Customer Churn Prediction Using Machine Learning

[1] Kim Min Joon, [2] Fatima Al Mansoori

[1] Pohang University of Science and Technology (POSTECH), Pohang, South Korea

[2] United Arab Emirates University, Al Ain, UAE

Corresponding E-mail: kim.joon@interviauniversity.com

Abstract

Customer churn is a common problem for businesses that depend on keeping customers for a long period of time. This includes industries such as telecommunications, banking, insurance, subscription services, and online platforms. Customer churn happens when an existing customer stops using a company's products or services. Losing customers can reduce revenue and can also increase the cost of attracting new ones. For this reason, being able to identify customers who may leave can help businesses take action before the customer actually churns. Machine learning provides a useful way to approach churn prediction because it can analyze large amounts of customer data and identify patterns linked with customer departure. This research compares four machine learning algorithms for predicting customer churn: Logistic Regression, Decision Tree, Random Forest, and XGBoost. The proposed process includes data collection, preprocessing, feature selection, model training, and evaluation. Accuracy, precision, recall, F1-score, and ROC-AUC are considered when comparing the models. The analysis focuses on the strengths and limitations of each algorithm and highlights the importance of customer-related features such as contract type, service usage, payment method, tenure, and monthly charges. Logistic Regression provides a simple and understandable baseline, while Decision Tree can identify nonlinear relationships. Random Forest improves stability by combining multiple trees, and XGBoost can learn more complex patterns through gradient boosting. Overall, the study shows that machine learning can help businesses identify customers who are more likely to leave and support the development of targeted customer-retention strategies.

Keywords: Customer Churn, Machine Learning, Customer Retention, Logistic Regression, Decision Tree, Random Forest, XGBoost, Predictive Analytics

1. Introduction

Customer retention is an important goal for businesses operating in competitive markets. Companies spend a significant amount of time and money attracting new customers, so keeping existing customers is also very important [1]. When customers stop purchasing a product or cancel a subscription, the business experiences customer churn. A high churn rate can have a negative effect on revenue, market share, and long-term growth.

There can be many reasons why a customer decides to leave. Some customers may be unhappy with prices, service quality, customer support, or available features. Others may receive a better offer from a competitor or experience a change in their personal situation[2]. Understanding the exact reason behind every customer's decision can be difficult, especially for organizations that have thousands or millions of customers [3].

Traditional churn analysis usually relies on historical statistics or manually created business rules. These methods can provide useful general information, but they may not be able to capture complicated relationships between different customer characteristics. Machine learning offers a more flexible approach because it can learn patterns directly from historical customer data.

A churn prediction model can use information such as customer age, tenure, service usage, contract type, monthly charges, payment method, customer support interactions, and previous complaints [4]. By analyzing these characteristics, the model can estimate whether a customer is likely to leave.

Different machine learning algorithms can be used for this task. Logistic Regression is a simple and interpretable classification method that provides a useful baseline. Decision Trees can identify nonlinear relationships and create understandable decision rules. Random Forest combines several decision trees to improve robustness, while XGBoost uses gradient boosting to learn more complex patterns[5].

The main objective of this research is to compare these four machine learning algorithms for customer churn prediction. The study focuses on data preprocessing, feature selection, model training, evaluation metrics, and the practical value of churn prediction. It also discusses how the predictions can help businesses develop more targeted customer-retention strategies[6].

2. Methodology

The research follows a supervised machine learning classification approach. The main goal is to predict whether a customer will leave the organization or continue using its services.

The target variable has two possible outcomes:

- Churn
- No Churn

The overall methodology includes the following stages:

1. Data collection
2. Data preprocessing
3. Exploratory analysis
4. Feature selection
5. Dataset splitting
6. Model training
7. Model evaluation
8. Comparative analysis

Four algorithms are considered in the study: Logistic Regression, Decision Tree, Random Forest, and XGBoost.

2.1 Research Approach

This research treats customer churn as a binary classification problem. The model learns from historical customer records that have already been labeled as either churned or retained. It then uses the learned patterns to make predictions for new customers [7].

The process starts with collecting and preparing the customer data. After preprocessing and selecting useful features, the dataset is divided into training and testing portions. The training data is used to develop the models, while the testing data is kept separate so that performance can be checked on customers the models have not previously seen[8].

2.2 Dataset

A suitable customer churn dataset can be used for experimental analysis. Such a dataset may contain information about customer demographics, account details, service usage, billing information, and interactions with customer support [9].

Typical features may include:

- Customer tenure
- Age
- Monthly charges
- Total charges
- Contract type
- Payment method
- Internet or service type
- Number of subscribed services
- Customer support interactions
- Previous complaints
- Churn status

The target variable shows whether the customer has left the organization.

The dataset should contain enough examples of both churned and retained customers [10]. If one class is much larger than the other, techniques such as class weighting or resampling may be considered so that the model does not become overly biased toward the majority class.

2.3 Data Preprocessing

Customer datasets may contain missing values, duplicate records, inconsistent categories, or information stored in formats that are not directly suitable for machine learning. Preprocessing is therefore an important step before training the models[11].

Missing numerical values can be handled using suitable statistical methods, while missing categorical values can be assigned an appropriate category. Categorical variables such as contract type, payment method, and service type need to be converted into numerical form. One-hot encoding is one common way of doing this.

Numerical variables such as monthly charges and total charges can have very different scales. Feature scaling may be useful for algorithms such as Logistic Regression. Tree-based

methods, including Decision Tree, Random Forest, and XGBoost, are generally less affected by differences in feature scale.

After preprocessing, the dataset should be divided into training and testing sets. The training portion is used to learn patterns related to customer churn, while the testing portion is used to evaluate how well the models perform on unseen customers.

2.4 Feature Selection

Feature selection involves identifying the customer characteristics that provide useful information for predicting churn. Not every variable in a customer dataset contributes equally to the prediction.

Features such as contract type, tenure, monthly charges, service usage, and customer support interactions may contain useful information about customer behavior. Removing irrelevant or highly repetitive features can reduce unnecessary model complexity and may also help reduce overfitting.

Feature importance can be examined using statistical methods and tree-based models. However, a feature being strongly associated with churn does not necessarily mean that it directly causes customers to leave. For example, a relationship between monthly charges and churn could also be connected to service quality, customer type, or other factors.

2.5 Logistic Regression

Logistic Regression is a commonly used classification algorithm for predicting the probability of a binary outcome. In this case, it can estimate the probability that a customer will churn.

One of its main advantages is that it is relatively easy to understand. Businesses can examine the effect of different variables and get an idea of which customer characteristics are associated with a higher probability of churn.

Logistic Regression is also computationally efficient, which makes it useful as a baseline model. However, it generally represents relationships in a relatively simple way. As a result, complicated interactions between different customer characteristics may not be fully captured unless additional feature engineering is performed.

2.6 Decision Tree

A Decision Tree predicts customer churn by repeatedly dividing customers into groups according to their characteristics. For example, the model might first divide customers according to contract type and then use tenure or monthly charges to make further decisions.

One advantage of a Decision Tree is that its predictions can be represented as a series of rules, which makes the model relatively easy to understand. This can be useful for business managers who want to see the factors that contribute to churn predictions.

However, a tree that becomes too deep can fit the training data too closely. This is known as overfitting and can reduce performance on new customers [12]. Limiting the tree depth and adjusting other model parameters can help reduce this problem.

2.7 Random Forest

Random Forest combines several decision trees to produce a more stable prediction. Each tree is trained using different subsets of the data and features, and the individual predictions are combined to produce the final result.

Random Forest can capture nonlinear relationships and interactions between customer characteristics. This is useful because customer churn is usually influenced by several factors rather than one variable alone[13].

The algorithm can also provide information about feature importance, which can help organizations identify customer characteristics that are strongly related to churn. However, Random Forest is more complex than a single Decision Tree and is generally harder to interpret. It can also require more computational resources when a large number of trees are used.

2.8 XGBoost

XGBoost is a gradient-boosting algorithm that builds decision trees one after another. Each new tree focuses on reducing the errors made by the earlier trees.

This approach allows XGBoost to learn complex relationships between customer characteristics and churn behavior. It is well suited to structured customer datasets and can provide strong predictive performance when it is properly tuned.

The main drawback is its complexity. XGBoost has more parameters than simpler models, so poor parameter choices can lead to overfitting. Cross-validation and careful tuning are therefore important when developing the model.

3. Results and Discussion

The comparison shows that the four selected algorithms approach customer churn prediction in different ways.

Logistic Regression provides a useful starting point because it is simple, efficient, and relatively easy to interpret. It can help organizations understand how customer characteristics are associated with the likelihood of churn. However, its simpler structure may not be enough to capture complicated interactions between multiple variables.

Decision Tree offers more flexibility because it can create nonlinear decision rules. For example, it may identify a combination of short tenure, high monthly charges, and a particular contract type as a pattern associated with greater churn risk. However, an individual tree can overfit the training data if it becomes too complex[14].

Random Forest addresses some of the weaknesses of a single tree by combining many trees. This can produce more stable predictions and allow the model to capture more complicated customer behavior. It can also provide information about important features. On the other hand, its ensemble structure makes it less transparent than Logistic Regression or a single Decision Tree.

XGBoost uses a sequential boosting approach and can identify detailed patterns in structured customer data. It can be particularly useful when churn depends on interactions among several variables. However, its greater complexity means that careful tuning and validation are necessary.

Several evaluation measures should be considered when assessing churn prediction models.

Accuracy represents the overall percentage of customers that are classified correctly. However, accuracy by itself can be misleading if only a small portion of customers actually churn.

Precision shows how many customers predicted to churn really do churn. High precision can help businesses avoid spending retention resources on customers who were not actually at high risk.

Recall measures how many of the customers who actually churn are successfully identified. High recall is important because failing to identify a customer who is genuinely at risk means the business may lose the opportunity to retain that customer.

F1-score combines precision and recall and provides a balanced measure when both types of errors are important. **ROC-AUC** can also be used to measure how well the model separates customers who churn from those who remain.

A simple comparison of the selected algorithms is shown below:

Algorithm	Main Strength	Main Limitation
Logistic Regression	Simple and interpretable	Limited for complex patterns
Decision Tree	Easy to understand	Can overfit
Random Forest	Robust and flexible	Less interpretable
XGBoost	Strong for complex patterns	Requires careful tuning

The analysis suggests that customer churn is usually influenced by several factors rather than one specific variable. Contract type can be important because customers with flexible contracts may have fewer barriers to leaving. Tenure can also provide useful information because newer customers may behave differently from long-term customers.

Monthly charges and service usage can also affect customer decisions. A customer who feels that the service is too expensive compared with its value may be more likely to leave. Repeated service problems or poor customer support can also increase dissatisfaction.

Churn predictions can be especially useful when they are connected to targeted retention strategies. Instead of offering discounts or incentives to every customer, an organization can focus its resources on customers who have been identified as having a higher risk of leaving.

At the same time, predictions need to be interpreted carefully. A high-risk prediction does not mean that a customer will definitely leave. It simply means that the customer's characteristics

are similar to patterns seen among customers who previously churned. Businesses should therefore combine model predictions with customer feedback and other relevant information.

Data quality is another major factor. Customer behavior changes over time, and a model trained on older data may become less accurate as pricing, competitors, market conditions, and customer preferences change. Regular evaluation and retraining can help keep the model useful.

Privacy and responsible data handling are also important. Customer datasets can contain personal and financial information, so organizations need to make sure that such data is collected, stored, and processed securely.

Overall, machine learning can help businesses move from reactive retention to a more proactive approach. Instead of waiting for customers to cancel their services, organizations can use prediction models to identify possible warning signs and address customer concerns earlier[15].

4. Conclusion

This research presented a comparison of four machine learning algorithms for customer churn prediction: Logistic Regression, Decision Tree, Random Forest, and XGBoost. The study covered the main stages of churn prediction, including data preprocessing, feature selection, model development, and performance evaluation.

Logistic Regression provides a simple and understandable baseline, while Decision Trees can create clear nonlinear decision rules. Random Forest improves prediction stability by combining multiple trees, and XGBoost can capture more complex relationships between customer characteristics and churn behavior.

The study also shows that accuracy alone is not enough when evaluating churn prediction models. Precision, recall, F1-score, and ROC-AUC should also be considered, especially when the number of customers who churn is much smaller than the number who remain.

Customer churn can be influenced by factors such as contract type, tenure, pricing, service usage, and overall customer experience. Machine learning can identify useful patterns in these variables and help businesses recognize customers who may be at greater risk of

leaving. Overall, machine learning-based churn prediction can be a valuable decision-support tool for customer retention. By identifying high-risk customers early, organizations can take more targeted actions, such as providing personalized offers, improving customer support, adjusting services, or introducing loyalty programs. Future research can improve these systems by including real-time customer behavior, customer feedback, social media information, and more advanced machine learning methods. Continuous monitoring and regular model updates will also be important so that predictions remain useful as customer behavior and market conditions change.

References

[1] J. Barach, "Federated Learning for Privacy-Preserving Employee Performance Analytics," *IEEE Access,* vol. 13, 2025, doi: 10.1109/ACCESS.2025.3591360.

[2] S. Malik, S. Harous, and H. El-Sayed, "Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women," in *International Symposium on Modelling and Implementation of Complex Systems*, 2020: Springer, pp. 95-106.

[3] A. M. Qatawneh, F. M. Aldhmour, and S. M. Alfugara, "The adoption of electronic payment system (EPS) in Jordan: case study of orange telecommunication company," *Journal of Business and Management,* vol. 6, no. 22, pp. 139-148, 2015.

[4] A. Mohammed, "Artificial Intelligence-Driven Cybersecurity Crimes: Emerging Threats, AI Bots, and Advanced Defense Strategies," *Euro Vantage journals of Artificial intelligence,* vol. 3, no. 1, pp. 46-65, 2026, doi: https://doi.org/10.65923/c2sxrx21.

[5] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing price prediction via improved machine learning techniques," *Procedia Computer Science,* vol. 174, pp. 433-442, 2020.

[6] R. Sailusha, V. Gnaneswar, R. Ramesh, and G. R. Rao, "Credit card fraud detection using machine learning," in *2020 4th international conference on intelligent computing and control systems (ICICCS)*, 2020: IEEE, pp. 1264-1270.

[7] A. Mohammed, "Transforming SOC Operations: Harnessing the Power of AI and ML for Enhanced Threat Detection," *INTERNATIONAL JOURNAL OF RESEARCH CULTURE SOCIETY Monthly Peer-Reviewed, Refereed, Indexed,* vol. 8, pp. 11-21, 2024, doi: 10.2017/IJRCS/ICCDMP01.

[8] S. Ara, A. Das, and A. Dey, "Malignant and benign breast cancer classification using machine learning algorithms," in *2021 international conference on artificial intelligence (ICAI)*, 2021: IEEE, pp. 97-101.

[9] S. Khairnar "EDGE COMPUTING FOR IOT DEVICES: A COMPREHENSIVE FRAMEWORK FOR DISTRIBUTED DATA PROCESSING AND REALTIME ANALYTICS," *Journal of Integrated Research,* vol. 2, no. 1, 2021.

[10] A. M. Qatawneh and A. M. Alqtish, "Critical examination of the impact accounting ethics and creative accounting on the financial statements," *International Business Research,* vol. 10, no. 6, p. 104, 2017, doi: 10.5539/ibr.v10n6p104.

[11] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," in *2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT)*, 2018: IEEE, pp. 1-4.

[12] J. Barach, "Enhancing Ransomware Resilience in Cloud-Based HR Systems through Moving Target Defense," *Computers, Materials and Continua,* vol. 86, no. 2, pp. 1-23, 2025, doi: https://doi.org/10.32604/cmc.2025.071705.

[13] S. Khairnar, "EXPLORING CORPORATE CLOUD ADOPTION: A COMPREHENSIVE MULTI-FACTOR EVALUATION," *International Journal of Data Science and IoT Management System,* vol. 1, no. 3, pp. 35-50, 2022.

[14] M. Beseiso, "Enhancing student success prediction: A comparative analysis of machine learning technique," *TechTrends,* vol. 69, no. 2, pp. 372-384, 2025.

[15] E. Ahmed, "Student performance prediction using machine learning algorithms," *Applied computational intelligence and soft computing,* vol. 2024, no. 1, p. 4067721, 2024.