Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.36484

Customer Churn Prediction Using Logistic Regression and Behavioral Features

Qirui Zhu

Qianweichang College, Shanghai University, Shanghai, China
zqr0310@shu.edu.cn

Abstract. Customer churn prediction is important across telecommunications, subscription platforms, retail, banking, and digital services because retaining existing customers is often less costly than acquiring new ones. This paper provides a review of customer churn prediction using logistic regression and behavioral features, aiming to demonstrate how statistical modeling can support customer risk identification and behavioral data interpretation. The review first clarifies the basic concepts of churn, behavioral features, and binary classification, then introduces the mathematical principles, modeling procedures, and evaluation metrics of logistic regression. Through a comparative analysis of four representative empirical studies across banking, telecommunications, retail, and digital subscription services, the paper finds that the applicability of logistic regression depends heavily on feature engineering quality and the degree of linearity in business contexts. Key challenges identified include inconsistent churn definitions, varying data quality, the confounding of correlation with causation, and privacy constraints on data usage. Future research directions such as dynamic behavioral modeling, causal inference, and privacy-preserving analytics are discussed to further enhance predictive reliability and interpretability. This review contributes to understanding how fundamental regression methods can serve as interpretable baseline models in business data science, offering a reference for both researchers and practitioners in customer retention and risk management.

Keywords: customer churn prediction, logistic regression, behavioral features, binary classification, interpretable machine learning

1. Introduction

Customer churn prediction is valuable across telecommunications, subscription platforms, retail, banking, and digital product operations because acquiring new customers costs more than keeping existing ones [1]. Customer churn occurs across many industries, although its operational definition varies by sector. Firms need to identify potential churn risks from multiple data sources (customer behavior, usage records, service experiences, and personal characteristics) to decide where to direct retention spending and marketing resources [1]. As transaction logs, service records, and user behavioral data accumulate, recent research has mainly focused on five areas: churn labeling rules, behavioral feature selection, class imbalance handling, predictive performance, and model interpretability [1]. Each matters for a different reason. Labeling rules shape the training samples

DOI: 10.54254/2753-8818/2026.36484

and what the model learns to recognize. Feature engineering determines which churn signals the model is able to capture. Class imbalance is common because churners often represent a minority of the customer base. Performance evaluation must balance overall accuracy against identifying at-risk customers. Interpretability also determines whether business teams trust and act on model outputs. In social commerce, churn diagnosis also needs to account for social interactions and review behaviors, not just transactions [2]. On interpretability, studies have started using Shapley Additive Explanations (SHAP) to interpret churn prediction models, making their outputs more credible and usable in real business decisions [3]. On class imbalance, sampling techniques and ensemble learning frameworks help, though the problem remains an open challenge [4].

Recent studies have compared the predictive performance of multiple machine learning methods across telecommunications, banking, retail, and subscription services using features such as contract type, tenure, usage frequency, billing amount, and payment method [5-7]. On public datasets such as IBM Telco Customer Churn and Kaggle Bank Churners, ensemble methods such as random forest and gradient boosting tend to achieve higher accuracy. Support vector machines are also competitive in some settings. Logistic regression and decision trees are relatively simple to interpret but have lower accuracy. However, they are still applicable in scenarios where interpretability is required over a slight performance advantage [5, 6]. Feature engineering quality significantly affects the performance of the model. In retail, adding behavioral transition features to recency, frequency, and monetary value (RFM) metrics can substantially improve logistic regression performance [8]. Gradient Boosting performs better on complex interaction data, and is therefore more suitable for subscription services [7]. No single way works in all cases. The selection of a model should consider the industry, characteristics of the data and prediction goals for the business.

This paper studies customer churn prediction with logistic regression and behavioral features. Its goal is to show how statistical modeling can support customer risk identification and behavioral data interpretation. The four parts of this paper are: basic concepts of churn and behavioral features, logistic regression and common prediction methods, representative studies, limitations and future directions. Part one provides the basic concepts. The second is the theory and application of logistic regression models. Part three introduces some empirical studies in several fields and suggests when to use logistic regression. Part four presents the current research deficiencies and prospects for future research. Through this analysis, the paper aims to clarify the role of basic regression methods as interpretable baselines in business data science, for both researchers and practitioners.

2. Basic concepts of customer churn, behavioral features, and binary classification

Customer churn refers to the termination of a customer's relationship with a firm's products or services. It is a central concept in customer relationship management. Behavioral data are the raw traces of customer behavior, such as transaction logs, service-use records and customer-service communications. Behavioral features are structured variables derived from the raw data, such as the recency of the last purchase, purchase frequency and cumulative monetary value. They are often summarized in the RFM model. These are patterns of customer behavior that directly serve as input for predictive models [1, 8]. Churn prediction is a binary classification problem. A classification model learns from past data to determine whether a customer is likely to leave and assigns a probability of leaving for each customer [1].

The two main kinds of churn are intentional and involuntary. Intentional churn refers to a customer leaving the service voluntarily because of dissatisfaction or changes in their circumstances. Among them, deliberate churn (switching to a competitor) is the main object of prediction work [1]. Involuntary churn occurs when the company ends the relationship because of non-payment or

DOI: 10.54254/2753-8818/2026.36484

breach of contract [1]. The definition of churn in practice varies by industry. Telecommunications companies generally label a customer as churned after 30-90 days of inactivity, and other sectors have different time limits or behavioral criteria [9]. In binary classification, features (such as consumption behavior) and the target variable (churn or no churn) are the primary inputs and outputs. Class imbalance is a persistent problem: churners are generally the minority, and when the sample ratio is skewed, the model will favor the majority class and overlook a disproportionate number of at-risk customers [1]. This will be dealt with in the model through special means [1]. Model training splits the data into a training set and a test set. The training set is used to learn churn patterns, and the test set evaluates the generalization performance on new data [1]. The model outputs a probability between 0 and 1 instead of a hard class label, and then, using a threshold, performs the final classification.

Data types for customer churn research include demographic information (age, income), contract and billing details (subscription type, payment history), product usage data (usage frequency, feature adoption patterns), service interaction logs, and complaint records [1, 2]. Unstructured data in social commerce are also relevant. Social network structures and user-generated reviews provide researchers with a more complete understanding of customer attrition [2]. The RFM model is the classic framework for transforming behavioral data into structured input. It captures customer behavior patterns in three dimensions: recency, frequency, and monetary value [8].

3. Logistic regression methods for customer churn prediction

Logistic regression is one of the most widely adopted statistical methods for customer churn prediction, valued for its computational efficiency and strong interpretability as a de facto industry baseline [1]. A logistic (sigmoid) function is used to map a linear combination of independent variables to the range [0, 1], and the result can be interpreted as the predicted probability of churn. For a given set of customer features, the churn probability is expressed as a function where the linear combination of feature values and their corresponding coefficients, plus an intercept term, is exponentiated and divided by one plus that same exponentiated value. The signs of the coefficients show the direction of association, and their magnitudes should be interpreted only after considering the scale of the features. Odds are defined as the ratio of the probability of a person churning to the probability of not churning, and they are convenient for interpretation. Taking the natural logarithm of the above odds results in a linear form, and the log-odds can be represented as the sum of an intercept and weighted feature values. Exponentiating the regression coefficients produces odds ratios, representing the multiplicative change in churn odds per unit increase in the corresponding feature. Maximum Likelihood Estimation (MLE) is employed to find the parameter values that maximize the probability of obtaining the given sample, and these are considered the model parameters. For each customer, the likelihood is either the predicted churn probability if that customer has actually churned, or one minus that probability if they have not. The total likelihood is the product of these individual terms for all customers. The optimal parameter estimates are obtained by maximizing the likelihood function. Finally, based on the probability threshold (usually 0.5), a classification is made: if it exceeds this value, then the customer is considered churned, otherwise, they are not. The threshold can be adjusted according to the business's tolerance for false positives and false negatives [7].

Preprocessing the raw data before building the model can ensure the validity and stability of the model. Missing values are generally handled using mean, median, or mode, or they are removed if the proportion of missing data is too large [1]. Categorical variables need to be converted into numerical ones: label encoding for binary categories and one-hot encoding for multi-class

DOI: 10.54254/2753-8818/2026.36484

categories. Continuous features may be standardized to reduce scale differences, particularly when regularization or scale-sensitive comparisons are used. The dataset is split into a training set and a test set (commonly 70/30 or 80/20), and K-fold cross-validation is used to evaluate the generalization ability of the model. To handle class imbalance without causing data leakage, oversampling methods such as the Synthetic Minority Over-sampling Technique (SMOTE) or undersampling should be applied only to the training data in each cross-validation fold [1, 4].

Model performance is evaluated using various indicators. The confusion matrix is the evaluation foundation. The four categories of a confusion matrix are: true positives, false positives, true negatives, and false negatives [4]. Accuracy measures overall prediction correctness but can be misleading in imbalanced data [1]. Precision is the ratio of predicted churners that actually churned, and recall is the proportion of actual churners correctly identified. As the harmonic mean of precision and recall, F1 score is a comprehensive metric suitable for dealing with imbalanced data [1]. The area under the receiver operating characteristic curve (AUC) indicates the degree of class discrimination at different thresholds [1]. When dealing with imbalanced data, precision, recall and F1-score should also be presented alongside it.

The coefficient analysis of logistic regression provides direct support for business interpretation. In practice, standardized coefficients or odds ratios should be interpreted together with uncertainty estimates. Practitioners can use these measures to assess associations between behavioral features and churn risk and design targeted retention strategies.

4. State-of-the-art studies of customer churn prediction with logistic regression

This section reviews four empirical studies covering banking, telecommunications, retail, and digital news subscription services [6-8, 10]. These four studies differ in industry background, sample sources, churn definitions, behavioral features, data processing methods, model selection, evaluation metrics, and key findings. Together, they provide rich material for understanding where logistic regression works well and where it encounters its constraints.

How churn is defined and which behavioral features are extracted determine what each model learns and what it can predict. Tran et al. designated customers as churned when their account status shifted from "Existing" to "Attrited" in banking, and used Kaggle credit card data with 10,127 records [6]. Their features were transaction-based: total transaction count, total transaction amount, and months on book. In telecommunications, Subramanian et al. defined churn as a voluntary switch to another provider, combining structured survey responses from multiple telecom companies with 352 unstructured social media comments [10]. The features were complaint records, subscription period, gender and age. Zelenkov and Suchkova applied k-means clustering in retail to divide customers into loyal, sleeping and churned groups, and considered migration to a lower-value cluster as churn [8]. Based on RFM features of 33,918 customers at a Russian retailer, their model added estimates for inter-cluster transition probabilities to the standard framework. O'Brien and Kunz defined churn as subscription cancellation before Q1 2021 and used about 2.8 million raw logs collected from a digital journalistic platform [7]. After cleaning, 69,308 login records remained, and from these, 703 subscriptions were selected for churn prediction. Key features were active months, login count, actions per visit and time per visit. These selections of churn labels and features set the information boundary in which each classifier needs to work.

The various preprocessing strategies in the studies directly affected model training and generalization. Tran et al. deleted the records with "Unknown" values, used label encoding and one-hot encoding for categorical variables, and standardized or normalized continuous features [6]. They applied SMOTE oversampling to address class imbalance, set a 70/30 train-test split, and used 10-

DOI: 10.54254/2753-8818/2026.36484

fold cross-validation. Subramanian et al. used a simpler way: removing missing values, scaling features, and encoding categorical variables, without explicit resampling [10]. Zelenkov and Suchkova first used k-means clustering to group customers, then estimated flow probabilities between clusters, and used the resulting transition probabilities as additional features [8]. O'Brien and Kunz cleaned about 2.8 million raw log entries, retained 69,308 valid login records, then applied label encoding, normalization, oversampling and a 75/25 train-test split [7]. Each of the above pre-processing steps shaped the data representation that each model ultimately learned from.

Model selection and evaluation metrics differentiate the four studies further, and their key findings reveal important trade-offs. Tran et al. compared five classifiers: k-nearest neighbors (KNN), logistic regression, decision tree, random forest, and support vector machine (SVM), and used accuracy, precision, recall and F1-score [6]. Random forest had the highest accuracy, followed by SVM and logistic regression. Segmentation of customers using K-means clustering showed only a slight improvement in accuracy. Subramanian et al. found that logistic regression was the most accurate model overall and had the highest sensitivity, outperforming KNN, Naive Bayes and deep learning methods [10]. The result shows that logistic regression is particularly valuable when features have roughly linear relationships with the log-odds of churn. Zelenkov and Suchkova used only logistic regression as the classifier [8]. Augmenting standard RFM features with cluster-flow transition probabilities improved AUC by more than 10% during stable periods. Their model showed degraded predictive ability during the COVID-19 lockdown of April 2020, revealing how fragile data-driven churn models can become when customer behavior shifts abruptly. O'Brien and Kunz found that gradient boosting significantly outperformed logistic regression (84.7% vs. 65% accuracy), and the latter only achieved a 35% recall rate for the churn class [7]. They also showed that by lowering the classification threshold, the recall of random forest rose to 90%, and thus threshold tuning can align model output with business objectives. Across the four studies, no single classifier dominates. Random forest performed best in the banking case [6], logistic regression was sufficient for the telecom study [10], and gradient boosting led in digital subscriptions [7]. Accuracy, AUC, recall, and F1-score each emphasize a different aspect of performance, and the choice among them should reflect specific business priorities.

The interpretation strategies used in the four studies highlight a fundamental distinction between logistic regression and ensemble methods. Subramanian et al. used feature weight analysis on the models to find that subscription plan, age and retention effort were influential churn predictors [10]. Zelenkov and Suchkova determined through permutation importance that recency and frequency were the most influential factors [8]. O'Brien and Kunz used feature importance rankings from ensemble methods, and the active months were at the top of all predictors [7]. A key difference is that ensemble methods generally do not provide interpretable coefficients and odds ratios directly, but rather feature-importance rankings. Logistic regression offers coefficients along with their standard errors and confidence intervals for formal statistical inference. Such inference includes testing whether a coefficient differs from zero, quantifying uncertainty around estimates, and reporting odds ratios. None of the ensemble-based studies in this review reported such inferential statistics. Therefore, logistic regression has a marked advantage in quantifying variable effects and conducting hypothesis-driven analysis.

Based on the above comparisons, it can be seen that the utility of logistic regression depends heavily on industry context and how features are constructed. If the behavioral features are well structured, then logistic regression will work well. Zelenkov and Suchkova, and Subramanian et al. verified this [8, 10]. When the behavior data are complex and the relationship is non-linear, ensemble methods perform better than logistic regression, as shown by O'Brien and Kunz and Tran

DOI: 10.54254/2753-8818/2026.36484

et al. [6, 7]. Logistic regression outputs conditional probability estimates, not deterministic classifications. Through the regression coefficients, practitioners can identify which customers are likely to churn and determine how much and in what direction each behavioral factor is related to this churn. Ensemble methods often use feature-importance rankings or post-hoc explanations rather than directly interpretable coefficients. When business decisions need to understand the quantitative link between behavioral changes and churn risk, logistic regression retains its value as an interpretable baseline model.

5. Limitations and prospects

Despite the interpretability and predictive capability of logistic regression in customer churn prediction, existing research faces several limitations. Firstly, there is no standard definition for the churn label. Studies use different time windows, from 30, 60, to 90 days of inactivity, to operationalize churn, and the choice of labeling rules directly affects the prediction results and model evaluation [9]. In the case of Gradient Boosting, for example, reducing the label window from 90 days to 30 days results in an accuracy drop of about 9 percentage points [9]. A small-seeming choice of definition can significantly change the model's performance. Moreover, the quality of behavioral data is inconsistent. Customer data often have missing values, recording biases or time lags that may not reflect the current situation of the customers [1]. The company has actual limitations in data collection, and different data formats in various systems further complicate data integration. When external events such as changes in customer behavior due to the COVID-19 pandemic occur, models trained on historical data can show a significant drop in performance [8]. In addition, churned customers are generally a small proportion, so class imbalance remains a problem. Models will be biased towards the majority class and have reduced detection capabilities for at-risk customers [1]. A model with high overall accuracy is of limited use for retention if it fails to identify the minority-class cases. More importantly, most studies focus on correlations rather than causal relationships between features and churn [3]. Some variables that are highly correlated with churn may have no actual causal effect [3]. For instance, total credit card revolving balance is strongly correlated with churn [3]. However, based on causal analysis, total credit card revolving balance does not directly cause churn. This distinction is critical when designing retention strategies. Besides, data distributions differ significantly among enterprises and sectors [1]. Thus, models have reduced generalizability. A model that performs well on one dataset may degrade substantially when transferred to another. Finally, privacy regulations strictly limit the use and sharing of personal data, thus restricting the collection of training data and cross-organizational cooperation in churn prediction research [1].

Several directions merit investigation for follow-up research. At the data level, time-series-based behavioral feature modeling could capture how customer behavior changes over time [7], and more effective sampling techniques are required to reduce the effect of class imbalance [1]. At the level of methodology, causal inference methods such as R-learner can distinguish correlation from causation and provide stronger support for targeted interventions [3]. Wider use of explainable AI methods, such as SHAP, can improve the transparency and enhance trust in decision-making [1, 3]. Differential privacy can be added to predictive models to balance data availability with regulatory compliance [11]. Fairness evaluation should be added into the model development process to check whether the prediction results differ significantly for different groups of customers [11], and systematically assess the performance of the model on these subgroups. Given the distribution shift in different industries and firms, more robust modeling strategies warrant exploration to improve cross-scenario generalization, such as domain adaptation or transfer learning. Finally, dynamic

DOI: 10.54254/2753-8818/2026.36484

customer risk modeling tailored to different business contexts will be a key direction for enhancing the practical utility of prediction research [1, 7], and businesses will be able to respond promptly to alterations in behavior rather than depending on periodic batch predictions.

6. Conclusion

Customer churn prediction using logistic regression and behavioral features provides an interpretable framework for identifying at-risk customers and examining how observed behaviors are associated with churn. The reviewed evidence shows that its practical value depends on consistent churn labels, careful feature engineering, appropriate treatment of class imbalance, and evaluation metrics aligned with retention objectives. Logistic regression remains useful as a transparent baseline, but its linear form may be insufficient when behavioral relationships are highly nonlinear or change over time. Reliable applications should therefore select appropriate classification thresholds, evaluate performance across multiple metrics, and validate models on data from different time periods or contexts before deploying predictions. These steps are necessary before model outputs are translated into targeted retention interventions. The comparison of empirical studies also indicates that model performance is highly context dependent. Differences in churn definitions, sampling procedures, feature construction, and classification thresholds can change reported results even when similar algorithms are used. Logistic regression is especially useful when the relationship between features and the log-odds of churn is reasonably stable and when coefficient interpretation is important, while ensemble methods may provide better predictive accuracy for complex nonlinear patterns. Class imbalance should be handled within the training process, and accuracy should be reported together with precision, recall, F1-score, and area-under-the-curve measures. The review also highlights that predictive association should not be interpreted as causal evidence. Causal analysis, fairness checks, and privacy-preserving methods can therefore complement predictive modeling when results are used to guide customer interventions. Future work should evaluate models under distribution shifts and across industries so that improvements in predictive performance can be distinguished from dataset-specific effects.

References

[1] Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A review on machine learning methods for customer churn prediction and recommendations for business practitioners. *IEEE Access, 12*, 70434–70463. https://doi.org/10.1109/ACCESS.2024.3402092

[2] Li, S. J., Yue, R., & Yu, B. (2025). Diagnosis and management of customer churn in social commerce scenarios. In *Proceedings of the 4th International Conference on Cyber Security, Artificial Intelligence and the Digital Economy (CSAIDE 2025)* (pp. 586–594). ACM. https://doi.org/10.1145/3729706.3729799

[3] Li, Y., & Yan, K. (2025). Prediction of bank credit customers churn based on machine learning and interpretability analysis. *Data Science in Finance and Economics, 5*(1), 19–34. https://doi.org/10.3934/DSFE.2025002

[4] Suguna, R., Suriya Prakash, J., Aditya Pai, H., Mahesh, T. R., Vinoth Kumar, V., & Yimer, T. E. (2025). Mitigating class imbalance in churn prediction with ensemble methods and SMOTE. *Scientific Reports, 15*, Article 16256. https://doi.org/10.1038/s41598-025-01031-0

[5] Mettle, H. N.-A., Ayerko, E. H., & Appiahene, P. (2026). Machine learning-based customer churn prediction in telecommunication industry. *Applied Computational Intelligence and Soft Computing, 2026*, Article 8496186. https://doi.org/10.1155/acis/8496186

[6] Tran, H. D., Le, N., & Nguyen, V.-H. (2023). Customer churn prediction in the banking sector using machine learning-based classification models. *Interdisciplinary Journal of Information, Knowledge, and Management, 18*, 87–105. https://doi.org/10.28945/5086

[7] O'Brien, D., & Kunz, R. E. (2025). Will they stay or will they go? Application of machine learning classifiers of purchase and churn prediction to a digital journalistic platform. *Journal of Media Business Studies, 22*(3), 237–265.

DOI: 10.54254/2753-8818/2026.36484

https://doi.org/10.1080/16522354.2024.2444075

[8] Zelenkov, Y. A., & Suchkova, A. S. (2023). Predicting customer churn based on changes in their behavior patterns. *Business Informatics, 17*(1), 7–17. https://doi.org/10.17323/2587-814X.2023.1.7.17

[9] Bugajev, A., Kriauzienė, R., Vasilecas, O., & Chadyšas, V. (2022). The impact of churn labelling rules on churn prediction in telecommunications. *Informatica, 33*(2), 247–277. https://doi.org/10.15388/22-INFOR484

[10] Subramanian, D., Ajitha, A., & Maidin, S. S. (2025). Unveiling hybrid model with Naive Bayes, deep learning, logistic regression for predicting customer churn and boost retention. *Journal of Applied Data Sciences, 6*(2), 1379–1391. https://doi.org/10.47738/jads.v6i2.675

[11] Nagpal, R., Usua, U., Palacios, R., & Gupta, A. (2025). FairRAG: A privacy-preserving framework for fair financial decision-making. *Applied Sciences, 15*(15), Article 8282. https://doi.org/10.3390/app15158282