Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Predictive and causal chain mining to discover actionable knowledge from stock markets

Harchana Bhoopathi, B. Rama

Department of Computer Science, KakatiyaUniversity, Warangal(U), Telangana, India

ARTICLE INFO

Article history:
Received 30 October 2020
Accepted 6 November 2020
Available online xxxx

Keywords:
Data mining
Cause-effect relationship
Causal chain mining
Causal Chain Miner (CCM)

ABSTRACT

In many scientific studies, discovering causal relationship is given high importance. Due to the availability of data in various domains, the research on mining causal relationships is growing steadily. Finding causal relationships from stock market datasets has its utility in that domain. Many researchers contributed towards developing algorithms for causal mining. In our prior work, we proposed algorithms for upstream and downstream causal relationships in stock market data. However, causal chain mining is relatively new research phenomenon where a set of inter-related actions are found and among the actions there exists causal relationship. Causal chains, in fact, reflect frequent occurrences of set of events with cause-effect relationship. In this paper we proposed a framework for mining causal relationships from stock market data. Two algorithms namely Stock Risk Prediction (SRP) and Causal Chain Miner (CCM) are proposed and implemented. An empirical study is made with a prototype application to validate the proposed framework. The experimental results showed the significance and usage of proposed causal chain mining and predictive algorithms which has comparable performance over the state of the art methods.

1. Introduction

Cause and effect is the theory used to understand the causal relationships. Every effect has its cause. It is the essence of this theory [7]. When causal relationships are found, it can help in understand how two variables are acting as dependent and independent variables and the relationships between them. Causal relationship discovery is useful in many applications like text mining, linguistic analysis, stock market analysis and evidence synthesis to mention few. It is used in healthcare in [1] for discovering evidence based cause and effects among variables of health information. The concept of Directed Acyclic Graphs (DAG) is employed for causal relationships discovery in [2]. Evidence synthesis is another important application in healthcare industry where causal chain analysis plays role as investigated in [4]. Causal structures are mined in [5] with scalable techniques.

Causality with Twitter data is studied in [6]. However, it does not consider causal chains. There are many researches as explored in [4,11-13,17,18]. From the literature, it is understood that causal relationships are found in many studies based on the events and associated verbs. This kind of theory is not possible in stock market data. Therefore, the approach to stock market analysis should be different. Towards this end, in this paper we proposed a framework and algorithms to achieve both stock risk analysis and also discovering causalchains. Our contributions in this paper are as follows.

- We proposed a framework for analysing stocks in general along with risk profiles besides discovering causal chains.
- We proposed an algorithm named Stock Risk Prediction (SRP) that provides valuable information in terms of risk profiles of given stocks.
- We proposed another algorithm named Causal Chain Miner (CCM) which is used to discover causal chains (causal structures) from given stock dataset.
- We built a prototype application using data science platform in python known as Anaconda.

The remainder of the paper is structured as follows. Section 2 provides review of literature on the causal chain mining. Section 3 presents the proposed framework and underlying methodology besides algorithms. Section 4 provides experimental results. Section 5 provides conclusions and gives future scope of the research.

2. Related work

Causal relationship exists between a dependent variable to be forecasted and an independent variable [9]. Causal system models

https://doi.org/10.1016/j.matpr.2020.11.120

Please cite this article as: H. Bhoopathi and B. Rama, Predictive and causal chain mining to discover actionable knowledge from stock markets, Materials Today: Proceedings, https://doi.org/10.1016/j.matpr.2020.11.120

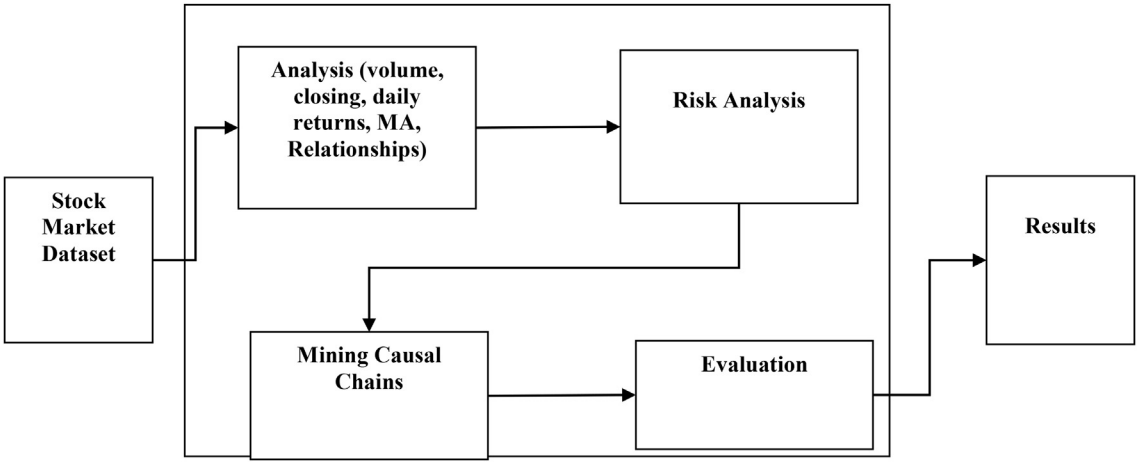

Fig. 1. Proposed framework.

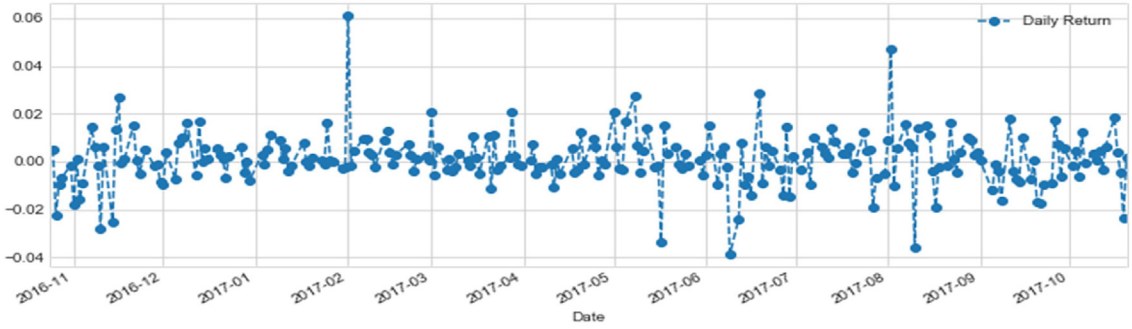

Fig. 2. Analysis of daily returns (Apple Stock).

can be extracted from information systems. These models can be used in different applications like discovering faults in a network system [10]. Underlying structures in datasets may have interesting structures and sub structures. However, they may have sources of bias as well. For instance, "V" shape structures may have endogenous bias while causal chains may have overcorrection bias [1]. Causal analysis may have different kinds of ideas such as causal forks ($A \leftarrow B \rightarrow C$), causal inverted forks ($A \rightarrow B \leftarrow C$) and causal chains ($A \rightarrow B \rightarrow C$). These relationships reflect particular correlation among the entities [2]. There are number of cases in which causal chains may exist. For instance, observations are made that an event denoted as e1 causes another event e2 which in turn may cause another event e3 [3]. Causal chain can be discovered in some areas in order to have synthesized evidence. In this aspect causal chain analysis makes sense. When causal chains establish evidence, it is possible that decision makers or policy-makers take expert decisions. Causal chain analysis provides other advantages such as understanding complex relations, improving quality of evidence and better utilization of evidence [4].

When Markov condition is assumed, causal chains can be understood. With prior knowledge let us say that we know A already and it has no causes. If there is dependency between B and A, B is caused by A but it may be done indirectly. B or some other variables causing A and B are to be ruled out as A has no causes. If there is a third variable like C which is having dependency on both A and B, there is formation of a causal chain [5]. Potential causation may exist in different scenarios [6,15]. Cause and effect relationships may form as a chain of related events. Causal connections may fall into the category of relationship known as causal chain which is also known as sequence of related events [7].

There are certain examples that are against co-location associated with events. These examples are interpreted as chain of causal connections. In this context, each causal link which is part of chain is expected to meet requirement of spatiotemporal co-location. Chain of unobserved events may also occur. There are atomic events, their associations and aggregate events [8]. In case of healthcare it is possible to have evidence based causalchains. They provide promising approaches to have comprehensive understanding of systems. Causal chains can be extracted from the data related to climate changes. It reveals the relationship between casual chains of climate changes and human security. There are different kinds of causalities such as simple causalities, resultatives causalities and instrumental causalities. Both qualitative and quantitative approaches can be used to analyse causal chains [13]. Time series data also can have causal relationships and causal chains [14]. Therefore, causal chains may be associated time domain as well. Causal connections may lead to having a chain of related events [16]. A dynamic causal topic modelling (DCTP) is explored in [18] based on a probabilistic model. Causal influence of each topic is ascertained effectively.

Causal chains are a kind of sequential patterns as studied in [2,17]. They show the order of occurrences apart from the relation-

ships. Causal chains may indicate events provided in a planning document, events in stock market data or any such thing based on domain data. There might be maximal causal chains.

3. Proposed methodology

A methodology is proposed to have stock market analysis, risk analyses and discovery of causal chains. The methodology includes collection of stock data, pre-processing it and then performs a series of operations as shown in Fig. 1. It includes analysis of volume of stocks, closing price, daily returns, moving averages and relationships. The framework also considers risk analysis that generates risk profiles for given stocks. It finally mines the causal chains and the evaluate results.

As can be seen in Fig. 1, the proposed framework discovers causal chains that are nothing but chains of related events. In order to realise this framework, two algorithms are proposed and implemented using Python data science platform known as Anaconda. Stock Risk Prediction (SRP) and Causal Chain Miner (CCM) are thetwo algorithms defined. The algorithms are as follows.

3.1. Stock risk prediction algorithm

Stock Risk Prediction (SRP) is the algorithm which takes stock dataset as input and returns risk profiles. It makes use of a Markov process with two terms such as drift and shock in order to achieve this.

Algorithm1: Stock Risk Prediction (SRP)

Input: Stock dataset D
Output: Risk profiles R
Start
Initialize stocks vector S
Initialize drift term vector D1
Initialize shock term vector S1
For each instance d in D
Update S
End For
S = Remove Duplicates(S)
Initiate Markov Process
For each stock s in S
For each instance in s
Compute drift
Compute shock
Update D1 with drift
Update S1 with shock
End For
Compute Value at Risk (VaR)
Add VaRalong with s to R
End For
Return R
Stop

As presented in Algorithm 1, risk analysis is made based on Markov process that exploits two terms like drift and shock for each stock in order to have the VaR to be computed. The results of this algorithm reveal the value at risk for every stock found in the dataset D.

3.2. Causal chain miner algorithm

Causal Chain Miner (CCM) is the algorithm used to discover causal chains from stock market data. It takes stock dataset as input and generates causal chain.

Algorithm2: Causal Chain Miner (CCM)

Input: Stock dataset D
Output: Causal chains vector C
Initialize causal patterns vector P
Initialize frequent patterns vector F
F = Get Frequnt Patterns(D)
P = Get Causal Patterns(F)
For each pattern p in P
If p is sequential pattern Then
If p has more than two events Then
Add p to C
End If
End If
End For
Return C

As shown in Algorithm 2, it has mechanisms to discover causal chains. First, it obtains causal patterns from the given dataset. It includes getting frequent patterns prior to obtaining causal patterns. Then it will start an iterative process to check whether the pattern is a sequential pattern and has more than two events involved. If the conditions are satisfied, it will consider that pattern as a causal chain and adds it to the output vector.

4. Experimental results

The proposed framework is realised with a prototype application built using data science platform known as Anaconda Python. The observations of the algorithms include general stock analysis in terms of price, closing price, volume, moving averages and daily returns. There are other important results observed. They include risk profiles of stocks and causal chains in non-technological stocks.

As presented in Fig. 2, the horizontal axis shows different dates while the vertical axis shows the daily returns of Apple stock. The results revealed that there are constant fluctuations in the performance of Apple stock from time to time.

As presented in Fig. 3, the horizontal axis shows daily returns while the vertical axis shows the statistical values of Apple stock. The results revealed that there are constant fluctuations in the performance of Apple stock from time to time.

As presented in Fig. 4, it is observed those different years along with months is given in horizontal axis while vertical axis shows closing price for analysis. The results revealed that that there is steady increase in the closing price from 2016 to 2017. There is gradual increase in the performance of Apple stock.

As presented in Fig. 5, it is observed those different years along with months is given in horizontal axis while vertical axis shows volume for analysis. The results revealed that that there is steady difference in the volume from 2016 to 2017.

As presented in Fig. 6, along with closing price, moving averages for 10 days, 20 days, 50 days and 100 days is analysed for Apple stock. The horizontal axis shows dates while the vertical axis shows moving averages and closing prices. There is increase in the moving averages from 2016 to 2017.

As presented in Fig. 7, there is correlation analysis among the four technology stocks considered. The positive and negative correlations can help us to understand the trends in stock markets.

As presented in Fig. 8 Monte Carlo analysis is presented for technology stocks like Apple, Amazon, Google and Microsoft. The results revealed that the temporal domain shows increase in closing price. When number of days is increased, the closing price is also increased.

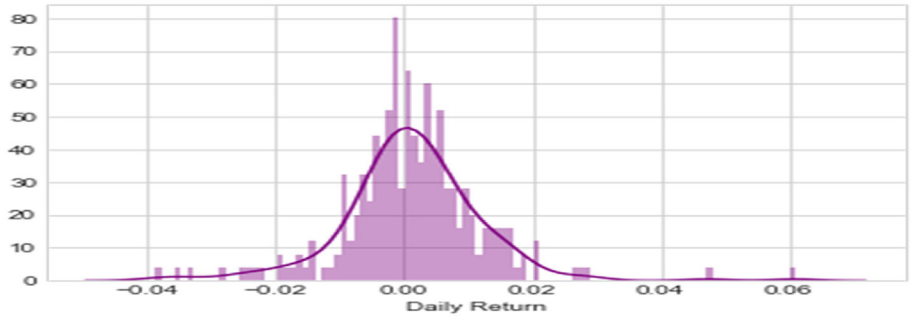

Fig. 3. Histogram reflecting daily returns of Apple stock.

Fig. 4. Historical view of closing price of Apple stock.

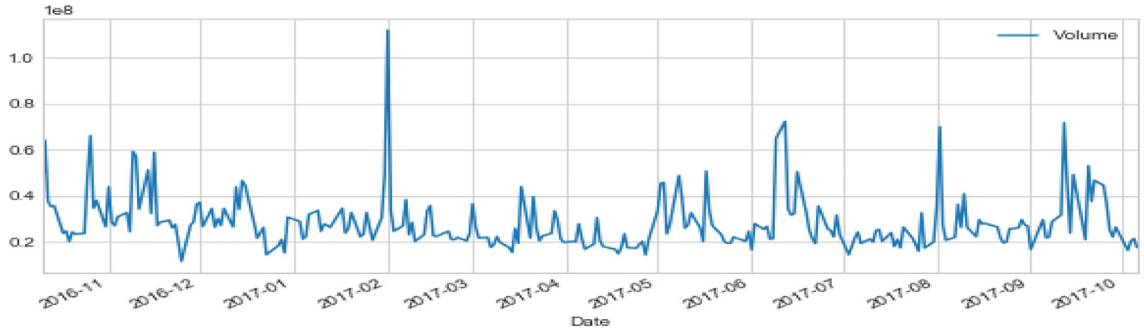

Fig. 5. Volume analysis of Apple stock.

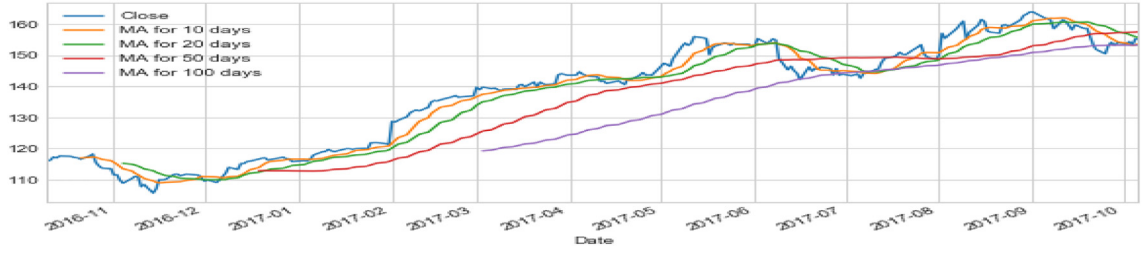

Fig. 6. Additional moving averages of Apple stock.

As presented in Fig. 9, there is correlation among all stocks in terms of daily returns. The correlations are studied among technology stocks like Google, Apple, Amazon and Microsoft.

As presented in Fig. 10, there is correlation among all stocks in terms of daily returns. The correlations are studied among technology stocks like Google, Apple, Amazon and Microsoft.

As can be seen in Fig. 11, the technology stocks are observed with risk profiles. The measure used to find risk is Value at Risk (VaR). The more is of VaR value, the more is the risk associated with the stock. The VaR of Google is $17.98, Amazon $18.13, Apple $2.48 and Microsoft $1.28. The results reveal the final price distribution, start price, mean final price and VaR.

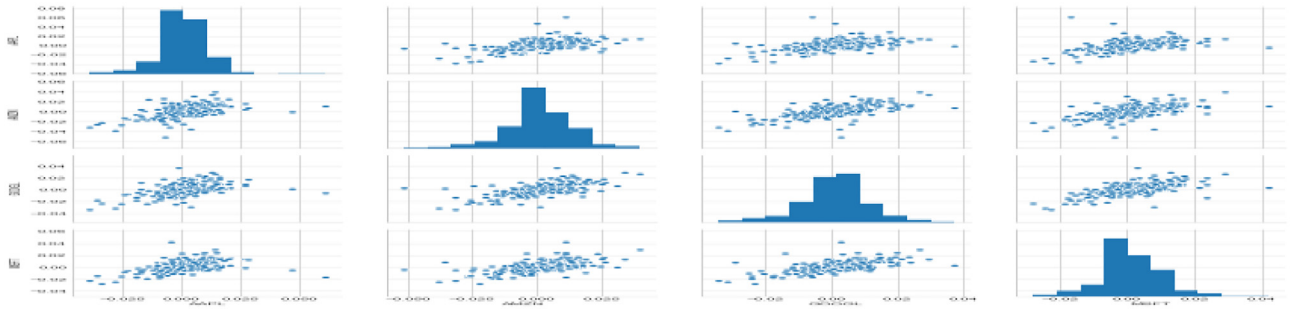

Fig. 7. Visual analysis of performance of all stocks.

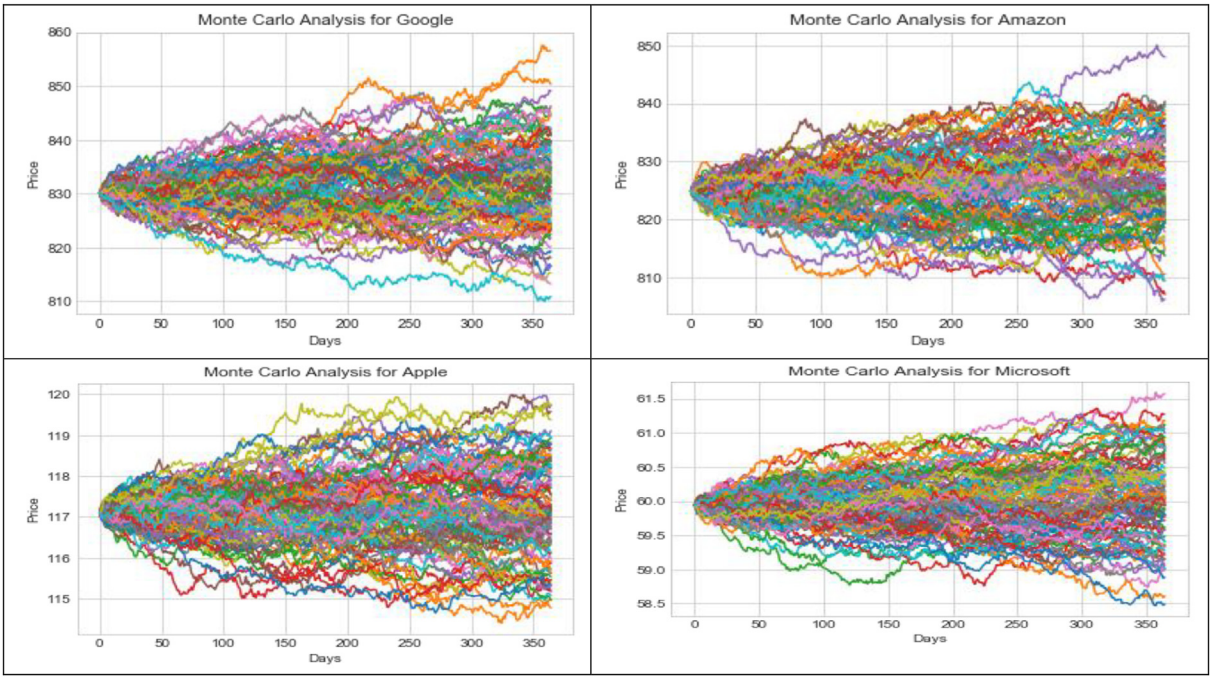

Fig. 8. Monte Carlo analysis for all stocks.

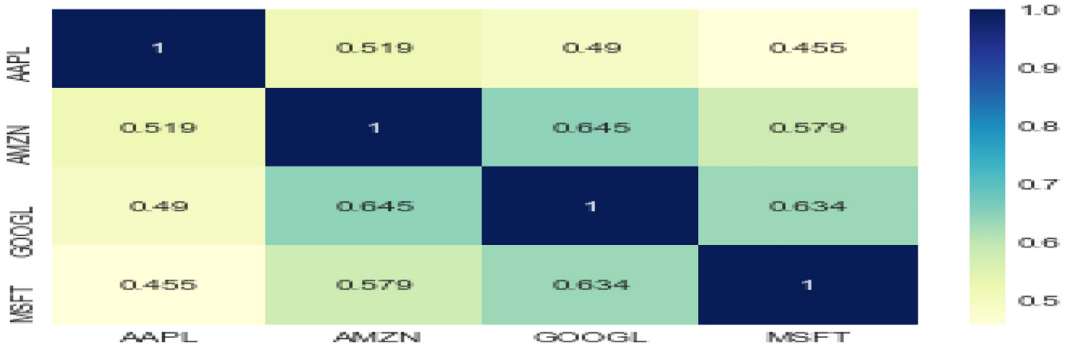

Fig. 9. Correlation among all stocks in terms of daily returns.

As presented in Fig. 12, there are many chains discovered from the given stock market data. The algorithm CCM is involved to mine causal chains. There are many stocks in the non-technological stock data. Three chains are shown in horizontal axis while the vertical axis shows quantitative results in terms of True Positives (TP), False Positives (FP) and precision. The causal chains discovered are I02D -->Hs02D -->M02D (shown as Causal Chain 1), Hs02D -->FTSE02D -->M02D (shown as Causal Chain 2) and Nikkei01U -->Nikkei02U -->M02U -->FTSE02D -->M02U (shown as Causal Chain 3). Different causal chains are discovered with high precision.

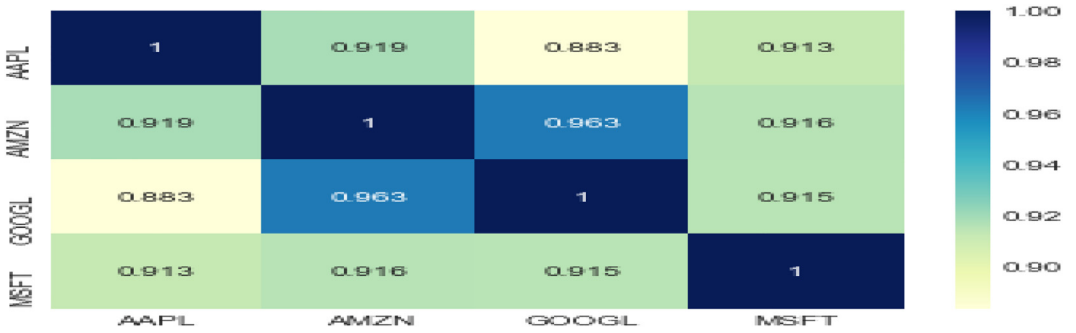

Fig. 10. Correlation among all stocks in terms of closing prices.

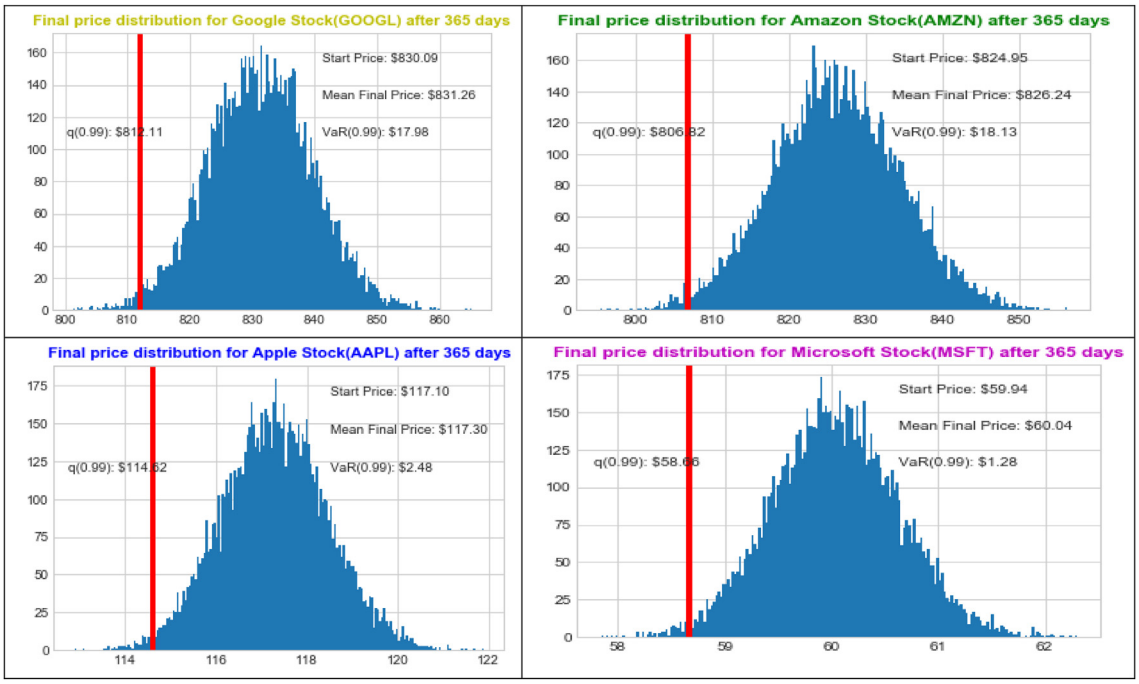

Fig. 11. Risk profiles of all the stocks.

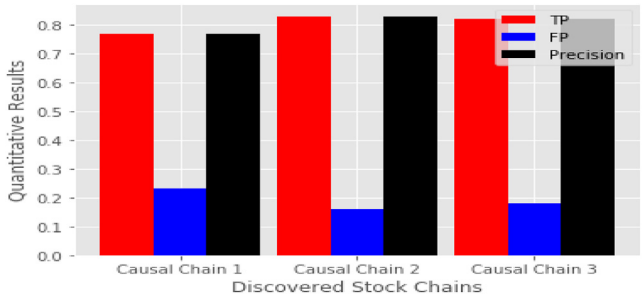

Fig. 12. Performance in causal chain discovery.

5. Conclusion

In this paper, we proposed a framework for stock performance analysis, risk analysis and discovery of causal chains using data mining approach. Technology stocks and other stocks are considered for empirical study. The analysis is related to stock prices, volume, closing price and risk prediction. Two algorithms are proposed to achieve this. They are known as Stock Risk Prediction (SRP) and Causal Chain Miner (CCM). The outcomes of these algorithms showed the importance of mining stock data. SRP produced risk profile of the given stocks while the CCM yielded causal chains that can provide many valuable insights about the stocks with causal effect and relationship among them. A Markov process named as Geometric Brownian Motion (GBM) is used to analyse historical data and establish Efficient Market Hypothesis (EMH). The notion of drift and shock are used to estimate risk profiles of stocks. Monte Carlo method is employed in order to simulate stocks and prices. The CCM on the other hand involves discovering causal sequential relationships that form as causal chains. The CCM's performance is evaluated and found that it is useful to ascertain the latent insights in stocks that lead to expert decision making. In future, we continue on causal chain mining with Artificial Intelligence (AI) that is machine learning approaches to improve the state of the art.

CRediT authorship contribution statement

Harchana Bhoopathi: Conceptualization, Data curation, Investigation, Methodology, Software, Visualization, Writing - original draft. **B. Rama:** Software, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Prandial Yadav, "Causal Pattern Mining in Highly Heterogeneous and Temporal EHRs Data",p1-111,2017.
[2] Krishna Murthy Inumala, Exploring causal relations in data mining by using directed acyclic graphs(Dag), Int. J. Adv. Comput. Eng. Netw. 3 (5) (2015) 1–17.
[3] Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, Abir Naskar, " Automatic extraction of causal relations from text using linguistically informed Deeo neural networks" , Proceedings of the SIGDIAL,p-306-316,2018.
[4] Kneale d, "Causal chain analysis and evidence synthesis" ,p1-7,2017.
[5] Craig Silverstein, Sergey Brin, Rajeev Motwani, " Scalable Techniques for mining Causal Structures ",p1-12,2016.
[6] D.E. O'Leary, Twitter mining for discovery, prediction and causality: Applications and methodologies, Intell. Syst. Account. Finance Manage. 22 (3) (2015) 222–247.
[7] S. Khan, S. Parkinson, " Causal Connections Mining Within Security Event Logs" , Proceedings of the Knowledge Capture Conference on –K- CAP, P1,2017.
[8] Bleisch, S. Duckham, M. Galton, A. Laube, Lyon, Mining candidate causal relationships in movement patterns, Int. J. Geograph. Inform. Sci. 28 (2) (2013) 363–382.
[9] S. Kamble, A. desai, P. Vatak "data Mining and Data warehousing for Supply chain Management", International Conference on Communivation , Information 7 Computing Technology(ICCICT), 2015.
[10] N. Ye, a reverse engineering algorithm for mining a causal system model from system data, Int. J. Prod. Res. 55 (3) (2016) 828–844.
[11] J. Qiu, E.T. Game, H. Tails, L.P. Olander, L. Glew, J.S. Kagan, S.K. Weaver, Evidence-based causal chains for linking health, development, and conservation actions, Bioscience 68 (2018) 182–193.
[12] M. Baumgartner "Detecting Causal Chains in Small-n Data Field Methods", 25 (1),3-24,2012.
[13] S. Alashri, J.-Y. Tsai, A.R. Koppela, H. Davulu, "Snowball: Extracting causal Chains from Climate Change Text Corpora" ,1st International Conference on Data Intelligence and Security(ICDIS).P1-8,2018.
[14] Yaseer Mohammad, Toyoaki Nishida, Mining Causal Relationships in Multidimensional Time series, Springer, 2010, pp. 309–338.
[15] M. Er Kara, S. UmitOktayFirat, A. Ghadge, A Data Mining-based framework for supply chain risk management, Comput. Ind. Eng. (2018) 1–32.
[16] Khan, Saad Parkinson, Simon, " Causal Connections Mining within Security Event Logs", In: Proceedings of the 9th International Conference on Knowledge Capture . ACM. P1-7,2017.
[17] Shreeram Sahasrabudhe, hector Munoz-Avila, " Discovering Causal Chains by Integrating Plan Recognition and Sequential Pattern Mining" , American Association for Artificial Intelligence,p1-6,2002.
[18] Y. Fan, Q. Zhou, W. Yue, W. Zhu, A dynamic causal topic model for mining activities from complex videos, Multimedia Tools Appl. 77 (9) (2017) 10669–10684.

Further Reading

[1] L. Wang, "Mining Causal Relationships among clinical variables for cancer diagnosis based on Bayesian analysis", BioDataMining ,8(1).P1-15,2015.