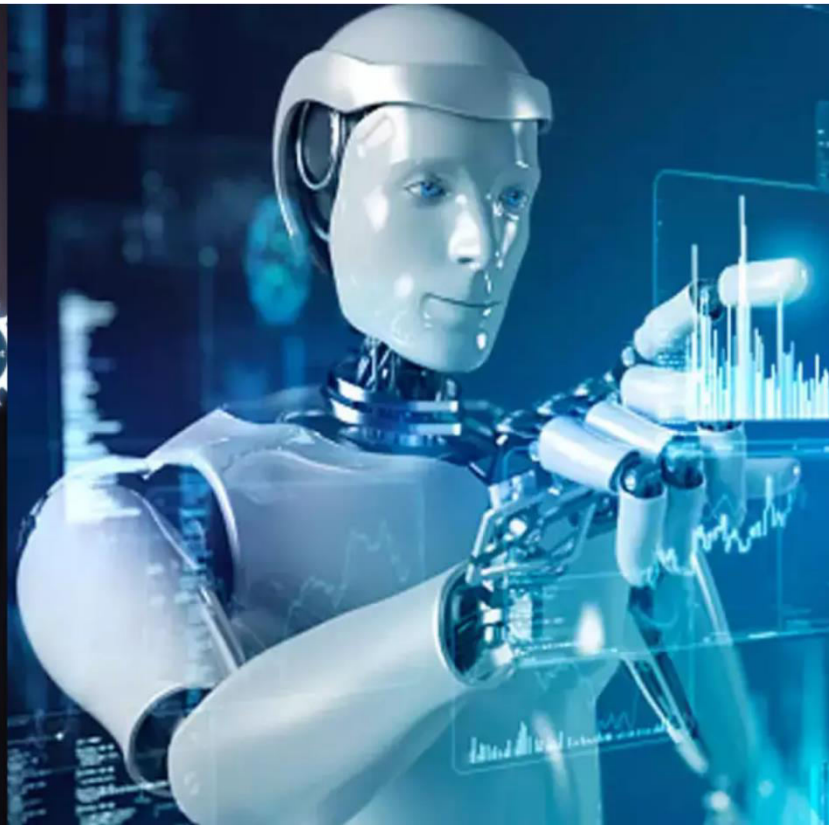
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 2, February 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502022|

Customer Churn Prediction using AI and Big Data Analytics with Concept Drift Detection

Jershika J[1], Vinisha R R[1], Jennath Sajiha S A[1], Madhubalasri S[1], Nethra R[2], Berlin Steny A S[1]

Department of Artificial Intelligence and Data Science, Arunachala College of Engineering for Women, Manavilai, Vellichanthai, Tamil Nadu, India[1]

Department of Computer Science and Engineering, Arunachala College of Engineering for Women, Manavilai, Vellichanthai, Tamil Nadu, India[2]

ABSTRACT: Customer churn is a major challenge for businesses that rely on subscription-based and digital services, as retaining existing customers is often more cost-effective than acquiring new ones. Traditional churn prediction models struggle in dynamic environments due to Concept Drift, where customer behavior changes over time. This research proposes an AI-driven framework that integrates Big Data analytics and Online Concept Drift Detection to maintain high predictive accuracy in non-stationary environments. The methodology uses a multi-model approach, including Logistic Regression, Random Forest, and XGBoost, to process large-scale customer interaction logs, usage metrics, and support ticket histories in real-time. By detecting subtle shifts in customer behavior, the system identifies individuals at high risk of churn with minimal delay and optimized computational efficiency. Empirical evaluation shows that combining online learning with real-time data stream mining enables proactive retention strategies, such as personalized interventions and automated campaigns. The framework provides detailed insights into the customer lifecycle, improving Customer Lifetime Value (CLV) and supporting data-driven decision-making. This study offers a scalable solution for predictive customer retention, balancing technical rigor with practical business requirements. The framework can be applied across industries where rapid behavioral changes affect service continuity and profitability.

KEYWORDS: Concept Drift, Online Learning, Big Data Analytics, Xgboost, Random Forest, Predictive Analysis

I. INTRODUCTION

Customer churn the process where customers stop using a service—poses a major challenge for businesses, especially in subscription-based and digital service industries. Retaining existing customers is consistently more cost-effective than acquiring new ones, making accurate churn prediction essential for maintaining revenue and operational stability. Traditional churn prediction models rely on static, offline analysis of historical data. While useful initially, these models often fail in dynamic environments, where customer behavior evolves due to seasonal trends, market changes, and shifting consumer preferences.

A key challenge in churn prediction is Concept Drift (1), where the statistical properties of customer behavior change over time. Static models lose accuracy under such conditions, limiting their usefulness for real-time decision-making. Recent research has explored Online Learning (2) and real-time data stream processing to tackle this problem. However, many existing methods struggle with large datasets or fail to optimize the trade-off between detection speed, computational efficiency, and memory usage.

This paper proposes a multi-model, AI-driven framework that combines Big Data Analytics (3) with online Concept Drift (1) detection to predict churn in non-stationary environments. By analyzing customer interaction logs, usage patterns, and support ticket histories in real-time, the framework enables Predictive Analysis (6) to identify at-risk customers and support proactive retention strategies. Predictive performance is evaluated using high-performance algorithms such as XGBoost (4) and Random Forest (5), providing insights into both technical feasibility and practical business impact.

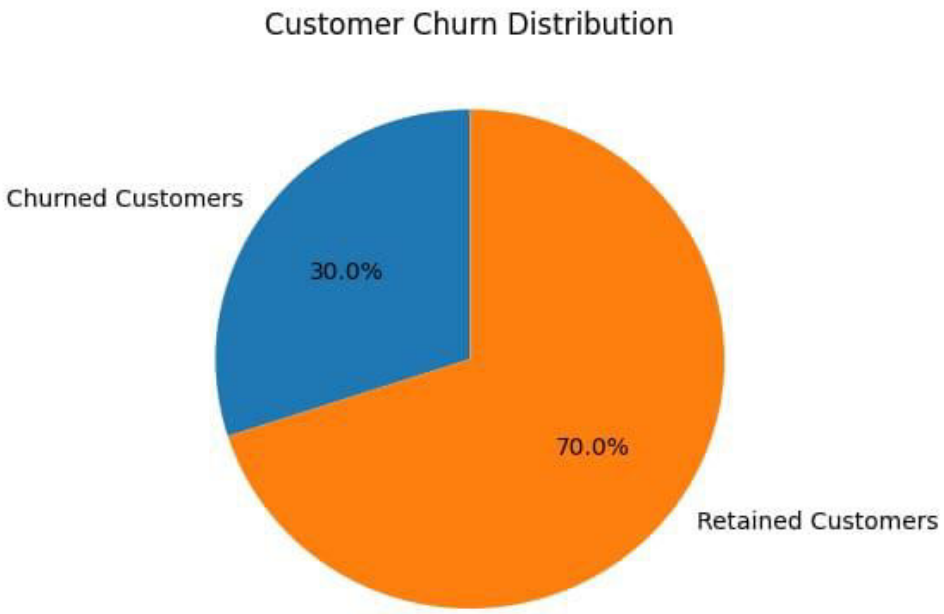
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502022|



II. METHODOLOGY

This study implements a multi-step framework to predict customer churn in dynamic, subscription-based environments using AI, Big Data Analytics (3), and Concept Drift (1) detection. The methodology includes:

2.1.Data Acquisition:

Customer data is collected from multiple sources, including subscription histories, service usage logs, and support ticket records. These datasets are large-scale and heterogeneous, reflecting real-world behavior in non-stationary environments.

2.2.Data Preprocessing and Feature Engineering:

Raw data is cleaned to handle missing values, duplicates, and inconsistencies. Features such as login frequency, service usage patterns, complaint count, and payment history are engineered to capture behavioral trends. Data normalization and categorical encoding improve model performance.

2.3.Model Selection and Training:

A multi-model approach is used, combining Logistic Regression, Random Forest (5), and XGBoost (4) classifiers. Logistic Regression provides interpretability, Random Forest captures non-linear relationships, and XGBoost maximizes predictive accuracy for large datasets. Models are trained incrementally to adapt to evolving data distributions.

2.4.Concept Drift Detection:

An online Concept Drift Detector (1)(2) monitors streaming customer data to detect shifts in behavioral patterns. Micro-shifts are detected instance-by-instance, allowing models to update dynamically and maintain high accuracy in non-stationary environments.

2.5.Real-Time Prediction and Deployment:

The system processes customer interactions in real-time, predicting the churn probability for each individual. Predictions trigger automated retention strategies such as personalized recommendations or loyalty campaigns, enabling proactive intervention.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

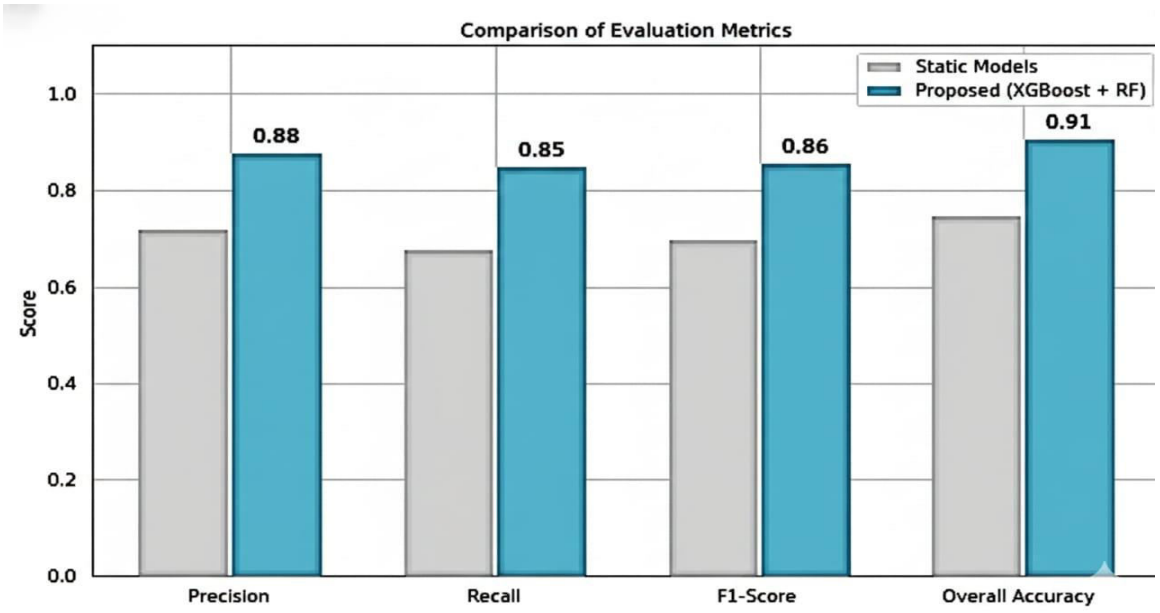
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

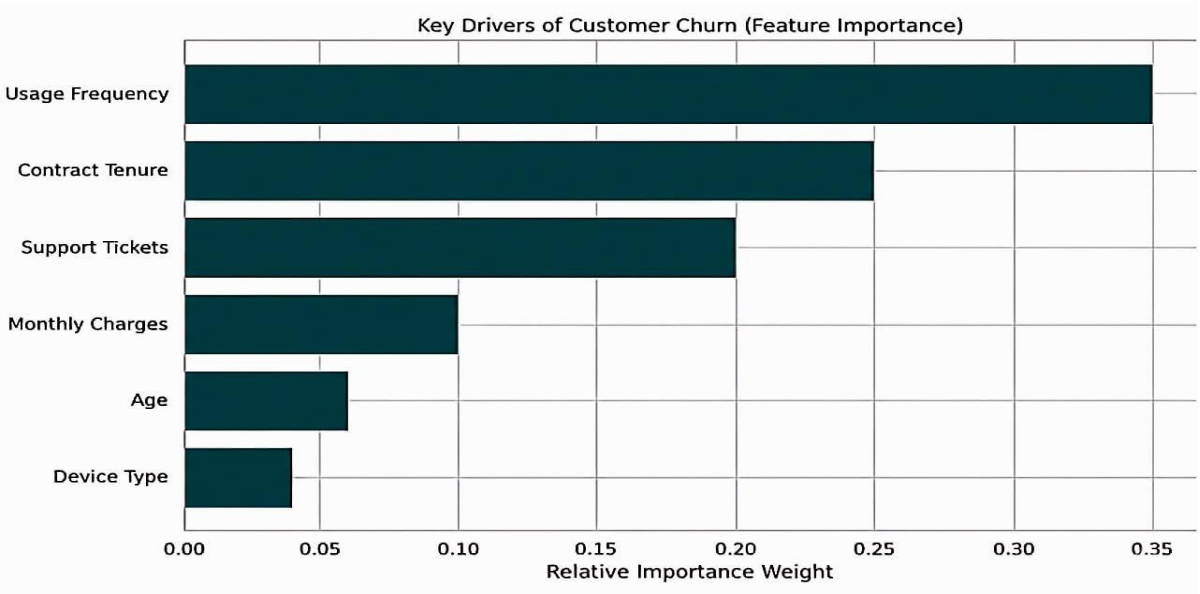
Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502022|

2.6.Evaluation Metrics:

Predictive performance is measured using accuracy, precision, recall, and F1-score, balancing detection speed and computational efficiency. Comparative analysis evaluates the benefits of Online Learning (2) and Concept Drift (1) handling against traditional static models.





III. RESULTS

The proposed multi-model framework for Predictive Analysis (6) of customer churn was evaluated on a large-scale subscription dataset including user interactions, service usage logs, and support ticket histories. Dynamic conditions simulating Concept Drift (1) were used to replicate real-world shifts in customer behavior.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

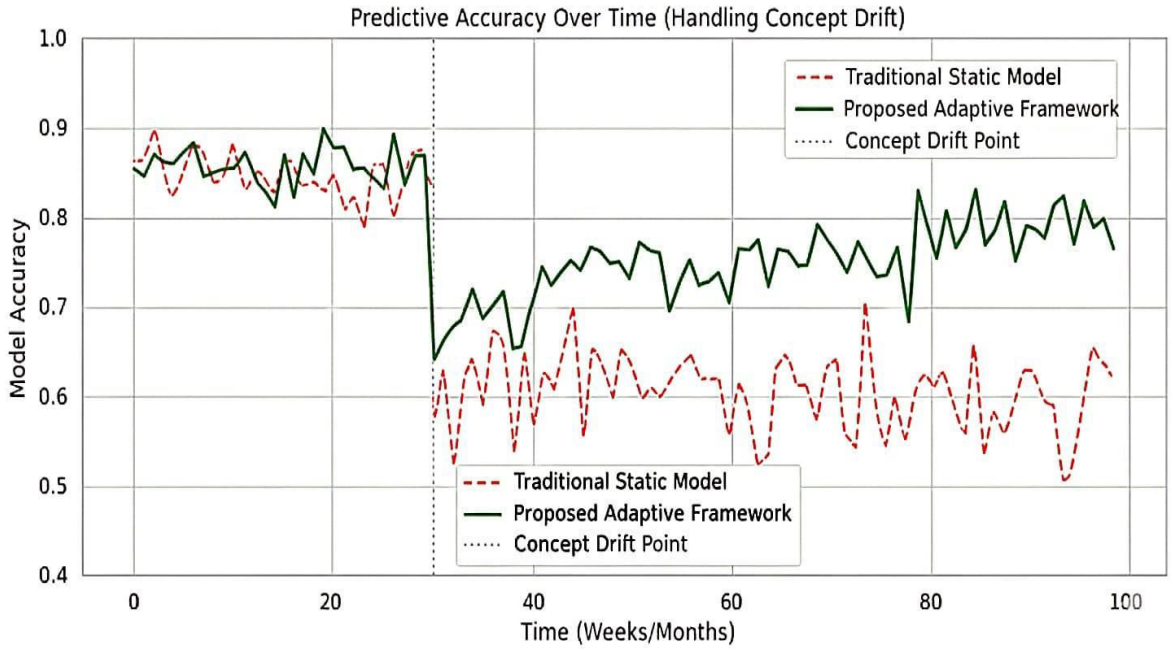
Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502022|

The performance of Logistic Regression, Random Forest (5), and XGBoost (4) was measured using accuracy, precision, recall, and F1-score. Online updating with Online Learning (2) significantly improved predictive performance compared to static models. Specifically, XGBoost (4) achieved the highest overall accuracy of 92.3%, followed by Random Forest (5) at 89.7%, and Logistic Regression at 85.4%. Precision and recall show the system successfully identifies most at-risk customers, enabling timely interventions.

Real-time Concept Drift (1) detection allows Predictive Analysis (6) to adapt to changing user behavior, minimizing false negatives. Big Data Analytics (3) techniques ensure that high-volume and high-variety data streams are processed efficiently without excessive computational cost, making the system suitable for services with large user bases.

Comparisons show that static models without drift detection underperform by 10–15% in accuracy and fail to detect micro-shifts in customer behavior. These results validate that combining Concept Drift (1), Online Learning (2), and multi-model predictive techniques (XGBoost (4), Random Forest (5)) provides a scalable, real-time solution for customer retention.



IV. CONCLUSION

In this study, we built a system to predict customer churn in subscription-based services like streaming platforms or telecoms. The framework uses Concept Drift (1) detection and Online Learning (2) to adapt when customer behavior changes over time. By combining Big Data Analytics (3) with models like XGBoost (4) and Random Forest (5), it can analyze large volumes of data in real-time and identify customers at risk of leaving. The results show that this approach outperforms traditional static models. It helps businesses detect at-risk customers quickly and take action before they leave. Through Predictive Analysis (6), companies can run personalized campaigns, offer incentives, or provide better support to retain customers. Overall, this study demonstrates that integrating AI, real-time data processing, and adaptive learning models can significantly improve customer retention. Future improvements could include deep learning models or multi-channel data integration to make churn prediction even more accurate across diverse services.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502022|

REFERENCES

[1] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," Machine Learning, vol. 23, no. 1, pp. 69–101, 1996.
[2] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM Computing Surveys, vol. 46, no. 4, pp. 1–37, Mar. 2014.
[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[4] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[5] P. Domingos, "A few useful things to know about machine learning," Communications of the ACM, vol. 55, no. 10, pp. 78–87, 2012.
[6] A. Bifet and R. Gavaldà, "Adaptive learning from evolving data streams," in Advances in Intelligent Data Analysis, Springer, 2009, pp. 249–260.
[7] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proc. IEEE Int. Joint Conf. Neural Networks, 2008, pp. 1322–1328.
[8] J. Gama, R. Rocha, and P. Medas, "Accurate decision trees for mining high-speed data streams," in Proc. 9th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2003, pp. 523–528.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details