International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

Decoding Customer Intent through Big Data–Driven Insights

Dr. P. Guhan [1], E. Yuvasri[2]

[1]*Principal, Jaya College of Arts & Science, Chennai.*
[2]*PG Student, Department of Computer Applications, Jaya College of Arts & Science, Chennai.*

Abstract - *Businesses can anticipate purchases, tailor recommendations, and optimize marketing strategies by leveraging big data analytics to predict customer behaviour. Traditional analytical methods are no longer adequate to capture intricate and dynamic customer patterns due to the exponential growth of digital footprints created by web logs, mobile applications, transactions, and social media interactions. This paper proposes a comprehensive framework for forecasting consumer actions in e-commerce by integrating large-scale data ingestion, preprocessing pipelines, automated feature engineering, scalable machine learning models, and production-grade deployment mechanisms. This architecture is based on contemporary big-data technologies, such as Apache Spark for distributed processing, Kafka for real-time streaming, HDFS for scalable storage, and sophisticated algorithms like XG Boost, Light GBM, Transformers, and deep neural networks. We describe modular system designs with a focus on real-time personalization, explainable AI (XAI), data drift monitoring, and continuous model retraining (MLOps), as well as a thorough review of previous research and limitations in current predictive systems. For comprehending the real factors influencing consumer choices, creating synthetic data to increase model resilience, and using ethical AI techniques to guarantee equity and openness in extensive e-commerce analytics.*

Key Words**: Big Data, Customer Behavior Prediction, E-commerce Analytics, Machine Learning,Clickstream Analysis, User Segmentation

1. INTRODUCTION

Predicting customer behaviour—including purchase intent, churn probability, click-through likelihood, and customer lifetime value—has become a critical capability for competitive e-commerce platforms. Conventional analysis For competitive e-commerce platforms, the ability to predict customer behaviour—including purchase intent, churn probability, click-through likelihood, and customer lifetime value—has become essential. Because businesses now produce enormous amounts of high-velocity data from web logs, mobile applications, transaction systems, and various third-party sources, traditional analytics techniques are no longer adequate. Big data analytics addresses these challenges by leveraging distributed storage and processing frameworks capable of handling terabytes or even petabytes of heterogeneous data. To extract meaningful patterns from such large-scale datasets, modern systems must integrate robust data engineering pipelines, advanced machine learning models, and real-time processing capabilities. This paper presents a scalable and practical methodology for building customer behaviour prediction systems that operate effectively across both batch and near-real-time environments, ensuring an optimal balance between predictive performance, latency, scalability, and interpretability.

2 Literature Review

[1]. Early efforts relied on logistic regression and decision trees on structured transactional data to predict churn and purchase probability. These approaches were effective on small to moderate datasets but suffered when scaled.

[2]. Ensemble methods (Random Forests, Gradient Boosting Machines) improved predictive power and robustness. Feature engineering—RFM (Recency, Frequency, Monetary), sessionisation, and user segmentation—became key to performance.

[3]. with richer behavioral signals (clickstreams, sequence data), RNNs, CNNs, and Transformer variants have been applied for session- and sequence-level predictions (e.g., next-item recommendation).

[4]. Big-data frameworks (Hadoop, Spark) and streaming platforms (Kafka, Flink) are widely used for scalable training and inference. Model serving at scale often involves feature stores and online model servers (e.g., TF Serving, Seldon)...

3. Methodology

3.1 Objectives

• With the high accuracy and minimal latency, forecast one or more customer behaviors (e.g., purchase within 7 days, churn in 30 days, next product category).
• Incorporate forecasts into recommendation engines and marketing automation.

3.2 Data Sources

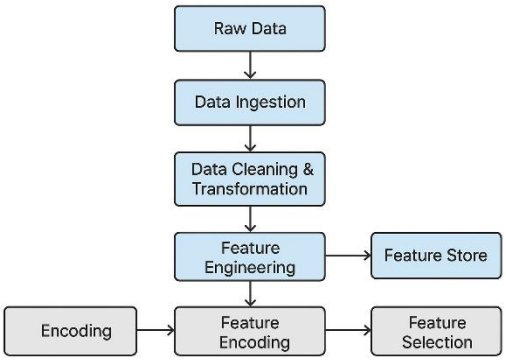
- **Transactional data**: Orders, returns, payment method, timestamps.
- **Behavioral logs**: Clickstream, page views, session durations, scroll/click events.

- **Profile data**: Demographics, location, loyalty tier.
- **Product/catalog data**: Category, price, availability.
- **External data (optional)**: Social signals, ad impressions, economic indicators.

3.3 Data Processing & Feature Engineering

Big data–driven customer behaviour prediction depends fundamentally on the quality, richness, and consistency of features engineered from raw, large-scale datasets. This phase converts massive, noisy, and multi-source information into structured, meaningful signals that machine-learning models can effectively use. To ensure scalability and fault tolerance, the workflow starts with data ingestion, where scheduled batch ETL jobs and real-time streams from Kafka/Kinesis are integrated into distributed storage systems like HDFS or Amazon S3. Once ingested, user activity logs and interaction traces are organized through essionization, which groups continuous events into coherent user sessions based on inactivity cutoffs.

High-level behavioural metrics like Regency–Frequency–Monetary (RFM) scores, popularity trends at the item and category levels, and time-decay–weighted aggregates to highlight recent actions are then produced by aggregation techniques.



(Fig 1: Big Data Processing Pipeline).

The pipeline also creates sequence-based features like last-N clicked or viewed items, inter-event time gaps, and temporal interaction patterns that capture evolving user intent. To represent deeper behavioural semantics, customer and product embeddings are learned using algorithms such as skip-gram, autoencoders, or sequence-based representation learning, enabling models to capture hidden relationships and similarities. The final stage includes comprehensive feature transformation steps such as normalization, scaling, one-hot and target encoding, frequency encoding, and systematic handling of missing or sparse values. All things considered, the pipeline for data processing and feature engineering guarantees that unprocessed big data transforms into excellent, information-rich representations that greatly improve predictive accuracy in customer behavior modeling.

3.4 Modeling Approaches

Modelling customer behaviour prediction in e-commerce requires selecting techniques capable of capturing complex feature interactions, sequential user patterns, and large-scale data variations. Important outcomes like purchase intent, churn risk, click-through probability, next-item recommendation, and total customer lifetime value are all predicted by these models.

Conventional baseline models, such as decision trees and logistic regression, offer interpretability and are good places to start when trying to understand important predictors. For richer tabular data, advanced gradient-boosting frameworks such as XG Boost, Light GBM, and Cat Boost are widely used due to their ability to handle heterogeneous features, non-linear relationships, and high-dimensional inputs. Deep learning architectures, such as RNNs, LSTMs, and Transformer-based models, are used to model sequential behaviour. These architectures learn temporal dependencies and recognize changing user interests from clickstreams or browsing sessions. By combining neural networks or gradient-boosting models with embedding layers for users, products, or events, hybrid approaches allow for the extraction of latent behavioural representations in addition to potent non-linear modelling. Finally, ensemble strategies combine multiple models to improve robustness, stability, and predictive accuracy across diverse customer segments and behaviour patterns.

3.5 Evaluation

To ensure both technical accuracy and practical impact, customer behaviour prediction models must be evaluated using a combination of statistical, machine-learning, and business-oriented metrics. For evaluating classification quality, standard performance metrics like AUC-ROC, Precision, Recall, F1-score, and PR-AUC are crucial, particularly in unbalanced environments where favourable events like purchases or churn may be uncommon. Furthermore, for decision-making systems like churn intervention programs or personalized recommendations, calibration metrics are used to confirm whether predicted probabilities accurately reflect true likelihoods. In order to replicate actual deployment behaviour, a strong validation strategy is equally crucial. Time-based splits, in which models are trained on earlier periods and validated on later windows, are frequently used, and cross-validation techniques are meticulously crafted to prevent temporal or user-level data leakage.

Beyond technical accuracy, models are ultimately evaluated through business metrics such as uplift in

conversion rate, increase in average order value (AOV), and improvements in customer retention measured through controlled A/B experiments. When taken as a whole, these evaluation dimensions guarantee that the predictive models are both statistically sound and able to produce quantifiable business value.

4. Existing System

Traditional e-commerce platforms rely on basic analytics pipelines that focus mainly on transactional reporting rather than predictive intelligence. In the existing system, customer behaviour analysis is typically limited to historical data summaries such as total sales, page views, or simple funnel metrics extracted through conventional reporting tools. Data is mostly processed through batch ETL pipelines, where interaction logs are aggregated overnight and stored in relational databases or basic data warehouses. These systems lack real-time processing capabilities, meaning customer actions cannot be analysed or responded to instantly. Sequential browsing behaviour, session patterns, or temporal dynamics are not captured by feature engineering, which is often limited to static attributes like user demographics or previous purchase counts. The use of machine learning is either restricted or depends on basic models that don't scale well with large datasets, such as logistic regression or rule-based recommendations. Furthermore, model deployment and monitoring mechanisms are either absent or manually managed, making it difficult to update models, track performance, or detect drift. The current system's capacity to deliver precise forecasts or customized user experiences at scale is severely limited by the absence of a unified feature store, automated pipelines, real-time event processing, and sophisticated modelling.

5. Proposed Modules

The proposed system introduces a modular, scalable, and intelligence-driven architecture designed to overcome the limitations of traditional e-commerce analytics. It begins with an Ingestion Module equipped with schema validation, deduplication, and fault-tolerant collectors to ensure high-quality event capture from multiple sources. For effective retrieval and subsequent processing, the Storage Module divides data into a raw zone and a curated zone, stores it in optimized parquet formats, and partitions it by date. A dedicated Feature Engineering Module leverages distributed Spark jobs and a centralized Feature Store to generate rich behavioural, sequential, and real-time features accessible across both offline training and online inference environments. The Modelling Module provides an end-to-end experimentation pipeline, supporting automated hyperparameter tuning, model versioning, and explainability frameworks to ensure transparency and reproducibility.

For deployment, the Serving Module enables low-latency inference through REST/gRPC endpoints, supported by intelligent caching and fallback strategies to maintain reliability under high traffic. A Feedback Loop Module, which gathers ground truth—such as conversions, user reactions, and online prediction performance—and feeds it into retraining cycles for model updates, is what propels continuous improvement. The Monitoring Module performs real-time data quality checks, tracks model accuracy, latency, drift, and fairness metrics, and triggers alerts when anomalies occur. Finally, the Privacy & Governance Module enforces strict access controls, anonymization methods, and compliance auditing to ensure secure and responsible handling of user data. These components work together to create a reliable, flexible, and future-ready system for predicting large-scale consumer behavior in e-commerce.

6. Implementation

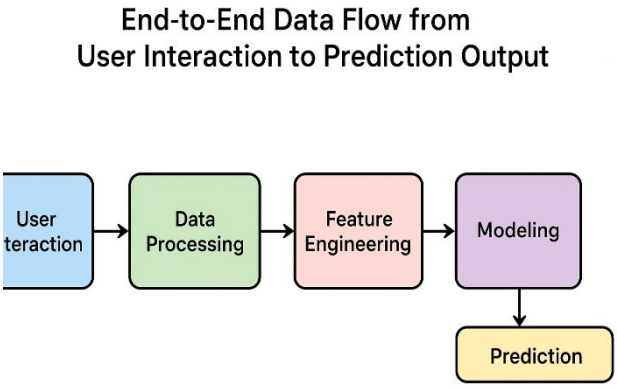


Fig 2: End to End Data Flow

The system is implemented using a modern, scalable big data and machine learning infrastructure tailored for real-time customer behaviour prediction. Data ingestion is handled through Kafka producers embedded in web and mobile applications, with Kafka Streams enabling real-time event transformation and routing. All incoming data is stored in an S3-based data lake, while Delta Lake layers provide ACID guarantees for reliable updates, versioning, and schema evolution. Large-scale feature engineering and distributed ETL workflows are supported by Apache Spark, which is used for processing and feature computation. A dedicated Feature Store—implemented using Feast or a custom design backed by Redis for online features and Parquet files for offline features—ensures consistency between training and inference.

Python frameworks like sickie-learn, XG Boost, PyTorch, and Tensor Flow are used in the modelling process, and Optima is used to optimize hyperparameters for better model performance. A lightweight FastAPI-based REST micro service that loads the most recent validated model

from ML flow's model registry is used to deploy models. Workflow orchestration and scaling are managed by Airflow and Kubernetes, ensuring reliable scheduling, execution, and distributed resource allocation. System health and performance are continuously tracked using Prometheus and Grafana dashboards, while custom data quality monitoring jobs detect anomalies, drift, and ingestion issues to maintain the integrity of the entire pipeline.

6.1 Example Experiment

1. **Dataset**: 6 months of user events and transactions. Create labeled examples where label=1 if purchase occurred within 7 days after the observation window.
2. **Features**: Last 7 days RFM, device type, channel, avg session length, last 5 viewed categories (embedded), time of day activity.
3. **Model**: XGBoost with class weighting; also train a sequence Transformer on last 50 events.
4. **Training**: Time-based split (train: months 1–4, val: month 5, test: month 6). Early stopping on validation AUC.

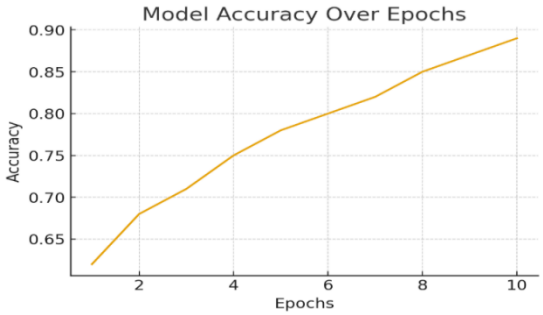


Fig 3: Accuracy Graph

5. **Evaluation**: Report AUC, Precision@10, calibration plot. Deploy best model to serving and A/B test against current recommender for 2 weeks.

```
# using pyspark
events = spark.read.parquet("s3://dl/raw/events/")
sessions = events.withColumn("session_id", generate_session_id(...))
rfd = sessions.groupBy("user_id").agg(
  last_event_time = max("timestamp"),
  session_count = countDistinct("session_id"),
  total_value = sum("price")
)
features = rfd.withColumn("recency_days", datediff(current_date(), col("last_event_time")))
features.write.parquet("s3://dl/curated/features/")
```

Fig 4: PySpark-Based Feature Engineering Pipeline

7. Conclusion

In this work, an end-to-end big data–driven framework for predicting customer behaviour in e-commerce has been presented, addressing the limitations of traditional analytics systems. The suggested architecture enables precise, real-time prediction of important customer actions like purchases, churn, and engagement by combining scalable data ingestion, distributed storage, sophisticated feature engineering, and cutting-edge modelling techniques. The incorporation of monitoring, governance, and feedback loops promotes continuous improvement and responsible data use, while the modular design guarantees flexibility, maintainability, and ease of integration across large-scale environments. Overall, the system provides a robust and production-ready solution that enhances personalization, improves decision-making, and drives measurable business value for modern e-commerce platforms.

REFERENCES:

[1] Ali, I., Mohammed, R., Nautiyal, A., & KumarSom, B. (2024). Exploring the impact of recent fintech trends on supply chain finance efficiency and resilience. https://doi.org/10.52783/eel.v14i1.1185

[2] Erdem, Ş., Durmuş, B., & Özdemir, O. (2017). Relationship between ad clicks and purchase intention: An empirical study of online consumer behaviour. European Journal of Economics and Business Studies, 9(1), 25–35. https://doi.org/10.26417/ejbes.v9i1.p25-35

[3] Gupta, S., & Israni, D. (2024). Machine learning–based customer behavior analysis and segmentation for personalized recommendations. https://doi.org/10.1109/icssas64001.2024.10760319

[4] Journal of Marketing Analytics. (2019). Palgrave Macmillan. https://doi.org/10.1057/41270.2050-3326

[5] Tokuc, A., & Dağ, T. (2024). Customer purchase intent prediction using feature aggregation on e-commerce data. E-Commerce Journal.

[6] Pandey, S., Aly, M., Bagherjeiran, A., Hatch, A., Ciccolo, P., Ratnaparkhi, A., & Zinkevich, M. (2011). Learning to target. Proceedings of the 20th Conference on Information and Knowledge Management, 1805–1814. https://doi.org/10.1145/2063576.2063837

[7] Punetha, N., & Jain, G. (2023). Game theory and MCDM-based unsupervised sentiment analysis of restaurant reviews. Applied Intelligence, 53(17), 20152. https://doi.org/10.1007/s10489-023-04471-1

[8] Widayati, C., Ali, H., Permana, D., & Riyadi, M. (2019). The effect of visual merchandising, sales promotion, and positive emotion on impulse buying behavior. Journal of Marketing and Consumer Research. https://doi.org/10.7176/jmcr/60-06

[9] Yerrineni, A., Ferri, B., Kokkili, S., Sirinelli, M., Martins, F. R., Riyasiman, S., & Kumar, R. (2023). Artificial intelligence–driven consumer behavioral analytics in marketing. Journal of Revenue and Pricing Management, 22(2), 171. https://doi.org/10.1057/s41272-022-00357-0