Progress in Application of Big Data Analytics Technology in E-Commerce Precision Marketing

Qixiang Sun
School of Business, Shanghai University of Technology, Xuhui District, Shanghai, 200030, China
1326876179@qq.com

ABSTRACT

As e-commerce has expanded globally and the volume of data generated by online users has grown substantially, big data analytics has become an important part of how companies approach precision marketing. This paper reviews the application of big data analytics in e-commerce precision marketing, focusing on four areas: customer segmentation, personalized recommendation systems, sentiment analysis and opinion mining, and dynamic pricing strategies. Drawing on studies published between 2015 and 2023, the review summarizes methodological developments, discusses the main technical and ethical challenges in the field, and points to directions for future research. The results show that machine learning methods (such as clustering algorithm, recommendation model based on deep learning, natural language processing (NLP) technology and price reinforcement learning) greatly improve the accuracy and return on investment (ROI) of finding suitable customers in online marketing. In other words, there are still problems in data protection, model transparency and cold start in recommendation system. This review aims to provide useful help to researchers and professionals, who want to know the progress and development direction in this field.

Keywords: Big Data Analytics; E-Commerce; Precision Marketing; Personalized Recommendation; Customer Segmentation; Sentiment Analysis; Dynamic Pricing.

1. INTRODUCTION

In the past ten years, e-commerce has developed rapidly. It produces a lot of different data, and it changes quickly. Transaction records, click logs, user comments, searches and social media activities together create a large amount of data that companies can use for marketing. Gandomi and Haider [1] use the famous volume, speed and diversity to describe these data-we call it "big data" phenomenon. Traditional marketing methods only use population segmentation and experience-based projects, and cannot keep up with this change. As Akter and Wamba [2] said in their systematic review, Big Data Analysis (BDA) is changing the way e-commerce companies create value. It enables them to talk to each customer, understand what they are going to do, and spend their money on marketing more effectively.

Precision marketing-when you target special customer groups with promotional information that matches what they like and do-becomes very important for large platforms such as Amazon, Alibaba and JD.com. Erevelles and others [3] say that consumer analysis makes this change possible because we can understand the differences between customers on a large scale and take action.

Although there are more and more researches on how to use BDA in e-commerce, there is still no comprehensive review covering all precision marketing functions. Previous comments usually focused on one thing-such as recommendation system or sentiment analysis-rather than putting them together. This limits the theoretical development and practical suggestions of those who want to use BDA in many marketing functions at the same time.

This review has three main objectives: (1) to explain the key BDA technologies used in e-commerce precise marketing and summarize the evidence supporting them; (2) Looking for common technical and ethical issues; And (3) suggest what to study in the future. In the next part, we will discuss the methods used in the review, the four main areas of using BDA, problems and ideas for the future, and then we will draw a conclusion.

2. LITERATURE REVIEW METHODOLOGY

This review follows an interpretive systematic approach adapted from established protocols in e-commerce and information systems research [2]. Literature was drawn from Web of Science, Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar. Search terms included combinations of “big data analytics,” “e-commerce,” “precision marketing,” “personalized recommendation,” “customer segmentation,” “sentiment analysis,” “dynamic pricing,” and “machine learning.” The review covers publications from January 2015 to December 2023, a period that captures both the maturation of deep learning methods and their widespread adoption in commercial applications. Only peer-reviewed journal articles, conference papers from major venues (IEEE, ACM, Springer, Elsevier), and high-quality technical reports were considered.

Studies were included if they (a) explicitly addressed at least one precision marketing function in an e-commerce context, (b) applied or evaluated an identifiable BDA technique, and (c) reported quantitative performance metrics or substantive qualitative findings. Studies that were purely conceptual without empirical grounding, or that focused exclusively on brick-and-mortar retail, were excluded. A two-stage screening process—title and abstract screening followed by full-text review—yielded a final corpus of 68 publications, which were organized into the four thematic domains discussed in the next section.

3. BIG DATA ANALYTICS APPLICATIONS IN E-COMMERCE PRECISION MARKETING

3.1 Customer Segmentation

Customer segmentation is a starting point for precision marketing: grouping customers into relatively homogeneous subsets lets companies allocate resources more efficiently and tailor messages to fit specific audiences. Traditional demographic segmentation has largely been replaced by behavior-based approaches enabled by BDA.

The Recency-Frequency-Monetary (RFM) model remains widely used in e-commerce segmentation, largely due to its simplicity and ease of interpretation. Ma [4] applied RFM to a large e-commerce dataset and found that K-Means clustering over RFM scores produces stable, commercially useful segments that outperform rule-based methods when identifying high-value customers. Rahim et al. [5] extended this by using RFM-derived features as inputs to a machine learning classification pipeline for repurchase prediction, achieving strong accuracy on a three-year transactional dataset. Their finding that recency is the dominant predictor of repurchase probability has direct implications for re-engagement campaign design.

Recent work has gone beyond RFM, using sequential behavioral data and deep learning architecture. By layering and grouping customers’ habits, we can see how the customer's value changes, and the simple model is not done well. Chen and Zhang [6] show that it is useful to mix multiple information sources (such as shopping, internet browsing and social media data) to understand the needs of customers, and this idea helps to create several mixed segmentation models. These multimodal methods make segmentation more accurate, but they also require a lot of data work and a lot of calculation, especially when you want to use them in real time. The good results of better segmentation can be seen in the activity. When we can identify very important customers ("champions") and separate them from those who are more sensitive to price or in danger of leaving, we can only reward those who need it, not everyone. This helps to make better use of promotional funds. Nowadays, with the help of cloud computing, it is easier to group a large amount of data in a complex way, although there are still many problems to understand how to explain and use these customer groups.

3.2 Personalized Recommendation Systems

Personalized recommendation systems (RS) are arguably the most studied and commercially significant application of BDA in e-commerce. By surfacing products or content aligned with individual users’ inferred preferences, RS directly influence conversion rates, basket size, and session depth. The literature distinguishes three main paradigms: collaborative filtering (CF), content-based filtering, and hybrid approaches that draw on both.

Collaborative filtering infers preferences from the behavior of similar users (user-based CF) or from items exhibiting similar co-purchase or co-rating patterns (item-based CF). Classical matrix-factorization CF has well-known limitations—particularly around data sparsity and the cold-start problem, which is the difficulty of generating useful recommendations for new users or items with limited interaction history. Deep learning integration helps to solve these problems. For example, Li [7] created a cross-border electronic commerce recommendation system. The system uses topic model to extract important words from text, and uses sequence-to-sequence learning to represent objects. This provides better suggestions for multilingual data sets that are not too large.

The chart method goes further. By representing users, objects and their relationships as heterogeneous graphs, the graphical neural network (GNN) model can find more complex patterns and social influences that plane CF cannot see. The knowledge graph model adds structural information about object attributes, and the results show that they are more accurate and easier to explain [8]. Explaining suggestions becomes very important for companies and rules. A system that can explain its suggestions can better win the trust of users and abide by the new transparency rules.

A practical problem of e-commerce recommendation system (RS) is to deal with real-time reasoning on a large scale. Platforms like Taobao handle hundreds of millions of recommendation requests every day, and the response time is only a few milliseconds. Distributed computing frameworks like Apache Spark MLlib help. For example, using model broadcasting in memory can reduce the reasoning cost of each request by about 40%, while the standard model loading of each task can be reduced by about 40% [9]. Embedding the streaming data about user interaction into online model updating is still an active research problem, because the model trained in batches may lose its prediction quality when the user's taste changes.

3.3 Sentiment Analysis and Opinion Mining

Comments left by users are a valuable source of unorganized marketing information in e-commerce. Emotional analysis (looking for opinions, attitudes and emotions in texts) enables companies to understand how people view products in near real time, find quality problems early and make personal suggestions according to emotional signals. In view of the number and variety of opinion texts, it has become a practical need to use automatic natural language processing (NLP).

At first, sentiment analysis used a dictionary-based method to provide a fixed polarity score for words. These methods work well in simple cases, but they are not clearly explained with words, denials and feelings in specific fields. Machine learning methods, such as support vector machine (SVM) and naive Bayesian classifier using bag representation or TF-IDF, have improved dictionary results, but they are still limited by simplistic representation. Yang et al. [10] put forward a hybrid China e-commerce review method, which mixes the emotion dictionary in a specific field with the deep learning encoder, and has better accuracy than either of the two methods. Their work shows that it is particularly difficult to analyze Chinese emotions, because word cutting and emotional expression are not direct, which makes normal methods more complicated.

The pre-training language model based on Transformers, especially the Bidirectional Encoder Representation (BERT) from Transformers and its subsequent versions, greatly improved the results of sentiment analysis. These models build rich contextual representations through large-scale pre-training and transfer well to downstream classification tasks with relatively little labeled data. Fine-tuned BERT variants now perform near human level on standard benchmarks, and in e-commerce settings they have enabled aspect-level sentiment extraction: identifying opinions about specific product attributes (e.g., "the battery life is disappointing but the display is excellent"). This level of granularity is useful for product teams and for recommendation engines that seek to match product attributes to user preferences.

Connecting sentiment signals to recommendation and pricing pipelines is a promising direction. Sustained positive review sentiment tends to correlate with price stability, while growing negative sentiment can signal an opportunity to adjust pricing or promotional activity before demand erodes. Real-time sentiment monitoring dashboards, built on streaming NLP pipelines, have been deployed by major platforms to put this connection into practice.

3.4 Dynamic Pricing Strategies

Dynamic pricing—algorithmically adjusting prices based on real-time signals including demand, competitor pricing, inventory levels, and estimated customer willingness-to-pay—is one of the more commercially sensitive BDA applications in e-commerce. Unlike the static, cost-plus pricing common in offline retail, dynamic pricing draws on continuous transaction data to optimize revenue at fine temporal and customer-level granularity.

Machine learning has largely replaced rule-based heuristic pricing engines in sophisticated e-commerce deployments. El Youbi et al. [11] compared gradient boosting machines (GBM), random forests, and regression baselines for dynamic pricing and found that GBM substantially outperformed linear models in capturing non-linear demand-price relationships. Incorporating competitor pricing signals and temporal features (day of week, proximity to holidays) further improved pricing accuracy on a real-world dataset. Reinforcement learning (RL) has attracted growing research interest as a method for sequential pricing optimization, where an agent learns a pricing policy through interaction with a simulated or real market environment. RL-based agents can in principle discover strategies that outperform manually designed heuristics by exploiting complex intertemporal price-demand relationships.

One ongoing tension in dynamic pricing is the balance between revenue optimization and customer fairness. Price personalization—offering different prices to different customers based on inferred willingness-to-pay—is technically feasible but raises equity concerns and may conflict with consumer protection regulations in some jurisdictions. Research has shown that customers who discover they paid more than others for the same product tend to lose trust in the platform and switch to competitors. This has led some companies to restrict personalization to discounts and promotional vouchers rather than applying differential base prices.

Inventory-aware dynamic pricing, which ties pricing decisions to stock levels to avoid both stockouts and overstock, is a major operational application. By combining the demand forecasting model-especially the series model based on long-term short-term memory (LSTM) that we trained with historical sales and promotion calendar data-with real-time inventory data, the pricing engine can automatically price or premium the market for each inventory holding unit (SKU). In promotional activities such as the Double Eleven (11.11), this is very important when the peak demand brings major problems to inventory management.

4. CHALLENGES AND FUTURE DIRECTIONS

4.1 Data Privacy and Regulatory Compliance

The collection and processing of granular behavioral data that BDA-driven precision marketing depends on has drawn increasing regulatory attention. The EU's General Data Protection Regulation (GDPR), enforced since 2018, imposes strict requirements around consent, data minimization, and the right to erasure that directly affect the data collection pipelines precision marketing relies on. Research has documented negative effects of GDPR enforcement on personalized marketing effectiveness and associated revenue [12]. Similar legislation—the California Consumer Privacy Act (CCPA) and its successor—has extended these constraints to other major markets.

For practitioners, the practical response involves developing privacy-preserving analytical techniques capable of delivering personalization without retaining individually identifiable data. Federated learning—a distributed approach in which models are trained across decentralized devices without centralizing raw data—offers one technically viable route. Differential privacy, which adds calibrated noise to model outputs to prevent individual re-identification, provides a complementary layer of protection. Both approaches carry costs in model accuracy, and integrating them into production recommendation systems remains technically demanding. Rather than treating compliance purely as a constraint, companies may find competitive advantage in being seen as responsible data stewards, given that users increasingly factor privacy reputation into their platform choices.

4.2 Algorithmic Transparency and Explainability

The use of increasingly complex models—deep neural networks, gradient boosted ensembles, graph neural networks—creates a real tension between predictive accuracy and interpretability. Regulators in the EU and elsewhere are moving toward requiring intelligible explanations for automated decisions; Article 22 of the GDPR already grants individuals a right to explanation in certain contexts. In marketing, explainability also has a commercial dimension: consumers tend to engage more with recommendations they can understand and find credible.

Existing methods like Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive explanations (SHAP) offer post-hoc attribution of predictions to input features, but they carry computational costs at inference scale and can be sensitive to hyperparameter choices. Developing architectures that are interpretable by design without giving up predictive performance remains one of the harder open problems in this field.

4.3 Cold-Start and Data Sparsity

The cold-start problem—the difficulty of generating useful recommendations for new users or items—persists as a fundamental limitation despite extensive research attention. New product launches are particularly affected: without enough interaction history, the recommendation engine struggles to surface those items to relevant audiences. Incorporating content-based features (product descriptions, images, categorical attributes) and external knowledge graphs helps to some extent, but a meaningful performance gap between new and established items in recommendation ranking remains in practice. Zero-shot and few-shot learning techniques, adapted from the NLP domain, may offer productive paths forward here.

4.4 Cross-Modality Data Integration

E-commerce interactions produce data across multiple modalities: text (search queries, reviews), images (product photos, user-submitted content), audio (voice search), video (live-streaming commerce), and structured transaction records. Most existing BDA approaches process these modalities separately, leaving significant signal unused. Learning joint representations that integrate multiple data types—cross-modal fusion—represents both a substantial technical challenge and a real opportunity for precision marketing. Early work on vision-language models adapted for fashion recommendation suggests the potential, and rapid progress in multimodal foundation models is expected to accelerate practical deployment in the near term.

5. CONCLUSION

This paper discusses the application of big data analysis in precision e-commerce marketing, including four aspects: customer segmentation, personalized recommendation, emotional analysis and dynamic pricing. Research shows that BDA has changed the way of precision marketing. It can aim at everyone in a way that has never been possible before. Machine learning methods, such as RFM clustering, deep collaborative filtering, transforming emotion model and RL pricing agent, push the boundaries that we can analyze.

At the same time, some thorny issues keep coming back. Data privacy rules prevent us from obtaining and using data in the way we want, and we must buy methods to protect privacy. The model is becoming more and more complex, which makes things less clear, which may lead people to no longer trust us. The rules will make us pay more attention to the problems of cold start and sparse data, which makes it difficult for us to provide good suggestions for new products and new platforms. These problems are not just technical problems-they are also related to the big problem in society, that is, to what extent we can use data to sell things.

Future research must give priority to personalized systems that protect privacy, such as federated and differential privacy, which are suitable for use in production. Interpretable AI (suitable for a large number of e-commerce data) is also important. Studying how to merge different data (cross-modal data fusion) and using real-time user behavior to update the model are other important directions. Since the rules are always changing, it is very important for researchers, companies and lawmakers to work together. Therefore, using big data analysis in precision marketing is technically good, good for money and correct for the company.

ACKNOWLEDGMENT

The author would like to thank the School of Business at Shanghai University of Technology for providing institutional support and access to academic databases. No external funding was received for this study.

REFERENCES

[1] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," International J ournal of Information Management, vol. 35, no. 2, pp. 137-144, Apr. 2015. DOI: https://doi.org/10.1016/j.iji nfomgt.2014.10.007

[2] S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," Electronic Markets, vol. 26, no. 2, pp. 173-194, May 2016. DOI: https://doi.org/10.1007/s12525-016-0219-0

[3] S. Erevelles, N. Fukawa, and L. Swayne, "Big Data consumer analytics and the transformation of marketing," Journal of Business Research, vol. 69, no. 2, pp. 897-904, Feb. 2016. DOI: https://doi.org/10.1016/j.jbusres.2015.07.001

[4] J. Ma, "E-commerce Customer Segmentation Based on RFM Model," in Frontier Computing. FC 2021, Lecture Notes in Electrical Engineering, vol. 827, Springer, Singapore, 2022. DOI: https://doi.org/10.1007/978-981-16-8052-6_118

[5] M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, "RFM-based repurchase behavior for customer classification and segmentation," Journal of Retailing and Consumer Services, vol. 61, p. 102566, Jul. 2021. DOI: https://doi.org/10.1016/j.jretconser.2021.102566

[6] C. L. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Information Sciences, vol. 275, pp. 314–347, Aug. 2014. DOI: https://doi.org/10.1016/j.ins.2014.01.015

[7] L. Li, "Cross-Border E-Commerce Intelligent Information Recommendation System Based on Deep Learning," Computational Intelligence and Neuroscience, vol. 2022, Article ID 3798701, 2022. DOI: https:// doi.org/10.1155/2022/3798701

[8] X. Cai, W. Guo, M. Zhao, Z. Cui, and J. Chen, “A knowledge graph-based many-objective model for explainable social recommendation,” IEEE Transactions on Computational Social Systems, vol. 10, no. 6, pp. 3021-3030, Dec. 2022. DOI: https://doi.org/10.1109/TCSS.2022.3161289
[9] M. Zaharia et al., “Apache Spark: A Unified Engine for Big Data Processing,” Communications of the ACM, vol. 59, no. 11, pp. 56-65, Nov. 2016. DOI: https://doi.org/10.1145/2934664
[10] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, “Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning,” IEEE Access, vol. 8, pp. 23522-23530, 2020. DOI: https://doi.org/10.1109/ACCESS.2020.2969854
[11] R. El Youbi, F. Messaoudi, and M. Loukili, “Machine Learning-driven Dynamic Pricing Strategies in E-Commerce,” in Proc. 2023 14th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2023, pp. 1-5. DOI: https://doi.org/10.1109/ICICS60529.2023.10330541
[12] G. A. Johnson, S. K. Shriver, and S. G. Du, “Consumer Privacy Choice in Online Advertising: Who Opts Out and at What Cost to Industry?,” Marketing Science, vol. 39, no. 1, pp. 33-51, Jan.-Feb. 2020. DOI: https://doi.org/10.1287/mksc.2019.1198