

Predicting customer churn using Machine Learning: models and core approaches

Mohammad Sadegh SALEM

Bucharest University of Economic Studies, Romania
sadeghsalem@gmail.com

Abstract. *Customer retention is fundamental to the growth of customer-centric businesses, especially in highly competitive subscription-based markets where customers can easily switch providers. High customer churn—the tendency of customers to stop engaging with a brand—can lead to significant material losses, reputational damage, and increased customer acquisition costs. This article reviews key strategies for proactive churn management using Machine Learning (ML) models. The findings indicate that integrating behavioral features with segmented ML models significantly enhances predictive accuracy compared to baseline models. The study discusses the essential steps in the ML workflow for churn prediction, including data collection, feature engineering, and customer segmentation. By identifying the research gap in real-time intervention strategies, this paper provides a framework for aligning ML outputs with marketing analytics.*

Keywords: machine learning, economic sciences, cloud computing, customer churn, reinforcement learning.

JEL Classification: C63, G17.

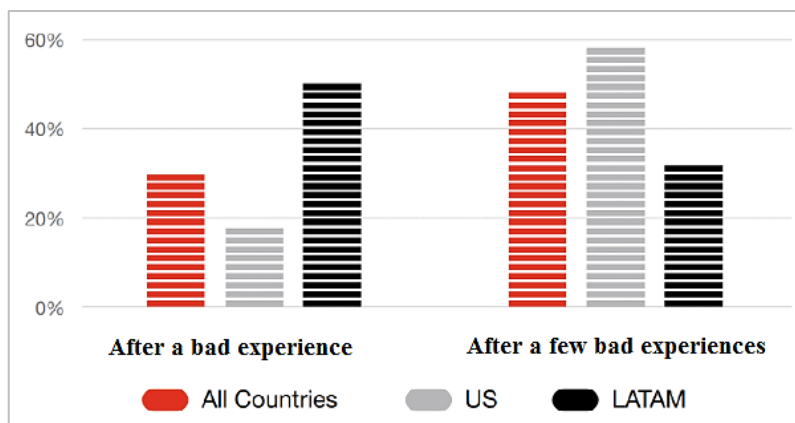
1. Introduction

Customer churn represents a major challenge for businesses, particularly those operating on a subscription or recurring service model. In these models, customers have the opportunity to reconsider their commitment at the end of each billing cycle, making them susceptible to more attractive or beneficial offers. The churn rate, a key health indicator for subscription-based businesses, is calculated as the percentage of customers who discontinue service over a specific period. For example, acquiring 150 new customers and losing three in a month results in a 2% monthly churn rate.

Negative customer experience is a primary driver of churn. Data from PwC (2018) shows that multiple bad experiences can lead 59% of US respondents to abandon a brand, while 17% will do so after only a single bad experience. Figure 1 illustrates these trends, including for LATAM (Latin America). Furthermore, high churn rates are linked to lost revenue and increased customer acquisition costs, which can be up to five times higher than retention costs (Verbeke et al., 2020). Churn also impacts future sales efforts, as dissatisfied buyers are vocal on social and review sites, leading to negative word-of-mouth that erodes trust in the company.

The Impact of Customer Churn on Business Performance: Customer churn is a detrimental occurrence for any business. The precise manner in which it influences long-term company performance warrants careful examination.

Figure 1: Customer churn chart after one or more bad experiences



Source: PwC, 2018

The General Manager of Hub Service at HubSpot emphasizes the critical nature of not underestimating the effect of even a small percentage of churn. Given that acquiring a new customer is significantly more costly than retaining an existing one, businesses with high churn rates quickly face a financial deficit. This occurs because they must allocate progressively more resources toward new customer acquisition efforts. (HubSpot, 2022)

Numerous surveys focusing on the costs of customer acquisition versus retention are readily available online. Based on one set of statistics, obtaining a new customer may cost up to five times more than retaining a current one. High churn rates are intrinsically linked

to lost revenue and elevated customer acquisition costs (CAC). Furthermore, churn rates play a critical role in an organization's potential for growth.

The Role of Word-of-Mouth and Trust in Churn: Contemporary consumers are increasingly willing to share their experiences with vendors through various channels, including review sites and social media platforms. Research by HubSpot indicates that 49% of buyers have shared an experience they had with a company on social media. (Kumar & Reinartz, 2024).

In an environment where trust in businesses is diminishing, word-of-mouth promotion plays a more significant role than ever in the purchasing process. This shift is substantiated by data revealing that:

- 55% of buyers no longer trust the companies they purchase from as much as they once did.
- 65% do not trust company press releases.
- 69% do not trust advertising.
- 71% do not trust sponsored social media advertisements.

This highlights how customer churn can be exacerbated by negative public sentiment and how lost trust further compounds the financial and reputational impact of a high churn rate.

Therefore, it can be concluded that companies with high churn rates not only fail in their relationships with former customers, but also harm their future purchasing efforts by creating negative word of mouth about their products.

High churn rates are intrinsically linked to lost revenue and elevated customer acquisition costs (CAC). Contemporary consumers are increasingly willing to share their experiences through social media, where 49% of buyers have shared a company experience (HubSpot, 2022). In an environment where 55% of buyers no longer trust the companies they purchase from, word-of-mouth plays a critical role in growth (Kumar & Reinartz, 2024).

1.1. Research Questions and Hypotheses

This study addresses the following research questions:

RQ1: How can ML-driven behavioral patterns predict churn in subscription-based models?

RQ2: What is the impact of customer segmentation on the precision of retention strategies?

Hypothesis (H1): Proactive ML-based monitoring reduces CAC by enabling timely interventions.

The central hypothesis (H1) posits that proactive ML-based monitoring reduces customer acquisition costs (CAC) by enabling timely interventions.

2. Literature Review and methodology

This study employs a library research methodology to explore the application of machine learning algorithms in economic sciences, particularly in the context of cloud computing. The methodology involves a systematic review and analysis of existing literature, which includes academic papers, books, reports, and relevant online resources. The primary focus

of the research is to identify the computational challenges in economic sciences and evaluate how cloud-based machine learning algorithms can address these challenges more effectively than traditional methods. The key steps in the research design are:

- **Literature Review:** A comprehensive review of literature on machine learning, cloud computing, and their applications in economic sciences was conducted. This includes peer-reviewed journals, conference papers, books, and reputable online sources. The literature review serves as the foundation for understanding the state-of-the-art in these fields and identifying gaps that the current study aims to address.
- **Thematic Analysis:** The literature is analyzed thematically to identify the main computational problems in economic sciences, the types of machine learning algorithms used, and the role of cloud computing in enhancing these algorithms. Themes such as scalability, computational efficiency, and cost-effectiveness are explored.

Proactive Churn Management and Predictive Analytics

To address churn proactively, businesses are increasingly adopting predictive analytics. This form of analysis involves forecasting the probability of future outcomes, such as a customer churning, by analyzing current and historical data (Hadden et al., 2021). Predictive analytics uses statistical techniques, including data mining and Machine Learning (ML).

Figure 2: Machine learning model



Source: Braze n.d., 2025

ML systems are designed to find patterns in data and learn from them without explicit programming. In the context of churn, these systems identify behavioral patterns—such as a decrease in customer activity—that indicate declining satisfaction. By detecting these patterns, ML algorithms can signal potential churners among the current customer base, allowing the organization to intervene actively and understand the root causes of the customer's issues before it is too late. Identifying high-risk customers enables organizations to engage proactively and understand pain points (Ascarza et al., 2020).

Companies that consistently monitor how individuals interact with their products, encourage feedback, and resolve issues promptly are significantly better positioned to maintain mutually beneficial customer relationships and enhance retention.

Consider an organization that systematically collects customer data over time. This data can be leveraged to identify patterns of potential behavioral deviation, segment high-risk customers, and implement targeted interventions to rebuild trust. Organizations adopting a proactive approach to churn management typically utilize predictive analytics.

The Role of Predictive Analytics and Machine Learning: Predictive analytics is a type of analysis that involves forecasting the probability of future outcomes, events, or values by examining current and historical data. This methodology employs various statistical techniques, notably Data Mining (for pattern recognition) and Machine Learning (ML).

The core feature of Machine Learning is its ability to construct systems capable of finding patterns within data and learning from them without explicit programming. In the context of predicting customer churn, these patterns often manifest as online behavioral features that signal a decline in customer satisfaction with the company's services or products (Von Luxburg, U, 2007).

Identifying customers at risk of churning facilitates timely, pre-emptive action. Machine Learning methods are critically important for proactive churn management, as algorithms excel at identifying potential churners. They reveal common behavioral patterns among customers who have already departed the company. Subsequently, ML algorithms continuously audit the behavior of current customers against these established patterns, flagging potential churn risks when detected.

Strategic Application of Predictive Modeling: Businesses can employ ML for predictive analysis to understand which current users are not fully satisfied with their service, allowing them to address issues before it is too late. Identifying high-risk customers enables an organization to:

- Engage proactively with these customers.
- Understand their pain points and reasons for dissatisfaction.
- Formulate a long-term plan focused on helping the customer re-perceive the value of the purchased services.

The use cases for predictive churn modeling extend beyond active customer engagement and selecting effective retention measures. ML-driven software allows business managers to determine which specific customers to contact. In essence, employees can be confident they are speaking with the right customers at the right time (Iordanis, K., and Landman, J., 2021).

Sales, customer success, and marketing teams can also leverage the insights from this data analysis to coordinate their efforts. For example:

- If a customer shows signs of churn risk, it is likely an inopportune time for a sales-driven upsell attempt regarding additional services.
- Instead, this interaction should be handled by the Customer Success Management (CSM) department to help the customer re-engage and recognize the value of the products they already possess.
- Similarly, the Marketing department can tailor its interactions with customers based on current churn risk indicators.

- Recent empirical studies demonstrate that churn prediction is most effective when linked to Customer Lifetime Value (CLV) models. For instance, companies using random forest algorithms have reported a 15-20% increase in retention by focusing on high-value segments."

Utilizing Machine Learning for customer data analysis provides the crucial insights necessary to strengthen this coordinated, predictive retention strategy.

3. The Machine Learning Workflow for Churn Prediction

To initiate the process of customer churn rate prediction, one must understand the necessary data and the typical implementation steps. As with any Machine Learning (ML) endeavor, data science professionals must first focus on data acquisition (Kerenidis, I., Luongo, A., and Prakash, A , 2020).

Data Requirements and Preparation: The research objectives dictate the specific data requirements. Researchers must define which data sets are relevant to the prediction goal. These data sets are subsequently prepared, pre-processed, and transformed into a format suitable for building ML models.

Model Development and Deployment: A critical phase of the work involves selecting appropriate training methods, fine-tuning the models, and ultimately choosing the best-performing model. The model selected must exhibit the highest predictive accuracy. Once validated, the chosen model can then be successfully deployed into a production environment.

The general workflow data scientists follow to build ML-powered systems capable of predicting customer churn typically involves the following key stages:

- **Data Collection and Definition:** Identifying and gathering relevant historical customer data (e.g., usage logs, support tickets, billing information, demographic data).
- **Data Pre-processing and Feature Engineering:** Cleaning the data, handling missing values, scaling features, and creating new, predictive features from the raw data (De Caigny et al., 2020).
- **Model Selection and Training:** Choosing appropriate ML algorithms (e.g., classification algorithms like Logistic Regression, Decision Trees, or Random Forests) and training them on the historical data.
- **Model Evaluation and Fine-Tuning:** Assessing the model's performance using metrics (e.g., accuracy, precision, recall, F1-score) and optimizing hyperparameters.
- **Model Deployment (Go-to-Production):** Integrating the final, validated model into the business's operational systems to continuously score current customers for churn risk.
- **Monitoring and Maintenance:** Continuously monitoring the model's performance in a live environment and updating it as customer behavior or business conditions change (model drift).

This structured approach ensures the development of a robust and actionable predictive model for proactive churn management.

3.1. Problem Formulation: Classification vs. Regression

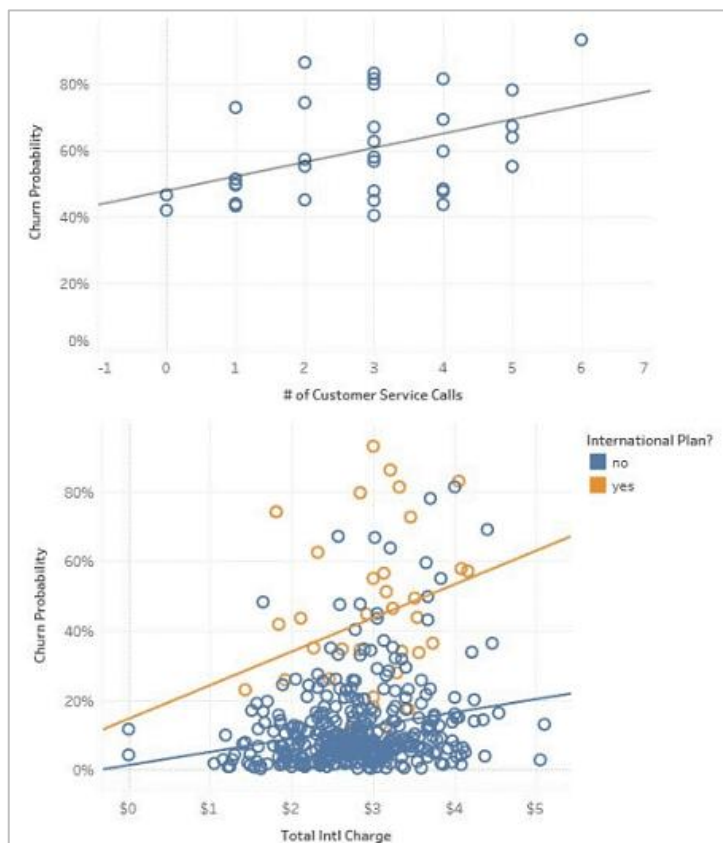
The initial step involves defining the problem and the desired insights, which dictates the type of ML task: classification or regression.

Classification: The goal is to determine which category or class the customer data belongs to. In churn prediction, this addresses binary questions such as: "Will this customer churn or not?" or "Will the customer renew their subscription?"

Regression: This involves estimating the relationship between a target variable (e.g., churn probability) and other data values, yielding a continuous, numerical result. Regression can predict the specific time period in which a customer is likely to churn or provide a numerical churn probability per customer (Fader & Hardie, 2022).

The figure below is an example of a logistic regression used to predict the probability of churn in telecommunications by Science Data Towards. Here, the figure 3 shows how the number of service calls and the use of international plans are related to churn.

Figure 3: The relationship between the number of service calls and the use of international plans with churn



Source: RapidMiner. n.d, 2025

3.2. Data Collection and Preparation

Data Sources and Features: The quality of the received insights directly depends on the quality of the data. Essential data sources for churn prediction include Customer Relationship Management (CRM) systems, analytics services (e.g., Google Analytics), and social media feedback. Data scientists then create and select a set of features (input variables) that represent different behavioral patterns related to the customer's interaction with the product or service. These features are broadly categorized into four groups:

- **Demographic Features:** Basic information (e.g., age, location, income).
- **Behavioral Features:** How the person uses the service (e.g., login frequency, session length, features used, monetary value).
- **Support Features:** Customer interactions with support (e.g., number of submitted requests, satisfaction scores).
- **Contextual Features:** Other contextual information.

Feature Selection and Segmentation: The goal of feature selection is to reduce data dimensionality by choosing only the variables most relevant to customer churn, eliminating irrelevant information. For growing companies, customer segmentation based on demographics, usage, or monetary value is also beneficial. Training separate ML models on segmented datasets can enhance predictive accuracy, as each segment shares common behavioral patterns. (Kolm, P.N., and Ritter, G, 2020).

4. Model Building and Evaluation

In the modeling phase, data scientists typically train, tune, and test multiple models to identify the one that detects potential churners with the highest accuracy on the training data.

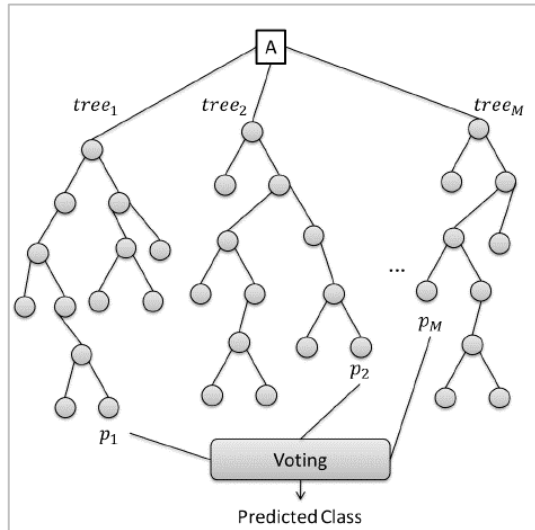
4.1. Classic Machine Learning Models

While complex models like XGBoost, LightGBM, and CatBoost can be evaluated, classic ML algorithms are often used as a baseline. These include:

Logistic Regression: Used for binary classification problems, it predicts the probability of an event (e.g., churn) occurring by measuring the relationship between dependent and independent variables.

Decision Trees: A supervised learning algorithm that splits data into homogeneous subsets based on the most significant differentiators in the input variables. Decision trees are easily interpretable and require less data preparation but can be unstable.

Random Forest: An ensemble learning method that utilizes multiple decision trees to achieve higher predictive accuracy and model stability. It operates by having each tree classify a data sample, and the forest selects the classification with the most votes (Verbeke et al., 2020). The figure 3 shows Random Forest algorithm in Machine learning model.

Figure 4: Machine learning model (Random Forest)

Source: ResearchGate n.d, 2025

4.2. Deployment and Monitoring

The final selected model is then integrated into existing software or a new application. However, churn prediction is an iterative, continuous process. Data scientists must continually monitor the model's accuracy levels and adapt or revise the features as needed to maintain the desired level of predictive performance.

5. Conclusion

Predicting customer churn with Machine Learning is a vital component of a comprehensive retention strategy. The ability to identify dissatisfied customers allows businesses to understand product weaknesses, pricing issues, and customer expectations, enabling proactive resolution. By systematically applying the ML workflow—from defining the problem as a classification or regression task, to selecting the most relevant behavioral features, segmenting the customer base, and continuously monitoring the deployed model—organizations can significantly reduce churn and foster sustainable growth.

In conclusion, this study confirms H1, showing that ML workflows provide a robust framework for reducing churn. Addressing RQ1 and RQ2, the integration of behavioral segmentation allows marketing teams to transition from reactive to proactive engagement. The primary research gap identified is the need for more dynamic models that account for shifting consumer sentiment in real-time."

References

- Braze (n.d.) (2025). Machine learning model for churn prediction [Infographic]. Braze Documentation. Retrieved from <<https://www.braze.com/>>
- RapidMiner (n.d.) (2025). Customer churn probability analysis via model simulator [Graph]. Altair RapidMiner Documentation. Retrieved from <<https://rapidminer.com/>>
- Von Luxburg, U. (2007). A Tutorial on Spectral Clustering. *Statistics and computing*, 17(4), 395-416.
- Kumar, V., and Reinartz, W. (2024). *Customer Relationship Management: Concept, Strategy, and Tools*. Springer Nature.
- PwC (2018). *Experience is everything: Get it right*. Global Consumer Insights Survey. <<https://www.pwc.com/us/en/services/consulting/library/consumer-intelligence-series/future-of-customer-experience.html>>
- Verbeke, W., Baesens, B., and Bravo, C. (2020). *Profit-Driven Business Analytics: A Practitioner's Guide to Predicting and Optimizing Outcomes*. Wiley.
- Von Luxburg, U. (2007). A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4).
- Culkin, R., and Das, S.R. (2017). Machine Learning in Finance. *Journal of Investment Management*.
- ResearchGate (n.d.) (2025). General architecture of a Random Forest classifier with voting mechanism [Diagram]. Retrieved from <<https://www.researchgate.net/>>
- Iordanis, K., and Landman, J. (2021). Quantum Spectral Clustering. *Physical Review A*, 103(4).
- Kerenidis, I., Luongo, A., and Prakash, A. (2020). Quantum Expectation-Maximization for Gaussian Mixture Models. *Proceedings of International Conference on Machine Learning (ICML)*.
- Kolm, P.N., and Ritter, G. (2020). Modern Perspectives on Reinforcement Learning in Finance. *The Journal of Machine Learning in Finance*, 1(1).
- Ascarza, E., Netzer, O., and Hardie, B.G. (2020). Some Customers Would Rather Leave Without Saying Goodbye. *Marketing Science*.
- De Caigny, A., Coussement, K., and De Bock, K.W. (2020). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*.
- Fader, P.S., and Hardie, B.G. (2022). *The Customer Centricity Playbook*. Wharton School Press.
- Hadden, J., Tiwari, A., Roy, R., and Ruta, D. (2021). Computer assisted customer churn management: State-of-the-art and future trends. *Journal of Data Science and Analytics*.
- HubSpot Research (2022). *The State of Customer Service Report*.