An Efficient Anomaly Detection System for E-commerce Pricing Using Machine Learning Techniques.

Faiz ur Rehman[a,], Wasiq Aslam[a], Fahad Aslam[a], Samina Nazar[a],Faheem Nawaz khan[b], Afshan Naseem[c]

[a]Department of Computer Science, MY University, Islamabad, 47350 Pakistan;
[b]Department of Computer Science, Shifa Tameere-e-Millat University, Park Road Campus Islamabad;
[c]Institute of Oceanography and Environment (INOS), Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

Abstract Anomaly detection plays a vital role for e-commerce businesses due to the enormous volume of numeric data they handle. Price data is a prominent example of such data, and detecting anomalies in numeric data is challenging due to potential errors and outliers, which can disrupt sales calculations and result in financial losses. Anomaly detection in e-commerce aims to identify the outliers within a given dataset automatically. This research study explores unsupervised and supervised machine learning approaches for detecting price anomalies in publicly available e-commerce datasets. This research proposed an anomaly detection framework that involves feature selection using the Z-score technique, followed by multiple corresponding analyses (MCA) of the same selected features. Machine learning techniques, including Isolation forest and DBSCAN, are employed to detect price anomalies. A threshold is then established for a voting scheme (Voting ZID) to enhance the accuracy of the anomaly detection technique. Different machine learning classifiers are evaluated after labelling the anomalies (anomalous and non-anomalous), including Gaussian Naïve Bayes, Random Forest, LDA, SVM and the proposed Gradient Boosting Machine technique. The algorithms are optimized by tuning different parameters. The proposed technique attains high performance in terms of precision 0.9713 % and accuracy 0.9751 % - the highest on the benchmark.

Keywords: Machine learning, Anomaly Detection, Algorithm, and E-commerce.

Introduction

Anomaly detection remains an open issue in the field of E-commerce. Usually, machine learning algorithms are used to detect incorrect and inconsistent data. Machine learning techniques have identified some issues and several solutions to the problems. Thus, Anomaly detection in E-commerce will become a way of life for business managers. Detecting anomalies is crucial but often provides critical data in various application domains. Anomaly detection is widely used in the real world, such as in cyber security, medical diagnosis, social media, fault detection, fraud prevention, finance, energy demand or consumption, and e-commerce [1]. Such types of applications need anomaly detection algorithms that accurately detect with the best performance and fast execution. Various techniques detect anomalies in real-world environments by considering internal data, i.e., input data, and external data, i.e., output data like Amazon, Walmart, Big Mart, and eBay [2]. Let's consider an example regarding price anomalies in an E-commerce data set. Studies have shown that price cannot be negative; it must range from 0 to any positive integer value.

Further, the quantity of items also cannot be negative. The preprocessing step removes negative price and quantity in this data set. Furthermore, this research utilizes approaches like interquartile range and z-score to detect anomalies in price, whether abnormal or not. Detection anomalies in such datasets are difficult for the machine due to huge variances or price fluctuations. Such variance causes and leads to anomalies that are very difficult for machine learning algorithms to detect. In retail companies, anomaly detection methods help the monitoring team if any changes or unexpected changes in several completed transactions [3]. Outliers, noise, and anomalies occur for many reasons, including holiday shopping or sudden weather changes. By detecting anomalies if prices increase and selling quickly, businesses can check stock levels and make important decisions, such as ordering surplus stock or checking alternative inventory that accurately meets customers' needs [4].

E-commerce has many anomalies, such as changes in cost structure, price structure, sales, profits, products, unexpected technical site issues, etc. This research section reviews some of the techniques already proposed from the perspective of anomaly detection in retail and e-commerce datasets[5]. This paper addresses e-commerce website challenges by introducing the SVM algorithm to improve accuracy in detecting price anomalies in real-time datasets [6]. However, the experimental results of the baseline framework show that they can detect anomalies with 94.02 % precision on the average. Our approach detected anomalies in price features using unsupervised methods using our hybrid approach (Z-Score + Isolation Forest + DBCAN), and then moved towards a supervised approach on the basis of proposed voting schemes to classify them anomalous and non-anomalous data. This research utilizes and implements different machine-learning algorithms. After deploying several models, our proposed GBM gained better accuracy and precision than previous approaches. It is valuable research and practical for large-scale retail companies to detect anomalies properly and efficiently. Therefore, this research will further investigate and develop an optimal framework to detect anomalies with improved performance.

Lastly, we will discuss the significance of anomaly detection. In business analytics, anomalies are removed in the last few decades before building models. Those outliers and anomalies are difficult to remove. Due to advancements in machine learning and deep learning technology, anomalies are identified automatically, and alerts are generated to support teams. Thus, advancements in hardware and software considerably reduced the anomaly detection cost and made it easy and affordable for retail companies. E-commerce websites generate a massive amount of data [6] because daily operations and transactions have made a collection of information. The collected data are used to make essential decisions for retail companies [8] in future circumstances. Thus, anomaly detection is beneficial for retailers when using data. Due to machine learning advancements, anomaly detection is effortless and does not need surplus work to maintain data collection. Real-time information is needed for e-commerce retailers [9] because they require faster responses to make changes during alerts generated by data in real time[7]—automatic real-time detection made possible for businesses with high speed and error-free analytics to their lives. When anomalies occur, the systems detect them automatically and correlate to scores, and alerts are generated to prevent profit and sales loss for their business. It allows e-commerce retailers to check effectively and efficiently. A dotcom retailer and business adopt anomaly detection as a way of life due to its fasteners, ease, cheapness, and effortlessness.

Related Work

The research article [8] shows three main machine-learning anomaly detection approaches. These approaches are supervised, semi-supervised, and unsupervised techniques, as shown in Figure 2. Adopting the current anomaly detection technique is crucial, depending on the availability of suitable labels in the data set. Supervised anomaly detection requires a data set with normal or abnormal labels for the algorithm. This type of approach involves training the model or classifier. Unsupervised anomaly detection techniques require a data set with neither classification nor labelling and permit the algorithm to work on information in the data set without guidance. Semi-supervised anomaly detection techniques are a series of data points, some for which labels are unknown. It is used to classify some unlabelled data using a label information set. Anomaly detection in E-commerce is a significant problem widely discussed in cyber security, medical diagnosis, social media, fault detection, fraud prevention, finance, energy demand or consumption, and E-commerce. A research article[9] describes that anomalies are divided into three categories: points, contextual, and collective anomalies. Anomalies are the data points, events, or observations that deviate from the expected behavior of the dataset. Three main types of outliers or anomalies are broadly divided and discussed.

1. Contextual Anomalies: The change in context-specific. These anomalies are single observations that mostly deviate from the expected observation. For example, the family expenditure is 100 $ every day, but it may be odd otherwise. Fraud detection of credit card in which amount is spent. Purchase with considerable transaction value.

2. Anomalous Subsequences: Anomalous subsequences are subsequences within a time series that deviate from other parts of the time series.

3. Time series Anomalies: Anomalous time series are time series that deviate from a collection of time series.

Anomaly detection is a crucial task for most machine-learning applications. With technological advancement, machine learning approaches have gained high popularity in the last decade. Thus, many researchers started working on similar machine-learning methods to identify and detect anomalies in multivariate data. For instance, the researcher worked on a clustering technique i.e. k-Mean, to detect outliers in univariate data.

In figure 1 describe Anomalous and non-anomalous data while figure 2 show approaches for anomaly detection using machine learning.

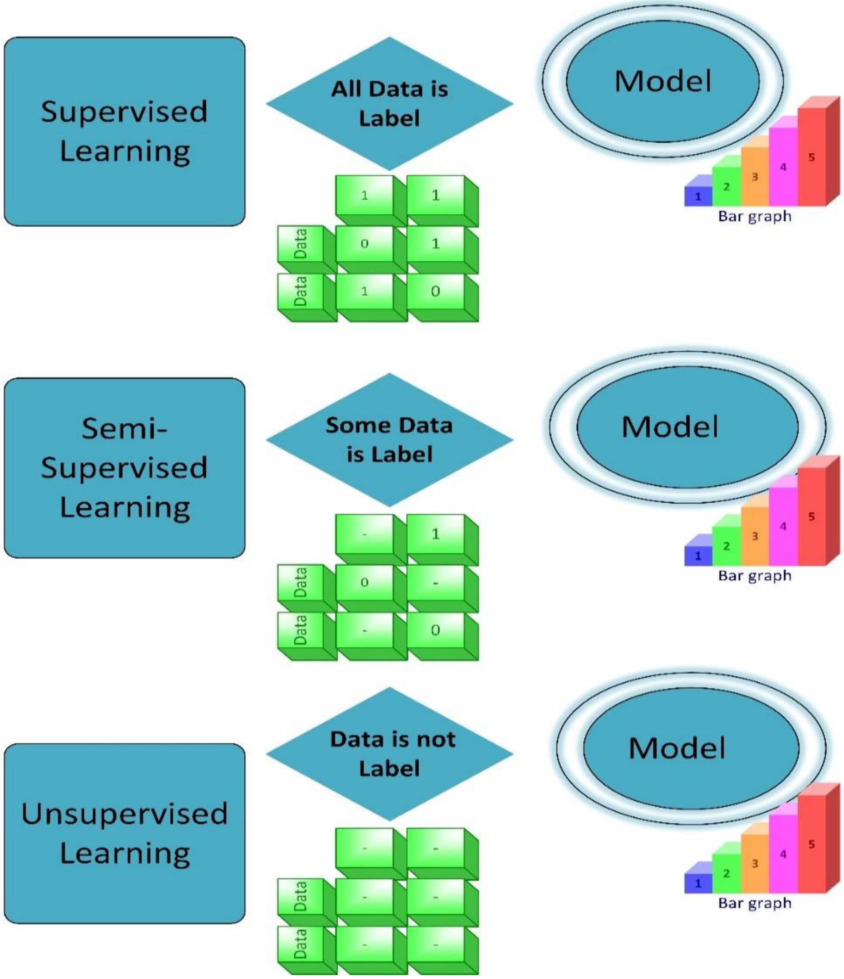


Figure 1. Machine learning Approaches for Anomaly Detection

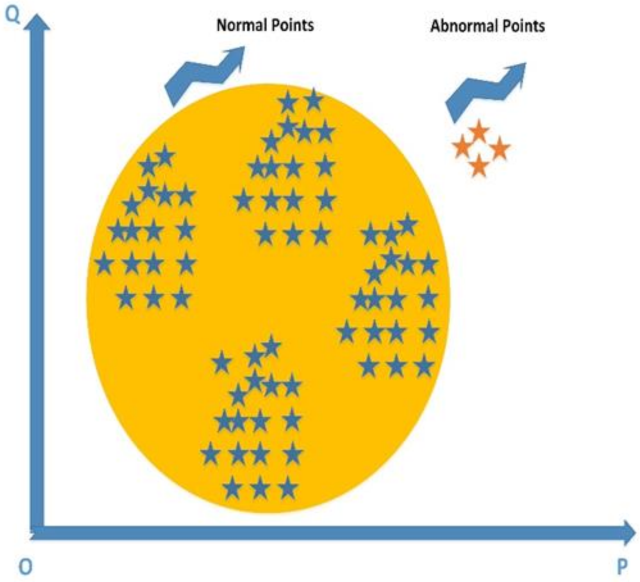


Figure 2. Anomalous and Non Anomalous Data

Machine Learning Approaches for Anomaly Detection

Machine learning techniques are widely used for solving complex nonlinear problems. There are many applications in which machine learning techniques are successfully applied, such as cyber security, medical diagnosis, social media, fault detection, fraud prevention, finance, energy demand or consumption and E-commerce. These techniques have been used for years in the anomaly detection domain.

In large-scale, online pricing systems, the timely detection and performance of anomalies are considered highly important. There is a lot of significant work on anomaly detection approaches, Welford's algorithm, a quartiles-based solution [10], a z-Score metric-based solution [11][12] and a machine learning-based solution called Half-Space Trees (HST)[13]. The first three algorithms are based on statistical indicators/metrics, while the last comes from machine learning. Most of the literature review on the unsupervised approach deals with real-world network environments and online streaming datasets, while the supervised approach is considered for training data sets. Such a type of training data set is difficult for online real-world networks. The research article [14] states that the decision tree technique, i.e., Isolation Forest, needs a lesser considerable amount of memory and works for linear time complexity. It is suitable for large data sets that consist of arbitrary characteristics. It detects anomalies fast and with high performance, while the distance-based approach takes enough time and performs poorly. However, techniques SVMs [8] support vector base algorithms to handle non-labelled data. In the research article [15], the author calculates the points having more weight and then computes them to K nearest Neighbour in search space using the Hilbert curve. In the research article [16], an auto encoder is an unsupervised approach that rebuilds the arbitrary or false input with an efficient architecture having an encoder and decoder. A research article [17] that implemented Graph Neural Network (GNN) detects hidden outliers or anomalies in social networking websites like Facebook, Twitter, etc., with efficient performance regarding accuracy. Time-series anomaly detection approaches [18] work with large online real data sets to detect anomalies quickly, accurately, and smoothly. In article [19], the Negative Selection algorithm operates in time linear complexity regarding the dataset size. In a research article [20] on social networking websites like Facebook, LinkedIn and huge enterprise, larger companies need a productive system that detects anomalies in real-time with efficient performance in less time. In literature [21] the authors designed such a system based on bio surveillance to detect false patterns or anomalies for retail data in the pharmacy. The system checks daily over a certain time frame which monitors and detects anomalies of sale daily. In research article [22] fast space-time metrics and statistics are used which enhance the pharmacy system more accurately and feedback of users was also achieved.

In recent years [23][24], machine learning has played an important role and produced good results in different fields. Techniques based on machine learning moved researchers towards anomaly detection in E-commerce [25][26]. In this paper, we propose machine learning techniques to find anomaly detection in the e-commerce dataset. The goal is to find anomaly detection in pricing of E-commerce data set products while reducing time, complexity, and accuracy.

The given input data's goal is to calculate and detecting anomalies in E-commerce pricing data set automatically. Still, many traditional methods used naïve based classifier[27] k-NN [28], local outlier factor and Fast Angle Based Outlier Detection (FABOD) [29] and quartile-based methods such as one-class SVMs [30][31]; however, their training time or prediction time were too long for scale. As discussed earlier, different methods have been proposed to calculate anomaly detection in E-commerce. However, accuracy still needs to improve.

Gradient Boosting Machine (GBM)[32] and neighbourhood-based approaches such as K Nearest Neighbourhood (KNN) [33], Local outlier factor (LOF) [34], and based Outlier Detection (ABOD) [35] are applied to different anomaly detection tasks from last many years. Tree-based approaches, i.e. the Random Forest and Gradient Boosting Machine, provided good performance and prediction times. At the same time, the Isolation Forest was not good, as we will see in the experiments section.

In a research article [36] four proximity-based unsupervised machine learning techniques are implemented in a power quality dataset. This contrasts standard classification tasks because these techniques are applied on unlabelled datasets, taking the values as input of power quality data into account. Finally, it detects anomalies and analyses them in a power quality dataset using proximity-based unsupervised machine learning techniques. Histogram-based outlier detection (HBOS), Cluster based Local outlier factor (CBLOF), Local Outlier factor (LOF) and K- K-nearest neighbourhood (KNN) methods are implemented to detect anomalies in the power quality dataset. Local outlier factor (LOF) does not detect anomalies as abnormal. Such methods are also implemented and observed in other datasets.

Neighbourhood-based anomaly detection techniques, in which KNN is the most popular algorithm, take an unsupervised approach when it has come to anomaly detection. This is because there is no real learning involved in the process and there is no pre-defined or pre-determined labelling of “anomalous” or “non-anomalous” in the dataset. It is fully based upon threshold values. Scientist and researchers randomly decide the cut-off values beyond which all points which are observed, after observation they classified as anomalies that is why there is no train test split of data or an accuracy report.

Interquartile range IQR is the simplest approach to detect irregular pattern, in which the data is to flag as input and data points that differ from common statistical characteristic distribution properties, including mean, median, mode and quartiles. Inter-quartile Ranges IQR is statistically used to analyse and measure the statistical dispersion, and the dataset is divided into quartiles. In other words, we can say that in a dataset in which anomalies are detected, we divide the ith feature or any set of observations into four defined intervals or quartiles based on the dataset's values.

Interquartile has some drawbacks in scenarios in which the pattern is based on seasonality. This is complicated and sophisticated technique, such as decomposing or diving the dataset into three or more quartiles or multiples trends to detect the changes in seasonality. The other drawback is that abnormal or normal may immediately or randomly change, as malicious adversaries constantly adapt themselves.

Unsupervised Anomaly Detection

There is a significant work on unsupervised anomalies detection in machine learning. In research article [37] unsupervised anomaly detection algorithms are mainly divided into (1) Nearest-neighbour approach, (2) Clustering approach, (3) statistical Approach. However researcher develops newly algorithms and compare their state of art techniques with results for instance local outlier factor (LOF) and k-nearest neighbour (KNN), unfortunately some of datasets are not present which make difficult and serious problem for finding metrics. Lazarevic et al. [37] analysed and implemented four techniques LOF, k-NN, PCA, and SVM for intrusion detection by taking data set of KDD-Cup99. The same approach was used on the same data set by applying and implementing a cluster-based technique i.e. KNN and one class SVM. Auslander et al. [37].implemented models of k-NN, Local outlier factor and clustering based techniques on maritime video surveillance data. Furthermore in research article [38] analysed Support Vector Data Description (SVDD), Gaussian Mixture Model (GMM) and k-NN for identifying and detecting outliers in 10 differ datasets. In these approaches [20][39] author implemented semi-supervised approach for training data set although author claims to implementation of unsupervised approaches. Comparative analysis of ten different data set of different anomalies approaches were used in by author Carrasquilla [37]. In research article [40] one unsupervised anomaly detection technique was implemented on different data set however its experimental results are compared to other outlier detection approaches.

There is a significant work and studies for implementation of single approaches only, however anomaly ensembles technique [41] is concatenation of two or more unsupervised anomaly detection approaches in order to enhance the performance of joint approaches. Although unsupervised anomaly detection doesn’t contain label data set, thus it is difficult to work on it and challenging task for simple combinations of approaches.

Supervised Anomaly Detection

In [42] classical approaches such Gradient Boosting Machine (GBM), Support Vector Machines (SVM), Decision Tree (DT) and Random Forest (RF) played very beneficial. In research article [43], European dataset is taken and different approaches Gradient Boosting Machine (GBM), Logistic Regression (LR), Reinforcement Learning (RL), Support Vector Machine (SVM) and concatenation of certain classifier were deployed that recall rate is increases up to 91 %. This is done by analysing and balance data by standardizing data, through this recall and Precision were increased. In these models as described above, Random Forest (RF) proved high accuracy rate of 95.5% than decision tree (DT) with 94.3 % accuracy and logistic Regression with 90 accuracy on European data set.

In [42] k-Nearest neighbours (KNN) and anomaly detection approaches are applied, this model work well and efficient in fraud detection system because in this work false rate is decreasing and detection rate of anomalies increasing. KNN approach also work in paper [44] author described and experimented this approach and comparing with other models. Furthermore some classical algorithms are analysed and compare with deep learning approaches. These approaches gain accuracy 80 percent approx. In paper [45] author compared the performance of KNN, NB, gradient boosting tree (XGBoost), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Decision Tree (DT), Multi-Layer Perceptron (MLP) and stacking classifier on European data set. After standardizing data, experimental result of classical approaches gain well accuracy however the stacking approaches gains higher accuracy than classical approaches. In research article [45] Auto-encoder and Restricted Boltzmann Machine approach were analysed and deployed in which experimental result show that these algorithms are highly considerable , best fitted and suitable for credit card anomaly detection. In research article [46] majority of work has been done on machine learning approaches for detection of anomalies in real time data set because these models are less time consuming and perform well in large scale data set. However deep neural network approaches on anomalies detection perform well in very large data set but such models are computational time consuming and expensive.

In literature review of this section we present detecting anomalies in E-commerce by comparison table 1. Different researchers applied different techniques to detect anomalies and tried to improve the Precision and accuracy. We have presented results, advantages and limitations of existing techniques. In the next research we will present our proposed methodology to find anomalies detection in E-commerce and improve its performance. In this research, the proposed system has been presented for detecting anomalies in E-commerce. The proposed architecture uses machine learning models. We will discuss model architecture and detail description of each module.

Table 1. Overview of Machine Learning Approaches and their Features

Paper	Year	Techniques	Dataset	Limitation
[47]	2023	Long short-term memory networks	Sale and Price 500 Index	A problem with LSTM is that it over fits the training data and generalization ability is lose particularly when there is noisy or erroneous data
[48]	2022	Robust covariance and Isolation Forest	Pump and dump in the Bit coins	Visualizing result is complicated. Taking consideration for correctly optimized because it requires a lot of computational power and implementation may be long.
[49]	2022	Principal Component Analysis and Neural Networks	Stock Price Data set	Decision tree, detecting anomalies based on generating I trees, Perform well if density of data is not same in dataset. While Random cur forest is good however running time is too large.
[50]	2023	Support Vector Machine	Credit Card Fraud Detection dataset	Powerful technique, easy to understand and implement however running time is too large depending upon the nature of data.
[51]	2023	Gradient Boosting Machine	Bank Transaction dataset	This approach is based on relationship between features and employing Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA).

[52]	2023	Extreme Boosting	E-commerce Transaction Dataset	IsGuestOrder feature is included in the dataset to increase the performance of Logistic regression and extreme boosting machine.
[1]	2022	Isolation Forest	Industrial Dataset	Isolation forest, no clear threshold for decision and parameter depend upon the size and dimension of dataset.
[53]	2023	XGBoost algorithm	Onion Price Dataset	The more complex data, it will not perform well because it is based on decision tree.
[54]	2022	AE- LSTM	Solar Power Generation Data	It becomes problematic in sequential data, it become over fitting due to learn long term dependencies.

Proposed Work

For the detection of our proposed technique our system comprises the following steps: Exploratory Data Analysis, Pre-processing and feature extraction, Detecting anomalies with Z score, Isolation Forest, and DBSCAN separately then passing through our proposed ZID method through voting system and finally calculate anomalies with labelled in form anomalous and non-anomalous data in shape of 1 and 0. Then data shuffling process is carried out. After that train test method is processed for splitting data. We applied different machine learning algorithms including Random Forest, Gradient Boosting Machine, Support Vector Machine, Gaussian Naïve Bias and LDA (Linear Discriminate Analysis) is used for detection of anomalies in data set. However all of algorithms produce reasonable results with quite good accuracy but Linear Discriminant Analysis (LDA) and Random Forest is not effecting in detection anomalies of E-commerce Data. Furthermore our proposed Gaussian Naïve Baise (GBM) detect anomalies with higher accuracy and precision and achieve best results in anomaly detection system. Our proposed system effectively detect false positive. The proposed methodology overview is shown in Figure 6.

PREPROCESSING

Selecting High Frequency Country

We first analyse type of currency in dollar. For that we count the number of countries in our dataset. In our data we have different countries ranging from C_1 to C_n as described in Eq. 1. Each country have transaction from t_1 to t_n as shown in Eq. 2. We have to calculate the percentage of each country i.e. P_c in which how much the transaction has been done by each countries as described in Eq. 3. In Exploratory data analysis and by using Eq. 3, 91.43 % transactional data is done by United Kingdom so we choose the type of currency in dollar. We ignore the other transactional data done by different countries as they are very less ranging from 0.1 to 1 % by using Eq. 4. Detect anomalies in unit price we are concern in dollar because 91 % of dataset contain United Kingdom so we opt dollar price.

$$C(x) = \sum_{i=1}^{C} (C_i, C_{i+1} \ldots . C_n) \qquad \text{Eq.1}$$

where $C(x)$ is countries in table

$$T(x) = \sum_{i=1}^{T} (T_i, T_{i+1} \ldots . T_n) \qquad \text{Eq.2}$$

where $T(x)$ is transactions in table

$$P_c = \frac{\{T(x) \partial C(x)\} * 100}{length(x)} \qquad \text{Eq.3}$$

Where $Percentage\, of\ Country = P_c$

$$T\left(x\right)=\begin{Bmatrix} 1 if\ P_c \geq 90\% \\ 0\,otherwise \end{Bmatrix} \quad \begin{pmatrix} keep=1 \\ ignore=0 \end{pmatrix} \text{Eq.4}$$

Remove Null Values

In second step of pre-processing we drop null values from our data set.

Drop Duplicate Values

In third step of pre-processing we drop duplicate rows in our data set.

Feature Extraction on EDA

We analyse the dataset in which country feature contains 91% data of United Kingdom. In fact our main target is to detect anomalies in price so to get rid of price conflict between different currencies we opt only UK Dollar that is 91% of whole data set. After exploratory data analysis we selected the features of stock code, stock description and price.

FEATURES SELECTION

In data set we have selected the features (FS) of stock code, description unit price and country. Here, the two categorical of qualitative variable i.e. stock code and description. So we used multiple corresponding analysis which are used to transform the qualitative variable as shown in table 3. In our study we're using "prince" using sklearn library of python for MCA model [55]. Component parameter is 2, iteration is 3, engine = auto and all remaining parameters are default parameter of sklearn. Symbols and descriptions are shown in table 2.

Table 2. Symbol and their Description

Symbols	Descriptions	Explanation
d	Data points	Referring to rows of data
D	Dataset	Entire collection of data
FS	Feature Selection	Stock code and description for MCA transformation
argmax	Argument Max	Variance in categorical data
F*	Optimal Subset	Stock code , description and price and possible country

Table 3. Transform Values of Categorical variable based on MCA

Stock code	Description	Unit Price	MCA 1	MCA 2
85123A	white hanging heart t-light holder	2.55	1.282533	-0.373759
71053	white metal lantern	3.39	-0.416247	-1.067324
84406B	cream cupid hearts coat hanger	2.75	-0.183455	-1.136621
84029G	knitted union flag hot water bottle	3.39	0.169823	-0.671747
84029E	Red woolly hottie white heart	3.39	-0.251174	0.906409

Let D as Dataset where each data points d belongs to D, so we maps the originals set of features to selected features (FS). Let denote this function as FS.

$$FS = D \rightarrow F \qquad \text{Eq.5}$$

Optimal subset F* that maximize certain objective function such information gain or classification accuracy. The optimal subset F^* can be determined as.

$$F^* = argmax\left(FS(D)\right) \qquad \text{Eq.6}$$

Before transformation our main task is to perform is pre-processing before analysing and identifying anomaly detection techniques that is standardizing. For standardizing data set if mean μ is 0 and its standard deviation σ is 1.

Many machine learning model works well on that data when features of that data is relatively similar scale or closely related to normal distribution. For data normalization we use standard scalar in standard scalar subtracting the mean and then scaling   a unit difference. Unit difference mean dividing all the value by standard deviation.

Thus, let D be the dataset and μ the mean of D and σ the standard deviation. The result of standardization is $x_{standardization}$ to rescale the features value between 0 and 1. Then, to standardize D by using Eq. 7

$$x_{stand} = \frac{x - \mu}{\sigma} \text{for all } x \in D \qquad \text{Eq. 7}$$

In research article [56] observed in their research support vector machine is good when data is to be standardized. Also other technique which are computed in this like isolation forest, random forest and K-NN benefit from.Standardisation is more robust to anomalies, and in many cases, it is good and preferable over normalization. Standardization is not equal to normalization, where D is changed, such that:

$$x_{stand} \text{ belongs } [0,1] \text{ for x ò D} \qquad \text{Eq.8}$$

After the transformation process has been done, we have train the data based on stock code and description to detect anomalies in price using unsupervised methods like isolation forest and DBSCAN in next session.

ANOMALY DETECTION FUNCTIONS

Calculating Z-Score

A z-score is defined as position of some raw score from distance its means, when it is measured in standard deviation units [57]. The positive z-score is values lies above the means and negative z-score means it lies beyond the mean.

After selected features we group the data based on stock code and description using group function and transform function with mean to calculate the mean of group data and similarly calculate standard deviation in this manner. After calculating means and standard deviation we calculate the Z score of ith feature i.e. price by using Eq. 9 and 10.

$$Z = \frac{x - \mu}{\sigma} \qquad \text{Eq.9}$$

$$Where\ Z = Z\ score$$

$$x = ith\ feature\ Score\ i.e UnitPrice$$

$$\mu = Means$$

$$\sigma = Standard\ Deviation$$

$$Prediction_{Zscore}(x) = P_z \qquad \text{Eq.10}$$

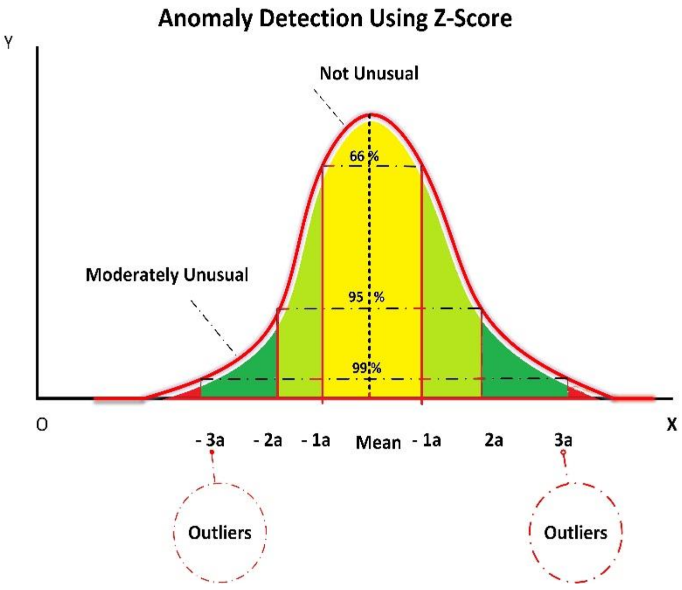


Figure 3. Z-Score graphical visualization

In this scenario we group data based on stock code and description. We calculate Z-score of each unit price of group data items. We setup threshold many time to adjust where anomalies actually exists, finally we setup a threshold 2 and -2. We have seen that an item having stock code " D ", where D means Discount and description "Discount" average historical prices 5 $ to 30 $ but suddenly price fluctuate from average historical prices i.e. 1867 $ as describe in table 4. This can cause be system error and other reason but our system detected anomaly as described in table 4.

$$Z = \frac{x-\mu}{\sigma} \qquad \text{Eq.9}$$

$$Where\ Z = Z\ score$$

$$x = 1867.86$$

$$\mu = 67.89$$

$$\sigma = 221.845$$

$$Prediction_{Zscore}(x) = P_z$$

In this scenario we have noted that an item having stock code " 20685 " and description "Doormat Red Retro spot" average historical prices 4 $ to 7 $ but suddenly price fluctuate from average historical prices i.e. 15.79 $. This can cause be system error and other reason but our system detected anomaly as shown in table 5.

Table 4. Calculation of Z-Score of an item D

Row ID	Stock Code	Description	Unit Price	Z-Score	Outlier Condition
479867	D	Discount	14.88	-0.238967	0
479868	D	Discount	19.82	-0.216699	0
479869	D	Discount	16.76	-0.230492	0
516221	D	Discount	20.53	-0.213498	0
516455	D	Discount	28.68	-0.176761	0
479867	D	Discount	14.88	-0.238967	0

479868	D	Discount	19.82	-0.216699	0
479869	D	Discount	16.76	-0.230492	0
150591	D	Discount	1867.86	+8.11358	1

Table 5. Calculation of Z-Score of an item Doormat Red Retro spot

ID	Stock Code	Description	Unit Price	Z-Score	Outlier Condition
454	20685	Doormat Red Retro spot	7.95	-0.111841	0
2124	20685	Doormat Red Retro spot	6.75	-0.453465	0
191466	20685	Doormat Red Retro spot	4.58	-1.071235	0
304634	20685	Doormat Red Retro spot	8.00	-0.097607	0
320829	20685	Doormat Red Retro spot	8.25	-0.026435	0
334085	20685	Doormat Red Retro spot	7.08	-0.359519	0
429598	20685	Doormat Red Retro spot	25.79	2.220103	1

Isolation Forest

Isolation tree is based on decision trees algorithm. The important thing about isolation forest is that is it measure some distances similar to K-mean algorithm as check whether it is anomalies in the data or not. The Isolation forest instead tries to isolates each data point and checks for data points that is different from other part of points in data set [58].

We first used multiple correspondence analysis to assign unique values of group data based on stock code and description. After assigning unique values using MCA model then we train our data and fitting data to detect anomalies using isolation forest. Isolation forest is one of mostly used and common recent model which was first proposed in 2008 and was later published in a paper 2012. Around 2016 it was taken incorporated in Python Scikit-Learn library. This isolation forest is tree based algorithm based on random forest and decision tree. In our data set the isolation forest model split the data into two part on the basis of random threshold value. The model process is continued until each data point is isolated completely. Furthermore when isolation forest model evaluate on complete data it filter the data points that is to be taken lessen step than other to be isolated. It returns score of each entity to detect anomalies in dataset by using Eq. 11. After detecting anomalies they are labelled with 1 and 0, 1 for anomalies and 0 for non-anomalous.

$$Prediction_{Isolation}(x) = P_i \qquad \text{Eq. } 11$$

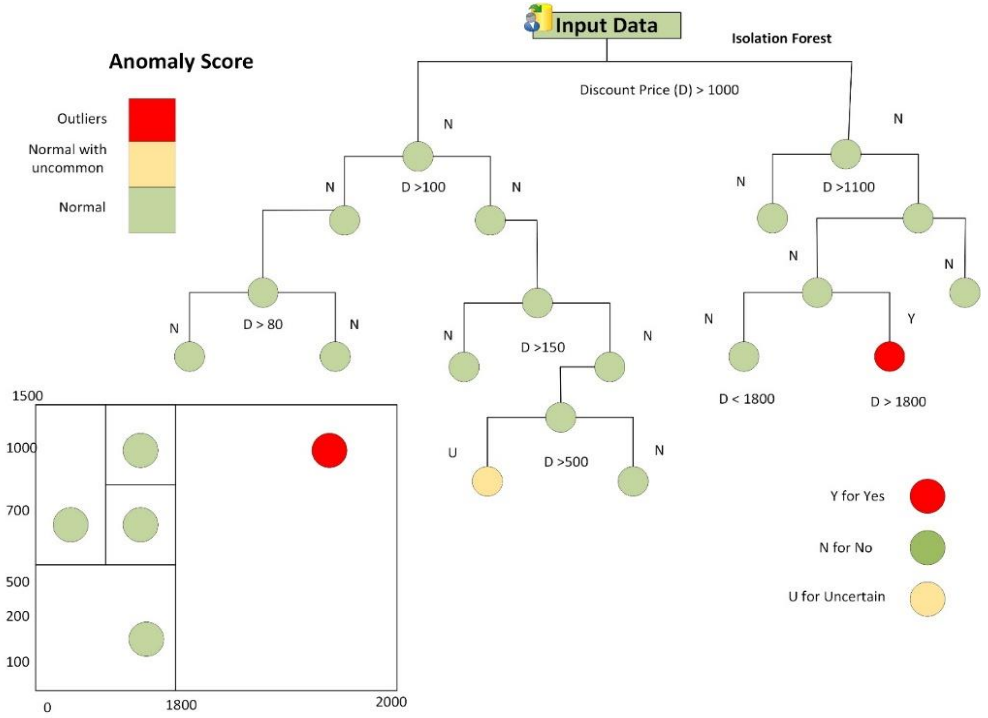


Figure 4. Isolation Forest graphical visualization and how it works in price anomaly

DBSCAN

DB-SCAN stands for density based Spatial Clustering of Application with noise. Densely grouped data are grouped into a single cluster. It is robust to outlier, this is very interesting feature DBSCAN clustering algorithm. In k-mean we have to tell the number of centroids beforehand, but in DBSCAN there is no need to tell number of clusters in advance [59]. It requires two parameter: one is epsilon and second one is min-points. Epsilon is the radius of circle that is to create for each data point whether its dense or and min-points is minimum number of data points which require inside of circle for data points to be classified as Core Point.

DBSCAN requires only two parameters: epsilon and min-points. Epsilon is the radius of the circle to be created around each data point to check the density and min-points is the minimum number of data points required inside that circle for that data point to be classified as a Core point.

In this algorithm, the category can be regarded as the sample dense area divided by the sample low-density area in the data space. Therefore, it can be used to detect anomalies in data samples by using Eq. 12.

$$Prediction_{Dbscan}\left(x\right)=P_{Db} \qquad \text{Eq.} \quad 12$$

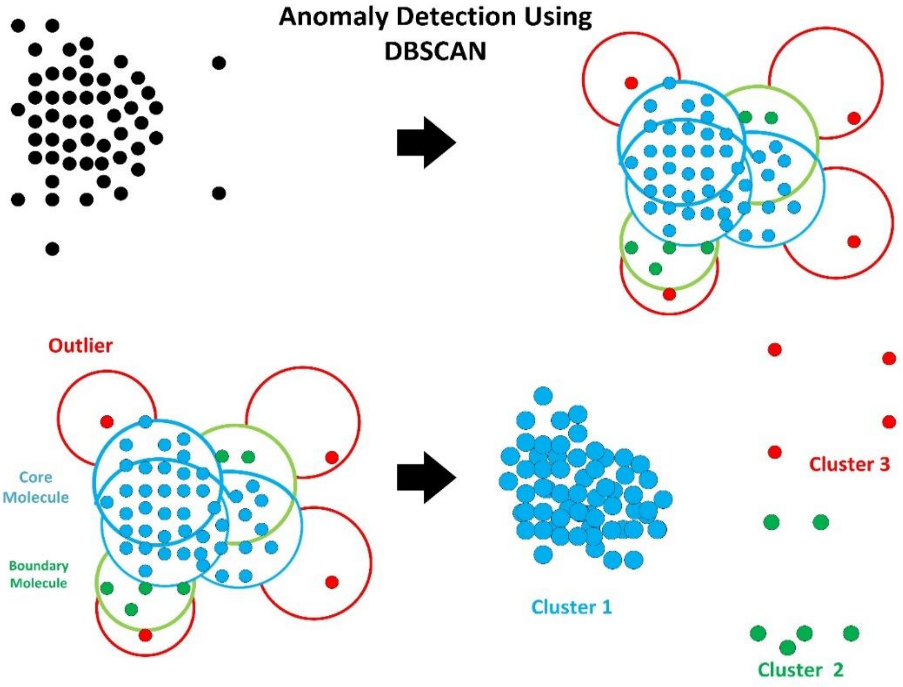


Figure 5. Example of DBSCAN graphical visualization

THRESHOLD SETTING

We calculate Z-score of each unit price of group data items. We setup threshold many time to adjust Where anomalies actually exists, finally we setup a threshold 2 and -2. We have seen that an item having Stock code" 2121 "and description "Vintage Union Jack Bunting" average historical prices 7 and 8 $ But suddenly price fluctuate from average historical prices i.e. 16.63 $. This can cause be system error And other reason but our system detected anomaly. Similarly we grouped data in same way to make Cluster for isolation forest and DBSCAN to detect anomalies by using Eq. 13 and 14.

$$Prediction_{Zscore}(x) = P_z(x)$$

$$Prediction_{Isolation}(x) = P_i(x) \qquad Eq.\ 13$$

$$Prediction_{Dbscan}(x) = P_{Db}(x) \qquad Eq.\ 14$$

$$Vote\ V = V(x)$$

$$Set\ S \in \{0,1\}$$

$$(P_z, P_i, P_{Db}) \in S$$

$$Let\ Zero\,Count = Z_c \quad and\ One\,Count = Z_o$$

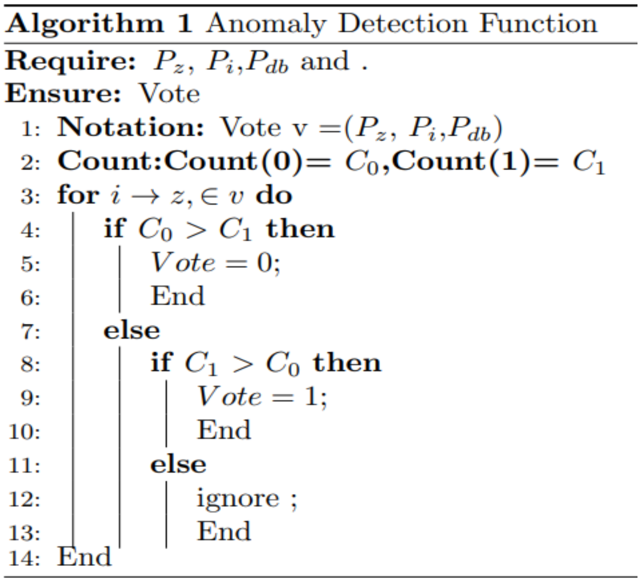
$$V(x) = \sum_{P=1}^{P} (P_z, P_i, P_{Db})$$

$$V\left(x\right)=1\ if\sum_{P=1}^{P}\left(P_z,P_i\ ,P_{Db}\right)\geq Z_c \qquad Eq.\ 15$$

$$V\left(x\right)=0if\sum_{P=1}^{P}\left(P_z,P_i\ ,P_{Db}\right)\geq Z_o \qquad Eq.\ 16$$

PROPOSED ZID METHOD TECHNIQUE

Detecting anomalies with Z score, Isolation Forest, and DBSCAN separately then passing through our proposed ZID method through voting system are defined in algorithm 1 and table 6. And finally calculate anomalies with labelled in form anomalous and non-anomalous data in shape of 1 and 0 by using Eq. 15 and 16



Algorithm 1 Anomaly Detection Function

```
Require: $P_z$, $P_i$,$P_{db}$ and .
Ensure: Vote
Notation: Vote v =($P_z$, $P_i$,$P_{db}$)
Count:Count(0)= $C_0$,Count(1)= $C_1$
for $i \rightarrow z, \in v$ do
    if $C_0 > C_1$ then
        $Vote = 0$;
        End
    else
        if $C_1 > C_0$ then
            $Vote = 1$;
            End
        else
            ignore ;
            End
End
```

Table 6. Working of ZID Voting System

Z-Score	DBSCAN	Isolation Forest	Proposed ZID voting System
0	1	1	1
0	0	1	0
0	1	0	0
1	1	1	1
0	0	0	0

Once the anomalies has been detected using proposed ZID technique system then we applied different machine learning algorithms in order to compute evaluation metrics.

Random forest is combination of decision trees structure. The risk of over fitting is handle and reduced because it is ensembles of decision tree. Random forest is able to handle non-linearity's and feature connections because it is not require feature scaling [60]. As decision trees, random forest also handle categorical data and extend to multi class classification setting. We generate several trees instead of a single tree due to this reason we used random forest algorithm which gives quite good accuracy results with large numbers of trees. Random forest randomly opt features and construct multiple decision trees upon the characteristic of dataset. In our scenario we have classification problem the random forest uses majority voting system. The prediction of each tree is treated as one vote for one class. It can be seen that the label will the class that get gains most votes. The random forest consist of several trees and it makes a prediction based on averaging prediction of each node of tree. This almost have better predictive model for accuracy than single decision tree algorithm. Random forest model gave better result and work efficient with default parameter. Deals well with multiple features that may associate. It is less variance then an ordinary decision tree. Default parameter of sci-kit learn library i.e. is No of tree=42, max depth is until all leaf are pure, best split is sqrt e.g. sqrt (# features) default is auto, criterion='entropy' by default is 'gini'.and all remaining parameter is default. Random Forest performs accurately for large number of datasets as it is flexible. On the other hands its computational cost is really much high.

Gaussian Naïve Bayes is probalistic model [61]. In our data set, the given observation it measure a probability distribution over set of classes. Conditional or posterior probility of any event is handle by Gaussian Naïve Bayes theorem. The Equation of Gaussian Naïve Bayes model is defined by equation 17.

$$P\ \frac{A}{B} = P\frac{B}{A} \times \frac{P\ (A)}{P(B)} \qquad \text{Eq.}\ \ 17$$

Naïve Bayes is easier to implement, and computing capacity or power of Naïve Bayes is so fast. Furthermore, it works well with high dimensional data. It depends stringently 20 on individualism molds and if that supposition is not met accurately this model will perform badly. Default parameters of the Naïve Bayes model are used using sklearn library of python.

Gradient boosting machine is part of machine learning boosting. It is very powerful technique for constructing nominative model [52]. Gradient boosting machine train many model in gradually and sequential fashion. Standard parameters of the Gradient Boosting model are used using sklearn library of python.

Linear Discriminant analysis is simple and efficient technique for determining of classification [62]. As it well understood easily and simple to learn, so it has a lot of variation and extensions. Default parameters of the Linear Discriminant analysis model are used using sklearn library of python

SVM Works efficiently for high dimensional data due to power of kernel trick [63]. Whereas, to choose the accurate kernel is a real trick. SVM works better on the data where dataset is larger as compared with the dimensions of the data. In our study we're using "linear" kernel using sklearn library of python for SVM model. Legalization parameter is 1.0, kernel is linear default kernel is rbf, and all remaining parameters are default parameter of sklearn.

The detection of anomalies in E-commerce need a lot of testing related to large amount data. Automatic detection method for E-commerce data usually depend on feature like size, shape, intensities. For the detection of our proposed technique our system comprises the following steps: Exploratory Data Analysis, Pre-processing and feature extraction, Anomaly detection functions including: 1) Calculating Z-score of group's items and set up threshold. 2) Detecting anomalies using Isolation Forest, 3) Detecting anomalies with DBSCAN, Proposed ZID voting system technique, data labelling, train test split methods and finally machine learning techniques including Random Forest, Gradient Boosting Machine, Support Vector Machine, Naïve Bias and LDA (Linear Discriminate Analysis) is used for detection. However all produce reasonable results with quite good accuracy but LDA and Random forest is not effecting in detection of E-commerce Data.

In this research, for anomalies detection in product pricing of E-commerce data set, we apply trained and tests model with enough data for automatically detection of anomalies.

The proposed method has the following key points:

1. Feature selection is first step that is carried out by extracting the price features by EDA. In pre-processing step we handle missing values and duplicate values.
2. Anomalies are detected using Isolation forest, DBSCAN and Z-score.
3. Then we proposed ZID voting scheme method to detect finalize anomalies according Isolation forest, DBSCAN and Z-score.
4. After that we label the data and shuffling or reset data process is carried out. To classify a machine learning approach applied for training and testing models for accurate detection.
5. We compare the results of our trained model with rest of the machine learning techniques to observe the good performance.

DATA LABELLING

In anomaly detection process, we have labelled data in 1 and 0 form when we are detecting anomalies using Z-score, the same procedure can be followed by DBSCAN and isolation forest and finally data is labelled in 1 and 0 (anomalous and non- anomalous) in proposed ZID voting system.

DATA SHUFFLING

In manual inspection, we have found that anomalous data 1 and non-anomalous data 0 are in form of sequences. We have shuffle the data to avoid conflict in sequence of 0 and 1. So we reset the data after labelling process.

TRAIN TEST SPLIT

To classify a machine learning approaches applied for training and testing models for accurate detection.

1. Training

For classification we use our training model to classify our data. Here are the some of the following steps in training.

i. Load the dataset.

ii. After labelling the data in term of true and false (1 and 0) labelled data set also loaded.

iii. 80% data used for training.

iv. For classification different machine learning techniques i.e. Random Forest, Gaussian Naïve Bias, LDA, SVM and proposed Gaussian Boosting machine , architecture used.

2. Testing

To evaluate the efficiency of the training model, we pass the test data that is 20 % of the whole data to the trained model and by employing accuracy measure, we evaluate the performance of trained data on test data. We can learn that how good a models trained to classify anomalies and non-anomalies. Our Proposed Gaussian Boosting machine architecture achieve higher Precision and accuracy with better results.

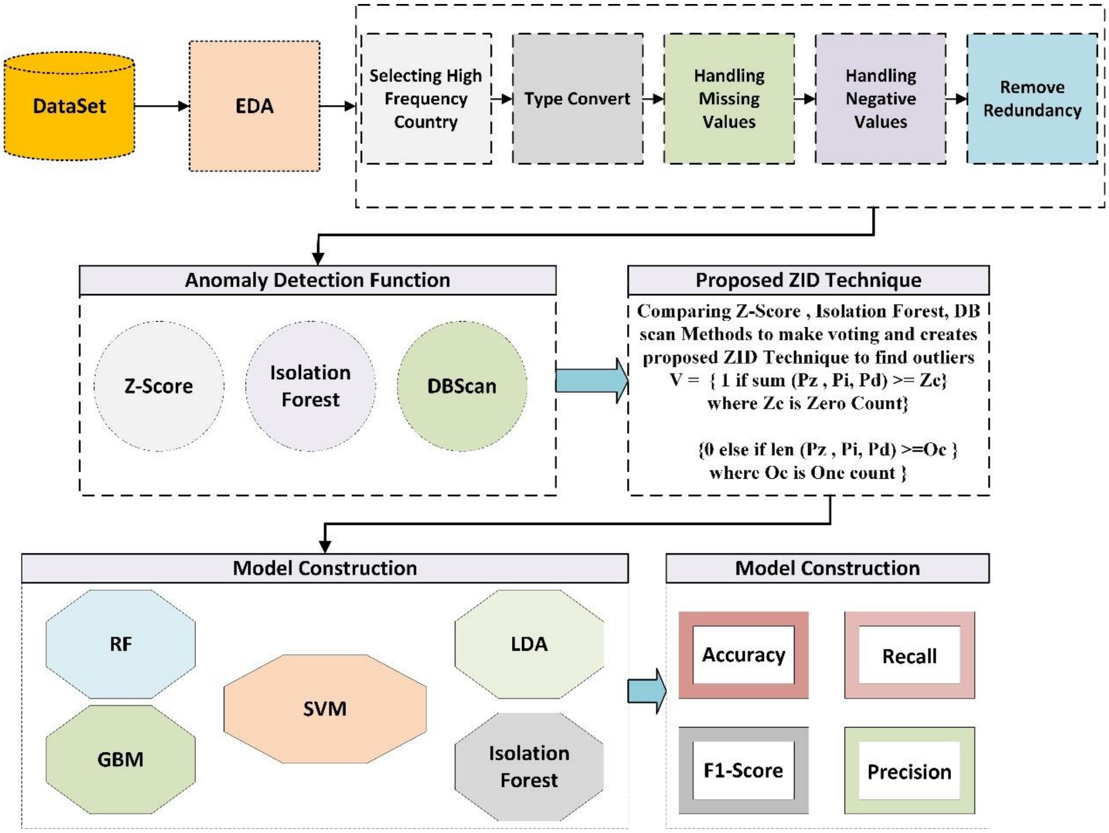


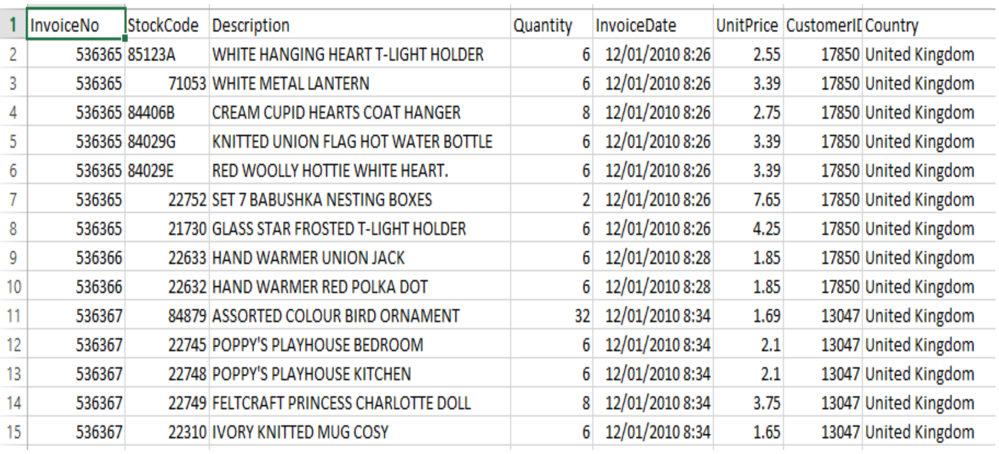
Figure 6. Proposed Model Framework

This research, we deployed machine learning -based model in which performance of our proposed model is best when analyzed for detecting anomalies in data. We implemented and evaluated many observation and experiments to gain the best accuracy results and to deployed the optimize approach for anomaly detection system. We evaluated our test and experiments on E-commerce dataset, in which anomalous data is considered as 1 labeled and non-anomalous data is consider as 0 labeled. Thus we experiment machine learning models on E-commerce dataset to detect outlier or anomalies. We analyzed best fit parameters for our model. Furthermore, upon this model, we will run our models on testing data to calculate evaluation results.

EXPERIMENTAL SETUP AND RESULT

Dataset Description

The proposed work is designed for detecting anomalies. To carry out this task, we used E-commerce dataset. The dataset contains invoice number, stock code, description, quantity, invoice date, unit price, customer ID, and country. The dataset contained 541909 data. First 15 rows are shown in Figure 7.



1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/01/2010 8:26	2.55	17850	United Kingdom
3	536365	71053	WHITE METAL LANTERN	6	12/01/2010 8:26	3.39	17850	United Kingdom
4	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/01/2010 8:26	2.75	17850	United Kingdom
5	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/01/2010 8:26	3.39	17850	United Kingdom
6	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/01/2010 8:26	3.39	17850	United Kingdom
7	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/01/2010 8:26	7.65	17850	United Kingdom
8	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/01/2010 8:26	4.25	17850	United Kingdom
9	536366	22633	HAND WARMER UNION JACK	6	12/01/2010 8:28	1.85	17850	United Kingdom
10	536366	22632	HAND WARMER RED POLKA DOT	6	12/01/2010 8:28	1.85	17850	United Kingdom
11	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/01/2010 8:34	1.69	13047	United Kingdom
12	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	12/01/2010 8:34	2.1	13047	United Kingdom
13	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	12/01/2010 8:34	2.1	13047	United Kingdom
14	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	12/01/2010 8:34	3.75	13047	United Kingdom
15	536367	22310	IVORY KNITTED MUG COSY	6	12/01/2010 8:34	1.65	13047	United Kingdom

Figure 7. First few rows of the dataset

Experimental setup

In this research we use anaconda3 by using Jupiter notebook for the implementation of this research. Seven different machine learning algorithms by using built in library of sklearn and keras. All experiments were run on an Intel (R) Intel(R) Core(TM) i5-3210M CPU @ 2.50GHz 2.50 GHz with installed 8 GB RAM, running on 64 bit Window Based Operating System. Python 3.7 with Numpy and Sklearn has been used for implementation of proposed model. Numpy stands for numerical python which is used for numerical.

Feature Subset Selection

Feature subset selection is the selection of relevant attributes that are going to be used in ML models. Main objective of feature subset selection is as given below;

1. Selection of best features help to train the ML model fast.
2. It makes the interpretation easy as best features help to minimize the complexity.
3. Right subset selection of features highly improves the performance of ML model.
4. It also minimizes over fitting

Evaluation Metrics

Accuracy represents the amount of TN (true negative) as well as TP (true positive) samples total number of samples and it is computed by using Eq. 18. Precision represents the proportion of predicted positive cases that are real positives and it is evaluated by Eq. 19. While recall is the proportion of actual positive cases that were correctly predicted as such and evaluated by Eq. 20. The harmonic mean of Precision and recall is known as F-measure and is calculated by Eq. 21

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Eq. 18}$$

$$\Pr ecesion = \frac{TP}{TP + FP} \qquad \text{Eq. 19}$$

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \qquad \text{Eq. 20}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad \text{Eq. 21}$$

Table 7. Standard evaluation metrics

Metrics	Description
True Positive (TP)	Are anomalous that are correctly classified as anomaly.
True Negative (TN)	are non-anomalous that are correctly classified non-anomalous samples
False Positive (FP)	are non-anomalous that are misclassified anomalous data
False Negative (FN)	are anomalous that are misclassified non-anomalous data

Validation and Testing

This section will validate our tests on E-commerce dataset. We applied different unsupervised machine learning models to end with anomalies are detected. Anomalies are detected using Z-score, Isolation forest and DBSCAN. After the deployment of three anomaly detection function we proposed ZID voting system to detect anomalies in efficient manner. For training we have used 80 percent of dataset while 20 percent data is used for testing.

Result and Discussion

We have detected anomalies using Z-score, Isolation Forest, and DBSCAN. Few anomalies are detected using Z-Score because it's statistical approach for anomaly detection. When we moved towards Isolation forest and DBSCAN, both machine learning technique detect anomalies accurately. We have found that some anomalies are not anomalous so we proposed ZID Voting System to create voting system to finalize anomalies in data. Details of anomalies are described in Table 8.

Table 8. Detection of anomalies using unsupervised and statistical approach

Z-Score	DBSCAN	Isolation Forest	Proposed ZID voting System
41	1075	1823	886

After the detection of anomalies, we have described the anomalous data and non-anomalous data in table 9. We have implemented Z-Score to find anomalies in price feature based on stock Code and description. In previous we have reported each stock and description. Now we have reported some of anomalies how actually anomaly exist that are stock code 'D' , '20685' ,'23243' and 21621. There are 886 anomalies, some of anomalies having different stock and description is shown in table 9 in order to ease for understanding. We have calculated the mean and standard deviation and set threshold 2 and -2 to calculate Z-score, if the greater than and equal to 2 and -2 classify as manual inspection. Table 10 shows 4 anomalies of different stock with description.

Table 9. Anomalous data and non-anomalous data

Anomalous Data	Non- Anomalous data
866	4,10800

Table 10. Statistical approach Anomaly Score for Price using Z-score

Stock Code	Description	Unit Price	Average Historical Price Range $	Z-Score	Outlier Condition
D	Discount	1867.86	14-281	+8.11358	1
20685	Doormat Red Retro spot	15.79	6-8	2.120103	1
23243	Set Of Tea Coffee Sugar Tins Pantry	1.71	5-6	-2.124714	1
21621	Vintage Union Jack Bunting	16.63	6-8	2.002533	1

In this research, when we implement the DBSCAN, the closely bounded region are non-anomalous data and separately bound region or isolated points are anomalous data. The graphical visualization of DBSCAN is described in Figure 8.

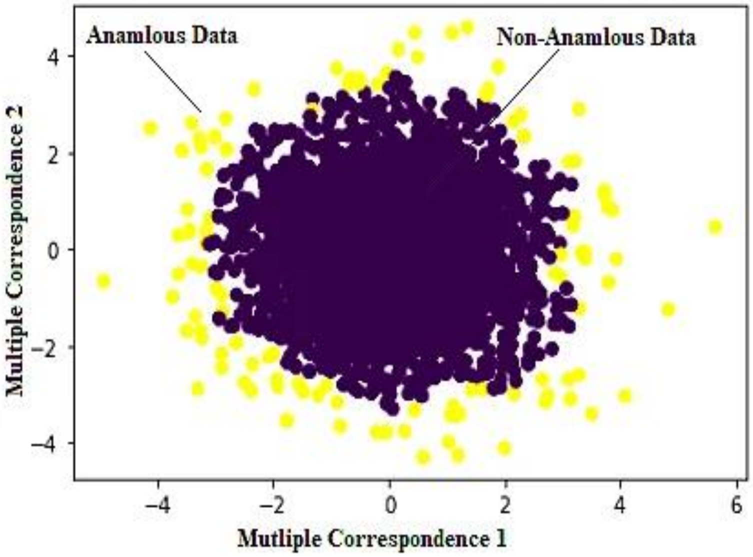


Figure 8. Anomaly Score and Outlier Region for Price using DBSCAN

When we deployed the isolation forest using unsupervised approach, the closely congested or bounded region are non-anomalous data and separately bound region or isolated points are anomalous data. The graphical visualization of isolation forest is described in Figure 9.

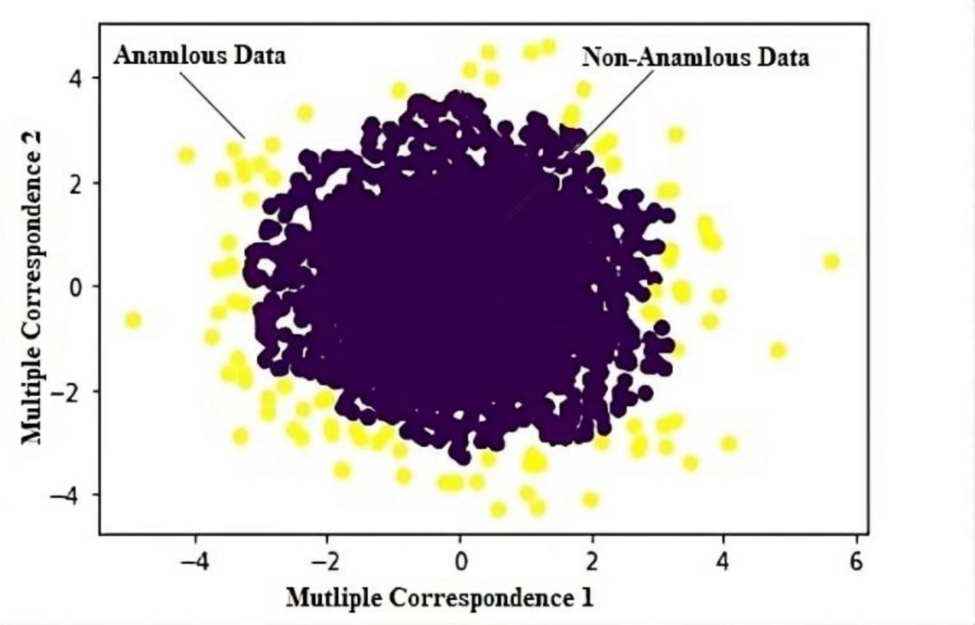


Figure 9. Anomaly Score and Outlier Region for Price using Isolation Forest

After the deployment of three anomaly detection function we proposed ZID voting system to detect anomalies in efficient manner. The graphical visualization of ZID voting system is shown in Figure 10.

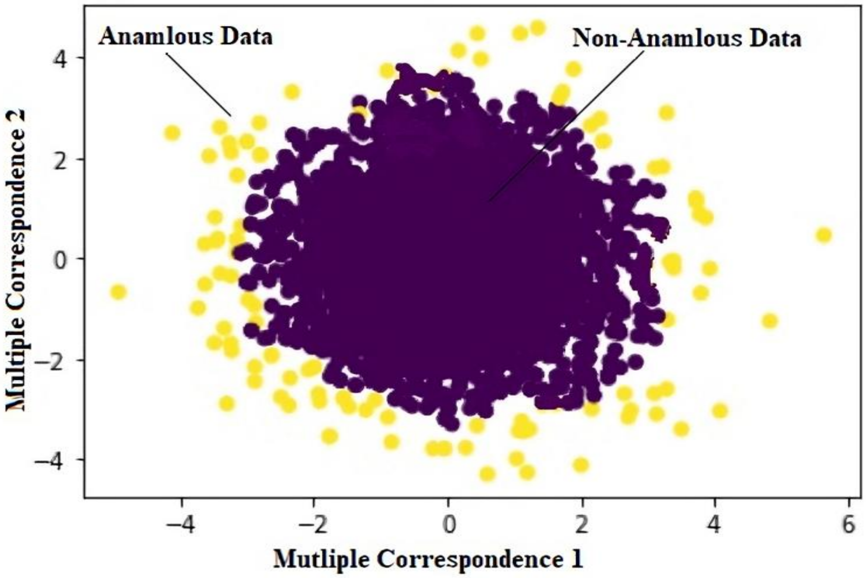


Figure 10. Anomaly Score and Outlier Region for Price using ZID voting system

The stated values are the average scores of all five runs in terms of Precision, recall, F1 and accuracy for both classes i.e. anomalous and non-anomalous. Results for the anomaly detection suggest that among all baseline classifiers, gradient boosting machine had the best performance on features dataset in terms of recall and F1 score, while Random Forest performed the best in terms of Precision and accuracy. This is consistent with the results reported in [64]. Table 11 and 12 also shows that Gradient Boosting Machine outperformed in terms of Precision and accuracy.

Table 11 and 12 shows that our proposed features had the best overall performance for anomalies detection, compared to baseline features classifiers in terms of Precision, and accuracy. These results suggest that our proposed model outperformed all baseline classifiers, including the base line features, in detecting anomalies in E-commerce dataset.

Table 11. Comparative results of anomaly Detection in terms of Precession (P), Recall (R) and F1 Score

Classifier	Baseline Anomaly Result			Proposed Anomaly Result		
	P	R	F1	P	R	F1
Random Forest	0.94	0.94	0.94	0.89	0.90	0.89
GNB	0.92	0.95	0.94	0.92	0.92	0.91
LDA	-	-	-	0.86	0.84	0.77
SVM				0.92	0.92	0.91
GBM	-	-	-	0.97	0.86	0.91

Table 12 Anomaly detection in term of Accuracy Results

Classifier	Proposed Results Accuracy
GNB	0.9222
Random Forest	0.8982
LDA	0.8547
SVM	0.9225
GBM	0.9751

We implemented and deployed a best approach to find anomalies detection in E-commerce. In this research we have compare our results with different existing approaches. Our proposed approach can be used to detect anomalies E-commerce data set. We take benefit of gradient boosting machine model and thus achieved Precision better than previous methods. Figure 11 presents the Precision of implemented models. We have implemented different supervised machine learning algorithms to obtain the Precision by using Eq. 19. Our proposed GBM model got higher Precision as compared to existing approaches.

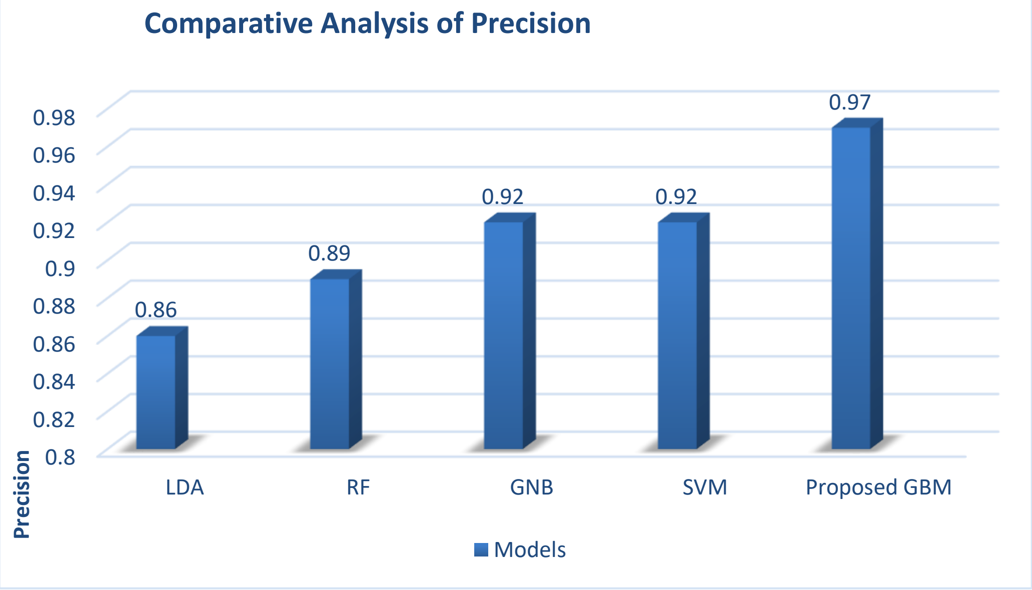


Figure 11. Comparative Analysis of proposed GBM with other model in term of Precision

Figure 12 presents the recall of implemented models. We have implemented different supervised machine learning algorithms to obtain the recall by using Eq. 20.

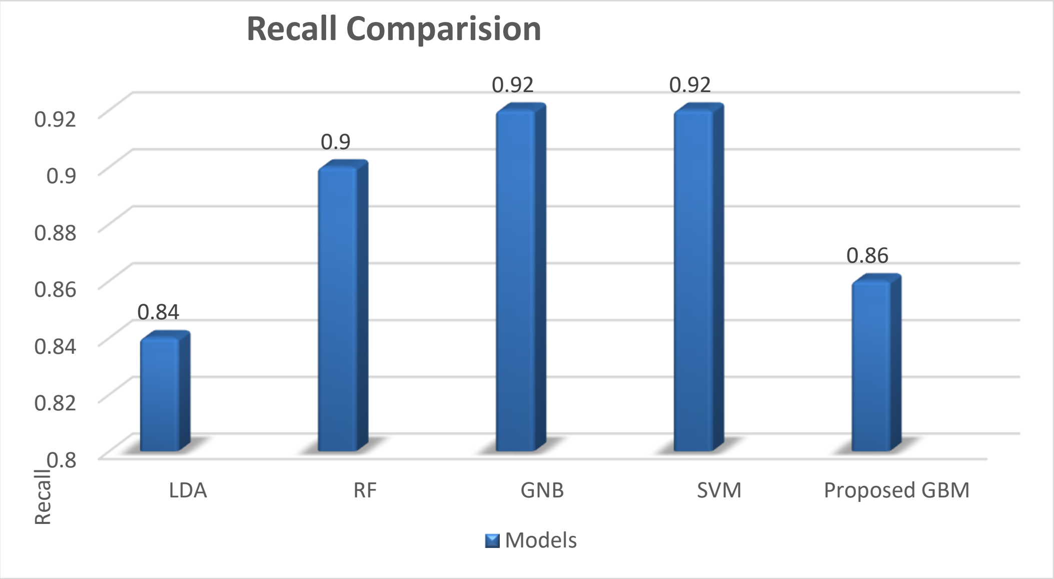


Figure 12. Comparative Analysis of proposed GBM with other approaches in term of recall

Figure 13 presents the F1-score of implemented models. We have implemented different supervised machine learning algorithms to obtain the F1-score by using Eq. 21.

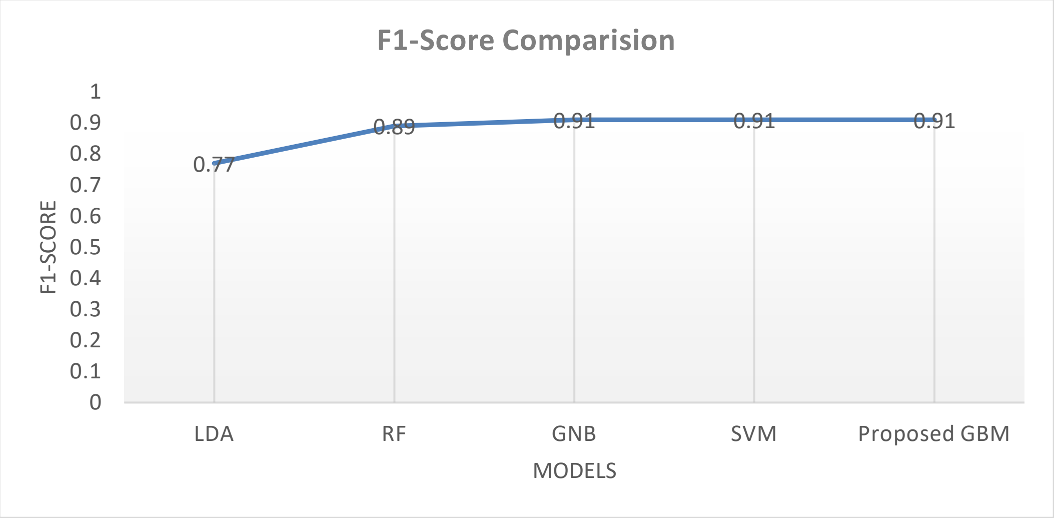


Figure 13. Comparative Analysis of F1-Score

Figure 14 presents the accuracy of implemented models. We have implemented different supervised machine learning algorithms to obtain the accuracy by using Eq. 18. Our experimental results on E-commerce dataset, as we noticed that the Precision of our proposed GBM model is 0.97 % and accuracy is 0.9751 % which is better than that of previous and existing approaches.

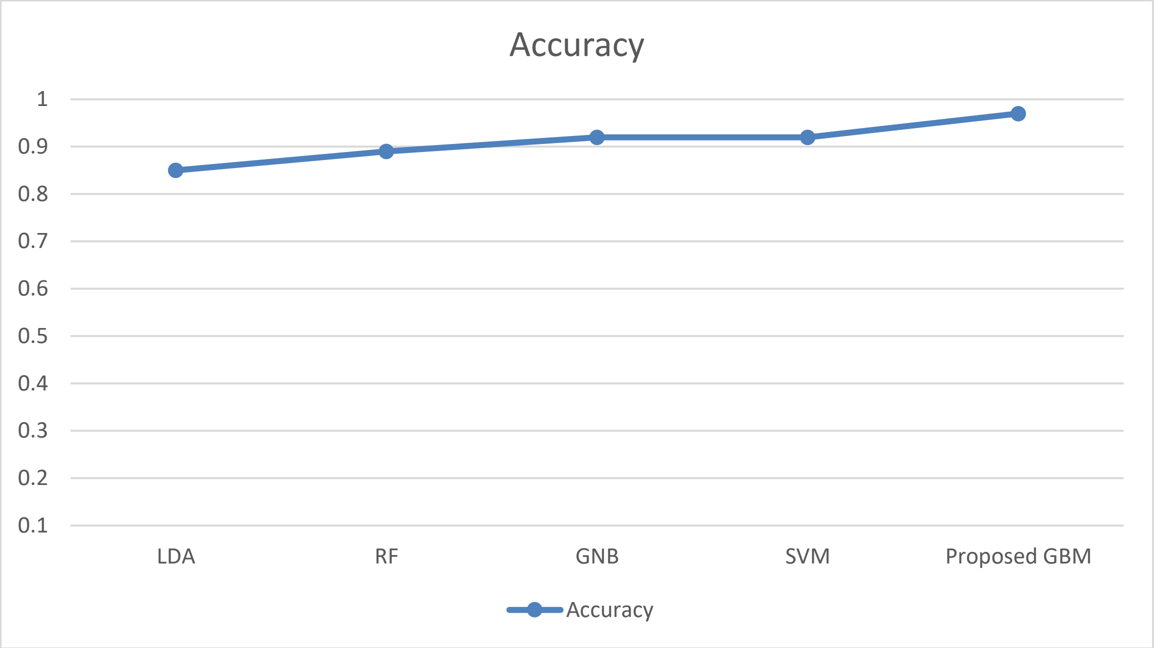


Figure 14. Comparative Analysis of accuracy of proposed GBM with other approaches

Execution Time

In this research, we have deployed five supervised anomaly detection algorithms. All of them take less time to predict the result. The detail of training testing and running time of model is illustrated in table 13. SVM and gradient boosting machine which is easier to deployed and implement however these two model take more time due to in every steps we have to calculate the polynomial function. Another drawback is that in during implementation of these models have to select the appropriate epsilon and measure F1 score. However the LDA has best time in term of training, testing and total running time. The detail description of running time is given figure 15.

Table 13. Training, Testing and Prediction times of anomaly detection model

Model	Training Time (s)	Testing Time (s)	Total Time
LDA	0.0154 seconds	0.0019 seconds	0.0173 seconds
GNB	0.0147 seconds	0.0116 seconds	0.0604 seconds
Random Forest	1.0478 seconds	0.0457 seconds	1.0935 seconds
SVM	1.2997 seconds	0.3523 seconds	1.6519 seconds
GBM	0.8956 seconds	0.0104 seconds	0.9060 seconds

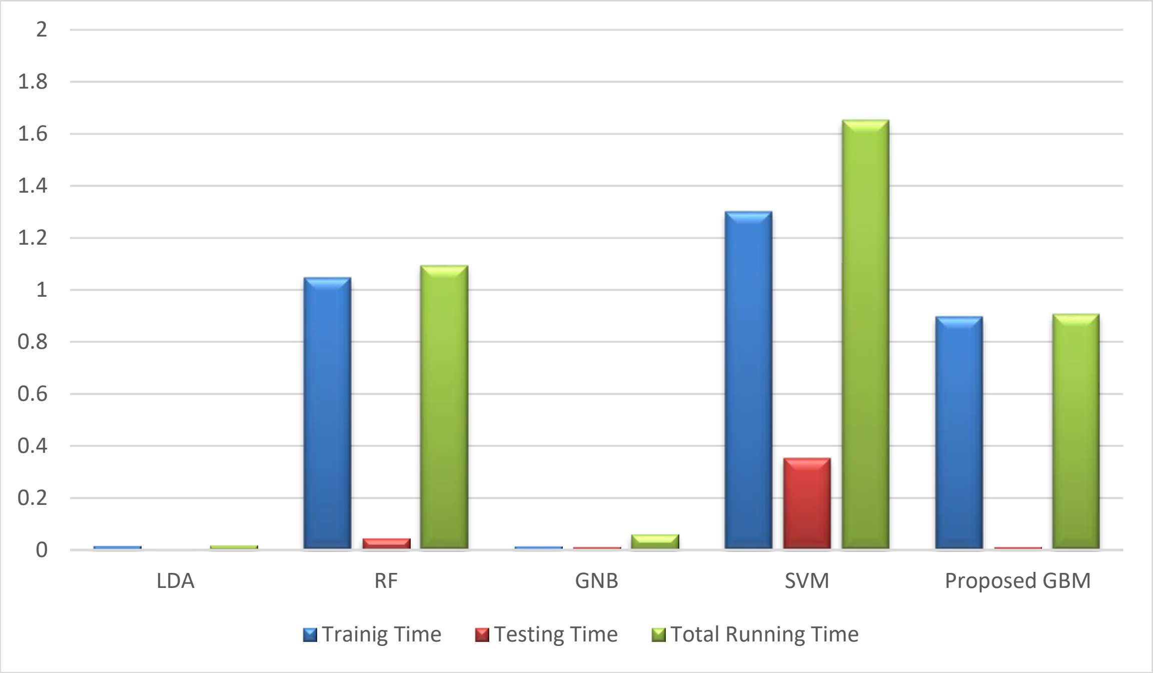


Figure 15. Running time of Models

CONCLUSION

Detecting anomalies in real time data set like E-commerce and retail store is a major challenge and complicated task for machine learning. In this research, we have deployed a different machine learning techniques to detect anomalies. We implemented Isolation Forest, DBSCAN, and Random forest, Gradient Boosting Machine, Gaussian Naïve Bayes, SVM and LDA. We take advantage of proposed model Gradient Boosting Machine to get and achieve and best Precision than existing methods. Our technique proposed Gradient Boosting Machine compared with other techniques evaluated in terms of Precision is 0.97, and accuracy is 0.9751 which is higher than existing methods and the highest on the benchmark. Our work on detecting anomalies using machine learning models performed efficient in very good scale. We gained and present better results than previous approaches. The proposed technique was implemented on E-commerce dataset. The major benefit of using machine learning is that data is to be taken as input and detecting anomalies as output which requires less preprocessing features and saves time by neglecting complicated and complex task. From past few years, real time and streaming data in E-commerce become more famous and got trend through different web platforms like Amazon, Daraz etc. The major reason beside this is, they are promising alternative to traditional web-search methods to sell and buy commodities. We will deploy in production of proposed Gradient Boosting Machine model, implemented models and deep learning models can be applied to the above- mentioned real time websites like Amazon, Daraz, and many others.

References

[1] M. Pincheira, M. Vecchio, and F. Antonelli, "An Adaptable and Unsupervised TinyML Anomaly Detection," 2023.

[2] Z. Wen, "Voucher Abuse Detection with Prompt-based Fine-tuning on Graph Neural Networks," pp. 4864–4870, 2019, doi: 10.1145/3583780.3615505.

[3] C. Poutré, D. Chételat, and M. Morales, "Deep Unsupervised Anomaly Detection in High-Frequency Markets," pp. 1–37, 2023.

[4] T. Delise, "Deep Semi-Supervised Anomaly Detection for Finding Fraud in the Futures Market arXiv : 2309 . 00088v1 [cs . LG] 31 Aug 2023," no. July, pp. 1–11, 2023.

[5] W. Wang *et al.*, "commerce orders," no. October, 2023, doi: 10.1117/12.3004538.

[6] M. Education, "Malicious url detection using machine learning," vol. 14, no. 02, pp. 537–552, 2023.

[7] A. Solehah, M. Taupit, and N. Azizan, "The Planning Process of the Online Transaction

Fraud Detection Using Backlogging on an E-Commerce Website," vol. 9, no. 1, pp. 56–62, 2023.

[8] Y. C A Padmanabha Reddy, P. Viswanath, and B. Eswara Reddy, "Semi-supervised learning: a brief review," *Int. J. Eng. Technol.*, vol. 7, no. 1.8, p. 81, 2018, doi: 10.14419/ijet.v7i1.8.9977.

[9] K. Shaukat *et al.*, "A Review of Time-Series Anomaly Detection Techniques: A Step to Future Perspectives," *Adv. Intell. Syst. Comput.*, vol. 1363 AISC, no. April, pp. 865–877, 2021, doi: 10.1007/978-3-030-73100-7_60.

[10] A. I. Tambuwal and D. Neagu, "Deep Quantile Regression for Unsupervised Anomaly Detection in Time-Series," *SN Comput. Sci.*, vol. 2, no. 6, pp. 1–16, 2021, doi: 10.1007/s42979-021-00866-4.

[11] G. Sandhya Madhuri and M. Usha Rani, "Review Paper on Anomaly Detection in Data Streams," in *Proceedings of the 2nd International Conference on Computational and Bio Engineering*, 2021, pp. 721–728.

[12] M. H. Fourati, S. Marzouk, and M. Jmaiel, "EPMA: Elastic Platform for Microservices-based Applications: Towards Optimal Resource Elasticity," *J. Grid Comput.*, vol. 20, no. 1, 2022, doi: 10.1007/s10723-021-09597-5.

[13] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised Anomaly Detection via Adversarial Training," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11363 LNCS, pp. 622–637, 2019, doi: 10.1007/978-3-030-20893-6_39.

[14] T. Barbariol and G. A. Susto, "TiWS-iForest: Isolation Forest in Weakly Supervised and Tiny ML scenarios," 2021, [Online]. Available: http://arxiv.org/abs/2111.15432

[15] Y. Ma and X. Zhao, "POD: a Parallel Outlier Detection Algorithm Using Weighted kNN," *IEEE Access*, pp. 81765–81777, 2021, doi: 10.1109/ACCESS.2021.3085605.

[16] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Recurrent EBM," *33rd Int. Conf. Mach. Learn. ICML 2016*, vol. 3, pp. 1742–1751, 2016.

[17] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: A literature review," *Expert Syst. Appl.*, vol. 197, no. April 2021, p. 116659, 2022, doi: 10.1016/j.eswa.2022.116659.

[18] M. Dancho, "Time Series Anomaly Detection - Lab18," *Data Sci.*, pp. 447–465, 2019, [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780128147610000137

[19] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2431 LNAI, pp. 15–27, 2002, doi: 10.1007/3-540-45681-3_2.

[20] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, 2017, doi: 10.1016/j.neucom.2017.04.070.

[21] M. Paolanti and E. Frontoni, "Multidisciplinary Pattern Recognition applications: A review," *Comput. Sci. Rev.*, vol. 37, p. 100276, 2020, doi: 10.1016/j.cosrev.2020.100276.

[22] M. Hu, Y. Wang, X. Feng, S. Zhou, Z. Wu, and Y. Qin, "Robust Anomaly Detection for Time-series Data," vol. XX, 2017.

[23] F. Xu and N. Wang, "Exploring Global and Local Information for Anomaly Detection with Normal Samples".

[24] Y. Zheng *et al.*, "From Unsupervised to Few-shot Graph Anomaly Detection: A Multi-scale Contrastive Learning Approach," vol. 14, no. 8, pp. 1–13, 2022, [Online]. Available: https://arxiv.org/abs/2202.05525

[25] A. Amellal, H. Seghiouer, and M. R. Ech-charrat, "Improving Lead Time Forecasting and Anomaly Detection for Automotive Spare Parts with A Combined CNN-LSTM Approach," vol. 16, no. 2, pp. 265–278, 2023.

[26] A. Alabrah, "An Improved CCF Detector to Handle the Problem of Class," 2023.

[27] C. Dong, "An Integrated System of Drug Matching and Abnormal Approval Number Correction," pp. 2–5.

[28] P. Bhattacharjee, A. Garg, and P. Mitra, "KAGO: an approximate adaptive grid-based outlier detection approach using kernel density estimate," *Pattern Anal. Appl.*, no. 0123456789, 2021, doi: 10.1007/s10044-021-00998-6.

[29] S. Ounacer, H. A. El Bour, Y. Oubrahim, M. Y. Ghoumari, and M. Azzouazi, "Using Isolation Forest in anomaly detection: The case of credit card transactions," *Period. Eng. Nat. Sci.*, vol. 6, no. 2, pp. 394–400, 2018, doi: 10.21533/pen.v6i2.533.

[30] H. Seo, S. Ryu, J. Yim, J. Seo, and Y. Yu, "Quantile Autoencoder for Anomaly Detection," 2020.

[31] D. Di, R. In, S. Statistiche, and C. Xxx, "Alma Mater Studiorum Università di Bologna with Quantile-based and other classi ers Coordinatore Dottorato :," 2018.

[32] E.-A. MINASTIREANU and G. MESNITA, "Light GBM Machine Learning Algorithm to

Online Click Fraud Detection," *J. Inf. Assur. Cybersecurity*, no. April, pp. 1–12, 2019, doi: 10.5171/2019.263928.

[33] M. Mca, "Credit Card Nearest Neighbor Based Outlier Detection Techniques," *Int. J. Comput. Tech. –*, vol. 5, no. 2, pp. 54–60, 2018, [Online]. Available: http://www.ijctjournal.org

[34] H. John and S. Naaz, "Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 4, pp. 1060–1064, 2019, doi: 10.26438/ijcse/v7i4.10601064.

[35] E. Klevak, S. Lin, A. Martin, O. Linda, and E. Ringger, "Out-Of-Bag Anomaly Detection," pp. 1–13, 2020, [Online]. Available: http://arxiv.org/abs/2009.09358

[36] R. Punmiya, O. Zyabkina, S. Choe, and J. Meyer, "Anomaly detection in power quality measurements using proximity-based unsupervised machine learning techniques," *2019 Electr. Power Qual. Supply Reliab. Conf. 2019 Symp. Electr. Eng. Mechatronics, PQ SEEM 2019*, no. June, 2019, doi: 10.1109/PQ.2019.8818236.

[37] A. Jhamb, D. Kumar, H. Chauhan, and E. Roorkee, "Robust and Unsupervised Anomaly Detection for Multivariate Dataset," pp. 3928–3937, 2020.

[38] V. Oleg, "Unsupervised anomaly detection in merchant vessel data Unsupervised anomaly detection in merchant vessel data Oleg Vlasovets," 2020.

[39] F. Tony Liu, K. Ming Ting, and Z.-H. Zhou, "Isolation Forest ICDM08," *Icdm*, 2008, [Online]. Available: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf%0Ahttps://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest

[40] S. Wibisono, M. T. Anwar, A. Supriyanto, and I. H. A. Amin, "Multivariate weather anomaly detection using DBSCAN clustering algorithm," *J. Phys. Conf. Ser.*, vol. 1869, no. 1, 2021, doi: 10.1088/1742-6596/1869/1/012077.

[41] Z. Zhao, "Ymir: A Supervised Ensemble Framework for Multivariate Time Series Anomaly Detection," 2021, [Online]. Available: http://arxiv.org/abs/2112.04704

[42] J. Frery, A. Habrard, M. Sebban, O. Caelen, and L. He-Guelton, "Efficient Top Rank Optimization with Gradient Boosting for Supervised Anomaly Detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10534 LNAI, pp. 20–35, 2017, doi: 10.1007/978-3-319-71249-9_2.

[43] P. Raghavan and N. El Gayar, "Fraud Detection using Machine Learning and Deep Learning," *Proc. 2019 Int. Conf. Comput. Intell. Knowl. Econ. ICCIKE 2019*, no. June, pp. 334–339, 2019, doi: 10.1109/ICCIKE47802.2019.9004231.

[44] J. Ren *et al.*, "Application of a kNN-based similarity method to biopharmaceutical manufacturing," *Biotechnol. Prog.*, vol. 36, no. 2, 2020, doi: 10.1002/btpr.2945.

[45] L. L. Li, X. Zhao, M. L. Tseng, and R. R. Tan, "Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm," *J. Clean. Prod.*, vol. 242, p. 118447, 2020, doi: 10.1016/j.jclepro.2019.118447.

[46] P. Bergmann, "MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," pp. 9592–9600.

[47] K. Gupta, M. A. Mohammed, and N. Jiwani, "ANOMALY DETECTION IN TIME SERIES USING DEEP LEARNING," vol. 7, no. 6, pp. 296–305, 2022.

[48] E. Akyildirim, M. Gambara, J. Teichmann, and S. Zhou, "Applications of Signature Methods to Market Anomaly Detection," pp. 1–32.

[49] S. Crépey, N. Lehdili, N. Madhar, and M. Thomas, "Anomaly Detection in Financial Time Series by Principal Component Analysis and Neural Networks †," pp. 1–38, 2022.

[50] W. Yu, Y. Wang, L. Liu, Y. An, B. Yuan, and J. Panneerselvam, "A Multi-perspective Fraud Detection Method for Multi-Participant E-commerce Transactions".

[51] S. Deshwal, "Anomaly detection in bank transactions using Machine Learning".

[52] B. Group and F. Make, "Fraud Detection on E-commerce Transactions Using Machine Learning Techniques," 2023.

[53] P. N. Saputri, F. Al Zami, F. O. Saputra, and P. Nurtantio, "Implementation Of Extreme Gradient Boosting Algorithm For Predicting The Red Onion Prices," pp. 18–27, 2021.

[54] P. Plants, "Power Plants," pp. 1–17, 2022.

[55] W. Ccoya and E. Pinto, "Comparative Analysis of Libraries for the Sentimental Analysis," 2023, [Online]. Available: http://arxiv.org/abs/2307.14311

[56] G. R. Ji, P. Han, and Y. J. Zhai, "Wind speed forecasting based on support vector machine with forecasting error estimation," *Proc. Sixth Int. Conf. Mach. Learn. Cybern. ICMLC 2007*, vol. 5, no. August, pp. 2735–2739, 2007, doi: 10.1109/ICMLC.2007.4370612.

[57] M. Zeitoun, "Analysis of Temperature Anomalies During the Spring Months in Jordan," *Int. J. Geoinformatics*, vol. 20, no. 1, pp. 88–98, 2024, doi: 10.52939/ijg.v20i1.3029.

[58] H. Xu, G. Pang, Y. Wang, and Y. Wang, "Deep Isolation Forest for Anomaly Detection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12591–12604, 2023, doi:

10.1109/TKDE.2023.3270293.

[59] A. Suryaputra Paramita, "Comparison of K-Means and DBSCAN Algorithms for Customer Segmentation in E-commerce," *J. Digit. Mark. Digit. Curr.*, vol. 1, no. 1, pp. 29–43, 2024.

[60] C. Mac, "Visualisation of Random Forest classification," 2024, doi: 10.1177/14738716241260745.

[61] N. Nayyer, N. Javaid, M. Akbar, A. Aldegheishem, N. Alrajeh, and M. Jamil, "A New Framework for Fraud Detection in Bitcoin Transactions Through Ensemble Stacking Model in Smart Cities," *IEEE Access*, vol. 11, no. August, pp. 90916–90938, 2023, doi: 10.1109/ACCESS.2023.3308298.

[62] Z. Ya, Z. Qingqing, W. Yuhan, and Z. Shuai, "LDA_RAD: A Spam review detection method based on topic model and reviewer anomaly degree," *J. Phys. Conf. Ser.*, vol. 1550, no. 2, 2020, doi: 10.1088/1742-6596/1550/2/022008.

[63] F. Rezaei, M. Afsharkazemi, and M. Keramati, "Detection of E-commerce Attacks and Anomalies using Adaptive Neuro-Fuzzy Inference System and Firefly Optimization Algorithm," *Int. J. Inf. Commun. Technol. Res.*, vol. 13, no. 1, pp. 32–39, 2021, doi: 10.52547/ijict.13.1.32.

[64] J. Ramakrishnan, C. Li, and E. Shaabani, "Anomaly detection for an e-commerce pricing system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 1917–1926, 2019, doi: 10.1145/3292500.3330748.