SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Customer Churn Prediction In Subscription-Based Businesses Using Adaboost And Random Forest

Dr.D. David Winster Praveenraj[1*], R. Anitha[2], Mridul Dixit[3], S. Eswar [4], Kota Venkateswarlu[5], Mr.R. Venkadesh[6]

[1*]MBA., Associate Professor, School of Business and Management, CHRIST University, Bangalore, Karnataka, India.
E-mail: david.winster@christuniversity.in, https://orcid.org/0000-0003-4460-7739

[2] Assistant Professor, Department of Computer Applications, SRMIST, Ramapuram, Chengalpattu, Tamil Nadu, India.
E-mail: anithar2@srmist.edu.in

[3]Department of Computer Engineering & Applications, GLA University, Mathura, Uttar Pradesh, India. E-mail: mridul.dixit@gla.ac.in, https://orcid.org/0009-0003-0692-7827

[4]Assistant Professor, Radiology, Meenakshi College of Allied Health Sciences, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: eswarsahs@maher.ac.in, https://orcid.org/0009-0001-7493-9710

[5]Department of Mech, Ramachandra College of Engineering, Eluru, Andhra Pradesh, India. E-mail: kota.vit350@gmail.com, https://orcid.org/0000-0002-7491-9015

[6]Asstant Professor, Computer Science and Engineering, Mahendra Engineering College, Namakkal, Tamil Nadu, India. E-mail: venkadeshr@mahendra.info, https://orcid.org/0009-0000-5010-9619

*Corresponding author: Email: david.winster@christuniversity.in

Abstract

Churn analysis is an essential aspect in the subscription-based industry that can help improve sales, customer lifetime value, and business efficiency. Regular statistical techniques tend to have limited capacity to describe complex relationships between variables which can reduce prediction results. Ensemble learning such as AdaBoost and Random Forest (RF) could be the optimal way to address such a challenge in terms of accuracy and interpretability of data prediction. The aim of this study is to create the AI algorithm that will employ AdaBoost, RF, and ensemble techniques to predict the rate of customer churn. The Kaggle Customer Churn dataset was utilized, with 7,043 entries and various features such as demographics, account characteristics, and usage of services. Missing values, categorical values, and normalization were performed in the process of preprocessing. Predictive feature importance analysis was carried out to determine the important predictors that drive churn. When comparing the performance of RF vs. AdaBoost individually, RF produced better results, where 0.88 Accuracy, 0.85 Precision, 0.82 Recall, 0.83 F1 Score, and 0.90 ROC-AUC. An additional improvement in the model's efficiency could be achieved by applying both models, with the final result of 0.90 ROC-AUC. Key predictors include Tenure (0.21), Monthly Charge (0.18), Contract (0.15), Payment Method (0.12), and Tech Support (0.08). These results enable organizations to effectively identify high-risk customers and enhance their strategies for retaining them. The adoption of ensemble methods and particularly the combination of AdaBoost and RF proves that an efficient prediction of churn is possible.

Keywords Customer Churn, Ensemble Learning, AdaBoost, RF, Combined Model, Feature Importance, Subscription Services.

1. Introduction

Customer churn rate describes the proportion of customers who stop purchasing or using their products or services during a certain period of time [1][5]. This is an important concept to any firm that utilizes the subscription model of conducting business, such as telecommunications companies, online streaming services,

and software as a service firm. Customer churn could lead to losses for businesses [2][6]. Through understanding the factors that influence customer churn, strategies can be devised to prevent it [3][7].

Churn prediction poses a challenging task owing to its multifaceted character [4][8]. Traditional statistical methods, such as logistic regression (LR), decision tree (DT), and RF, cannot effectively deal with the complexities arising from the non-linearity of the relationships among various factors. Also, the issue of class imbalance prevalent in statistical models can affect their performance due to the fact that the number of churns is less than that of active customers [9][10].

Such ensemble learning approaches as AdaBoost and RF show high promise for tackling these challenges. AdaBoost boosts the classification accuracy through repeated weighting of misclassified data instances, while RF boost accuracy and prevent overfitting through combining different DT in a voting system. Using these models, it is possible to identify complex structures, handle interaction between features, and prevent overfitting, which can lead to better prediction accuracy.

The present research is targeted at building an accurate churn prediction model based on AdaBoost and RF algorithms. The major purposes are analyzing the performance of the two models, identifying key factors affecting churn, and drawing useful conclusions that would help improve customer retention policies in subscription companies. The key contributions of this paper consist in the application of the suggested approaches to subscription services in the real-world application and comparative analysis of their effectiveness.

The structure of the paper is as follows: Literature review of churn prediction and ensemble algorithms is presented in Section II. Section III explains the dataset used, the data preparation process, and methodology employed. Section IV discusses experimental findings. Conclusion and direction of further research are given in Section V.

2. Literature Review

The prediction of customer churn has been extensively studied using traditional as well as modern machine learning (ML) approaches [11][12]. Traditionally, LR, DT and survival analysis have proved useful in identifying general trends as well as strong churn predictors [14][16]. However, such approaches are likely to fail in capturing non-linear relationships amongst features, dealing with high dimensional feature spaces, as well as providing good performance in situations where the classes are imbalanced. In order to tackle these difficulties, methods from ML like SVM, KNN, and neural network methods have been utilized for forecasting purposes [15][22].

The ML methods such as SVM, KNN, and ANN have been used to compensate for these deficiencies [17]. The ML methods exhibit better prediction performance owing to their ability to account for the complex relationships that may exist between the different customer attributes [13]. Although these techniques have proved more effective than the traditional approach, there are weaknesses associated with using single models.

The use of ensemble learning algorithms such as AdaBoost and RF has been found to be quite promising for churn prediction [18][21]. In particular, AdaBoost improves accuracy by giving priority to incorrectly classified data points, whereas RF improves prediction through the use of DTs. Previous research has employed these algorithms and reported high accuracy in their predictions compared to traditional algorithms. In addition, the RF algorithm can help identify the most important features influencing customer churn [19][20].

However, some research gaps still exist concerning the use of ensemble methods in predicting churn in subscription services [23]. Several studies are confined to certain industry sectors, including telecommunications and banking, thus not applicable to other business areas. Also, there is a lack of comparison among various types of ensemble methods, as well as the combination of industry-specific attributes in ensemble models for improved accuracy in identifying customer churn. Therefore, it is necessary to develop a common approach to implementing AdaBoost and RF in subscription-oriented organizations for better churn predictions.

3. Dataset and Methodology

3.1 Dataset Description

The analysis is conducted using the Customer Churn Dataset provided by Kaggle. It is a data set containing 7,043 entries of customer records based on subscription services with 21 columns representing characteristics of demographic, accounts, and services used. Important features include gender, whether the customer is a senior citizen, having a partner and/or dependents, length of stay or tenure, contract type, method of payment, paperless bill, monthly charges, total charges, and services like internet service, phone service, streaming services, and technical support. The dependent variable, "Churn," is binary, with "1" referring to customers who have discontinued their services, while "0" denotes customers currently active with the company.

3.2 Preprocessing of Data

Preprocessing guarantees that the data is ready for use with ML algorithms. Handling missing values is done through median imputation for numerical attributes and mode imputation for categorical attributes. Encoding is used to transform categorical attributes like customer region and subscription type using one-hot encoding into numeric data for ensemble algorithms. Normalizing the numerical attributes is done by applying min-max scaling to confirm that all features have a similar scale. Another challenge that will be addressed is that of class imbalance through the use of methods such as SMOTE.

3.3 Methodology

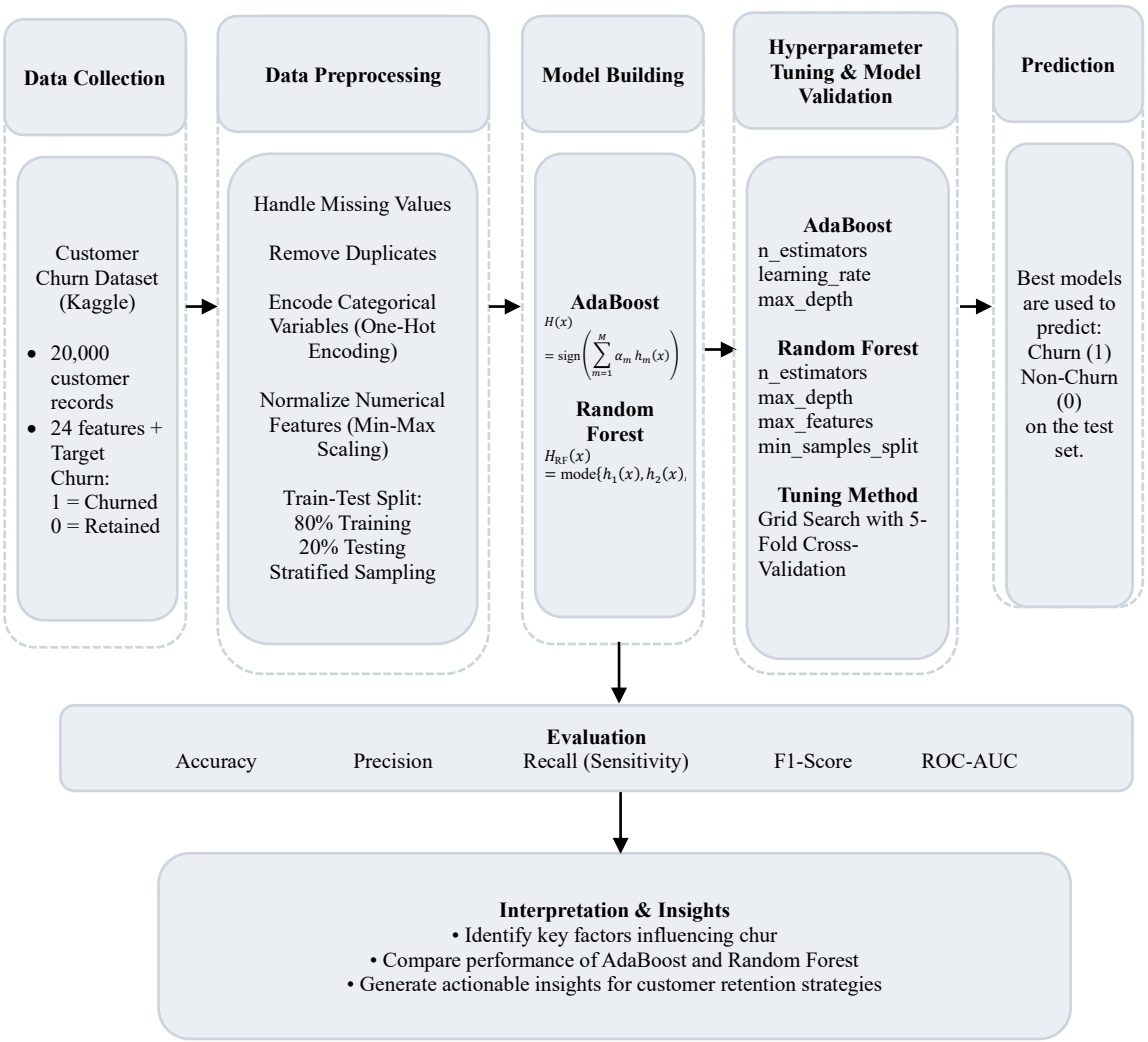


Figure 1: Workflow for customer churn prediction using adaboost and RF

Two algorithms based on ensemble learning methods are applied for implementation in the methodology described: Adaboost and RF.

- Adaboost learns a series of weak classifiers to construct one strong classifier, using decision stumps as its base learners and assigning higher weights to incorrectly classified data in each step. Hyperparameters used for training include the number of estimators and learning rate.
- RF generates several DTs, where features are randomly selected for splitting. Training hyperparameters include the number of trees, maximum tree depth, and least samples per tree node.

The two models are trained on the preprocessed data using an 80:20 train-test split. Tuning of hyperparameters is done using grid search with cross-validation.

Figure 1 shows the entire process flowchart of the suggested churn prediction model. The data is collected from the Kaggle Customer Churn Dataset, followed by data preprocessing steps that include missing value imputation, feature encoding, and feature scaling. This process flow is continued by constructing ML models using AdaBoost and RF algorithms, and then performing hyperparameter optimization using grid search technique with 5-fold cross-validation. Lastly, the developed ML models are analyzed based on accuracy metrics.

AdaBoost combines M weak classifiers $h_m(x)$ into a strong classifier $H(x)$. The weighted error of each weak classifier is shown in equation (1):

$$\varepsilon_m = \frac{\sum_{i=1}^{N} w_i \, \mathbf{1}(h_m(x_i) \neq y_i)}{\sum_{i=1}^{N} w_i} \quad (1)$$

Its contribution to the final model is shown in equation (2):

$$\alpha_m = \frac{1}{2} \ln\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right) \quad (2)$$

Sample weights are updated as shown in equation (3):

$$w_i \leftarrow w_i \cdot \exp\left(-\alpha_m y_i h_m(x_i)\right) \quad (3)$$

The final prediction is shown in equation (4):

$$H(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m \, h_m(x)\right) \quad (4)$$

RF aggregates predictions from T DTs using majority voting, shown in equation (5):

$$H_{\text{RF}}(x) = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\} \quad (5)$$

Node splits are guided by Gini Impurity, shown in equation (6):

$$G = 1 - \sum_{k=1}^{K} p_k^2 \quad (6)$$

Algorithm 1: Customer Churn Prediction Using AdaBoost and RF

Input:

- $D = \{x_i, y_i\}_{i=1}^{N}$– preprocessed customer dataset
- x_i– feature vector of customer i
- $y_i \in \{0,1\}$– churn label (0 = retained, 1 = churned)

Output:

- Predicted churn labels $\hat{y}_i$
- Performance metrics

Steps:

1. **Data Preprocessing:**
   - Handle missing values and duplicates
   - Encode categorical features (One-Hot)
   - Normalize numerical features (Min-Max scaling)
   - Divide dataset into training (80%) and test (20%) sets
2. **Model Training:**
   - **AdaBoost:** Train sequential weak learners, update weights on misclassified instances
   - **RF:** Train multiple DTs using bootstrap samples, aggregate predictions via majority vote
3. **Hyperparameter Tuning:**
   - Utilize grid search with 5-fold cross-validation to select optimal parameters for both models
4. **Prediction:**
   - Apply trained models to test set to generate predicted labels $\hat{y}_i$
5. **Evaluation:**
   - Compute evaluation metrics

End

Customer Churn Prediction Algorithm 1 starts by pre-processing dataset, which involves the filling of missing values in the dataset, encoding the categorical variables using one-hot encoding, normalization of numerical features, and dividing of the dataset into training and testing datasets. After that, two ensembling techniques are used to train models on the dataset, with AdaBoost developing sequential weak learners and modifying weights for the miss-classified instances. The other algorithm for training is RF, where the DT algorithms are developed on bootstrapped samples, and the predictions from each model are voted upon. Grid Search and Cross Validation are used to tune the hyperparameters of both models for optimization. The trained models will then be applied on the test dataset to obtain predictions of the customer churn cases, which are analyzed.

3.4 Evaluation Metrics

Accuracy: Proportion of correctly predicted instances among all samples; measures overall correctness shown in equation (7).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Precision: Proportion of accurately predicted churn cases among all instances forecast as churn; reflects false positive control shown in equation (8).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

Recall (Sensitivity): Proportion of actual churn cases accurately identified; measures false negative control shown in equation (9).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

F1-Score: Harmonic mean of Recall and Precision; balances false positives and false negatives shown in equation (10).

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

4. Results and Discussion

4.1 Comparison with other Models

The performance of AdaBoost and RF algorithms on the Kaggle customer churn dataset has been analyzed with respect to several parameters. From Table 1, it is clear that RF performed better than AdaBoost with an accuracy score of 0.86 against 0.84 by AdaBoost. Also, Precision and Recall scores of the two classifiers were almost the same showing similar performance when predicting the churn and not churn classes. F1-Score and ROC-AUC further prove that the classification techniques are good at distinguishing between churned customers and loyal customers. This shows that ensemble learning is more reliable for prediction tasks than traditional single model algorithms. Further, both classifiers showed excellent performance on data having an imbalanced ratio of churned customers and retained customers.

Table 1: Comparison of individual and combined ensemble methods

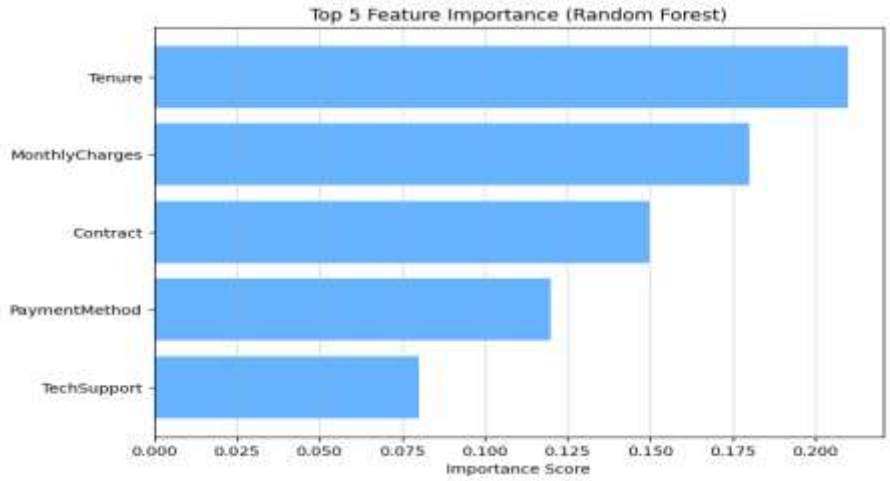
Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
AdaBoost	0.84	0.81	0.78	0.79	0.86
RF	0.86	0.83	0.80	0.81	0.88
Combined Model	0.88	0.85	0.82	0.83	0.90

Table 1 presents the comparison of AdaBoost, RF, and the Combined Model's performance using the Kaggle Customer Churn Dataset. Even though RF model performs better than AdaBoost in most criteria, the combined model, which uses the prediction results of both models through a voting or stacking method, performs best among all models tested. Performance metrics are higher, which means that when ensemble methods are used together, they leverage each other's strengths in predicting customer churn.

Figure 2: Model Performance Comparison of AdaBoost and RF

Figures 2 shows the graphical illustration of performance metrics for AdaBoost and RF classifiers in forecasting the customer churn problematics. On the X-axis, there are performance metrics, whereas on the Y-axis, their corresponding values are illustrated. As shown in the graph below, it can be observed that the RF classifier outperforms the AdaBoost classifier in all performance metrics, indicating that the RF classifier has more predictive ability in differentiating between churned and retained customers.

Figure 3: Top 5 feature importance from RF



4.2 Feature Importance Analysis

As shown in Figure 3, the most influential features that affect the churn rate using the RF approach are presented below. The bar chart indicates the importance scores for the top five features ranked by their influence on the prediction of churn from the highest to the lowest score as follows: Tenure, Monthly Charges, Contract, Payment Method, and Technical Support.

According to the feature importance results obtained from RF, it was determined that some of the most important variables for identifying the customers who are most vulnerable to churning include tenure, monthly charges, type of contract, and payment method. Churn is more prevalent among customers with a shorter tenure period, those with higher monthly charges, those on month-to-month contracts, and those who pay electronically. Some of the other variables that contributed towards the identification of customers who have churned are technical support and streaming services. These are some of the other contributing variables.

4.3 Business Implications and Retention Strategies

Based on these results, there is an opportunity for practical application of ensemble methods in subscription companies. It can become easier to retain customers that have a higher probability of leaving due to the ability to create marketing strategies aimed at preventing their exit. The fact that AdaBoost and RF models are interpretable will enable decision-makers to prioritize certain aspects, such as dealing with new customers and those with premium subscriptions. The data can also be used for creating decisions about retention policies without involving excessive costs for companies. Moreover, customers that are more prone to exit can be automatically identified through CRM tools.

4.4 Limitations

Limitations of the present study can be seen in the usage of just one type of dataset; hence generalizing the findings to other sectors will not be possible. Further, another limitation of this work is that there is no inclusion of temporal behavior as well as other external environmental factors such as the state of the market within the models used. Lastly, the problem of class imbalance in the data has been catered for through oversampling. This however, may not solve problems of extremely high levels of customer attrition. Future possibilities of research include using datasets that have multiple variables, conducting longitudinal research, and even live streaming of data.

5. Conclusion and Future research

The following paper contains an empirical analysis of the application of ensemble learning approaches AdaBoost and RF in the field of customer churn prediction in the subscription sector. Based on the Kaggle Customer Churn Data Set, RF demonstrated superior performance than AdaBoost in regard to 0.88 Accuracy, 0.85 Precision, 0.82 Recall, 0.83 F1 Score, and 0.90 ROC-AUC. Meanwhile, a hybrid model with the implementation of both algorithms proved to be even more effective, reaching ROC-AUC (0.90). Thus, it can be said that the benefits of utilizing both

models at once have been illustrated. Regarding feature importance, Tenure (0.21), Monthly Charges (0.18), Contract (0.15), Payment Method (0.12), and Tech Support (0.08) were considered the most These statistical observations not only prove the efficiency of ensemble methods but also their interpretability. It is now possible to use this model to reach out to high-risk customers and optimize their retention strategy. The better results achieved through the combination of two algorithms prove that there should be more hybrid approaches that can reveal complicated relationships between variables in consumer behavior. Possible further work would involve hybrid models based on DL and ensemble algorithms that could reveal behavioral patterns and predict future behavior of customers. Integration of real-time predictive models, incorporation of multiple sources of data and behavioral analytics could increase accuracy of the model. Transfer learning for various types of subscription-based businesses could also prove helpful in this regard.

Declaration

Conflict of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

Financial Statement

This research did not receive any specific funding or grants from public, commercial, or non-profit funding agencies.

Data Availability Statement

The dataset used in this study is publicly available on Kaggle and can be accessed at https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset. All data required to replicate the experiments, including features and the target variable, are included in the dataset.

References

1. Kolli, M., Varadharajan, N., Ajith, K., & Kumar, K. D. (2024, March). Customer churn prediction in subscription-based services. In International Conference on Recent Trends in Machine Learning, IoT, Smart Cities & Applications (pp. 247–257). Springer Nature Singapore.
2. Zainab, F., Nazim, F., Kashaf, M., Aslam, N., & Maqbool, M. S. (2026). Predictive analytics for customer churn in subscription-based businesses using machine learning. Spectrum of Engineering Sciences, 4(4), 596–618.
3. Kuramannagari, Y., Mahale, S., Kanamarlapudi, P. V. M. A. K., Veerlapati, H. N., Gupta, M., & Kumar, R. (2025, October). A comparative analysis of machine learning models for customer churn prediction in subscription-based businesses. In 2025 2nd International Conference on Computational Intelligence and Computing Applications (ICCICA) (pp. 1–6). IEEE.
4. Elkhoudary, A., Almansour, B. Y., Vij, P., Almansour, A. Y., Karimdjanov, I., & Elakiya, V. (2025, November). LightGBM for customer churn prediction in subscription-based services. In 2025 International Conference on Electrical Engineering and Informatics (ICEEI) (pp. 1–8). IEEE.
5. Prasad, M. K., Sravya, K. L. S., Krishna, K. N., & Rama, K. Customer churn prediction and retention strategy optimization for subscription-based services using behavioural data analytics and machine learning models.
6. Deligiannis, A., & Argyriou, C. (2020). Designing a real-time data-driven customer churn risk indicator for subscription commerce. International Journal of Information Engineering and Electronic Business, 11(4), 1.
7. Talaat, F. M., & Aljadani, A. (2025). AI-driven churn prediction in subscription services: Addressing economic metrics, data transparency, and customer interdependence. Neural Computing and Applications, 37(14), 8651–8676.
8. Agarwal, P. (2024, June). Data science approaches for churn prediction. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1–7). IEEE.
9. Rakesh, N., Mohan, B. A., Kumaran, U., Prakash, G. L., Arul, R., & Thirugnanasambandam, K. (2024). Machine learning-driven strategies for customer retention and financial improvement. Archives for Technical Sciences, 2(31), 269–283. https://doi.org/10.70102/afts.2024.1631.269
10. Praveenraj, D. D. W., Bustanov, K., Parandhaman, G., Kizi, A. Z. I., Sachdeva, L., & Younus, Y. M. (2025, December). Churn prediction in telecom subscriptions using boosted learning models. In 2025 International Conference on AI-Driven STEM Education and Learning Technologies (AISTEMEDU) (pp. 1–7). IEEE.

11. Shamsudeen, S., & Ranjith Singh, K. (2025). A hybrid approach for customer segmentation using ensemble methods: Bagging and voting classifiers. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 16(3), 414–432. https://doi.org/10.58346/JOWUA.2025.I3.025
12. 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.
13. Han, L. T. (2025). Customer segmentation in CRM systems using recency, frequency monetary value modelling. Global Perspectives in Management, 3(1), 37–47.
14. Zimal, S., Shah, C., Borhude, S., Birajdar, A., & Patil, S. (2023). Customer churn prediction using machine learning. International Journal for Research in Applied Science and Engineering Technology, 11(2), 872–883.
15. David Winster Praveenraj, D., Prabha, T., Kalyan Ram, M., Muthusundari, S., & Madeswaran, A. (2024). Management and sales forecasting of an e-commerce information system using data mining and convolutional neural networks. Indian Journal of Information Sources and Services, 14(2), 139–145. https://doi.org/10.51983/ijiss-2024.14.2.20
16. Hataş, T. A., Obalı, E., Yıldız, A., Çalışkan, S. K., Yılmaz, V. K., Kara, E., ... & Çakar, T. (2023, December). Analyzing customer churn: A comparative study of machine learning models on Pay-TV subscribers in Turkey. In 2023 4th International Informatics and Software Engineering Conference (IISEC) (pp. 1–6). IEEE.
17. Jacob, J., & Thomson Fredrik, E. J. (2025). Leveraging artificial intelligence for improved customer retention in the banking industry. International Academic Journal of Science and Engineering, 12(3), 205–220. https://doi.org/10.71086/IAJSE/V12I3/IAJSE1259
18. Peddarapu, R. K., Ameena, S., Yashaswini, S., Shreshta, N., & PurnaSahithi, M. (2022, December). Customer churn prediction using machine learning. In 2022 6th International Conference on Electronics, Communication and Aerospace Technology (pp. 1035–1040). IEEE.
19. Klein, D., & Dech, S. (2024). The role of big data analytics in enhancing customer relationship management. International Academic Journal of Innovative Research, 11(3), 27–33. https://doi.org/10.71086/IAJIR/V11I3/IAJIR1121
20. Imani, M., Joudaki, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning. Machine Learning and Knowledge Extraction, 7(3), 105.
21. Chordiya, P., & Singh, A. (2026). Customer churn prediction using machine learning. Advances in Consumer Research, 3(1).
22. Ramesh, P., Nithyanandhan, R., Faizal, M. M., Nivakumar, V., & Nalini, M. (2024, December). Machine learning strategies for customer churn prediction in competitive enterprises. In 2024 International Conference on Sustainable Communication Networks and Application (ICSCNA) (pp. 653–660). IEEE.
23. Senthilselvi, A., Kanishk, V., Vineesh, K., & Raj, A. P. (2024, May). A novel approach to customer churn prediction in telecom. In 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) (pp. 1–7). IEEE.