Journal of Computer Sciences

Original Research Paper

A Novel Anomaly Detection Approach for Nifty Stocks using Machine Learning for Construction of Efficient Portfolio to Reduce Losses and Protect Gains

Virrat Devaser and Priyanka Chawla

Department of Computer Science and Engineering, Lovely Professional University, Jalandhar, Punjab, India

Article history
Received: 17-02-2022
Revised: 10-05-2022
Accepted: 12-05-2022

Corresponding Author:
Priyanka Chawla
Department of Computer Science and Engineering
Lovely Professional University,
Jalandhar, Punjab, India
Email: priyankachawla.cse@gmail.com

Abstract: Machine Learning is the most essential and widely utilized methodological approach these days for performing, organizing, and analyzing data using a variety of approaches to produce correct and efficient results. Machine learning methodologies are also applicable to stock markets, which have grown multi-fold in recent years, and with the amount of money involved, the possibility of manipulation is always present and has increased. Machine learning techniques can be used to detect anomalies in price behavior or price movements that are out of the ordinary. We investigated how to detect abnormalities utilizing a combination of fundamental and technical aspects in this study. We used multiple machine learning approaches to detect various types of abnormalities using fundamental and technical factors integrated into stock market data in the current proposed study work. The results have been produced utilizing a bagging strategy that included the use of a class One support vector machine, a local outlier factor, and other techniques. Companies tend to have varying valuations across sectors, but generally follow a range for a specific industry type; hence the data sets have been divided by sector. The results were segmented into the industry type, such as banking, cement, and energy. The portfolio has been built using anomaly scores. The method assisted in the removal of equities from the portfolio, avoiding losses and preventing profits from eroding. The mean absolute error has been determined to be 10.22%, which enhances the whole system's ability to detect anomalies.

Keywords: Machine Learning, Support Vector Machine, Data Mining, Stock Markets, Anomaly Detection, Class One Support Vector Machine, Local Outlier Factor, Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, Portfolio Building

Introduction

Machine learning is a branch of Artificial Intelligence that provides way to learn and adapt patterns in data; machines can learn how to tackle a specific problem without the help of anybody else, Agrawal and Agrawal (2015); Hussein *et al*. (2015). Machine learning allows machines to conduct intelligent tasks that would otherwise require human contact by applying intricate scientific and factual devices. The focus and zeal have always been on making systems learn from data and this process of automating complex learning schemes has given rise to high optimism among systems analysts, who hope that a few tasks associated with the structure, activity, and correspondence systems can be executed on specific machines by making them learn efficiently from data. Among different systems and approaches, Chu *et al*. (2008), right now the focus of the most of study methods is on using Machine Learning, Sheta *et al*. (2015); Fasanghari *et al*. (2010); Jabez and Muthukumar (2015); Kim (2003). A few applications of machine learning methods in various systems include the area of anomaly detection in stock markets, which is our current area of the proposed research. In particular, the type of the work is of two types, to be specific, (i) we give an initial instructional exercise on the utilization of ML techniques and their application in the different systems, commonly known as supervised learning techniques, and (ii) we study the current data and let the machine learn on its own and then understand the learning and try to refine it using additional approaches, commonly known as an unsupervised form of learning, Wang *et al*. (2019); Luo *et al*. (2018); Pyo *et al*. (2017). The anomalies in general can be of different types in context to stock markets, but for the current research, we are focusing on the perspective of a long-term investor who plans to stay invested for many years for maximum returns. The stock market behavior itself is quite unpredictable so the biggest challenge is to differentiate between anomalies and regular price hiccups.

In the proposed study we focus on anomaly detection from the perspective of fundamental factors and not entirely on technical factors since the latter itself is very tricky to predict accurately in the short term. However, we have taken one technical attribute along with fundamentals to bring in dynamic decision-making. This difference specifies the depth of current research work rather than breadth, which itself can be very fast specifically in the context of stock markets, Parikh and Shah (2015); Spronk *et al*. (1997); Zopounidis *et al*. (1995).

Objectives of Current Research Work

1. The current research work identifies the anomalies in stock prices based on fundamental and technical factors combined
2. The proposed model uses class-one SVM and local outlier factor approaches on the transformed data which has been extracted sector-wise after cleaning the data set
3. The results are extracted using python and verification is done by constructing the portfolio. The built-in portfolio is cross-checked using the current approach and stocks found under anomalous behavior are moved out from the portfolio

Novelty

The approaches used for anomaly detection have mostly worked on price anomalies in the stock price of a company. The focus had been mostly on the technical movement of price trends. These days volatility could be high because of large amounts of money being involved in secondary markets so the probability of large movements on either side of prices is not ruled out. The involvement of algorithm-based trading has further made the issue complicated. The current approach tries to solve the above issue by combining fundamental and technical parameters, Bermúdez *et al*. (2007), further extracting the data set sector-wise as different sectors of the market tend to perform differently concerning changes in inflation, monetary policies, and other macro factors affecting the stock prices from time to time. The use of these two things makes this current research work bring novelty and has more scope to explore further for future research work which can be carried out in this field.

The proposed research takes a data set of stocks listed on the national stock exchange and a cleaning mechanism is applied to data for changes in face values, splits or bonus issue concerns, Andrysiak *et al*. (2018); Samaras *et al*. (2008); Sood and Devaser (2015); Yunusoglu and Selim (2013). The fundamental factors including one factor for technical factors are combined to study the anomalies which will be detected using the class one support vector machine and local outlier factor. The anomaly detection will be applied to the data set sector-wise as different sectors tend to be behaving in trend, if the detection approach is applied to combined data, in that case, the mean absolute error will be more, so to improve the accuracy in the current research and to eliminate false positives (false anomalies in the current context).

Background and Terminology

Machine learning techniques are these days being used to solve various types of system problems and no doubt they provide a lot of accuracy in comparison to other popular traditional ways of doing research like statistical-based methods. The machine learning computations can be implemented in various ways and we separate the calculations in the form of fundamental classifications, supervised learning, unsupervised learning, and semi-supervised learning. Machine Learning (ML) calculations have been effectively executed on different issues. The techniques it has been used already in intrusion detection systems, credit card fraud detection systems, and also for anomaly detection in stock markets where they can be used in a wide variety of ways, Wang *et al*. (2019); Sycara *et al*. (1996). In the recent past studies related to fraud detection systems have been extensively used for financial applications.

Supervised Learning

Supervised learning is a technique in which samples are classified proir to the application of learning technique. The objective is to find the estimation of at least one yield factor given the estimation of a variable of info factors x. The yield variable can be a nonstop factor or then again a discrete variable. A preparation informational index contains N tests of the information factors. These learning strategies develop a capacity $y(x)$ that permits to anticipate the worth of the yield factors about another estimation of the sources of info, Hegazy *et al*. (2014). This learning is separated into two main types of models, depicted beneath: A parametric model, where some number of variables is used in the model is set, and in the non-parametric model, where the number of variables is subject to the preparation.

Parametric and non-parametric models: Given a data set, it is prepared by cleaning the data for various issues depending upon the type of data set and its behavior, normally could be requiring the removal of certain incomplete tuples or completing the incomplete tuples by finding the normalized or actual values. In the current research work, it was a challenge to clean the data as different stocks got face value split or bonus issues, etc.

For parametric-based models, distribution should be normalized, which gives an effective technique to assess the hypotheses based on some parameter. They are statistically more powerful as compared to their counterparts.

In non-parametric techniques, the different parameters rely upon the input data set. They do not rely on distribution and are quite effective in their domain of usage. Both can be utilized for relapse and order issues. On account of k-closest neighbor strategies, all preparation information tests are put away (preparing stage). One can choose the best estimation of k, cross-approval may be utilized, Ahmed *et al*. (2017a).

Unsupervised Learning

The unsupervised learning approach can be described as working like an informal community investigation where data labels before learning are not put. On account of unaided learning, the preparation data set comprises just a lot of info vectors (data samples). While unaided learning can improve various errors, bunching or group examination is the most widely recognized. In this, we rely on two calculations, k-means and the semi-supervised model as tests of parceling approaches and model-based methodologies, individually. k-means is maybe the most notable clustering algorithm. It is an iterative calculation beginning with an underlying allotment of the information into k groups, Ahmed *et al*. (2017b). At that point, the focal point of each group is registered and information focuses are relegated to the bunch with the nearest focus. In the wake of introducing the parameters and assessing the underlying estimation of the log probability, the calculation shifts back and forth between two stages.

Other Techniques

Semi-supervised learning techniques need further input as discussed above as they are not complete always in which a large portion of the preparation tests are not labeled, whereas just a few marked information focuses exist. Semi-supervised learning is utilized for a similar sort of utilization as supervised learning. Self-preparing is the most established type of semi-managed learning, Ahmed *et al*. (2016). It is an iterative procedure; during the principal arrangement just labeled information focuses are utilized by a supervised learning calculation. At that point, at each progression, a portion of the unlabeled focuses are labeled as per the expectation coming about for the prepared choice work.

Reinforcement Learning (RL) permits operators to investigate accessible activities and improve their conduct. It differs from supervised learning in a way that we don't label the data in input/output pairs. The approach works to make a balance between unknown and known. They can have their applications in game theory, control theory, operations theory, Information theory, and many other related streams.

Anomaly Types

Various techniques under anomaly detection can be used point anomalies: An object of data is an outlier if it is lying away from the others. For instance: Detecting excessive rise or fall in price without any reason.

Context-based anomalies: These types of anomalies are context-specific. This type of anomaly can be commonly found in time-stamp-related data e.g., Spending 5000 rs on shopping for things every day during a specific tour can be normal, but not otherwise.

Collective anomalies: These types of anomalies occur in related data instances concerning the entire data set, for example, ECG time series plot.

In context to stock markets the point anomalies can be abrupt price movement of a stock, whereas there is no detected reason or cause nor there any negative or positive news flow. The context-based anomalies can be abrupt price movements on the high volatility days when you see large price movements in stock prices on either side. It is very important to note here that it is natural to observe these price movements but they cannot be considered anomalies, Li *et al*. (2017). Similarly, collective anomalies will be observed in behavior or pattern on regular days or the days of large price movements and observing the collective anomalies. Table 3 illustrates the different types of learning approaches and the significance of data labels. However, it must be noted that it should not be taken as a hard and fast rule as the same depends upon a lot of other factors including the type of application.

Materials and Methods

Classification-wise anomaly detection: These methods depend on learning a classifier utilizing some preparation information to recognize outliers from others. The anomalous class is expected to be very uncommon and One Class Classifier (OCC) is used in typical examples. The new information point is contrasted and if it is altogether different it would be pronounced peculiar, Golmohammadi and Zaiane (2015). The classifier is found by picking a piece and utilizing a boundary to set the nearby space delimiting the shape of extraction of anomalies in the element space.

Clustering-based anomaly detection: These techniques expect ordinary occurrences to be close to the nearest centroid; along these lines information examples that are far off from the centroids are anomalous. Initially, a bunching calculation is utilized to distinguish centroids, second the separation of each datum occasion with the nearest centroid is determined. This separation is the anomaly score of each example, Golmohammadi and Zaiane (2015).

Nearest Neighbor-based anomaly detection: The standard thought is closest neighbor-based abnormality identification is that typical information occurrences happen in thick neighborhoods, along these lines information occasions that are far off from their closest neighbors are irregular. Utilizing kth Nearest Neighbor: In these strategies, the abnormality score of each occurrence is determined dependent on the separation to its kth closest neighbor.

The class-one support vector machine is another unsupervised learning approach that can be used in implementation which has been used in the proposed research work. They are effective for binary type classifications where most of the samples belong to one main class and the rest will be classified as outliers. They can also be used in cases where the minority class and the samples are comparatively less. Based on the idea of the methodology, class-one characterizations are generally appropriate for those assignments where the positive cases don't have a reliable example or design in the component space, making it hard for other arrangement calculations to gain proficiency with a class limit. All things being equal, regarding the positive cases as exceptions, it permits one-class classifiers to overlook the errand of separation and rather center around deviations from typical for sure is normal, Al-Hnaity *et al*. (2016). One should recall that the upsides of one-class classifiers include some significant downfalls of disposing of all of the accessible data about the larger part class. Thus this arrangement ought to be utilized cautiously and may not fit some particular applications. The scikit-learn library in python gives a modest bunch of normal one-class order calculations proposed for use in the exception or abnormality recognition and change location, like one-class SVM, Isolation Forest, Elliptic Envelope, and Local Outlier Factor. The Local Outlier Factor (LOF) calculation registers a score mirroring the level of irregularity of the vectors. It estimates the nearby distance deviation of a given information point concerning its neighbors. The thought is to distinguish the examples that have a considerably lower distance than their neighbors. The nearby distance is acquired from the k-closest neighbors. The LOF score of perception is equivalent to the proportion of the normal nearby distance of his k-closest neighbors and its neighborhood distance. The number k of neighbors considered, is more noteworthy than the base number of items a bunch needs to contain, so different samples can be nearby anomalies compared with this group, such information is for the most part not accessible, and taking neighbors = 15 to 20 seems to function admirably overall. At the point when the extent of exceptions is high, n_neighbors ought to be more prominent. The strength of the LOF calculation is that it takes both neighborhood and worldwide properties of data sets into thought: It can perform well even in data sets where unusual examples have distinctive fundamental densities. Outlier detection and Novelty detection can both be used for anomaly detection, Outlier detection can be unsupervised whereas Novelty detection itself can be semi-supervised. The one-class SVM has been presented for that reason and carried out in the Support Vector Machines module in the SVM. One-class SVM object. It requires the decision of a kernel and a scalar boundary to characterize the outskirts. The RBF portion is generally picked even though there exists no careful equation or calculation to set its data transmission boundary. This is the default in the scikit-learn execution. The nu boundary, otherwise called the edge of the one-class SVM, compares to the likelihood of finding anomalous observations.

Proposed Work

In the current work data set was taken for companies listed in Nifty and the data set consisted of technical as well as fundamental factors for the years 2000 to 2017. The dataset was taken from kaggle and nseindia.com. The parameters which have been explained in Table 2, were taken for the dataset. If a particular parameter is related to the technical movement of stock prices like simple moving averages or exponential moving averages, support or resistance concerning a trading range then these types of the parameter are classified as technical parameters. On the other hand, if a parameter is providing the information on fundamentals of a company it is classified as a fundamental parameter. In our research work, we have taken both types of parameters for the identification of anomalies present in the data set. As stocks belong to different types of categories like financial, Information technology, Oil and Gas, etc., a category attribute is added for all the stocks. Table 6 shows the industry categories.

For a particular sector, parameters may be having a separate set of values which may be classified as an outlier for some other sector. Therefore the category of a stock is also taken in the identification of outliers or anomalous values. The previous approaches have used the same types of parameters for all types of stocks overlooking the needs of a particular sector.

The class-one SVM approach is based on a support vector machine where the distance between points is maximized and the data points are separated in planes.

In SVM classification distance is computed using the equation, where the distance of a hyperplane equation.

$w^T\Phi(x)$ +b = 0 from a given point vector Φ (x_0). The purpose is to maximize the distance:

$$dH\left(\Phi(X_0)\right) = \frac{\left|W^T\Phi(X) + b = 0\right|}{\|W\|_2} \quad (1)$$

Here $d_H((\Phi(x0))$ is the distance of the plane from a given point vector where $\|w\|_2$ is the Euclidean distance for the length of w given by:

$$\|W\|_2 = \left(W1^2 + W1^2 + W1^2 + ...W1^2\right) \quad (2)$$

The Local outlier Factor (lof) works on the equation as shown:

$$\begin{aligned} &reachability - distance_k\left(A,B\right) \ \arcsin\theta \\ &= max\left\{k - distance\left(B\right), d\left(A,B\right)\right\} \end{aligned} \quad (3)$$

Algorithm1: Anomaly detection algorithm

Input: Datasets sector wise
Output: Anomaly scores
1. Identification of factors affecting the fundamentals of a company for long-term investing. (set of factors related to fundamental factors S1)
2. Identification of factors for Technical trends. (set of factors related to technicals S2)
3. Obtaining the data set for listed companies on parameters identified. (S1 and S2)
4. Performing the cleaning of the data set for parameters S1 and S2.
5. Adding the attribute category for each of the companies.
6. Extracting of the anomaly results by using the model for calculating the anomaly score by using the modified approach first through local outlier factor and then transforming the dimensionality and cross verify by applying Isolation forest and other approaches by bagging approach.
7. The combined anomaly score would be used for the selection of stock to become eligible in the portfolio, the more the score lesser weightage will be used for allocation.
8. Construct the portfolio for selected Nifty 50 stocks and verify the anomalous behavior using the proposed model-based implementation.
9. Reconstructing the portfolio after the anomaly score-based allocation to maximize the returns and prevent the existing profits from eroding.
10. Validate the approach with US ETF data

The initial cleaning process was applied to the dataset where the bonus/split and other related issues in price were normalized. The initial process of data cleaning was required and the data set was cleaned considering the following points. The face value split in prices was adjusted and corrected. Similarly, adjustments were required for the issue of bonus shares by companies, if this type of cleaning process is not carried out then wrong results are likely to be obtained while computing anomalies in data. The data set is taken in raw form before the actual cleaning of data. The results were extracted using scikit in python, using class one support vector machine, local outlier factor, isolation forest, and other methods. Figure 2 shows the position of data for companies in the energy sector. A support vector machine is a classification algorithm and using it in the form of a One-Class support vector machine, it is widely used to find anomalies in data sets as a single class is taken as main and others are classified as outliers.

The function has the ‘nu’ argument which controls the ratio of outliers in the data set. As achieving high accuracy in the field of stock market prediction is always a challenge hence achieving MAE of around 10 will also be considered better as extracted anomalies will allow saving hard-earned money of investors.

When we are doing the modeling using a one-class support vector machine, it captures the density of the main class.

The results achieved need to be segregated as per sector to which a particular company belongs as already explained above and also shown through results. The reason being in the Information technology sector companies were having very high Price to earnings ratios as compared to other sectors making them probable cases for getting detected as outliers. Another approach followed here can be used as leader-follower where sector-wise a leader can be identified as its parameter value can be taken as normalized. The sales growth parameter helps in identifying future winners as even if they enjoy higher price to earnings ratios through sales growth figures it can be normalized. Anomalous recognition can distinguish uncommon occasions, for example, occasions that happen once in a while and henceforth, of which you have almost no examples. The issue is at that point, that the typical method of preparing a classifier won't work. An anomalous score compares to those examples where the thickness of likelihood is "low" in other words we can classify this as an outlier. Models incorporate the checking of daily stock market data checking the fundamentals and techniques for predicting stock market winners, still unable to do so with accuracy, but the issue always is the safety of investment and hence to safeguard investment the mechanism to find anomalous behavior of stocks be correctly identified so that overall portfolio is saved. The price movements on either side can always be large on certain days but classifying these as anomalous will be false positive as this type of behavior could be normal when looking at regular stock price movements. To overcome all these types of issues and to address the issue of safety of long-term investors, our focus of current research is the fundamentals behind the listed stocks which are earnings, book value, and other related factors. The proposed research is not taking the data of tick trade or very short-term trade but is more focused on fundamental factors rather than technical factors. The recent fall of the stock market due to the COVID-19 impact has also strengthened this point as the recovery was fast and also good returns were there for long-term investors. The Algorithm1 which is used in the current work first identifies the factors related to stock markets and price movements, the factors affecting are fundamental factors for stock selection in the portfolio. The technical factors which control the short-term price movements ranging from a few days to a few weeks are selected in the next step of the algorithm. The moving averages of like 20-day moving average, 50-day moving average, 100-day moving average, and 200-day moving average are taken. In the current study, simple moving averages based on closing price are taken, however exponential moving averages can also be taken. After the cleaning of the dataset, the category attribute is added to identify the category of stock. The anomaly scores are computed by applying the class-one SVM and local outlier factor approach along with other approaches. The combined anomaly scores were used to construct a portfolio for the selected stocks. The constructed portfolio would be validated concerning price movements, which in a way would be more dominated by the technical attributes. A high anomaly score would lead to rebalancing the portfolio. The rebalancing exercise can be done weekly or monthly depending upon the fund managers or when the averages would signal trend changes. This aspect would lead to saving the occurred profits from vanishing away and makes the current approach dynamic in nature.

Table 1 shows the general flow of machine learning approaches being followed. As illustrated in the table the utility of each approach along with the types of output we can expect is illustrated. Table 2 specifies the parameters used in the current study along with their general description, as visible the current research takes into account both fundamental as well as technical parameters in the current study. The preparation data set as visible in the form of an array was transformed as a variable in the implementation tool which is python. Figure 1 shows the Nifty dataset statistics. The dataset used is for multi-years (2000-2017). Figure 2 shows the results of local outlier factor results for data of energy sector stocks. The classification of stock categories has been a major achievement of the proposed study as the parameter value ranges could be different for stocks depending upon their type of category. The results of local outlier factor results for data of pharmaceutical sector stocks were extracted along with class-one SVM. Figure 3 shows the results of class-one SVM results for pharmaceutical sector stocks. Figure 4 shows the results of class-one SVM results for data of metal sector stocks. The results of local outlier factor results for data of Industrial and manufacturing sector stocks were found consistent with the index. Figure 5 shows the results of local outlier factor results for data of services sector stocks. The outliers were cross verified with anomaly score obtained as it will be used as input for portfolio construction and rejig. Figure 6 shows the Cluster-based outlier detection results. Figure 7 shows the KNN-based outlier detection results.

Table 1: Machine learning techniques

Machine learning techniques	Supervised learning (labeled)	Classification	
		Discrete and qualitative	Input is data, output is class
		Regression continuous and quantitative	Input data, output is a number
	Un supervised learning (un labeled)	Clustering discrete	Input data, to find input regularities
		Dimensionality reduction	Input data, to find best lower dimensional representation
	Reinforcement learning	Policy	Agent learners (reward/penalty)

Table 2: Dataset parameters

Parameter	Description
Industry type	Sector of company e.g., financial or oil and gas sector of the company e.g., financial or Oil and Gas
Market cap	total market cap of the company
Sales growth	Average sales growth of the organization
Book value	Book value of the company
Price-Earnings P/E	Industry price to earnings ratio
Dividend yield	dividend yield of the company
Return on capital employed	Average return on capital employed
Return on equity	Average return on equity

Table 3: Data labels for types of learning

Type of learning	Features
Supervised anomaly detection	Labels are available for both normal data and anomalies, similar to skewed (imbalanced) classification
Supervised anomaly detection	Labels are available for both normal data and anomalies, similar to skewed (imbalanced) classification
Unsupervised anomaly detection	No labels are assumed, based on the assumption that anomalies are very rare compared to normal data

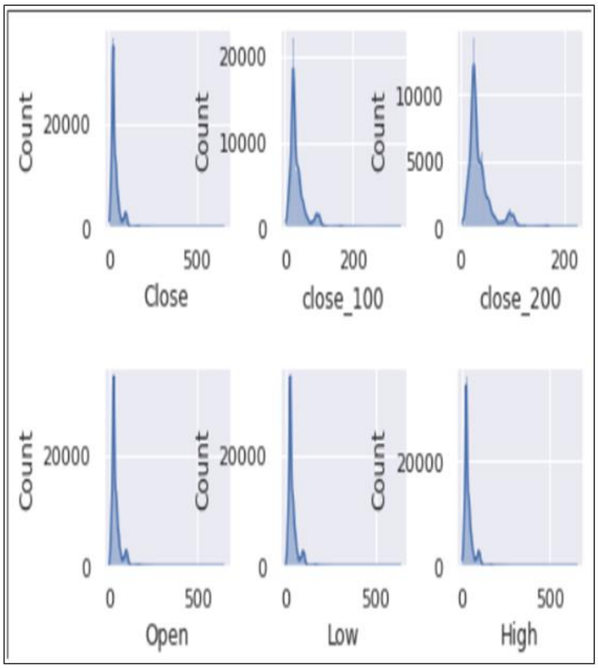

Fig. 1: Nifty dataset statistics

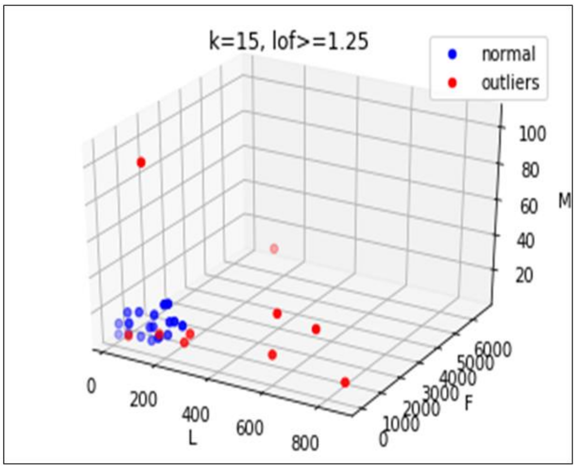

Fig. 2: Local outlier factor for companies in the energy sector

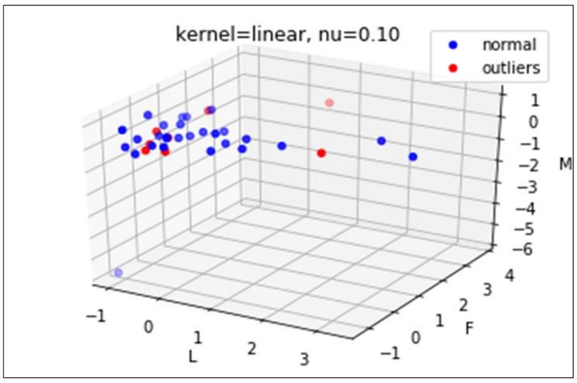

Fig. 3: Class-one SVM results for companies in the pharma sector

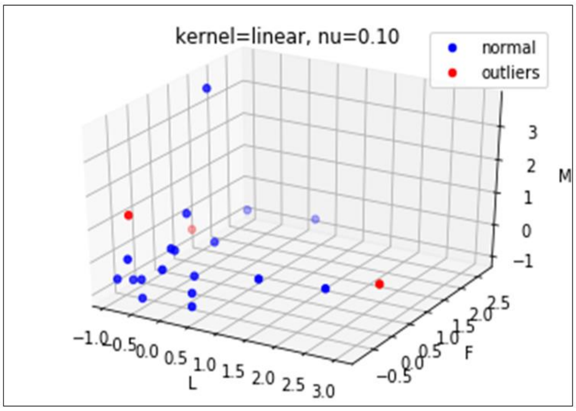

Fig. 4: Class-one SVM results for companies in the metal sector

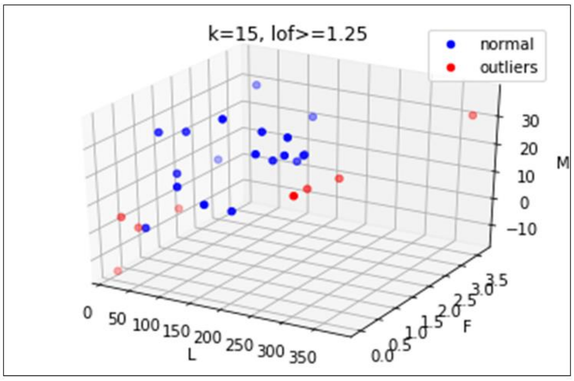

Fig. 5: Local outlier factor for companies in the services sector

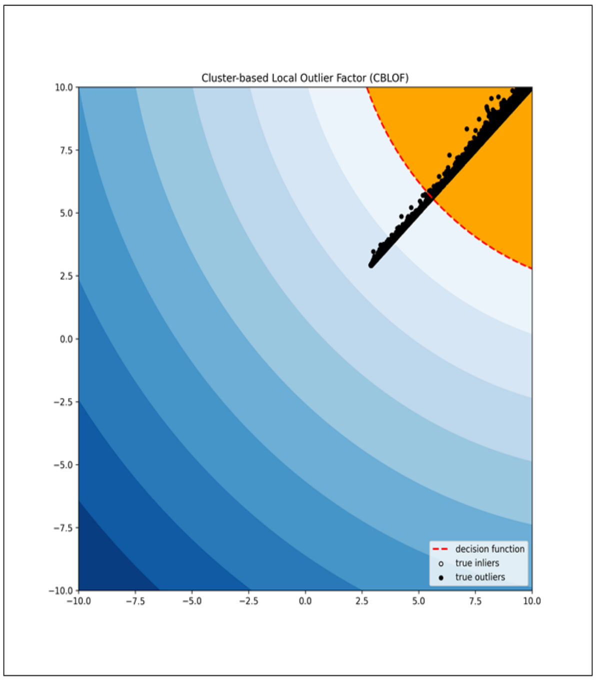

Fig. 6: Cluster-based local outlier factor

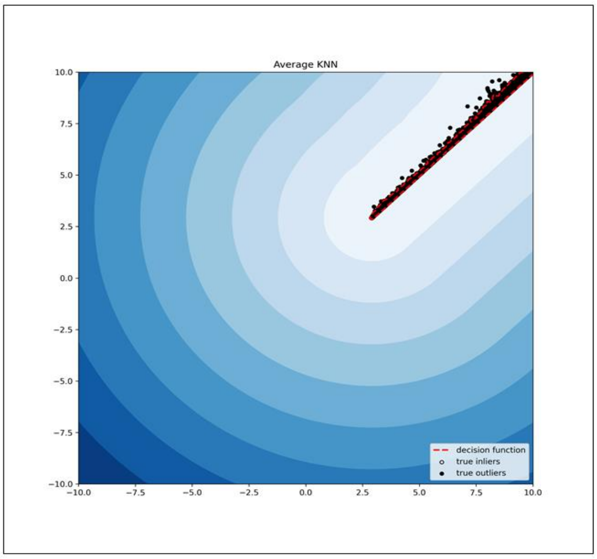

Fig. 7: Average KNN based outliers

Results

The data cleaning was done considering various aspects like stock price split due to bonus share issue or face value split and other aspects. These aspects continue to dominate the majority of the cleaning process. The categorization into different sectors is a key here as otherwise, it becomes very difficult to find anomalies as one parameter value may not be applicable for other categories of stocks. Further, as the current approach tries to find outliers it becomes essential to compare based on the same or similar parameters. As we know that stock market depends on a lot of parameters and often news flows tend to be carrying the momentum, but the focus of our study would be concerning the long-term investment. Considering that the impact of short-term movements on either side of the range generally does not bother long-term investors. The anomaly scores computed on different algorithms would identify the outliers and the portfolio constructed can be saved from the impact using anomaly scores by following rejig process.

Portfolio Construction and Rejig

Given the anomaly scores of all stocks eligible to be added to the portfolio the fund manager has a choice to review the portfolio allocation based on anomaly scores computed. A portfolio can be initially constructed as per the fund manager's discretion of allocation of funds to given stocks. The rejig of the portfolio can be done by reducing the stake or completely selling the stocks present in the portfolio whose anomaly score is high. Thus the current research work prevents the losses and also helps to prevent the gains as the stocks having high anomaly scores can be offloaded as desired. The current research work portfolio is constructed for some sectors as shown in Table 7. The funds are allocated based on a percentage as shown in Table 7. This initial percentage can be decided by the fund manager as per market capitalization and nature of the mutual fund scheme as if the scheme is targeting large-cap, mid-cap, or small-cap funds. The sectors selected for portfolio can be a subset of set of sectors. Table 4 shows the mean absolute error computed sector wise. Table 5 shows the list of parameters for cross verifying anomalies after the computation. Table 8 shows the updated portfolio after the rejig. As the scope of current research work is constrained to anomaly detection, our focus is on the rejig process after getting the anomaly scores. The rejig process will take the anomaly scores and the scores will be presented to the fund manager for taking a further decision on whether to continue or de-allocate funds to the stocks having high anomaly scores. The updated percentages, as shown above, are after the rejig due to anomaly score as some stock allocation gets changed. As our current study is sector-specific so it improves the efficiency of the overall portfolio.

Discussion

There has been a lot of research on anomaly detection in the field of intrusion detection, fraud detection, and related fields. The approaches of all have been different concerning the algorithm as some approaches have used support vector machines, time-series based predicting and forecasting can also be used for these types of problems where expected values against calculated values can be used for outlier case considerations. The proposed model in our research works on calculating the anomaly score based on outlier values and helps in building a portfolio for validation. The stocks having high anomaly scores will be removed from the portfolio and help save the gains and reduce further losses. The existing approaches have not constructed the portfolio for verification purposes, instead just shown the effectiveness of the approach on different data sets and standard data sets. The next level comparison is done on the ETF data of the US as the volume of the ETF traded in the US is quite large as compared to other economies of the world. Figure 8 shows the Angle based outlier detection. Figure 9 shows the ETF average close with respect to the number of days as the comparison is done with the US ETF. Figure 10 shows the ETF normal vs. anomalous using a Random forest ensemble for comparison. Figure 11 shows the ETF open vs. close using a Random forest ensemble. Figure 12 shows the Isolation forest-based outlier detection. Figure 13 shows the K Nearest Neighbors-based outlier detection results. Figure 14 shows Feature Bagging based outlier detection.

Table 4: Achieved mean absolute error sector wise for subset of sectors

Sector	Achieved MAE
Industrial manufacturing	1.364
Energy	2.832
Chemicals	4.794
Pharmaceuticals	6.063
Financial	10.895
Cement	3.058
Consumer goods	11.821
Construction	3.566

Table 5: List of parameters considered for cross verifying anomalies

Company name
Market cap
Current Price
Book value
Price/earnings ratio
Return on equity
Sales growth

Table 6: Industry categories

IT
Services
Financial
Oil and Gas
Chemical
Energy
Pharmaceuticals
Metals

Table 7: Portfolio construction by selecting few sectors

Sector	Allocation of funds in percentage
IT	15
Services	15
Financial	15
Oil and Gas	15
Chemical	10
Energy	10
Pharmaceuticals	10
Metals	10

Table 8: Portfolio after rejig by applying anomaly score

Sector	Allocation of funds in percentage
IT	20
Services	12
Financial	10
Oil and Gas	12
Chemical	10
Energy	10
Pharmaceuticals	11
Metals	15

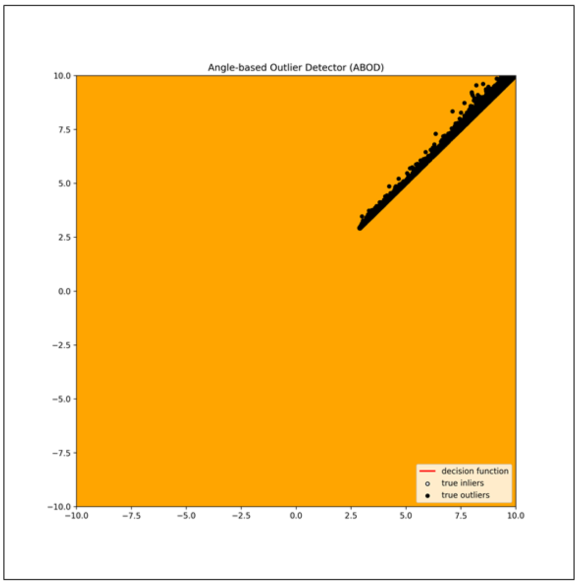

Fig. 8: Angle-Based Outlier Detector (ABOD)

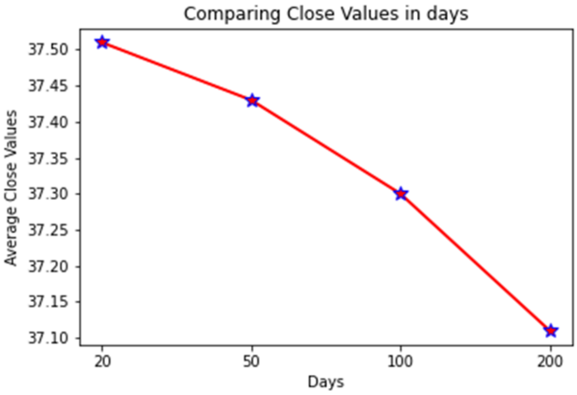

Fig. 9: ETF average close with respect to the number of days

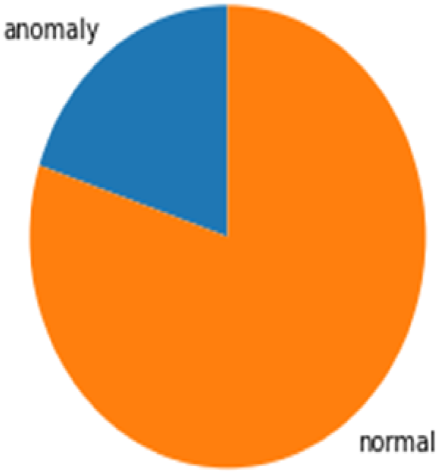

Fig. 10: ETF normal vs anomalous using random forest ensemble

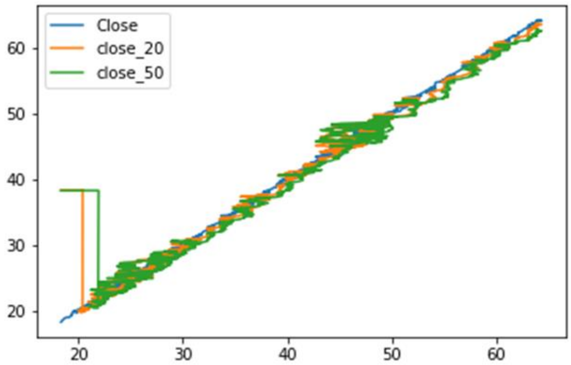

Fig. 11: ETF open vs close using random forest ensemble

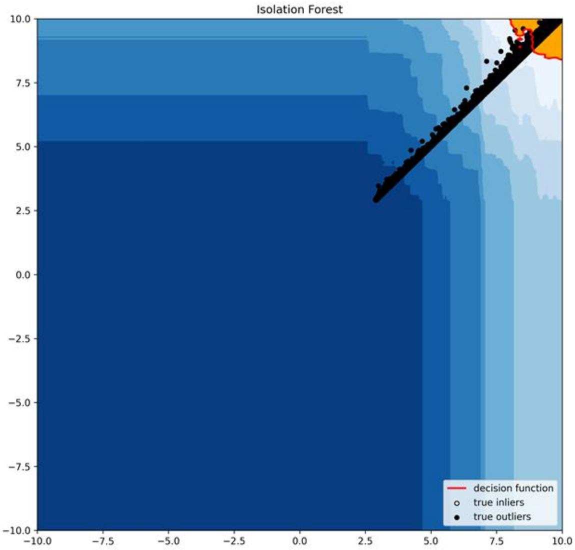

Fig. 12: Isolation forest-based outliers detection

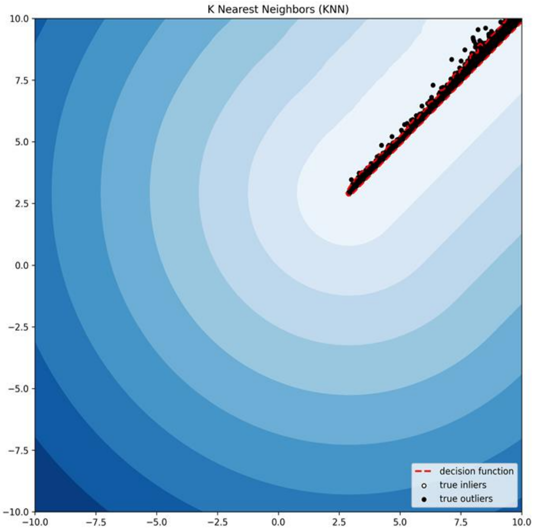

Fig. 13: K Nearest neighbors KNN based outlier detection

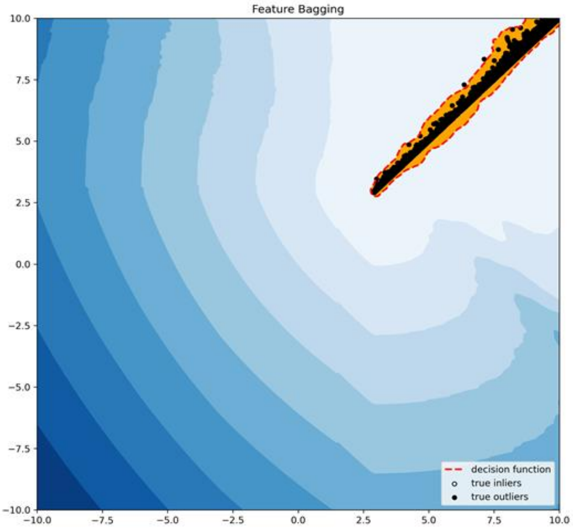

Fig. 14: Feature bagging based outlier detection

Conclusions and Future work

Over the previous decade, machine learning was used for various areas like stock market price prediction machines, learning calculations can utilize the huge amount of information accessible from a range of observing components to make them 'learn' for a fact and make the systems more spry and versatile. In the current work, we extracted anomalies using class one support vector machine for Nifty stocks by categorizing them in different categories as per sector thereby improving the accuracy of outlier detection as one of the biggest challenges in the field of outlier detection is that a particular rule if applicable to one sector may not be applicable for all sectors, for example, the interest rate sensitive stocks may behave differently in comparison to others. Further, that parameter applicable can also change considering a particular sector, if one company is having more sales growth and hence can be trading at a higher price-earnings ratio than companies in another sector. The outlier detection approach used here by applying class one support vector machine and others aim to reduce the risk associated with a portfolio as a significant amount of portion allocated to an anomalous stock can be reallocated to save losses to the overall portfolio. There are approaches available in portfolio allocation which tries to reduce losses by systematically distributing funds but our approach differs from them in that in a way our approach detects outliers or anomalous stocks which can significantly erode the gains of the portfolio for the long term investors. The current approach was tested against US stocks and ETF datasets to figure out the efficiency of the approach. The current approach is not for short-term traders as most of the factors under our study evaluate a stock from a long-term perspective like sales growth and identifies fundamental anomalies. Further, the concept of hierarchical anomalies can be used as it has shown good results for multi-dimensional data for dimensionality was very high.

Acknowledgment

The authors are thankful to the colleagues at Lovely Professional University for their valuable insights provided.

Author's Contributions

Virrat Devaser: Participated in data processing, data cleaning, literature survey, design of the methodology, and writing the manuscript.

Priyanka Chawla: Coordinated the experiments, result in analysis, and writing the manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and that no ethical issues are involved.

References

Agrawal, S., & Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. Procedia Computer Science, 60, 708-713. doi.org/10.1016/j.procs.2015.08.220

Ahmed, S., Lavin, A., Purdy, S., & Agha, Z. (2017a). Unsupervised real-time anomaly detection for streaming data. Neurocomputing, 262, 134-147. doi.org/10.1016/j.neucom.2017.04.070

Ahmed, M., Choudhury, N., & Uddin, S. (2017b). Anomaly Detection on Big Data in Financial Markets. Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017-ASONAM '17, 998-1001. doi.org/10.1145/3110025.3119402

Ahmed, M., Mahmood, A. N., & Islam, M. R. (2016). A survey of anomaly detection techniques in the financial domain. Future Generation Computer Systems, 55. doi.org/10.1016/j.future.2015.01.001

Al-Hnaity, B., & Abbod, M. (2016). Intelligent Systems andApplications.650,19-42. doi.org/10.1007/978-3-319-33386-1

Andrysiak, T., Saganowski, Ł., & Maszewski, M. (2018). International Joint Conference SOCO'17-CISIS'17-ICEUTE'17 León, Spain, September 6-8, 2017, Proceeding. 649. doi.org/10.1007/978-3-319-67180-2

Bermúdez, J. D., Segura, J. V., & Vercher, E. (2007). A fuzzy ranking strategy for portfolio selection is applied to the Spanish stock market. IEEE International Conference on Fuzzy Systems, 3–6. doi.org/10.1109/FUZZY.2007.4295466

Chu, H. C., & Hwang, G. J. (2008). Elicitation of time scale-oriented expertise from multiple experts. Second International Conference on Innovative Computing, Information and Control, ICICIC 2007, October. doi.org/10.1109/ICICIC.2007.281

Sheta, A. F., Elsir, S., & Faris, H. (2015). A Comparison between Regression, Artificial Neural Networks and Support Vector Machines for Predicting Stock Market Index. International Journal of Advanced Research in Artificial Intelligence, 4(7), 55-63. doi.org/10.14569/IJARAI.2015.040710

Fasanghari, M., & Montazer, G. A. (2010). Design and implementation of fuzzy expert system for Tehran Stock Exchange portfolio recommendation. Expert Systems with Applications, 37(9), 6138–6147. doi.org/10.1016/j.eswa.2010.02.114

Golmohammadi, K., & Zaiane, O. R. (2015). Time series contextual anomaly detection for detecting market manipulation in the stock market. Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015. doi.org/10.1109/DSAA.2015.7344856

Hegazy, O., Soliman, O. S., & Salam, M. A. (2014). A Machine Learning Model for Stock Market Prediction. 4(12), 7. doi.org/2047-3338

Hussein, A. S., Hamed, I. M., & Tolba, M. F. (2015). Intelligent Systems'2014. 322, 871–882. doi.org/10.1007/978-3-319-11313-5

Jabez, J., & Muthukumar, B. (2015). Intrusion detection system (ids): Anomaly detection using outlier detection approach. Procedia Computer Science, 48(C), 338–346. doi.org/10.1016/j.procs.2015.04.191

Kim, K. J. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55(1–2), 307–319. doi.org/10.1016/S0925-2312(03)00372-2

Li, A., Wu, J., & Liu, Z. (2017). Market Manipulation Detection Based on Classification Methods. Procedia Computer Science, 122, 788–795. doi.org/10.1016/j.procs.2017.11.438

Luo, J., Hong, T., & Yue, M. (2018). Real-time anomaly detection for very short-term load forecasting. Journal of Modern Power Systems and Clean Energy, 6(2), 235–243. doi.org/10.1007/s40565-017-0351-7

Parikh, V., & Shah, P. (2015). Stock Prediction and Automated Trading System. IJCS, 6, 104-111. http://csjournals.com/IJCSC/PDF6-1/21.%20Vishal.pdf

Pyo, S., Lee, J., Cha, M., & Jang, H. (2017). Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the Korean stock markets. PloS one, 12(11), e0188107. doi.org/10.1371/journal.pone.0188107

Samaras, G. D., Matsatsinis, N. F., & Zopounidis, C. (2008). A multicriteria DSS for stock evaluation using fundamental analysis. European Journal of Operational Research, 187(3), 1380-1401. doi.org/10.1016/j.ejor.2006.09.020

Sood, P. K., & Devaser, V. (2015). Stock price prediction for a sectorial leader in NSE using neural network. International Journal of Applied Engineering Research, 10(55), 2067-2074.

Spronk, J., & Hallerbach, W. (1997). Financial modeling: Where to go? With an illustration for portfolio management. European Journal of Operational Research, 99(1), 113–125. doi.org/10.1016/S0377-2217(96)00386-4

Sycara, K., Pannu, A., Williamson, M., Zeng, D., & Decker, K. (1996). Distributed intelligent agents. IEEE Expert-Intelligent Systems and Their Applications, 11(6), 36–46. doi.org/10.1109/64.546581

Wang, C., Liu, Z., Gao, H., & Fu, Y. (2019). Applying Anomaly Pattern Score for Outlier Detection. IEEE Access,7,16008-16020. doi.org/10.1109/ACCESS.2019.2895094

Yunusoglu, M. G., & Selim, H. (2013). A fuzzy rule-based expert system for stock evaluation and portfolio construction: An application to Istanbul Stock Exchange. Expert Systems with Applications, 40(3), 908–920. doi.org/10.1016/j.eswa.2012.05.047

Zopounidis, C., Godefroid, M., & Hurson, C. (1995). Designing a multicriteria decision support system for portfolio selection and management. In Advances in stochastic modeling and data analysis (pp. 261-292). Springer, Dordrecht. doi.org/10.1007/978-94-017-0663-6_17