Intelligent System for the Detection of Insider Trading in Indian Stock Market

Amosh Sapkota[1], Anand Kumar[2], Anjali Mathur[3]
{amoshsapkota@gmail.com[1], anand170030045@gmail.com[2], anjali_mathur@kluniversity.in[3]}

UG Student[1,2], Associate Professor[3]
Department of Computer Science and Engineering, KoneruLakshmaiah Education Foundation,Guntur(A.P), India, 522502.

Abstract. Insider trading is a pervasive stock market malpractice that has existed since the inception of the security market. Insider trading is notoriously difficult for regulators worldwide to crack down on. India has a dismal track record when it comes to prosecuting insider traders. In the last three decades of Sebi's existence, there hasn't been a single conviction for insider trading. In this study, we mainly inspectthe features of insider trading by examiningkey indicators during the time length before the release of insider information. In our investigation, we proposed a methodology for detection of insider trading in Indian stock market. To start with, the insider trading cases that happened in the Indian financial exchange we collected corporate filing data from NSE website for each company of NIFTY 50, which has different columns related to price, action and person or organization doing that action from 1stJanuary to a day prior to the publicationof financial results of December quarter offiscal year 2020-21. On doing as such, we have seen that enormous exchange have been done prior to publicationof financial results in some companies, which can be suspected as insider trading. At that point, themachine learning algorithms were utilized for preparing and for foreseeing Insider trading. Then, the algorithms were used for training and for predicting insider trading. Finally, their performance wasmeasured, compared and accuracy was calculated.Experiments revealed that the recommended method successfully achieved the best accuracy. This could be amazingly helpful for detecting insider trading in future, not only in Indian stock market, but also in other stock exchanges. The proposed approach and results in this examination is of incredible importance for market controllers to improve their oversight proficiency and precision on insider trading..

Keywords: Insider trading, security market, Sebi, corporate filing, NSE, NIFTY 50, financial results, machine learning, deep learning..

1 Introduction

The stock market is a marketplace for the buy, sale, and issue of publicly traded companies stock. As security markets have become larger financial market, more security frauds have also emerged. Among them one is insider trading, which is a practice of trading in a publicly held company's share by a person having non-public, material information about that company. It has a detrimental effect on investor's confidence and that impartiality in market hinders the imperishable development of security markets. The Securities and Exchange Board of India (SEBI) is India's securities and commodity market regulator,

ICCAP 2021, December 07-08, Chennai, India

DOI 10.4108/eai.7-12-2021.2314574

reporting to the Ministry of Finance. Sebi begins by determining who is an insider, which is typically the listed company's top executives, the company's board of directors, the auditors, the financial or information management staff, the promoters, and those associated with the promoters. Even these officials' close relatives are considered to be internally connected and thus considered insiders. The second critical factor is having a firm grasp on what constitutes unpublished price information. This could be anything from securing a large contract to achieving favourable financial results. Finally, it determines who traded using that data.

Insider trading is not illegal in India; some is legal, but it should be reported to the regulator, the SEBI. According to Regulation 7(2) of the SEBI (Prohibition of Insider Trading) Regulations, 2015, upon becoming a promoter or being appointed as key managerial personnel and director of a company, he or she is required to file a corporate filing of the transactions if the aggregate value of the securities traded, whether in a single transaction or a series of transactions over any calendar quarter, exceeds ten lakh rupees or such other value as may be specified within two trading days from the date of the transaction. This information is made available on the websites of the stock exchanges under corporate actions.

We gathered data for our research from the National Stock Exchange's (NSE) corporate actions section. We extracted data on all NIFTY 50 companies separately from 1st January to a day prior to the publication of those companies' December quarter financial results for fiscal year 2020-21. Principal Component Analysis (PCA) was used to reduce the dimension of the data and extract the features due to its size. The obtained data was then used to train various classification algorithms such as Random Forest (RF), Naive Bayes, and Decision Tree. Additionally, it was fed into a Dense Neural Network, which predicted whether or not the transaction involved insider trading.

2 Literature Survey

For the detection of insider trading, research[1] employs Extreme Gradient Boosting and Multi-Objective Optimization. To begin, an integrated system of XGboost and NSGA-II was used to automatically derive insider trading cases that occurred in the Chinese stock market in the past, as well as their relevant indicators. The proposed approach then used the NSGA-II to optimise the XGboost parameters via multi-objective functions, followed by training the XGboost model. Then, using XGboost with configured parameters, the test samples were identified. Efficiency and accuracy were evaluated across multiple time windows. The experiment demonstrated that the best accuracy was obtained with a 90-day time window. Here, the accuracy varied according to the time window length. As a result, this model cannot be relied upon to detect insider trading.

In the other work [2], an intelligent system was proposed with an integration of Principal Component Analysis and Random forest to detect insider trading in Chinese stock market. The proposed method began by collecting twenty-six relevant indicators for samples of insider trading that occurred between 2007 and 2017 and comparable samples of non-insider trading in the Chinese stock market. The indicator dimension was then reduced using PCA, and the principal components were extracted. The RF algorithm then developed an understanding of the relationship between insider trading samples and principal components. Finally, the PCA-RF model was used to categorise the samples of insider trading and non-insider trading. While the proposed method performed well within the 60-day time window, the sensitive period (time window preceding the release of insider information) should be taken into account when predicting insider trading.In the other work [3], patterns were discovered from large scale exploratory analysis of insider filings and related data, based on the complete Form 4 fillings from the U.S. Securities and Exchange Commission (SEC). Here, temporal and network-centric aspects of the trading behaviours of insiders were explored and made different

discoveries. 12 million transactions by 370 thousand insiders spanning 1986 to 2012 were analysed and studied how the trading behaviours of insiders differ based on their roles in their companies, the transaction types, the company sectors, and their relationships with other insiders.An unwarranted assumption like Form 4 filings can extract the hidden relationships between insiders was made. Second, the author claims that the insiders from same family ends to trade similarly. Due to reliance on many unwarranted assumptions, the insider network is flawed and lacks accuracy.

In the other work [4], a method was proposed to predict and detect illegal insider trading proactively (before releasing of official news) by analysing different sources of structured and unstructured data. A list of companies was identified by tree-based visualization and also by past data of SECas prominent illegal insider trading cases. From those targeted companies, a pattern of illegal insider trading was discovered. Later, LSTM RNN was used to forecast the stock transaction volume. Then their proposed algorithm ANOMALOUS was used to see whether it matches with the discovered anomalous patterns or not, to identify insider trading. This work figure on the concept history repeats, which may not be true every times.

3 Theoretical Analysis

3.1 Principal Component Analysis

PCA is a dimensionality-reduction technique for reducing the dimensionality of large datasets by converting a large collection of variables into a smaller one that retains the majority of the information in the large set.

While accuracy suffers when the number of variables in a dataset is reduced, the secret to dimensionality reduction is to trade off some accuracy for simplicity. Smaller datasets are easier to examine and visualise, and machine learning algorithms can analyse data much faster without having to account for irrelevant factors.

The mathematical modelling of principal component analysis may be broken down into six steps:

- Take the whole dataset with d+1 dimensions and remove the labels, resulting in a d-dimensional dataset.
- Calculate the mean of each dimension over the whole dataset.
- Calculate covariance matrix for the entire dataset.

So, using this method below, we can find the covariance of two variables X and Y.

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n} (Xi - \bar{x})(Yi - \bar{y}) \quad (1)$$

- Calculate the Eigenvectors and Eigenvalues.

From the covariance matrix we have above, we can simply compute eigenvalues and eigenvectors.

If A is a square matrix, v is a vector, and λ a scalar satisfies $Av = \lambda v$, then λ is an eigenvalue associated with eigenvector v of A.

The roots of the characteristic equation below are the eigenvalues of A.

$$\det(A-\lambda I)=0 \quad (2)$$

- To construct a $d \times k$ dimensional matrix W, sort the eigenvectors by decreasing eigenvalues and pick k eigenvectors with the biggest eigenvalues.
- Transfer the samples into new subspace.

Using the equation, $y = W' \times x$where W' is the transpose of the matrix W, we can convert our samples into a new subspace.

Finally, two principal components have been computed and projected onto the new subspace.

3.2Decision Tree

A Decision Tree is a type of machine learning model that can be used to solve problems involving classification and regression. In this tree-structured classifier, internal nodes represent dataset features, branches represent decision rules, and each leaf node represents the conclusion.

Decision nodes are used to make any decision and contain multiple branches, whereas Leaf nodes are used to represent the outcomes of the decisions.

3.3 Dense Neural Network

The neurons in a network layer connect all of the layers together densely. Each neuron in a layer receives information from all of the neurons in the layer before it, making them tightly linked. To put it another way, the dense layer is a completely linked layer, which means that all of the neurons in one layer are connected to those in the next.

A densely connected layer learns features from all of the preceding layer's combinations, but a convolutional layer focuses on consistent features with a limited repeating field.

3.4 Naive Bayes

Naive bayes algorithm works on Bayes theorem which is one of the classification technique of machine learning model.The existence of one feature in a class is assumed to be independent to the presence of any other feature by a Naive Bayes classifier. The Naive Bayes model is simple to construct and is especially effective for big data sets.

The Naive Bayes Classifier was motivated by the Bayes Theorem, which states the following equation:

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)} \tag{3}$$

Where, X is an input variable and y is an output variable.

Given the class, variables are independent. We can rewrite above equation as:

$$P(X|y) = P(x_1|y) * P(x_2|y) * \dots * P(x_n|y) \tag{4}$$

P(X) is a constant while solving for y, which means we can take it out of the equation and replace it with proportionality.

$$P(y|X) \propto P(X|y) * P(y)$$

-or-

$$P(y|X) \propto P(y) * \prod_{i=1}^{n} P(x_i|y) \tag{5}$$

Naive Bayes objective is to select the class y with the highest probability, which is calculated as:

$$y = argmax_y[P(y) * \prod_{i=1}^{n} P(x_i|y)] \tag{6}$$

3.5 Random Forest Classifier

Random forest is a method for supervised learning. It is capable of both classification and regression. Additionally, it is the most adaptable and user-friendly algorithm available. A forest is made up of trees. The more trees a forest has, the more robust it is believed to be. Random forests construct decision trees from randomly selected data samples, obtain predictions from each tree, and then vote on the best solution. Additionally, it serves as a fairly accurate indicator of the feature's importance.

Random forests have a wide variety of applications, including recommendation engines, image classification, and feature selection. It can be used to detect fraudulent activity, classify dependable loan applicants, and forecast illnesses. It is the foundation of the Boruta algorithm, which selects significant features from a dataset.

4 Experimental Investigation

In this research the input values arethe data related to transactions on stocks of different NIFTY 50 companies from 1st January to a day before publication of financial results of 2020-21 December quarter of those companies. On doing so, we have noticed that big transactions have been done on some companies, before publication of financial results, which can be suspected as Insider trading.For identifying the insider trading activities, a large number of related variables of quoted companies could be employed. But the data available wasso largethat it may cause an issue of overfitting while training our model, so we scaled down the data using PCA. Thus, different columns with negligible importance were removed and two columns Principal Component 1 and Principal Component 2 were used for training different machine learning algorithms.First Decision Tree was used which is a tree-structured classifier for prediction of insider trading whose accuracy came out to be 0.85. Then Random Forest(RF) was used which is a tree-based method used for classification or regression, where an individual tree from a group of decision trees vote to give an output decision.Its accuracy came out to be 0.86. Again, Naive Bayes was used whose accuracy was observed as 0.99. Further, a deep learning method called Dense Neural Network was used whose accuracy of prediction came out to be 0.92. This shows that Naive Bayes does the classification work well. TheOutput Value is Insider Trading ‘YES’ or ‘NO’.

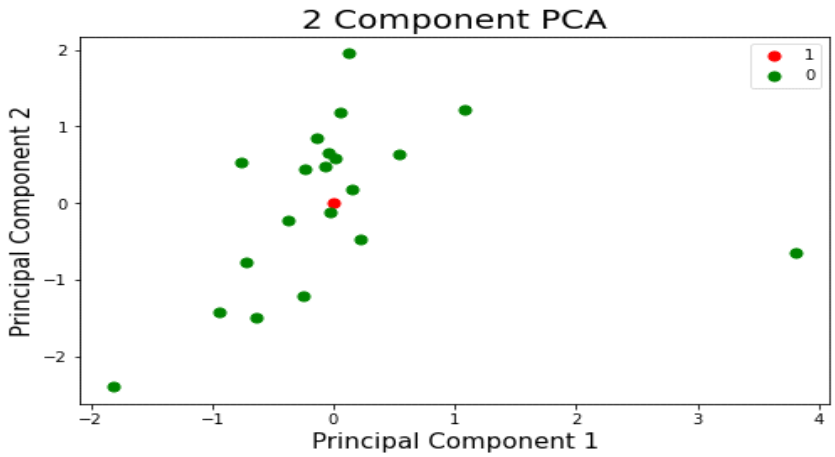

Figure 1. Visualization of Principal Component Analysis [1= " YES" AND 0= "NO"].

In the case of stratification, the dataset is used for dividing the given information into Yes or NO. This dataset was acquired from NSE website under corporate filing actions for different NIFTY 50 companies from 1st January to a day before publication of financial results of 2020 December quarter of those companies.

Here input values: -

'Symbol', 'Series', 'Date', 'Prev Close', 'Open Price', 'High Price', 'Low Price', 'Last Price', 'Close Price', 'Average Price', 'Total Traded Quantity', 'Turnover', 'No. of Trades', 'Deliverable Qty', '% DlyQt to Traded Qty', 'REGULATION', 'NAME OF THE ACQUIRER/DISPOSER','NO. OFSECURITIES (ACQUIRED/DISPLOSED)', 'NOTIONAL VALUE(BUY)', 'NOTIONAL VALUE(SELL)'

Target:-

Insider Trading - YES / NO

```
[34] df = pd.DataFrame({'Real Values':y_test, 'Predicted Values':y_pred})
     df
```

	Real Values	Predicted Values
0	NO	NO
1	NO	NO
2	NO	NO
3	NO	NO
4	NO	NO
...	...	...
295	NO	NO
296	NO	NO
297	NO	NO
298	NO	NO
299	NO	NO

300 rows × 2 columns

Figure2. Screenshot of real values and predicted values using Naive Bayes Model.

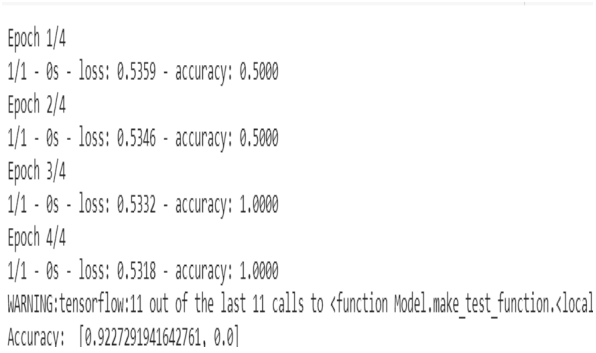

Figure 3. Visualization of Deep Neural Network.

5 Experimental Results

We have used metrics like accuracy, precision, recall and f1-score.Here in this project the metrics we have used are given below.

Accuracy: Accuracy is a fraction of the predictions our model got right. Mathematical representation of accuracy is denoted as:

Accuracy = (Number of correct predictions)/(Total number of predictions)

Precision: Precision is the ratio of true positive value to all positive values.

Mathematically precision is denoted as:

Precision = (True positive values)/ (total positive values)

Recall: Recall is the measure of correctly identifying true positive value.

Mathematically recall is denoted as:

Recall = (True positive values)/ (True positive values + False negative values)

F1-score: F1-score is also a type of metric which is used when there is equal importance of both precision and recall to our model, it is calculated by harmonic mean of precision and recall.

Mathematically F1-score is denoted as:

F1-score = (precision * recall)/ (precision + recall)

Table 1. Comparison of Experimental Results.

S.N.	Methodology	Precision	Recall value	F1-score	Accuracy
1.	Decision Tree Classifier	0.86	1.00	0.92	0.8571
2.	Random Forest Classifier	0.86	1.00	0.95	0.8671
3.	Naive Bayes Classifier	1.00	0.99	1.00	0.9933
4.	Dense Neural Network	0.92	0.91	0.91	0.9227

6 Discussion of Results

After the implementation of the different machine learning and deep learning models, we have reached thefollowing results. Beginning with the accuracy of the Decision Tree Classifier is 0.8571.Its precision is 0.86, recall value is 1.00 and f1-score is 0.92. Secondly, the accuracy of the Random Forest classifier model is0.8671. Its precision is 0.86, recall value is 1.00and f1-score is 0.95. Similarly, the accuracy of Naive Bayes Classifier is 0.9933. Its precision is 1.00, recall value is 0.99 and f1-score is 1.00.Finally, the accuracy of the Dense Neural Network model is 0.9227.Its precision is 0.92, recall value is 0.91 and f1-score is 0.91.

7 Conclusion

From this model accuracies obtained from various machine learning algorithms, we can tell that the Naive Bayes classifier gives the best results, which is 0.99 followed by Dense Neural Network, then Random Forest and Decision Tree model to the least accurate respectively. Hence, we conclude that our work on Insider trading prediction has been completed, which is used to reduce the time and cost of SEBI or investigating team to find out susceptible Insider trading. The task to predict Insider trading is a difficult task because of limited data available. Insider trading doesn't happen only in equity section, it may happen in derivative section too. As, no data was available for derivative section this work is limited to detect insider trading only in equity section. The limitation of this project is that because of limited data available it's difficult to say 100% insider trading. Another drawback is that we were not able to compare our result with any similar work because to the best of our knowledge our work is the first in detecting illegal insider trading using corporate action data, transaction data and machine-learning and deep learning techniques.

References

[1] Shangkun Deng , Chenguang Wang , Jie Li , Haoran Yu , Hongyu Tian , Yu Zhang , Yong Cui , Fangjie Ma and Tianxiang Yang Identification of Insider Trading Using Extreme Gradient Boosting and Multi-Objective Optimization. Information, 10(12), p.367

[2] Shangkun Deng, Chenguang Wang, Zhe Fu, Mingyue Wang An Intelligent System for Insider Trading Identification in Chinese Security Market.Comput Econ 57, 593–616 (2021)

[3] Tamersoy, A., Xie, B., Lenkey, S. L., Routledge, B. R., Chau, D. H., &Navathe, S. B. (2013). Inside insider trading: Patterns & discoveries from a large scale exploratory analysis. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013 (pp. 797-804).

[4] Sheikh Rabiul Islam, Sheikh Khaled Ghafoor, William Eberle Mining Illegal Insider Trading of Stocks: A Proactive Approach

[5] K. Golmohammadi, O. R. Zaiane and D. Díaz, "Detecting stock market manipulation using supervised learning algorithms," 2014 International Conference on Data Science and Advanced Analytics (DSAA), 2014, pp. 435-441, doi: 10.1109/DSAA.2014.7058109.

[6] M. L. Huang, J. Liang and Q. V. Nguyen, "A Visualization Approach for Frauds Detection in Financial Market," 2009 13th International Conference Information Visualisation, 2009, pp. 197-202, doi: 10.1109/IV.2009.23.

[7] K. V. Nesbitt and S. Barrass, "Finding trading patterns in stock market data," in IEEE Computer Graphics and Applications, vol. 24, no. 5, pp. 45-55, Sept.-Oct. 2004, doi: 10.1109/MCG.2004.28.

[8] K. Golmohammadi and O. R. Zaiane, "Data Mining Applications for Fraud Detection in Securities Market," 2012 European Intelligence and Security Informatics Conference, 2012, pp. 107-114, doi: 10.1109/EISIC.2012.51.
[9] Xuan Zhang, "Empirical analysis of the insider trading's characteristics in China stock market," 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011, pp. 6628-6631, doi: 10.1109/AIMSEC.2011.6010631.
[10] A. Tamersoy, B. Xie, S. L. Lenkey, B. R. Routledge, D. H. Chau and S. B. Navathe, "Inside insider trading: Patterns & discoveries from a large scale exploratory analysis," 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), 2013, pp. 797-804, doi: 10.1145/2492517.2500288.
[11] A. B. Kasgari, M. T. Taghavifard and S. G. Kharazi, "Price manipulation fraud detection by Intelligent Visual Fraud surveillance system," 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), 2019, pp. 1646-1651, doi: 10.1109/CoDIT.2019.8820499.
[12] Alpaydin E. Introduction to machine learning [M]. MIT press, 2010.
[13] https://www.nseindia.com/companies-listing/corporate-filings-actions
[14] https://www1.nseindia.com/products/content/equities/equities/eq_security.htm
[15] https://www.livemint.com/market/stock-market-news/how-india-cracks-down-on-insider-trading-11580199120367.html