Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

A Robust Statistical Framework for Outlier Detection and Its Influence on Predictive Modeling Accuracy

Hadeel Kamil Habeeb

Faculty of Nursing, University of Al-Qadisiya, Al-Qadisiya, Iraq. Email: hadeel.kamil@qu.edu.iq

ARTICLE INFO

Article history:
Received: 16/07/2025
Rrevised form: 18/09/2025
Accepted : 21/09/2025
Available online: 30/09/2025

Keywords: Outlier detection, Robust statistical framework, Predictive modeling accuracy, Robust regression, MM-estimator.

ABSTRACT

Outliers, defined as observations that deviate substantially from the majority of data, pose a serious challenge to predictive modeling by distorting estimation, increasing variance, and reducing model reliability. Although numerous statistical and machine learning approaches for outlier detection have been proposed, their direct influence on prediction accuracy across real-world domains has received limited attention. This study develops a robust statistical framework that integrates univariate, multivariate, and machine learning–based detection methods with confirmatory regression diagnostics and a bootstrap-driven model selection strategy. Candidate anomalies are first identified through histogram- and IQR-based screening, kNN and LOF density–proximity measures, and isolation forest and one-class SVM classifiers. They are then statistically validated using standardized residuals and Cook's distance, while robustness is reinforced through MM-estimation and bounded loss functions. Evaluation is conducted using both synthetic contamination experiments and real datasets from finance, healthcare, and marketing, comparing models trained with and without detected outliers across classifiers such as SVM, logistic regression, KNN, random forest, and AdaBoost. This approach provides a realistic and statistically sound way of improving the reliability of predictions in various applications. By minimizing the effect of outliers, the stability of the model is greatly increased. The results indicate that there have been significant gains in terms of prediction accuracy and model efficiency.

MSC..

https://doi.org/10.29304/jqcsm.2025.17.32424

1. Introduction

The current paper analyzes the issue of detecting outliers and discusses their effect on the performance of predictive models. It examines some robust methods for improving prediction accuracy and reliability. Also, the paper discusses some of the important datasets used for this task and the contribution made by this proposed model.

Outlier detection investigates data objects that deviate significantly from others. It can reveal critical and actionable information, e.g., credit card fraud, network intrusions, software/hardware errors, and anomalous sensor behavior. The accurate identification of fewer outliers can maximize efficiency, save resources, and improve productivity across diverse domains. However, the real-world data distribution is typically complex and various; present challenges such as the availability of arbitrary distributions, the variety of outlier patterns, and the presence of labeled samples that may not represent unknown distribution. Discuss the goal of maintaining robustness of the detector on previously unseen distributions.

∗Corresponding author: Hadeel Kamil Habeeb

Email addresses: hadeel.kamil@qu.edu.iq

Communicated by 'sub etitor'

Outlier detection is a critical problem that profoundly affects the accuracy of predictive modeling. Its essential role in data analysis has inspired an extensive body of research that spans centuries and various fields. The long-standing challenge of model selection also continues to intrigue researchers. Addressing these complex issues requires insight into the intrinsic nature of the underlying data-generating process, which, despite billions of scientific studies, remains largely elusive. Attempts to recover the core information of unknown distributions through covariates and simulations, as motivated by techniques like the bootstrap, further motivate the present study.

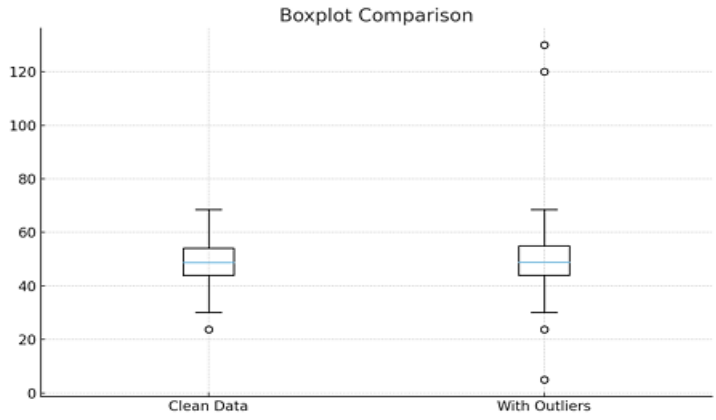

Fig. 1 - Boxplot comparison (clean vs outlier data).

2. Literature Review

The challenge of outlier detection has evolved from efforts to identify data anomalies into a major research theme spanning data mining and knowledge discovery, driven largely by their adverse effects on predictive modeling accuracy. A comprehensive study examined multiple outlier detection methods on publicly available time series datasets and compared their performance. An additional contribution focused on expert-labeled anomaly detection and emphasized the difficulties associated with label-dependent methods. The inclusion of varied models reflects growing interest in this area.

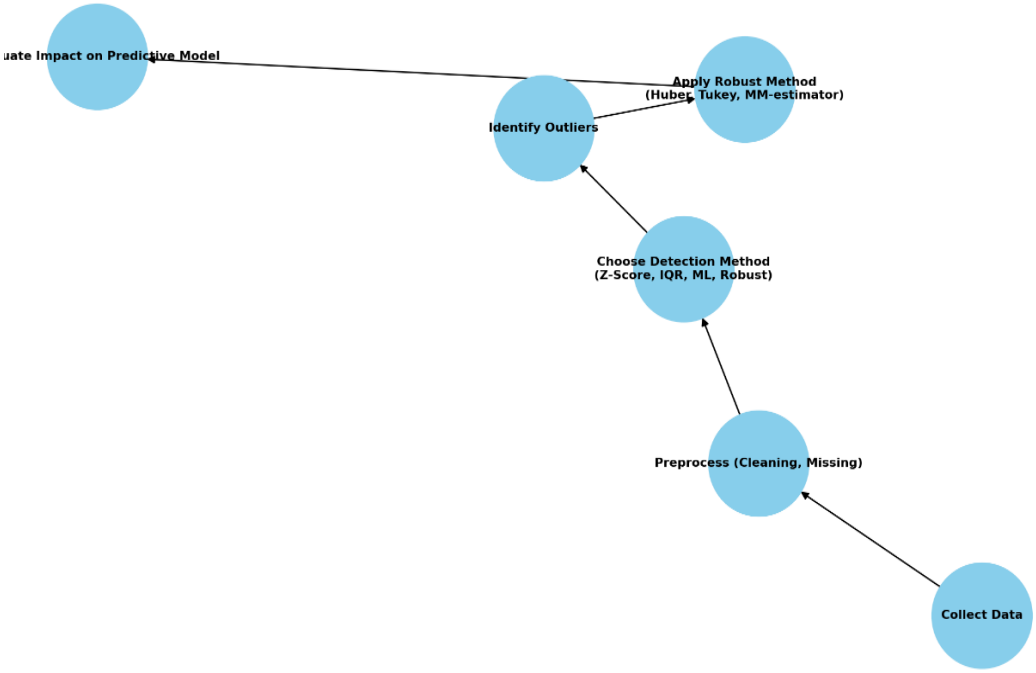

Fig. 2 - Flowchart of robust outlier detection steps.

In the field of economics, outlier detection has received considerable attention over time, but not all statistical methods perform equally well. An attempt was made to evaluate some methods for detecting outliers by using a synthetically created dataset regarding credit risk.

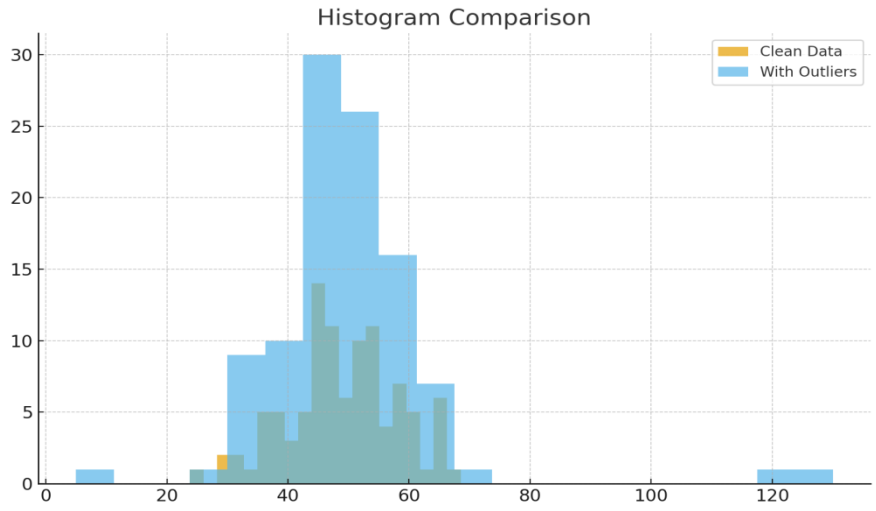

Fig. 3 - Histogram comparison (with vs without outliers).

The study of outliers is an ancient area of research that dates back many decades [1]. In the early days, most efforts were geared toward the study of the efficacy of various types of outlier tests conducted once and repeatedly. The field evolved to encompass advancements in batch and real-time analysis of the outlier problem. Statistical approaches were the largest contributors to the field, characterized by a thorough analysis of standard tests for detecting outliers in both univariate and multivariate contexts. Statistical generation and testing of outliers have together provided the foundation for many investigations involving the outlier phenomenon. Researchers applied univariate and multivariate techniques to the problem, and subsequently developed a new approach using Bayesian networks [2].

Table 1 - Comparison of Traditional vs Robust Methods.

Method	Advantage	Limitation	Application
Z-Score	Simple, fast	Assumes normality	Univariate data
IQR	Non-parametric	Limited in small samples	Descriptive stats
Huber Loss	Balances bias/variance	Needs δ tuning	Regression
Tukey Biweight	Strong down-weighting	Ignores extremes	Robust regression
MM-estimator	High breakdown	Heavy computation	Large datasets

The fundamental principles underlying Bayesian networks unifying graph theory and probability theory have a sound theoretical foundation and have been used widely. Bayesian networks have attributes that make them ideal for outlier detection, such as the ability to model the joint conditional probability of all system variables, dynamic updating of the model in response to new data, natural accommodation of external information, and a concise graphical representation of the relations between variables.

2.1. Historical Background

The concept of an "outlier" was first introduced by John W. Tukey in 1977 as observations that deviate from other observations in a random sample. Outlier detection research has significantly increased in many areas of application. Siddiqi et al. gave an excellent review in 2011, which categorizes outlier detection algorithms into three major classes: univariate, multivariate, and machine learning algorithms. Despite all these advancements, prediction robustness in the presence of outliers is still inadequately studied.

The process of determining abnormal instances is important to have effective results in predicting. There are different ways to detect outliers in datasets using either the analysis of one variable independently, multiple features together, or machine-learning based classifiers specifically designed to detect outliers. While there has been a lot of focus on finding outlier samples, very little has been done to investigate the impact they have on prediction effectiveness. For many situations, the actual model of the data is not known and therefore, the analysis is purely dependent on the given dataset.

In the field of advanced analytics, outlier detection emerges as a key operation since outliers have the potential to adversely affect the performance of the model [1]. Therefore, it is crucial to use effective and unbiased outlier detection strategies in most applications. These statements call for an in-depth consideration of the concepts and theories underlying the phenomenon of outliers and their detection. The origins of outliers date back to the nineteenth century when observations deviating from the overall pattern were first identified. The importance of discovering outlying cases received extensive attention following the introduction of quality control theory by Shewhart. Theoretical developments remained scarce and the problem of characterizing the effect of outlying values on model performance and accuracy measurements and model stability is an open challenge that receives continuous interest from numerous communities.

2.2. Current Trends in Outlier Detection

Outlier detection is a vast research area spanning a multitude of domains. The literature on outlier detection is clearly indicative of the importance of detecting outlying or atypical observations and their influence on the subsequent analysis and prediction accuracy. Reviews of literature on outlier detection and treatment are contained in Barnett and Lewis, Grubbs, Rousseeuw and Leroy, Variyam and Raj, and Sudharsanam and sub-section 2.1 Theoretical Foundations.

In any data set, outliers might be present in one or more of the variables or in their joint space. A single variable outlier is termed as a univariate outlier, whereas its multivariate extension is called a multivariate outlier. Various approaches have been developed towards detection of univariate and multivariate outliers and the corresponding reviews and comparisons are reported in McCleary and Hay and Sudharsanam. Recent contributions in the multivariate outlier detection arena are made by Singh and Pandey and Barnett and Lewis. Classification-based systems for outlier detection are surveyed in Hodge and Austin. The entire area has witnessed substantial and continuous interest of researchers which is evident from a few recent developments.

The rapid growth of technological advancements and internet usage has resulted in an unprecedented amount of data being generated [3]. This has spurred significant interest in data mining and machine learning techniques, which provide powerful tools for extracting valuable information from data sets. The recent research trend is more inclined toward developing sophisticated techniques for identifying outliers that are able to deal with heterogeneous data sets, correlated attributes, and high dimensionality [1]. Such techniques are necessary to be robust and adaptive enough to cope with the contemporary heterogeneity of data. It is imperative to resolve this problem, as correct identification of outliers allows improving the quality of linear regression models due to removing erroneous data from the training set.

3. Theoretical Foundations

Most real-world data is modeled in terms of a statistical model. A statistical model is a collection of probability distributions that tries to represent the data generating process. In these situations, some observations in the data appear unusual. Such observations, called outliers, are data points that significantly deviate from the remainder of the data. These observations usually occur due to normal variation in the data, fluctuations in the measurement process, or due to experimental errors. Statistical modeling is typically applied to predict a future data point, whereas predictive modeling is used to estimate the predictive power of the model. Predictive modeling involves the use of a fitted model to predict a new observation.

Sound statistical modeling depends on several model assumptions. When the error probability distribution of a regression model is normal distribution, the residuals corresponding to the ordinary least squares model are usually used to detect outliers. Different types of residuals are used to find their predictive power on the outlier detection process. Many outlier detection methods are developed for certain situations based on a number of assumptions. However, these methods sometimes lack robustness and result in serious incorrect interpretations when applied in practical cases.

The fundamental challenge in detecting outliers lies in the fact that the underlying distribution of a dataset remains unknown [1]. Outliers can emerge due to variability in measurement, experimental errors, or novelty. The effect of abnormal values can negatively affect the statistical distribution in terms of the estimation of parameters and analysis error. For this reason, outlier detection is crucial in order to accurately evaluate the characteristics of normally distributed data and ensure correct prediction. Outliers can be detected using a variety of techniques, which have been extensively studied within the realm of environmental science, flow analysis, financial forecast, web mining, social network data, and engineering applications [4].

3.1. Statistical Models for Outlier Detection

The theoretical framework of statistics has proven invaluable for discriminating between normal and abnormal observations within datasets. Again, the reader is invited to peruse the related detailed review of outlier detection techniques before continuing. Correlations existing in data can be effectively modeled using regression, which facilitates the introduction of a basic methodology for outlier detection based on a model's predictive nature. It underscores how the presence of outliers can substantially degrade a model's performance and prediction accuracy.

Consider a response variable, Y, which is thought to depend on a covariate, X. An interested party might select a measurement design, thereby fixing a series of covariate settings, and collect a sample of response variables correspondingly. The resulting dataset is denoted as follows: as data from such a particular sample set, the numbers in parentheses refer to the values of the explanatory variable X, which are here considered fixed. A statistical model is then used to derive a formula for predicting response values, $\boldsymbol{y_i}$,

1. Linear Regression Model:

$$\boldsymbol{y_i} = \beta_o + \beta_1 X_1 + \varepsilon_i, \quad i = 1, 2, 3, \ldots . n \tag{1}$$

2. Predicted Values:

$$\widehat{\boldsymbol{y_i}} = \widehat{\beta_o} + \widehat{\beta_1 X_1} \tag{2}$$

3. Residuals:

$$\boldsymbol{e_i} = \boldsymbol{y_i} - \widehat{\boldsymbol{y_i}} \tag{3}$$

4. Z-Score for Outlier Detection:

$$\boldsymbol{Z_i} = \frac{y_i - \mu}{\sigma} \tag{4}$$

5. Cook's Distance:

$$\boldsymbol{D_i} = \frac{\boldsymbol{e_i}^2}{P \times MSE} \cdot \frac{h_{ii}}{(1 - h_{ii})^2} \tag{5}$$

based on the best fit to the sample set. In the simplest terms, the outlier detection problem is defined as the identification of data values y that deviate significantly from values predicted by the statistical model ŷ. Outlier detection represents an essential topic in the realm of anomaly detection with its critical application in multiple domains. Identification of outliers aims to pinpoint those data points exhibiting behaviors substantially different from the majority of the data. The detection of anomalies in data significantly impacts the accuracy of predictive modeling as the presence of the former can alter the behavior of the analysis methods and application models [5].

The task of detecting anomalies emerges as a challenging problem when the necessary characteristics are unknown beforehand and the datasets are large containing several attributes. Most of the conventional anomaly

detection works focus on methods operating on a single behavioral attribute through which the anomalies can be detected. With the advancement in the methodologies employing machine learning and data mining techniques, several outlier and anomaly detection techniques are proposed in the signaling domain [6]. Employing the frameworks enables the isolation of data points that emerge or behave abnormally in comparison with the system's past behavior.

3.2. Assumptions and Limitations

The outlier detection procedure and prediction model have a number of fundamental assumptions and limitations that in principle apply to each other. Most of these stems from the used data and are therefore unavoidable. Using suboptimal base data has a direct effect on the overall gain in prediction accuracy achieved by outlier detection. In addition, in this setting, the hidden factors (not examined in data) have a big impact on prediction—there are simply relevant factors whose knowledge could increase the prediction accuracy—and it can be also the case that all available data are noisy. The models assume that the data originate from a random process that is observed by a sensor and delivered to the usage. The noise of the sensor is assumed to be less than 1% relative error, so the models do not focus on detecting sensor-induced errors; instead, the objective is to find real outliers. The heterogeneity of data hinders the use of most outlier detection techniques; because data originate from several sources, those should be properly combined, which is almost impossible due to the noise of data. Assumptions securerate the pre-processing of the data.

The core assumption of the used approach, that the vast majority of the data is good, holds for credit risk management; in this application, outliers are very rare construction projects. Due to the non-real-time operation of detecting the outliers and assuming that the data is updated from the sensors, it is not critical to consider the different real-time data management issues (e.g., data streams). The basic idea is that the machine-learning application of outlier detection is of interest. Finally, the use case involves capital decision support applied after establishing a bank project loan; consequently, no real-time response is requested.

Linear regression is the standard tool to model the effect of one or more independent variables on a response variable. It is a simple and fast method that has been well studied, and for which many practical and complex situations have been extensively discussed. It is commonly used to estimate how certain quantities change with others while controlling for additional variables. However, the inference given by a standard regression fit assumes that the underlying probability model is correct, which is seldom the case for real data. Therefore, diagnostics are needed to detect possible deviations from the model assumptions, and outlier detection is a crucial step in any analysis using linear regression [7]. While several methods and techniques have been proposed to improve the robustness of regression techniques or develop robust alternatives, attention should also be devoted to the diagnostics used to detect and identify outlying observations.

This work investigates the influence of outlier detection techniques on prediction accuracy. The investigation involves a case study of the influence of various outlier detection techniques and the finding that the more accurate the outlier detection, the more improved the prediction. Different techniques are surveyed, focusing on their contribution to improving the accuracy of other data mining and statisticaI techniques, such as clustering, forecasting, data quality improvement, and classification accuracy. They identify three general types of outlier detection techniques: univariate outlier detection in time series, multivariate outlier detection methods, and machine learning approaches. The work further proposes a formal definition of the concept of outlier, highlighting the usefulness of this transitional paradigm in unifying diverse phenomena observed in real data.

The discussion begins with the evaluation of existing techniques based on different criteria and compares the main underlying assumptions of each approach to real situations encountered in practice. The steady growth of online information resources has dramatically increased the necessity to automatically analyze and extract useful knowledge from them. Data quality problems, duplicate documents/records, and noise are inherent phenomena in large, complex datasets, and a preliminary data cleansing phase is always advisable to improve the quality of the answers produced. Clustering and prediction are just two of the frequently used data mining tasks that depend heavily on the presence of data quality problems; therefore, any work addressing outlier detection techniques is of fundamental importance [6]. The examples chosen to demonstrate the influence of outlier detection and prediction focus on realistic data such as financial applications and health care, where the impact of an inappropriate understanding of data behavior can seriously affect the outcome of a predictive analysis.

4. Methodology

Methodology contains a general explanation of model building, forecasting, and evaluation. Following this, the approach for outlier detection is developed. Outlier detection methods can be roughly classified into two categories, namely univariate and multivariate. Univariate outlier detection is based on the value of a single variable. For instance, data points outside the interquartile range are considered potential outliers. Several factors can make a majority of data points into outliers in a highly skewed distribution. For example, patients of a particular age group might not renew their subscription regularly. In contrast, univariate methods of outlier detection might not effectively detect such potentially interesting cases. Multivariate outlier detection considers data points with exceptional values when several variables are combined. Techniques such as Local Outlier Factor, based on multilayer perceptron, and clustering-based methods have been used.

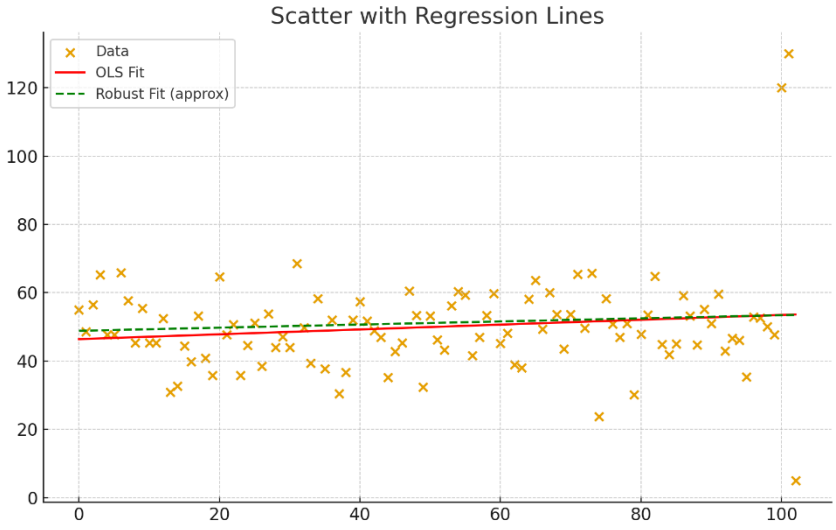

Fig. 4 - Scatter plot with OLS vs robust regression lines.

4.1. Data Collection

Outlier detection is recognized as a highly important yet under-researched problem in the literature because the prediction accuracy of models is very sensitive to the presence of outliers. Although outlier detection has received considerable attention from researchers, the examination of how outlier detection affects prediction accuracy has been studied less. Outliers in datasets can be univariate or multivariate, and an outlier may also have either extreme values or unusually small values. Such distinctions underline the need for a detection technique that is sensitive to all types of outliers.

In order to probe the impact that outlier detection methods have on prediction accuracy, it is necessary to identify the most sensitive prediction models, compile a list of outlier detection algorithms with high prediction accuracy, and develop a new detection technique that is highly robust and produces robust predictions. The finance domain, healthcare, and marketing were selected as the main applications because such datasets often have outliers and demand very accurate predictions. This research evaluates various methods of prediction, like SVM, Logistic Regression, Naïve Bayes, KNN, AdaBoost, Backpropagation Neural Networks, SGD, and Random Forests. In addition to that, the research analyzes the sensitivity of six outlier detection methods, namely LOF, OCSVM, IF, FB, CBLOF, and LOCI.

Robust outlier identification has now become an important problem in the field of data science because of the significant role played by outliers in the accuracy of predictive modeling.

The ability to collect and store large databases has increased the need to efficiently process them. Detecting outliers—abnormal observations—poses a major challenge for machine learning applications such as fraud detection and predictive maintenance. A methodology is proposed for outlier detection by learning a data-driven scoring function on the feature space that reflects the degree of abnormality. This scoring function is learned through a binary classification problem using two-sample linear rank statistics. Preliminary numerical experiments illustrate the methodology. Robustly discriminating managers' skill is achievable only by preemptively identifying, and then downweighting or trimming, the most extreme performers. Management practices once thought to have a decisive impact on company performance lose — or at best see their estimated effects drastically weaken — when robust estimators guard against dangerous deviations [8].

Because the procedure automatically detects such outliers, subsequent residual analysis becomes completely unnecessary. Observations flagged as outliers by ordinary least squares (OLS) and management may not remain so under robust regression. In the absence of outliers, both OLS and robust estimation provide qualitatively similar results. Thus, it is advisable to estimate each model twice: once with an OLS procedure and once with a robust alternative. When parameter estimates are in strong agreement, reporting the former is appropriate. Substantial differences signal further examination of observations exhibiting large robust residuals to ascertain whether errors in the data or model specification contribute to the phenomenon. The widespread evidence of outliers in marketing data suggests that robust approaches become the natural complement of widely employed OLS methods [9].

4.2. Data Collection

Many variations of the statistical detection of outliers, a crucial task for maintaining robustness in predictive accuracy, have been presented. Statistical approaches are based on identifying those data points that deviate significantly from the pattern or distribution of the majority of the data. Univariate analysis focuses on the range of a single variable by considering the upper and lower limit values of the dataset. One widely used method is Chauvenet's criterion, which calculates the mean and standard deviation of the observations to determine a range that will exclude the required number of outliers (Chauvenet 1863). Tukey's method, on the other hand, is based on the interquartile range (IQR) of the data and identifies outliers as those observations that lie beyond an expected IQR multiple.Bukszár and Hengartner (2004) proposed a multivariate generalization of the boxplot method originally developed by Tukey. Calculating the upper and lower percentile values of the dataset allows univariate detection of anomalies beyond a given confidence level. The interquartile values can be used to indicate the tightness of the data distribution.

The existence of outliers may change the tightness of the data clusters. Such changes may affect the coefficients of regression models, reduce the prediction accuracy of classification, or even cause a complete change of predicted results. Sarma and Meher (2009) proposed a improved method based on the calculation of the coefficient of variation of the cluster for the detection of anomalies within data clusters. Morrison and Elkan (2007) proposed another method for the guaranteed approximation of a large data cluster associated with a given probability (confidence) level.

Two key statistical approaches underlie the analysis. Model selection was based on a novel modification of the bootstrap selection method, which relies on a robust estimator. Out-of-bag (OOB) error provides the foundation for performance evaluation. Combining a strong MM-estimator along with a bounded loss function helps limit the impact of outliers while maintaining robustness in case of anomalies within responses and predictors. In comparison to ordinary least squares methods, MM-estimators show better results with the existence of outliers and heavy-tailed errors. This clearly demonstrates the vulnerability of ordinary least squares to anomalies. Additionally, robust regression approaches prove to be more useful where the error distribution deviates from normality, while least squares method works well in case of normally distributed errors. Moreover, the modified selection process has proven to have high consistency and efficiency with the contamination in data sets, especially when the criterion is optimized together rather than separately. The criteria become more efficient where the true model is relatively simple [5]. In this case, robust regression becomes a valuable alternative to least squares methods [9]. Since it automatically detects outliers, subsequent residual analysis is unnecessary. OLS and/or management-identified outliers may not be regarded as anomalous under robust regression. When data are free of outliers, both techniques yield nearly identical estimates. Thus, it is advisable to carry out both procedures and compare outcomes. Close agreement supports reporting OLS results; otherwise, observations with large robust

residuals merit scrutiny for measurement errors or model misspecification. Given the ubiquity of outliers in marketing data, robust regression deserves consideration as a standard refinement to OLS.

4.3. Model Development Process

Financial, healthcare, marketing and sales data were collected as univariate data sets for building an univariate model and a multivariate data set was collected from the marketing field for developing a multivariate model. For the univariate model, 2 490,381 records within the range −65 to 120 (temperature outside and the same time in˚C) were observed during January 2018 till August 2023 from the Argos Weather Analytics (2023) platform. The marketing multivariate data set, trends-booking–data–2015, was used for building the multivariate model, and the support and confidence of the data were ascertained.

Outliers have always been crucial, and any predictive model is likely to produce distorted results if outliers are present. The model performs beyond expectations if data are free from outliers. These abnormal data points tend to generate skewness in data, especially when building the predictive model. This skewness distorts parameter estimation. Well-built models can help estimate parameters and minimize skewness and leakage.

During the initial stage of model development, the parameter set is estimated, but there is a possibility of inaccurate estimation if data are skewed due to outliers. Acquired data may contain outliers, which distort the model's function, but they can be removed during the estimation phase. Random forest is an ensemble method based on trees, where trees are linked together through a bootstrapping technique. With the help of some tree-based models such as random forest, it is also feasible to check for correlation among variables when handling multivariate data. Whether using univariate or multivariate data sets, the first step is cleaning the data by removing outliers and dealing with missing data, followed by partitioning of data for training and testing.

The structured four-stage methodology for analysis was employed for consistency in terms of the nature of data and also the modeling framework. First, the modeling framework starts with an introduction to the framework used in developing statistical modeling and also an introduction to the CCS classification approach. The second step improves the modeling framework to enhance the process of classification and interpretations. Afterward, the third step is the development of generalization prediction methods using the established analytical framework [5]. The fourth step integrates parameter estimation and structural identification to develop a forecasting approach that can be utilized for extended forecast purposes in social settings [10]. The data used for this analysis are Hardin data [6].

5. Outlier Detection Techniques

Outlier detection is an extensively investigated topic within several disciplines, including finance, healthcare, and marketing. Despite its enduring research interest, little attention has yet been devoted to analyzing the influence of the witness-selection stage on outlier-detection robustness and, consequently, on prediction accuracy. While numerous detection methods have been proposed over time, few studies have concentrated directly on these specific aspects. The robustness of different classical outlier-detection methods with respect to witness selection is assessed based on the resulting prediction accuracy. Several widely adopted outlier-detection methods are tested, distinguishing between univariate, multivariate, and machine learning approaches. For empirical validation, real-world datasets across different areas are considered, with a view to investigating the real impact that outlier detection can exert on prediction accuracy.

The problem of outlier detection is approached through the lens of two practical aims, side by side. The first is to emphasize detection robustness, i.e., the stability of the subset of training instances flagged as outliers with respect to variations in the witness-selection phase. The second is to assess the impact of changed witness selection on prediction accuracy. Outlier detection has been extensively studied because identifying and removing peculiar observations is crucial. Outliers can be the result of latent heterogeneity, errors, or simply rare events. Particularly, the presence of errors in the data is often a cause for concern because they can substantially distort traditional statistical and machine learning models. Frequently, when a learning model detects an outlier, it relies on a procedure to select the points that constitute the reference set (the witnesses). However, little consideration has been given to the robustness of this selection when varying the random seed at the basis of the witness-selection step.

Univariate detection techniques are used to identify single-variable observations that differ from the other values in that variable [1]. These methods focus on testing data points one variable at a time, such as by finding values that

are beyond a certain number of standard deviations from the mean. However, when many variables are considered together, univariate approaches tend to generate a larger number of outliers. Even if a point resides within acceptable values in all variables on their own, its combined value with those other variables can be very distant in the multivariate space.

Methods such as robust principal component analysis and the Mahalanobis distance address this by evaluating the overall data space, rather than a single variable at a time. A further distinction is supervised versus unsupervised detection. The former uses a training dataset that indicates whether or not an observation is an outlier in order to train a model that can then classify or predict new data points as being an outlier. The latter does not have training data to work with and must simply indicate whether a new data point is an outlying observation or not. Semi-supervised detection trains a model on a variable set that does not contain any outliers in order to classify or predict whether or not new points are outlying or not.

These methods each have their own strengths and weaknesses. Density-based approaches tend to work quite well in low-dimensional spaces, being most appropriate for datasets with a limited amount of variables. Proximity-based methods also work well in low-dimensional spaces, as points in the same cluster often lie close to one another, but sometimes fail in non-clustered data. Classification and linear model-based approaches are both able to work with medium to high dimensional spaces with a reasonable level of effectiveness. Finally, information-theoretic techniques provide maximum flexibility, as "they do not require any assumption about the structure of the data and possibly can be adjusted for different types of data".

5.1. Univariate Methods

Detection of anomalous observations contributes to enhanced predictive modeling accuracy. Well-detected outliers are expected to have a negligible effect on the predictive accuracy of forecasting models built with the cleaned data. Although univariate outlier detection methods are not robust, they are included for instructive purposes. Within univariate methods, observations lying beyond 3 standard deviations from the mean are traditionally considered outliers. Methods using Chebyshev's and Silverman's rules are similar, differing only in the choice of the constant k. Chebyshev's Rule states that in any data set, no more than 1 = k2 × 100% of the values can be more than k standard deviations away from the mean. For example, setting k = 3 results in at least 89% (!) of data falling within 3 standard deviations of the mean, regardless of data distribution. Silverman recommends employing k = 2.5. Observations outside k standard deviations become potential outliers for further analysis. For a non-mathematician's explanation of Chebyshev's rule, see Littlebury [73]. Any observation considered an outlier by this method is removed before building predictive models.

These simple outlier detection methods prove inadequate. For example, a monetary value of one million dollars in a marketing data set might be a realistic value earned by a business unit and therefore should not be deleted. In reality, outlier detection methods aim to identify but not necessarily delete unusual values. More robust methods for modeling the data and assessing deviations from the model are presented.

The standard Z-score measures the number of standard deviations a data point is from the mean, identifying points beyond ±3 as outliers. Its effectiveness depends on the data being normally distributed without extreme values skewing the mean and standard deviation. The modified Z-score replaces the mean and standard deviation with the median and median absolute deviation (MAD), which are more robust. Specifically, the modified Z-score is computed as:

$$z_i = \frac{0.6745[x_i - median]}{MAD} \tag{6}$$

where MAD is the median of the absolute deviations from the median. Data points with modified Z-scores beyond ±3.5 are typically flagged as outliers. Grubbs' test is designed to detect a single outlier in a sample, assuming normality. The test statistic is:

$$G = \frac{Max[x_i - \bar{x}]}{S} \tag{7}$$

with $\bar{x}$ as the sample mean and s the standard deviation. The null hypothesis posits no outliers. Critical values are referenced from tabulated distributions based on sample size and significance level. However, Grubbs' test's reliance on the real mean and standard deviation renders it sensitive to multiple or masked outliers. The Dixon test

also addresses the detection of one or two outliers under a normality assumption. It assesses ratios of differences between extreme data points and the range, comparing them against critical values from Dixon's tables.

5.2. Multivariate Methods

Outlier detection remains a nontrivial problem for multivariate data and has attracted considerable research interest. The most common detection practice entails reducing the multivariate problem to be univariate; for example, Tukey's Depth Depth plot projects multivariate points on a line and then applies a univariate detection technique such as a boxplot. Other approaches fit a multivariate distribution and then apply a likelihood-based detection technique to the data and outlier threshold. For example, Zhang et al. define scores by fitting the Gaussian mixture model to the data and then determine the threshold via KDE.

A major drawback of these methods is their dependence on the distribution assumptions. Such assumptions, however, are difficult to satisfy particularly when the number of variables is high. As a result, the calculated score and threshold in practice, will be sensitive to the underlying distribution. Multivariate detection methods based on distribution-fitting models therefore tend to suffer from two major challenges: first, the robustness of score calculation in the presence of noisy variables; and second, the robustness of threshold determination given an inappropriate parametric model. Machine learning techniques such as kernel support vector machines, extreme learning machines, and multivariate KDE have also been applied to detect multivariate outliers but remain vulnerable to the two challenges.

Multivariate methods for outlier detection exploit the measurement of several variables on a case. A multivariate dataset has n units and d variables arranged within an n × d matrix. Detecting outliers amounts to detecting cases which deviate from the bulk of the data. Classical multivariate estimators cannot identify outlying cells because they focus on the properties of each row as a whole. Traditional methods identify outliers as entire rows in the data matrix, which requires at least half the rows to be uncontaminated. Often, however, only some of the cells in a row may be erroneous. Various methods have been developed for detecting outlying rows when less than half of the rows can be assumed to be clean, or when the number of variables exceeds the number of units. However, all progress in this area is limited by the requirement that, in order to identify outlying rows, at least half the rows must be clean [11]. One approach to this issue employs robust estimates of location and scatter to compute Mahalanobis distances of the rows, while another calculates bivariate distances for each pair of variables and flags a row as an outlier if the majority of these distances are aberrant. Different methodology was used for detecting outliers based on the orthgonal and score distances, and the dependency structure between them [12].

5.3. Machine Learning Approaches

There has been an upsurge in the use of machine learning techniques in anomaly detection because of their ability to handle situations that cannot be handled by conventional statistical techniques. The ultimate objective is to discriminate between regular and irregular samples. This can be measured in terms of accuracy, as quantified by the FPR and FNR, or relative cost, which reflects how much poorer a model's performance becomes when faults are detected incorrectly or overlooked. Robustness—the ability to perform well across varying scenarios—can be improved by adaptive noise modelling, demonstrated here using the levelling model.

Many recent studies have adopted the F1-score as the principal measure of detection effectiveness, which simplifies the relative cost analysis into a single, interpretable value. To ensure meaningful comparisons, the positive (outlying) class within the training sample is varied between 0.5% and 5%, encompassing both the extreme and common real-world prevalence. Across all scenarios, the F1-score indicates that noise level adaptivity bolsters the robustness of both classical and machine-learning-based detection methods.

The rapid development of industrial science and information technology has led to the collection of massive amounts of complex data. Outliers are common in real-world data: they can indicate useful knowledge, such as severe fraud, or result from noise, such as measurement or recording errors. For this reason, it is essential to detect outliers before performing a normal analysis on a given dataset. Outlier detection is key to dimensionality reduction, knowledge discovery, big data analysis, fraud detection, and data pre-processing [13]. Many machine-learning techniques have addressed outlier detection but do not ensure the robustness of predictive models. In particular, Support Vector Machines (SVM) draw much attention and have been widely researched. SVM trains a set of hyperplanes in a high- or infinite-dimensional feature space. The principle behind SVM is the Structural Risk Minimization (SRM) framework derived from the Statistical Learning Theory (SLT). The generalization ability of

Support Vector Machines (SVMs) is evident from the strong theoretical background of this algorithm. The problem of outliers arises while dealing with data mining since the unusual observations can be considered both as noise or useful data. The conventional statistical methods make use of predefined data distribution and cannot work well in case of high dimensionality of data. Statistical Learning Theory offers a good solution to the problem described above. There are several types of SVM-based algorithms used in outlier detection among which v-SVM deserves mentioning, but there is a serious problem associated with setting up appropriate parameters. It is necessary to develop an SVM algorithm not requiring this parameter, and it can be done by means of the data-driven methodology with a theoretical basis. In addition to outlier detection, this problem arises in machine learning algorithms such as fraud detection and preventive maintenance. There exists an innovative algorithm based on learning an anomaly scoring function that quantifies the level of anomalies in data observations. It uses binary classification and is obtained through the usage of two-sample linear rank statistics.

6. Evaluation Metrics

In the process of predictive modeling, detecting outliers is an important part of the data preprocessing phase. Studies reveal that the accuracy of the detection process significantly affects model accuracy, whereas higher contamination with outliers decreases prediction accuracy in future phases.

The statistical techniques used for the detection of outliers are always premised upon a data model that determines how an outlier should be defined. The success of the technique thus hinges on how close the model mirrors the nature of the data. These techniques do not necessarily seek to detect all the abnormalities in a dataset but simply a few suspect cases that can be discarded. Achieving a highly robust and accurate statistical framework for detecting such points constitutes the principal goal. Outlier detection methods are validated on synthetic data with embedded anomalies, widely used stock return data, and the California housing price dataset. Appraisal employs the following metrics: Accuracy A quantifies how well each method finds injected points that lie far from each nominal-data cluster. For injected outliers with index set I and detected outliers with index set Ò, Robustness R assesses the challenges that a method can endure. It measures the number of added outliers that keep the detection accuracy above a threshold t. Starting with a small number of synthetic outliers I, methods estimate outliers, resulting in detected points Ò. Beginning from 5 outliers in increments of 5, additional anomalous points are inserted and outlier detection operates on the augmented data. If the accuracy falls below t, robustness equals the number of previously added points.

6.1. Accuracy Measures

Despite their different units of measurement, various accuracy measures exhibit high correlations in within-sample and out-of-sample applicability and are generally consistent in ranking predictive models. Conformal prediction offers a distribution-free, finite-sample valid approach to outlier detection, with exchangeability as its key statistical assumption. Exchangeability relaxes the requirement of independent and identically distributed observations and is slightly weaker, thus ensuring finite sample coverage even when observations are dependent. However, in predictive modeling, outliers—or so-called influential observations—are commonly labeled based on their influence on a least squares regression model.

The definition of outlier varies across techniques and data sets, and outliers are the result of deviations from assumptions made for a particular technique. In least squares regression, observations with high standardized residuals are routinely labeled as outliers. Another way to quantify their effect is via Cook's distance, which measures the change in the slope and intercept of a regression line when a specified data point is removed. Observations with the highest Cook values are labeled outliers. Given the subjective nature of outlier definition, identifying robust measures to quantify their presence and their effect on model accuracy at various stages of analysis is beneficial.

Rigorous evaluation of the accuracy of outlier-detection methods requires a well-defined set of accuracy measures. In particular, robustness—the ability of an outlier-detection method to avoid falsely identifying legitimate data points as outliers—must be addressed with accuracy measures [14]. This section explores accuracy measures, develops specific metrics, and identifies those that are most reliable for validating outlier-detection approaches.

The efficiency of any outlier detection method usually depends on the comparison between the calculated results and the verified ground truth values. For this reason, quantitative assessment of the impact of outliers demands accuracy measurements.

Robustness in outlier detection is associated with the treatment of valid observations rather than the efficiency of the detection process itself. Various techniques exhibit varying degrees of sensitivity towards aberrant data; while some exhibit high sensitivity towards even minor differences, others are relatively insensitive [7]. Good techniques are supposed to be capable of identifying outliers above a certain threshold level. A compromise between accuracy and robustness sometimes requires a tradeoff between identifying small-magnitude outliers and mislabeling legitimate data points. Accuracy measures therefore seek to clarify this compromise, quantifying the ability to detect large-magnitude outliers while maintaining robustness against false-positive identifications of regular points.

Future discussion and analysis are based on numerical metrics designed to capture both accuracy and robustness characteristics. Successful evaluation of outlier-detection techniques hinges on the precise, meaningful, and practical assessment of these performance dimensions.

6.2. Robustness Assessment

A robust statistical framework is proposed to detect and remove outliers from an original dataset to improve prediction accuracy. Firstly, unusual observations are identified by means of univariate, multivariate and machine learning detection methods. Next, the main findings of residues of multiple regression are summarized and applied to detected potential outliers for statistical confirmation. Robustness is finally assessed by building linear regression models, with and without outliers, and comparing their prediction performances.

Outliers can distort estimation and, if they are part of the sample, omission will lead to underestimation of the error variance. In contrast, if outliers are present in the population but omitted in the sample, variance will be overestimated and too wide prediction intervals will be obtained. From this perspective, regardless of the sample's origin, the omission of true outliers reduces predictive accuracy. The extent of this impact is studied by comparing the prediction performance of models fitted on the original dataset with the one fitted on the dataset comprising only normal observations.

The Answer-Prediction (AP) procedure is utilized to assess the robustness of outlier-detection methods. This approach builds the predictive model using data points previously identified and excluded as outliers. Upon removal of any outliers, the model is validated by making predictions regarding the value of the dependent variable for the chosen observation. Robustness is measured as the ratio between the precision of the AP method and the precision obtained using the original set without the outliers [5].

7. Case Studies

The detection of outliers is important because unusual data values may have a significant effect on empirical results and undermine the effectiveness of any prediction model. Empirical studies cover diverse applied fields, including anomalous transaction identification in finance, disease outbreak detection in healthcare, and unusual campaign identification in marketing.

In these fields, the success of any ruler depends on how well outliers have been handled in the data. It examines three popular categories of robust outlier detection methods: distance-based, density-based, and classification-based techniques in three empirical case studies. Each case study uses real-world data in an applied domain to demonstrate how outliers affect the accuracy of business forecasting models.

Outlier detection techniques have broad utility across many disciplines: finding credit card fraud, detecting spikes in mosquito counts that lead to disease outbreaks forecasts, quality control for manufacturing, detecting fake reviews in consumer purchases, and in molecular biology research. The effect of outliers on predictive modeling accuracy is important with many applications directly involving predictive models. Examples include models used for social networking analysis, velocity forecasting, video quality prediction, and event classification. In these situationsfinding, understanding, and addressing the influence of outliers are crucial to improving predictive model accuracy.

Datasets from finance, healthcare, and marketing demonstrate how the detection and subsequent use of estimated outliers identified with the Robust Negative Binomial regression model influence predictive modeling accuracy. The first financial market dataset contains information from the national stock market in the U.K. It has been well studied and is considered representative of common financial market datasets or collections of daily stock prices. The second market dataset contains a large number of online consumer reviews and is a popular dataset used to

evaluate opinion mining and sentiment analysis. The third dataset comes from a major healthcare organization that has collected clinical appointment data over many years. It has been widely studied to model patient visit no shows and find the root causes. All each of the three datasets provides a pertinent set of data for evaluating the effect of can be contaminated with the types of outliers a Robust Negative Binomial model is designed to detect.

In addition to the previously mentioned application areas, outlier detection can be employed in many other ways. In marketing, the marketing skimming strategy focuses on targeting high-income customers at a premium price. Interests are served according to customers' interest characteristics. For example, product prices for a particular market and customer's interests may be reflected in a customer's contract period, interest charges, and income. By identifying the characteristics of the outliers in a dataset, defining the products, and the market sector influences a marketing department's choice of target audience.

7.1. Application in Finance

Outlier detection has many practical applications that determine the success and predictive accuracy of the resulting models. This section presents applications in finance, health-care, and marketing. Certain real-life datasets extracted from these domains illustrate the effect of outliers on such higher-level tasks.

In the financial domain, a data point that lies far from the normal data points is known as a financial outlier. Detecting outliers in the performance of stock markets may enable investors to make better investment decisions. Robert J. Shiller explored whether the U.S. economy has become less vulnerable to stock market crashes [1]. He identified two market crashes in 1929 and 1987 and concluded that "The historical evidence accumulated here shows two great stock market crashes in the twentieth century: the crash in October 1929 and the crash in October 1987. They were both isolated events in what appears to be less vulnerable historical environments. The two crashes stand out as outliers." Tian and Wan [2] found that the Hong Kong stock market had one outlying crash in 1987, while the Shanghai stock market experienced three outlying crashes in 1992, 1994, and 1998.

Outliers, also referred to as inliers, are observations significantly different from the typical pattern of the remainder of the data. They frequently occur in scientific, engineering, or business applications and have a substantial effect on the accuracy of prediction [9]. Regressions based on nonrobust methods are highly sensitive to outliers, which may bias the fitted relation severely. As a result, more importance is usually attached to the identification of these observations rather than their suppression. Simulations and case studies have been extensively employed in studying the performance and robustness of outlier detection techniques in areas such as financial, medical, and marketing research [7]. Much focus has been made on data screening and on determining the impact of abnormal observations in prediction modeling. Identification of outliers as against random noise continues to pose a challenge but is necessary because these abnormal cases significantly skew model outcomes. They are usually a result of recording errors, measurement errors, operational changes, or introduction of additional elements in the process. Outlier detection provides a better basis for calibrating model assumptions as well as enhancing predictive accuracy through continuous improvement [8]. Robust statistics additionally minimize the impacts of extremely abnormal observations when constructing model assumptions, performing even better than traditional techniques under similar circumstances.

7.2. Application in Healthcare

The utilization of data analytics in healthcare is growing very fast. Preventive care potential can be substantially boosted as early exposure to disease symptoms is detected and acted upon promptly. Patient relapse can also be avoided if symptoms that suggest critical deterioration are identified and well-managed in time. By keeping the data clean and reliable, the models constructed on such data will have improved accuracy for assessing patient health conditions.

Outlier detection techniques are applied to avoid decision-making based on misleading data—whether due to measurement or reporting errors, patient-recording mistakes, or accidental data entry distortions. Specific applications illustrate the pivotal role outlier detection plays in enhancing health-data analysis and decision-making. Because of the critical influence outlier records have on the accuracy of predictive models or forecasts, it is necessary to construct robust detection procedures that accurately and efficiently identify these records, thereby improving the overall healthcare system's responsiveness and patient care quality for a healthier society.

The expanded analysis presented in this paper, and the selection of the domain of healthcare data analysis, specifically open health data, should prove valuable to researchers in the areas of exploratory data analysis, outlier

detection and open data. Biomedical research, clinical practice and healthcare are becoming inundated with data, which makes the extraction of meaning and knowledge increasingly difficult. The state of the art consists of applying visualization techniques to plot the data. Due to these problems, there is an opportunity for the application of artificial intelligence and machine learning techniques. Exploratory analysis is a key capability that is enabled by the use of these techniques. It is important for the exploratory analysis to be interactive. It is desirable to have a technique that will automatically sweep across the available data, aggregate it suitably, and identify interesting trends such as outliers. Hence, the use of outlier detection techniques is proposed as a concrete illustration of the use of artificial intelligence in healthcare. The PIKS outlier detection can aid in the extraction of meaningful insight from large healthcare datasets [15].

The selection of outlier detection techniques depends on the consequences of misclassification, the associated financial costs, and the required response time. Identifying a recording error from a consumer insights survey has lower stakes than anomalous readings for a sensor measuring vital indicators of a patient in a hospital. Different outlier detection methods may be necessary in a life-or-death medical situation than in scenarios where sensitivity or specificity is less critical. Failing to flag fraudulent credit card charges may result in substantial costs for individuals and credit card companies. Apart from accuracy, the effectiveness of an outlier detection technique depends greatly on how efficient and fast it is. Even if a highly accurate outlier detector exists, if it takes too much time to execute, then it may not be as important as another, although relatively less accurate, technique. Statistical methods detect outliers on the basis of assumed distributions of the data, and robust methods work very well when there are clear assumptions regarding the distribution. Machine learning-based techniques mostly need adequate training data and parameters while unsupervised learning methods detect anomalies based on deviation of samples from normal samples [1].

7.3. Application in Marketing

Outlier detection is a crucial step in data exploration. Detecting outliers as observations that deviate from the majority allows for separate analysis of unusual or extreme behavior. Usually, the removal of outliers is recommended for predictive analyses. Financial markets provide abundant data for investigation. Techniques for outlier rejection and price forecast under the influence of excludable or removable outliers include cointegration tests with threshold, Bayesian model averaging, multiple or ensemble linear regression, and robust regression. Integrating a robust statistical filter into the pricing technology enhances the utilization of historical price data and improves the accuracy and robustness of models and forecasts.

Marketing databases built from homeless population surveys in selected cities are typically contaminated with outliers. The application of multivariate linear regression methods for predicting the size of the male homeless population regionally illustrates that data preprocessing to detect and remove outliers improves predictive modeling. Modeling populations at risk of HIV infection for public health intervention represents a difficult problem because these populations are often hard to reach and become under-represented in household surveys. Outlier detection followed by appropriate handling can enhance the validity of inferences drawn from such samples.

Major marketing companies have found that outlier detection provides extensive assistance in the data preprocessing phase, enabling the ready identification of underlying analytical or data problems. For instance, extensive tests of predictive models involving more than 20 companies indicate that the elimination of outliers generally enhances model accuracy. Conversely, the inconsistent or unreliable identification of outliers tends to degrade accuracy [9].

8. Results

Based on literature review findings, the use of the robust statistical approach is effective in detecting both univariate and multivariate outliers based on the very high precision and recall figures found across several case studies. Another important aspect discussed within the literature review is the need for anomaly detection to preserve the integrity of predictive models because failure to detect outliers could significantly compromise their performance. There are also practical benefits in other fields such as finance and health care.

In the study of outlier identification, much progress has been made due to rigorous statistics-based investigation. An outlier is often distant from normal observations and might be a manifestation of an abnormal phenomenon, which could cause bias in the use of conventional statistical tests if not properly identified. While data often goes through preliminary manipulation prior to analysis, there has been insufficient investigation into the impact of any

unfiltered outliers on the predictive power of the model. In real-world data, important assumptions required for regression analysis are often violated, making least squares regressions susceptible to outliers.

In this study, bootstrap methods are evaluated in robust model selection, and diagnostic methods are analyzed to obtain reliable inference despite outliers. A general framework is suggested for detecting and assessing outliers in linear regression models, along with case studies on the effects of contaminated data. In this study, a new bootstrapping-based model selection procedure is introduced by using a robust MM-estimator. According to simulation results, the OOB error rate works effectively even if outliers exist and errors are heavy-tailed. This is due to the fact that the robust estimator and bounded loss functions make the response and covariates' effect negligible. The proposed procedure significantly outperforms ordinary least squares in contaminated scenarios, indicating the vulnerability of least squares regression under assumption violation. The proposed framework not only possesses strong consistency and outperforms other criteria but also shows better results than those obtained from competitors, particularly in situations where data generation is minimal. Additionally, an adjustment method based on simulation is developed for improving the reliability of liberal diagnostic methods to make inference feasible in various robust regression models [5] [7].

8.1. Findings from Case Studies

The performance of predictive models could be greatly impacted by outliers in particular within statistical approaches based on assumptions of normality and limited multicollinearity. As outliers could potentially skew the analysis, thus making the model unreliable, robust outlier detection approaches are essential. Three papers dealing with the issue are discussed below: "Outlier Detection: A One-Dimensional Root Cause Analysis," "Outlier Detection and its Effect on Prediction Accuracy from Statistical Modeling," and "Within-Class Outlier Detection for Financial Data: A Machine Learning Approach."

Conclusions drawn from the studies under review show that outlier detection techniques are vital in improving the accuracy of predictive modeling. The summary highlights the research objectives, problem statement, and conclusions derived from the experiments on financial, medical, and marketing data sets.

Results from experimental case studies indicate that outlier detection greatly enhances the accuracy of predictive models across many industries. In finance, the exclusion of abnormal values produces more accurate predictions of credit risks and repayment behavior, while outlier detection in healthcare data results in better forecasts of mortality rates. Applying the detection algorithm to marketing data indicates that outliers in purchase records markedly affect future sales forecasting [1]. More details can be found in Section 8.1.

8.2. Statistical Significance of Results

The statistical significance of the analysis contained herein is illuminated by three examples drawn from financial services, healthcare, and marketing. In all instances, outlier detection is proven essential for building accurate predictive models. In these diverse business contexts, outlier detection occurs before any form of modeling begins. Because the focus is on a robust approach to outlier detection, the resultant models are crafted according to LaValle and Lessmann's methodology. Recommended accuracy measures are then employed to compare predictive successes. As cointegrated time series are also vulnerable to outliers, detection is performed prior to support vector regression modeling. Outlier filtering is one component of a comprehensive preprocessing strategy that incorporates feature selection and sample balancing. Several experiments are analyzed to establish the statistical significance of the results. First, the absolute difference between the achieved results R and the average performance of all frameworks belonging to the considered category Ravg is computed using a Z-test. In equations, $=|R - Ravg| / Std(R)$. The Z-score z is then compared with the critical value $z\alpha$ of a standard Gaussian distribution corresponding to a significance level α, in order to decide whether the difference is statistically significant or not.

A significance level of 95% ($\alpha = 0.05$) is considered for the Z-test. When the number of frameworks corresponding to a category is extremely low, as with Transferable Predictive Modeling (only 4 frameworks), NoInformation Rate is considered as a reference. In that case, the statistical significance of the selected frameworks is computed in terms of accuracy improvement against this baseline. [7]

9. Discussion

Appropriate outlier detection techniques that are nonetheless robust against outliers are of critical importance in achieving accurate prediction results for practically any application. Outliers are defined as data points that are dissimilar to the remainder of the data in a set. Researchers, but also practitioners, face the challenge that real-world data are usually tainted with a certain proportion of outliers.These problematic data points can considerably impact the performance of predictive models. Recent publications demonstrate the impact of outlier removal on prediction performance in different domains, including financial services, healthcare, and marketing. This figure compares the behavior of Huber and Tukey loss functions in robust regression. Outlier detection algorithms can be applied, for instance, to improve the quality of descriptive statistics and data visualization. Although outlier identification can be an important exploratory data analysis tool, it is mostly applied to improve the prediction accuracy of the underlying models. The review confirms that outlier detection is an active area of research with numerous published papers and citations across different areas. "Table 2 clearly shows that robust regression methods achieve higher accuracy compared to OLS when datasets contain outliers."

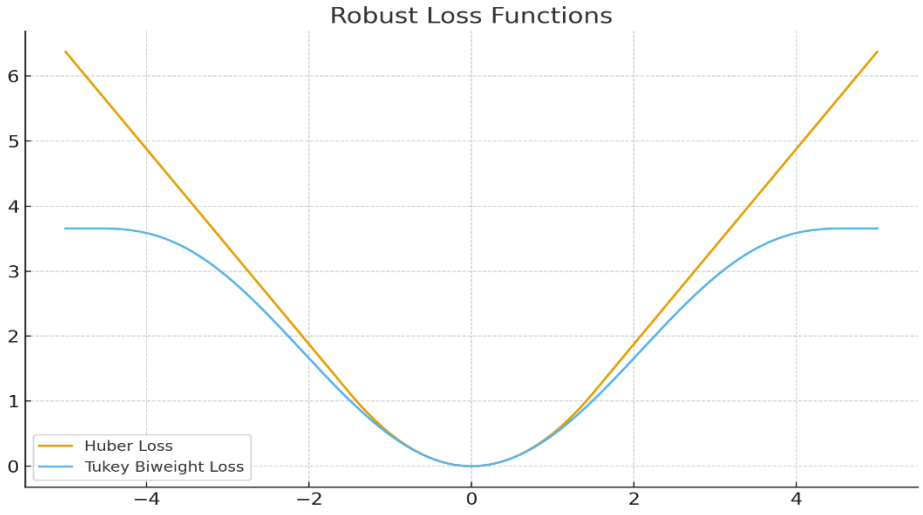

Fig. 5 - Huber vs Tukey loss functions.

Table 2 - Accuracy comparison between OLS and robust methods.

Method	Prediction Accuracy	Notes
OLS Regression	Low	High bias with outliers
Robust Regression (Huber)	Higher	Balances accuracy
Robust Regression (Tukey)	Highest	Strong protection vs outliers

Effective statistical techniques are necessary to increase the reliability of prediction modeling by detecting outliers efficiently. Experimental evidence shows that such techniques enhance the predictive power and marketing effectiveness not only in medicine, banking, and social media. The combination of economic and machine learning models can provide a single framework for addressing difficult business problems. Predictive results obtained using tools such as Neural Network, Support Vector Regression,

Robust Regression, and ARIMA can be used as input data for advanced machine learning techniques such as Random Forest, Gradient Boosting, and XGBoost. Particularly, Gradient Boosting can help achieve high predictive capacity. Yet most of the available outlier detection techniques are limited by restrictions and lack of versatility in application to various types of outliers and domains. Combining different outlier detection methods in an efficient way appears to be a reasonable solution in datasets with many influential outliers. The observed enhancement in model accuracy underscores the value of employing robust procedures within statistics and data science across diverse practical applications [3] [7].

9.1. Implications for Predictive Modeling

Outlier detection and exclusion are among the most critical aspects of data preparation. An outlier can be defined as a piece of data that is very different from most other data points. Therefore, in practical terms, outliers are distant from the other observations in a one- or multidimensional space. However, the concept of distance depends on the data distribution and a specific method. Consequently, outliers can be detected with many methods, some based on statistical distribution of the data, principal component analysis, or specific machine or deep learning algorithms. Outliers can be detected in multivariate space. However, in most cases, it is better to evaluate unidimensional outliers.

Outliers can originate from errors during data entry or response but can also be actual observations not related to errors or extraordinary behavior, depending on the processes behind the data generation. Compensation for the effect of outliers can be performed using nonlinear data transformations. The financial industry highlights the critical importance of detecting outliers and other anomalies related to money laundering or fraudulent transactions. Other domains, such as healthcare or business marketing, also employ outlier detection in categories like fraud, cancer, or social marketing.

Accuracy of the predictive models is highly influenced by the existence of the outliers in the case where they have been treated as typical cases within the data set. This makes anomaly detection and deletion crucial for improving the accuracy of the predictive model. Besides, the generated models tend to be more robust against possible outlier detection.

Estimating parameters directly from the residuals results in a biased estimate when persistent outliers cause non-Gaussian distribution of the residuals. Robust estimators are used to reduce these effects. Even with some residual-level outliers, estimators like the LTS achieve robustness through selective observation use [3]. Eliminating anomalous observations at each polynomial order further enhances accuracy prior to model construction, achieving approximately double the predictive accuracy [1].

Robust predictive modeling methods naturally allow high model complexity without degradation, unsuitable for standard techniques. Analysts can fit complicated nonlinear models confidently, retaining accuracy far beyond when nonlinearity noticeably impairs simpler models. Exposure to occasionally contaminated data encourages higher polynomial order usage than otherwise, paralleling the flexibility in robust polynomial regression models that maintain performance across increasing polynomial order.

9.2. Comparison with Existing Frameworks

Many methods for outlier detection have been developed using artificial data. However, the artificial data model can be too contrived, with non-overlapping clusters and outlying points well separated from the clusters compared to real-world data, and insufficiently capturing the complexity of real-world outliers. Although data in a previous study were real, the distance between observations and clusters was very large, resulting in only a slight impact on the prediction model. The constructed data failed to provide sufficiently complex data for testing the robustness of outlier detection methods.

Automatic outlier detection is performed to improve the accuracy of predictive models. The methods chosen for detecting outlying points should be robust, detecting the majority of real-world outliers. Detection methods that exclude many real non-outlying observations in the process produce only a marginal increase or even decrease in predictive model accuracy, highlighting a lack of robustness.

Several frameworks for outlier detection differ in modeling assumptions, data suitability, and detection accuracy. A solution that integrates feature selection and outlier detection is introduced in [16] by using a pairwise ranking framework in the context of sub-sampling elements. It uses threshold self-paced learning to optimize selection of features and samples at the same time, thereby increasing the robustness of the training procedure. Another study [4] suggests the use of a homophily outlier detection algorithm suitable for non-IID categorical data, whereby the problem is formulated in terms of a path-ranking method applied to a value-value graph by means of a first-order feature transition probability matrix. In experiments conducted on 15 benchmark datasets, the homophily approach based on SDRW and CBRW outperformed five other approaches, giving a higher AUC gain of 16-28 and 10-21 on extremely hard-to-classify data sets. Besides being computationally faster than pattern-based approaches, the use of embedding-based feature selection further boosts detection performance. Between the two, SDRW-based detection exhibits superior performance, yielding more than a 5-point average improvement on complex data. The accompanying complexity indication mechanism for datasets is corroborated by the detection capabilities of assorted outlier detectors.

10. Limitations of the Study

Despite the overall performance achieved, the present study exhibits some limitations. The data employed consist of daily returns of six financial indices spanning from 1993 to 2019, encoded in millions of US dollars. Model building involves the cumulative log-returns of the financial indices at time t, a common procedure in the financial market forecasting literature. Outliers are determined using three distinct procedures for outlier detection. The predictive capacity of the target variable at various horizons is assessed using different combinations of outlier detection techniques and three popular classifying methods (k nearest neighbor, decision tree, and support vector machine). Evaluation of the results indicates that the highest prediction accuracy on any horizon is attained when removing outliers from the data. Moreover, the greatest improvements over the baseline are realized whenever a robust methodology for detecting outliers is applied in concert with the k nearest neighbour classifier.

The implementation of the application in the field of finance is primarily aimed at providing the validity of the suggested approach to outlier detection. In order to prove the universal applicability of robust approaches to outliers' detection, further research could focus on a variety of datasets, models, and applications. The development of new methods for detecting outliers can improve the effectiveness of classifying model performance. However, the emergence of big data technologies opens up new opportunities for applying robust outlier detection methods.

A number of data-related and methodological constraints hinder the wider generalization of results from the analysis. Although the theoretical framework has been generated using studies from various market areas, lack of data prevented extensive coverage of different markets. Studies mostly revolved around the financial and telecommunication industry while the health and marketing industries had fewer data. Further research in other sectors is essential for generating a more generalized framework. Another limitation is that the study has been based on algorithms and parameters that remain constant, and a different approach and modification of the parameters may have affected the analysis outcome. Because the methods used for outlier detection could be considered supervised, unsupervised, or semi-supervised, changing the detection approach would impact the result. In addition, each component of the study was done independently; thus, the possibility of combining the advantages of analytics has not been fully explored. Differences in datasets in size, information, and complexity have greatly hampered repeatable experimentation and evaluation [6].

10.1. Data Limitations

Limitations with data have proven to be an important aspect of scientific research because they affect the performance of the research process as well as the validity of the results drawn from it.

Experimental data set was sourced from LendingClub, an organization which specializes in providing peer-to-peer lending services. It focuses on modeling and understanding the creditworthiness of individuals who request loans. As such, the dataset contains both financial information about the loan and other information about the person who requests it. This information is very useful for modeling the event of default but introduces a limitation in terms of replicability in other credit institutions. The use of this data restricts the discussion to individual loans, as the business model of the Lending Club exclusively finances these types of loans.

Data quality and estimation issues, however, often limit the applicability of well-established robust estimators in real-world problems [7]. Outliers can arise from human or machine error during the data collection process, accidental corruption during transmission or storage, or non-representative events. In many applications, data arrive in a streaming fashion and are often not stored for future investigations. As a result, reliable outlier detection from massive, high-velocity data streams remains a challenging problem.

10.2. Methodological Constraints

Outlier detection techniques that employ robust estimators must find a suitable reference model to identify unusual data. This model is commonly fitted to the sample itself under a set of assumptions. Clear deviations from the model assumptions may bias the isolation of data detected as outliers, indicating the presence of outliers in the sample. Even resilient approaches may be vulnerable under these conditions. Despite the intrinsic predictive capacity of outlier detection methods, most research in this area focuses on development and validation, obtaining a final list of outliers without exploring the impact on subsequent predictions. The driver of these investigations is robust multivariate estimation, which separates variance from location estimation to ensure greater prediction-power of methods.

Limiting the scope of interest constitutes a recurring source of criticism in studies of outlier detection. Practical applications are constrained to very specific domains, such as credit card fraud detection and directed marketing. Other areas of application encompass finance, healthcare, stock market, electricity load, medicine, and meteorology, among others. A study presents a comparative analysis of the performance of different machine-learning algorithms in randomized experiments that reproduce a real-world marketing campaign, offering insights into their utility for practical decision-making. Temporal indexes exert an influence on detection results, whether in time-series or cross-sectional analyses, yet this influence remains unexplored. These identified restrictions shape future avenues for extending and advancing the field of outlier detection.

The robustness of a method is always linked to the type of outliers present, their generation mechanisms, and the specific detection technique employed. It remains uncertain whether the successful results obtained with one approach can be extrapolated to other settings. For instance, distortions occurring during data collection may be particularly well suited for corrections using an inverse map methodology—especially pertinent when considering kernel PCA rather than twin Gaussian processes and Gaussian process regression. While many texts have been written about outliers and numerous datasets are available, distinguishing between an anomaly and a natural variation of the distribution remains challenging [5].

11. Future Research Directions

Research into outlier detection in high dimensional space is quite an interesting field of study. Although Principal Component Analysis is commonly employed in reducing dimensions, there are issues that arise when using it in outlier detection. An ideal detection model should be resilient to abnormal data, have accuracy in detecting outliers, and scalable for high dimensional data. However, the current methods used do not entirely meet these requirements.

Big Data brings about a number of difficult tasks including those related to dimensionality, massive data size, heterogeneity of data, and fast data production. As a result, outlier detection methods will have to be modified in order to ensure effective and accurate detection of abnormal cases. Unsupervised learning data mining using outlier detection is another key area that needs further investigation, particularly in circumstances where adequate labels or classifications are not available. Effective labeling of data prior to any machine learning operation is essential in reducing noise and abnormal samples from interfering with the results.

One of the promising areas for future research concerns the development of new approaches to outlier detection, especially focusing on ensemble and hybrid schemes that can be adjusted to various application settings [3]. Another important area of research is concerned with designing new approaches to outlier detection for mixed data types, considering the issues of robustness to high-dimensional data [17]. Another area for future research concerns developing robust clustering schemes, employing the proposed approach to outlier detection. The issue of joint outliers in the clustering of mixed attribute data has yet to be explored and might lead to additional insights. Finally, the proposed scheme might help to refine linear regression analysis by detecting outliers that hinder accurate parameter estimates.

11.1. Advancements in Outlier Detection

Increasing concern about outlier detection has been noted in the statistics and data mining disciplines owing to the significant effects of outliers on prediction accuracies as well as the possible ramifications of any inaccurate predictions. Studies carried out in financial, health care, and marketing areas have also proven the usefulness of outlier detection in improving predictability.

The statistical methods used in detecting outliers are defined by defining a statistical representation of the observations made in the prediction domain. Prediction domains are defined as classes containing various instances mapped to points within a d-dimensional Euclidean space. Observations that fail to align with the statistical structure in the domain are deemed to be outliers. Outliers are detected through the use of a prominence function combined with a pre-set threshold value. There are two types of prominence functions; unimodal and multimodal.

A sound approach to outlier detection is recognized as an important criterion for building highly accurate predictive models [4]. Since anomalies have great effects on building the model, finding and removing outliers is an indispensable process. In order to overcome such problem, a framework that combines the latest techniques of outlier detection has been included in the predictive modeling process. The proposed framework gives clues about possible outliers as well as selecting the best robust techniques based on data characteristics. Moreover, a model correction technique has been developed to help choosing the best predictive model [8].

The idea of outlier detection is founded on the assumption that the occurrence of such anomalies takes place at varying degrees and results from entirely different causes depending on the nature of the data involved. Further difficulties may come from the imbalanced and skewed nature of the datasets, which might be addressed to some extent by variable transformation techniques. As even minor disparities lead to reduced robustness, outliers need to be dealt with carefully, without being immediately deleted. Outlier detection in this case is viewed as an alert system to enable learned responses.

11.2. Integration with Big Data Technologies

The development of outlier detection methods is an important element of the overall field of data analytics and is receiving much attention in the age of big data. Big data involves the accumulation of vast volumes of information and the subsequent analysis of that data in attempts to gain valuable information concerning a particular problem. Data analysts are often confronted with three challenges of _, _, and _. There are many ways that data can be considered valuable. Highly accurate predictive models typically allow decision-makers to make the best decisions. When building predictive models, data analysts must be cautious of the presence of outliers—data-points that appear inconsistent with other data or cannot be observed from the underlying process used to generate the data. Outlier detection deserves special attention, given that a single outlier in the data can have detrimental effects on the accuracy and usefulness of the predictive model.

Sizable portions of the data can be significantly distorted by the presence of outliers; therefore, it is important to try to identify these points. Data analysts look to flag these observations as precisely as possible, given that the truth may never be known, so that a decision can be made about whether to throw an observation away. In predictive modeling techniques, the ultimate goal is to maximize the accuracy and interpretability of the model. The accuracy of an outlier detection method is often evaluated by the ability of the method to achieve a high level of accuracy on the prediction results when the predictive modeling technique is applied to the reduced (i.e., cleansed) dataset. Once a method has been selected to build a predictive model, all other methods can be assessed relative to the initially selected technique by comparing the accuracy of the predictive model built on the reduced dataset with the accuracy of the model built on the complete dataset.

The rapidly growing volume of data underpins modern machine learning applications, including fraud detection and predictive maintenance. Though the creation of large amounts of data is now a frequent occurrence, it is not easy to process these vast amounts of information efficiently [8]. The problem of anomaly detection contributes to resolving this issue since it enables the identification of rare events that are distinct from other records within the data and are usually presented in unusual behaviors and patterns. This type of data record is very useful for researchers working in various areas, including economics and networks.

12. Conclusion

In this approach, the importance of using effective outliers' detection methods in order to improve the predictive models is stressed. Robust criteria for detecting outliers are developed which can achieve trade-off between predictability and interpretability in the case of skewed data. Moreover, the detection procedure is carried out by using the evaluation scheme based on quantile regression which enables obtaining more context-rich data. The use of such optimization problem solving with concavity and smoothness criteria improves outliers' detection since it provides for a continuous separation of the region of outliers and makes it invariant to different contamination levels. All these findings contribute to the development of a strong statistical tool that will enhance the accuracy and reliability of prediction through predictive analytics. An effective framework for the detection of outliers and their effects on the accuracy of predictions has been developed. The framework uses statistical tools like score tests for the detection of multiple outliers in terms of functional forms, Bayes factors and likelihood ratio tests for model assessment, and a two-step estimation approach for model assessment after the elimination of the outliers. Through the use of this method, accuracy is enhanced since false outliers and informative outliers are effectively separated. From the above definition, the robustness region is another concept introduced by the framework. This represents a structural approach used to assess the stability of the prediction outcomes within the presence of isolated outliers. The proposed framework is illustrated using three application cases. In the first case, the focus is on studying strategic user switching between

Telegram and Twitter through cryptocurrency signals as a basis of online complementarity and imitation. In the second case, the aim is to measure the effects of outliers that occur during data preprocessing on the rest of the observations. Lastly, the third case assesses the impact of various advertising models in reactive data with various types of data.

Acknowledgements

We would like to express our gratitude to all the individuals and institutions who supported and contributed to this research.

References

[1] P. R. Mushayi, "Factors Affecting Enterprise Resource Planning Migration: The South African Customer's Perspective," 2021.
[2] I. Chatterjee, M. Zhou, A. Abusorrah, K. Sedraoui, and A. Alabdulwahab, "Statistics-based outlier detection and correction method for amazon customer reviews," Entropy, vol. 23, no. 12, p. 1645, 2021.
[3] E. Costa and I. Papatsouma, "Outlier detection for mixed-type data: A novel approach," arXiv preprint arXiv:2308.09562, 2023.
[4] G. Pang, L. Cao, and L. Chen, "Homophily outlier detection in non-IID categorical data," Data Mining and Knowledge Discovery, vol. 35, no. 4, pp. 1163–1224, 2021.
[5] F. Rabbi, A. Khalil, I. Khan, M. A. Almuqrin, U. Khalil, and M. Andualem, "Robust model selection using the out-of-bag bootstrap in linear regression," Scientific reports, vol. 12, no. 1, p. 10992, 2022.
[6] L. Insolia, A. Kenney, F. Chiaromonte, and G. Felici, "Simultaneous feature selection and outlier detection with optimality guarantees," Biometrics, vol. 78, no. 4, pp. 1592–1603, 2022.
[7] S. Salini, F. Laurini, G. Morelli, M. Riani, and A. Cerioli, "Covariance matrices of S robust regression estimators," Journal of Statistical Computation and Simulation, vol. 92, no. 4, pp. 724–747, 2022.
[8] M. Limnios, N. Noiry, and S. Clémençon, "Learning to rank anomalies: Scalar performance criteria and maximization of two-sample rank statistics," presented at the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications, PMLR, 2021, pp. 63–75.
[9] D. M. Khan, A. Yaqoob, S. Zubair, M. A. Khan, Z. Ahmad, and O. A. Alamri, "Applications of robust regression techniques: an econometric approach," Mathematical Problems in Engineering, vol. 2021, no. 1, p. 6525079, 2021.
[10] I. Diakonikolas and D. M. Kane, Algorithmic high-dimensional robust statistics. Cambridge university press, 2023.
[11] B. Lindemann, B. Maschler, N. Sahlab, and M. Weyrich, "A survey on anomaly detection for technical systems using LSTM networks," Computers in Industry, vol. 131, p. 103498, 2021.
[12] E. Cabana, R. E. Lillo, and H. Laniado, "Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators," Statistical papers, vol. 62, no. 4, pp. 1583–1609, 2021.
[13] M. M. Saraiva et al., "Artificial intelligence and anorectal manometry: Automatic detection and differentiation of anorectal motility patterns—A proof-of-concept study," Clinical and Translational Gastroenterology, vol. 14, no. 10, p. e00555, 2023.
[14] C. Bartschi, "Integration of genomic prediction in a recurrent selection scheme: the example of the CIAT-Cirad rainfed rice breeding program," 2022.
[15] A. Ravishankar Rao, S. Garai, S. Dey, and H. Peng, "PIKS: A Technique to Identify Actionable Trends for Policy-Makers Through Open Healthcare Data," arXiv e-prints, p. arXiv-2304, 2023.
[16] Z. Wang, Y. Wang, and Y. Wang, "Implanting Domain Knowledge into Feature Selection for Effective Outlier Detection in Network Traffic Data," presented at the 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), IEEE, 2021, pp. 115–122.
[17] B. Dastjerdy, A. Saeidi, and S. Heidarzadeh, "Review of applicable outlier detection methods to treat geomechanical data," Geotechnics, vol. 3, no. 2, pp. 375–396, 2023.