INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

Received: 30-08-2024

Accepted: 18-10-2024

Published: 18-11-2024

Citation: Singh U, Singh S, Rathee T, Vaish M (2024) Customer Retention Modeling over the OTT Platform using Machine Learning. Indian Journal of Science and Technology 17(42): 4365-4371. https://doi.org/10.17485/IJST/v17i42.2619

* **Corresponding author**.

er.sandeep85@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Customer Retention Modeling over the OTT Platform using Machine Learning

Upma Singh[1], Sandeep Singh[1]*, Tripti Rathee[2], Manav Vaish[2]

1 Department of Electronics and Communication Engineering, Maharaja Surajmal Institute of Technology, 110058, Delhi, India
2 Department of Information Technology, Maharaja Surajmal Institute of Technology, 110058, Delhi, India

Abstract

Objectives: Customer retention, a multifaceted issue that plagues the digital entertainment industry, particularly within the realm of Over-the-Top (OTT) platforms, poses a significant challenge, impacting revenue and sustainability. With the discontinuation of subscriptions or usage, retention not only impacts immediate revenue streams but also threatens the long-term sustainability of these platforms. Recognizing this challenge, this paper undertakes a comprehensive examination of predictive modeling methods tailored explicitly for forecasting customer retention within OTT platforms. **Method:** To construct predictive models, a diverse array of datasets is harnessed, encompassing a wide spectrum of customer-related variables. These datasets include demographic information, viewing history, subscription patterns, and various engagement metrics. Leveraging these datasets, advanced machine learning algorithms are deployed to develop robust models capable of predicting customer retention. **Findings:** The scrupulous study evaluates the implementation of a range of machine learning algorithms, including logistic regression, random forests, AdaBoost classifier, Decision Tree, and K nearest neighbor (KNN) classifier. Assessment metrics such as F1-score, recall, precision, and accuracy are employed to show the effectiveness of the employed models for customer retention modeling within OTT platforms. The results reveal that the highest accuracy of 80.40% is obtained using the AdaBoost classifier. **Novelty:** The research uses attribute significance analysis as a means of identifying the fundamental factors that impact client retention. OTT providers can receive important insights into the elements causing subscriber attrition by identifying these main drivers, which will help them develop tailored retention strategies.

Keywords: Neural networks; Logistic regression; Random forests; Decision trees; KNN; AdaBoost classifier; Retention prediction

1 Introduction

Nowadays, organizations in various industries face a substantial financial burden due to the competitive nature of today's business environment, client retention, and the phenomenon of customers discontinuing their service or subscription. Accurate retention estimates are essential for optimizing profitability and promoting long-term growth because acquiring new customers frequently turns out to be far costlier than keeping present ones[1–3]. This paper carefully traverses through the employment of machine learning (ML) as an efficient tool for proactively recognizing customers at the peril of retention. ML algorithms can identify important retention-related behavioral patterns by evaluating vast volumes of past customer data. This enables companies to classify their consumers and create customized interventions before they become disengaged. Customer retention analysis, a crucial task for businesses across industries, relies on ML techniques to speculate and alleviate customer attrition[4,5].

Though research in this field has showcased the efficacy of diverse methodologies and algorithms, still there is a gap that must be encountered to retain the customer[6].

Throughout the telecommunications industry, earlier studies on predicting customer retention primarily focused on a finite set of machine learning classifiers[7]. To contrast and recognize the most potent predictive model, a multiobjective-cost-sensitive ant colony optimization (MOC-ACO-Miner) methodology was presented by Özmen et al.[8] it combines multiobjective ACO-based cost-sensitive learning with cost-based nondominated sorted genetic algorithm feature selection. One of the top 100 IT firms in Turkey uses MOC-ACO-Miner to forecast customer attrition. Joy et al.[9] suggested a huge data-driven hybrid methodology that combines a deep neural network and a machine-learning model in order to precisely predict client attrition. Moreover, the optimal set of features for our suggested model is found using feature selection methods including Chi-squared testing and Sequential Feature Selection (SFS). The aim of Kumar et al.'s[10] investigation was to give service providers proactive tactics to lower churn rates and boost customer retention by using advanced machine learning techniques to forecast customer attrition in the telecom sector. In a computational experiment using bank customer attrition data, Szeląg et al.[11] demonstrated the explanatory and predictive powers of monotonic decision rules. The dataset instances are balanced by employing the SMOTE (Synthetic Minority Over-Sampling Technique) with KNN, Naïve Bayes, C4.5, Random Forest, AdaBoost, and ANN[12]. With an AUC of 91.10%, Random Forest was found to function the best.

The researchers also used Naïve Bayes and Random Forest, finding Random Forest to be more effective with an accuracy of 71.99%[13]. Adhikary Gupta examined classifiers exceeding 100 for the telecom sector's retention forecasting[14]. Regularized random forest had the best results, with an accuracy of 73.04%, while Bagging Random Forest exceeded the rest in terms of AUC, coming in at 67.20%. Pasquadibisceglie et al.[15] formulated the topic of customer churn prediction as a Predictive Process Monitoring (PPM) problem to be solved in potentially dynamic retail data contexts. In order to provide greater accuracy in predictive modeling, Sikri et al. presented an innovative strategy: The Ratio-based data balancing method. This strategy handles data skewness as an initial processing phase[16]. Saleh et al. investigate potential churn drivers in the Danish telecom market and their relationship to retention tactics. Although the number of service providers has expanded dramatically in recent years, the Danish telecom market is currently saturated in terms of users[17]. Singh et al. used bank data to forecast which customers are most likely to cease using the bank's services and begin paying for them[18].

Though research in this field has showcased the efficacy of diverse methodologies and algorithms, there is still a gap that must be encountered to retain the customer. Some of the key research gaps observed are as follows:

1. Lack of Contextual Understanding of User Behaviour.
2. Personalization at Scale
3. Cross-Platform User Behaviour Analysis
4. Explaining Model Predictions (Interpretability)
5. Dynamic Retention Strategies

While ML models can predict churn based on user activity, many existing models fail to incorporate deeper contextual information, such as mood, social influence, or real-time events (e.g., content release timing). These external factors can significantly impact user engagement. Many models are trained on aggregate behavior without capturing the nuances of individual preferences in real-time. This paper tries to overcome the shortcomings in the field of churn prediction by applying ML algorithms.

The authors explore the field of ML to provide new insights into the dynamics of customer attrition and to challenge the limits of current retention prediction techniques. The authors have evaluated prediction systems using a variety of performance metrics, including F1 score, recall, accuracy, and precision, using logistic regression, AdaBoost classifier, K-Neighbors classifier, Gradient Boosting classifier, and Random Forest classifier approaches. To balance the instances in the dataset, the training set is subjected to the Synthetic Minority Oversampling Technique (SMOTE). Findings are acquired and both with and without SMOTE are discussed.

The main contribution of the paper is as follows:

1. The applied logistic regression and random forest model uses contextual features like sentiment analysis of user reviews, social media engagement, or external events (e.g., new content releases) which improve the retention rate of customers.
2. The paper uses the AdaBoost classifier which significantly improves the performance of weak learners, making them highly effective when combined. Also, this model is nonparametric and adaptable.
3. Random forest handle complex, non-linear relationships and is extended to include cross-platform user activity as a feature. They help identify patterns where user behavior on one platform impacts retention on another.

Paper Organization: The remaining portions of the paper are arranged as follows: The churn prediction system's techniques are introduced in Section 2. Section 3 presents the results, while Section 4 concludes the paper.

2 Methodology

2.1 Data acquisition and preprocessing:

The first step in the methodology involved is the acquisition of relevant data sources containing historical customer information. These sources may include transactional data, consumer profiles, usage trends for services, and retention labels. The data is gathered from different references for instance customer relationship management (CRM) systems, billing databases, call detail records (CDRs), and online platforms. The collected data undergoes extensive preprocessing to address problems like inconsistencies, outliers, and missing values. Data cleaning tasks, including standardization, normalization, and encoding categorical variables into numerical representations, are performed to certify the quality and consistency of the datasets. Additionally, the authors explore the distribution of target labels (retention vs. non-retention) to assess class imbalance and apply appropriate sampling techniques if necessary.

2.2 Attribute engineering:

Once the data preprocessing is complete, we proceed with attribute engineering to extract meaningful predictors of customer retention. Exploratory Data Analysis (EDA) techniques are employed to achieve perception into the aspects of the dataset and identify latent attributes relevant to retention prediction. We generate new attributes by transforming or combining existing variables to capture meaningful patterns and relationships. These attributes may include customer demographics, behavioral metrics, engagement levels, and satisfaction scores. Furthermore, we leverage domain knowledge and business expertise to select informative attributes with predictive power for retention prediction. An entire numerical attribute histogram is shown in Figure 1. Figure 2 shows the tenure distribution with retention.

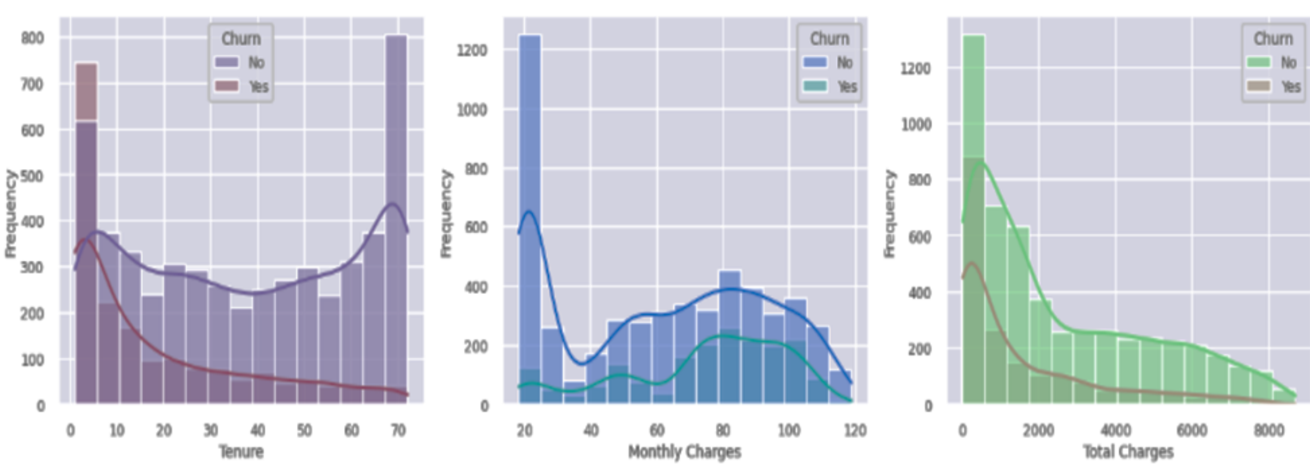

Fig 1. Histogram of all Numerical attributes

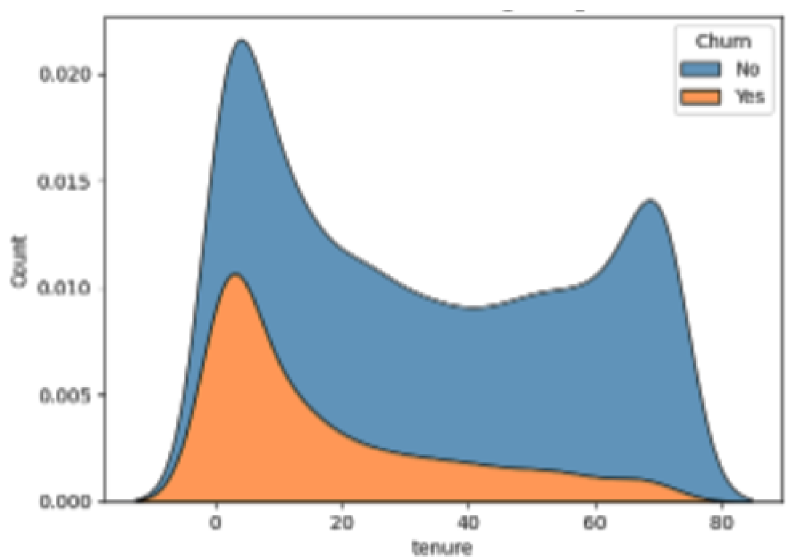

Fig 2. Distribution of tenure regarding retention

2.3 Selection and Assessment of Models:

A wide range of ML techniques appropriate for retention prediction tasks are taken into consideration throughout the model selection and evaluation phase. Among these methods are Logistic Regression, Random Forests, AdaBoost Classifier, Gradient Boosting, and K-Neighbors Classifier. Training, validation, and test sets are created from the pre-processed dataset using appropriate random or time-based splitting techniques. To prevent overfitting and get optimal performance, we employ cross-validation methods to optimize the hyperparameters of the models we train on the training data. Table 1 shows the attributes of the dataset without preprocessing.

The trained models are evaluated using a variety of performance metrics, given in Equations (1), (2), (3) and (4)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$PRECISION = \frac{TP}{TP + FP} \tag{2}$$

$$RECALL = \frac{TP}{TP + FN} \tag{3}$$

$$F_1\,SCORE = 2 * RECALL * \frac{PRECISION}{RECALL + PRECISION} \tag{4}$$

By counting the true positive rate (TPR) and false positive rate (FPR) from classification models and applying them as the vertical and horizontal axes, respectively, the ROC curve is produced. According to Equations (5) and (6), the FPR and TPR are computed. What's beneath the ROC curve is called the area under the curve (AUC). Likewise, a higher AUC value indicates superior results. The configuration of hyperparameters is presented in Table 2.

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

Table 1. Attributes of Dataset without preprocessing

S.No	Attributes	Description	Data Format
1	Customer ID	Customer ID	Int
2	TV Subscriber	Whether the customer is a TV subscriber or not	Int
3	Movie Subscriber	Whether the customer has movie subscribed or not	Int
4	Subscription age	Time of the subscription	Float
5	Bill average	Bill Average of the customer	Int
6	Remaining contract	Time remaining of the contract	Float
7	Service failure count	Count of the service failure	Int
8	Download Average	Average download of the customer	Float
9	Upload Average	Average upload of the customer	Float
10	Download Over Limit	Limit crossed while downloading	Int
11	retention	Whether the customer retained or not	Int

Table 2. Hyper Parameter Configuration

Classifier	Hyper Parameters	Dataset (without SMOTE)	Dataset (SMOTE)
Logistic Regression	Penalty	11	12
	Solver	Lib linear	lbfgs
Random Forest Classifier	Max_ eatures	sqrt	auto
	Max_Leaf_Nodes	20	20
AdaBoost Classifier	lr	0.1	0.01
	Solver	Adam	Adam
Gradient Boosting Classifier	Max _attributes	Sqrt	Sqrt
	Max_Leaf_Nodes	20	20
K Neighbors Classifier	None	None	None

2.4 Ensemble methods and model stacking:

To improve prediction even further, we explore ensemble learning strategies including stacking, boosting, and bagging. Ensemble approaches combine predictions from multiple base models to improve overall accuracy and durability. We implement ensemble methods e.g., random forest ensembles and Gradient Boosting Machines (GBMs) to leverage the heterogeneity of individual models and mitigate biases. We also experiment with model stacking approaches, in which the final predictions are made by a meta-learner using the input attributes from a variety of models. By combining complementary strengths of different algorithms, ensemble methods, and model stacking offers a powerful framework for improving retention prediction performance.

2.5 Model Interpretability and Explainability:

In addition to predictive accuracy, we prioritize model interpretability and explainability in retention prediction. To explain model predictions and identify the main mechanisms influencing customer retention, we use methods like SHAP (Shapley Additive explanations), partial dependence plots, and attribute importance analysis. By elucidating the contributions of individual attributes to retention prediction, we provide actionable insights for decision-makers to devise targeted retention strategies and allocate resources effectively.

2.6 Validation and Deployment:

To verify robustness and generalization performance, we assess the final retention prediction model's performance on the validation set once it has been constructed. We conduct sensitivity analysis and stress testing to evaluate model stability and resilience to variations in input data. The validated model is then prepared for deployment in operational environments, considering factors such as scalability, latency, and integration with existing systems. It is crucial to regularly recalibrate the model and monitor its performance in order to adjust to shifting consumer trends and shifting business environments.

3 Results and Discussion

3.1 Result of Dataset 1 without SMOTE

resents the results without SMOTE. KNN and AdaBoost Classifier in our model shows the highest Accuracy (83.23% and 81.87%), F1 value (80.57% and 81.91%), and AUC (86.54% and 85%), indicating strong performance in both overall prediction correctness and the balance between precision and recall.

Logistic Regression and Random Forest models in our case show moderate performance in terms of recall and F1 score, though the AUC remains high, indicating good classification ability.

Compared to the benchmarks, our models consistently outperform in AUC, showing better overall discriminatory power across most algorithms. However, Lalwani et al.'s model outperforms in terms of precision and recall in certain cases like Logistic Regression and Random Forest

3.2 Result of Dataset 1 with SMOTE

resents the results with SMOTE.Our models tend to outperform Wu et al.'s models across most metrics, especially in AUC, indicating stronger classification power and model discrimination. AdaBoost and Decision Tree models are highly competitive, with Wu et al.'s AdaBoost having a slight edge in accuracy, while your Decision Tree outperforms Wu et al. in almost all metrics. It is noticed from Table 4 that the AdaBoost classifier has a maximum accuracy of 80.98%. The KNN Classifier got the best precision of 56.08% in comparison to other classifiers. Logistic Regression got the highest F_1 score and Recall value of 64.23% and 79.52% respectively.

We found that when it came to predicting accuracy and robustness, ensemble techniques like gradient-boosting machines and random forests consistently beat individual algorithms. Overall, our findings demonstrate how machine learning-based strategies may effectively tackle the problems of predicting client retention and managing retention in the fast-paced corporate world of today.

Table 3. Results without SMOTE

Algorithms	Comparison	Accuracy	Precision	Recall	F1 Value	AUC
Logistic Regression	Our Model	81.34	66.84	55.28	62.12	85.04
	Wu et. Al [19]	80.19	65.17	54.57	59.37	84.36
	Lalwani et. Al [20]	80.45	79.11	80.23	78.89	82
Random Forest	Our Model	80.05	65.48	50.32	57.12	85.66
	Lalwani et. Al [20]	78.04	78.68	77.54	77.91	82
	Wu et. Al [19]	79.55	66.1	47.51	55.25	83.79
AdaBoost Classifier	Our Model	81.87	75.44	83.87	81.91	85
	Wu et. Al [19]	80.08	65.39	53.24	58.61	84.51
	Lalwani et. Al [20]	81.71	81.21	80.14	80.28	84
KNN	our model	83.23	80.32	81.26	80.57	86.54
	Lalwani et. Al [20]	79.64	79.71	78.38	77	80
Decision Tree	our model	80.21	81.48	80.33	80.69	82.5
	Lalwani et. Al [20]	80.14	80.1	78.81	78.89	83

Table 4. Results with SMOTE

Algorithms	Comparison	Accuracy	Precision	Recall	F1 Value	AUC
Logistic Regression	Our Model	75.45	53.76	79.52	64.23	85.25
	Wu et. Al [19]	74.82	51.74	78.76	62.43	84.39
Random Forest	Our Model	77.89	56.01	73.37	63.44	84
	Wu et. Al [19]	76.99	55.14	73.25	62.86	83.8
AdaBoost Classifier	Our Model	80.98	56.03	72.34	62.09	85
	Wu et. Al [19]	77.19	55.44	73.35	63.11	84.52
KNN	our model	75.37	56.08	74.28	60.3	86.22
	Wu et. Al [19]	NR*	NR	NR	NR	NR

Continued on next page

Table 4 continued

Decision Tree	our model	77.23	56.21	72.87	61.89	82.98
	Wu et. Al [19]	76.74	54.97	72.67	62.26	82.83

NR-Not Reported

4 Conclusion

It is more advantageous for operators to suggest retention methods to customers who are going to quit the OTT platform because customers there are always and always inclined to be saturated. Thus, developing a system that can forecast customer attrition over the OTT platform was the aim of this research. Following a thorough analysis of several models and performance measures, the authors concentrated on optimizing the Ada boost Classifier (Accuracy 81.87%) and Logistic Regression (Accuracy 81.34%) models. These models perform well, especially when it comes to recall, which is in line with our main goal of correctly detecting possible retention. By giving recall a higher priority than precision, the chance of false negatives is reduced and consumers who are at risk of retensioning were identified.

To address the unbalanced datasets, additional under-sampling, and oversampling methods can be investigated in the future. Furthermore, to enhance prediction accuracy, ROC analysis can be examined in greater detail to determine a more sensible threshold for churn prediction.

References

1) 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;14:100342–100342. Available from: https://doi.org/10.1016/j.rico.2023.100342.
2) Haddadi SJ, Farshidvard A, Silva S, Reis F, Reis JCDS, M. Customer churn prediction in imbalanced datasets with resampling methods: A comparative study. . *Expert Systems with Applications*. 2024;246:123086–123086. Available from: https://doi.org/10.1016/j.eswa.2023.123086.
3) Poudel SS, Pokharel S, Timilsina M. Explaining customer churn prediction in telecom industry using tabular machine learning models. . *Machine Learning with Applications*. 2024;17:100567–100567. Available from: https://doi.org/10.1016/j.mlwa.2024.100567.
4) Ahmed N, Umair M. Churn prediction using machine learning: A coupon optimization technique. *World Journal of Advanced Engineering Technology and Sciences*. 2024;12(2):332–54. Available from: https://doi.org/10.30574/wjaets.2024.12.2.0310.
5) Mouli KC, Raghavendran CV, Bharadwaj VY, Vybhavi GY, Sravani C, Vafaeva KM, et al. An analysis on classification models for customer churn prediction. . *Cogent Engineering*. 2024;11:2378877–2378877. Available from: https://doi.org/10.1080/23311916.2024.2378877.
6) Sam G, Asuquo P, Stephen B. Customer churn prediction using machine learning models. *Journal of Engineering Research and Reports*. 2024;26(2):181–93. Available from: https://doi.org/10.9734/jerr/2024/v26i21081.
7) Amin A, Al-Obeidat F, Shah B, Tae MA, Khan C, Durrani HU, et al. Just-in-time customer churn prediction in the telecommunication sector. *The Journal of Supercomputing*. 2020;76:3924–3972. Available from: https://doi.org/10.1007/s11227-017-2149-9.
8) Özmen M, Aydoğan EK, Delice Y, Toksarı MD. Churn prediction in Turkey's telecommunications sector: A proposed multiobjective-cost-sensitive ant colony optimization. . *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2020;10. Available from: https://doi.org/10.1002/widm.1338.
9) Joy UG, Hoque KE, Uddin MN, Chowdhury L, Park SB. A Big Data-Driven Hybrid Model for Enhancing Streaming Service Customer Retention through Churn Prediction Integrated with Explainable AI. . *IEEE Access*. 2024. Available from: https://doi.org/10.1109/ACCESS.2024.3401247.
10) Kumar KP, Kanishkar P, Raja VD, Kumar TA, Gopal SB, Gunasekar M. Telecom Churn Movement Prediction Using Machine Learning. In: International Conference on Intelligent Systems Design and Applications. Springer Nature Switzerland. 2023;p. 235–243. Available from: https://doi.org/10.9734/jerr/2024/v26i21081.
11) Szeląg M, Słowiński R. Explaining and predicting customer churn by monotonic rules induced from ordinal data. *European Journal of Operational Research*. 2024;317(2):414–438. Available from: https://doi.org/10.1016/j.ejor.2023.09.028.
12) Pamina J, Raja B, Sathyabama S, Sruthi MS, A V. An effective classifier for predicting churn in telecommunication. *Journal of Advanced Research in Dynamical & Control Systems*. 2019;11. Available from: https://ssrn.com/abstract=3399937.
13) Tijah SK. Retention Prediction. 2020. Available from: https://www.kaggle.com/khotijahs1/retention-prediction.
14) Adhikary DD, Gupta D. Applying over 100 classifiers for churn prediction in telecom companies. . *Multimedia Tools and Applications*. 2021;80:35123–35167. Available from: https://doi.org/10.1007/s11042-020-09658-z.
15) Pasquadibisceglie V, Appice A, Ieva G, Malerba D. TSUNAMI-an explainable PPM approach for customer churn prediction in evolving retail data environments. *Journal of Intelligent Information Systems*. 2024;62(3):705–738. Available from: https://doi.org/10.1007/s10844-023-00838-5.
16) Sikri A, Jameel R, Idrees SM, Kaur H. Enhancing customer retention in telecom industry with machine learning driven churn prediction. . *Scientific Reports*. 2024;14:13097–13097. Available from: https://doi.org/10.1038/s41598-024-63750-0.
17) Saleh S, Saha S. Customer retention and churn prediction in the telecommunication industry: a case study on a Danish university. *SN Applied Sciences*. 2023;5(7):173–173. Available from: https://doi.org/10.1007/s42452-023-05389-6.
18) Singh PP, Anik FI, Senapati R, Sinha A, Sakib N, Hossain E. Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. . *Data Science and Management*. 2024;7:7–16. Available from: https://doi.org/10.1016/j.dsm.2023.09.002.
19) Wu S, Yau WC, Ong TS, Chong SC. Integrated churn prediction and customer segmentation framework for telco business. *IEEE Access*. 2021;9:62118–62154. Available from: https://doi.org/10.1109/ACCESS.2021.3073776.
20) Lalwani P, Mishra MK, Chadha JS, Sethi P. Customer churn prediction system: a machine learning approach. . *Computing*. 2022;104:271–94. Available from: https://doi.org/10.1007/s00607-021-00908-y.