Theory and Practice of Science and Technology
2025, VOL. 6, NO. 6, 24-29
DOI: 10.47297/taposatWSP2633-456905.20250606

Machine-Learning Models for Customer-Behavior Analytics: A Review

Haixiang Sun

School of AI and Advanced Computing, XJTLU Entrepreneur College (Taicang), Xian Jiaotong-Liverpool University, Suzhou Jiangsu, 215400, China

ABSTRACT

Customer-behavior analytics has rapidly evolved with the advent of machine learning, enabling businesses to predict and influence customer actions. This review surveys recent machine-learning models applied in customer-behavior analytics, including predictive models for churn, lifetime value, and purchase behavior, as well as clustering techniques for customer segmentation. We compare the performance of different approaches (from simple classifiers to deep learning and ensemble methods) and discuss their relative strengths and weaknesses. A key theme is the trade-off between model complexity and interpretability: while advanced models can improve accuracy, their opaque nature raises concerns addressed by explainable AI. We also identify practical challenges in deploying these models, such as data quality, class imbalance, and evolving consumer trends. Finally, we outline future research directions, highlighting the need for hybrid modeling techniques and improved interpretability to build more trustworthy and effective customer analytics systems.

KEYWORDS

Machine learning; Customer-behaviour analytics; Churn prediction; Customer Lifetime Value (CLV); Customer segmentation; Ensemble learning; Deep learning; Explainable AI (XAI); Model interpretability; Data mining

1 Introduction

Customer-behavior analytics involves using data-driven methods to understand and predict how customers act and make decisions. With the growth of digital commerce and big data, machine-learning (ML) techniques have become essential for extracting insights from customer data and informing business strategies. Companies now routinely apply ML models to a range of customer-focused tasks, from predicting churn (which customers are likely to leave) to estimating customer lifetime value (CLV) and personalizing marketing offers. For instance, ML algorithms are widely used to forecast customer attrition in subscription services and telecom, helping firms identify high-risk customers and intervene to improve retention [1]. In e-commerce, learning algorithms can analyze browsing and purchase histories to predict future buying behavior and tailor recommendations [2]. Even niche domains such as agricultural e-business are leveraging big data and ML for precision marketing, using consumer insights to target promotions more effectively [3]. These examples illustrate the broad impact of ML on customer analytics across industries.

This review focuses on supervised learning models (for prediction tasks like churn, CLV, and purchase propensity) and unsupervised models (for tasks like customer segmentation) that have been applied to customer-behavior data. We draw on recent studies (2023 – 2025) that demonstrate state-of-the-art approaches in this domain. Key questions addressed include: Which ML models are most commonly used for different customer analytics tasks? How do these models compare in terms of predictive performance and applicability? What are the challenges in interpreting and deploying such models in practice? To answer these, we examine representative studies from the literature, compare model outcomes, and discuss emerging themes such as model interpretability and hybrid modeling.

The remainder of this paper is organized as follows. The next section provides an overview of the main types of ML models used in customer analytics and their typical applications. We then present a comparative analysis of these models, highlighting performance benchmarks reported in recent studies. Following that, we discuss interpretability and explainability, which are crucial for adopting ML in customer-facing decisions. We outline current challenges in the field, from technical issues (data quality, class imbalance) to practical deployment concerns. Finally, we suggest future research directions and conclude with a summary of findings.

2 Model overview

2.1 Supervised learning models

A variety of supervised ML algorithms have been employed to predict customer behaviors. These include classic statistical learners like logistic regression and decision trees, more complex ensemble methods (e.g. random forests and gradient boosting), and recent deep learning models. For example, to predict customer churn, researchers have tested algorithms ranging from logistic regression and k-nearest neighbors to ensemble tree-based methods on telecom datasets [1]. Similarly, classification models are used to anticipate post-purchase behaviors such as repeat purchasing or product returns. One study applied four common algorithms – logistic regression, decision tree, random forest, and XGBoost (gradient boosting) – to predict after-sales customer actions in an online retail setting [4]. In the context of customer lifetime value, which often involves regression to estimate monetary value, more advanced techniques like stacked ensembles have been proposed. For instance, Haddadi and Hamidi[5] introduced a stacking ensemble that integrates multiple base learners to improve CLV prediction accuracy.

Deep learning has also made inroads, particularly for tasks involving unstructured data or complex patterns. Geetha et al. [6] present a two-step framework combining deep learning with traditional ML to analyze e-commerce textual reviews for customer satisfaction analytics. In their approach, a deep neural network (e. g. a CNN or LSTM) first learns representations from review text, and then a gradient boosting machine uses those features to predict satisfaction levels. This hybrid deep+ML model capitalizes on deep learning's strength in feature extraction while retaining the predictive power of ensemble methods, yielding better performance than either approach alone [6].

2.2 Unsupervised learning models

Unsupervised techniques, especially clustering algorithms, are crucial for customer segmentation. By grouping customers with similar attributes or behaviors, businesses can tailor strategies for each segment (e.g. premium vs. budget shoppers). Traditional clustering methods like K-means, hierarchical clustering, and density-based clustering (DBSCAN) have all been explored for segmenting retail customers. For example, John et al [7]. conducted an exploration of multiple clustering algorithms to segment customers in the UK retail market. Their study compared the quality of segments obtained by different methods, illustrating how algorithm choice can affect the insights gained (e.g. K-means yielding clear but spherical clusters vs. DBSCAN identifying irregular clusters and outliers) [7]. Unsupervised models do not provide predictive outcomes, but they help reveal latent groupings and patterns in customer data, which can inform targeted marketing and personalized services.

2.3 Context-specific models

Some ML models are tailored to specific contexts of customer analytics. In e-commerce and digital marketing, recommendation systems (collaborative filtering, etc.) and sequence models (for clickstream analysis) come into play, though these are beyond the primary scope of this review. It is worth noting that hybrid approaches are emerging which blend different model types to address complex customer behavior phenomena. One example is a meta-modeling approach that integrates multiple algorithms (ensemble + neural networks) to improve churn prediction for online platforms [8]. Such hybrid models aim to capture diverse patterns in customer data by leveraging the complementary strengths of various algorithms.

Table 1 summarizes selected studies and the ML techniques they employ for key customer-behavior analytics tasks:

Table 1 REPRESENTATIVE ML APPLICATIONS IN CUSTOMER-BEHAVIOR ANALYTICS

Study	Application	Techniques	Key Finding
[1]	Customer churn (telecom)	Logistic reg., Random Forest, KNN, XGBoost	Gradient boosting achieved highest accuracy (~84%), slightly outperforming logistic regression.
[4]	After-sales behavior (e-commerce)	Logistic reg., Decision Tree, Random Forest, XGBoost	Random Forest yielded the best performance (highest F1/AUC), after addressing class imbalance with SMOTE.
[5]	Customer lifetime value (online retail)	Stacking ensemble (multiple regressors)	Stacking ensemble outperformed individual models, improving CLV prediction accuracy.
[7]	Customer segmentation (retail)	K-means, Hierarchical, DBSCAN clustering	Different algorithms produced varying segment patterns; highlighted the importance of algorithm selection for meaningful segments.
[6]	Satisfaction prediction (e-commerce reviews)	CNN-based deep learning + XGBoost	Hybrid deep learning + ML model improved predictive accuracy over using either model type alone.
[8]	Customer churn (online services)	Meta-model (ensemble + neural nets)	Integrated approach demonstrated better churn prediction performance than single classifiers.
[2]	Online consumer behavior (browsing)	Machine learning analysis of browsing metrics	Showed that metrics like time spent reading product info can predict purchase intent, providing insights for conversion optimization.
[9]	Explainability in AI (general)	Survey of XAI methods	Emphasizes the need for explainable models to ensure trust and transparency in AI-driven customer analytics.
[3]	Precision marketing (agri e-commerce)	Big data analytics, data mining	Demonstrated use of big data techniques to personalize marketing for agricultural e-commerce, improving targeting effectiveness.

3 Model comparison

The performance and suitability of ML models in customer analytics can vary widely depending on the task and data characteristics. Predictive accuracy is often a primary metric for model comparison. Across studies, more complex models (ensemble methods and deep learning) tend to achieve higher accuracy than simpler models, albeit sometimes by modest margins. For example, in telecom churn prediction, a gradient boosting model slightly outperformed a logistic regression (about 84% vs 83% accuracy) in one experiment [1]. Similarly, an e-commerce study found that a random forest achieved better predictive metrics (F1 score, AUC) for post-purchase behavior than a single decision tree or logistic model [4]. These findings indicate that ensemble techniques, which aggregate the predictions of multiple learners, can capture customer behavior patterns more effectively than any single classifier.

That said, no one model is universally best. Interestingly, boosting did not always win out; in the after-sales behavior study, random forest slightly surpassed XGBoost, suggesting that model performance can depend on the dataset and tuning [4]. Deep learning models, known for their capacity to model complex, non-linear relationships, have shown improved performance in scenarios with rich data (e.g. text reviews, high-dimensional features). Geetha et al [6]. report that incorporating a deep neural network for feature learning led to better accuracy than using gradient boosting on raw features alone. However, deep models typically require large datasets and can be prone to overfitting if data is limited or not representative. In some cases, traditional ML methods can rival deep learning on tabular customer data. For instance, a well-calibrated logistic regression or decision tree might perform competitively on structured customer records, especially after careful feature engineering.

When comparing models, it is also important to consider speed and scalability. Simpler models like logistic regression and decision trees are fast to train and easy to implement, making them attractive for real-time or on-device analytics. Ensembles and deep networks demand more computational resources and may not be necessary if incremental accuracy gains are small. In customer analytics contexts where thousands or millions of predictions must be made regularly (such as daily churn scoring for a large subscriber base), the trade-off between accuracy and efficiency becomes relevant. Some studies opt to use simpler models if they offer comparable performance, due to easier deployment. For example, if a logistic model achieves nearly the same churn prediction accuracy as a complex ensemble, practitioners might prefer the logistic model for its simplicity and maintainability [1].

Evaluation criteria: Beyond raw accuracy, other evaluation criteria influence model choice. These include precision/recall trade-offs (important in churn to minimize false positives vs. false negatives), cost-sensitivity (e.g. weighting high-value customers more in CLV predictions), and stability over time. Model comparison in research often involves cross-validation and statistical tests to ensure differences are significant. For instance, Chen et al [4]. used 5-fold cross-validation to compare models for after-sales behavior, giving a robust basis to declare random forest the top performer. In unsupervised modeling like clustering, evaluation is more qualitative: John et al [7]. assessed clustering outcomes by their business usefulness (how well segments differentiated customer groups in spend and preferences) rather than a single numeric metric.

Another aspect of comparison is model complexity vs. interpretability, which bridges into the next section. Some companies may favor slightly less accurate models if they are far more interpretable and thus actionable. In summary, recent literature suggests that ensemble methods often provide strong performance for predicting customer behaviors, and hybrid models can push the accuracy frontier further, but the benefits must be weighed against practical considerations like interpretability, computational cost, and data availability.

4 Interpretability

While machine-learning models can greatly improve the accuracy of customer-behaviour analytics, their adoption in business decisions hinges on interpretability. Marketing and customer relationship teams need to understand why a model makes a particular prediction (e.g. why a certain customer is flagged as high churn-risk) in order to trust and act on it. However, many high-performing models are black boxes; for example, an ensemble of hundreds of trees or a deep neural network with many layers does not readily explain its outputs. This lack of transparency can be a barrier in domains where explanations are required for decision-makers or to meet regulatory standards.

The emerging field of Explainable AI (XAI) addresses this issue by developing techniques to make model predictions more understandable. Ali et al [9]. provide a comprehensive overview of XAI methods and emphasize their importance for achieving trustworthy AI. They argue that explainability is not just a technical add-on, but a foundational requirement for AI systems in finance, healthcare, and by extension, customer analytics. Common XAI approaches include post-hoc explanation tools like LIME and SHAP, which can highlight which features (e.g. customer age, contract length, number of support tickets) most influenced a specific prediction. By applying such techniques, even a complex model's decision can be broken down into human-intelligible factors. For example, a churn prediction model's output might be accompanied by a ranked list of reasons ("customer tenure 1 year" and "high complaint count" as top factors), helping managers devise appropriate retention strategies.

Another strategy to enhance interpretability is to use inherently interpretable models when feasible. Simple decision trees, linear models, or rule-based classifiers allow direct insight into how inputs relate to predictions. In customer segmentation, clustering results can be made interpretable by profiling each cluster (e.g. Cluster A: young urban shoppers with high online engagement). John et al. [7] effectively interpret their segmentation by describing distinguishing attributes of each cluster, which makes the insights actionable for marketing plans. However, relying solely on simple models may sacrifice accuracy, as noted earlier. A practical compromise is the use of hybrid modeling: for instance, using an accurate black-box model to make predictions but pairing it with a surrogate interpretable model for explanation. Some recent churn studies adopt this approach, where a complex model flags at-risk customers and a simpler model is used to approximate the decision boundary for explanation purposes [9].

The push for interpretability is also driven by ethical and legal considerations. Regulations (like GDPR in Europe) give individuals the right to an explanation of algorithmic decisions, which can apply to scenarios like credit scoring or personalized pricing in e-commerce. Being able to justify why one customer is offered a discount and another is not is crucial for fairness and accountability. Therefore, customer analytics solutions increasingly must incorporate interpretable mechanisms. Researchers suggest that future customer-behaviour models will need to be "glass box" to some degree – either interpretable by design or augmented with robust explanatory tools – to be fully accepted in operational use [9].

5 Challenges

Despite significant progress, there are several challenges in applying machine-learning to customer-behavior analytics.

Customer data often come from disparate sources (sales transactions, web analytics, customer service logs, social media, etc.). Integrating these into a cohesive dataset is non-trivial, and the resulting data may contain inaccuracies or missing values. Models are highly sensitive to data quality; poor data can lead to spurious patterns. Studies frequently spend considerable effort on data preprocessing. For example, Prabadevi et al. [1] emphasize techniques for handling missing data and outliers in churn datasets before model training. Ensuring data cleanliness and combining structured and unstructured data (like numerical purchase histories with text reviews) remain ongoing challenges.

Many predictive customer analytics tasks suffer from imbalanced data. In churn prediction, the proportion of customers who churn can be relatively small, meaning models may become biased toward the majority (non-churn) class. Similar issues arise in fraud detection or in predicting rare events. Researchers address this with strategies like SMOTE oversampling and class weighting [4], but achieving good recall on the minority class without too many false alarms is difficult. Imbalance can lead to models that look accurate overall yet miss the most crucial cases (e.g. failing to identify truly at-risk customers), so careful evaluation metrics (precision, recall, AUC) are essential.

Customer preferences and market conditions change over time. A model trained on last year's data may become less accurate as new products emerge or competitors enter the market. This concept drift means models require periodic retraining or online learning to stay current. Capturing evolving behavioral patterns is challenging; for instance, a surge in digital engagement post-pandemic can invalidate assumptions made by earlier models. Few of the surveyed studies explicitly address drift, often working on static datasets. Developing models and pipelines that adapt to real-time data streams is still an area for improvement in customer analytics.

Customer data is sensitive, and using ML on personal data raises privacy issues. Stricter data protection regulations restrict how data can be used and shared. This can limit the richness of data available for modeling (for example, not being able to use certain identifiable attributes). Moreover, some ML models might inadvertently encode biases present in historical data, leading to unfair outcomes (e.g. systematically excluding a demographic group from high-value offers). Ensuring ethical AI – models that are both privacy-preserving and bias-mitigated – is a challenge that intersects with interpretability. Ali et al. [9] note that transparency can help detect and correct bias, but proactive measures (like bias audits and using privacy-enhancing techniques) are needed as well.

Bringing ML models from the research stage into a production business environment poses practical difficulties. Deployed models need to be robust, scalable, and integrate with existing customer relationship management systems. There can be a gap between the promising results reported in academic papers and the performance of those models in a live setting with noisy data and feedback loops. Maintenance involves monitoring model performance over time, recalibrating models as needed, and managing the feedback loop (where model-driven actions influence future customer behavior, complicating the data). These operational challenges mean that organizations must invest in MLOps (Machine Learning Operations) for customer analytics, which not all companies are prepared for.

6 Future directions

Looking ahead, research and practice in customer-behavior analytics are likely to focus on several promising directions.

Hybrid and Ensemble Techniques: As seen in recent studies, combining multiple modeling approaches often yields superior results. Future work will explore more sophisticated hybrid models – for example, blending deep learning with gradient boosting, or ensemble frameworks that incorporate both supervised and unsupervised components. The goal is to capture the multifaceted nature of customer behavior (spanning quantitative metrics and qualitative insights) in one predictive system. For instance, an ensemble could integrate a neural network (capturing complex feature interactions) with a simpler model (adding stability and interpretability), leveraging the strengths of each [5-6].

Explainability and Human-in-the-Loop AI: There will be continued emphasis on making customer analytics models transparent and actionable. Future ML systems may include built-in interpretability, allowing marketers to query "Why did the model recommend this action?" on the fly. Interactive, human-in-the-loop approaches are expected to gain traction: analysts could adjust model inputs or rules and see how outcomes change, thereby blending human expertise with AI insights. Ali et al [9]. suggest an agenda for explainability activism, where stakeholders demand clarity from AI – this aligns with business needs in marketing to justify decisions to executives and customers alike. We can anticipate new techniques that produce explanations as part of the prediction output, not as an afterthought.

Real-Time and Streaming Analytics: Customer behaviors can be transient; capturing real-time signals (like current session activities on a website or recent transaction sequences) is crucial for timely interventions. Future models will increasingly handle streaming data and perform online learning, updating their predictions as fresh data arrives. This is important for applications such as dynamic pricing or instant churn alerts based on live customer actions. Advancements in algorithms that learn incrementally, as well as infrastructure for low-latency data processing, will support this shift. Studies might evolve from batch-training models on historical data to evaluating how models can continuously learn and adapt in production environments.

Domain Adaptation and Generalization: A practical direction is developing models that generalize across different domains or markets. For example, a model trained on e-commerce retail data might be adapted to work for a subscription service context with minimal retraining. Techniques like transfer learning could allow knowledge learned about one type of customer behavior to inform predictions in a new setting. This would be valuable for companies expanding into new regions or product lines with limited data. Research could focus on what fundamental patterns in customer analytics are transferable and how to calibrate models to local nuances.

Integrated Customer Analytics Platforms: We may see holistic platforms that combine descriptive, predictive, and prescriptive analytics. Rather than treating each model (churn, CLV, segmentation, etc.) in isolation, future systems could unify these into a 360-degree customer analytics framework. Such a system might use one model's output as another's input (for instance, using segment labels as features in a churn model) to improve overall understanding. This integrated approach can benefit from multi-task learning – training models jointly to predict several related customer metrics, thereby sharing information across tasks. Early research in multi-objective optimization for customer analytics hints at potential gains in efficiency and consistency of insights.

In summary, the future of machine-learning in customer-behavior analytics lies in more intelligent, transparent, and adaptable models. The continued collaboration between domain experts (marketing analysts) and data scientists will be essential to ensure that these advanced techniques remain grounded in real business needs and constraints.

7 Conclusion

Machine-learning models have become indispensable tools in deciphering customer behavior and guiding customer-centric decision-making. This review has examined the landscape of ML approaches used in customer-behaviour analytics, covering predictive models for churn, lifetime value, and purchase propensity, as well as clustering methods for customer segmentation. Recent contributions show that advanced models – from ensemble learners to deep neural networks – can significantly enhance prediction accuracy and provide actionable insights, especially when applied to rich customer datasets. At the same time, we observed that simpler models still hold value, particularly when interpretability and ease of deployment are paramount considerations.

A critical insight is that no single model suits all scenarios. Effective customer analytics often requires a blend of methods and a careful balance between accuracy and transparency. The rise of explainable AI techniques [9] underlines the importance of making models' inner workings accessible to stakeholders, ensuring that predictive insights lead to trust and meaningful actions. We also highlighted ongoing challenges such as data integration, class imbalance, and model adaptability in dynamic market conditions. Addressing these challenges is crucial for translating technical model improvements into sustained business impact.

In closing, machine learning offers powerful capabilities to predict and influence customer behavior, but its successful application rests on more than just algorithms. It requires high-quality data, thoughtful model selection (and often combination), interpretability, and continuous learning mechanisms. The research trends and future directions identified – including hybrid modeling and real-time analytics – indicate a maturing field moving towards comprehensive, human-aligned AI solutions for customer analytics. By building on these developments, organizations can better anticipate customer needs and foster stronger, longer-lasting customer relationships through data-driven strategies.

About the Author

Haixiang Sun, Haixiang.Sun22@student.xjtlu.edu.cn.

References

[1] B. Prabadevi, R. Shalini, and B. R. Kavitha, Customer churning analysis using machine learning algorithms [J].International Journal of Intelligent Networks, vol. 4, pp. 145–154, 2023.

[2] H. H. Chang and A. Mukherjee, Machine learning and consumer data [J].arXiv preprint arXiv:2306.14118, 2023.

[3] J. Du, Predictions for consumer behaviour of e-commerce sales data 2023–2024 based on the LightGBM model [J].In Proceedings of the 1st international conference on e-commerce and artificial intelligence (ECAI 2024), Beijing, China: Science; Technology Publications, 2024, pp. 109–116.

[4] Z. Chen, The investigation of machine learning approaches for customer segmentation, in Proceedings of the 1st international conference on e-commerce and artificial intelligence (ECAI 2024). [J].South Bridge, Edinburgh, U.K.: Science; Technology Publications, 2024, pp. 71–76.

[5] A. M. Haddadi and H. Hamidi, A hybrid model for improving customer lifetime value prediction using stacking ensemble learning algorithm [J].Computers in Human Behavior Reports, p. 100616, 2025.

[6] M. Geetha and D. Karthika Renuka, Deep learning architecture towards consumer buying behaviour prediction using multitask learning paradigm [J].Journal of Intelligent Fuzzy Systems, vol. 46, no. 1, pp. 1341–1357, 2024.

[7] J. M. John, O. Shobayo, and B. Ogunleye, An exploration of clustering algorithms for customer segmentation in the UK retail market [J]. Analytics, vol. 2, no. 4, pp. 809–823, 2023.

[8] B. A. Maulana and N. Hidayati, Churn prediction in credit customers using random forest and XGBoost methods [J].Indonesian Journal of Data and Science, vol. 6, no. 1, pp. 81–89, 2025.

[9] S. Ali et al., Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence [J].Information fusion, vol. 99, p. 101805, 2023.