Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/22/20231208

Review on the influence of machine learning methods and data science on the economics

Zhekai Liu

School of Science, Rensselaer Polytechnic Institute, Troy, New York, United State, 12180

liuz28@rpi.edu

Abstract. In this era of extraordinary accessibility to information, business faces both unprecedented obstacles and opportunities. Constantly accumulating data encompasses everything from consumer behavior to market trends. However, the question of how to extract useful information from this enormous quantity of data and apply it to economic decision-making becomes crucial. Complex non-linear relationships and high-dimensional data frequently render conventional statistical methods and economic models ineffective. Integration of data science and machine learning techniques has enabled economists to extract valuable insights from large-scale and complex economic data. By examining the role of data science and machine learning in economics and tracing its historical development from the refinement of statistics to the era of big data with advanced computational power, this paper will discuss the significance of data-driven decision making and forecasting in the economy with specific algorithm in supervised and unsupervised learning and focus on future challenges and developments.

Keywords: economics, data science, supervised learning, unsupervised learning.

1. Introduction

Contemporary economics frequently focuses on the prices and news surrounding the entire market, particularly the stock and futures markets. As a result of the tendency of globalization for more countries and businesses to have extensive connections to global markets, they generate a substantial quantity of data daily. To extract valuable information from these data, economic analysts must always perform numerous tasks, including cleaning, comparing, and organizing. In recent years, the explosion of computer-based data science has significantly altered analysis techniques. Individuals were able to construct models and process various types of data using computers. Programmable computer processing is rapid, efficient, and accurate [1]. Since the end of the 20th century, the discipline of computing has experienced rapid growth. Improvements in algorithms and semiconductor technology provide the computing industry with enormous benefits. In the field of data science, computers are able to manage increasingly complex data with a variety of specialized processing methods, whereas humans cannot. In addition, the advent of machine learning could be of great assistance with data classification, prediction, and numerous other tasks that meet the needs of economic analysts. Today, machine learning plays a crucial role in data science, particularly in the economics field: financial institutions and hedge funds use machine learning algorithms to make high-frequency trading decisions, banks and lending institutions use machine learning models to assess creditworthiness and manage risk, and E-commerce

DOI: 10.54254/2755-2721/22/20231208

companies and retailers use machine learning to segment customers based on their preferences, behavior, and purchasing patterns [2].

By examining the role of data science and machine learning in economics and tracing its historical development from the refinement of statistics to the era of big data with advanced computational power, this paper will discuss the significance of data-driven decision making and forecasting in the economy with specific algorithm in supervised and unsupervised learning and focus on future challenges and developments. This paper provides a comprehensive analysis of the evolution of machine learning in data science and its application to economics, as well as a discussion of the prospects and limitations of machine learning in the context of contemporary big data.

2. Roles of data science and machine learning

Although data science's popularity seems to have emerged out of nowhere, the field in fact has a rich and lengthy history as an interdisciplinary field. The evolution of statistics marks the commencement of data science. In the 19th century, statisticians such as Gauss and Pearson pioneered the fundamental theories and methods of statistics, including probability theory, regression analysis, and testing of hypotheses. These elements laid the foundation for data science. During the first half of the 20th century, computer scientists conducted research and developed computer programming, algorithms, and data processing techniques as a result of the invention and development of computers. This led to the rise in popularity of data science. Later, at the end of the 20th century and the beginning of the 21st century, significant amounts of data began to be generated and stored due to the popularity of the Internet and the increase in computing power [3]. The global expansion of the Internet provided data science with vast data resources, which became the propelling force behind the field's rapid growth. The era of big data has officially arrived with the advent of mobile devices, sensor technology, and cloud computation. Big data consists of vast, high-velocity, and diverse data, and data scientists have begun to investigate methods and tools for processing and analyzing big data, including distributed computing, Hadoop, and Spark. Data science has progressively utilized machine learning and artificial intelligence in recent years. The application of data science is expanding to a variety of disciplines, such as finance, healthcare, energy, transportation, and other areas closely related to the lives of individuals. Data science has become a fundamental competency in many industries and organizations due to its increasing significance in solving real-world problems and supporting decision making [4].

In discussing the history of data science, it is clear that data science is developing and evolving, with machine learning as one of its most important components. Since the explosion of data and the rise in computing power, data scientists require efficient methods to extract valuable information and insights from massive and complex datasets. In this environment, machine learning techniques have evolved. This technology, which is founded on the training and self-improvement of large data sets, enables computers to learn from data using a variety of algorithms and models, and then apply this learning to tasks such as prediction, classification, and clustering. This approach to learning based on data makes machine learning ideally suited for analyzing economic data and solving economic problems.

The study of economics always entails vast quantities of data, such as macroeconomic indicators, market data, consumer behavior data, etc. These data frequently contain features with complex correlations that are difficult to capture using conventional statistical methods and models [5]. Compared to conventional economic theories and models, machine learning is more adept at handling large-scale and high-dimensional data and discovering latent relationships between features. Using specific algorithms, models can automatically discover these relationships to aid economists in identifying underlying economic patterns and trends. In addition to rapid data processing and analysis, machine learning can assist with tasks such as prediction and policy decision making. After the model has been fitted to the historical data, machine learning algorithms can perform accurate forecasting and trend analysis, advising economists to make more trustworthy economic forecasts and decisions. Moreover, economists could use it to calculate the theoretically optimal allocation of resources and risk management solution. Risks and opportunities can be discovered through analysis and modeling,

DOI: 10.54254/2755-2721/22/20231208

allowing for the creation of plans and actions that increase the efficiency with which resources are used and decrease the likelihood of adverse events [6].

3. Application of machine learning in economic data

3.1. Supervised learning on prediction and classification

An essential subset of machine learning is supervised learning. It seeks to discover relationships by utilizing labeled training data to make accurate predictions or classifications. Its fundamental concepts are features with well-defined labels, where features are attributes or variables used to describe and represent data and labels are the variables to be predicted or classified. Constructing a mathematical model, employing learning algorithms to learn from the training data, and optimizing the model's predictions by minimizing the loss function are the steps of supervised learning. Regression, classification, and sequential prediction are prevalent supervised learning algorithms. People frequently use supervised learning to perform prediction and classification tasks in economics [7].

When governments formulate economic policies or investors devise investment strategies, the economic environment and industry's future is always the top priority. When companies or markets want to analyze the performance of their customers, analysts are always required to categorize individuals into distinct patterns. There are recurring patterns and cycles in human and economic activity. Trends such as economic growth, recessions, and recoveries tend to exhibit certain regularities over extended periods of time, and so would people's behavior if they possessed the same characteristics. Through the analysis of historical data, cyclicality and similarity can be identified, and predictions can be made based on them. There are numerous applications for supervised learning in economic and market forecasting. Due to the stock market's non-linear, dynamic, stochastic, and unreliable nature, forecasting familiar phenomena is not a straightforward task [8]. Nonetheless, there are numerous supervised learning tools available to facilitate this task. Commonly employed techniques include support vector machines (SVM), k nearest neighbors (kNN), artificial neural networks (ANN), decision trees, fuzzy time series (FTS), and evolutionary algorithms (EAs). SVM seeks to identify a hyperplane that optimally separates data classes by maximizing the margin between them. SVM is renowned for its capacity to manage high-dimensional data, limit errors, and generalize effectively to new data. kNN is a straightforward and efficient classification algorithm. It classifies a data point by contemplating the class labels of its closest training set neighbors. The "k" value chosen determines the number of neighbors considered. The data point's class label is determined by a majority vote among its k nearest companions. ANN is a computational model that is based on the structure and operation of the human brain. It consists of interconnected, layered, artificial neurons. ANN can learn intricate relationships between inputs and outputs by adjusting the weights and biases of its connections. It is utilized for a variety of tasks, including classification, regression, and pattern recognition. Decision Trees are a prominent algorithm for classification and regression tasks in machine learning. They develop a tree-like representation of decisions and their potential outcomes. The tree is constructed by recursively splitting the data based on various characteristics, optimizing at each stage for the best separation or predictive power. Decision Trees can manage both numerical and categorical data and are interpretable. Fuzzy time series analysis integrates fuzzy logic and time-series analysis. It is utilized for forecasting and predicting time-dependent data in the presence of uncertainty and ambiguity. FTS employs fuzzy logic to develop linguistic rules based on fuzzy sets, which are subsequently used to formulate forecasts or make decisions. De-fuzzification is used to transform ambiguous outputs into clear outputs. Evolutionary Algorithms are a class of optimization algorithms inspired by natural selection and biological evolution. They are utilized to solve intricate optimization issues by iteratively seeking the optimal solutions. EAs typically employ mechanisms such as mutation, crossover, and selection to increase the fitness of a population of candidate solutions over successive generations [9].

DOI: 10.54254/2755-2721/22/20231208

3.2. Unsupervised learning and information mining

Unsupervised learning is a possible component of machine learning. Supervised learning is a machine learning technique that uses labeled training data to train a model and then uses the features of new data to infer new data labels. Unsupervised learning, on the other hand, is a method that discovers patterns, structure, and correlations in the data without the need for labeled training data. Unsupervised learning can be utilized in economic data science for tasks such as data clustering, anomaly detection, downscaling, and association rule mining. Supervised learning is appropriate for tasks requiring prediction, whereas unsupervised learning is appropriate for tasks involving data exploration and the discovery of latent relationships [10].

Numerous successful applications of unsupervised learning exist in market basket and consumer segmentation. For instance, if a grocery store applies unsupervised learning to its products for sale, it may discover connections between ostensibly unrelated items. Therefore, redesigning the store and positioning related products next to one another could help them comprehend the consumer mindset and increase revenue. Basic algorithms such as clustering algorithms, dimensionality reduction algorithms, and association rule mining algorithms are extremely useful for displaying detailed information when analyzing common datasets. The objective of clustering algorithms such as k-means, Hierarchical and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is to group data points with similar intrinsic properties. Dimensionality reduction techniques, such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE), reduce the number of variables or features in a dataset while conserving its fundamental structure. They facilitate the visualization of high-dimensional data and the elimination of irrelevant or redundant features. Association rule mining algorithms, such as the apriori algorithm and the FP-growth algorithm, can unearth interesting relationships or associations between distinct items in a dataset. They are utilized to identify frequent patterns or principles [11]. People have developed numerous other advanced systems and algorithms that are common in the real world. Anomaly detection techniques are used to identify instances in a dataset that deviate from the expected patterns and are therefore anomalous or outliers. They protect individuals from anomalies, fraud, and unusual economic events. And generative models such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) could learn the data's underlying distribution and generate new samples that resemble the original data distribution. These techniques can now be used in conjunction with artificial intelligence to produce convincing simulations or create synthetic data [12].

4. Difficulties and future developments

Despite the fact that data science in economics has progressed over many years, there are still many obstacles to surmount in light of the current economic data environment and demand. The efficiency and precision of machine learning algorithms are contingent on the input data's quality. In economic data science, data are frequently absent, incorrect, or inconsistent, which can result in skewed or inaccurate model training results. In contrast, economic data frequently includes confidential information such as personal financial data, transaction records, etc. People are frequently reluctant to reveal their complete economic and privacy information. Future developments will necessitate improved data collection, cleansing, and validation techniques to enhance data quality, as well as improved data encryption, secure sharing, and privacy protection technologies to ensure the security and confidentiality of economic data. Another issue is that economics majors may be unfamiliar with computer algorithms that require knowledge of codes and mathematics. They frequently view machine learning models as black-box models, which makes it difficult to articulate their decision-making process. Transparency, interpretability, and interdisciplinary decisions are crucial for policymakers and stakeholders in the economic field. Future advancements will necessitate more researchers in both economics and data science, and these two disciplines will also converge in order for machine learning models to provide more explicable outcomes and decision processes [13].

DOI: 10.54254/2755-2721/22/20231208

5. Conclusion

The rise of data science and machine learning has had a profound effect on the field of economics, allowing economists to extricate valuable insights from massive and complex economic data. In economic data analysis, machine learning techniques have become indispensable tools, supplying solutions to challenges faced by traditional statistical methods and models. Economists may generate precise forecasts and classifications through supervised learning, which aids in economic forecasting, market analysis, and policy decisions. Unsupervised learning enables the exploration of data, the discovery of latent patterns, and the enhancement of customer segmentation and market basket analysis. As data science and machine learning continue to advance, their influence on the economy will increase. The application of these technologies will result in more efficient resource allocation, enhanced decision-making procedures, and the identification of new economic opportunities and threats. By addressing the obstacles and fostering inter-disciplinary collaborations, we can uncover the full potential of these technologies, thereby contributing to a more data-driven, efficient, and well-informed economic ecosystem.

References

[1] Provost, F. and Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making, Big Data. https://doi.org/10.1089/big.2013.1508.

[2] Paruchuri, H. (2021). Conceptualization of Machine Learning in Economic Forecasting, Asian Business Review. VOL. 11 NO. 2. https://doi.org/10.18034/abr.v11i2.532.

[3] Romein, C. Annemieke And Kemman, Max And Birkholz, Julie M. and Baker, James And De Gruijter, Michel and Meroño-Peñuela, Albert And Ries, Thorsten Aand Ros, Ruben And Scagliola, Stefania (2020). State of the Field: Digital History, History. VOL. 105 Number. 365 Pages. 291-312. https://doi.org/10.1111/1468-229X.12969.

[4] Brady, Henry E. (2019). The Challenge of Big Data and Data Science, Annual Review of Political Science, Vol. 22:297-323. https://doi.org/10.1146/annurev-polisci-090216-023229.

[5] Liran Einav and Jonathan Levin (2014). Economics in the age of big data, Science. VOL. 346, NO. 6210. DOI: 10.1126/science.1243089.

[6] Cao, L., Yang, Q. & Yu, P.S. (2021). Data science and AI in FinTech: an overview, Int J Data Sci Anal 12, 81–99. https://doi.org/10.1007/s41060-021-00278-w.

[7] Athey, S. (2018). The impact of machine learning on economics, The economics of artificial intelligence: An agenda, University of Chicago Press, 507-547.

[8] Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., & Kim, H. C. (2021). Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions. Electronics, 10(21), 2717. https://doi.org/10.3390/electronics10212717.

[9] Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. Expert Systems with Applications, 124, 226-251.https://doi.org/10.1016/j.eswa.2019.01.012.

[10] Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2019). Supervised and unsupervised learning for data science, Springer Nature.

[11] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and unsupervised learning for data science, 3-21. https://doi.org/10.1007/978-3-030-22475-2_1.

[12] Bagad, P., Mitra, S., Dhamnani, S., Sinha, A. R., Gautam, R., & Khanna, H. (2021). Data-Sharing Economy: Value-Addition from Data meets Privacy, In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (pp. 1105-1108). https://doi.org/10.1145/3437963.3441712.

[13] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects, Science, 349(6245), 255-260. https://doi.org/10.1126/science.aaa8415.