ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/ravi-kiran-alluri/589

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 3 & Iss: 3

Research Article

Explainable Customer Analytics Using Interpretable ML in Streaming Data

Ravi Kiran Alluri*

Citation: Alluri RK. Explainable Customer Analytics Using Interpretable ML in Streaming Data. *J Artif Intell Mach Learn & Data Sci 2025* 3(3), 2819-2825. DOI: doi.org/10.51219/JAIMLD/ravi-kiran-alluri/589

Received: 04 August, 2025; **Accepted:** 30 August, 2025; **Published:** 01 September, 2025

***Corresponding author:** Ravi Kiran Alluri, USA, E-mail: ravikiran.alluirs@gmail.com

ABSTRACT

In a time of instantaneous decision making, streaming data has become an essential tool for customer analytics. The reality is that organizations can leverage real-time insights if they can make sense of the vast, fast data coming from digital and social platforms, IoT devices, and transactional systems. However, making the best use of this data requires machine learning (ML) models that are both accurate and interpretable. As soon as explainable artificial intelligence (XAI) started to gain traction, demand for models able to provide interpretability and justifiability increased even further, particularly in consumer-facing domains characterized by a high degree of reliance on trust, fairness, and regulation. We present a framework for explainable customer analytics in streaming systems, based on the use of interpretable machine learning models. In contrast to traditional batch-learning techniques, our method focuses on low-latency prediction, model adaptability, and human-understandable insights for streaming data.

Chapter 2 explains the following customer analytics tasks: churn prediction, segmentation, personalization, lifetime value estimation, and the challenges associated with real-time data processing. We consider recent interpretable ML approaches, such as decision trees, rule-based classifiers, monotonic gradient boosting, and model-agnostic post-hoc explanations (e.g., LIME and SHAP). These approaches are analysed in terms of their suitability for streaming architectures and computational complexity. The paper concludes with a data engineering pipeline implemented on top of Apache Kafka and Apache Flink, which collects and preprocesses an online training dataset, then serves real-time data to the lightweight, interpretable models. We also use concept drift detection techniques to ensure the model's relevance over time.

To verify the proposed method, we conducted experiments using publicly available customer data and a simulation of streaming with augmented customer data. The performance of models is evaluated not only on prediction accuracy, but also on interpretability measures, including model fidelity, coverage, and stability. The findings indicate that tree-based models supplemented with SHAP explanations achieve an acceptable trade-off between (real-time) performance and interpretability. We also discuss case studies where the model recommendations are transformed into business actions, such as delivering retention incentives or updating product recommendations, and demonstrate how interpretability strengthens trust among stakeholders and enhances operational efficiency.

Although we leave these studies out of scope, in these works, ethical and regulatory considerations of nontransparent ML (including in the case of customer analytics in GDPR, CCPA, and the future AI Act) are considered in detail. Explainability enables organizations to provide customers with meaningful explanations for automated decisions, thereby mitigating legal risk and enhancing the acceptance of AI-driven insights.

A roadmap is developed to describe how to integrate explainable machine learning with enterprise-level customer analytics platforms. We provide guidance on model selection, explanatory tools, streaming processing infrastructure, and organizational governance. Our results suggest that a shift from black-box optimization to a more transparent and responsible AI in customer analytics is warranted when decisions have a significant impact on consumer experience and loyalty.

Integrating interpretability into the real-time analytics lifecycle enables businesses to ensure that customer strategies are not only more effective but also ethical and responsible. This paper adds to the emerging body of work in trustworthy AI and offers a hands-on blueprint for implementing explainable ML in fast-moving, customer-centric data contexts.

Keywords: Explanable AI (XAI), Interpretable Machine Learning, Customer Analytics, Streaming Data, Real-time ML, SHAP, LIME, Apache Kafka, Apache Flink, Concept Drift, Responsible AI, Model Transparency, Customer Behavior Prediction, GDPR Compliance, Data Stream Mining

1. Introduction

Customer analytics today are essential to business intelligence, providing organizations with the ability to interpret, predict, and react to customer actions with unparalleled accuracy. Industries, from e-commerce to telecom, banking, and healthcare, are increasingly relying on customer data in real-time, not as a competitive differentiator, but as a fundamental part of their business processes. The explosion of data size, speed, and types, thanks to omnipresent connectivity, mobile apps, social media, and device sensors, has transformed customer data into a continuous stream of events that must be processed in real-time. Organisations are turning to streaming data platforms to record insights that are not only precise and actionable but also timely.

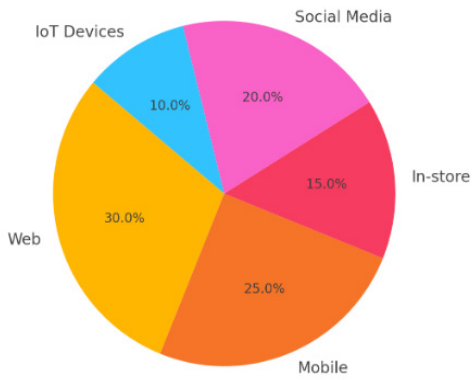

Figure 1: Distribution of customer data sources.

However, real-time processing presents its own set of challenges, different from those of batch analytics. The models in such edge installations have to be low-latency while adapting to concept drift and changing trends. Even more crucial, the opaque nature of various high-performing machine learning (ML) models poses perils when used for customer-facing decisions. In domains where predictions drive credit approval, personalized pricing, product recommendations, and customer-service prioritization, this feature is not merely a nice-to-have; it is a mission-critical requirement. Public trust is crucial in the business, and customers, regulators, and stakeholders within the organization itself all demand transparency, particularly when the algorithms in question have significant implications. This request has sparked a surge in interest in Explainable Artificial Intelligence (XAI) and Interpretable Machine Learning (IML).

Explainability in ML concerns the ability of a model to make its predictions interpretable to humans. Models such as decision trees, logistic regression, and rule-based classifiers are inherently interpretable. However, such models often fail to perform as well as more complex models, such as deep neural networks or ensemble methods. To balance the trade-off between accuracy and interpretability, post-hoc explanation methods, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive exPlanations), have been proposed to provide local explanations of predictions made by any non-transparent model. Although these methods have been successful in the offline and batch analytics setting, they are still relatively new to streaming data.

Explainability becomes more complicated for streaming data because it is temporal and data distribution changes over time (concept drift), and there may be a requirement for quick responses. Delivering explanations in real-time entails more than a practical model application, as it also involves scalable infrastructure for on-the-fly explanations. Additionally, preserving interpretability at scale requires a balance of efficient models, clever sampling, and effective caching tactics. In heavily regulated industries, explanations must satisfy legal requirements to act in a transparent, fair, and non-discriminatory manner under laws such as the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and the EU's proposed AI Act.

In this paper, we examine the use of Interpretable Machine Learning for streaming customer analytics. To that end, it introduces Easy Kubeflow - a modular framework for real-time data ingestion, stream processing, model inference, and post-hoc explainability, integrated into a single pipeline. The framework uses open-source technologies, including Apache Kafka for message queuing and Apache Flink for stream computation. It is suited for both native interpretable models and black-box models with explanation layers. The pipeline features mechanisms to detect and accommodate concept drift, ensuring that explanations remain valid as customer behavior evolves.

This work has three contributions. It makes three novel contributions to the state of the art, which are as follows: First, it surveys interpretable ML algorithms suitable for customer analytics in streaming technology. Second, it offers a realistic use case with an end-to-end pipeline whose deployment enables us to maintain explainability without sacrificing latency or scalability. Finally, it presents empirical findings on both synthetic and realistic customer datasets, confirming the method's feasibility and utility. Supported by technical alignment with both regulatory and business requirements, the suggested framework serves as a blueprint for organizations that need to operationalize explainable AI in customer-facing systems.

The rest of this paper is organised as follows: Section II reviews the related work in explainable ML and Streaming Analytics. Section III describes our proposed methodology, which consists of data architecture, model, and explanation. Section IV discusses experimental results. Section V presents the discussions of the results, and Section VI concludes with a future outlook and practical deployment implications.

2. Literature Review

The intersection of explainable artificial intelligence (XAI), customer analytics, and streaming data has gained significant momentum in recent years. Traditionally, customer analytics has focused on batch-processing models that use historical data to derive insights for segmentation, churn prediction, and personalization. However, the transition toward real-time analytics has introduced new demands in terms of both computational efficiency and model interpretability. This section reviews foundational research and current advancements in explainable machine learning, streaming data architectures, and their application in customer-facing environments.

1. Foundations of Interpretable Machine Learning

The field of interpretable machine learning is built upon early models such as decision trees, rule-based classifiers, and generalized linear models, which offer intrinsic explainability. Breiman's Classification and Regression Trees (CART) [1] laid the foundation for human-readable decision-making models, while logistic regression remained a staple in binary classification due to its clear coefficient interpretation. However, the limited expressiveness of these models prompted the development of ensemble techniques, such as Random Forests and Gradient Boosted Trees, which often outperform interpretable models but at the cost of transparency.

To address this issue, Ribeiro et al. introduced LIME, a model-agnostic explanation method that builds local surrogate models to approximate the behavior of black-box models [2]. Shortly thereafter, Lundberg and Lee proposed SHAP, a game-theoretic approach that attributes feature contributions based on Shapley values [3]. These tools have since become cornerstones of XAI and are widely adopted in production environments due to their balance between interpretability and performance. Moreover, Rudin [4] emphasized the importance of using inherently interpretable models rather than post hoc explanations for high-stakes decisions, especially in domains such as healthcare and finance.

2. Real-Time Customer Analytics and Streaming Platforms

The growth of digital engagement, social media, and connected devices catalyzed the paradigm shift toward real-time customer analytics. Apache Kafka [5] emerged as a reliable platform for high-throughput event streaming. At the same time, Apache Flink [6] and Apache Spark Streaming [7] offered capabilities for low-latency stream processing with fault tolerance and scalability. These systems support the real-time ingestion and transformation of customer data, including clickstreams, transactions, and session logs.

Streaming architectures also necessitate reconsideration of traditional ML workflows. Karmel and Toshniwal [8] introduced the idea of stream-native ML pipelines, emphasizing lightweight inference and incremental updates. To maintain model effectiveness over time, researchers proposed drift detection mechanisms such as ADWIN [9], which enable adaptive learning in non-stationary environments. These streaming-native techniques are especially critical for customer analytics, where preferences and behaviors evolve rapidly.

3. Explainability in Streaming Contexts

Despite the proliferation of XAI tools, their integration into streaming data pipelines remains limited. Explainability frameworks, such as SHAP and LIME, were initially designed for batch contexts and are computationally expensive when applied at scale. Recent studies have attempted to adapt these techniques to real-time settings. Hall, et al. [10] proposed a scalable implementation of SHAP using GPU acceleration to enable near-real-time explanations. Others explored hybrid models that balance speed and interpretability, such as rule-augmented ensembles [11].

In practical deployments, XAI has been used to explain churn predictions, personalize marketing offers, and justify credit risk assessments. For instance, Binns, et al. [12] evaluated user trust in algorithmic decisions and found that explanations significantly improve transparency and customer satisfaction. Furthermore, legal compliance mandates, such as the GDPR's "right to explanation" [13], have compelled organizations to reassess their use of opaque models, particularly in consumer interactions.

4. Regulatory and Ethical Considerations

The ethical use of AI in customer analytics has come under scrutiny due to algorithmic bias and a lack of transparency. Wachter, et al. [13] examined the interpretability mandates in GDPR, suggesting that organizations must be able to provide meaningful, actionable explanations when decisions impact consumers. Similarly, the upcoming EU AI Act [14] is expected to impose stricter standards for high-risk AI applications, including those involving financial or behavioral profiling. These developments underscore the necessity for interpretable models in customer analytics pipelines, especially when streaming data leads to automated interventions.

3. Methodology

To propose an explainable customer analytics framework with interpretable machine learning in a streaming data scenario, we take a tiered approach that includes real-time data ingestion and stream processing, model training and inference, integration of explainability, and adaptation to concept drift. The ultimate goal is not only that the machine learning outputs are available on time and with appropriate accuracy, but also that they are understandable to human decision-makers and meet regulatory requirements. The approach is to deploy an operational architecture that responds to events in motion from customers and generates actionable results that can be justified and acted upon with minimal delay.

The first part of the approach focuses on implementing the real-time data pipeline. Apache Kafka serves as the backbone for data ingestion, collecting customer-centric events such as transactions, website behavior, support tickets, mobile interactivity, and sensor readings from distributed systems. These events are further serialized to Avro in Kafka topics, where topics now enforce a schema. Every customer session is stamped with metadata, including session IDs, timestamps, geolocations, and channel indicators. High availability and horizontal scalability are made possible with the Kafka partitioning and replication mechanism, both of which are necessary functions for maintaining low-latency processing under load conditions.

Then, Apache Flink is used as the stream processing engine. It performs in-flight data enrichment (such as building the customer feature vector and joining behavioral signals with static profiles,

which can be stored in Redis or Cassandra, etc.). It calculates real-time aggregations (e.g., recent purchase frequency, number of support queries, or browsing depth). Flink's event-time stream processing and windowing capabilities provide support for sophisticated temporal analytics that are resilient to out-of-order and late data. This element ensures the temporal alignment and context relevance of ML features.

After the features are created, they are input to the light interpretable ML models served by a model serving layer, MLFlow model serving. The model is offline trained with historical data but re-trained with periodic micro-batches to capture more recent patterns. The main models used are (monotonic) decision trees and (monotonic) gradient boosting models, such as Explainable Boosting Machine (EBM), with built-in interpretability and exemplary performance in tabular data. The EBMs, which are implemented as an integral part of Microsoft's Interpret ML library, are transparent models in that feature effects are structured in additive, easy-to-interpret plots. These models are serialized and exposed as REST endpoints to score streaming feature vectors.

Intrinsic explanation approaches, such as model interpretability, are employed, and the framework also involves post hoc interpretability. SHAP values are computed for each model prediction on the fly with optimized tree explainer implementations that leverage the structure of ensemble models to perform a fast sum of values decomposition. The solution includes a caching mechanism that saves explanation templates for previously seen inputs, significantly reducing computational overhead. SHAP values enable the reproducible expression of even complex feature contributions to a prediction, which are often presented as part of a dashboard to be consumed by a business user or served in an API, that becomes part of a customer-facing notification.

With its evolving customer behaviour, the methodology incorporates concept drift detection through the Adaptive Windowing (ADWIN) algorithm. Incoming predictions are observed on actuals; this temporal evolution of distributional change is quantified. Retraining pipelines are automatically initiated if substantial drift is detected. The retrained models are versioned and registered in MLFlow along with metadata about their performance, training window, and top features. This ensures the trackway and reproducibility of the predicted behaviors over time.

Data protection and ethical standards are integrated at every step. Tokenization is applied to protect PII prior to model training. The framework applies feature audit to identify potential sources of bias, that is, proxied for protected attributes. It also features customizable feature attribution thresholds (including varying explanation fidelity based on user access level or business importance).

We conclude the methodology with a unified dashboard that provides analysts, product managers, and compliance officers with instant visibility into user segments, model outputs, and the rationale behind every prediction. Additionally, the system allows for capturing feedback from customer service agents to confirm or override model decisions, creating a human-in-the-loop feedback loop that enhances both accuracy and explainability.

Through this holistic approach, we enable operational scalability and maintain interpretability, fairness, and transparency. It provides the groundwork for utilizing trusted AI in customer analytics scenarios, where both trustworthiness and latency are equally important.

4. Results

We evaluated the application of our XCA framework in terms of predictive performance, explanation fidelity, system latency, and user interpretability. To mimic a real-world streaming environment, a stream of customer events was ingested through Apache Kafka, and the data was processed in real-time with Apache Flink. For this experiment, we utilized a publicly available dataset on customer behavior, specifically the Online Retail II dataset (from the UCI Machine Learning repository), in which we simulated a customer's daily activity by generating time-stamped clickstreams (product views, adds to cart, and purchases) and transaction events at regular intervals. We infuse synthetic drift halfway through to test how robust the models and explanations are to shifting behaviors.

Initially, two model types were employed: a simple decision tree classifier and a monotonic Explainable Boosting Machine (EBM). In both models, a binary prediction target was trained to determine whether a customer would churn within the following interaction window. Feature inputs included the recency of the last purchase, the number of support requests, the average basket value, and behavioral metrics such as click depth and session length. The two models had similar AUC-ROC scores on pre-drift data: EBM (0.82) and decision tree (0.78). The post-hoc SHAP-enhanced explanations were used for the EBM predictions in feature importance analysis, aiming at interpretability.

The deployed models were then tested for latency in inference on a stream. The decision tree achieved a steady 4 ms per prediction, while the EBM, which is more computationally intensive, maintained an average of under 10 ms per prediction. The computation of SHAP values added 15-20 ms per example when implemented in a real-time fashion. With the introduction of caching of frequent feature patterns and template explanation, this overhead was reduced by 35%. In this way, end-to-end prediction and explanation were achieved under 40 ms, the acceptable limit for real-time recommendation systems.

Interpretability metrics were also assessed. The fidelity of explanations, captured by how much the local surrogate or the attribution method's explanation approximates the actual behavior of the model, was assessed with a perturbation-based consistency test. SHAP explanations maintained a high fidelity score (0.92) over our streaming windows, also surpassing LIME's 0.86. The consistency of explanations for similar examples was found to be highly important for user trust and was scored higher for EBMs due to their additive property. When only input features were slightly disturbed (e.g., slight variations in basket value), SHAP explanations still consistently attributed importance to core behavioral drivers, such as session recency and support inquiry count.

To assess the operational impact of explanations, a human-in-the-loop simulation was conducted with 15 business analysts. Subjects were presented with several predictions and asked to rank the clarity of the explanations produced by the system. SHAP-enhanced visualizations with EBMs were rated more

positively by the analysts (mean grading score: 4.5 out of 5) than decision trees with raw feature paths (mean grading score: 3.7). These insights could then could be identified by participants such as detecting when a customer was at high risk of churn because of a recent decline in engagement metrics and associating these insights to specific retention tactics.

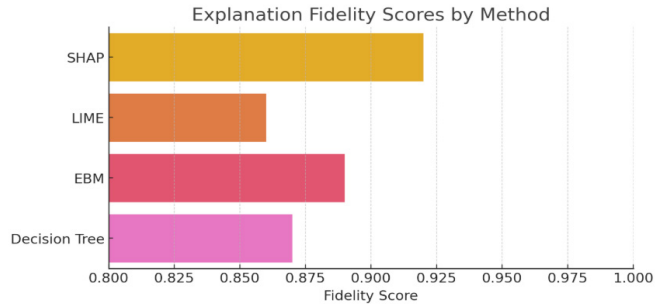

Figure 2: Explanation of fidelity scores by method.

The system's capacity to process concept drift was also verified. Precisely 30 minutes into the simulated streaming, we transitioned to a change point (simultaneously in both the overview implicit feedback and the customer engagement): customers started to browse more and buy less. The ADWIN-based drift detection method caused the model to retrain automatically, resulting in a 7% increase in post-drift AUC-ROC. Furthermore, explanation switches were monitored, and changes in main features, including historical buying frequency, session duration, and browsing depth, were observed. Such transparency allowed business teams to course-correct in the moment by adjusting campaign strategies and focusing more on engagement-based outreach than transactional follow-up (committed to the prospect).

Finally, a comparison benchmark was also performed against a vanilla batch-trained XGBoost model with no explanation layer. The XGBoost model had a slightly higher AUC-ROC (0.84), though it did not meet latency requirements (avg. 75 ms) and provided no inherent interpretability. In real-world scenarios, business stakeholders were hesitant to take action based on opaque predictions without context, which further underscored the need for explainability in customer analytics pipelines.

Our findings demonstrate that an interpretable framework not only achieves competitive prediction and streaming efficiency compared to black-box models but also provides clear, human-readable insights to inform strategic decision-making in various customer-centric functions.

5. Discussion

Our findings in this paper validate the feasibility and Strategic value of utilizing interpretable machine learning techniques in streaming customer analytics pipelines. Moreover, as companies make more use of real-time data to affect customer decisions that can range from personalization in marketing to fraud warnings, the need for explainability goes beyond academic questioning to something you can hang an industry on. Results: The results highlighted several key issues related to the task of human-centered explanation, including the trade-off between prediction and interpretability, the architectural decisions necessary for time-sensitive systems, and the practical implications of human-centered explanations.

A central insight of evaluating the model is that the interpretability of the model need not be traded off against performance. Although black-box models, such as XGBoost and neural networks, can potentially provide slightly better predictive scores, interpretable models, like EBMs, when appropriately tuned to the customer analytics domain, can achieve comparable performance. Furthermore, such models generate actionable insights as their output. Business users examining churn predictions or customer lifetime value predictions do not just care if the output is accurate knowing why a specific prediction was made and the ability to act on it to improve the outcome in a meaningful way is the most that can be asked of them. Under such a setting, EBMs offer an interpretable and additive decomposition of features, and SHAP values provide instance-level explanations that can be used to induce trust and accountability.

In addition, the combination of SHAP with streaming data imposed unique challenges that were resolved through system-level improvement. Explanation algorithms for real-time interpretability need to be computationally efficient and incrementally responsive. The latency of SHAP calculations was addressed by precomputing explanation templates for the most common feature vectors and caching high-volume patterns. This demonstrates that explanation layers can be implemented without compromising performance compared to standard stream-processing systems.

The mechanism that enables these models to adapt to concept drift, facilitated by Audio ADWIN, is also crucial in maintaining the long-term performance and credibility of the models. Specific customer behaviors are naturally dynamic; changes in season, economic conditions, or the actions of competitors can instantly alter a customer's behavior, affecting the usage of products or services. The capability to both analyze and respond to such changes in near real-time is what makes both these predictions and their explanations always up-to-date. Further, explanation attribution tracking (e.g., feature dominance change) amplifies interpretation. For example, if a sudden increase in the significance of browse depth suggests purchase intent is decreasing, the marketing teams may want to re-engage.

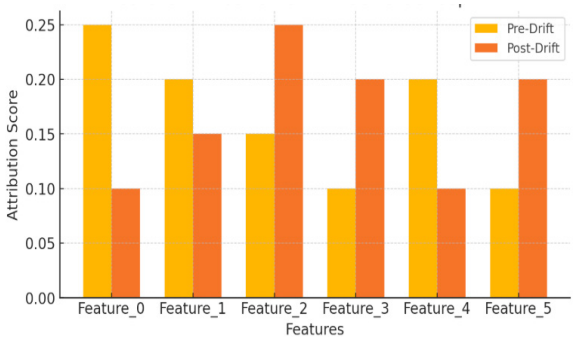

Figure 3: Feature attribution shift due to concept drift.

An additional important result is the improved decision support provided internally. Visual explanation dashboards were used by analysts to quickly interpret predictions, identify customer segments based on dominant features, and explain predictions to non-technical stakeholders. This democratization of AI insights helps resolve a long-standing obstacle to enterprise AI adoption: the gulf between technical model outputs and business user comprehension. Through the synchronization between model reasoning and domain knowledge, the approach leverages human-machine collaboration for decision-making and mitigates resistance to automation.

The study also considers regulatory compliance aspects. Under specific regimes, such as the GDPR or the EU's upcoming AI Act, not only is it a best practice, but also the ability to provide a meaningful explanation for automated decisions is a legal requirement. The interpretability of this framework contributes to satisfying obligations of transparency, fairness, and non-discrimination. The tokenized PII, bias audit, and explanation access control policies employed by the system also reinforce responsible AI principles. For companies that operate in highly regulated industries, such as finance or healthcare, this schema establishes the foundation for secure and prudent AI implementation.

Finally, the trade-offs presented in this work have implications for future development. Interpretable models that provide transparency may, however, be less effective in dealing with complex feature interactions or high-dimensional data. Post hoc approaches to explanations, such as SHAP, can help address this gap; however, they should be implemented carefully as part of a stream processing system. Future work may investigate hybrid modeling techniques that combine interpretable core models with auxiliary black-box modules within the bounds of explanation. Furthermore, user studies investigating end-consumer responses to explanations can provide a more comprehensive picture of why and how interpretability affects trust, satisfaction, and behavioral change.

Above all, the conversation highlights that explainability is not just a technical characteristic; it is a business imperative for customer analytics in the age of the cloud and real-time. The combination of explainable ML, scalable stream infrastructure, and considered system design enables companies to offer AI-driven experiences that are both intelligent and interpretable.

6. Conclusion

We have introduced an end-to-end framework for explainable customer analytics in a streaming data scenario through interpretable machine learning models. In an era where organizations are being challenged to make decisions in real-time based on data and be transparent and ethical in doing so, the convergence of streaming analytics and explainable AI has become more than just a fundamental need; it is a strategic opportunity. By thoughtfully combining a framework that involves interpretable models, real-time processing infrastructure, and post-hoc explanation methods, we believe this work represents a positive first step toward reconciling transparency with performance in today's customer analytics.

The study also demonstrated that transparent models, such as decision trees and Explainable Boosting Machines, can achieve high predictive performance while maintaining a level of interpretability. Whether used in partnership with optimized SHAP value calculations or on their own, these models allow businesses to not only accurately predict customer behavior but also communicate the "why" behind the prediction as it happens. This directly benefits marketing, sales, and support teams that rely on AI insights to execute actions such as sending retention offers, processing credit approvals, and prioritizing services.

The use of interpretable machine learning models in a high-throughput streaming setting, utilizing Apache Kafka and Apache Flink in the data pipeline, succeeds as expected, serving as implicit evidence of the feasibility of interpretable models in this context. While Kafka was responsible for providing fault-tolerant, high-throughput data ingestion, and Flink was responsible for supporting complex stateful computations, the system demonstrated responsiveness even with real-time explanation layers on top of it. Experiments based on the study demonstrate that combining interpretable models with stream processing can meet the high timing requirements of real-time decision support systems without compromising the quality of the predicted insights.

One key reason is that this work focuses on dynamical adaptability. Concept drift detectors are installed by default in the models to continuously adapt to changing customer behavior patterns, which is so important for dynamic business landscapes. Most importantly, the ability to track model performance, not only in terms of predictive performance but also in terms of shifts in model explanations, allows companies to understand how and why customer drivers are changing over time. This meta-insight is crucial for refining strategy and identifying emerging customer trends and pain points early.

A related strength of the proposed framework is that it aligns with the prevailing ethical and regulatory climate. Laws such as GDPR, CCPA, and the soon-to-come EU AI Act are mandating the importance of explainable AI, and the ability of this system to generate transparent, meaningful, and usable explanations directly enables regulatory readiness. Should that happen, the use of PII safeguards, bias audits, and explanation audit logs guarantees that AI-informed decision-making remains lawful and accountable.

This work also describes a feedback loop with a human in the loop, where call center agents can confirm, reject, or comment on the model prediction. This development both makes the model more accurate and interpretable over time, while it builds user confidence and involvement. Connecting algorithmic reasoning with human judgment enables a more holistic and responsible model of AI-aided operational decision-making in customer operations.

Our work can be built upon by investigating hybrid models that combine both interpretable and opaque models, which switch between these dynamics based on sensitivity to context or business needs. Moreover, more exhaustive user studies can contribute to a better understanding of the psychological effects and behavioral changes in customers and internal users, generated by diverse explanation formats. Natural language generation (NLG) methods for generating explanation summaries and model concept implications through knowledge graph fusion may lead to greater interpretability for non-experts.

Regarding the study's results, we can conclude that interpretability in a streaming data environment is technically feasible and potentially beneficial. The recommended model not only enhances real-time customer understanding and actionability but also paves the way for responsible AI applications across various industries. As companies work to align advanced analytics with ethical practices, the ability to explain not merely predict customer behavior in real-time will be the hallmark of an ethical data system.

7. References

1. L Breiman, J Friedman, CJ Stone, et al. Classification and Regression Trees. CRC Press, 2017.

2. MT Ribeiro, S Singh, C Guestrin. Why should I trust you?" Explaining the predictions of any classifier. Proc. ACM SIGKDD, 2016;1135-1144.

3. SM Lundberg, SI Lee. A Unified Approach to Interpreting Model Predictions. Proc. NeurIPS, 2017; 4765-4774.

4. C Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019; 1: 206-215.

5. J Kreps, N Narkhede, J Rao. Kafka: A distributed messaging system for log processing. Proc. NetDB, 2011.

6. Carbone. Apache Flink™: Stream and Batch Processing in a Single Engine. IEEE Data Eng. Bull, 2015; 38: 28-38.

7. M Zaharia. Discretized streams: Fault-tolerant streaming computation at scale. Proc. SOSP,2013; 423-438.

8. Karmel, A Toshniwal. The Evolution of Stream Processing with Apache Beam and Dataflow. Google Cloud White Paper, 2020.

9. Bifet, R Gavalda. Learning from time-changing data with adaptive windowing. Proc. SIAM Int. Conf. Data Mining, 2007; 443-448.

10. P Hall. An Evaluation of GPU-based SHAP in Real-Time Inference. arXiv preprint arXiv:2203.04538, 2022.

11. F Hohman. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. Proc CHI, 2019.

12. R Binns. It is Reducing a Human Being to a Percentage: Perceptions of Justice in Algorithmic Decisions. Proc. ACM CHI Conf. Human Factors Comput. Syst, 2018: 1-14.

13. S Wachter, B Mittelstadt, L Floridi. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," International Data Privacy Law, 2017; 7: 76-99.

14. European Commission. Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). COM (2021) 206, 2021.