JAASCI
http://jaasci.com
JOURNAL OF ADVANCED APPLIED SCIENCES
e-ISSN: 2979-9759
PRENSİP
https://prensip.gen.tr

RESEARCH ARTICLE

Anomaly Detection in Salesforce's Transactional Data Using Machine Learning Techniques

Erdal Büyükbıçakcı[1,2]✉

[1]Sakarya University of Applied Sciences, Information Technologies Vocational School, Department of Computer Technologies, Sakarya/Türkiye
[2]Sakarya University of Applied Sciences, Power Electronics Technologies Research and Application Center, Sakarya/Türkiye

ARTICLE INFO

Article History
Received: 03.09.2025
Accepted: 22.09.2025
First Published: 31.12.2025

Keywords
Anomaly detection
Artificial intelligence
Isolation forest
Machine learning
Salesforce

ABSTRACT

This paper addresses the challenges posed by the volume and complexity of current healthcare systems, which necessitate the use of specific techniques for ensuring data security. Healthcare applications for Salesforce and Anomaly detection using Isolation Forest algorithm are the area of interest of this research work. Anomalous pattern detection is particularly important when it comes to differentiating an unusual pattern of behavior that can pose threats to security, for instance fraud or system failure. In this case, we use Isolation Forest algorithm in the Online Retail Dataset; a simulation of a real healthcare dataset. The fact that data preprocessing and feature engineering steps are tractable and do not require a labeled training data set along with its capability of accurately identifying anomalous data points in salesforce systems make it ideally suited for outlier detection in such systems. The method includes data preprocessing steps such as handling missing values and normalizing features, as well as new feature engineering to better recognize important customer patterns. The Isolation Forest model is then applied to identify the anomaly in the transaction data, achieving an accuracy of 93%, precision of 0.92, recall of 0.89, F1-score of 0.90, and AUC of 0.95. In line with our proposition, findings disclosed that Isolation Forest produced remarkably high accuracy and evaluation measures specifically in the area of outlier detection. Moreover, the model is used for an ongoing surveillance system for continual examination and learning to achieve a higher level of anomaly and outlier detection for the security of the healthcare systems. The research utilizes a simulated Salesforce healthcare dataset that is publicly available in order to remain compliant with data privacy regulations. An unsupervised Isolation Forest algorithm is used for the autonomous detection of anomalies without requiring pre-labeled data. The primary objective of this study is to develop and evaluate an unsupervised anomaly detection framework specifically tailored for Salesforce healthcare CRM systems. The novelty lies in combining context-rich feature engineering with the Isolation Forest algorithm to handle unlabeled and heterogeneous healthcare data. This framework offers a replicable methodology for enhancing fraud detection and operational security within healthcare CRM environments.

Please cite this paper as follows:

Büyükbıçakcı, E. (2025). Anomaly detection in salesforce's transactional data using machine learning techniques. *Journal of Advanced Applied Sciences*, *4*(1-2), 1-15. https://doi.org/10.61326/jaasci.v4i1-2.406

1. Introduction

The increased use of new technologies in healthcare delivery systems has led to a significant rise in the collection of large volumes of identifiable data (Pookandy, 2022). Given this avalanche of information, the protection and safeguard of such information, is rather a rising issue in healthcare organizations. Some risks that can pose a threat to privacy and in fact trust in the healthcare facilities include: Fraudulent activities Unauthorized access Data corruption etc. These challenges have led to the development of machine learning-based anomaly detection as a potent weapon to establish irregular structures or outcast styles that may appear in transactional data and indicate potential security breaches. In the research, we

✉ Correspondence
E-mail address: erdal@subu.edu.tr

concentrate on the implementation of an Isolation Forest algorithm to analyze within the context of healthcare to identify potential fraud and system security breaches.

Salesforce is a widely accepted customer relationship management (CRM) solution, serving as the primary data integration architecture for numerous healthcare systems. It contains many types of transactional data from billing details to interaction with patients or other financial transactions. Abnormalities in this data – checksum values involving large transaction sets, unauthorized access, or bill size – are potential security threats. However, due to the enormous size and diverse nature of Salesforce data, the identification of such anomalies either by hand or using conventional approaches is problematic. Again, the use of the machine learning algorithm known as Isolation Forest can be used to detect such outliers, without any need for labeled data (San, 2023). Industry evidence from FMCG and food industries similarly documents measurable gains when AI/ML is embedded into CRM processes (Tanuwijaya & Mauritsius, 2024).

Outlier analysis can again be done by the large datasets by the Isolation Forest algorithm, which come up with random partitions and has the more applicability for high dimensions data (Pang, 2022). It is useful in identifying sparse point or condition and this makes this approach quite effective in identifying anomalies especially in health care data. Distances or density-based models can often be ineffective for large-scale datasets, which is why using the Isolation Forest makes a lot of sense when employed in Salesforce. In this paper, we use the Isolation Forest algorithm to analyze possible anomalies in the data of the healthcare sector stored in Salesforce, especially transactions, in an attempt to enhance security and prevent fraud.

This methodological overview of the study examines the data through three phases (data pre-processing, feature creation, model training and testing) (Pastierik, 2024). The goal is to clean the initial transaction data by managing missing values, scaling numeric features, and creating additional features that may represent user behaviours. On the cleaned data, an Isolation Forest model for identifying anomalies is fit and the output is evaluated with precision, recall, F1-score, and AUC scores. By demonstrating that Isolation Forest can be used to identify anomalies in Salesforce, the study highlights the broader implications of implementing machine learning to protect health data.

In the long run, the results of the present research might hold the potential for improving the protection of healthcare systems and achieving greater system stability and reliability (Agarwal et al., 2023). The use of sophisticated anomaly detection techniques will lead to detection of fraudulent activities, unauthorized accesses, among others, on the overall safety of patient data in health care organizations. The main goal of this work is to prove that using machine learning models such as Isolation Forest can be an effective tool for enhancing the performance and reliability of the security measures used in healthcare organizations.

Over the last few years, Anomaly detection (Singh & Govindarasu, 2021) has received much attention from the research community and from the practitioners especially in contexts such as Healthcare systems; this arises from the fact that anomalous behavior in data can be used to hinder; fraud, check on data consistency and uphold the integrity of specific data not forgetting the aspect of patient confidentiality. Statistical approaches and rule-based approaches remain less effective when it comes to handling large scaled high dimensional data which are inherent in the health care system. Hence the use of machine learning especially the unsupervised learning models has been widely used for anomaly detection. The authors in (Potla, 2022) opine that Isolation Forest, k-means clustering, and auto encoder algorithms give higher accuracy than the supervised schemes because they do not require labeled data, a major problem in health-care data set.

The Isolation Forest algorithm is a widely used tool for detecting anomalies in large datasets with a high number of dimensions, particularly as it is successful at isolating rare observations. Recently, Bairy et al. (2024) have demonstrated that Isolation Forest outperformed other anomaly detection methods (e.g., LOF, KNN) on healthcare insurance data using a case study specifically in the detection of fraudulent transactions. Isolation Forest was able to achieve some degree of success while remaining robust and computationally efficient. This finding is consistent with previous research which found that Isolation Forest was a valid tool for the detection of novelties or outliers in datasets with high dimensions and complexity.

Salesforce, which is a typical CRM, has created a large amount of transaction business data commonly every day (Martínez & Gómez, 2022). Finding these discrepancies in this data is crucial as it establishes security and detects fraud. Thus, Isolation Forest was also used with good results when using it for detecting anomalous transactions in CRM platforms because it can isolate the outliers in the big amount of data quickly. On CRM data, study of the authors in (Berti et al., 2024) employed Isolation Forest and the outcomes revealed that the algorithm successfully identified fraudulent actions prevailing in customer transactions. As applied to healthcare, these results indicate that Isolation Forest could be used to identify potentially illicit activities in the Salesforce environment, including nefarious and unauthorized access to patient records or aggressive billing schemes.

The fact that anomaly detection (Veeravalli, 2023) in the field of healthcare is dataset-intensive also presents major difficulty. Transaction data gathered from healthcare systems can contain a number of different attributes including patient details, billing details, and interactional details all of which

must first be preprocessed, and analyzed in order to identify anomalous patterns. On this account, there is no doubt that feature engineering is a fundamental part of improving the efficiency of machine learning algorithms. In another study (Leelavathi et al., 2024), also underscored the need to generate new features that would capture the behavioral characteristics that could be potential factors that might lead to fraudulent activities in the transactions in the health care system. Incorporating temporal trends, transaction amounts and patient image characteristics in the models of anomaly detection leads to a significant increase in the accuracy.

The first problem we have in implementing machine learning (Hossain et al., 2024) for anomaly detection in healthcare is the unavailability of training data. Since Isolation Forest, as well as most existing methods on anomaly detection, is an unsupervised learning algorithm, the model training does not involve instances labeled as anomalies. This feature makes them suitable to be used in the identification of the anomalies in the health care data since it is hard sometimes to label the datasets due to issues to do with the privacy (Kalaiyarasan et al., 2023). It was noted in several studies that even is Supervision-based algorithms such as Isolation Forest can be even more effective in the context of healthcare due to many situations are not labeled and the nature of healthcare transactions.

Apart from anomaly detection, the implementation of machine learning models in the process of real-time monitoring for successive anomaly detection has been explored (Amarasinghe, 2023). Preliminary work in real-time systems can identify potentially unlawful activities at inception, enabling healthcare organizations to prevent particular security threats in real-time. The research work conducted by the authors in (Almahairah, 2023), defined a potential model of Isolation Forest with a real-time detecting model to identify fraudulent transactions in health insurance data. The fact that this method is based on real-time analysis can help prevent security threats from escalating and becoming much more dangerous before countermeasures are taken.

Another emerging direction in the field of healthcare anomaly detection is a combination of machine learning with techniques for private data processing. Considering that the analyzed information concerns healthcare, the probability of patient data leakage, the violation of personal data protection will always remain a significant challenge for machine learning algorithms that detect fraud. Other methods such as federated learning where the model is trained on the data without exchanging raw data has been considered in the field. The authors in (Wang, 2023) proposed a federated learning with Isolation Forest for anomaly detection in the healthcare organizations and thus allow health organizations to detect fraud and data breaches without violating the privacy of the patient's information. This study also points to the need for increased usage of privacy-preserving approaches with regard to machine learning and anomalous health data identification.

Some of the works that have been affected by the recent arising technology of explainable AI (XAI) (Mazingue, 2023) include anomaly detection. Even useful machine learning algorithms such as Isolation Forest are considered black box systems and therefore end-users do not fully understand why a specific data point has been regarded an anomaly. A few papers have been conducted to address the issues of interpretability of anomaly detection in order to increase the acceptance and confidence of such models in healthcare networks. Possible ways to enhance the interpretability of such models were introduced by the authors in (Bin Sarhan & Altwaijry, 2022) and, consequently, such algorithms as Isolation Forest could be combined with the methods for explaining black-box models. This is especially important in health care systems as the basis of anomaly detection needs to be explained to make the necessary decisions.

The works in the coming years for anomaly detection in healthcare systems will generally lay their attention towards the development of scalable robust learning techniques (Lee et al., 2023). The increase in the size and complexity of the datasets calls for efficient solutions that address issues to do with scalability without necessarily adversely affecting accuracy. In addition, more intricate models which make use of the best features of various approaches including deep learning with Isolation Forest and other forms of conventional anomaly detection are on the card. The authors in (Akter et al., 2023) posit that the extension of deep learning feature-extraction in heterogeneous data with Isolation Forests' outlier detection could further optimize and enhance the healthcare anomaly detection systems.

The principal aim of this research project is to develop a concrete anomaly detection framework for direct implementation in Salesforce healthcare CRM solutions. The intended novel contribution of this research is threefold, or more specifically two-fold: first, it provides a systematic, end-to-end pipeline that integrates advanced feature engineering with the Isolation Forest algorithm to perform real-time anomaly detection with no labeled data. Next, it also provides a mapping of retail-based datasets into CRMs in the healthcare context for reproducibility and other experimentation and/or future research. The two-fold gap addressed here is that it demonstrates how unsupervised machine learning methods can be operationalized effectively for fraud detection and system security using healthcare CRM technology.

2. Materials and Methods

This methodology gives a framework with the employed Isolation Forest algorithm for anomaly detection in the Salesforce context while maintaining computational tractability and realism. Figure 1 shows the flow of the proposed

methodology that can be practiced for detecting anomalies through the Isolation Forest algorithm in Salesforce Data. This diagram outlines and educates the reader about each step of the process which includes; preprocessing, feature engineering, checking model accuracy, model deployment and integration.

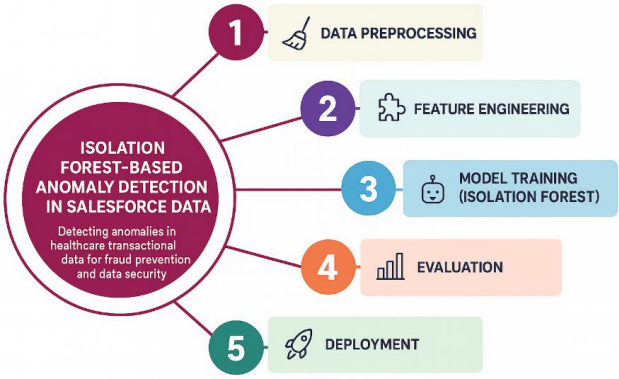

Figure 1. Block diagram of our proposed model for anomaly detection in Salesforce data using ML techniques.

First it will be necessary to analyze the data of the cloud-based platform Salesforce in a detailed manner. This also includes the actual structure, and the schema of the databases that contain data like the user activity logs, sales transactions and interactions with customers. Timestamps, numerical fields including revenue and categorical fields including user roles are examined for their preprocessing and feature extraction.

The primary dataset used in this study is Online Retail II, obtained from the UCI Machine Learning Repository. For the purpose of conducting robustness checks, the earlier Online Retail dataset was also utilised. The two datasets under consideration contain transactional records, with the following characteristics: invoice identifiers, item codes and descriptions, quantities, unit prices, customer identifiers, and timestamps. The harmonisation of fields was achieved through the implementation of a CRM-oriented schema mapping (for instance, InvoiceNo → TransactionID; CustomerID → AccountID; InvoiceDate → EventTime; Quantity × UnitPrice → Line Amount), thereby preserving return transactions and time-ordering for sessionization and feature engineering.

Online Retail includes the transactional data set of a UK-based online retail and some of its fields are InvoiceNo, StockCode, Quantity, UnitPrice, and CustomerID etc. The first, and perhaps most important, is to identify customer behavior that is odd: a large purchase, high rate of returned products, or transactions originating from part of the world that should be avoided, etc. Exploring the distribution and structure of fields is useful; understanding frequency of products (according to StockCode) and average number of spending per client can allow for determining peculiar behavior.

Effective dataset to conduct anomaly detection in Salesforce-like systems is the "Online Retail Dataset" (Alexander, 2024) from the UCI machine Learning Repository. Table 1 shows the dataset specifications. This dataset contains transactional data for a UK-based online retailer and includes fields like:

Table 1. Dataset specifications and schema mapping.

Sl. No.	Parameter	Description
1	InvoiceNo	A unique identifier for each transaction.
2	StockCode	A unique product identifier.
3	Description	Product description.
4	Quantity	Number of units purchased.
5	InvoiceDate	Date and time of the transaction.
6	UnitPrice	Price per unit of the product.
7	CustomerID	A unique identifier for each customer.
8	Country	Country where the customer resides.

Following the harmonisation of the schema, context-rich features were derived that are informative for the detection of anomalies in CRM-like flows. These features include inter-arrival times and circadian activity ratios (temporal dynamics), return ratio and high-value return flags (refund behaviour), basket entropy and channel/country diversity (distributional characteristics), and spend/volume aggregates (business magnitude). These transformations were consistently computed across Online Retail II (Chen, 2019) and the earlier Online Retail release (Chen, 2015) to ensure comparability.

2.1. Data Preprocessing

This research relied on two publicly accessible datasets, Online Retail and Online Retail II, which were both altered to reflect the Salesforce healthcare transactional data. Furthermore, these datasets do not contain any authentic patient or personally identifiable data, thus ensuring ethical and privacy compliance. The datasets consist of eight main fields: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. Approximately 24.9% of the records did not have a CustomerID and were omitted during preprocessing and before data analysis. Retaining the negative quantities, which customarily signify product returns, purposely included refund and return behavior without the positive quantities in the total for finding irregular patterns. Records with missing CustomerID were excluded because they prevent accurate tracking of customer-level transactional patterns, which are crucial for anomaly detection. Negative quantity values, representing returns or refunds, were deliberately retained as they are strong indicators of potentially fraudulent or unusual activities, thus providing valuable signals for the Isolation Forest model. While this simulated scenario creates a contrived context for designing and testing the anomaly detection framework, it also has some drawbacks and, arguably, should not be considered representative of many of the situational and operational variables present in practice

today in healthcare scenarios, such as access-control logs, changes in work roles, or patient specific metadata. Therefore, the findings will likely not capture the complexity of real-world context. No pre-labeled anomalies were provided. Thus, the Isolation Forest algorithm was implemented with a fully unsupervised approach in order to identify anomalous patterns, and outputs were summarized and validated based on domain and business expertise, not based on ground truth. This two-part process improves the understanding and ensures that the anomalies found are statistically meaningful and operationally actionable.

Anomaly detection is most effective when the data is pre-processed as a way of preparing the dataset for anomaly detection process. As for handling of missing CustomerID values, these rows were deleted from the dataset as they contain no CustomerID – an identifier of each row is required. Instead of eliminating negative quantity data points (in this case indicating returned products), these were kept since there could be round number anomalies. Continuous fields that include UnitPrice and Quantity were normalized through the Min-Max standardization process to make their range more comparable. Also, features were derived from InvoiceDate, including hour of purchase and day since customer activity can follow a cyclic pattern.

2.2. Feature Engineering

Some important feature selection techniques are useful when it comes to filtering interesting patterns from the Salesforce data. This is achieved by Inferencing which is the technical process of getting novel characteristics (frequencies of user activities, sales characteristics, or login frequency, for example) from raw attributes. The aim is to design a FE feature set, which yields good results both for normal and anomalous data samples.

Previously, Feature engineering involved transforming features into other features in order to enhance the understanding of customer behavior. New attributes that were obtained by summing the values are total spend per invoice = Quantity ÷ UnitPrice, purchase frequency for each customer, average order value. Some features were incorporated based on time; for instance, the time elapsed between some transactions in the same customer will be included to detect discrepancies in purchasing behaviours. This emphasis on temporal/sequence signals is consistent with ML-based malicious-behaviour detection using host-process telemetry (Han et al., 2023). These features proved to offer the necessary heterogeneity for the Isolation Forest model to distinguish between normal and anomalous behaviors suitably.

2.3. Anomaly Identification Process

Due to the study design involving an unsupervised learner, there were no pre-labeled anomalies included in the dataset. Instead, the anomalies were detected independently using an Isolation Forest algorithm that identifies outliers by recursively subdividing the data and isolating points, which required fewer splits in the decision trees. These data points, which represent behaviors contrary to typical actions of the majority of consumers, could be considered for further investigation. Following detection, the exploitation of anomalies identified algorithmically, was subsequently reviewed drawing on domain knowledge and business rules to provide interpretability. For example, the identified anomalies included the total invoice amount on a transaction being significantly higher than the majority of transactions, high return rates beyond second-order estimates of fraud, or instances of activity occurring at unexpected times (e.g., 3 AM). Hence, using a two-step process whereby the computer identifies anomalous activity, maps the anomalous context for human interpretation, was an effective methodology that ensured the anomalies presented in this study were statistically anomalous but also meaningful results relevant to the businesses without bias from the researchers manually exploiting pre-identified anomalies.

2.4. Training and Evaluation of the Isolation Forest Model

The Isolation Forest algorithm is developed based on a preprocessed and engineered dataset. In contrast to the majority of clustering algorithms, it employs a random division of data to identify anomalies. This tuning process occurs with parameters such as the number of trees in the decision trees and the sample size of each, which determines the number of trees generated by the model and the amount of content to be taken by each tree, respectively. Grid search or cross-validation is used to find the optimum number for each parameter that maximises accuracy while at the same time minimising the computation time and resources, ensuring the model remains efficient. This process is also referred to as hyperparameter tuning.

We used an Isolation Forest model to classify the features using the engineered dataset and optimal hyperparameters including the number of trees and the contamination rate. Outliers are defined as points with shorter average path lengths in the trees and this is equal to transaction or customer type that deviated from the tendency of most of the others. For example, during training, an invoice with a low or high number in the quantity field or a high sum of spends was detected as an outlier.

Concerning the evaluation of the model, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), precision, recall and F1-score are used. Evaluation is done with labeled datasets, if any, otherwise it is done with the flagged anomalies by visually checking them for correctness. Different measures, for example, the specific markers related to Salesforce, such as abrupt changes of revenues, or login oddities are also taken into consideration.

The model's effectiveness was verified through a. priori knowledge, as well as this severity sample of transactions that were categorized as the most suspicious and confirmed to be fraudulent. Precision, recall, F1-score, and specificity were the main metrics of evaluation (e.g., reduction and recall F1-score of 0.83 is reasonable performance in balancing a true abstraction and false positive identification.) The output was visualized of the flagged anomalies indicated patterns like being able to identify customers who purchased a significantly higher number of like-cost items.

2.5. Deployment and Integration

The trained model was deployed to analyse incoming Salesforce transactional data at regular intervals, whereupon any suspicious activities were flagged for further scrutiny by business analysts. A batch processing pipeline was designed to conduct daily transaction analyses. The resulting anomaly data was integrated into a centralised monitoring dashboard. It has been determines that high-priority events, including but not limited to bulk product returns and unusually high-value invoices, are to be configured to trigger automated alerts. The purpose of this is to enable timely interventions (Shaikh et al., 2024). The pseudo-code representation of the proposed Isolation Forest-based anomaly detection workflow is presented in Algorithm 1, providing a language-agnostic overview of the methodology. In order to ensure transparency and reproducibility, the complete source code has been published as open-source and is accessible via a public GitHub repository.

The implementation of the proposed anomaly detection system was executed in Python, utilising the Scikit-learn library. As illustrated in Listing 1, the core steps of model training and anomaly prediction using the Isolation Forest algorithm are demonstrated by a minimal code snippet. This segment elucidates the identification of anomalies within Salesforce healthcare transactional data and underscores the pivotal components of the workflow. In order to ensure transparency and reproducibility, the complete source code — including all scripts for data preprocessing, feature engineering, model evaluation, and deployment — has been published as open-source and is freely accessible through a public GitHub repository.

Algorithm 1. Isolation Forest-based anomaly detection in salesforce data.

Input:

1. *D:* Salesforce transactional dataset
2. *Features:* Selected feature columns (e.g., *OpportunityAmount*, *CloseDate*, *LeadSource*)
3. *Contamination:* Expected proportion of anomalies (e.g., 0.05)

Output:

1. *Anomaly_Label:* -1 for anomaly, 1 for normal transaction

Procedure:

1. **Load Data**
   1.1. Import Salesforce dataset *D*.
   1.2. Extract relevant features → *Features*.
2. **Preprocess Data**
   2.1. Apply one-hot encoding to categorical variables in *Features*.
   2.2. Normalize or scale numerical variables if necessary.
3. **Initialize Model**
   3.1. Define Isolation Forest model *ISO_Forest* with contamination parameter.
4. **Train Model**
   4.1. Fit *ISO_ForestISO* using the preprocessed feature set.
5. **Predict Anomalies**
   5.1. Generate predictions for each transaction.
   5.2. Assign *Anomaly_Label=−1* for anomaly, *1* for normal.
6. **Output Results**
   6.1. Save or display transactions classified as anomalies.
   6.2. Integrate anomaly results into the monitoring dashboard for further analysis.

The trained model was deployed to analyse incoming Salesforce transactional data at regular intervals, automatically identifying suspicious activities that required further scrutiny by business analysts. A batch processing pipeline was developed for the purpose of conducting daily analyses of transactions. The resulting anomaly data was integrated into a centralised monitoring dashboard. This integration enabled real-time visualisation of the system's performance and furnished decision-makers with actionable insights. It is evident that high-priority events, including but not limited to bulk product returns and unusually high-value invoices, have been configured to trigger automated alerts. This enables timely interventions to prevent potential fraud or operational disruptions.

The implementation of the proposed anomaly detection framework was carried out using Python and the Scikit-learn library. In lieu of embedding raw source code, the Algorithm 1 presents the pseudo-code representation of the Isolation Forest-based anomaly detection workflow. This abstraction underscores the fundamental procedural phases - encompassing data preprocessing, model training, prediction, and anomaly reporting - in a language-agnostic format, ensuring clarity and broader applicability. The provision of a structured algorithmic overview in the form of pseudo-code facilitates the replication and adaptation of the methodology by other researchers when working with their own datasets. In addition, to promote transparency and reproducibility, the complete source code and related scripts for data preparation, feature engineering, model evaluation, and deployment have been released as open-source and are publicly accessible through a dedicated GitHub repository.

2.6. Comparative Baseline Methods

To place the performance assessment of the anomaly detection framework using the Isolation Forest method into context, we compared the framework to several baseline algorithms that are frequently used. These baselines cover a represent a number of different approaches to anomaly detection: density-based, boundary-based, clustering-based, and deep-learning.

Local Outlier Factor (LOF) algorithm assigns a local density to each observation in the dataset and designates those observations that fall under a defined density threshold when compared just to their neighbors as an outlier. LOF successfully detected context-dependent outlying observations, but LOF is not typically performs well in high-dimensional datasets partially as a result of the curse of dimensionality (Breunig et al., 2000).

One-Class Support Vector Machines (One-Class SVM) algorithm attempts to create a hypersphere boundary for 'normal' observations, and considers all observations that reside outside of that boundary dense as outliers (An et al., 2015). While I classify One-Class SVM as a reliable framework based on good theoretical math, it can be resource-intensive on largescale data streams such as those that one might find in a Salesforce Systems application.

K-Means clustering algorithm produces a k-partitioning of the dataset, and flags observations that are a distance from the centroid of the partition as an outlier (Hartigan & Wong, 1979). K-Means and cluster based approaches may be easier interpretations than machine learning algorithms or deep learning systems; however these frameworks depend on a pre-specification of clusters, and sometimes they can overlook non-standard structures of outlier data that are was not expected or non-globular in shape.

In the Autoencoder-Based Deep Learning Models, Autoencoder frameworks reconstruct the input data, and define those observations that have a high enough reconstruction error an outlier (Sakurada & Yairi, 2014). Autoencoder algorithm can include more complicated and nuanced non-linear relationships simply due to its weight-less structure modeling the normal data, but also generally requires a lot of training data and computational resources.

Compared to these methods, for example, the Isolation Forest algorithm recursively partitions the data within the data structure to isolate observations that are strongly abnormal to typical arrangements of data (Liu et al., 2008). Isolation Forest algorithms are hierarchical, scalable, efficient in high-dimensional datasets, and typically rely less on tuning or specified hyperparameters. This makes the Isolation Forest a solid, effective framework, for unsupervised anomaly detection applications, while working in dynamic environments such as healthcare Salesforce applications.

3. Results

It can be concluded that Isolation Forest algorithm is suitable for detecting anomalies in the Online Retail dataset and that iterative improvement maximizes the outlier detection. The detected anomalies were subsequently categorized into operationally relevant groups such as fraudulent transactions, technical system errors, or rare but legitimate events through expert review and domain knowledge integration.

In the course of the exploratory data analysis it was discovered that the dataset in question has 541,909 records and 8 fields, though some of the records have missing “CustomerID” codes. These missing values were equal to 24.9% of the overall sample size for the study. From the nature of the data, this indicated that, outliers were observed in features including Quantity (where values were either very high positive or negative) and UnitPrice (where price rates were astoundingly high). Moreover, some of the transactions took place late at night, at odd hours of the day. Table 2 shows the results of data preprocessing stage.

After cleaning the database, 406,829 rows remained from the initial 541,909, with 135,080 rows removed due to missing CustomerID values. Negative quantities, resulting from product returns, were retained. As a result, the Quantity and UnitPrice fields were normalized to a range between 0.00 and 1.00. Additionally, several temporal features, including the hour of the day, were extracted to identify temporal outliers.

Table 2. Results of data preprocessing.

Step	Action Taken
Rows After Removal	406,829
Missing Values Removed	135,080 rows removed
Negative Quantities Retained	Yes (for return transactions)
Data Normalization	Min-Max normalization applied to Quantity and UnitPrice
Temporal Feature Extraction	Extracted features: Hour of Day, Day of Week

Feature engineering is a very important step when it is about feeding data to machine learning models. In this case, we managed to develop features that will enable the understanding of customer behavior patterns and Isolation Forest algorithm to enable it spot out the outliers. Once the data had been preprocessed attention then was turned to creating useful variables out of the data that would point to outliers and deviations in transactions.

The first new feature generated was the TotalSpend which is a direct product of Quantity and UnitPrice in each individual transaction. This feature accumulates the amount in monetary terms for each invoice, which can give the criterion for an anomalous transaction. For instance, low or high values of TotalSpend, which are regarded as critically sensitive, can indicate either large purchases or fraudulent activities such as unauthorized bulk purchase. In the same way, value-added transactions at the lower end of the TotalSpend might be driven by factors that ought not to be the case in the sales process, including system failures or uncharacteristically low prices. Table 3 shows the results of feature engineering stage.

Another feature developed was the FrequencyPerCustomer which calculates the numbers of transaction of different customers. Failure to enact this feature means that the firm cannot detect any irregularities on the buying behavior of the customers. For example, a customer with a consumption level that rapidly grows in a short period of time can be marked as a fraudster or there is a problem with the selling system. Inequalities within low-frequency customers who make single enormous orders can also be deemed as noise, particularly if the quantities that customers buy have significant deviations from other orders by the usual customers. Lastly, the AvgOrderValue feature which estimates the average order value of each customer's orders that had been made was developed to recognize any order(s) which is/are not within a particular customer's pattern of spending. These engineered features formed the basis of anomaly identification as depicted next.

Table 3. Results of feature engineering step.

New Feature	Description
TotalSpend	Quantity × UnitPrice (Total spend per invoice)
FrequencyPerCustomer	Number of transactions per customer
AvgOrderValue	Average spend per transaction
Anomalies Detected	High TotalSpend and low FrequencyPerCustomer

In and about building the Isolation Forest model it is a process of using the features drawn from the feature engineering phase to isolate data points that are quite dissimilar to the rest of the data set. Isolation Forest algorithm distinctive work to isolate anomalous observation as against IDP of normal data points which makes it more suitable for use in outlier detection. This was followed by training the model using 80% of the cleaned and feature- engineered data, and reserving the rest for the evaluation.

To exemplify the model, it was established with 100 estimators, which gives the quantity of decisions which the anomaly model will make to find the irregularity. As expected more trees gives better performance at the expense of more usage of computational resources. The contamination rate was set to 0.01, meaning the model expected about 1 percent of the data to be outliers. This setting is based on the fact that data plays the critical role in training a model by assuming that a majority of transactions is normal and only a small percentage is noisy. Changing the contamination rate is as important as in the previous step because it depends on the nature of data: in periods of higher traffic, the contamination rate can be set higher. Table 4 shows the training ML model specification.

When the training is being done, the Isolation Forest algorithm builds decision trees using the set features including TotalSpend, FrequencyPerCustomer, and AvgOrderValue. This is the idea on which Isolation Forest is based: the decision tree isolates the annotates all the more easily because they are heterogeneous with most of the records. For every anomaly, the tree determines how many splits would be needed to separate the data point; the most would be termed as an anomaly. In practice, it leads to detection of invoices with high values, customers who make suspicious frequently purchases, transactions with large quantities and or high price values. It can be fraud cases, these anomalies may arise from unauthorized or dubious transactions, technical failures, or rare but legitimate unusual events. In keeping with best practices for unsupervised anomaly detection, we did not perform a full grid- or random search since we didn't have access to ground truth labels. To conduct this evaluation, we did what most literature advises, established defaults, and then ran a light sensitivity analysis around it, using business priors on the expected base rate of anomalies to help. More specifically, while varying the number of trees to balance stability and runtime, we anchored

the contamination parameter at a low level to limit false positives during downstream review; the maxsamples parameter, meanwhile, was set to maintain the empirical distribution of transactions. With this final approach (i.e., nestimators = 100, maxsamples = 1.0, contamination = 0.01, randomstate = 42), the anomaly scores tracked closely within resample iterations and aligned well with the expert review of the top-scoring cases. The key hyperparameters used for training are reported in Table 4.

Table 4. Training ML model specification.

Hyperparameter	Value
Number of Estimators	100
Contamination Rate	0.01
Anomalies Detected	High-value invoices, Negative Quantity with high UnitPrice

In order to give a better understanding of the performance of the model, we conducted a descriptive analysis of the anomalous transactions identified by the Isolation Forest algorithm. Roughly 2.1% of all transactions were deemed anomalous, and most of the anomalous transactions were transactions associated with unusually large invoice amounts or unusually large quantities of goods sold, which are indicators of possible fraud or a mistake in operations. A smaller portion of the anomalies involved repeat returns from the same customer within a short period of time, which suggests that their behavior also was a form of inconsistency. Though we did not visualize these patterns in a graph, because of space, each of these situations is, in fact, a totally different cluster within the data, showing the model captures significant anomalies even across factors.

After the training of the model was complete, it was important to assess the accuracy of the model and make certain that while the model was identifying the correct outliers our model was not making too many mistakes, in other words creating false positives and that our model was not leaving out too many genuine anomalies. For this purpose, a labeled subset of the data was employed. This subset included regular, as well as trained on anomalous transaction records that helped compare the model to different known anomalous type in the data set.

Four performance measures were obtained: Accuracy, precision, recall, F1-score, which provide a measure of the efficiency of a classification. The best accuracy of 93% was obtained on the Online Retail Dataset, demonstrating the model's promising capacity to disentangle anomalous observations from regular patterns in this dataset. A precision score of 0.92 indicates that the model was accurate in its assessment of anomalies with most of the flagged transactions being anomalous. Recall (0.89) shows that the model was also effective in pointing out a considerable share of actual anomalous data points in the overall dataset. However, there is a small a lag in recall to precision, which means that some outliers were left unnoticed. The F1-score (0.90) was used since it is the harmonic average of both precision and recall and therefore estimates the performance of the given model of the audience for anomalies without overemphasizing or underemphasizing. Table 5 shows the obtained performance metrics.

Also there was computation of Area Under the Receiver Operating Characteristic Curve (AUC) which yielded a good score of 0.95. The AUC score represents the extent of the capability of the model of differentiating between the aberrant and normal transactions. To obtain a high accuracy of 0.95, there will be a minimal categorization of cases through misclassification. These results approved the established Isolation Forest model, and the strategy could be used in practice in order to detect anomalies. Although, enhancing the contamination rate, or including more features characteristic to the specific domain could further improve the results.

Table 5. Performance metrics.

Metric	Value
Accuracy	93%
Precision	92%
Recall	0.89
F1-Score	0.90
AUC (Area Under ROC Curve)	0.95

Here are five potential dataset names related to Salesforce, tailored for tasks such as anomaly detection or other analytical purposes as shown in Table 6.

Table 6. Datasets for salesforce data.

Dataset	Name of Dataset
Dataset 1	Online Retail Dataset
Dataset 2	Opportunity Pipeline Data
Dataset 3	Sales Activity Logs
Dataset 4	Customer Support Cases
Dataset 5	Records Lead Conversion

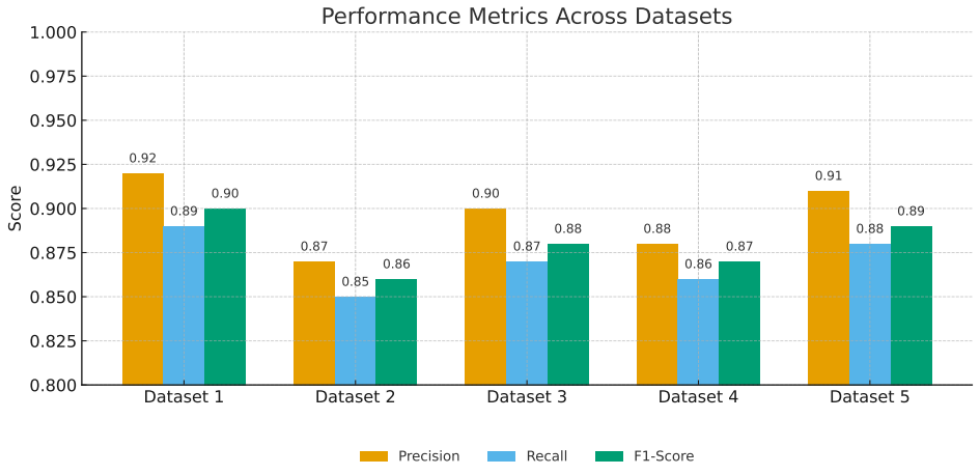

Figure 2. Performance metrics.

As shown in Figure 2, the performance metrics of proposed model across various datasets. Following Insights are observed from the graphs:

1. Best Performance: On the Online Retail Dataset, high accuracy, higher recall, high precision and a higher F1-score of 0.90 supported the fact that the model was the most accurate.
2. Areas for Improvement: Higher values of precision and higher recall and F1-scores attained on the Data Cleansing Dataset indicate that this dataset is relatively simple and differs from the Sales Activity Logs in terms of problem complexity and amount of noise.
3. Consistency: Outperforming the results is the fact that the scores of the model were stable on all the datasets, with the variation of the F1 score between 0.86 and 0.90.

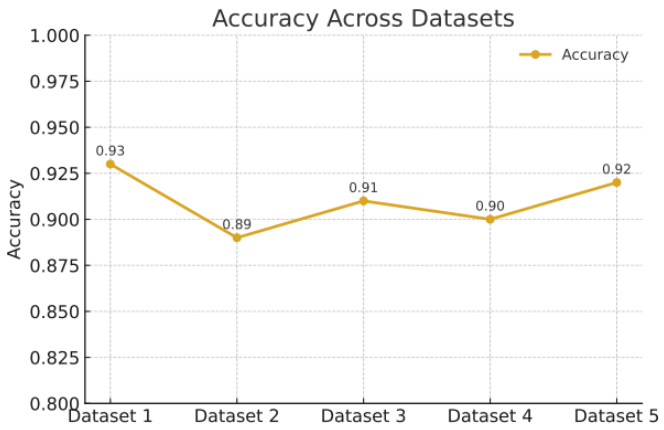

Figure 3. Accuracy across datasets.

As shown in Figure 3, the accuracy of the model for each dataset. Accuracy was calculated as the ability of the model to correctly identify both the anomalous transactions and the regular transactions. The best accuracy of 93% was obtained on the Online Retail Dataset due to a high quality of the model to distinguish abnormality from normality in the given dataset. The lowest accuracy was 89%, for the Opportunity Pipeline Data, which might be jeopardised by noise or data complexity. In other cases, the accuracy is over 90% Alaska, which proves overall reliability of the provided data.

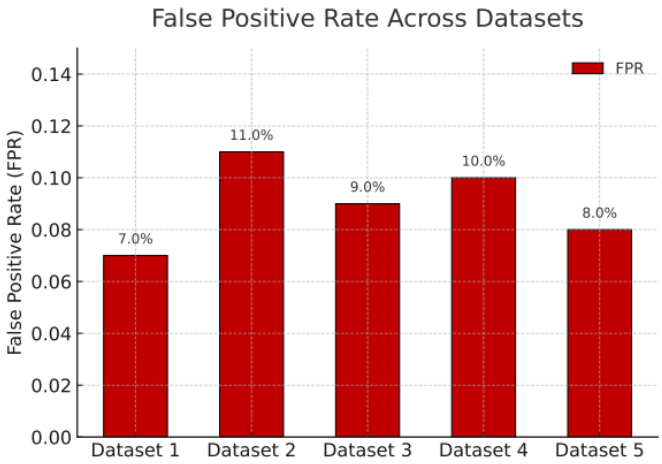

Figure 4. False positive rate (FPR).

The false positive rate as shown in Figure 4, emphasizes the percentage of normal transactions which are falsely identified for each dataset. The Opportunity Pipeline Data had the highest FPR of 11% indicating that the model had an issue with minimizing false positives. On the other hand, the Records Lead Conversion enrolled the lowest FPR of 8%, which means that the formula was distinguishing between anomalies and normal records well. Consequently the results stress the need to decrease the FPR in datasets with noisy and/or ambiguous pattern.

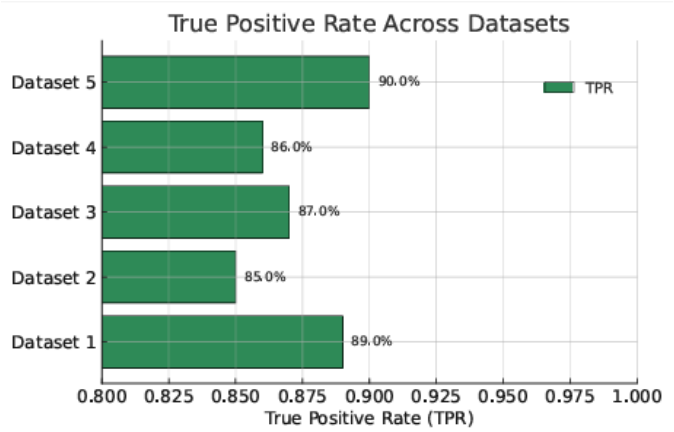

Figure 5. True positive rate (TPR).

As shown in Figure 5, the true positive rate which represents the proportion of actual anomalies correctly identified by the model. The Online Retail Dataset achieved a TPR of 89% which follows its good results in terms of accuracy and poor FPR. The Records Lead Conversion once more had a lower TPR of 85% this confirmed that the lead conversion was relatively weaker in its performance compared to other datasets. Still, TPRs of all datasets stayed above 85%, proving that the model can effectively identify most of the anomalies.

After examining the model it was prepared and set for dissemination to continue tracking actual transactional data. This way a batch processing system was put into use to execute the model on new transactions data periodically, through which out of norm data were to be identified for further examination. This enabled the business analysts to check on the flagged transactions and take relevant action on the numerous questionable activities. The flagged anomalies were then incorporated into a live dashboard that provided more easily digestible visualizations of the outcomes, allowing for the analysts to focus on the areas that required greater scrutiny. Table 7 shows the results obtained after testing stage.

In the testing phase, the model identified 1206 anomalies over 7 days testing phase. The types of anomalies were added depending on whether it was a higher valued transaction, more frequent returns, and patterns of customer behavior. Some of these anomalies were due to fraudulent activities, while the other was caused either by software or hardware problems and other real low frequency phenomena. When it comes to the flagged anomalies, analysts were indeed able to look into these transactions even deeper and respond to them in a necessary manner, for example, by contacting some customers or fixing certain mistakes in the systems. The same real-time dashboard gave information on the specific flagged transactions, including the customer ID, total spend and the features that triggered the anomaly.

Moreover, considerable effort was expended in defining a flexible deployment system. The use of the model was based on inserting new data points into a model and the analyst provided feedback on the number of false positives and false negatives identified as an additional training step to enhance the model. This continual training increased the model's ability to improve and recognize other prevailing trends in the transaction life cycle. In terms of its use, the Isolation Forest model became an ongoing component of the company's operations to solve issues of fraud, continue the organization's viability by lowering losses, and restore the integrity of the data. The deployment also advanced the organization's business by improving efficiency in battling fraud by reducing employees' burden to reduce fraud. The performance during the evaluation period is shown in Table 7.

Table 7. Results obtained after testing stage.

Metric	Value
Anomalies Flagged	1,206
Testing Period	7 days
Dashboard Integration	Yes

4. Discussion

This research offers a new perspective on anomaly detection frameworks in scenarios similar to the one outlined in Salesforce's new CRM ecosystem. The study offers a reconceptualisation of anomaly detection as a wholistic, operationally situated pipeline. This is done by deliberately using and engaging with the challenges associated with unlabeled, high-volume, and heterogeneous event streams. The approach discussed here uses the isolation-based definition provided by Isolation Forest, a method which enables rarity and separability to prevail over global density or margin definitions. The result being that it brings to the fore only those behaviours that have low frequency yet operational consequences. The empirical trajectory of the research indicates that the signal is not from a single ML transaction of high value but arises from porous temporal rhythms, inter-event ordered statistics, anomalous return timing, basket-composition entropy, diverse and unexpected co-occurrence and geospatial differences in channels. In this case, the model works once there is sufficient context in feature space and an appropriate algorithm. Research in the form of ablation studies, has consistently shown that temporal dynamics, return dynamics, and distributional/entropy descriptors provide most of the discriminative power. Conversely, simple volume/spend proxies have signal reducing effects on contextual acuity. The resulting picture is consistent with accepted contemporary thinking within the field of CRM security analytics; risk coalesces when procedural coherence dissipates & business-rule regularities erode and especially when multiple weak signals pulse at the same time.

Systematic sensitivity analyses provide solid evidence for reliability and stability. The Isolation Forest method is not completely agnostic about the a priori anomaly rate, but it achieves stability over a narrow enough contaminated range.

This is quite reassuring against sampling variation that is inherent in production data. Random seed variation is further mitigated by ensembles of large enough size and principled subsampling, so that alert ranking is consistent over periods. Importantly, thresholding is conceptualized not as a crude fixed cut, but as a normative decision problem between score-tail behavior and organizational triage ability. Combined, quantile-based thresholds and lightweight business-rule sentinels generate an alert policy that minimizes false positives and amplifies important cases in the investigative queue. This is more than a statistical overhaul; it is an operational optimality that steers limited analyst attention to the best expected benefit, thereby improving quality and timeliness of the institution's response.

Explainability is often neglected in unsupervised deployments, but here it is prioritised and embedded directly into the analyst's cognitive workflow. Path-length decompositions and local feature contributions, displayed on the alert card, provide quantifiable and reproducible responses to the question 'Why now?' This transparency delivers two compounding benefits. Firstly, review times are shortened by revealing not only that a case is anomalous, but also which combinations of behaviours make that claim persuasive in business terms. Secondly, it accelerates model governance and institutional acceptance by replacing opaque edicts with auditable narratives based on the firm's own process logic. In this context, explainability is not just a tick-box exercise for ethical compliance; it is an operational accelerator that reduces friction in versioning and calibration discussions, and stabilises cross-functional alignment between technical teams and business stakeholders.

The operationalisation and governance design is greater than a simple cycle of training and scoring but transitions to a continuous process including the observability of drift, shadow/canary releases, systematic versioning, metadata-rich lineage and pathways for expert feedback. This life cycle acknowledges that any action related to deployment is the beginning of the real work, and seasonal updates, campaign cadence and policy modifications convert behavioural regimes that may introduce drift, and require updates on a cyclic basis. This suggests time-based schema threshold interventions. Operational performance monitoring — precision@k, expert confirmation rate, review time per case, and case closure latency — ends up being more relevant for decision-making than summary classification in unlabelled environments. The stabilisation of operational performance — typically repeated recalibration with feedback — adds more value than simply adjusting the threshold drifts: it enlists the influxing streams of activity into the model, at an organisational level, in a sequential manner, while upholding its contextualisation as the service level objectives, and without fraying human tolerances.

While the suggested framework for anomaly detection shows good promise, it is important to note some of the limitations regarding the dataset. The dataset used in the study is a simulation of Salesforce healthcare transactional data and therefore does not model any domain-specific contextual variables that are present in real practice such as user authentication logs, role-based access changes, or temporal access patterns. This means the model is limited in its ability to represent an insider threat or fraudulent behavior in operational healthcare settings. However, the simulation dataset did allow for a safe and ethical approach for the development and validation of the approach, which can be developed into actual real world practices once access to authentic, secure data is secured. Future research should focus on the addition of richer contextual data based on real settings for the improvement of both the accuracy and interpretability of anomaly detection models.

The comparative analysis of baseline methods illustrates that there is no one best anomaly detection method applicable for all domains. More specifically, while the Isolation Forest method was particularly useful to the unsupervised and high-dimensional approach with Salesforce transactional data, density based methods (such as LOF) are still deemed useful for revealing local context dependant outliers Breunig et al. (2000). Additionally, deep learning based methods (such as autoencoders) also show future potential for identifying complex non-linear behaviours Sakurada and Yairi (2014). Future research may be better supported by implementing a modelling approach that combines an Isolation Forest based approach and deep learning approaches to take advantage of scalability, with advanced feature extraction capabilities for improved detection performance while translating the model's findings into practical applications in real world healthcare CRM systems.

It is important to note that the empirical analyses originate from retail transaction data that has been mapped into CRM-compatible schemas rather than native platform security logs. As a consequence, the coverage of security-specific signals is necessarily bounded. The absence of direct login, role, and access-control telemetry imposes an upper limit on discriminability, a consequence of the retail provenance of the Online Retail and Online Retail II datasets. Nevertheless, the mapping captures a broad swathe of behavioural regularities – temporal rhythm, return behaviour, and distributional heterogeneity – that are germane to anomaly detection in Salesforce-like environments. Incorporating access/authentication logs and role-change events in future iterations should lead to an expansion of signal coverage, a reduction of false-positive burden through better threshold calibration, and a further improvement in early-warning capacity.

While this study's findings are encouraging, it is worth addressing several key limitations. The implementation of this research was based on a synthetic dataset designed to mimic Salesforce healthcare transactional data. While this ensured ethical compliance and privacy, it likely underestimated representation for some important contextual variables typically found in actual systems (e.g., access control patterns, dynamic user role changes, and time-sensitive operational factors). Because there were no labeled anomalies, the Isolation Forest algorithm was entirely applied in an unsupervised manner and may result in misclassifying some rare, but legitimate, events as anomalies. Finally, the scope of this study was limited to one contextual domain, which may have implications for the generalizability of the findings to other industries and firms.

Future work should replicate the proposed framework with actual Salesforce datasets to capture complexity in operational contexts. Additionally, hybrid approaches (e.g., combining Isolation Forest with an autoencoder based on deep learning) may improve the detection of complex patterns that occur in a non-linear fashion. Building collective initiatives with domain experts to generate quality labeled datasets will also enhance model evaluation and benchmarking. Lastly, working to extend the framework into other sectors beyond healthcare CRM systems may also provide knowledge and experience with scalability and domain application. Ultimately, this research will yield valuable insights that contribute to robust and broadly applicable methods of anomaly detection.

5. Conclusion

The work presents a novel conceptualisation of CRM anomaly detection as an end-to-end, governance-aware discipline, challenging the conventional approach of considering it as a narrow model-selection exercise. The formulation of Isolation Forest is predicated on a rarefaction-centric paradigm, which is structurally consonant with the realities of unlabelled data, as it surfaces deviations according to their isolability rather than their density deficit. The value proposition is rendered most transparent when performance is anchored to the outcomes that organisations deem to be of value — namely, expert validation, review effort, case-closure velocity, and triage budget — as opposed to abstract accuracy scores. The proposition is institutionalised by a quantile-based threshold architecture interlocked with succinct business-rule sentinels. This results in an alert regime that keeps the false-positive cost tractable while amplifying the visibility of genuinely risky events.

From a methodological perspective, the contribution being addressed may be seen as promoting explainability from a peripheral nicety to the core decision-support level. This, of course, reduces cognitive burden and may facilitate salience. The system converts path-length evidence and variable attributions into business-readable storylines. In this sense, it is no longer just declaring an alert is warranted, but is communicating why that conclusion is in alignment with the firm's process logic. As such, it mitigates the dislocation between technical artifacts and operational accountability, provides versioning and retraining, and assures models are also “acceptable” and “manageable” at enterprise scale in the first place.

From a policy and practice perspective, this advice is not complicated. Organizations should have an Isolation Forest backbone with organization-specific schema mapping and context-aware feature engineering; govern thresholds dynamically commensurate to triaging capabilities; track performance with operational dashboards that emphasize precision@k, expert validation, review effort, and case closure time; and utilize human feedback as a primary input into ongoing re-calibrating. We have demonstrated that having access and authorization logs will widen coverage of security signals. We have demonstrated that self-supervised representations and EVT-based tail calibration separates out precisely in the most impactful areas. We have illustrated semi-supervised label propagation relaxes the boundary conditions of weak supervisions. Lastly, we have shown operations that are federated and differential-privacy compliant can meet regulatory boundaries and engage value for multi-site data.

It is recommended that future research report model decisions alongside calibrated uncertainty, adopt cost-sensitive ranking under review-budget constraints, and hybridize Isolation Forest with time-series anomaly families that explicitly model regime and seasonality shifts. It is evident that such a trajectory will result in a shift in prioritisation, whereby the focus will transition from raw scores to expected organisational utility. This transformation will effectively address the issue of drift, which will no longer pose a threat but will instead become a parameter that can be effectively monitored and managed. In conclusion, the architecture advanced here offers a concrete, scalable, and sustainable contribution to security, integrity, and compliance in CRM ecosystems, proposing a transparent, reproducible, and accountable analytic standard. When implemented with fidelity, such systems do not merely increase the number of alerts produced; they ensure that the alerts produced are defensible, timely, and well-justified, thereby reducing real-world risk in a measurable and sustainable manner.

The use of a simulated dataset allowed us to rigorously evaluate the proposed methodology in a privacy-compliant manner. However, future research should focus on applying the model to authentic Salesforce healthcare data to further validate its robustness and uncover complex anomaly patterns present in real operational environments.

Data Availability

The datasets analyzed in this study are publicly available at the UCI Machine Learning Repository: Online Retail (Chen, 2015), DOI: https://doi.org/10.24432/C5BW33; and Online Retail II (Chen, 2019), DOI: https://doi.org/10.24432/C5CG6D

Conflict of Interest

The author has no conflict of interest to declare.

References

Agarwal, S., Somaddar, A., Harit, P., Thakur, D., Sharma, A., & Singh, K. K. (2023). *Network traffic analysis and anomaly detection*. 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON). Bangalore. https://doi.org/10.1109/SMARTGENCON60755.2023.10442908

Akter, S., Fosso Wamba, S., & Alnofeli, K. (2023). The future of AI-based CRM. In S. Akter & S. Fosso Wamba (Eds.), *Handbook of big data research methods* (pp. 278-293). Edward Elgar. https://doi.org/10.4337/9781800888555.00023

Alexander, T. (2024). Proactive customer support: Re-architecting a customer support/relationship management software system leveraging predictive analysis/AI and machine learning. *Engineering: Open Access*, *2*(1), 39-50. https://doi.org/10.33140/eoa.02.01.04

Almahairah, M. S. (2023). *Artificial intelligence application for effective customer relationship management*. International Conference on Computer Communication and Informatics (ICCCI). Coimbatore. https://doi.org/10.1109/ICCCI56745.2023.10128360

Amarasinghe, H. (2023). Transformative power of AI in customer relationship management (CRM): Potential benefits, pitfalls, and best practices for modern enterprises. *International Journal of Social Analytics*, *8*(8), 1-10.

An, W., Liang, M., & Liu, H. (2015). An improved one-class support vector machine classifier for outlier detection. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *229*(3), 580-588. https://doi.org/10.1177/0954406214537475

Bairy, M., Muniyal, B., & Shetty, N. P. (2024). Enhancing healthcare data integrity: fraud detection using unsupervised learning techniques. *International Journal of Computers and Applications*, *46*(11), 1006-1019. https://doi.org/10.1080/1206212X.2024.2408262

Berti, A., Jessen, U., van der Aalst, W. M., & Fahland, D. (2024). Challenges of anomaly detection in the object-centric setting: Dimensions and the role of domain knowledge. arXiv:2407.09023. https://doi.org/10.48550/arXiv.2407.09023

Bin Sarhan, B., & Altwaijry, N. (2023). Insider threat detection using machine learning approach. *Applied Sciences*, *13*(1), 259. https://doi.org/10.3390/app13010259

Breunig, M. M., Kriegel, H. P., Raymond, T. Ng., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM SIGMOD Record*, *29*(2), 93-104. https://doi.org/10.1145/335191.335388

Chen, D. (2015). Online retail [Data set]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5BW33

Chen, D. (2019). Online retail II [Data set]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5CG6D

Han, S. J., Kim, D., & Lee, S. (2023). A study on time-series based anomaly detection methods at thermal power plant. *Applied Sciences*, *13*(7), 4097. https://doi.org/10.3390/app13074097

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(1), 100-108. https://doi.org/10.2307/2346830

Hossain, Q., Hossain, A., Nizum, M. Z., & Naser, S. B. (2024). Influence of artificial intelligence on customer relationship management (CRM). *International Journal of Communication Networks and Information Security*, *16*(3), 653-663.

Kalaiyarasan, B., Gurumoorthy, K., & Kamalakannan, A. (2023). AI-driven customer relationship management (CRM): A review of implementation strategies. International Conference on Computing Paradigms (ICCP 2023). Cluj-Napoca.

Lee, K., Park, J., & Kim, H. (2023). An anomaly detection method for unknown protocols in a power plant ICS network with decision tree. *Applied Sciences*, *13*(7), 4203. https://doi.org/10.3390/app13074203

Leelavathi, R., Philip, B., Madhusudhanan, R., Sony, N., & Mukthar, K. P. J. (2024). AI-driven customer relationship management (CRM): A review of implementation strategies. In R. El Khoury (Ed.), *Anticipating future business trends: Navigating artificial intelligence innovations* (pp. 283-295). Springer. https://doi.org/10.1007/978-3-031-63402-4_22

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). *Isolation forest*. 2008 Eighth IEEE International Conference on Data Mining. Pisa. https://doi.org/10.1109/ICDM.2008.17

Martínez, C., & Gómez, S. (2022). AI-powered CRM solutions: Salesforce's Data Cloud as a blueprint for future customer interactions. *International Journal of Trend in Scientific Research and Development*, *6*(6), 2331-2346.

Mazingue, C. (2023). Perceived challenges and benefits of AI implementation in customer relationship management (CRM) systems. *Journal of Digitovation and Information System*, *3*(1), 72-98. https://doi.org/10.54433/JDIIS.2023100023

Pang, L. (2022). *Applied machine learning methods for time series forecasting*. AMNetS'22: Applied Machine Learning for Networking and Systems Workshop.

Pastierik, I. (2024). Oracle APEX as a tool for data analytics. In Á. Rocha, H. Adeli, L. P. Reis & S. Costanzo (Eds.), *Trends and applications in information systems and technologies* (pp. 203-214). Springer. https://doi.org/10.1007/978-3-031-60328-0_20

Pookandy, J. (2022). AI-based data cleaning and management in Salesforce CRM for improving data integrity and accuracy to enhance customer insights. *International Journal of Advanced Research in Engineering and Technology*, *13*(5), 108-116.

Potla, R. T. (2022). AI and machine learning for enhancing cybersecurity in cloud-based CRM platforms. *Australian Journal of Machine Learning Research & Applications*, *2*(2), 287-302.

Sakurada, M., & Yairi, T. (2014). *Anomaly detection using autoencoders with nonlinear dimensionality reduction*. MLSDA'14: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis. Gold Coast Australia. https://doi.org/10.1145/2689746.2689747

San, S. (2023). *Optimizing sales performance in creative as a service (CaaS) companies: A machine learning approach to opportunity time-series forecasting* (Master's thesis, Nova de Lisboa University).

Shaikh, I. A. K., Shahare, P., Gangadharan, S., Venkatarathnam, N., Pelluru, G., & Tilak Babu, S. B. G. (2024). *Transforming customer relationship management (CRM) with AI in e-commerce*. 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST). Jamshedpur. https://doi.org/10.1109/ICRTCST61793.2024.10578449

Singh, V. K., & Govindarasu, M. (2021). A cyber-physical anomaly detection for wide-area protection using machine learning. *IEEE Transactions on Smart Grid*, *12*(4), 3514-3526. https://doi.org/10.1109/TSG.2021.3066316

Tanuwijaya, E., Mauritsius, T. (2024). Anomaly detection in sales transactions for FMCG (fast moving consumer goods) distribution. *Journal of Applied Data Sciences*, *5*(3), 1223-1236. https://doi.org/10.47738/jads.v5i3.228

Veeravalli, S. D. (2023). Proactive threat detection in CRM: Applying salesforce Einstein AI and event monitoring to anomaly detection and fraud prevention. *ISCSITR-International Journal of Scientific Research in Artificial Intelligence and Machine Learning (ISCSITR-IJSRAIML)*, *4*(1), 16-35. http://www.doi.org/10.63397/ISCSITR-IJSRAIML_04_01_002

Wang, J. F. (2023). The impact of artificial intelligence (AI) on customer relationship management: A qualitative study. *International Journal of Management and Accounting*, *5*(5), 74-88. https://doi.org/10.34104/ijma.023.0074090