INTERNATIONAL REVIEW OF ARTIFICIAL INTELLIGENCE IN MARKETING

NAWALA

Volume 1, Number 2, 2022

# Machine Learning Approaches to Customer Churn Prediction in Subscription Markets

**Nurul Muthmainna Zainuddin**[1*]

[1] Universitas Fajar, Makassar, Indonesia

## Abstract

**Article history:**
Received: August 30, 2022
Revised: September 17, 2022
Accepted: October 28, 2022
Published: December 30, 2022

**Keywords:**
Churn Prediction, Customer Retention, Machine Learning, Predictive Analytics, Subscription Markets.

**Identifier:**
Nawala
Page: 117-130
https://nawala.io/index.php/iraim

This article examines how machine learning has been applied to customer churn prediction in subscription markets such as telecommunications, media streaming, Software as a Service (SaaS), and financial services, where recurring revenue makes customer retention critical. It asks which algorithms and data practices are most commonly used, how they handle issues like class imbalance and high-dimensional features, and whether models are evaluated in terms of business value as well as statistical accuracy. The study employs a systematic review of peer reviewed articles published between 2017 and 2021, synthesizing evidence across telecom, broadband, banking, and other subscription contexts. The findings show that tree-based ensembles, stacking, and deep learning generally outperform traditional statistical models, especially when paired with targeted feature engineering and imbalance handling techniques. The article discusses these patterns by comparing sectors, methodological choices, and evaluation criteria, and concludes that future research should integrate economic objectives, temporal dynamics, and interpretability to make churn models more actionable for subscription businesses.

*Corresponding author:
(Nurul Muthmainna Zainuddin)

©2022 The Author(s).
This is an open-access article under CC-BY-SA license (https://creativecommons.org/licence/by-sa/4.0/)

## 1. Introduction

Subscription-based business models have become central to value creation in telecommunications, media streaming, Software as a Service (SaaS), gaming, and digital financial services. In these markets, revenue is generated through recurring payments rather than one-off transactions, which makes long-term customer relationships and stable subscriber bases critical for profitability. Losing existing subscribers is typically more costly than acquiring new ones, both because of sunk acquisition costs and the high marginal value of retained customers over their lifetime (Lemmens & Gupta, 2020). As competition intensifies and switching barriers decline, customer churn has emerged as a strategic risk factor that can erode revenue, increase marketing expenses, and destabilize growth trajectories if not effectively managed.

To address this challenge, firms increasingly rely on predictive analytics and machine learning to identify customers at high risk of churn and to inform targeted retention interventions. A substantial body of research has examined churn prediction in telecommunication and broadband services, treating attrition as a supervised classification problem where algorithms distinguish between churners and non-churners based on behavioral, contractual, and demographic features (Amin et al., 2017; Ahmad et al., 2019; Dhini & Fauzan, 2021). More recently, advanced methods such as ensemble learning, gradient boosting, and deep architectures have shown improved performance over traditional statistical models by capturing nonlinear relationships and high-order feature interactions (Ahmad et al., 2019; Xu et al., 2021). Parallel developments in banking and other subscription-

like financial services underscore the potential of explainable machine learning to support both predictive accuracy and managerial interpretability in churn models (Guliyev & Tatoğlu, 2021).

Despite this proliferation of studies, the literature remains fragmented along three important dimensions. First, existing work tends to focus on specific sectors, particularly telecommunications, with limited synthesis across subscription markets such as SaaS, streaming, and broadband that share similar economics and contract structures (Dhangar & Anand, 2021). Second, while individual papers demonstrate the superiority of particular algorithms or feature-engineering strategies, there is no consolidated view of which machine learning approaches perform best under different data characteristics, imbalance levels, or operational objectives such as profit maximization rather than accuracy alone (Ahmad et al., 2019; Lemmens & Gupta, 2020; Xu et al., 2021). Third, relatively few reviews explicitly connect technical modeling choices to actionable customer lifecycle management, such as segmentation of risk, personalization of retention campaigns, and integration with business rules and regulatory requirements (Dhangar & Anand, 2021; Guliyev & Tatoğlu, 2021).

This article addresses these gaps by conducting a systematic literature review of machine learning approaches to customer churn prediction in subscription markets, focusing on peer-reviewed studies published between 2017 and 2021. The review synthesizes evidence on algorithmic choices, feature types, data-preprocessing strategies, evaluation metrics, and sectoral contexts, with particular attention to imbalanced data handling and model interpretability. By mapping how

different methods perform across industries and problem settings, the study aims to clarify the state of the art, highlight unresolved methodological and managerial issues, and identify promising directions for future research and practice in subscription-based churn management.

## 2. Literature Review

Research on customer churn prediction in subscription markets has developed along several interrelated streams that collectively highlight the value and limitations of machine learning approaches. Early work in telecommunications and broadband treated churn prediction primarily as a binary classification task, comparing traditional methods such as logistic regression with more advanced machine learning models that exploit behavioral, contractual, and demographic features (Amin et al., 2017; Ahmad et al., 2019). More recent studies extend these comparisons to ensemble techniques such as random forests, gradient boosting, and bagging, generally finding that ensemble models outperform single learners in terms of accuracy, lift, and area under the curve when predicting churn in telecom and fixed broadband settings (Ahmad et al., 2019; Dhini & Fauzan, 2021; Xu et al., 2021). In parallel, work in banking and other subscription-like financial services illustrates how machine learning can yield both improved predictive performance and more granular insight into risk drivers when models are designed with managerial interpretability in mind (Guliyev & Tatoğlu, 2021).

A substantial body of research addresses two recurring methodological challenges in churn prediction: severe class imbalance and high-dimensional feature

spaces. Imbalance arises because churners typically represent a small minority of the customer base, which can bias models toward the majority non-churn class if not properly handled. Comparative studies show that the choice of resampling and cost-sensitive techniques has a strong impact on model performance and profitability measures (Zhu et al., 2017). Building on this, Salunkhe and Mali (2018) proposed a hybrid ensemble and under-sampling approach tailored to churn data, reporting gains in detecting minority churners without excessive loss of specificity. Feature selection studies such as Sivasankar and Vijaya (2019) demonstrate that systematic reduction of irrelevant or redundant variables can improve both accuracy and interpretability of churn models, particularly in telecommunication datasets with numerous usage and contract attributes. Together, these contributions underscore that algorithm choice, imbalance handling, and feature engineering need to be considered jointly rather than in isolation.

More recently, deep learning and representation learning have been explored as means of capturing complex patterns in subscriber behavior over time. Cenggoro et al. (2021) introduced a vector-embedding-based deep learning model for telecom churn that achieves competitive F1 performance while offering an interpretable latent representation of customer states. Such approaches suggest that it is possible to combine predictive power with explainability, which is crucial for operational use in customer relationship management and for regulatory compliance. However, reviews of churn prediction studies point to fragmentation across sectors, data characteristics, and evaluation metrics, with limited integration of profit-based measures, intervention cost, or long-term customer value into model assessment

(Lemmens & Gupta, 2020; Dhangar & Anand, 2021). This motivates a systematic synthesis that compares machine learning approaches across subscription markets, clarifies how they address imbalance and interpretability, and relates technical design choices to managerial objectives in churn management.

## 3. Methods

This study adopted a systematic literature review approach to identify and synthesize peer reviewed work on machine learning approaches to customer churn prediction in subscription markets. The search focused on articles published between 2017 and 2021 in English language journals. Major databases such as Scopus, Web of Science, IEEE Xplore, ScienceDirect, Google Scholar, and other publisher platforms relevant to marketing, information systems, computer science, and applied statistics were queried. Search strings combined terms related to churn and attrition with terms referring to subscription or contractual settings and machine learning, for example "customer churn", "subscription", "telecom", "banking", "machine learning", "ensemble", "deep learning", and "classification". All retrieved records were exported to a reference manager and duplicates were removed prior to screening.

Inclusion criteria were restricted to empirical or methodological studies that (a) addressed churn or attrition in subscription type markets such as telecommunications, broadband, media streaming, SaaS, or subscription like financial services, (b) applied at least one machine learning method for prediction or modeling of churn, and (c) were published in peer reviewed journals within the 2017

to 2021 window. Conference papers, dissertations, book chapters, non-academic reports, and studies focused solely on conceptual frameworks without empirical modeling were excluded. Screening proceeded in two stages: first, titles and abstracts were examined for relevance to churn prediction and subscription contexts, then full texts of potentially eligible articles were assessed against the inclusion criteria. For each included study, a structured coding template captured information on domain, data characteristics, feature types, handling of class imbalance, modeling techniques, evaluation metrics, and any links to profit, customer value, or managerial decision making. The coded material was synthesized narratively and thematically, with studies grouped according to sector and dominant methodological approach.

## 4. Results and Discussion

The systematic review identified a clear pattern in how machine learning is applied for churn prediction in subscription markets. Across telecom, broadband, e-commerce, and other recurring revenue contexts, most studies relied on supervised learning models built from transactional, usage, and demographic data, typically framed as a binary classification task (Ahmad et al., 2019; Dhini & Fauzan, 2021). Tree-based ensembles and hybrid architectures consistently outperformed linear baselines such as logistic regression, achieving higher accuracy, AUC (Area Under the Curve), and recall, particularly for the minority churn class (Ahmad et al., 2019; Cenggoro et al., 2021; Xu et al., 2021). Random forest and gradient boosting models were especially prominent, with evidence that combining them with domain-specific feature engineering yields further performance gains (Amin et al., 2017; Dhini &

Fauzan, 2021). This pattern is reinforced by work on telecom churn, where random forest provided strong results and enabled factor identification for retention strategies (Ullah et al., 2019), and by hybrid approaches that integrate logistic regression with tree-based methods, which improved predictive power and stability across multiple churn datasets (De Caigny et al., 2018).

Feature selection and representation learning emerged as central levers for improving model quality. Studies that used rough-set based selection, clustering-assisted feature construction, or deep representation learning generally reported better generalization and more robust performance under class imbalance compared with naïve inclusion of all available variables (Amin et al., 2017; Sivasankar & Vijaya, 2019; Cenggoro et al., 2021). Handling skewed churn rates remained a recurring methodological challenge: resampling techniques such as SMOTE, under-sampling, and cost-sensitive learning were often employed, and models were evaluated using AUC, F1, and recall rather than accuracy alone (Salunkhe & Mali, 2018; Ullah et al., 2019; Dhini & Fauzan, 2021). Nevertheless, many articles still prioritized marginal gains in global performance metrics rather than systematic analysis of error trade-offs along the churn decision threshold, which limits the ability to translate predictive improvements into actionable retention policies.

When the review focused on subscription business logic, a key divide emerged between accuracy-oriented and profit-driven studies. Only a subset of the literature explicitly integrated customer lifetime value, retention costs, or campaign profitability into model design and evaluation, even though marketing analytics research suggests that profit-based criteria can materially change which customers

should be targeted and how models are tuned (Lemmens & Gupta, 2020). Profit-driven work using advanced neural architectures showed that optimizing directly for expected profit can lead to different decision boundaries than those implied by accuracy or AUC, and that these models can substantively increase incremental retention revenue in telecom settings (Jafari-Marandi et al., 2020). Similarly, some empirical studies in banking and subscription services highlighted that churn prediction models that are modestly less accurate but better aligned with value-based targeting can outperform purely statistical champions in business terms (Lemmens & Gupta, 2020; Guliyev & Tatoğlu, 2021). This evidence suggests that the dominant focus on conventional performance metrics in the churn literature may understate the true strategic potential of machine learning for subscription markets.

Sectoral comparisons within the review further emphasized contextual nuances. Telecom and broadband studies generally operated on large, structured datasets with relatively rich usage and contract features, which favoured tree-based ensembles and gradient boosting (Ahmad et al., 2019; Ullah et al., 2019; Dhini & Fauzan, 2021). In contrast, e-commerce and digital subscription platforms often combined order history with behavioural traces such as browsing events, returns, and service interactions. Here, hybrid models that linked logistic regression with XGBoost, or that coupled interpretable linear components with non-linear learners, performed well, achieving high accuracy while preserving some degree of interpretability for managers (De Caigny et al., 2018; Li & Li, 2019). At the same time, relatively few studies explicitly addressed issues such as temporal dynamics of churn risk over contract cycles, multi-product subscriptions, or cross-channel

engagement, even though these are characteristic of modern subscription ecosystems. Overall, the findings indicate that machine learning can materially enhance churn prediction in subscription markets, but that the existing evidence base is still biased towards telecom contexts, accuracy metrics, and short-term binary labels, leaving significant room for future work on value-based, longitudinal, and cross-sector modelling.

## 5. Conclusion

Machine learning has become a central approach for customer churn prediction in subscription markets, particularly in telecommunications, broadband, and subscription-like financial services. Across the reviewed studies, models such as random forests, gradient boosting, stacking ensembles, and deep learning architectures consistently outperform traditional methods on common performance metrics, especially when predicting the minority churn class. Work on feature engineering, feature selection, and representation learning shows that model performance depends not only on the chosen algorithm but also on how behavioural, contractual, and demographic signals are constructed and filtered. Approaches that explicitly address class imbalance and high-dimensional feature spaces tend to deliver more stable and actionable churn scores for retention planning.

At the same time, the evidence base and this review have important limitations. Most empirical studies are concentrated in telecom and broadband, so conclusions may not fully capture the dynamics of SaaS, media streaming, or multi-

product subscription platforms. Many models are trained and evaluated on static datasets with binary churn labels, which underrepresents the temporal nature of churn risk and the impact of customer lifecycle stages. In addition, evaluation is often framed in terms of accuracy, AUC, or F1, while practical aspects such as intervention costs, contact channel capacity, and organizational constraints are rarely built into the modelling process. This review is also limited to peer reviewed journal articles in English published between 2017 and 2021, which means that recent developments and industry practice reflected in conference papers or practitioner reports are not fully represented.

Overall, the findings suggest that the next generation of churn prediction research and practice should move beyond pure accuracy benchmarking toward value-oriented and operationally grounded designs. Future work could integrate profit, customer lifetime value, and campaign costs directly into objective functions and decision thresholds, use longitudinal data to model how churn risk evolves over time, and study the impact of specific retention interventions. There is also a need for more attention to interpretability, so that data driven churn scores can be translated into clear reasons and targeted actions that make sense to managers and comply with regulatory expectations. By combining strong predictive performance with economic relevance, transparency, and temporal insight, machine learning based churn models can evolve from technical scorecards into strategic tools for managing long term customer relationships in subscription markets.

## References

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data, 6*(1), 1-24.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing, 237*, 242-254.

Cenggoro, T. W., Wirastari, R. A., Rudianto, E., Mohadi, M. I., Ratj, D., & Pardamean, B. (2021). Deep learning as a vector embedding model for customer churn. *Procedia Computer Science, 179*, 624-631.

De Caigny, A., Coussement, K., & 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, 269*(2), 760-772.

Dhangar, K., & Anand, P. (2021). A review on customer churn prediction using machine learning approach. *International Journal of Innovations in Engineering Research and Technology, 8*(05), 193-201.

Dhini, A., & Fauzan, M. (2021). Predicting customer churn using ensemble learning: Case study of a fixed broadband company. *International Journal of Technology, 12*(5), 1030-1037.

Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. *Journal of Applied Microeconometrics, 1*(2), 85-99.

Jafari-Marandi, R., Denton, J., Idris, A., Smith, B. K., & Keramati, A. (2020). Optimum profit-driven churn decision making: innovative artificial neural networks in telecom industry. *Neural Computing and Applications, 32*(18), 14929-14962.

Lemmens, A., & Gupta, S. (2020). Managing churn to maximize profits. *Marketing Science, 39*(5), 956-973.

Li, X., & Li, Z. (2019). A hybrid prediction model for e-commerce customer churn based on logistic regression and extreme gradient boosting algorithm. *Ingénierie des Systèmes d'Information, 24*(5), 525-530.

Salunkhe, U. R., & Mali, S. N. (2018). A hybrid approach for class imbalance problem in customer churn prediction: A novel extension to under-sampling. *International Journal of Intelligent Systems and Applications, 14*(5), 71.

Sivasankar, E., & Vijaya, J. (2019). A study of feature selection techniques for predicting customer retention in telecommunication sector. *International Journal of Business Information Systems, 31*(1), 1-26.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE access, 7*, 60134-60149.

Xu, T., Ma, Y., & Kim, K. (2021). Telecom churn prediction system based on ensemble learning using feature grouping. *Applied Sciences, 11*(11), 4742.

Zhu, B., Baesens, B., & vanden Broucke, S. K. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information Sciences, 408*, 84-99.