Data-Driven Customer Retention for SMEs: Predicting Repeat Purchase and Customer Value

Ilknur Ozturk[1,*]

[1]Faculty of Economics, Administrative and Social Sciences, Nisantasi University, Istanbul, Turkey

Email: ilknur.ozturk@nisantasi.edu.tr

Abstract

The strategic importance of customer retention in small and medium-sized enterprises (SMEs) is due to the fact that the resources are limited, and the indiscriminate customer acquisition and customer retention campaigns are economically inefficient. However, the descriptive reporting used by many SMEs does not have the advantages of transaction-driven analytics that allows differentiating between high-value and low-yield customer relationships. This paper creates a repli-cable customer-analytics pipeline in SME-type retail environments, using publicly available transactional data. In con-trast to the macro-level forecasting research, the paper integrates customer value segmentation with the future-oriented repeat-purchase prediction and translates the results into retention actions explicitly. The customer-level features were based on invoices, quantities, prices, product variety, and return behavior and were derived using the public Online Retail dataset. Observation windows on a monthly were transformed into a repeat-purchase 90-day problem. Three predictive models—logistic regression, random forest, and gradient boosting—were compared after customer segmentation based on recency, frequency, and monetary behavior. The findings indicate that random forest model had the highest discrimination (ROC-AUC = 0.750; PR-AUC = 0.821), followed by logistic regression, which was only slightly less than it and more interpretable. Segment analysis also showed a very concentrated revenue base with *Champions* having 27.5 percent of the customers but 67.2 percent of recent revenue and 81.0 rate of repeat purchasing. The paper provides a submission-ready, transparently reproducible, and managerially understandable design that is particularly applicable in SMEs that want low-cost retention analytics, customer ranking, and allocation of marketing resources.

Keywords: Customer retention; SMEs; Repeat purchase; Customer value; Business data analytics; Retail analytics

1. Introduction

In the case of SMEs, retaining customers is not a marketing concern but rather a resource management matter. Restricted budgets, minimal staffing designs, and reduced data infrastructures imply that SMEs are frequently not able to focus on all customers to a similar level of degree. Practically, this brings a requirement of customer analytics systems capable of responding to two managerial questions concurrently: who are the customers that are *w*orth protecting, and who are the customers that are likely to repurchase. That two-fold question has increasingly become urgent as digital channels are producing more detailed records of transactions even as more competition is increasing the cost of acquiring customers broadly and undifferentiated [2, 5, 6].

Some recent studies in business intelligence and business data analysis have demonstrated that firms that rely on data outperform others by aligning data-driven analytic outputs with actual managerial activities like segmentation, targeting, pricing, and demand planning [13, 14]. Nonetheless, the real issue faced by SMEs is not merely to get the right models. The more basic question is how to convert raw transaction histories into a retention portfolio of action. Recent articles provide valuable components of that puzzle: some of them see the adoption of analytics in SMEs, others focus on purchase or churn prediction within an online context, and a third wave looks at the relationship marketing, customer lifetime value, and repurchase behavior. But these streams are not usually joined. Studies of adoption-oriented SME tend to end at a capability and performance relation, predictive studies tend to focus on algorithmic precision more than on managerial parsimony, and retention-oriented studies not necessarily incorporate a customer value segmentation with model-driven future-buying.

The aim of this paper is to fill that gap with the purpose of having a different design as compared to the previous macroeconomic forecasting manuscript. Instead of constructing a forecasting model based on external cues, the current study considers the use of *customer intelligence based on transactions*. It analyses the public invoice-level retail data to construct customer snapshots with numerous observation windows. The snapshot is characterized by recency, frequency, monetary value, basket properties, product diversification, tenure and the behavior of the returns. Two uses of these features are then made. This is because they first create customer value segments that show the existing revenue structure of

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026

the portfolio. Second, they assist in a progressive classification task which forecasts a repeat purchase in the next 90 days of a customer. These two blocks can be integrated to enable managers to integrate value and propensity within a single decision logic.

The study contributions are four-fold. First, it provides a completely reproducible analytical design with publicly available data as opposed to corporation-specific data. Second, it builds a retention-driven design that fits better in SME situations than designs that need intricate clickstream, subscription, or multi-source enterprise data. Third, it shows that interpretable analytics can still compete with more flexible machine-learning models, which are tested. Fourth, it is not limited to predictive accuracy but suggests an action-focused customer portfolio perspective, where the revenue concentration, repeat-purchase probability, and retention treatment can be evaluated together.

The remainder of the paper will be structured in the following way. Section 2 presents the new literature and positions the study amongst at least fifteen new empirical and review articles. Section 3 will outline the research propositions and conceptual model. Section 4 presents the public retail dataset, feature engineering, segmentation reasoning and predictive design. The empirical findings are reported in section 5. Section 6 formulates managerial implications of SME retention and revenue increment. The theoretical contribution is described in section 7. Section 8 addresses the study limitations and future studies. Section 9 concludes.

2. Background and Related Studies

bfseries2.1. SME customer retention as an analytics problem

The recent SME literature increasingly treats customer retention as an analytics-enabled strategic capability rather than a purely relational or promotional concern. Babalghaith et al. [2] show that analytics adoption is associated with stronger financial, market, and business-process performance in SMEs, suggesting that the value of data capabilities extends well beyond reporting. In a broader review of SME digital transformation, Sagala et al. [3] identify information systems, organizational readiness, and financial discipline as core success factors, emphasizing that small firms require structured digital mechanisms rather than ad hoc digital tools. Similarly, Trinh et al. [4] argue that analytics in SMEs often depends on boundary-spanning managerial roles that translate operational data into decision-relevant formats. Taken together, these studies imply that SME customer retention is not only a marketing execution issue but also an information-processing challenge.

A related stream emphasizes customer-facing digital capability. Alkhasoneh et al. [5] find that social media use among SMEs contributes to brand awareness and customer engagement, while Seo et al. [6] show that customer acquisition and customer retention strategies both support SME growth in B2B supplier markets. These findings are important because they shift attention from isolated digital tools to the broader question of how customer-facing capabilities become embedded in growth processes. Still, the mechanisms in these studies remain largely survey-based and do not demonstrate how transaction records themselves can be transformed into customer-level retention intelligence.

bfseries2.2. Recent work on purchase, repurchase, and churn prediction

The literature on purchase and churn prediction has developed rapidly. Chou et al. [7] compare buy-till-you-die modeling with machine learning and show that a stable and interpretable Lasso model can outperform recurrent neural networks for repurchase prediction. Kim et al. [9] combine RFM-based customer characteristics with browsing patterns and show that predictive performance improves when behavioral and session signals are jointly modeled. In a systematic review, Chen et al. [8] synthesize 98 studies and highlight the field's growth, methodological diversity, and the increasing managerial relevance of customer purchase prediction in B2C e-business.

Other recent studies push toward more specialized or algorithmically complex designs. Du et al. [10] propose an entropy-based approach for community e-commerce repurchase prediction, while Haddadi et al. [12] compare resampling strategies for imbalanced churn datasets across telecommunications, online retail, and banking. Poudel et al. [11] bring interpretability to churn modeling by combining tabular machine-learning models with local and global explainability, and Zaghloul et al. [15] integrate RFM analysis, K-means, and deep neural networks to achieve very high churn-prediction performance on e-commerce datasets. Boozary et al. [17] further show that ensemble methods such as XGBoost and Random Forest can outperform classical classifiers in churn settings. These studies are analytically valuable, but many either assume richer data than SMEs usually possess, prioritize algorithmic novelty, or treat retention as a pure classification task rather than a customer-portfolio management problem.

bfseries2.3. Customer value and relationship analytics

Another branch of the recent literature focuses on customer value, relationship marketing, and lifetime value. Dogan et al. [13] review business-analytics approaches to customer lifetime value (CLV) and show that current research increasingly combines probabilistic, machine-learning, and hybrid models. Wong et al. [16] compare probability-based and machine-learning approaches to customer lifetime value prediction and find that the simpler probabilistic approach can sometimes outperform machine learning when structural assumptions hold. On the strategic side, Roy et al. [14] argue that AI-capable relationship marketing can strengthen customer relationships when organizations possess the capabilities to deploy

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026

analytics adaptively. These studies underline a recurring theme: value-based customer management depends on both predictive performance and interpretability.

Still, a practical disconnect remains. CLV and relationship-marketing studies often speak to long-horizon strategic value, whereas repeat-purchase and churn studies tend to focus on short-horizon classification. SMEs require an intermediate managerial format that translates transaction behavior into *near-term retention prioritization*. That implies combining segmentation (who matters now) with prediction (who is likely to buy next) in a way that supports action rather than merely descriptive reporting.

bfseries2.4. Synthesis of the gap

The review reveals four unresolved issues. First, SME-focused studies largely explain *whether* digital or analytics capabilities matter, but seldom operationalize a customer-level retention pipeline using public, reproducible data [2–4]. Second, purchase and churn prediction studies provide strong modeling insights, yet many require platform data, clickstream data, or sector-specific structures that are difficult for SMEs to replicate [9, 10, 15]. Third, the retention literature often separates customer value from repeat-purchase likelihood instead of treating them jointly [13, 16]. Fourth, highly accurate machine-learning studies frequently provide less guidance on how managers should prioritize customer groups under resource constraints [12, 17].

The present study addresses these gaps by combining segmentation and predictive modeling in a single retail-analytics design, keeping the data and workflow publicly reproducible, and emphasizing managerial interpretation suitable for SMEs.

Table 1: Comparison of recent studies relevant to SME customer retention analytics

Study	Research focus	Data context	Method	Main contribution	Remaining gap relative to this study
Babalghaith and Aljarallah (2024)	SME analytics adoption	233 Saudi SMEs	PLS-SEM	BDA adoption strongly relates to financial, market, and process performance	Focuses on adoption and performance, not transaction-level retention
Sagala and Ori (2024)	SME digital transformation	305-paper review	SLR and thematic analysis	Digital transformation success depends on organizational, IS, and financial factors	Does not operationalize customer-level analytics
Trinh (2024)	SME analytics capability	Vietnamese SMEs	Qualitative/managerial analysis	Management accountants can act as analytics translators in SMEs	No predictive customer model
Alkhasoneh et al. (2025)	SME social media use	290 SMEs in Jordan	UTAUT2 + PLS	Social media usage improves brand awareness and customer engagement	Engagement outcomes are not tied to transaction records
Seo and Lee (2025)	SME customer strategy	279 Korean manufacturing SMEs	SEM	Acquisition and retention strategies both support firm growth	No customer-level behavioral scoring
Chou et al. (2022)	Repurchase prediction	Large online retail data	BG/BB + Lasso + ML	Interpretable models can outperform recurrent architectures	Not focused on SME decision workflows
Chen et al. (2024)	Purchase prediction review	98 studies	Bibliometric + thematic review	Research is growing fast, with strong emphasis on prediction pipelines	No integrated SME action logic
Kim et al. (2024)	Online purchase prediction	1.19M sessions	RFM + graph metrics	Combining transaction and browsing behavior improves prediction	Requires rich clickstream data unavailable to many SMEs
Du and Chen (2024)	Community e-commerce repurchase	Consumer behavior logs	Entropy-based prediction	Information search behavior improves repurchase prediction	Platform-specific and less suitable for simple SME deployment
Poudel et al. (2024)	Explainable churn prediction	Telecom churn data	Tabular ML + SHAP	Interpretability strengthens decision usefulness	Sector-specific and subscription-oriented
Haddadi et al. (2024)	Imbalanced churn prediction	Telecom, online retail, banking	Resampling + ML comparison	Resampling choices materially affect churn performance	Focuses on imbalance handling more than customer value
Dogan et al. (2025)	Customer lifetime value analytics	Cross-domain review	Overview analysis	CLV research increasingly blends probability models with business analytics	Review does not offer a retention action matrix
Roy et al. (2025)	AI-capable relationship marketing	Theory-led business study	Conceptual + dynamic capabilities	AI can strengthen customer relationships when tied to capabilities	Lacks open transaction-level validation
Zaghloul et al. (2025)	Online retail churn	Olist, Instacart, Online Retail	RFM + K-means + LSTM/GRU	Hybrid deep learning can achieve very high churn accuracy	Less emphasis on managerial simplicity and SME usability
Wong et al. (2025)	Customer lifetime value	Health service portfolio	Probability vs. ML	Probability models can outperform ML when structural assumptions hold	Not aimed at non-contractual SME retail
Boozary et al. (2025)	Customer retention with ML	Comparative churn modeling	Classical vs ensemble ML	Ensemble models delivered strongest accuracy metrics	Does not combine segmentation and retention prioritization

3. Research Propositions and Conceptual Model

The logic of this study is intentionally framed as a set of propositions rather than formal hypotheses tied to latent survey constructs. The empirical design relies on public transactional data and seeks to establish a decision-oriented customer analytics workflow for SMEs.

Proposition 1. Customers with stronger recency, frequency, monetary, and behavioral profiles are more likely to make repeat purchases within the next 90 days.

Proposition 2. Customer segments differ materially in both revenue contribution and future purchase propensity, which makes uniform retention treatment inefficient.

Proposition 3. Explainable customer analytics can provide practical retention guidance for SMEs, even when slightly more accurate black-box alternatives are available.

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026

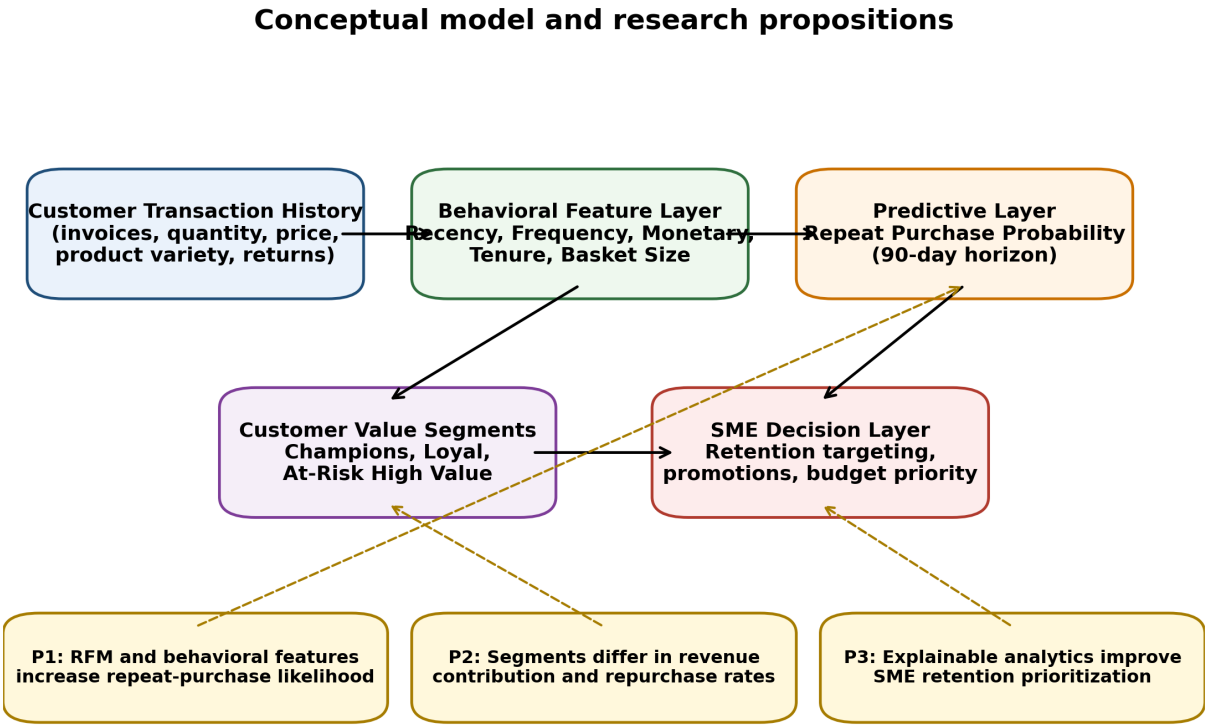

Figure 1: Conceptual model and research propositions

Figure 1 presents the conceptual model. The model begins with customer transaction history, transforms it into a behavioral feature layer, and then feeds two outputs: customer value segments and repeat-purchase probabilities. Those outputs jointly inform the SME decision layer.

4. Retail Transaction Data and Analytical Procedure

bfseries4.1. Data source and business context

The empirical setting is based on the public *Online Retail* transaction dataset [1]. The source describes the data as all transactions occurring between 1 December 2010 and 9 December 2011 for a UK-based non-store online retailer that mainly sells unique all-occasion gifts and serves many wholesale customers. Because the dataset is public, the full workflow can be reproduced without access to proprietary systems.

After cleaning, the study retained 406,789 line items with non-null customer identifiers, non-zero quantities, and positive prices or valid return records. Of these, 397,884 lines represented purchases and 8,905 represented returns, spanning 4,371 identifiable customers, 18,532 purchase invoices, and 37 countries. Although the true size class of the firm is not explicitly stated by the source, the business context is highly relevant to SME retention problems because it reflects a single digital retailer operating with invoice-level transactional data rather than enterprise-scale customer data warehouses.

bfseries4.2. Feature engineering and retention target

The analytical design follows a rolling customer-snapshot logic rather than a single static customer table. Seven monthly observation windows were constructed from March 2011 to September 2011. For each month-end, customer behavior was summarized over the prior 180 days. The dependent variable was whether the customer made at least one purchase within the next 90 days.

The feature set was intentionally chosen to be *SME-feasible*. It includes:

- recency (days since last purchase),
- frequency (orders in the recent window),
- monetary value (recent revenue),
- average basket value,
- average items per order,
- product variety,

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026

- tenure,
- total orders and total spend before the cutoff,
- return transaction ratio,
- a simple country indicator (United Kingdom vs. non-United Kingdom).

These variables do not require website tracking, ad-platform logs, CRM subscription data, or third-party enrichment. That keeps the design closer to what many SMEs can plausibly implement.

bfseries4.3. Descriptive overview

Table 2 reports descriptive statistics for the final observation window used for customer segmentation. The customer base is heterogeneous. For example, recent monetary value ranges from less than 3 monetary units to 134,880, while product variety ranges from 1 to 3,771 distinct products. Such dispersion suggests that treating all customers uniformly would obscure meaningful differences in both value and retention potential.

Table 2: Descriptive statistics for the final customer observation window

Variable	Mean	SD	Min	Median	Max
recency days	57.33	51.15	0.00	42.00	179.00
frequency	2.86	4.19	1.00	2.00	93.00
monetary	1353.47	5203.95	2.90	542.20	134880.00
avg basket value	408.58	638.43	2.90	304.80	21535.90
avg items per order	247.06	382.76	1.00	166.00	9014.00
product variety	57.18	118.63	1.00	31.00	3771.00
tenure days	110.26	54.41	0.00	126.00	179.00
return txn ratio	0.18	0.37	0.00	0.00	5.00

bfseries4.4. Customer segmentation procedure

Customer value segmentation was built from recency, frequency, and monetary behavior. Customers in the final observation window were scored and grouped into six interpretable segments: *Champions*, *Loyal*, *Promising*, *Needs Attention*, *At-Risk High Value*, and *Low Value*. The segmentation logic was designed for managerial readability rather than cluster opacity. This choice is consistent with the paper's emphasis on actionability over technical novelty.

bfseries4.5. Predictive models and evaluation

Three commonly used classification models were estimated:

1. Logistic Regression,
2. Random Forest,
3. Gradient Boosting.

The training sample included customer snapshots through July 2011, while August and September 2011 snapshots formed the chronological test set. This split preserves temporal ordering and avoids leakage from future purchase activity. Model performance was assessed with ROC-AUC, PR-AUC, accuracy, precision, recall, and F1 score. In addition, a separate logit model was estimated using the same feature family to examine statistical significance and directionality for the most important explanatory variables.

5. Findings

bfseries5.1. Customer portfolio segmentation

The segmentation results already reveal why retention should be selective. Figure 2 shows the distribution of customers across the six segments. Although the portfolio appears numerically broad, the value structure is highly concentrated. Figure 3 shows that the *Champions* segment contributes the majority of recent revenue.

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026

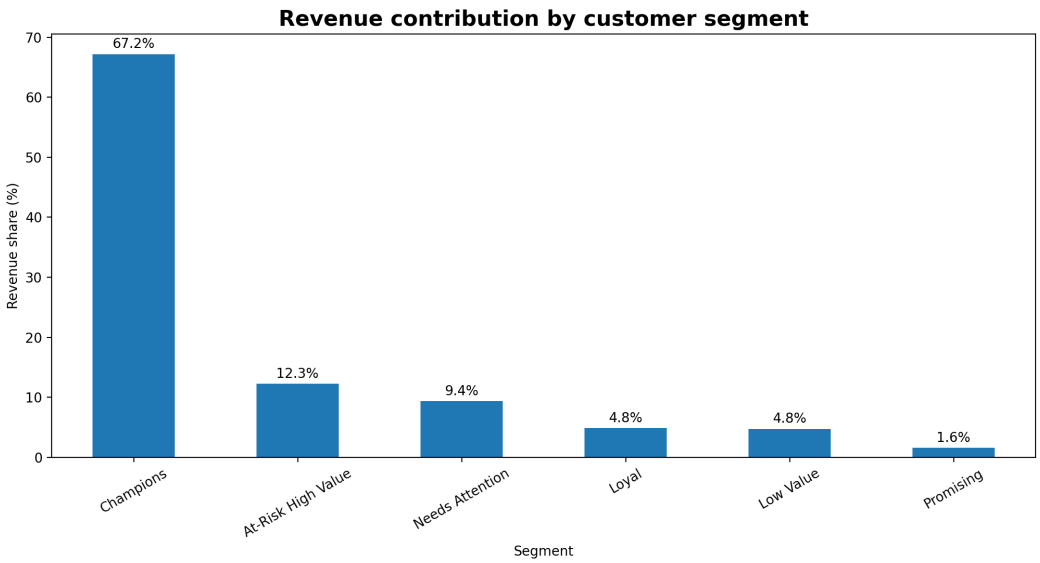

Figure 3: Revenue contribution by customer segment

Table 3 quantifies these differences. Champions account for 800 customers and 67.2% of recent revenue, with an 81.0% repeat-purchase rate. In contrast, Low Value customers represent 777 customers but only 4.8% of revenue and a 37.2% repeat-purchase rate. At-Risk High Value customers are especially important from a managerial standpoint: they contribute 12.3% of recent revenue yet show markedly weaker recency than Champions. This profile makes them prime candidates for targeted reactivation.

Table 3: Segment-level customer value and repeat-purchase patterns

Segment	Customers	Mean revenue	Repeat rate (%)	Avg. recency	Avg. frequency	Revenue share (%)
Champions	800	3,311.59	81.0	14.6	6.10	67.2
At-Risk High Value	313	1,544.73	66.1	73.9	3.22	12.3
Needs Attention	466	791.98	49.6	76.8	1.72	9.4
Loyal	299	638.49	54.5	15.4	1.88	4.8
Low Value	777	241.27	37.2	112.7	1.04	4.8
Promising	258	242.06	44.2	16.6	1.03	1.6

These results support Proposition 2. Customer segments are not only different in descriptive terms; they differ materially

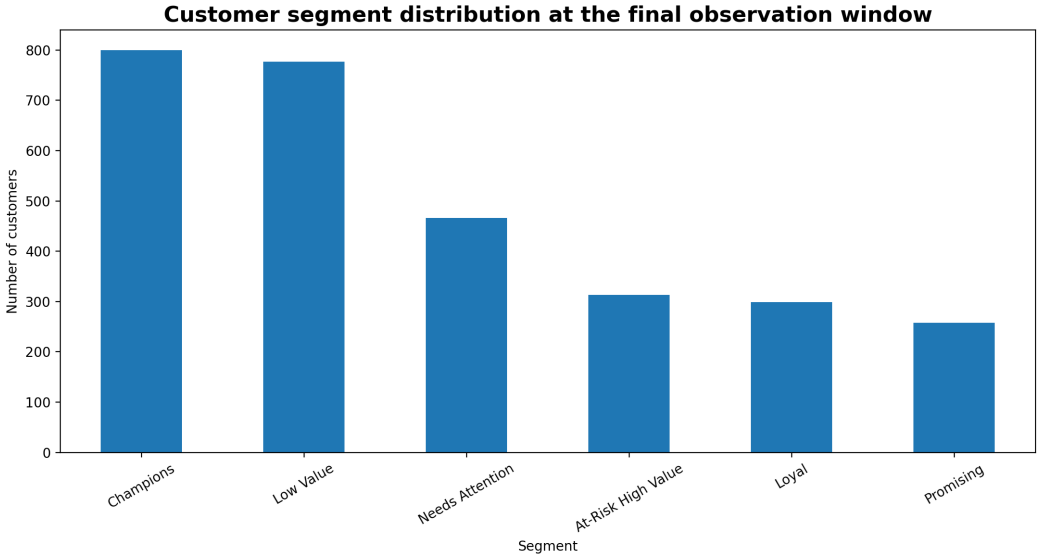

Figure 2: Customer segment distribution at the final observation window

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026

in economic relevance and in their likelihood of buying again. For SMEs, this matters because retention budgets can be focused where financial upside is highest rather than spread thinly across the portfolio.

bfseries5.2. Repeat-purchase prediction

Table 4 reports predictive performance. The random forest model achieved the strongest discrimination with ROC-AUC = 0.750 and PR-AUC = 0.821. Gradient boosting was close behind, while logistic regression trailed slightly at ROC-AUC = 0.740 and PR-AUC = 0.816. Importantly, the gap between the best-performing model and the interpretable baseline was modest.

Table 4: Out-of-sample performance for repeat-purchase prediction

Model	ROC-AUC	PR-AUC	Accuracy	Precision	Recall	F1
Random Forest	0.750	0.821	0.675	0.752	0.665	0.706
Gradient Boosting	0.746	0.817	0.672	0.734	0.691	0.712
Logistic Regression	0.740	0.816	0.672	0.739	0.681	0.709

Figure 4 visualizes the ROC curves for the three models. The performance pattern suggests that non-linear methods capture some additional signal, but not enough to make interpretable methods irrelevant.

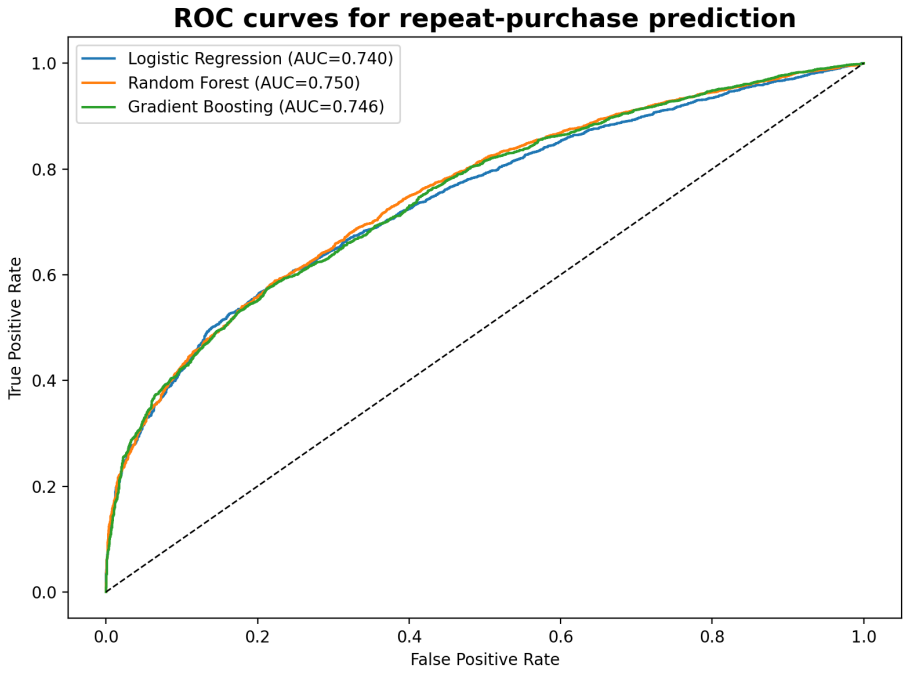

Figure 4: ROC curves for repeat-purchase prediction

Figure 5 shows the confusion matrix for the best-performing model (Random Forest). The model correctly classified 1,622 non-repeaters and 2,218 repeat purchasers in the test set, with fewer false negatives than false positives. For retention planning, this profile is operationally acceptable because missing a likely repeater can be more costly than reviewing a false positive when promotional budgets are limited but not negligible.

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026

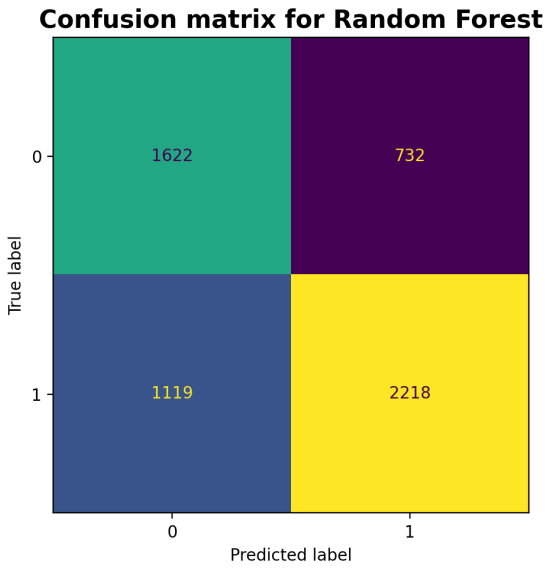

Figure 5: Confusion matrix for the best-performing repeat-purchase model

The prediction results support Proposition 1. Transaction-derived behavioral features provide useful forward-looking information about repurchase. The discriminatory performance is not extreme, which is consistent with real managerial environments where customer decisions are only partially observable, but it is sufficiently strong to improve prioritization over intuition or descriptive reporting.

bfseries5.3. Explainability and proposition assessment

To assess Proposition 3, the study examined coefficient-level interpretability alongside model accuracy. Figure 6 plots the strongest logistic-regression coefficients. Table 5 reports the statistically most informative features from the logit specification.

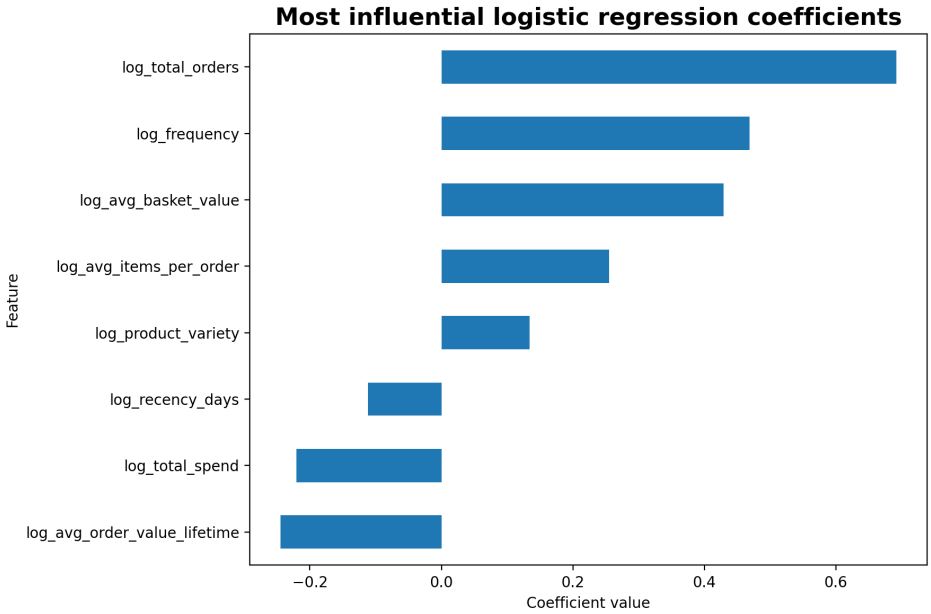

Figure 6: Most influential logistic-regression coefficients

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026

Table 5: Most informative logit coefficients for repeat-purchase likelihood

Feature	Coefficient	p-value
log avg items per order	0.263	0.0000
log product variety	0.121	0.0000
log recency days	-0.105	0.0018
return txn ratio	0.129	0.0094
uk customer	0.104	0.1332
log frequency	1.733	0.2222
log avg basket value	1.151	0.2345
log tenure days	-0.038	0.3593

Three points are notable. First, recency enters with the expected negative sign: as the time since last purchase increases, repeat-purchase likelihood declines. Second, average items per order and product variety are both positive and statistically significant, indicating that broader and deeper shopping baskets are associated with stronger repeat behavior. Third, return behavior is positive in the fitted model. This may appear counterintuitive, but in retail contexts it can signal customer engagement rather than dissatisfaction alone; active customers naturally have more opportunities both to purchase and to return.

Although the random forest model achieved slightly stronger predictive power, logistic regression remained competitive. This matters because SMEs often need transparent models they can explain in meetings, embed in spreadsheet-like dashboards, and trust when deciding whom to target. In that sense, Proposition 3 is supported in a *practical* rather than absolute-performance sense. Explainable analytics did not outperform the best black-box model here, but they delivered nearly equivalent predictive quality together with superior interpretability.

6. Managerial Insights for SME Retention and Revenue Growth

The findings suggest that SME retention strategy should be organized around a portfolio mindset. The first managerial implication is *do not retain everyone equally*. The segment results show that a relatively narrow set of customers accounts for most of the recent revenue. For an SME, this means that broad loyalty campaigns can be inefficient because they dilute scarce marketing resources. Champions and At-Risk High Value customers warrant more attention than Promising or Low Value groups, even if those larger groups appear numerically attractive.

The second implication concerns *retention sequencing*. A useful SME order of action is as follows. First, protect Champions through continuity-oriented actions such as stock reliability, preferred communication, and relationship maintenance. Second, reactivate At-Risk High Value customers through targeted offers, reminders, or service recovery because the upside is large and the customer already has a proven spending history. Third, develop Loyal and Promising customers through cross-sell and basket-expansion tactics. Finally, handle Low Value customers with low-cost automated interventions rather than expensive manual treatment.

The third implication is that *prediction alone is insufficient*. A customer with high repurchase probability but low economic value should not receive the same treatment as a customer with both high value and moderate attrition risk. Figure 7 illustrates this logic with an action matrix that jointly positions segments by value and repeat-purchase likelihood. The matrix allows SME managers to interpret analytics outputs in a language that matches budgeting and targeting decisions.

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026

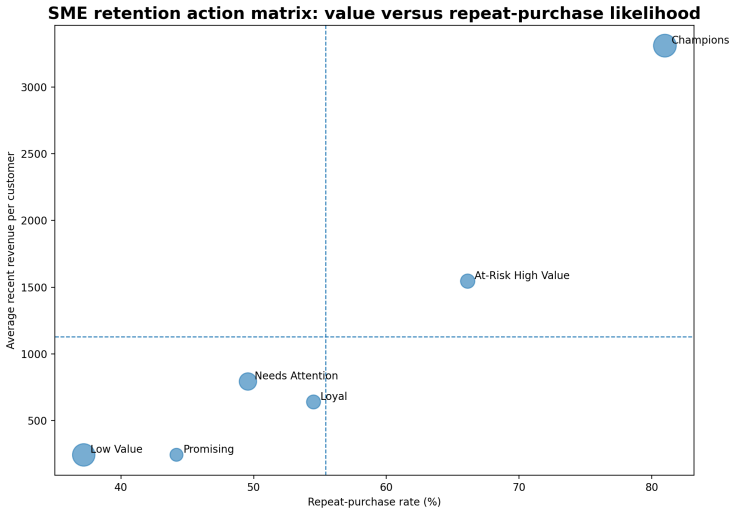

Figure 7: SME retention action matrix: value versus repeat-purchase likelihood

The fourth implication concerns *analytical simplicity*. Many SMEs do not need highly elaborate customer-intelligence infrastructures to obtain useful retention guidance. The present design uses only invoice-level data and still produces meaningful segmentation and prediction results. This lowers the barrier to adoption. A small retailer can implement a version of this workflow in a relational database, spreadsheet pipeline, business-intelligence dashboard, or cloud notebook without building a full enterprise AI stack.

The fifth implication concerns *returns and product breadth*. The models suggest that deeper baskets and broader product variety are positively associated with future purchasing. This points to practical interventions such as assortment-based recommendations, replenishment reminders, and personalized bundles. At the same time, the return ratio result should prompt caution. Managers should not interpret returns uniformly as negative; rather, they should distinguish between high-value engaged customers with occasional returns and structurally unprofitable customers who generate costly reverse-logistics patterns.

7. Theoretical Implications and Contribution

The paper contributes to business intelligence and business data analysis in three ways. First, it links SME analytics capability research with transaction-level customer modeling. Existing SME studies tend to analyze adoption drivers, digital transformation, or broad performance effects, whereas this study shows how a manageable set of behavioral indicators can be operationalized into retention intelligence [2–4].

Second, it bridges two often separate analytical traditions: customer value segmentation and repeat-purchase prediction. Much of the literature emphasizes one or the other. By placing both inside one empirical design, the study argues that customer analytics for SMEs should be evaluated not only by predictive performance but also by its ability to support differentiated treatment across the customer portfolio.

Third, the paper contributes to the interpretability debate. Recent retention studies often demonstrate strong results with ensembles or deep learning [15, 17], while review work highlights the growing methodological diversity of the field [8, 13]. The present findings suggest a more nuanced view: for SME deployment, a slightly less accurate but more transparent model may be strategically preferable when the interpretability gains improve adoption, communication, and implementation quality.

8. Study Boundaries and Future Research

The study has several limitations. First, although the dataset reflects a single digital retailer and is highly relevant to SME-style decision problems, the true size class of the focal business is not explicitly documented by the public source. The paper therefore makes an *SME-relevance* claim rather than a verified SME-identification claim. Second, the data contain transactions but not marketing exposure, browsing depth, customer service interactions, or promotional histories. As a result, the models do not capture all of the mechanisms that may influence repurchase.

Third, the retention target is defined as repeat purchase within 90 days. That horizon is managerially reasonable for tactical interventions, but different retail categories may require alternative windows. Fourth, the segmentation logic is deliberately interpretable and managerial; a more complex clustering or latent-state approach might reveal finer behavioral granularity. Fifth, the analysis does not estimate formal customer lifetime value, profitability net of service costs, or campaign response elasticity.

These limitations create a useful agenda for future work. Future studies could integrate promotion records, customer

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026

service logs, or clickstream data where available; compare alternative retention windows; add uplift modeling to estimate intervention effects; or combine customer-level predictions with inventory and margin information. Another promising direction would be cross-sector replication in food retail, specialty e-commerce, and B2B distribution environments where SME customer structures differ substantially.

9. Conclusion

The current paper created a customer analytics architecture of SMEs based on transaction data to segment customers and predict repeat-purchase with the use of publicly available retail data. The structure is actually not like aggregate forecasting studies, which focus on managerial action on a customer level. The findings indicate that customer value is very concentrated and that behavioral variables derived through transactions can be used to predict repeat purchase with a useful level of accuracy. Although random forest model provided the best predictive discrimination, the logistic regression was not far behind and provided better understanding of management.

The bigger picture here is that SMEs do not require the scale of the enterprise data infrastructures to start practicing meaningful customer analytics. Better targeting, budget prioritization, and revenue protection can already be facilitated by a retention system based on recency, frequency, monetary behavior, basket characteristics, and product variety. The study, by integrating segmentation and prediction in a reproducible workflow, provides a useful business-data-analysis contribution that can be applied in practice, transparently, and adapted further to suit the context of SMEs.

References

[1] UCI Machine Learning Repository (2015). Online Retail. Available at: https://archive.ics.uci.edu/dataset/352/online+retail

[2] Babalghaith, R. and Aljarallah, A. (2024). Factors Affecting Big Data Analytics Adoption in Small and Medium Enterprises. *Information Systems Frontiers*, 26(6), 2165–2187. https://doi.org/10.1007/s10796-024-10538-2

[3] Sagala, G. H. and Ori, D. (2024). Toward SMEs Digital Transformation Success: A Systematic Literature Review. *Information Systems and e-Business Management*, 22(4), 667–719. https://doi.org/10.1007/s10257-024-00682-2

[4] Trinh, H. T. (2024). An SME Approach to Data Analytics by Management Accountants in the Transition Economy of Vietnam. *Journal of Science and Technology Policy Management*. https://doi.org/10.1108/JSTPM-12-2023-0222

[5] Alkhasoneh, O. M., Jamaludin, H., Bin Zahar, A. R. I., and Al-Sharafi, M. A. (2025). Drivers of Social Media Use Among SMEs and Its Impact on Brand Awareness and Customer Engagement. *Asia-Pacific Journal of Business Administration*, 17(3), 595–615. https://doi.org/10.1108/APJBA-02-2024-0102

[6] Seo, E. and Lee, E. (2025). Linking SMEs' Customer Strategy to Firm Growth: The Case of Manufacturing Suppliers in South Korea. *Asia Pacific Journal of Marketing and Logistics*, 37(3), 782–799. https://doi.org/10.1108/APJML-03-2024-0313

[7] Chou, P., Chuang, H. H.-C., Chou, Y.-C., and Liang, T.-P. (2022). Predictive Analytics for Customer Repurchase: Interdisciplinary Integration of Buy Till You Die Modeling and Machine Learning. *European Journal of Operational Research*, 296(2), 635–651. https://doi.org/10.1016/j.ejor.2021.04.021

[8] Chen, S., Xu, Z., Xu, D., and Gou, X. (2024). Customer Purchase Prediction in B2C E-Business: A Systematic Review and Future Research Agenda. *Expert Systems with Applications*, 252, 124261. https://doi.org/10.1016/j.eswa.2024.124261

[9] Kim, S., Shin, W., and Kim, H.-W. (2024). Predicting Online Customer Purchase: The Integration of Customer Characteristics and Browsing Patterns. *Decision Support Systems*, 177, 114105. https://doi.org/10.1016/j.dss.2023.114105

[10] Du, J. and Chen, W. (2024). Enhancing Community E-Commerce Repurchase Prediction Through Information Entropy Analysis. *International Journal of e-Collaboration*, 20(1). https://doi.org/10.4018/IJeC.349897

[11] Poudel, S. S., Pokharel, S., and Timilsina, M. (2024). Explaining Customer Churn Prediction in Telecom Industry Using Tabular Machine Learning Models. *Machine Learning with Applications*, 17, 100567. https://doi.org/10.1016/j.mlwa.2024.100567

[12] Haddadi, S. J., Farshidvard, A., Silva, F. dos S., dos Reis, J. C., and Reis, M. M. (2024). Customer Churn Prediction in Imbalanced Datasets with Resampling Methods: A Comparative Study. *Expert Systems with Applications*, 246, 123086. https://doi.org/10.1016/j.eswa.2023.123086

[13] Dogan, O., Hiziroglu, A., Pisirgen, A., and Seymen, O. F. (2025). Business Analytics in Customer Lifetime Value: An Overview Analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(1), e1571. https://doi.org/10.1002/widm.1571

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026

[14] Roy, S. K., Nazaritehrani, A., Pandit, A., Apostolidis, C., and Ray, S. (2025). AI-Capable Relationship Marketing: Shaping the Future of Customer Relationships. *Journal of Business Research*, 192, 115309. https://doi.org/10.1016/j.jbusres.2025.115309

[15] Zaghloul, M., Barakat, S., and Rezk, A. (2025). Enhancing Customer Retention in Online Retail Through Churn Prediction: A Hybrid RFM, K-Means, and Deep Neural Network Approach. *Expert Systems with Applications*, 290, 128465. https://doi.org/10.1016/j.eswa.2025.128465

[16] Wong, A., Viloria Garcia, A., and Lim, Y.-W. (2025). A Data-Driven Approach to Customer Lifetime Value Prediction Using Probability and Machine Learning Models. *Decision Analytics Journal*, 16, 100601. https://doi.org/10.1016/j.dajour.2025.100601

[17] Boozary, P., Sheykhan, S., GhorbanTanhaei, H., and Magazzino, C. (2025). Enhancing Customer Retention with Machine Learning: A Comparative Analysis of Ensemble Models for Accurate Churn Prediction. *International Journal of Information Management Data Insights*, 5(1), 100331. https://doi.org/10.1016/j.jjimei.2025.100331

DOI: https://doi.org/10.54216/AJBOR.140201
Received: December 22, 2025, Revised: March 08, 2026 Accepted: April 02, 2026