to train certain models (Mohri et al., 2018), such as the fact that an orange is categorised as citrus fruit. Training data are used to train the machine and develop the data model; validation data are used to fine-tune the parameters of the algorithm to refine results; test data tests the accuracy of the model's output and inference data are the actual data used to make predictions (Mohri et al., 2018; Theobald, 2017).

- **Processing:** Inputted datasets are processed by the machine through the algorithm and machine learning model (Oosthuizen, 2021). An algorithm can be defined as a sequence of transactions that transform input into output (Alpaydin, 2020). The machine develops the model based on the algorithm and what it has learnt from the training and validation data. The algorithms that the machine uses to process the input data, are built on statistical methods and theories. They can be broadly categorised into the following three categories based on the features and labels of datasets used to train the machine, as well as how the data are received and how test data are used to evaluate the algorithm (Alpaydin, 2020; Mohri et al., 2018; Theobald, 2017):
    - *Supervised machine learning* – The machine is trained through labelled datasets where the relationship between the features and the outcome is known, and the machine uses patterns in the data to create a model;
    - *Unsupervised machine learning* – The machine is trained with unlabelled datasets where not all features and data patterns are classified in the training data, to find hidden patterns; and
    - *Reinforcement learning* – The machine is trained with input data where the outputs are not labelled but graded. The machine interacts with the environment and receives a reward based on the accuracy of each action with the objective to maximise the reward.

Irrespective of the category in which the algorithm falls, all machine learning systems function on the fundamental principle of generalisation (Mohri et al., 2018). Generalisation means that the model can use what it has learned from the training data and accurately predict the outcome of the new dataset even though the features might differ from the training data (Alpaydin, 2020):

- **Output:** After the processing is completed, the machine learning system generates an output which is then used by businesses for decision-making purposes (Oosthuizen, 2021).

In order to process input data into usable output, different components are required to form a complete machine learning system comprising data, infrastructure and other resources. These components should be able to function together effectively and align with organisational objectives and requirements. To develop an effective machine learning system, an organisation should follow the machine learning life cycle.

### The machine learning life cycle

As a result of the complexity and data-driven nature of machine learning systems, the system development life cycle associated with machine learning is not as straightforward as that of traditional software solutions and information technology (IT) systems (Laato et al., 2022). Although the key stages of system development, known as design, development and implementation, remain the same with the development of a machine learning system, the components of these stages differ. The machine learning life cycle as proposed by Laato et al. (2022) consists of the three key stages of the development life cycle with the different components under each of these stages explained in further detail below:

- **System design:** The components within this stage of the machine learning life cycle consist of requirement gathering, data resources and environmental analysis. During the system design stage of the life cycle, the requirements of the system needed to achieve the objectives and strategy of the organisation are determined. An environmental analysis evaluates the resources currently available within the organisation to determine whether they are available and sufficient. It is in this stage of the life cycle that the amount and types of data required for the project are determined and sources are identified from where they can be obtained. This stage of the machine learning life cycle is data-centric compared to a normal system development life cycle with a focus on obtaining the right amount and types of data, data quality and preparing the data for processing. Furthermore, machine learning projects require collaboration between highly skilled individuals such as data scientists, domain experts, those charged with governance and other stakeholders to ensure that the system design stage is successful. Once the organisation has designed the requirements and objectives for the machine learning project and the resources required have been determined, the system can now be developed.
- **System development:** The system development stage consists of the data management, model development and model testing components. At this stage, the various classes of input data are obtained and preprocessed. Furthermore, the machine is trained and the model developed, refined and tested. Once the model functions in alignment with requirements and objectives, it can be implemented. The repeated training and evaluation of the machine learning system in this stage set this life cycle apart from a normal system development life cycle. The unpredictability of machine learning systems leads to an increased need for training with multiple datasets to refine the model.
- **System implementation:** This stage is made up of two components, namely deployment and monitoring. The trained and tested machine learning model is deployed and integrated into the IT environment and inference data is used to make customer demand predictions. The performance of the system, infrastructure, controls and human resources are continuously monitored. The constant monitoring, re-evaluation and possible retraining of the machine learning system sets this stage of the life cycle apart from a normal system development life cycle.

While this section provided some much-needed background and understanding of the principles on which machine learning operates and how it is developed, it is important for the organisation to understand how machine learning is used for customer demand forecasting in retail.

## Data governance

To identify the data-related risks associated with the use of machine learning for customer demand forecasting, an understanding of data governance is required.

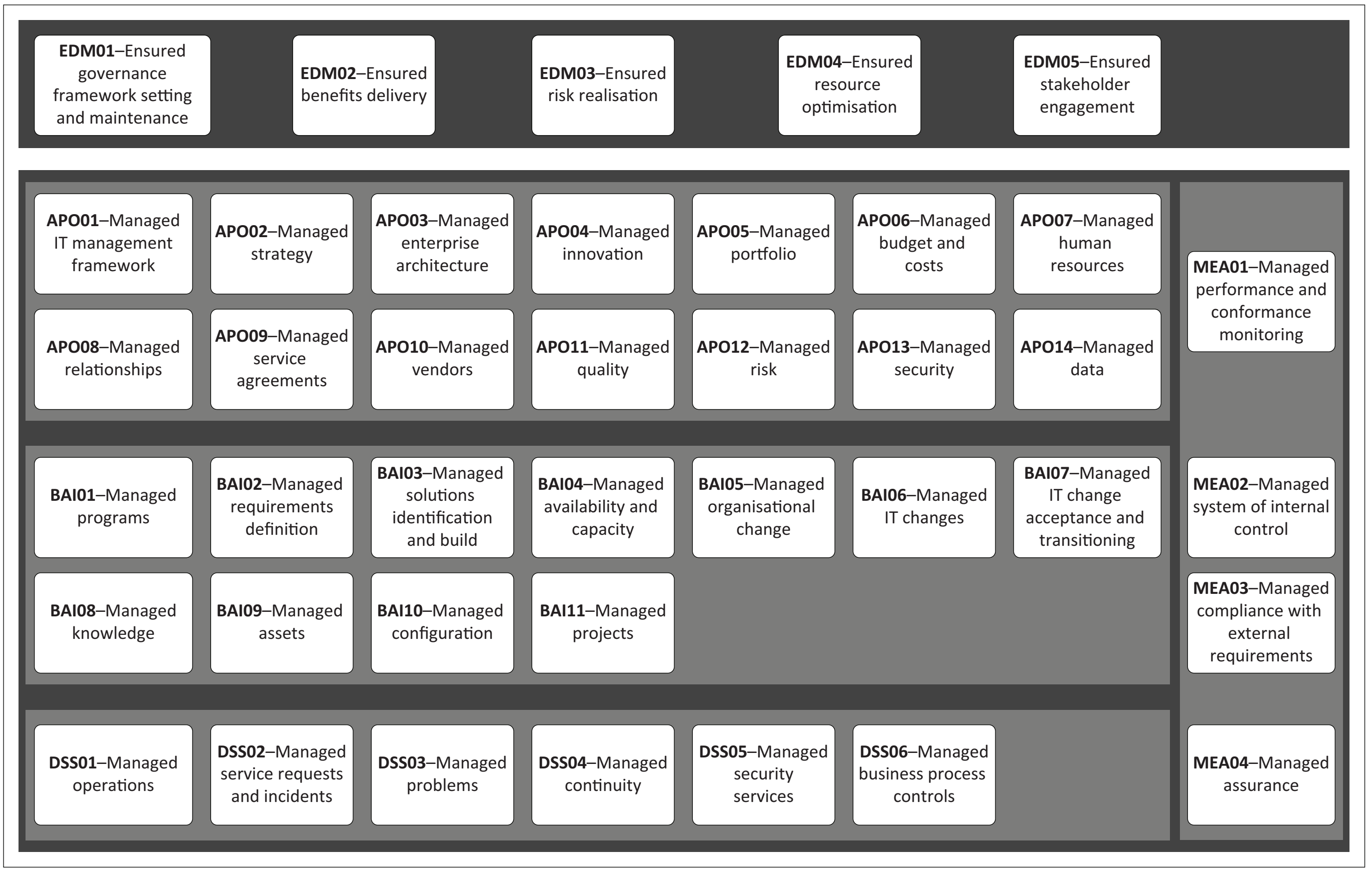
Data governance provides a framework for decision-making and accountability around data assets in the form of standards, policies and procedures (Brous et al., 2016). When data is governed appropriately, it ensures that all parties involved follow the organisation-wide agenda in terms of the strategic objectives towards data governance, that the value of data is maximised and that all data-related risks are sufficiently managed and addressed (Abraham et al., 2019).

Using a framework to provide structure, facilitates data governance and is essential when using a machine learning system. A governance framework ensures that comprehensive evaluations have been conducted to identify all risks in order to appropriately respond to them (Hendrawan et al., 2022). Two comprehensive frameworks were investigated to form an understanding of what data governance entails and how it can be used to identify data-related risks. The two frameworks are the COBIT-2019 framework developed by ISACA (2018) and the DAMA DMBOKv2 developed by DAMA International (2017). Each of these frameworks is discussed in more detail in the sections that follow.

### COBIT-2019 governance framework

COBIT-2019, focused on enterprise and IT governance, is aimed at being more flexible and tailored to a specific organisation (De Haes et al., 2020; ISACA, 2018). COBIT-2019 is a useful tool to ensure that the significant data-related risks identified are valid, accurate and complete. COBIT-2019 provides a comprehensive, internationally recognised framework which is well-known and frequently used in business (De Haes et al., 2020). The 40 governance and management objectives of COBIT-2019 are summarised in Figure 1.

Although COBIT-2019 does have objectives that specifically address the data used and produced by an organisation, it is not sufficient to use this framework alone for data governance. COBIT-2019 does not have the necessary detail on specific data considerations to ensure its quality, integrity and security. As it is such a comprehensive and well-known governance framework, it is still useful in conjunction with a more detailed data governance framework to ensure that data-related risks are identified. One such framework which



**EDM01**–Ensured governance framework setting and maintenance
**EDM02**–Ensured benefits delivery
**EDM03**–Ensured risk realisation
**EDM04**–Ensured resource optimisation
**EDM05**–Ensured stakeholder engagement

**APO01**–Managed IT management framework
**APO02**–Managed strategy
**APO03**–Managed enterprise architecture
**APO04**–Managed innovation
**APO05**–Managed portfolio
**APO06**–Managed budget and costs
**APO07**–Managed human resources
**APO08**–Managed relationships
**APO09**–Managed service agreements
**APO10**–Managed vendors
**APO11**–Managed quality
**APO12**–Managed risk
**APO13**–Managed security
**APO14**–Managed data

**BAI01**–Managed programs
**BAI02**–Managed requirements definition
**BAI03**–Managed solutions identification and build
**BAI04**–Managed availability and capacity
**BAI05**–Managed organisational change
**BAI06**–Managed IT changes
**BAI07**–Managed IT change acceptance and transitioning
**BAI08**–Managed knowledge
**BAI09**–Managed assets
**BAI10**–Managed configuration
**BAI11**–Managed projects

**DSS01**–Managed operations
**DSS02**–Managed service requests and incidents
**DSS03**–Managed problems
**DSS04**–Managed continuity
**DSS05**–Managed security services
**DSS06**–Managed business process controls

**MEA01**–Managed performance and conformance monitoring
**MEA02**–Managed system of internal control
**MEA03**–Managed compliance with external requirements
**MEA04**–Managed assurance

*Source*: Edmead, M.T. (2020). *Using COBIT 2019 to plan and execute an organization's transformation strategy*. ISACA. Retrieved from https://www.isaca.org/resources/news-and-trends/industry-news/2020/using-cobit-2019-to-plan-and-execute-an-organization-transformation-strategy

IT, information technology.

**FIGURE 1:** The governance and management objectives of COBIT-2019.

provides more detail on data governance and management, is the DAMA DMBOKv2 framework.

### DAMA DMBOKv2 framework

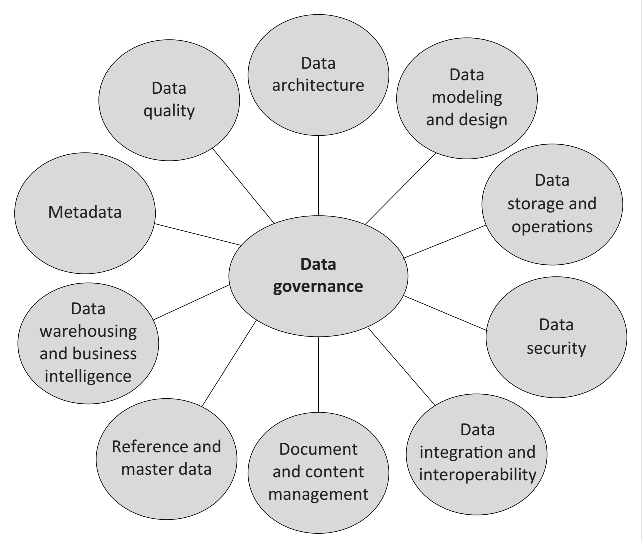
DAMA DMBOKv2 does not only provide a comprehensive approach to data governance, but it also includes practical principles, practices, methods, techniques and metrics for the implementation of successful governance and management of data (DAMA International, 2017; Hendrawan et al., 2022). Understanding these principles aids in the comprehensive identification of significant data-related risks.

The DAMA DMBOKv2 framework approach to data governance involves 10 knowledge areas which are all centred around data governance. They all form part of a mature management function and are all equally important. Governance is central to all these areas as without appropriate governance, there would be no consistency and balance between the different areas (DAMA International, 2017; Hendrawan et al., 2022). Figure 2 shows the different knowledge areas of the DAMA DMBOKv2 centred around data governance.

To aid with the identification of significant data-related risks, the knowledge areas of the DAMA DMBOKv2 were linked to the COBIT-2019 objectives for which they provide clearer guidance with a specific focus on data governance.

## Data-related risks

The data management knowledge areas of the DAMA DMBOKv2 and the COBIT-2019 objectives are used to identify the significant data-related risks through the application of each process and knowledge area to the use of machine learning in each stage of the retail customer demand forecasting life cycle. A detailed mapping is available from the authors on request. A risk matrix (Table 2) which contains the significant data-related risks to the various stages of the machine learning life cycle summarises the results. The 'blacked-out areas' indicate where the risks originate within the machine learning life cycle. Table 2 can be used by those charged with governance and other management functions (such as the Chief Information Officer, others charged with the governance of the machine learning system or the data used, or the risk committee and audit committee) in a retail environment to identify and evaluate the data-related risks in the various stages and components of the machine learning life cycle when machine learning is used for retail customer demand forecasting. Furthermore, the risks can be used by internal and external auditors as a basis for the planning of their audit and to obtain a better understanding of the risks involved in using machine learning for customer demand forecasting.


Data quality
Data architecture
Data modeling and design
Metadata
Data storage and operations
Data governance
Data warehousing and business intelligence
Data security
Reference and master data
Document and content management
Data integration and interoperability

*Source*: Adapted from DAMA International. (2017). *DAMA-DMBOK: Data management body of knowledge* (2nd ed.). Technics Publications, LLC

**FIGURE 2:** The knowledge areas of the DAMA DMBOKv2.

The risks are categorised under three categories: risks at a *strategic level*, which are high-level risks arising from inadequate or ineffective governance and management of data; risks at an *operational level* arising from inadequate or ineffective processes and systems for the management of resources; and risks at a *technical level* that are specific to the data and components of the machine learning system and data life cycle.

Figure 3 summarises the risks identified under each of the three categories. These risks are further discussed in the sections that follow.

## Risks at a strategic level

Risks at a strategic level are high-level risks which are present because of inadequate or ineffective governance and management of data used in the machine learning system. The various origins of risks arising from inadequate or ineffective data governance will be discussed in further detail below.

### Inadequate data governance policies

Inadequately documented, or incomprehensive and ineffective data governance policies could lead to the system not achieving the business objectives (KPMG, 2018).

### A lack of involvement by the governing body

The governing body not being involved in the development and integration of policies and procedures to address the adoption of the new machine learning system, and the governance and management of data, negatively impacts the governance culture of and acceptance by the rest of the organisation (IoDSA, 2016; Van Wyk & Rudman, 2019).

### Inadequate stewardship and oversight

If a sufficiently qualified Chief Data Officer is not assigned as a steward of the data, it may lead to insufficient data governance and miscommunication between the organisation and its stakeholders (DAMA International, 2017).

**TABLE 2:** Risk matrix linking the significant data-related risks to the various stages of the machine learning life cycle.

| Significant data-related risks | Stages of the machine learning life cycle | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | System design | | | System development | | | System implementation | |
| | Requirement gathering | Data resources | Environmental analysis | Data management | Model development | Model testing | Deployment | Monitoring |
| **Strategic risks** | | | | | | | | |
| Inadequate governance and management | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| **Operational risks** | | | | | | | | |
| Human resources | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Escalating cost | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| • Design and development | ■ | - | ■ | - | ■ | ■ | - | - |
| • Infrastructure, data and resources | - | ■ | - | ■ | ■ | ■ | ■ | - |
| • Monitoring and management | - | - | - | ■ | - | - | - | ■ |
| Service providers | ■ | - | ■ | ■ | ■ | - | ■ | ■ |
| **Technical risks** | | | | | | | | |
| *Risks affecting the acceptability of and access to data* | | | | | | | | |
| Data quality | ■ | ■ | - | ■ | - | ■ | - | ■ |
| Bias | ■ | ■ | - | ■ | - | ■ | - | ■ |
| Availability | ■ | ■ | - | ■ | - | ■ | - | ■ |
| Accessibility | ■ | ■ | - | ■ | - | ■ | - | ■ |
| *The effective functioning of components responsible for data management* | | | | | | | | |
| Interoperability | ■ | ■ | - | ■ | - | ■ | - | ■ |
| • Technical | ■ | - | ■ | - | ■ | ■ | ■ | ■ |
| • Semantic | ■ | ■ | ■ | - | ■ | ■ | ■ | ■ |
| • Syntactic | ■ | ■ | ■ | ■ | - | - | - | ■ |
| Integration | ■ | ■ | ■ | ■ | ■ | ■ | - | ■ |
| Scalability | ■ | - | ■ | - | ■ | - | - | ■ |
| *Compliance with legal and regulatory requirements* | | | | | | | | |
| Security | ■ | | ■ | | ■ | | | ■ |
| • Compliance | ■ | ■ | - | ■ | - | - | - | ■ |
| • Unauthorised access | - | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| • Insider breaches | - | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Privacy | | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| • Data subject rights violations | ■ | ■ | - | ■ | - | - | - | ■ |
| • Manipulation of data | - | ■ | - | ■ | ■ | ■ | - | ■ |
| • Security breaches | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Ownership | ■ | ■ | - | ■ | - | - | - | ■ |

### A lack of strategic and operational alignment

If a data strategy is not in place to ensure alignment between the strategy of the technology and data, and the goals and objectives of the business or if the strategy does not align with business objectives, there is a risk that the business might not achieve its business objectives (DAMA International, 2017; KPMG, 2018). Moreover, insufficient operational alignment between the business and IT processes can lead to the system being used inappropriately (KPMG, 2018).

### Inadequate stakeholder engagement

Without engagement with all relevant stakeholders, the business requirements for machine learning and data could be incomplete and stakeholders could be more resistant to change (DAMA International, 2017).

### Inadequate management of value-delivery

The cost of the investment in the system and data could exceed the return on investment with no value being added by the system (KPMG, 2018). It could also prevent stakeholder buy-in and impair the development and implementation of the system (Arunachalam et al., 2018).

### Inadequate change management

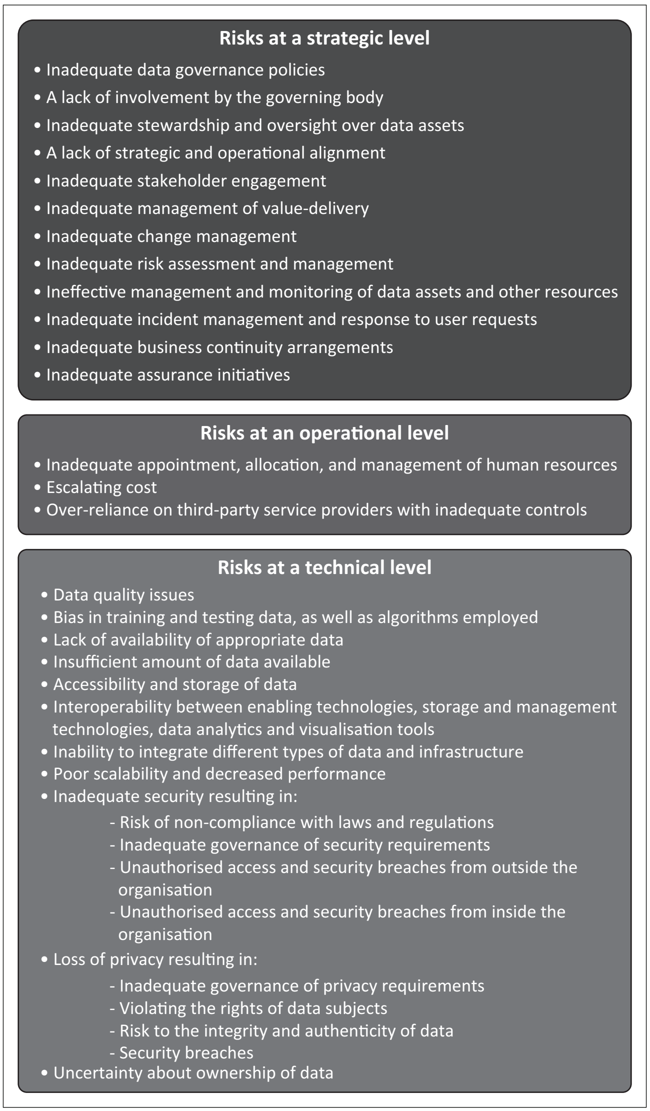
Stakeholders who are not prepared for change may be resistant to it. This could lead to a lack of buy-in from stakeholders and derail the strategy, policies and procedures of the organisation (DAMA International, 2017; Kusi-Sarpong et al., 2021).

### Inadequate risk assessment and management

Inadequate risk assessment procedures could lead to not all risks being identified in a timely manner or not at all. This would lead to risks exceeding the risk tolerance levels set by the governing body (IoDSA, 2016) and expose the organisation to operational, security and privacy problems (KPMG, 2018).

### Ineffective management and monitoring of data assets and resources

The system may not be functioning in line with business objectives and strategies, data may be at risk and business



**Risks at a strategic level**

- Inadequate data governance policies
- A lack of involvement by the governing body
- Inadequate stewardship and oversight over data assets
- A lack of strategic and operational alignment
- Inadequate stakeholder engagement
- Inadequate management of value-delivery
- Inadequate change management
- Inadequate risk assessment and management
- Ineffective management and monitoring of data assets and other resources
- Inadequate incident management and response to user requests
- Inadequate business continuity arrangements
- Inadequate assurance initiatives

**Risks at an operational level**

- Inadequate appointment, allocation, and management of human resources
- Escalating cost
- Over-reliance on third-party service providers with inadequate controls

**Risks at a technical level**

- Data quality issues
- Bias in training and testing data, as well as algorithms employed
- Lack of availability of appropriate data
- Insufficient amount of data available
- Accessibility and storage of data
- Interoperability between enabling technologies, storage and management technologies, data analytics and visualisation tools
- Inability to integrate different types of data and infrastructure
- Poor scalability and decreased performance
- Inadequate security resulting in:
  - Risk of non-compliance with laws and regulations
  - Inadequate governance of security requirements
  - Unauthorised access and security breaches from outside the organisation
  - Unauthorised access and security breaches from inside the organisation
- Loss of privacy resulting in:
  - Inadequate governance of privacy requirements
  - Violating the rights of data subjects
  - Risk to the integrity and authenticity of data
  - Security breaches
- Uncertainty about ownership of data

**FIGURE 3:** Significant data governance risks associated with the use of machine learning for retail customer demand forecasting.

opportunities could be overlooked. It could also lead to the misallocation of resources or resources being overutilised or underutilised (KPMG, 2018).

### Inadequate incident management

Failure to respond to user requests and detect and respond to data-related incidents or errors in the operation of machine learning components, escalates the issue and increases the risk of unauthorised changes, loss or corruption of data and data storage or transmission infrastructure being unavailable when needed (DAMA International, 2017; KPMG, 2018). Furthermore, incidents which are not effectively managed and responded to, are at risk of re-occurring (KPMG, 2018).

### Inadequate business continuity arrangements

There is a risk of financial losses and reputational damage if business processes are unable to continue (IoDSA, 2016; KPMG, 2018), especially in the case of perishable food products where inventory management is time-sensitive.

### Inadequate assurance initiatives

The governing body and other stakeholders would not be able to monitor whether the machine learning system or data governance mechanisms implemented, perform in line with expectations (Van Wyk & Rudman, 2019).

While the risks identified in this section are high-level, the processes and systems implemented to manage resources, data and related components of the machine learning system and data life cycle, lead to more detailed risks at an operational level.

## Risks at an operational level

Risks at an operational level arise from internal inefficiencies in the processes and systems that manage the resources of an organisation and can be divided into inadequate appointment and management of human resources, cost and using third-party service providers, which are further discussed.

### Inadequate appointment and management of human resources

The use of a machine learning system requires staff with unique skills which are typically advanced and scarce (KPMG, 2018; Kusi-Sarpong et al., 2021). Most companies implementing machine learning systems do not have access to these resources (Arunachalam et al., 2018; KPMG, 2018; Kusi-Sarpong et al., 2021). This creates a risk that processes implemented do not align with the business objectives and values; and in the event of major incidents, the organisation may not overcome and recover from this, which may damage the reputation of the retailer, or lead to legal action (KPMG, 2018; Van Wyk & Rudman, 2019).

Moreover, misallocation of these resources leads to underutilisation or overutilisation of resources (KPMG, 2018; Van Wyk & Rudman, 2019), while a lack of training on usage and interpretation could lead to the corruption of data or errors being made (Kusi-Sarpong et al., 2021). When roles and responsibilities are unclear or are not properly assigned, the risk of unauthorised or inappropriate access and changes to the system and data increases (KPMG, 2018).

### Escalating costs

Implementing a machine learning system together with the components required as part of the data life cycle has significant cost implications for an organisation (Arunachalam et al., 2018). Costs are incurred throughout the machine learning and data life cycle and consist of costs related to infrastructure; collection and processing of data; human resources; security and privacy and continuous monitoring and management of the system, data and resources (Arunachalam et al., 2018; Kusi-Sarpong et al., 2021). Considering all the costs related to the data and machine learning life cycle discussed above, the risk exists that these costs are excessive and not in line with the budget if these costs are not managed effectively (Kusi-Sarpong et al., 2021).

### Using third-party service providers

The organisation may need to outsource certain services where the organisation does not have the necessary skills or infrastructure in-house, such as cloud storage or cloud computing (Rawat et al., 2021). Tracking the different service providers, their roles and responsibilities and controls implemented may be difficult, which increases the risk of non-compliance with laws and regulations and could impact controls around business continuity (KPMG, 2018). Furthermore, the organisation could become too reliant on a service provider or the service or solution that it provides (KPMG, 2018; Miller, 2013).

While these risks are at an operational level, technical risks are introduced because of the specific infrastructure used in the machine learning and data life cycle, as well as the characteristics of data used by the retailer for customer demand forecasting of perishable food.

## Risks at a technical level

Risks at a technical level are specific to the data and infrastructure used in the machine learning and data life cycle and can be categorised into three groups, relating to risks affecting the acceptability of and access to data; the ability of components of the machine learning system responsible for the management of data to function effectively and compliance with legal and regulatory requirements.

### Quality

Without quality data being used as input, the predictions generated by the machine learning system will not be accurate, leading to inappropriate decision-making by the retailer. Quality risks arise from invalid data; data not reflecting the true reality; errors in the capturing of data; inconsistencies in the formatting or attributes of data; not all relevant details being included in data; missing values or incomplete datasets and outdated data (DAMA International, 2017; KPMG, 2018; McGraw et al., 2020; Miller, 2013; Waterman & Bruening, 2014).

### Bias

Bias occurs when there are unreasoned judgements or prejudices present in a dataset and can statistically be represented as deviations from expected values (DAMA International, 2017). Using unstructured data and domains exacerbates the risk of bias. This is especially true when there are no standards available to understand the data, as is the case when using machine learning (Hurwitz et al., 2015). When looking at the data life cycle, bias can be introduced at different levels, for example, the collection of data to satisfy a predefined conclusion, non-statistical collection, the use of historic demographical data which lacks diversity and a lack of representation in collected data (Alpaydin, 2020; DAMA International, 2017; McGraw et al., 2020; Oosthuizen, 2021). Bias in training and testing data leads to the model including similar biases in its predictions, which leads to inaccurate and misleading predictions (Lotfian et al., 2021).

### Availability

While retailers have vast amounts of historical sales data available, this data may not provide the desired patterns as the retail market is constantly changing (Miller, 2013; Zhu et al., 2021). Furthermore, there is a risk that non-demand or contextual data (such as economic, regional, social media and preferential data) are not sufficiently available, which could impact the accuracy of customer demand predictions (Miller, 2013).

Another consideration with regard to available data is the amount of data required by the machine learning system to build an accurate and trustworthy model (Alpaydin, 2020; Theobald, 2017). If too little data is available to train the system, there is a risk of underfitting, as the model is incapable of capturing the patterns and relationships between different variables. When too much data are used, on the other hand, there is a risk of overfitting because of the model being too complex. This causes the machine to make accurate predictions on training data, but then not be able to generalise on test and production data (Alpaydin, 2020; Theobald, 2017).

### Accessibility

As a result of the high volumes of complex data used in machine learning systems, storage facilities may be unable to handle the volume of data and provide access to the data as and when it is required (Alpaydin, 2020; Kusi-Sarpong et al., 2021).

In addition, when a data storage facility is unavailable or a malfunction occurs, the system would not have timely access to the data required to make predictions. This may delay inventory management decision-making and lead to the decay of perishable food. Furthermore, the risk exists that the entity has no alternative storage facility in place to ensure the continued and uninterrupted operation of the system (IoDSA, 2016; KPMG, 2018).

### Interoperability

Machine learning systems are dependent on enabling technologies, storage and management technologies, data analytics and visualisation tools and other relevant solutions (Angarita-Zapata et al., 2021). Furthermore, the large volumes of data generated and collected from various sources are presented in different formats, structures and programming languages (Kusi-Sarpong et al., 2021; Miller, 2013). The main areas of concern with regard to data in machine learning systems are technical interoperability, where infrastructure components are unable to communicate with one another; semantic interoperability, where components are unable to comprehend data transferred from various sources and syntactic interoperability, where transferred data is not understood because of a lack of consistency between format, structures and programming languages (Maciel et al., 2017; Van Wyk & Rudman, 2019).

### Integration

By using different types of data from various sources, there is a risk that the data cannot be integrated and used together without significant data integration mechanisms (Gartner, 2023). Using data integration mechanisms introduces additional risks, one of the most significant being the loss of data lineage. Without the preservation of the data lineage, the organisation would no longer be able to pinpoint where the data comes from, how it has changed, and how it is used by the organisation. This would lead to the organisation not being able to prove that the predictions from the data are valid and representative (DAMA International, 2017). Furthermore, where functions or components are outsourced to service providers, there is a risk of ineffective integration between the components of the service provider and that of the organisation (Géczy, 2014).

### Scalability

When we consider the increase in data volumes from most sources, such as years' worth of daily historical sales or weather data, certain components pose significant risks of decreased performance if they are not scalable, which may slow down the entire system (Chen & Zhang, 2014). These components include those responsible for data storage, processing and transmission, as well as machine learning algorithms (Arunachalam et al., 2018; Chen & Zhang, 2014; Géczy, 2014; Kusi-Sarpong et al., 2021; Van Wyk & Rudman, 2019).

### Security

As the retailer places trust in the decisions made by their machine learning system, it is important that the data used to make these decisions are not subjected to unauthorised access or changes that could lead to inaccurate decisions being made (DAMA International, 2017; Gartner, 2023). The risks that threaten the security of data in a machine learning system can be categorised as follows:

- **Risk of non-compliance with laws and regulations:** Non-compliance with privacy laws and regulations or not identifying all the regulatory requirements that must be met may lead to reputational damage, legal implications or ethical repercussions for the organisation (De Bruyn, 2014; KPMG, 2018).
- **Inadequate governance of security requirements:** Security risks may be aggravated by inadequate governance practices implemented by management, such as a lack of adequately documented and enforced data security policies, insufficient monitoring and management of security and ineffective security policies and controls implemented by service providers (Hurwitz et al., 2015; KPMG, 2018; Kshetri, 2014; Miller, 2013; Van Wyk & Rudman, 2019).
- **Unauthorised access and security breaches from outside the organisation:** Security breaches from outside the organisation occur because of unauthorised access to the data of an organisation without authentication of the user and through malicious attacks (DAMA International, 2017). Apart from traditional IT security challenges, such as hacking, malware, denial of service and phishing, there are attacks that are specific to the data used in machine learning systems, for example, poisoning attacks focusing on changing training data to make predictions inaccurate, and adversarial examples focused on testing or inference data where small changes in the data lead to mistakes in the output of the machine learning model (Liu et al., 2018).
- **Unauthorised access and security breaches from inside the organisation:** Ineffective access and user rights within the organisation lead to users that should not have access to certain data obtaining access and using, changing or sharing that data without the necessary authority (KPMG, 2018).

Security breaches also lead to risks to the privacy of sensitive information and PII used in the machine learning model.

### Privacy

Using customer information containing PII and sensitive company information increases the need for the protection of the privacy of this data, as well as compliance with laws such as the GDPR and the POPI Act (De Bruyn, 2014; McGraw et al., 2020), because of the following risks:

- **Inadequate governance of privacy requirements:** Privacy risks may be aggravated by inadequate governance practices implemented by management, as explained under security risks above (Hurwitz et al., 2015; KPMG, 2018; Miller, 2013; Van Wyk & Rudman, 2019).
- **Violating the rights of data subjects:** The violation of data subject rights can take many forms, such as using data without the subject's consent, using data for a purpose for which the subject did not consent to, or not removing the data subject's information from a dataset (Bardi et al., 2014; Kshetri, 2014; Lotfian et al., 2021; McGraw et al., 2020; NDPA, 2018). Furthermore, it may be difficult to provide explanations of the algorithm or the business may not be willing to reveal business secrets, resulting in the data subject not understanding how the algorithm works and how their data will be used (Jensen, 2013; NDPA, 2018). This not only leads to the risk of non-compliance with laws and regulations (GDPR, 2016; POPI Act, 2013) but also increases the reputational risks to the organisation and possible legal actions and loss of trust that could ensue from the individual data objects (Lotfian et al., 2021).
- **Risk to the integrity and authenticity of data:** As machine learning systems use large volumes of data from various sources, there is a risk that the data is not accurate. Data obtained from some sources may be manipulated, falsified or outdated (Jensen, 2013). The large volumes and wide variety of data combined with the complex sources make it difficult to verify the integrity and authenticity of data, in addition to removing false and malicious data from the dataset (Sun et al., 2021).
- **Security breaches:** When security breaches occur, there is a risk that personal or sensitive data is accessed, changed or shared by unauthorised users, which may lead to reputational damage and legal liability (Hurwitz et al., 2015). In addition to the security breaches discussed in

the section titled Security, other security breaches present a privacy risk. These include membership inference attacks, where the attacker tries to find out if a specific person is included in training data (Xue et al., 2020); and re-identification risk, where new, previously unknown connections are formed during the transformation and integration of data (Bardi et al., 2014).

### Ownership

In most cases, where data are created by the organisation and that contain PII, it is considered that the organisation has shared ownership together with the data subject (Durn, 2021). Because of the uncertainty about who the data owner is, it creates the risk of unauthorised or inappropriate access, data sharing, changes being made or inappropriate use of the data (Durn, 2021; KPMG, 2018).

## Conclusion

When a machine learning system is introduced into a retail environment, the complexities of the system, as well as the types and inherent characteristics of data used by the machine, lead to data-related risks. These risks include high-level risks at a strategic level because of inadequate or ineffective governance and management of data. The next level of risks is risks at an operational level, which arise from inefficiencies in internal processes and systems that manage the resources of an organisation. Lastly, risks at a technical level arise which are specific to the data and infrastructure used in the internal processes and systems used for customer demand forecasting. These risks should be identified and evaluated to determine the likelihood of occurrence, as well as the impact on the retailer if they were to occur. The challenge is that most retailers do not understand the technology and are therefore unable to identify all the risks. Furthermore, limited guidance is available to those charged with governance in a retail environment to help them identify and understand the risks. The current literature on data-related risks associated with the use of machine learning is not comprehensive as it focuses on the use of machine learning in a general sense and is not specifically contextualised within the customer demand forecasting or retail environment or the data used for this specific purpose.

In order to identify all the risks, governance frameworks are valuable tools which can be used to identify all the risks. The objectives of COBIT-2019 and the knowledge areas of the DAMA DMBOKv2 were used to benchmark and identify significant data-related risks when using machine learning for customer demand forecasting of perishable food by retailers. These risks are listed in Figure 3 and can be used by internal or external auditors or those charged with governance and other management functions within an organisation to identify the risks when implementing machine learning. It can be used as a complete list or it can be utilised to rate the risks that require attention while being compliant with two internationally known governance frameworks. Moreover, this study contributes to academic research by using internationally known governance frameworks to ensure completeness and rigour. This ensures that the risks identified are comprehensive and not on an *ad hoc* basis.

Conducting research in an IT field poses a limitation in terms of information that is available at the time of the study. As the IT field is fast-paced, new technologies and advancements may become available in future which was not available to the researcher at the time of the research study. This research study does not cover the futuristic viewpoint of AI and machine learning in retail but only considers information available to the author up to the date of submission of the study. Furthermore, the study does not aim to be a comprehensive technical study of the technology, algorithms or mathematical and statistical workings of the technology and algorithms. Although a broad overview of the workings of the technology, architecture and models is needed to understand and formulate the significant data-related risks and mitigating controls, the study does not provide detailed technical or statistical guidance. Another limitation of this study pertains to the research methodology employed, as the selection of data sources and their interpretation may introduce potential subjectivity on the part of the authors.

Further research into the optimal algorithms and machine learning models for accurate customer demand forecasting may be of value, as this research study did not investigate the mathematical and statistical functioning of the technology. Furthermore, while this research was conducted to identify the significant data-related risks through a structured literature review and the use of governance frameworks, a future area of research includes the empirical testing of the list of risks identified within a retail environment where machine learning is used for customer demand forecasting of perishable foods. This would ascertain the value that the list adds to those charged with governance and assurance. Further research is also required to formulate mitigating controls to address the identified significant risks.

## Acknowledgements

This article is partially based on the author, L.P.’s Master’s dissertation entitled, ‘Data-related risks and mitigating controls for the use of machine learning in retail customer demand forecasting’, towards the degree of Master of Commerce, in the School of Accountancy, Stellenbosch University, South Africa, with supervisors R.J. Rudman and G.F. Nel, received March 2024.

### Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

### Authors’ contributions

L.P. conducted the research and documented the findings. Both L.P. and R.J.R. contributed to the final version of the manuscript. R.J.R. supervised the project.

## Funding information

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

## Data availability

The data that support the findings of this study are available from the corresponding author, R.J.R. upon reasonable request.

## Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency or that of the publisher. The authors are responsible for this article's results, findings and content.

# References

Abraham, R., Schneider, J., & Vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, *49*, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

Alpaydin, E. (2020). *Introduction to machine learning* (4th ed.). The MIT Press.

Angarita-Zapata, J.S., Alonso-Vicario, A., Masegosa, A.D., & Legarda, J. (2021). A taxonomy of food supply chain problems from a computational intelligence perspective. *Sensors*, *21*(20), 6910. https://doi.org/10.3390/s21206910

Arunachalam, D., Kumar, N., & Kawalek, J.P. (2018). Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice. *Transportation Research Part E: Logistics and Transportation Review*, *114*, 416–436. https://doi.org/10.1016/j.tre.2017.04.001

Bardi, M., Xianwei, Z., Shuai, L.I., & Fuhong, L. (2014). Big data security and privacy: A review. *China Communications Supplement*, *11*(14), 135–145. https://doi.org/10.1109/CC.2014.7085614

Brous, P., Janssen, M., & Vilminko-Heikkinen, R. (2016). Coordinating decision-making in data management activities: A systematic review of data governance principles. *International Federation for Information Processing*, *9820 LNCS*, 115–125. https://doi.org/10.1007/978-3-319-44421-5_9

Chen, C.L., & Zhang, C. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, *275*, 314–347. https://doi.org/10.1016/j.ins.2014.01.015

DAMA International. (2017). *DAMA-DMBOK: Data management body of knowledge* (2nd ed.). Technics Publications, LLC.

De Bruyn, M. (2014). The Protection of Personal Information (POPI) Act – Impact on South Africa. *The International Business & Economics Research Journal*, *13*(6), 1315–1340. https://doi.org/10.19030/iber.v13i6.8922

De Haes, S., Van Grembergen, W., Joshi, A., & Huygh, T. (2020). COBIT as a framework for enterprise governance of IT. In *Enterprise governance of information technology: Achieving alignment and value in digital organizations* (3rd ed., pp. 125–162). Springer Nature Switzerland AG.

De Moraes, C.C., De Oliveira Costa, F.H., Pereira, C.R., Da Silva, A.L., & Delai, I. (2020). Retail food waste: Mapping causes and reduction practices. *Journal of Cleaner Production*, *256*, 120124. https://doi.org/10.1016/j.jclepro.2020.120124

Domingo, J. (2023). *Why and how retailers can stop overstocking*. Inventoro. Retrieved from https://inventoro.com/retail-overstocking/

Durn, E. (2021). *Data ownership: What's in a name?* Deloitte Blog: Risk. Retrieved from https://www2.deloitte.com/uk/en/blog/risk-powers-performance/2021/data-ownership-whats-in-a-name.html

Edmead, M.T. (2020). *Using COBIT 2019 to plan and execute an organization's transformation strategy*. ISACA. Retrieved from https://www.isaca.org/resources/news-and-trends/industry-news/2020/using-cobit-2019-to-plan-and-execute-an-organization-transformation-strategy

Gartner. (2023). *Information Technology glossary*. Retrieved from https://www.gartner.com/en/chat/information-technology/glossary

GDPR, Regulation (EU) 2016/679 (2016). General Data Protection Regulation. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Géczy, P. (2014). Big data characteristics. *The Macrotheme Review*, *3*(6). Retrieved from https://macrotheme.com/yahoo_site_admin/assets/docs/8MR36Pe.97110828.pdf

Guo, Z.X., Wong, W.K., & Li, M. (2013). A multivariate intelligent decision-making model for retail sales forecasting. *Decision Support Systems*, *55*(1), 247–255. https://doi.org/10.1016/j.dss.2013.01.026

Hendrawan, F.R., Kusumasari, T.F., & Fauzi, R. (2022). Analysis of design implementation guidelines for data governance management based on DAMA-DMBOKv2. In *7th international conference on informatics and computing, 08–09 December 2022.* IEEE.

Hurwitz, J., Kaufman, M., & Bowles, A. (2015). *Cognitive computing and big data analytics* [Book]. John Wiley & Sons. Retrieved from https://web-s-ebscohost-com.ez.sun.ac.za/ehost/ebookviewer/ebook/bmxlYmtfXzk3ODEzMF9fQU41?sid=f9fe109c-aa9c-4573-910843749ed35919@redis&vid=0& format=EB&rid=1

IoDSA (Institute of Directors Southern Africa). (2016). *King IV Report on corporate governance for South Africa*. Retrieved from https://cdn.ymaws.com/www.iodsa.co.za/resource/collection/684B68A7-B768-465C-8214-E3A007F15A5A/IoDSA_King_IV_Report_-_WebVersion.pdf

ISACA. (2018). *COBIT® 2019 Framework: Governance and management objectives*. Retrieved from https://store.isaca.org/s/store#/store/browse/detail/a2S4w000004Ko9ZEAS

Jensen, M. (2013). Challenges of privacy protection in big data analytics. In *IEEE International Congress on Big Data, 27 June 2013 – 02 July 2013* (pp. 235–238). IEEE.

Kaleva, H., & Småros, J. (2023). *The complete guide to machine learning in retail demand forecasting*. RELEX Solutions. Retrieved from https://www.relexsolutions.com/resources/machine-learning-in-retail-demand-forecasting/

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of Artificial Intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

KPMG. (2018). *AI risk and controls matrix*. Retrieved from https://assets.kpmg.com/content/dam/kpmg/uk/pdf/2018/09/ai-risk-and-controls-matrix.pdf

Kshetri, N. (2014). Big data's impact on privacy, security and consumer welfare. *Telecommunications Policy*, *38*(11), 1134–1145. https://doi.org/10.1016/j.telpol.2014.10.002

Kumar, A., Shankar, R., & Aljohani, N.R. (2020). A big data driven framework for demand-driven forecasting with effects of marketing-mix variables. *Industrial Marketing Management*, *90*, 493–507. https://doi.org/10.1016/j.indmarman.2019.05.003

Kusi-Sarpong, S., Orji, I.J., Gupta, H., & Kunc, M. (2021). Risks associated with the implementation of big data analytics in sustainable supply chains. *Omega*, *105*, 102502. https://doi.org/10.1016/j.omega.2021.102502

Laato, S., Birkstedt, T., Mantymaki, M., Minkkinen, M., & Mikkonen, T. (2022). AI governance in the system development life cycle: Insights on responsible machine learning engineering. In *Proceedings – 1st international conference on AI engineering – Software engineering for AI, CAIN 2022, 16–17 May 2022* (pp. 113–123). IEEE.

Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V.C.M. (2018). A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE Access*, *6*, 12103–12117. https://doi.org/10.1109/ACCESS.2018.2805680

Lotfian, M., Ingensand, J., & Brovelli, M.A. (2021). The partnership of citizen science and machine learning: Benefits, risks and future challenges for engagement, data collection and data quality. *Sustainability*, *13*(14), 8087. https://doi.org/10.3390/su13148087

Maciel, R.S.P., David, J.M.N., Claro, D.B., & Braga, R. (2017). Full interoperability: Challenges and opportunities for future information systems. In Brazilian Computer Society, *Grand research challenges in information systems in Brazil 2016-2026* (pp. 107–118). Brazilian Computer Society.

McGraw, G., Figueroa, H., Shepardson, V., & Bonett, R. (2020). *An architectural risk analysis of machine learning systems: Toward more secure machine learning*. Berryville Institute of Machine Learning. Retrieved from https://www.garymcgraw.com/wp-content/uploads/2020/02/BIML-ARA.pdf

Miller, H.E. (2013). Big data in cloud computing: A taxonomy of risks. *Information Research*, *18*(1). Retrieved from http://InformationR.net/ir/18-1/paper571.html

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning* (2nd ed.). The MIT Press.

NDPA. (2018). *Artificial intelligence and privacy*. Retrieved from https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf

Oosthuizen, K. (2021). *Artificial Intelligence in retail: The AI-enabled value chain.* Stellenbosch University. Retrieved from http://hdl.handle.net/10019.1/124902

POPI Act, Pub. L. No. 4 (2013). Retrieved from https://www.gov.za/sites/default/files/gcis_document/201409/3706726-11act4of2013protectionofpersonalinforcorrect.pdf

Rawat, D.B., Doku, R., & Garuba, M. (2021). Cybersecurity in big data era: From securing big data to data-driven security. *IEEE Transactions on Services Computing*, *14*(6), 2055–2072. https://doi.org/10.1109/TSC.2019.2907247

Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management*, *38*(1), 187–195. https://doi.org/10.1016/j.ijinfomgt.2017.07.008

Sahd, L.-M. (2016). Significant risks relating to mobile technology. *Journal of Economic and Financial Sciences*, *9*(1), 291–309. https://doi.org/10.4102/jef.v9i1.43

Sankaran, A. (2021). *Finding the data: How to avoid AI and analytics project failures*. Forbes. Retrieved from https://www.forbes.com/sites/forbesbusinesscouncil/2021/10/15/finding-the-data-how-to-avoid-ai-and-analytics-project-failures/?sh=588c8a3f5a66

Sharma, R., Kamble, S.S., Gunasekaran, A., Kumar, V., & Kumar, A. (2020). A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. *Computers and Operations Research*, *119*, 104926. https://doi.org/10.1016/j.cor.2020.104926

Sun, L., Zhang, H., & Fang, C. (2021). Data security governance in the era of big data: Status, challenges, and prospects. *Data Science and Management*, *2*, 41–44. https://doi.org/10.1016/j.dsm.2021.06.001

Sylvester, A., Tate, M., & Johnstone, D. (2013). Beyond synthesis: Re-presenting heterogeneous research literature. *Behaviour and Information Technology*, *32*(12), 1199–1215. https://doi.org/10.1080/0144929X.2011.624633

Tarallo, E., Akabane, G.K., Shimabukuro, C.I., Mello, J., & Amancio, D. (2019). Machine learning in predicting demand for fast-moving consumer goods: An exploratory research. *IFAC-PapersOnLine*, *52*(13), 737–742. https://doi.org/10.1016/j.ifacol.2019.11.203

Theobald, O. (2017). *Machine learning for absolute beginners* (2nd ed.). Scatterplot Press.

Tsoumakas, G. (2019). A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review*, *52*, 441–447. https://doi.org/10.1007/s10462-018-9637-z

United Nations. (2022). *The sustainable development goals report*. Retrieved from https://unstats.un.org/sdgs/report/2022/The-Sustainable-Development-Goals-Report-2022.pdf

Vandeput, N. (2021). *How to: Machine learning-driven demand forecasting*. Towards Data Science. Retrieved from https://towardsdatascience.com/how-to-machine-learning-driven-demand-forecasting-5d2fba237c19

Van Wyk, J. (2018). Cognitive computing governance and risk management. *Southern African Journal of Accountability and Auditing Research*, *20*(1), 71–87. Retrieved from https://hdl.handle.net/10520/EJC-11fd67a2a5

Van Wyk, J., & Rudman, R. (2019). COBIT 5 compliance: Best practices cognitive computing risk assessment and control checklist. *Meditari Accountancy Research*, *27*(5), 761–788. https://doi.org/10.1108/MEDAR-04-2018-0325

Waterman, K.K., & Bruening, P.J. (2014). Big data analytics: Risks and responsibilities. *International Data Privacy Law*, *4*(2), 89–95. https://doi.org/10.1093/idpl/ipu002

Xue, M., Yuan, C., Wu, H., Zhang, Y., & Liu, W. (2020). Machine learning security: Threats, countermeasures, and evaluations. *IEEE Access*, *8*, 74720–74742. https://doi.org/10.1109/ACCESS.2020.2987435

Yang, L., Li, J., Elisa, N., Prickett, T., & Chao, F. (2019). Towards big data governance in cybersecurity. *Data-Enabled Discovery and Applications*, *3*(1), 10. https://doi.org/10.1007/s41688-019-0034-9

Zhu, X., Ninh, A., Zhao, H., & Liu, Z. (2021). Demand forecasting with supply-chain information and machine learning: Evidence in the pharmaceutical industry. *Production and Operations Management*, *30*(9), 3231–3252. https://doi.org/10.1111/poms.13426