Dean&Francis

ISSN 2959-6157

Machine Learning Models for Customer Churn Prediction: Current Progresses and Challenges in the Telecommunications and SaaS Industries

Ruiying Zhu

Warrington Business School,
University of Florida, Florida,
United States

ruiying.zhu1@gmail.com

Abstract:

With the ongoing economic growth and the rise of new companies, customer churn has become a renewed focus for many businesses, particularly in the telecommunications and SaaS sectors. Accurately identifying lost customers using artificial intelligence technology is crucial for helping companies develop effective strategies to retain these valuable clients. This paper provides a comprehensive overview of recent research on the application of various machine learning models, including logistic regression, random forests, decision trees, XGBoost, and advanced deep learning models like Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN), to different types of datasets in the telecommunications and SaaS sectors. The study examines the workflow of each model, offering an in-depth analysis of their effectiveness and performance in predicting customer churn. In addition to evaluating the strengths and weaknesses of these models, the paper also addresses key challenges in machine learning, such as interpretability, applicability, and privacy concerns. It highlights current solutions employed by researchers, including SHapley Additive exPlanations (SHAP) for model interpretability, transfer learning, domain adaptation, and federated learning, to overcome these challenges. Furthermore, this paper emphasizes the ongoing need for further research to enhance the robustness and practical deployment of churn prediction models in these industries. In conclusion, this paper offers a comprehensive overview of current research on customer churn prediction models, highlighting the application of various machine learning techniques and addressing key challenges and solutions within the telecommunications and SaaS sectors.

Keywords: Customer churn prediction; machine learning models; telecommunication; Software as a Service

1. Introduction

Customer Churn refers to a phenomenon in which customers end their relationship with a company, either by canceling subscriptions, shifting to other services or products, or moving to a new competitor. This phenomenon can be ongoing or may have already occurred. In many industries with highly competitive pressure and market dynamics, retaining customers is challenging even for large companies. For instance, one of the largest e-commerce companies, Amazon, lost 2.6 million active users with the rise of competitors Temu and Shein in 2023 [1]. A significant portion of a company's revenue for most businesses comes from their existing customers. As a result, many companies have increasingly focused their marketing efforts on customer retention rather than solely on acquiring new customers in recent years [2]. Accurately identifying customers who have been lost plays a crucial role in helping companies refine and develop policies aimed at retaining these valuable customers. Artificial intelligence possesses advanced data handling capability that allow computers to analyze company's historical data and customer behavior, enabling accurate prediction of customer churn.

With the rise of artificial intelligence and its great progress, both traditional machine learning and deep learning technologies have been widely used in many fields including healthcare, education, finance, etc., especially in the field of business analysis, people have carried out a lot of work. For instance, Li et al. applied a new K-means based Adaptive Learning Particle Swarm Optimization (KM-ALPSO) algorithm in order to avert the dependence of k-means on the initial centers to improve prediction in the customer segmentation area [3]. Furthermore, Murugan et al. examined three models — K-Nearest Neighbor (KNN), cluster based Logistic Regression (LR), and cluster based XG boost — to assess their effectiveness in predicting loan defaults and the likelihood of occurrence in the field of financial risk management [4]. Apart from these popular areas, customer churn prediction remains a key focus for data analysts and has been the subject of extensive research, with numerous studies and investigations conducted in this field. Numerous machine learning models have been utilized to identify the most accurate predictors of customer churn. The emergence of new deep learning models has expanded the possibilities for improving prediction accuracy. The effectiveness of these models may vary depending on the region in which they are applied. For example, in the telecommunication sector, Lalwani et al. reported that Adaboost and XGboost Classifier achieved the highest accuracy of 81.71% on their dataset following advanced data engineering approaches [5]. Kolomiiets et al. reported that a deep neural network with 32 hidden layers demonstrated high accuracy in predicting outcomes for B2B software subscriptions within IT companies in the Software-as-a-Service (SaaS) sector [6]. Given the significance of customer churn and the continuous evolution of algorithms developed to address customer churn predictions, it is essential to provide a comprehensive overview of this area.

The remainder of this paper is structured as follows: the method section will detail how artificial intelligence algorithms are applied to address customer churn prediction issues across various fields, especially in the fields of telecommunications, SaaS and E-commerce, along with an in-depth look at the specific algorithms proposed. The discussion section will analyze the strengths and weaknesses of these algorithms, explore recent challenges in customer prediction models, and suggest future research directions. Finally, the conclusion will summarize the key points of the paper.

2. Method

2.1 Introduction of Machine Learning Workflow

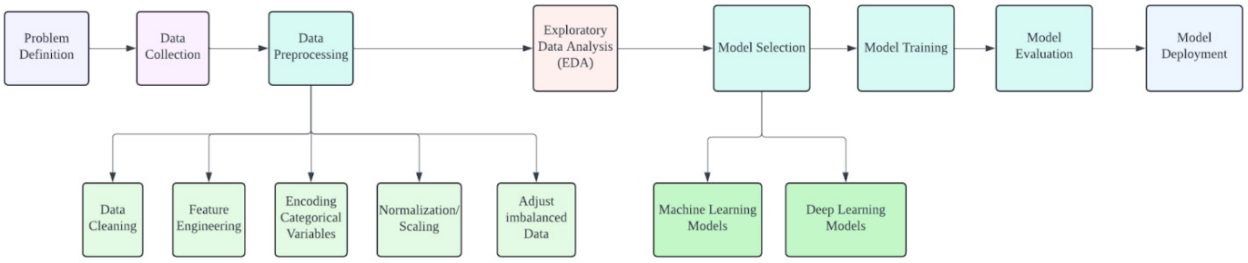

Fig. 1 The Schematic of Machine Learning Workflow (Photo/Picture credit: Original)

Predicting customer churn using Machine Learning (ML) methods typically involves several critical steps, from problem definition to model deployment. Fig. 1 is included to illustrate this machine learning workflow. The goal is to identify early indicators suggesting a customer's intention to terminate their relationship with a company [7]. In industries such as SaaS, telecommunications, and e-commerce, data collection includes historical customer information, including demographics, interactions, payments, and usage. The preprocessing phase involves data cleaning, feature engineering, encoding categorical variables, and normalization, with techniques such as Synthetic Minority Oversampling Technique (SMOTE) or class weighting used to address class imbalance. Exploratory Data Analysis (EDA) is essential for visualizing features, selecting significant variables, and choosing appropriate machine learning or deep learning models based on the dataset's characteristics. The model is then trained and tested, and evaluated against a threshold to ensure its suitability for deployment. Generally, the metrics to evaluate the models are F1-Score, Precision, Recall, Accuracy and Area Under Curve (AUC). Due to variations in data reconstruction across these industries, the applied models and their predictions can produce different results. The following sections will introduce how different models are applied and performed in the SaaS and telecommunications fields.

2.2 Telecommunication

2.2.1 Logistic regression

In machine learning, customer churn is typically approached as a binary classification task, with logistic regression frequently employed to model the probability of churn based on various features [7]. For instance, Melian et al. [8] employed K-means clustering to segment customers on a dataset of a major telecommunication company in Romania and then applied a logistic regression model within each cluster. They further refined the model by selecting features with a p-value less than 0.05. To assess the model's performance in each cluster, they utilized AUC to evaluate the model's accuracy rates in customer prediction.

2.2.2 Decision tree

Melian et al. utilized decision tree classifiers across three customer groups to identify key factors driving churn. The analysis revealed that the number of months was the primary decision point for predicting churn. They assessed the model's performance using AUC [8].

Similarly, Wagh et al. used decision tree classifiers on customer data to identify those likely to churn. Initially, the data was split into training and testing sets. Due to the imbalance between churn and non-churn customers, the training data was up-sampled using SMOTE and cleaned with Edited Nearest Neighbor (ENN) to achieve a balanced dataset. The Decision Tree classifier, applied after this balancing process, achieved significantly improved performance [9], a notable enhancement compared to the model's performance on the original unbalanced dataset.

2.2.3 Random forest algorithm

The Random Forest classifier extends the Decision Tree algorithm by combining multiple trees into an ensemble to boost performance. Wagh et al. [9] applied the random forest algorithm, following similar data preprocessing steps as the decision tree.

Melian et al. [8] used the Random Forest model to predict customer churn and identify key predictive features, splitting the dataset 60/40 for training and testing. The model, configured with 500 trees, performed well on the training set, but its accuracy and sensitivity (Recall) declined on the test set, likely due to data imbalance. To address this, Melian et al. [8] applied the Balanced Random Forest (BRF) algorithm, which is tailored to handle imbalanced datasets. While the overall accuracy decreased compared to the standard Random Forest, the model's sensitivity (Recall) improved significantly.

2.2.4 XGBoost

In the research by Shrestha et al. [10], a comparative approach using the XGBoost algorithm was applied to address dataset imbalance in customer churn prediction in telecommunications. Two datasets—one public and one from Nepal—were used. Shrestha et al. illustrated the overall process in Fig. 2.

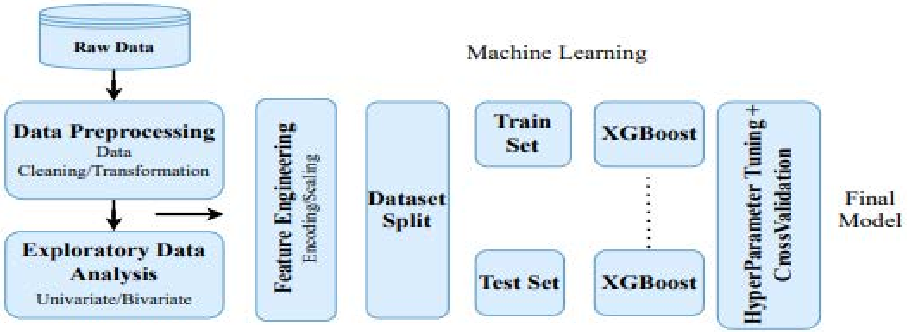

Fig. 2 The Framework of the XGBoost Model proposed by Shretha et al. [10]

After completing data preprocessing and EDA, feature engineering steps, such as encoding categorical variables and scaling numerical features, were implemented to prepare the dataset for the XGBoost model. The dataset was then split into training and testing sets in an 80:20 ratio.

To optimize the model's performance, hyperparameter tuning was conducted using GridSearch Cross-Validation (CV), a method that systematically searches for the best combination of hyperparameters. The model's performance was further validated through 10-fold StratifiedK-Fold Cross Validation, ensuring that each fold maintained the original distribution of the target variable. The F1-score and accuracy metrics were used to evaluate the model's effectiveness on both the training and testing datasets.

2.2.5 Deep learning Models

In recent years, deep learning models like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Deep Neural Networks (DNN) have gained popularity in customer churn prediction, particularly for managing imbalanced data. Almufadi et al. used a combination of under-sampling, oversampling with SMOTE, one-hot encoding, and data normalization for preprocessing. After hyperparameter tuning, they assessed the performance of three models using accuracy, F1-score, precision, and recall metrics [11].

Building on existing techniques, Fujo et al. [12] proposed the Deep-BP-ANN model for telecommunication datasets, as illustrated in Fig. 3. They used Lasso Regression and variance thresholding for feature selection and random oversampling for dataset balancing. Evaluated using holdout and 10-fold Cross Validation, the Deep-BP-ANN model was assessed by five metrics: F1-score, precision, recall, accuracy and AUC.

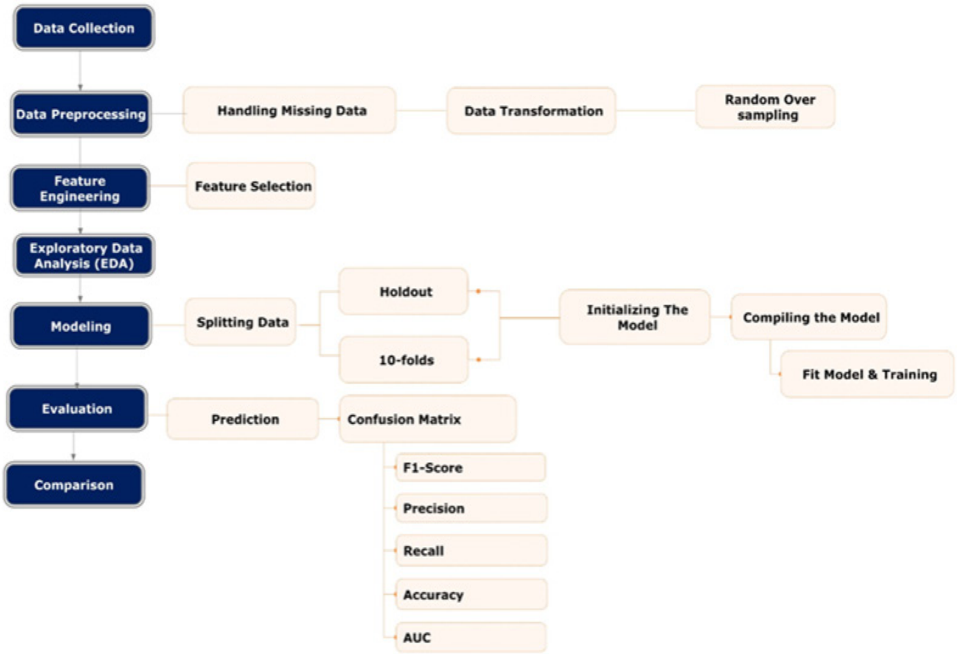

Fig. 3 The Schematic of Deep-BP-ANN workflow proposed by Fujo et al. [12]

2.3 Software as a Service (SaaS)

2.3.1 Logistic regression

In the SaaS field, the target variable in the Logistic Regression model is analogous to that in the Telecommunications field. Dias et al. [2] used a Logistic Regression model as a baseline for comparing customer churn prediction in the SaaS industry. The dataset underwent preprocessing similar to approaches in the telecommunications field. Implemented using Scikit-learn, the model's hyperparameters were fine-tuned with GridSearchCV. A moderate regularization (C = 1) ensured good generalization, while Lasso regularization highlighted key features. The 'liblinear' solver was chosen for its efficiency with small to medium-sized datasets, making the model practical for SaaS churn prediction.

However, the hyperparameter tuning process is not commonly used in the Logistic Regression model. Calli et al. [13] employed various feature selection techniques, including Chi-square (for categorical variables), information gain, gain ratio, and the Gini index, instead of focusing on hyperparameter tuning for customer churn prediction on

SaaS as a business to business (B2B) industry.

2.3.2 Random forest

The Random Forest algorithm is frequently used to identify the key factors influencing customer churn.Dias et al. [2] applied Random Forest due to its ability to handle small datasets and high-dimensional features in the SaaS industry. The model ranked feature importance, optimizing performance through hyperparameter tuning. Effectiveness was demonstrated using accuracy on a preprocessed dataset from a Portuguese software house.

Calli et al. [13] applied a Random Forest Model to predict customer churn for an ERP company, focusing on important features. Instead of hyperparameter tuning, they employed four feature selection methods: Chi-Square, Gini Index, Information Gain, and Gain Ratio, ultimately selecting ten features for analysis. They also normalized the data to reduce variations.

2.3.3 Boosting models

In SaaS applications, boosting models like AdaBoost, GBM, and XGBoost are widely used for predicting customer application status. AdaBoost adjusts weights to enhance weak learners, while GBM focuses on reducing errors. XGBoost improves on these by incorporating parallelization and regularization for better performance. Dias et al. [2] applied these models, using GridSearchCV for hyperparameter tuning, and combined features from customer application status and related issues after removing irrelevant ones. They focused on a categorical target variable, evaluating model performance with Recall to identify potential churn.

Similarly, Dahlen et al. [14] explored churn factors in the B2B SaaS sector over 4 and 8 months using boosting models. They employed comprehensive data handling techniques, including cleaning, filtering, normalization, splitting, and balancing. Hyperparameters were tuned with GridSearchCV, and feature selection was done via RFECV. They tested XGBoost and LightGBM, using AUC-ROC for evaluation, and employed SHapley Additive exPlanations (SHAP) to explain model outputs.

2.3.4 Deep learning models

Deep learning methods have become increasingly popular for customer churn prediction in the SaaS sector due to their ability to identify complex patterns within large, diverse datasets. For instance, Kolomiiets et al. [6] utilized a Deep Neural Network, which consists of multiple layers, to predict customer churn in B2B Software by Subscription. Each layer of neurons processes the output from the previous layer, forming a hierarchy that enhances abstraction and enables the testing of complex hypotheses. They employed scikit-learn and TensorFlow for their deep learning implementation, creating a two-dimensional tensor and normalizing the data, with missing values set to 0. After preprocessing, the dataset was split into training and testing sets. The ReLU activation function was chosen for its superior performance in preparing data for deep networks, compared to logistic and hyperbolic tangent functions. The model was trained using the Adam optimizer, with binary cross-entropy used to measure error, and accuracy was the primary metric for evaluation.

3. Discussion

In terms of the advantages and disadvantages of these models mentioned above, logistic regression model is commonly used as a baseline in customer churn prediction due to its simplicity and effectiveness in binary classification. However, it often underperforms because it assumes a linear relationship between features and the target and struggles with imbalanced datasets. Decision trees offer simplicity and handle non-linear relationships but can overfit and become biased towards the majority class in imbalanced datasets. Random forests, which combine multiple decision trees, reduce overfitting and are more robust, yet are complex and harder to interpret. Boosting models enhance predictive accuracy by iteratively improving weak learners but are complex and prone to overfitting without careful tuning. Deep learning models excel in capturing complex patterns but suffer from interpretability issues and require extensive tuning to avoid overfitting.

Although many progresses have been achieved in this field, there are still some challenges. Interpretability remains one of the key challenges in the field of Customer Churn Prediction, particularly because many of the models with the highest accuracy—such as ensemble machine learning techniques or deep learning methods—are inherently complex and require extensive tuning. Current research often prioritizes identifying the most accurate models rather than focusing on interpreting their outcomes.

In such situations, some promising methods can be considered for enhancing model interpretability. One widely used approach is SHAP, which is model-agnostic and provides transparent explanations for model predictions. For example, Noviandy et al. explored this method, using SHAP to generate graphs that clearly identify which features contribute the most to predictions. Additionally, Laiwani et al. [5] introduced an expert system to further enhance the interpretability and effectiveness of machine learning models. This expert system helps guide the feature selection process by leveraging domain knowledge to pinpoint significant features, thereby making the model's outcomes more understandable and actionable.

Applicability issues also remain a significant challenge in customer churn prediction, primarily due to difficulties in generalizing models across different domains, maintaining model interpretability and relevance, and overcoming technical and resource constraints that may hinder broad deployment. For example, a random forest classifier might perform exceptionally well on a particular dataset after extensive preprocessing, but its effectiveness could diminish when applied to other datasets due to differences in data characteristics.

To address these applicability issues, current research increasingly employs domain adaptation techniques. For instance, Noviandy et al. [15] utilized transfer learning to adapt a churn prediction model from one geographic region to another, accommodating the distinct customer behaviors in different areas. Similarly, Fujo et al. [12] acknowledged the significant variation in data distribution across regions and customer segments. They applied transfer learning to allow deep learning models to adjust to new data from different regions by leveraging pre-trained models on one dataset and fine-tuning them with a smaller dataset from a new domain. This approach ensures that the models can be effectively applied across different markets.

Privacy is another significant concern in customer churn prediction, especially due to the sensitive nature of the data involved, such as personal information, behavioral data, and purchase history. While techniques like data anonymization and pseudonymization (where data is replaced with artificial identifiers) are commonly employed to protect customer information, several privacy issues still persist, including data security, data sharing, and potential biases. To address these challenges, federated learning emerges as a highly effective solution, as it keeps data on local devices and only shares model updates with a central server, thereby safeguarding the raw data. However, further research is needed to explore and enhance the capabilities of federated learning in this context.

Despite significant efforts by many researchers to address these challenges, progress remains limited. In the future, further research will explore this field, including the development of new machine learning models that can better tackle these challenges.

4. Conclusion

This paper thoroughly examines customer churn prediction model algorithms in the telecommunications and SaaS industries, focusing on recent developments. It discusses key models, including logistic regression, decision trees, random forests, XGBoost, and advanced deep learning techniques such as DNN, CNN, and DEEP-BP-ANN. The study evaluates the effectiveness and performance of these models by analyzing their application across various datasets, data preprocessing methods, and the overall process of how each model is applied to the data. Through comparative analysis, the paper highlights the strengths and weaknesses of each model in the context of customer churn prediction.

Additionally, the paper identifies three major challenges facing current churn prediction models: interpretability, applicability, and security. To address these issues, it explores advanced techniques like SHAP for enhancing model interpretability, transfer learning and domain adaptation for improving applicability, and federated learning for bolstering security and privacy. The paper emphasizes the need for further research, particularly in model deployment, to better address these challenges and improve the practical implementation of customer churn prediction models in these industries.

References

[1] Cohan P. Down by 26 million users, Amazon could keep losing customers to Temu. Forbes [Internet]. 2024 Mar 21.

[2] Dias JR, Antonio N. Predicting customer churn using machine learning: A case study in the software industry. Journal of Marketing Analytics. 2023 Dec 2:1-7.

[3] Li Y, Chu X, Tian D, Feng J, Mu W. Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. Applied Soft Computing. 2021 Dec 1;113:107924.

[4] Murugan MS. Large-scale data-driven financial risk management & analysis using machine learning strategies. Measurement: Sensors. 2023 Jun 1;27:100756.

[5] Lalwani P, Mishra MK, Chadha JS, Sethi P. Customer churn prediction system: a machine learning approach. Computing. 2022 Feb;104(2):271-94.

[6] Kolomiiets A, Mezentseva O, Kolesnikova K. Customer churn prediction in the software by subscription models it business using machine learning methods. InCEUR Workshop Proc 2021 (Vol. 3039, pp. 119-128).

[7] Dias JR, Antonio N. Predicting customer churn using machine learning: A case study in the software industry. Journal of Marketing Analytics. 2023 Dec 2:1-7.

[8] Melian DM, Dumitrache A, Stancu S, Nastu A. Customer churn prediction in telecommunication industry. A data analysis techniques approach. Postmodern Openings. 2022 Mar 14;13(1 Sup1):78-104.

[9] Wagh SK, Andhale AA, Wagh KS, Pansare JR, Ambadekar SP, Gawande SH. Customer churn prediction in telecom sector using machine learning techniques. Results in Control and Optimization. 2024 Mar 1;14:100342.

[10] Shrestha SM, Shakya A. A customer churn prediction model using XGBoost for the telecommunication industry in Nepal.

Procedia Computer Science. 2022 Jan 1;215:652-61.

[11] Almufadi N, Qamar AM. Deep Convolutional Neural Network Based Churn Prediction for Telecommunication Industry. Comput. Syst. Sci. Eng.. 2022 Jan 1;43(3):1255-70.

[12] Fujo SW, Subramanian S, Khder MA. Customer churn prediction in telecommunication industry using deep learning. Information Sciences Letters. 2022 Jan;11(1):24.

[13] Çallı L, Kasım S. Using Machine Learning Algorithms to Analyze Customer Churn in the Software as a Service (SaaS) Industry. Academic Platform Journal of Engineering and Smart Systems. 2022;10(3):115-23.

[14] Dahlén D, Mauritzon W. Machine Learning-based Prediction of Customer Churn in SaaS.

[15] Noviandy TR, Idroes GM, Hardi I, Afjal M, Ray S. A Model-Agnostic Interpretability Approach to Predicting Customer Churn in the Telecommunications Industry. Infolitika Journal of Data Science. 2024 May 27;2(1):34-44.