Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 197–206
https://doi.org/10.54941/ahfe1008254

Customer Churn Prediction for Consumer-Facing Delivery Apps

Muhammad Jawad Mufti[1], Omar Hammad[1,3], Haitham Saleh[3,4], and Muqaddas Gull[2]

[1]Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
[2]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
[3]Industrial and Systems Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
[4]Interdisciplinary Research Center for Smart Mobility and Logistics (IRC-SML), King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

ABSTRACT

Customer disengagement in non-contractual on-demand services must be inferred from activity rather than explicit cancellation, motivating repeated risk assessment. This study proposes a rolling-window prediction and decision-support framework using 401,164 bookings from 101,765 customers of a doorstep car-wash platform. Matched 30/30-, 15/15-, and 7/7-day observation/evaluation windows generated customer states with 54 behavioral and operational features. Logistic Regression, Random Forest, XGBoost, LSTM, GRU, and Temporal ResNet were tuned and evaluated using customer-grouped partitions, with recall prioritized. LSTM achieved the highest recall at 30, 15, and 7 days (96.0%, 96.4%, and 90.9%), while the highest ROC-AUC values were 0.944, 0.951, and 0.954. Retrospective 30-day analysis identified engagement-depth-specific behavioral patterns, including service failure (65.8%), promotion dependence (52.7%), inactivity (62.0%), and subscription loss with price burden (50.2%), which informed candidate retention actions. External Instacart replication achieved 94% recall and 0.9777 ROC-AUC with 30-day XGBoost, providing preliminary evidence of cross-dataset applicability.

Keywords: Customer churn, Rolling window, Behavioral modeling, Decision support, Customer retention, Machine learning

INTRODUCTION

Customer churn is a major challenge for consumer-facing, usage-driven services because profitability depends on retaining existing customers as well as acquiring new ones. In non-contractual delivery services, customers may simply stop using a service, so disengagement must be inferred from transactional behavior rather than formal cancellation (Neslin et al., 2006; Ascarza, 2018). Machine-learning methods can identify complex churn-related patterns, but operational value depends on detecting risk early enough to support retention action (De Caigny et al., 2018; Verbraken et al., 2013).

Received September 18, 2026; Revised October 24, 2026; Accepted October 30, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

Two limitations remain prominent. First, static customer snapshots provide limited support for reassessing risk as behavior changes over time (De Caigny et al., 2020; Lalwani et al., 2022). Temporal studies increasingly use recurrent or sliding-window representations, yet observation periods and future inactivity horizons are not always specified within a single reproducible prediction process (Mena et al., 2024; Ahlstrand et al., 2025). Second, churn scores alone provide limited guidance about the behavioral patterns associated with risk or the actions that may be appropriate; retention support is often generic rather than relationship-stage specific (Gattermann-Itschert and Thonemann, 2022; Sikri et al., 2024).

This study develops a rolling-window prediction and decision-support framework for a doorstep car-wash platform, enabling repeated churn-risk assessment in non-contractual services. Beyond prediction, it links at-risk customers to engagement-depth-specific behavioral patterns and retention actions. From a human-factors perspective, the DSS translates predictive outputs into interpretable cues and candidate actions for retention decision-makers. External replication on Instacart examines applicability beyond the original context.

This study makes four contributions: (1) an explicit rolling-window formulation that captures evolving customer behavior while separating observed behavior from future inactivity; (2) comparison of feature-based and sequence-based models across 30-, 15-, and 7-day horizons; (3) an engagement-depth-aware decision-support layer that converts predicted risk into interpretable behavioral patterns and domain-informed retention actions without causal claims; and (4) external replication on Instacart to assess methodological transferability across a second transaction-based customer service context.

RELATED WORK

Customer churn prediction is commonly formulated as binary classification using aggregated customer features and models such as Logistic Regression, Random Forest, gradient boosting, and ensembles. Static representations can overlook behavioral change before disengagement, especially in non-contractual services where churn is inferred from inactivity (Lalwani et al., 2022; De Caigny et al., 2020). Recent research addresses temporality through recurrent time-varying RFM measures (Mena et al., 2024), inactivity-based sliding windows (Bugajev et al., 2025), transformer windows (Ahlstrand et al., 2025), and RFM/deep-learning hybrids (Zaghloul et al., 2025). These studies establish the value of temporal information but do not consistently integrate observation length, future horizon, and repeated customer-state generation into one explicit rolling process.

A parallel literature extends prediction toward retention decision support. Churn-risk estimates have been used to trigger interventions (van Leeuwen et al., 2025), while prediction has been combined with segmentation and recommendation (Jahan and Sanam, 2026). Explainability methods such as SHAP improve transparency (Peng et al., 2023; El Attar and El-Hajj, 2026), but explanations and recommendations are often global, segment-level, or

weakly linked to engagement-depth-specific behavioral evidence. The present study integrates repeated rolling-window assessment with observation-window engagement-depth segmentation, candidate behavioral risk-pattern assignment, and targeted action mapping, while testing transferability on a second transaction-based dataset.

METHODOLOGY

Data and Preprocessing

The primary dataset (October 2024–January 2025) contains 401,164 bookings from 101,765 doorstep car-wash customers with transactional, operational, temporal, subscription, monetary, and spatial attributes (Figure 1). Preprocessing corrected data types, encoded categorical variables, handled structural missingness, derived success-dependent attributes from completed bookings, represented windows without successful bookings, and discretized Haversine distances into ordinal groups. Numeric scaling was fitted on training data only.

Rolling-Window Churn Formulation

Risk was modeled at the customer-window level (Figure 2). For customer i on reference day t, behavior in $[t-W+1, t]$ predicts activity in $[t+1,t+W]$ as defined in equation 1. At least one successful booking denotes activity; otherwise, the instance is inactive. The 30/30-, 15/15-, and 7/7-day windows advance daily using only observation-window features, with identical consecutive summaries collapsed to reduce redundancy. Thus, 7- and 15-day targets represent short-horizon inactivity, whereas 30 days reflects longer disengagement.

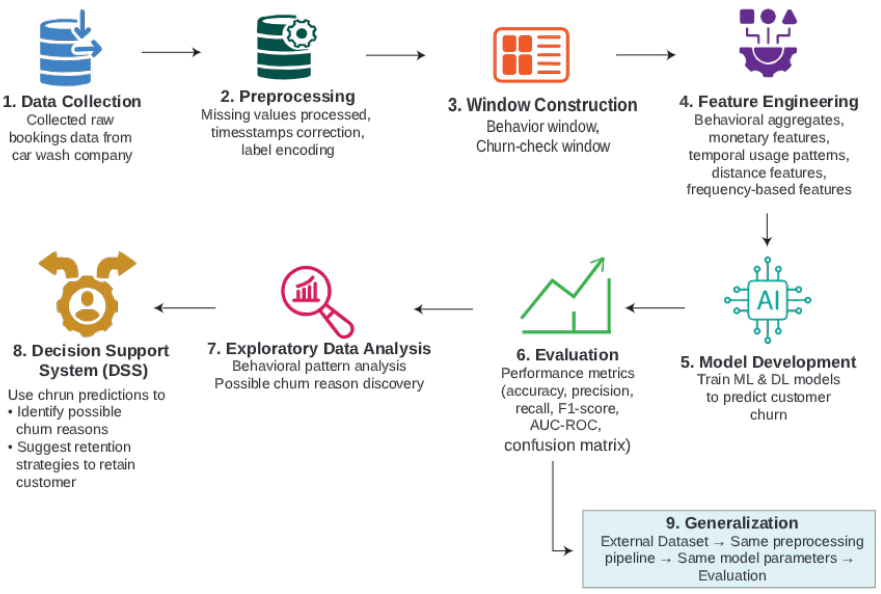


Figure 1: Proposed rolling-window churn prediction and decision-support framework.



Figure 2: Rolling observation and future churn-evaluation window formulation.

$$y_{i,t} = \begin{cases} 1, \\ 0, \end{cases} \text{if no qualifying activity occurs during} [t+1, t+W_c], \text{otherwise.} \quad (1)$$

Feature Engineering and Predictive Modeling

Each observation window contains 54 features covering booking activity/outcomes, recency, monetary/payment behavior, subscription activity, spatial/service characteristics, and categorical preferences; identifiers and window-boundary fields were excluded. Models include Logistic Regression (Hosmer et al., 2013), Random Forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), LSTM (Krichen and Mihoub, 2025), GRU (Cho et al., 2014), and Temporal ResNet (Koh et al., 2021; Mohammadi Foumani et al., 2024). Hyperparameters were selected using 5-fold GridSearchCV on the training data, with recall as the scoring metric. LSTM, GRU, and Temporal ResNet used sequence lengths of 10 windows respectively, with training-median left-padding for shorter histories. The 30-day EDA informs the DSS rules. Class distributions were near balanced at 30 and 15 days, whereas the 7-day dataset was more skewed as shown in Table 2. Accordingly, the 7-day training pipeline incorporated SMOTE (k = 5) after customer splitting to balance the training classes approximately 1:1; validation and test sets remained unchanged (de Vargas et al., 2023).

Evaluation Protocol

For each horizon, one customer-grouped 70/15/15 train/validation/test split was shared across all models. Sequence construction followed splitting, and scaling and balancing were fitted only on training data. Inactivity was the positive class, and precision, recall, and F1 used binary averaging. Test predictions used an a priori 0.50 probability threshold across all models and external replication. The threshold was not optimized and served only to standardize comparison; deployment should select thresholds on validation data according to intervention capacity and retention costs. ROC-AUC was additionally reported to provide a threshold-independent comparison of model discrimination.

Decision-Support Framework

The DSS analyzes at-risk instances by observation-window engagement depth, defined by the number of successful bookings within the current observation window: 1, 2, 3, or ≥4. EDA and rule derivation used observed 30-day labels and were conducted separately from predictive model fitting; the resulting priority-based rules are summarized in Table 1. Lower-engagement-depth instances are assessed for service failure, promotion dependence, and inactivity,

while higher-engagement-depth instances are assessed for subscription loss, price burden, and discount dependence. These retrospectively derived rules are intended for exploratory decision support rather than validated causal attribution. Unmatched cases remain unclassified, and intervention effectiveness is not tested. The rule cutoffs were heuristic thresholds used to operationalize the retrospective EDA patterns and were not optimized for predictive or intervention performance.

Table 1: Engagement-depth-specific rules for assigning candidate behavioral risk patterns.

Stage	Priority-Based Candidate Risk Patterns and Conditions
1 booking	(1) Subscription-service failure: failed_booking_count > 0 and subscription_booking_count > 0; (2) booking/service failure: failed_booking_count > 0 and subscription_booking_count = 0; (3) promotion dependence: failed_booking_count = 0, promo_used = 1, and promo_code_discount_type = 1; (4) inactivity: failed_booking_count = 0, promo_used = 0, subscription_booking_count > 0, and days_since_last_booking ≥ 10.
2 bookings	(1) Subscription-service failure: failed_booking_count > 0 and subscription_booking_count > 0; (2) promotion dependence: promo_used = 1 and promo_code_discount_type = 0; (3) booking/service failure: failed_booking_count > 0, subscription_booking_count = 0, and promo_used = 0; (4) inactivity: failed_booking_count = 0, promo_used = 0, subscription_booking_count > 0, and days_since_last_booking ≥ 5.
3 bookings	Same as 2 bookings, except inactivity requires days_since_last_booking ≥ 4.
≥4 bookings	(1) Subscription loss + price burden: has_subscription_booking = 1, last_booking_is_subscription = 0, and avg_real_total_non_sub ≥ 1.10 × avg_real_total_sub; (2) discount-supported subscription loss: same subscription-loss conditions, no ≥10% price burden, and pct_discounted_bookings > 0; (3) subscription loss without price burden: same subscription-loss conditions, no ≥10% price burden, and pct_discounted_bookings = 0; (4) discount-dependent non-subscription behavior: has_subscription_booking = 0 and pct_discounted_bookings > 0.
All depths	Unclassified when no higher-priority rule is satisfied.

External Methodological Replication

External replication used a four-month Instacart subset (102,127 transactions; 11 source features) (Zaghloul et al., 2025). The 30/30-day formulation produced 96,790 customer-window instances with 25 behavioral predictors; models were retrained using the same evaluation protocol and fixed 0.50 threshold.

RESULTS AND DISCUSSION

Datasets and Experimental Setup

Rolling windows generated 280,756, 355,099, and 346,615 customer-window instances at 30-, 15-, and 7-day horizons, respectively, each with 54 features. The inactive/active distributions were 39.5%/60.5%

at 30 days, 48.7%/51.3% at 15 days, and 66.9%/33.1% at 7 days, indicating greater class imbalance at the shortest horizon. All six models were hyperparameter-tuned using recall as the selection metric because false negatives represent missed retention opportunities. To address the stronger 7-day imbalance, SMOTE was applied only to the 7-day training data.

Churn Prediction Performance

Table 2 reports held-out performance at the fixed 0.50 threshold. LSTM achieved the highest observed recall at 30, 15, and 7 days (96.0%, 96.4%, and 90.9%). At 7 days, LSTM and GRU shared the highest F1-score (91.9%), while GRU achieved the highest ROC-AUC (0.954); XGBoost led the classical models in ROC-AUC at every horizon. On Instacart, XGBoost achieved 94% recall and 0.9777 ROC-AUC, while LSTM and GRU reached 91.3% and 92.0% recall, respectively, providing preliminary evidence of cross-dataset applicability.

Table 2: Hyperparameter-tuned test performance of the six models at 30-, 15-, and 7-day horizons (SMOTE-balanced training at 7 days), fixed 0.50 threshold. F1 is the harmonic mean of the tabulated precision and recall.

Horizon	Model	Test Acc (%)	Precision (%)	Recall (%)	F1 (%)	AUC
30-day	XGBoost	89.7	87.0	86.9	86.9	0.944
30-day	Logistic Regression	87.9	84.8	84.5	84.6	0.916
30-day	Random Forest	88.5	85.5	85.3	85.4	0.929
30-day	LSTM	94.2	90.0	96.0	92.9	0.944
30-day	GRU	93.7	89.9	94.3	92.0	0.944
30-day	Temporal ResNet	93.7	89.9	94.6	92.2	0.941
15-day	XGBoost	80.8	80.3	80.2	80.3	0.889
15-day	Logistic Regression	76.1	75.5	75.5	75.5	0.837
15-day	Random Forest	78.4	77.9	77.7	77.8	0.859
15-day	**LSTM**	**94.3**	**92.2**	**96.4**	**94.3**	**0.944**
15-day	GRU	91.7	92.6	90.2	91.4	0.951
15-day	Temporal ResNet	91.7	91.5	91.5	91.5	0.939
7-day (SMOTE)	XGBoost	68.3	76.0	77.0	76.5	0.836
7-day (SMOTE)	Logistic Regression	72.5	82.0	75.5	78.6	0.786
7-day (SMOTE)	Random Forest	67.9	76.0	76.0	76.0	0.832
7-day (SMOTE)	LSTM	89.3	93.0	90.9	91.9	0.932
7-day (SMOTE)	GRU	89.3	93.6	90.2	91.9	0.954
7-day (SMOTE)	Temporal ResNet	88.7	92.5	90.5	91.5	0.943

Behavioral Risk Patterns and Retention Support

Exploratory analysis of 30-day inactive-labeled instances showed that behavioral patterns varied with observation-window engagement depth. Churn incidence declined from 48.3% at one booking to 6.9% at ≥4 bookings. Booking/service failure dominated at one booking (65.8%), consistent with the failed-booking differences in Figure 3(a). Promotion dependence was most frequent at two bookings (52.7%), matching the promo-code usage pattern in Figure 3(b). At three bookings, inactivity/recency drift accounted for 62.0% of assigned patterns, consistent with Figure 4(a). Among ≥4-booking instances, subscription loss with price burden was most frequent (50.2%), followed by discount-dependent non-subscription behavior (32.6%) (Table 3). These patterns are retrospective associations rather than verified causes of churn.

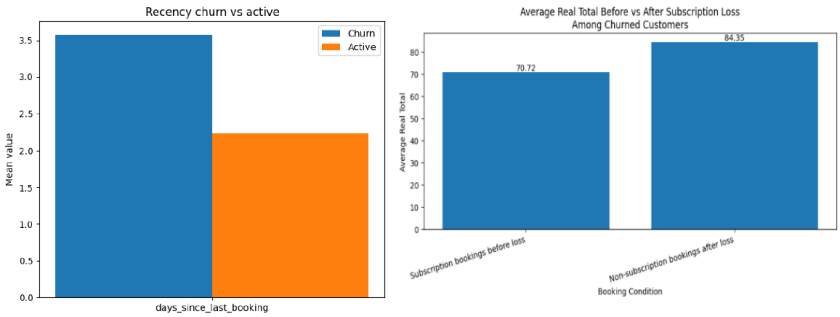


Figure 3: (a) failed-booking incidence at 1 booking; (b) promo-code usage rate at 2 bookings.

Among ≥4-booking churn-labeled instances, mean real total was SAR 67.13 versus SAR 58.78 for active instances. For churn-labeled customers who lost subscription pricing, mean real total increased from SAR 70.72 to SAR 84.35, with 93.46% paying more on average, as shown in Figure 4(b). Together with Table 3, these findings motivate differentiated candidate retention actions: recovery for service failure, package or upgrade messaging for promotion dependence, re-engagement for inactivity, and renewal or value-focused outreach for subscription-loss patterns. Unclassified cases receive generic reactivation and feedback prompts. These remain decision-support suggestions, not experimentally validated treatments.

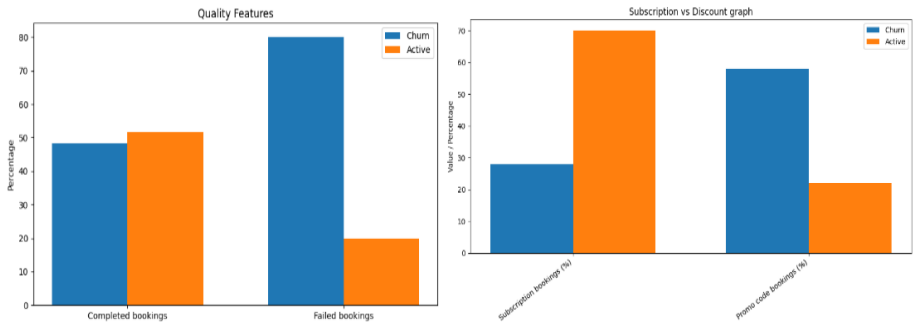


Figure 4: (a) days-since-last-booking at 3 bookings; (b) average real total before vs. after subscription loss at 4+ bookings.

Table 3: Behavioral risk-pattern shares among 30-day inactive-labeled instances by engagement depth.

Assigned Risk Pattern	1 Booking	2 Bookings	3 Bookings	4+ Bookings
Booking/service failure	65.8%	21.9%	17.1%	—
Subscription service failure	3.2%	6.7%	18.1%	—
Promotion dependence	7.6%	52.7%	—	—
Long inactivity / recency drift	22.4%	12.1%	62.0%	—
Subscription loss + price burden	—	—	—	50.2%
Discount-dependent non-subscription behavior	—	—	—	32.6%
Unclassified	1.0%	6.6%	2.8%	17.2%

DISCUSSION

The findings indicate that rolling windows support repeated assessment of customer disengagement across 30-, 15-, and 7-day horizons. At the fixed 0.50 threshold, LSTM achieved the highest recall across horizons (96.0%, 96.4%, and 90.9%), while the highest ROC-AUC values were 0.944, 0.951, and 0.954. This extends prior static churn modeling and complements temporal approaches using time-varying RFM, sliding windows, and Transformers (Lalwani et al. (2022); Mena et al., 2024; Bugajev et al., 2025; Ahlstrand et al., 2025).

Beyond prediction, the DSS links flagged risk to engagement-depth-specific behavioral patterns and candidate actions. Prior work combines churn prediction with segmentation, explainability, intervention triggering, or recommendation (Peng et al., 2023; Zaghloul et al., 2025; van Leeuwen et al., 2025; El Attar and El-Hajj, 2026; Jahan and Sanam, 2026) , whereas the proposed DSS conditions interpretation and action mapping on observation-window engagement depth. The observed patterns are retrospective associations rather than verified causes. Instacart replication achieved 94% recall and 0.9777 ROC-AUC with 30-day XGBoost, providing preliminary evidence of cross-dataset applicability rather than numerical superiority.

CONCLUSION

This study presented a rolling-window churn-risk prediction and decision-support framework for an on-demand service. Matched 30-, 15-, and 7-day windows enabled repeated inactivity assessment, with LSTM achieving the highest recall across all horizons. The DSS linked flagged risk to engagement-depth-specific candidate patterns and actions, while Instacart replication provided preliminary cross-dataset evidence. Limitations include the four-month dataset, dependence among overlapping windows, limited seasonal and pricing information, retrospective DSS-rule derivation without threshold-sensitivity analysis, no paired significance testing of model-performance differences, and no prospective evaluation of interventions or DSS usability. Future work should evaluate DSS rules and interventions prospectively and incorporate confidence intervals, paired model comparisons, usability, trust, workload, and decision quality.

ACKNOWLEDGMENT

The authors acknowledge Sweater for providing access to the real-world operational dataset used in this study.

REFERENCES

Ahlstrand, J., Borg, A., Grahn, H. and Boldt, M. (2025). "Using Transformers for B2B Contractual Churn Prediction Based on Customer Behavior Data", in: *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS)*, Vol. 1, pp. 562–571.

Ascarza, E., Neslin, S.A., Netzer, O., Anderson, Z., Fader, P.S., Gupta, S., Hardie, B.G.S., Lemmens, A., Libai, B., Neal, D., Provost, F. and Schrift, R. (2018). "In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions", *Customer Needs and Solutions*, Vol. 5 No. 1–2, pp. 65–81.

Breiman, L. (2001). "Random Forests", *Machine Learning*, Vol. 45 No. 1, pp. 5–32.

Bugajev, A., Kriauzienė, R. and Chadyšas, V. (2025). "Realistic Data Delays and Alternative Inactivity Definitions in Telecom Churn: Investigating Concept Drift Using a Sliding-Window Approach", *Applied Sciences*, Vol. 15 No. 3, Article 1599.

Chen, T. and Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System", in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014). "Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation", in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734.

Coussement, K., Lessmann, S. and Verstraeten, G. (2017). "A Comparative Analysis of Data Preparation Algorithms for Customer Churn Prediction: A Case Study in the Telecommunication Industry", *Decision Support Systems*, Vol. 95, pp. 27–36.

De Caigny, A., Coussement, K., De Bock, K.W. and Lessmann, S. (2020). "Incorporating Textual Information in Customer Churn Prediction Models Based on a Convolutional Neural Network", *International Journal of Forecasting*, Vol. 36 No. 4, pp. 1563–1578.

de Vargas, V.W., Aranda, J.A.S., Costa, R.S., Pereira, P.R.S. and Barbosa, J.L.V. (2023). "Imbalanced Data Preprocessing Techniques for Machine Learning: A Systematic Mapping Study", *Knowledge and Information Systems*, Vol. 65 No. 1, pp. 31–57.

De Caigny, A., Coussement, K. and De Bock, K.W. (2018). "A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees", *European Journal of Operational Research*, Vol. 269 No. 2, pp. 760–772.

El Attar, A. and El-Hajj, M. (2026). "Explainable AI-Driven Customer Churn Prediction: A Multi-Model Ensemble Approach with SHAP-Based Feature Analysis", *Frontiers in Artificial Intelligence*, Vol. 9, Article 1748799.

Gattermann-Itschert, T. and Thonemann, U.W. (2022). "Proactive Customer Retention Management in a Non-Contractual B2B Setting Based on Churn Prediction with Random Forests", *Industrial Marketing Management*, Vol. 107, pp. 134–147.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013). *Applied Logistic Regression*, 3rd ed. John Wiley & Sons.

Jahan, I. and Sanam, T.F. (2026). "A Comprehensive Framework for Customer Retention in E-Commerce Using Machine Learning Based on Churn Prediction, Customer Segmentation, and Recommendation", *Electronic Commerce Research,* Vol. 26, pp. 1–44.

Koh, B.H.D., Lim, C.L.P., Rahimi, H., Woo, W.L. and Gao, B. (2021). "Deep Temporal Convolution Network for Time Series Classification", *Sensors*, Vol. 21 No. 2, Article 603.

Krichen, M. and Mihoub, A. (2025). "Long Short-Term Memory Networks: A Comprehensive Survey", AI, Vol. 6 No. 9, Article 215.

Lalwani, P., Mishra, M.K., Chadha, J.S. and Sethi, P. (2022). "Customer Churn Prediction System: A Machine Learning Approach", *Computing*, Vol. 104, pp. 271–294.

Mena, G., Coussement, K., De Bock, K.W., De Caigny, A. and Lessmann, S. (2024). "Exploiting Time-Varying RFM Measures for Customer Churn Prediction with Deep Neural Networks", *Annals of Operations Research*, Vol. 339, pp. 765–787.

Mohammadi Foumani, N., Miller, L., Tan, C.W., Webb, G.I., Forestier, G. and Salehi, M. (2024). "Deep Learning for Time Series Classification and Extrinsic Regression: A Current Survey", *ACM Computing Surveys*, Vol. 56 No. 9, pp. 217:1–217:45.

Neslin, S.A., Gupta, S., Kamakura, W., Lu, J. and Mason, C.H. (2006). "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models", *Journal of Marketing Research*, Vol. 43 No. 2, pp. 204–211.

Peng, M., Peng, Y. and Li, H. (2023). "Research on Customer Churn Prediction and Model Interpretability Analysis", *PLOS ONE*, Vol. 18 No. 12, Article e0289724.

Sikri, A., Jameel, R., Idrees, S.M. and Kaur, H. (2024). "Enhancing Customer Retention in Telecom Industry with Machine Learning Driven Churn Prediction", *Scientific Reports*, Vol. 14, Article 13097.

van Leeuwen, R., Do, H. and Koole, G. (2025). "Churn Management in Hospitality", Journal of Big Data, Vol. 12, Article 145.

Verbraken, T., Verbeke, W. and Baesens, B. (2013). "A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25 No. 5, pp. 961–973.

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*, Vol. 290, Article 128465.