



An Efficient Data Mining Framework for Customer Behavior Analysis in E-Commerce Applications

Aurelio Zuberiwala

Department of Mechatronics Engineering, Sundarban College of Technology Studies, Bangladesh
aurelio.zuberiwala@scts-bd.net

Peer Review Information	Abstract
<i>Submission: 02 Jan 2018</i> <i>Revision: 17 Feb 2018</i> <i>Acceptance: 04 March 2018</i> Keywords <i>E-Commerce, Customer Behavior, Data Mining, Customer Segmentation, Purchase Intention</i>	E-commerce platforms generate large volumes of transaction, clickstream, session, and contextual data, but these records become valuable only when they are converted into interpretable customer-behavior patterns. This methodology paper proposes an efficient data mining framework that integrates preprocessing, behavioral feature construction, customer segmentation, association-pattern discovery, and purchase-intention classification within a single analytical pipeline. The framework is designed to combine descriptive and predictive analysis: clustering identifies behaviorally similar customer groups, association rules reveal co-occurring actions or product interests, and supervised classification estimates whether an active session is likely to end in a purchase. A 2018 public benchmark, the Online Shoppers Purchasing Intention Dataset, is used to define the validation setting. It contains 12,330 independent sessions, with 1,908 purchase sessions and 10,422 non-purchase sessions, making class imbalance an important methodological consideration. Efficiency is improved through early data reduction, selective feature engineering, standardized transformation, and evaluation metrics that emphasize minority-class detection rather than raw accuracy alone. The proposed approach provides a reproducible structure for e-commerce analytics and supports customer segmentation, personalized offers, retention planning, and website optimization without depending on a single mining technique.

Introduction

Digital commerce systems continuously record behavioral traces such as page visits, browsing duration, product interactions, referrals, cart activity, orders, and repeat visits. These records allow customer analysis to move beyond demographic description toward evidence based on observed actions. However, e-commerce data are heterogeneous and highly imbalanced: most sessions do not produce a transaction, while a relatively small proportion of visitors complete a purchase. Consequently, an effective analytical framework must discover meaningful customer groups and patterns while also identifying the small number of high-intent sessions.

Early clickstream research showed that browsing depth, page-view duration, and repeat visits contain information that is not captured by aggregate website statistics. Bucklin and Sismeiro (2003) demonstrated that individual-level clickstream behavior changes dynamically during a visit, while Moe (2003) showed that navigation sequences can distinguish buying, searching, browsing, and knowledge-building visits. These findings establish a foundation for customer-behavior analysis in which the sequence and intensity of website interactions are treated as behavioral signals rather than simple traffic counts.

Data mining extends this perspective by combining customer value estimation, segmentation, rule discovery, and prediction. Chen, Chiu, and Chang (2005) integrated behavioral variables with transaction information to identify changes in customer behavior, and Chen, Sain, and Guo (2012) demonstrated how recency, frequency, and monetary value can be used for customer-centric online retail analytics. Later e-commerce studies combined web-usage mining, customer clustering, and predictive models to support website design and targeted service (Carmona et al., 2012; Suchacka & Chodak, 2017). The objective of the present paper is therefore to define a compact methodology that unifies these tasks into one efficient framework suitable for e-commerce applications and datasets available up to 2018.

Literature Review

Moe (2003) analyzed page-to-page navigation and proposed that online visits can be differentiated by underlying shopping objectives. The study is important because two visitors who view a similar number of pages may still exhibit different intent depending on the types and order of pages visited. For a data mining framework, this implies that session-level features should represent both activity volume and behavioral context. Bucklin and Sismeiro (2003) similarly modeled continuation and exit decisions from clickstream logs and reported that aggregate website metrics may diverge from individual-level behavior. Together, these studies motivate the use of session-level observations and careful feature construction.

Chen et al. (2005) focused on changes in retail customer behavior over time. Their approach combined RFM-type customer value variables with transaction information and association-rule-based pattern comparison. The study demonstrates that customer behavior is dynamic; therefore, a useful framework should not only form static segments but should also allow periodic re-mining of patterns. Chen et al. (2012) further showed that RFM-based segmentation provides a practical basis for grouping online retail customers and translating transactional histories into actionable customer categories.

Carmona et al. (2012) presented a web-usage-mining process for an e-commerce website that included data collection, preprocessing, clustering, association rules, and subgroup discovery. Their work directly supports a multi-stage architecture rather than a single-algorithm

solution. Suchacka and Chodak (2017) extended this line of research by using association rules to identify session features associated with purchase probability, demonstrating that rule mining can provide interpretable behavior patterns that complement predictive classifiers.

Recent work up to 2018 strengthened the role of segmentation and integrated analytics in online environments. Ballestar, Grau-Carles, and Sainz (2018) applied data science to large-scale e-commerce transactions and showed that customer segments can be linked to differences in commercial activity and customer journeys. Wong and Wei (2018) integrated customer segmentation with customized-service prediction, emphasizing that descriptive customer groups and predictive service decisions can be developed together. An et al. (2018) also demonstrated that behavioral and demographic data from online platforms can be aggregated to isolate actionable customer segments. These studies support the central design choice of the proposed framework: descriptive mining and predictive modeling should share a common preprocessing and feature layer.

Methodology

1. Research Design and Data Sources

The proposed framework follows a sequential knowledge-discovery design. Data are first collected from e-commerce transaction records, web analytics or server logs, customer profiles, and session context. A transaction table may include customer identifier, order time, product, quantity, price, and order status. A clickstream or analytics table may include page categories, page counts, time spent, bounce rate, exit rate, referral source, visitor type, and device or browser information. Customer identifiers are used only when permitted by the application; otherwise, session-level analysis can be performed with anonymized identifiers.

For benchmark validation, the methodology adopts the Online Shoppers Purchasing Intention Dataset released in 2018 by Sakar and Kastro (2018). The dataset contains 12,330 sessions and records page counts and durations for administrative, informational, and product-related pages together with bounce rate, exit rate, page value, special-day proximity, month, operating system, browser, region, traffic type, visitor type, weekend status, and the binary Revenue outcome. UCI reports no missing values and states that each session corresponds to a different user during the observed one-year period, reducing repeated-user campaign bias.

2. Data Preprocessing

Preprocessing is performed before any model is fitted. Duplicate sessions and invalid transaction records are removed; cancellations and returns are separated from completed purchases when transactional data are available. Numerical variables are checked for extreme values and are log-transformed when highly skewed. Categorical variables such as visitor type, month, browser, and traffic source are encoded using one-hot or frequency coding. Continuous features are standardized for distance-based clustering. To prevent leakage, transformations are estimated only on the training portion of the data and then applied to validation or test records.

Class imbalance is handled explicitly because non-purchase sessions dominate typical e-commerce datasets. Instead of evaluating models only through accuracy, the framework uses precision, recall, F1-score, balanced accuracy, and ROC-AUC. When resampling is required, oversampling or class weighting is applied only to training folds so that the test distribution remains realistic. This design prevents an apparently high accuracy from masking poor detection of actual buyers.

3. Behavioral Feature Engineering

The feature layer converts raw activity into compact behavior indicators. For transaction histories, RFM variables are calculated as Recency = time since the latest purchase, Frequency = number of completed orders in the observation window, and Monetary = total or average spending. Additional variables include average basket value, product diversity, discount usage, return rate, and purchase interval. For web sessions, engagement is represented using product-page count and duration, informational-page count and duration, bounce and exit rates, page value, repeat-visitor status, weekend flag, and traffic source. The framework may also derive ratios such as product-page time per product page and product-related activity as a share of all viewed pages.

4. Customer Segmentation

Customer or session segmentation is performed using K-means when variables are continuous and standardized. The number of clusters is selected using the elbow criterion together with silhouette analysis and managerial interpretability. Cluster profiles are described using medians or standardized means rather than only cluster labels. In transaction-rich applications, segments may represent high-value loyal customers, frequent low-value purchasers,

recently acquired customers, or inactive customers. In session-level applications, clusters may represent focused product explorers, high-engagement researchers, quick bouncers, and returning comparison shoppers. The use of clustering is consistent with the segmentation findings of Chen et al. (2012), Ballestar et al. (2018), and An et al. (2018).

5. Association and Sequential Pattern Mining

Association-rule mining is used to identify interpretable co-occurrence patterns. For market-basket data, items or product categories are represented as transactions and rules of the form $X \rightarrow Y$ are evaluated using support, confidence, and lift. For behavioral sessions, discretized conditions such as high product-page duration, returning visitor status, or low exit rate can be associated with purchase outcomes. Minimum support is kept sufficiently high to avoid unstable rare patterns, while lift greater than one is used to identify positive associations beyond chance. The interpretability of this stage is important because marketing or website teams can translate rules directly into cross-selling, navigation, or targeting actions, as illustrated by Chen et al. (2005) and Suchacka and Chodak (2017).

6. Purchase-Intent Classification

The predictive stage estimates the probability that a current session will end in a purchase. A decision tree is used as an interpretable baseline, while logistic regression or a nonlinear classifier can be added when predictive performance is more important. The data are divided using a stratified split so that the purchase ratio is preserved. Hyperparameters are selected through cross-validation on the training data. The target is the Revenue variable or an equivalent completed-purchase flag. Feature importance or coefficient direction is reported to connect predictions back to behavioral drivers, avoiding a purely black-box output.

7. Efficiency and Decision Integration

Efficiency is addressed at three levels. First, unnecessary raw fields are discarded after validation, reducing memory and computation. Second, a shared feature table is reused by clustering, rule mining, and classification instead of repeating transformation steps. Third, results are stored as compact customer or session scores: segment label, dominant behavior pattern, purchase probability, and recommended action. Batch segmentation may be executed daily or weekly, while purchase-intent scoring can be applied to active sessions in near real time.

This layered design reflects the integrated analytics approach discussed by Wong and Wei

(2018) while retaining the web-usage focus of Carmona et al. (2012).

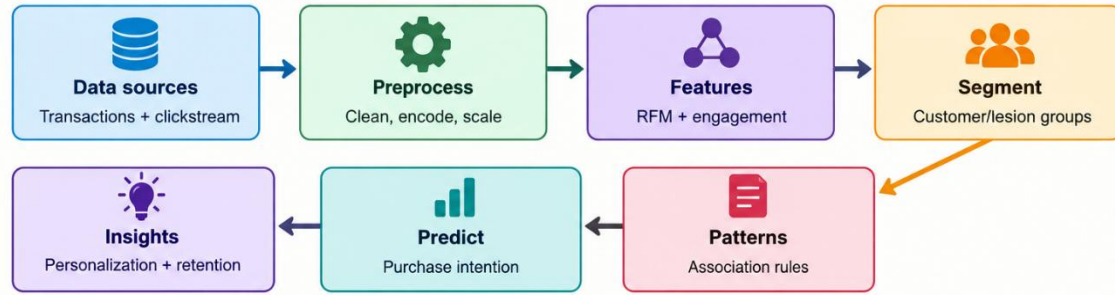


Figure 1. Proposed efficient data mining framework for e-commerce customer behavior analysis.

Results

The benchmark data profile demonstrates why a multi-stage framework is necessary. The 2018 UCI dataset contains 12,330 independent sessions, of which 10,422 (84.5%) did not end in a purchase and 1,908 (15.5%) did end in a purchase (Sakar & Kastro, 2018). This corresponds to approximately 5.46 non-purchase sessions for every purchase session. A trivial classifier that predicts “no purchase” for every visit would therefore obtain about 84.5% accuracy while identifying none of the actual buyers. This result confirms that accuracy alone is not an adequate evaluation criterion for e-commerce purchase-intent analysis.

The framework converts the raw feature space into three analytical outputs. First, clustering

creates behaviorally coherent customer or session groups that can be compared by engagement and value. Second, association rules expose interpretable relationships between browsing conditions, product interests, and outcomes. Third, classification produces a purchase-probability score for each session. These outputs are complementary: segmentation explains broad behavior types, rules explain recurring combinations of actions, and classification supports prioritization of individual sessions. The combination is consistent with prior evidence that e-commerce behavior benefits from both descriptive and predictive mining rather than a single technique (Carmona et al., 2012; Wong & Wei, 2018).

Table 1. Benchmark data profile and methodological interpretation

Measure	Observed value	Framework implication
Benchmark sessions	12,330	Sufficient for classification and clustering experiments
Purchase sessions	1,908 (15.5%)	Minority class; recall and F1-score are essential
Non-purchase sessions	10,422 (84.5%)	Dominant class; raw accuracy can be misleading
Class ratio	$\approx 5.46 : 1$	Use stratification, weighting or training-only resampling
Predictor structure	Behavioral, traffic and contextual variables	Supports segmentation, rule mining and prediction
Missing values	None reported by UCI	Preprocessing focuses on coding, scaling and leakage control

From a decision perspective, the descriptive results indicate that e-commerce analytics should be optimized for ranking and identifying a small group of likely buyers while still retaining interpretable customer segments. For example, a high-intent session may combine sustained product-page engagement, low exit behavior, returning-visitor status, and positive page-value signals, while a low-intent cluster may exhibit short product visits and high bounce behavior. These are hypotheses to be tested within the

proposed validation process rather than fixed rules. The framework is therefore designed to produce site-specific segments and rules rather than assuming that behavior patterns transfer unchanged between retailers.

Conclusion and Discussion

This paper presents an efficient methodology for analyzing customer behavior in e-commerce applications through an integrated data mining pipeline. The framework begins with transaction,

clickstream, and contextual data; applies cleaning, encoding, scaling, and feature construction; then combines segmentation, association-rule mining, and purchase-intent classification. The literature up to 2018 shows that each of these analytical perspectives contributes different information: clickstream studies explain visit behavior, RFM and clustering identify customer groups, association rules reveal interpretable relationships, and predictive models support session-level action (Moe, 2003; Chen et al., 2012; Suchacka & Chodak, 2017). The 2018 UCI benchmark also highlights a practical challenge: only 15.5% of sessions represent purchases. Consequently, a method that appears accurate can still be ineffective if it fails to detect the minority purchasing class. The proposed validation strategy therefore emphasizes stratified evaluation and precision, recall, F1-score, balanced accuracy, and ROC-AUC. This makes the evaluation more aligned with the business objective of finding likely buyers rather than simply describing the majority of non-purchasing visitors. The main contribution is methodological integration and efficiency rather than a claim that one algorithm is universally superior. The framework can be adapted to different e-commerce settings by changing the feature layer and model family while keeping the same workflow. Future implementations should additionally test temporal drift, customer privacy constraints, product-season effects, and model calibration. Periodic re-segmentation is especially important because customer behavior changes over time, a concern already emphasized in retail behavior mining (Chen et al., 2005). With these safeguards, the framework provides a practical foundation for personalized marketing, customer retention, cross-selling, and website optimization.

References

- An, J., Kwak, H., Jung, S.-g., Salminen, J., & Jansen, B. J. (2018). Customer segmentation using online platforms: Isolating behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining*, 8, 54. <https://doi.org/10.1007/s13278-018-0531-0>
- Ballestar, M. T., Grau-Carles, P., & Sainz, J. (2018). Customer segmentation in e-commerce: Applications to the cashback business model. *Journal of Business Research*, 88, 407–414. <https://doi.org/10.1016/j.jbusres.2017.11.047>
- Bucklin, R. E., & Sismeiro, C. (2003). A model of web site browsing behavior estimated on clickstream data. *Journal of Marketing Research*, 40(3), 249–267. <https://doi.org/10.1509/jmkr.40.3.249.19241>
- Carmona, C. J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M. J., & García, S. (2012). Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. *Expert Systems with Applications*, 39(12), 11243–11249. <https://doi.org/10.1016/j.eswa.2012.03.046>
- Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3), 197–208. <https://doi.org/10.1057/dbm.2012.17>
- Chen, M.-C., Chiu, A.-L., & Chang, H.-H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, 28(4), 773–781. <https://doi.org/10.1016/j.eswa.2004.12.033>
- Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, 13(1–2), 29–39. https://doi.org/10.1207/S15327663JCP13-1&2_03
- Sakar, C., & Kastro, Y. (2018). Online Shoppers Purchasing Intention Dataset [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5F88Q>
- Suchacka, G., & Chodak, G. (2017). Using association rules to assess purchase probability in online stores. *Information Systems and e-Business Management*, 15, 751–780. <https://doi.org/10.1007/s10257-016-0329-4>
- Wong, E., & Wei, Y. (2018). Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction model. *International Journal of Retail & Distribution Management*, 46(4), 406–420. <https://doi.org/10.1108/IJRDM-06-2017-0130>