APPLICATIONS OF DATA MINING IN FINANCE

Naveen Kunnathuvalappil Hariharan
University of the Cumberlands, United States

ABSTRACT

Data mining as a discipline of computer science has been widely employed in several domains as a result of the need to find methods to evaluate fast expanding data. Finance is one of the most appealing data mining application areas in these new technologies. As a means of managing large data, enterprise efficiency, and business intelligence, data mining and machine learning are critical. In the financial business, data mining is extremely valuable. To date, data mining has shown to be a viable solution for detecting financial data dynamics and linkages. It's been used in a variety of financial situations. In this work, we concentrate on the use of data mining in stock forecasting, portfolio management, and investment risk analysis, as well as the prediction of bankruptcy and foreign exchange rates and the identification of financial fraud.

Keywords: Bankruptcy, Data mining, Exchange rates, financial fraud. Portfolio management

Introduction

Many organizations, such as banks, stock market authorities, taxation authorities, large accounting and auditing firms, specialist databases, and so on, collect financial data, which is sometimes made public. The use of data mining techniques on financial data can help solve categorization and prediction challenges while also making the decision-making process easier. Corporate bankruptcy, credit risk estimation, going concern reporting, financial difficulty, and corporate performance forecast are all examples of financial classification issues (Kudyba and Hoptroff 2001). Data mining as a discipline of computer science has been widely employed in several domains as a result of the need to find methods to evaluate fast expanding data. Finance is one of the most appealing data mining application areas in these new technologies.

With the growing amount of data saved in files, databases, and other repositories, it is becoming increasingly vital to build sophisticated tools for analyzing and extracting meaningful knowledge from such vast amounts of data. Data Mining (DM) was proposed in the late 1980s, accompanied by the development of computer and database technology, in order to discover valid, complex, and not obvious information from large amounts of data (Pyle 2003), using concepts and methods from the fields of Artificial Intelligence, Pattern Recognition, Database Systems, and Statistics.

Financial institutions that provide financial services to their clients or members collect financial data. Financial institutions are divided into three categories: Banks, building societies, credit unions, trust companies, and mortgage loan companies are deposit-taking institutions that accept and manage deposits and make loans; insurance companies and pension funds; and brokers, underwriters, and investment funds are deposit-taking institutions that accept and manage deposits and make loans (Bose and Mahapatra 2001). In comparison to other businesses, data in the finance industry has its own characteristics, such as big volumes, completeness, reliability, and high frequency, among others. As a result, financial institutions generate massive databases as the basis for using data mining technology to tackle immensely complicated and dynamic finance problems.

Data mining and machine learning are crucial for the management of vast data, corporate efficiency, and business insight. In the financial business, data mining is extremely valuable (Dean 2014). Financial firms discover hidden patterns in massive sets of data in order to keep track of the information in their databases. Personal data can be used to explain a client's financial situation and behavior before and after he or she accepts credit. Customers can get a number of services from most financial institutions, including data monitoring and the creation of a business savings account. Customers are given credit in transactions such as mortgage business, vehicle loan services, investment services, insurance services, and stock investment services, according to the timetable. Prediction of future financial events, such as stock markets, foreign exchange rates, bankruptcy, credit rating of the bank's customer information, predictive financial and investment analysis, trading futures, and understanding and managing financial risk in banks, are some of the other financial applications of data mining and machine learning.

DATA MINING

The phrase "data mining" refers to innovative technologies for analyzing large amounts of data intelligently. Information systems, machine learning, artificial intelligence, data engineering, and knowledge discovery are just a few of the disciplines where these technologies have arisen (Soares and Ghani 2010). Finance, which is becoming increasingly conducive to data-driven modeling as enormous quantities of financial data become available, is one of the most enticing application areas for these developing technologies. Data mining, also known as knowledge discovery or data discovery, is a process that entails studying and analyzing data from various sources, evaluating, and combining it into more useful and important information - information that can be used to increase revenue and profits, reduce costs, or do both. Data mining has grown increasingly significant in the lives of corporations and governments in recent years.

Data mining is also utilized for searching for consistent patterns and/or systematic correlations between variables in vast volumes of data (generally business or market-related data), and then validating the conclusions by applying the discovered patterns to new subsets of data. Prediction is the ultimate goal of data mining, and predictive data mining is the most popular sort of data mining with the greatest direct business implications. The data mining procedure is divided into three stages (Maheshwari 2014). 1) The preliminary investigation. 2) Model construction or pattern recognition with validation and verification. 3) Implementation (that is, the application of the model to new data in order to generate predictions).

EXISTING APPLICATIONS OF DATA MINING IN FINANCE

a) Prediction of the Stock Market

Market investors strive to maximize their profits by buying and selling their investments at the right time. Predicting the future trend (i.e., rise, fall, or remain constant) of a stock is a difficult task since stock market data is highly time-variant and typically follows a nonlinear pattern. Prior research has shown that the growth rates of a number of fundamental characteristics, such as revenues, earnings per share, capital investment, debt, and market share, can be used to predict future returns of individual companies (Papajorgji and Petraq 2013).

Stock market fluctuations have historically been modeled using regression models. Those models, on the other hand, can only anticipate linear trends. Neural network modeling, which includes back-propagation (BP) networks, probabilistic neural networks, and recurrent neural networks, has been the most widely utilized data mining technique in stock market prediction thus far (Xu 2012).

The majority of neural network models that attempt to predict particular stocks only take data from the relevant marketplaces. Some studies seek to evaluate the effect of more established markets whose values affect the performance of smaller emerging markets by using not only the current stock index value, but also trading volumes from all indexes in neural network models. With the addition of key external market indicators, index forecasting performance is expected to progressively improve. The input values in each BP neural network (each for a specific market) in a study to predict a five-day future value for several market indexes were the set of one-, two-, and five-day lags of the closing value, as well as the respective one-, two-, and five-day normalized average trading volumes for the respective index markets (Desai 2012). The single output value was the value of the relevant market's five-day future index. The results showed that when more external knowledge was supplied to the neural network, prediction performance improved.

B. Portfolio Management

Portfolio management is a significant topic in the investment world. It is about how people choose which securities to hold in their investment portfolios and how money should be divided across different asset classes, such as equities versus bonds and domestic versus overseas securities (Stewart et al. 2019). The basic goal is to select a collection of risk assets from which to build a portfolio in order to maximize return while minimizing risk or to attain a certain return while minimizing risk. In order to help financial decision-making, investment models must offer the best projection of projected returns and risks. The standard deviation of return is a measure of return variability that is employed in the conventional Markowitz "Mean-Variance" model for effective portfolio (Peng et al. 2008).

The business factor models like the Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) have proved effective in explaining how assets are priced in financial markets by connecting their performance to risk (Ali Khan and Sun 1997). Although these models do not forecast future asset returns, they do give valuable risk management information. The APT has been the most common portfolio management strategy in recent years. The APT model's main assumption is that investment risks may be divided into systematic and idiosyncratic risks. To get the best asset pricing, the APT method advises employing quadratic programming (Nguyen and Others 2010). It aids in the identification of a few risk characteristics among a group of candidates, as well as the selection of securities based on their relative risks and returns. In portfolio management, APT has been used with neural networks. An APT model can be used to set pricing in this hybrid approach, and then a neural network can forecast the future trend of each risk component. Following the generation of investment choices, the best portfolio is chosen depending on the investor's preferences.

Hidden input-output Conditional hidden Markov models (IOHMMs) allow the emission (and possibly the transition) probabilities to be conditioned on an input sequence. They can be linear, logistic, or nonlinear conditional distributions. As a result, IOHMMs have also been utilized to predict financial time series. According to early study, when forecasting the conditional density of returns of market and sector indices, IOHMMs can produce much higher performance than historical average estimates, as measured by the out-of-sample likelihood (Bengio et al. 2001); (Christensen et al. 2020).

C. Bankruptcy Prediction

The prediction of corporate bankruptcy is a well-known problem in the financial literature. Indeed, bankruptcy is one of four generic terms for corporate distress that can be defined as the condition in which a business is unable to meet its debt obligations and petitions a court for either debt reorganization or asset liquidation (Wu et al. 2010). The following factors contribute to business failure and bankruptcy: economic, financial, neglect, fraud, disaster, and others. Economic factors such as industry weakness and poor location are examples of economic factors, while financial factors include excessive debt and insufficient capital. Investors, for example, want to minimize credit risk and avoid non-profitable investments (Wilson and Sharda 1994). As a result, several authors have previously researched this topic. The impact of bankruptcy predictions on financial markets is significant.

Because of the necessity of making correct and timely strategic business choices, bankruptcy prediction has been a focus of research in business analytics. Even while the accuracy of the prediction model is critical, the model's understandability and portability are equally vital (Bellovary et al. 2007). For shareholders, creditors, policymakers, and business managers, accurately predicting bankruptcy has been a critical issue.

(Mansouri et al. 2016) compared neural network modeling to two statistical methods for predicting bankruptcy, logistic regression and audit analysis, and designed a neural network model for predicting the bankruptcy of manufacturing companies in addition to introducing neural network models. The data used is from the Kerman province between 1985 and 1997. The ratios of liquid assets to total assets, profit before interest expenditure and total taxes investments, equity to debt, capital investments, profit before interest expense, and sales tax were used in this study. Their findings revealed that the designed neural network model outperforms two other statistical methods in terms of predictability.

(Ohlson 1980) employed a Logit model with no previous assumptions about bankruptcy probability or predictor variable distribution. Hold-out sample tests, on the other hand, are prone to being upwardly biased (Grice and Dugan, 2001), because variations in macroeconomic factors are time-sensitive.

(Boyacioglu et al. 2009)) Predicted bankruptcy using a neural network, a support vector machine, and statistical models. They used equity to total assets, equity to debt load, total debt to total assets, net income to average assets, and current assets to total assets in this study. The findings of their study revealed that training and establishing valid data, as well as achieving a unique technique to solve problems and anticipating insolvency by neural networks, are difficult on their own. As a result, it is suggested that it be combined with other patterns.

In a study, (Jeong et al. 2012) coupled the artificial neural network with collective patterns and the genetic algorithm. They used a combined algorithm to predict a company's bankruptcy and collective models to determine the input variables. In this model, they used the sales cost-to-goods-sold ratio, current debt-to-total-assets ratio, interest expense-to-sales ratio, and current debt-to-total-assets ratio to predict company bankruptcy. When compared to decision trees, collective models, multivariate analysis, and various linear models, the combined algorithm outperforms them all.

D. Foreign Exchange Market

Global financial competitiveness encourages countries to liberalize their markets and attract foreign investment. Foreign exchange rates have gotten more volatile as company activities have become more multinational. Forecasts of future exchange rates are used to make many business choices. The exchange rate fluctuates constantly, and in order to buy products and services produced in another country, one must first buy the currency of that country. Traders benefit by purchasing a currency at a cheap rate and then selling it at a higher rate as demand rises. Researchers have used different methodologies to estimate foreign exchange rates, ranging from statistical regression methods to novel so-called data mining techniques such as neural networks and support vector machines, to help with these decision-making (or risk management) processes. Neural Network (NN) approaches, particularly Multilayer Perceptron (MLP) models and other statistical techniques, are used in almost all contemporary financial modeling research (Akkaya and Uzar 2011).

There has been a substantial rise in study into the usage of NN models in recent years. A number of academics have built financial neural network models for predicting inflation, stock prices, exchange rates, bond ratings, and bankruptcy (Dhanani 2003).

E. Fraud Detection

Securities fraud is a broad term that relates to fraudulent actions in the offering and sale of securities. The following are the several types of securities fraud:

a) High-Yield Investment Fraud: These schemes usually promise guaranteed returns on low- or no-risk securities investments. In order to operate their funds, the perpetrators take advantage of the investors' trust and promise enormous profits. Pyramid scams, Ponzi schemes, Prime Bank Schemes, Advance Fee Fraud, Commodities Fraud (foreign currency exchange and precious metals fraud), and promissory notes are the most common high yield investments (Neisius and Clayton 2014).

b) Broker Embezzlement: This technique involves taking unauthorized and illegal steps to profit from a client's investment. This could include illegal trading or fabricating paperwork (Rezaee 2002).

c) Market Manipulation: This scheme entails an individual or a group of individuals attempting to profit by interfering with a fair and orderly market (Dorminey et al. 2012).

Pattern Recognition

The purpose of pattern recognition is to find patterns that are comparable to trends that have previously been associated with fraudulent activity. This goal can be achieved on two levels: a) recognizing suspicious traders who engage in fraudulent behavior, and b) detecting securities linked to fraudulent activity (which is desired since regulators can halt trading on such securities to maintain a fair market for all participants). In this instance, a real-time (online) data mining procedure is required. Data: a set of historical trading data for each trader account (in the case of “a”) or for each securities (in the case of “b”), as well as a list of fraud patterns/trends (labels) (Ngai et al. 2011).

Outlier Detection

Outlier Detection seeks to identify observations that are outliers in comparison to the rest of the data. This can aid in the discovery of previously unknown fraud practices. In addition, rather than utilizing a threshold to filter out spikes, this scenario allows for effective detection of spikes based on market conditions. This, like the previous group, can be done at both the security and trader levels (West and Bhattacharya 2016). Data refers to a trader's or a security's past trading information. Clustering algorithms are commonly used to discover anomalies, and labeled data is not required.

Rule Induction

The purpose of Rule Induction is to extract rules that can be reviewed and used by securities market auditors and regulators. Data includes historical trading information for each trader account, as well as trader accounts that have been flagged as potentially fraudulent (Blanton and Others 2012). Unlabeled data can also be used to extract rules that identify undiscovered patterns and abnormalities (unsupervised learning methods).

Social Network Analysis

The objective of social network analysis is to find out which trader accounts are working together to manipulate the market. Data: each trader account's historical trading information. Additional data sources concerning traders' work histories and relationships must be integrated into the dataset (for example, the NASD utilizes a Central Registration Depository (CRD) that keeps track of federally registered brokers' information) (Lipner 2013).

Visualization

Visualization's purpose is to provide representations that go beyond traditional charts, allowing auditors to interact with market data and identify dangerous tendencies. Market data visualization is critical for both real-time monitoring and off-line research. Auditors can use visualization to spot questionable activity in securities and trading transactions (Dilla and Raschke 2015). Data used for visualization includes historical trading data or a real-time stream of securities/trader transaction data.

CONCLUSION

The usage of data mining has accelerated in the Big Data era. With their power and automation, data mining algorithms can deal with large amounts of data and extract value. Data mining techniques are applicabale in various financial applications such as loan risk analysis and payment prediction, mortgage scoring, and real estate services, in addition to the applications we've described. Despite the fact that data mining has been used in finance for many years, there are still a number of outstanding concerns and obstacles that must be properly addressed in order to accomplish effective financial management for both individuals and institutions. Massive datasets, accuracy, privacy, the complexity of performance measurements, market impacts, and regulator training are among the issues. Data mining techniques are evolving, and they have shown significant promise in financial applications, and they will continue to thrive in the emerging knowledge-based economy.

REFERENCES

1) Akkaya, G. Cenk, and Ceren Uzar. 2011. “Data Mining: Concept, Techniques and Applications.” GSTF Business Review (GBR) 1 (2): 47.
2) Ali Khan, M., and Yeneng Sun. 1997. “The Capital-Asset-Pricing Model and Arbitrage Pricing Theory: A Unification.” Proceedings of the National Academy of Sciences of the United States of America 94 (8): 4229–32.
3) Bellovary, Jodi L., Don E. Giacomino, and Michael D. Akers. 2007. “A Review of Bankruptcy Prediction Studies: 1930 to Present.” Journal of Financial Education 33: 1–42.
4) Bengio, Y., V. P. Lauzon, and R. Ducharme. 2001. “Experiments on the Application of IOHMMs to Model Financial Returns Series.” IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council 12 (1): 113–23.
5) Blanton, Kimberly, and Others. 2012. “The Rise of Financial Fraud.” Center for Retirement Research Brief, no. 12-5. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1059.922&rep=rep1&type=pdf.
6) Bose, Indranil, and Radha K. Mahapatra. 2001. “Business Data Mining — a Machine Learning Perspective.” Information Management 39 (3): 211–25.
7) Boyacioglu, Melek Acar, Yakup Kara, and Ömer Kaan Baykan. 2009. “Predicting Bank Financial Failures Using Neural Networks, Support Vector Machines and Multivariate Statistical Methods: A Comparative Analysis in the Sample of Savings Deposit Insurance Fund (SDIF) Transferred Banks in Turkey.” Expert Systems with Applications 36 (2, Part 2): 3355–66.
8) Christensen, Hugh, Simon Godsill, and Richard E. Turner. 2020. “Hidden Markov Models Applied To Intraday Momentum Trading With Side Information.” arXiv [q-fin.TR]. arXiv. http://arxiv.org/abs/2006.08307.
9) Dean, Jared. 2014. Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. John Wiley & Sons.

10) Desai, Samir. 2012. Stock Market Pattern Recognition Algorithm. Lap Lambert Academic Publishing GmbH KG.
11) Dhanani, Alpa. 2003. “Foreign Exchange Risk Management: A Case in the Mining Industry.” The British Accounting Review 35 (1): 35–63.
12) Dilla, William N., and Robyn L. Raschke. 2015. “Data Visualization for Fraud Detection: Practice Implications and a Call for Future Research.” International Journal of Accounting Information Systems 16 (March): 1–22.
13) Dorminey, Jack W., Arron Scott Fleming, Mary-Jo Kranacher, and Richard A. Riley Jr. 2012. “Financial Fraud.” New York 82 (6): 61–65.
14) Jeong, Chulwoo, Jae H. Min, and Myung Suk Kim. 2012. “A Tuning Method for the Architecture of Neural Network Models Incorporating GAM and GA as Applied to Bankruptcy Prediction.” Expert Systems with Applications 39 (3): 3650–58.
15) Kudyba, Stephan, and Richard Hoptroff. 2001. Data Mining and Business Intelligence: A Guide to Productivity. Idea Group Inc (IGI).
16) Lipner, S. E. 2013. “The Expungement of Customer Complaint CRD Information Following the Settlement of a FINRA Arbitration.” Fordham J. Corp. & Fin. L. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/fjcf19&section=6.
17) Maheshwari, Anil. 2014. Business Intelligence and Data Mining. Business Expert Press.
18) Mansouri, Ali, Arezoo Nazari, and Morteza Ramazani. 2016. “A Comparison of Artificial Neural Network Model and Logistics Regression in Prediction of Companies’ Bankruptcy (A Case Study of Tehran Stock Exchange).” International Journal of Advanced Computer Research 6 (24): 81–92.
19) Neisius, Jens, and Richard Clayton. 2014. “Orchestrated Crime: The High Yield Investment Fraud Ecosystem.” In 2014 APWG Symposium on Electronic Crime Research (eCrime), 48–58. ieeexplore.ieee.org.
20) Ngai, E. W. T., Yong Hu, Y. H. Wong, Yijun Chen, and Xin Sun. 2011. “The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature.” Decision Support Systems 50 (3): 559–69.
21) Nguyen, Tho Dinh, and Others. 2010. “Arbitrage Pricing Theory: Evidence from an Emerging Stock Market.” Development and Policies Research Center Working Paper Series, no. 2010/03. https://www.academia.edu/download/54451819/090909_APT_Thailand.pdf.
22) Ohlson, James A. 1980. “Financial Ratios and the Probabilistic Prediction of Bankruptcy.” Journal of Accounting Research 18 (1): 109–31.
23) Papajorgji, and Petraq. 2013. Enterprise Business Modeling, Optimization Techniques, and Flexible Information Systems. IGI Global.
24) Peng, Hui, Min Gan, and Xiaohong Chen. 2008. “A Mean-Variance Model for Optimal Portfolio Selection with Transaction Costs.” IFAC Proceedings Volumes 41 (2): 1627–32.
25) Pyle, Dorian. 2003. Business Modeling and Data Mining. Elsevier.
26) Rezaee, Zabihollah. 2002. Financial Statement Fraud: Prevention and Detection. John Wiley & Sons.
27) Soares, C., and Rayid Ghani. 2010. Data Mining for Business Applications. IOS Press.
28) Stewart, Scott D., Christopher D. Piros, and Jeffrey C. Heisler. 2019. Portfolio Management: Theory and Practice. John Wiley & Sons.
29) West, Jarrod, and Maumita Bhattacharya. 2016. “Intelligent Financial Fraud Detection: A Comprehensive Review.” Computers & Security 57 (March): 47–66.
30) Wilson, Rick L., and Ramesh Sharda. 1994. “Bankruptcy Prediction Using Neural Networks.” Decision Support Systems 11 (5): 545–57.
31) Wu, Y., C. Gaunt, and S. Gray. 2010. “A Comparison of Alternative Bankruptcy Prediction Models.” Journal of Contemporary Accounting & Economics 6 (1): 34–45.
32) Xu, Feifei. 2012. Data Mining in Social Media for Stock Market Prediction. Dalhousie University.