A Comprehensive Survey in ANN Based Customer Churn Prediction

Authors: Laksh Krishna Sharma, Rupesh Singh Karki, Prachi Dahiya and Umang Kant

1) Abstract

Customer churn, defined as the attrition of valuable customers to rival services, continues to pose a significant challenge for both telecommunications companies and subscription-based service providers. In recent years, the rapid pace of digital transformation, market liberalization, and the advent of new technologies have reshaped these industries, intensifying competition and complicating customer retention efforts. The repercussions of churn are considerable; even a minor reduction in the customer base can lead to substantial revenue losses and destabilize service operations. This study explores churn prediction using two datasets: one from the telecom sector and another from a subscription-based platform. The telecom dataset includes customer demographics, usage statistics, and account attributes, whereas the subscription dataset captures the billing details, service engagement, and content preferences. A diverse range of machine-learning techniques have been applied, ranging from ensemble and linear classifiers to tree-based, probabilistic, and hybrid approaches. Performance comparisons were conducted before and after hyperparameter tuning, with results demonstrating that feature engineering and optimization significantly enhance predictive accuracy. These insights contribute to the development of more effective customer retention strategies.

2) Introduction

Traditional research on churn prediction has primarily focused on fundamental usage metrics, such as voice call records, billing information, and general demographic data, particularly in the telecommunications sector [17]. However, as consumer behavior evolves, there is an increasing necessity to understand the broader array of factors influencing churn[18] [20]. This study broadens the scope of churn analysis by employing two comprehensive datasets: one from the telecommunications industry and another from a subscription-based service.

The telecommunications dataset provides detailed customer profiles encompassing demographics (e.g., gender, age, state, and city), service usage (calls made, SMS sent, and data usage), and account specifics (registration date, number of dependents, estimated salary, and churn status).

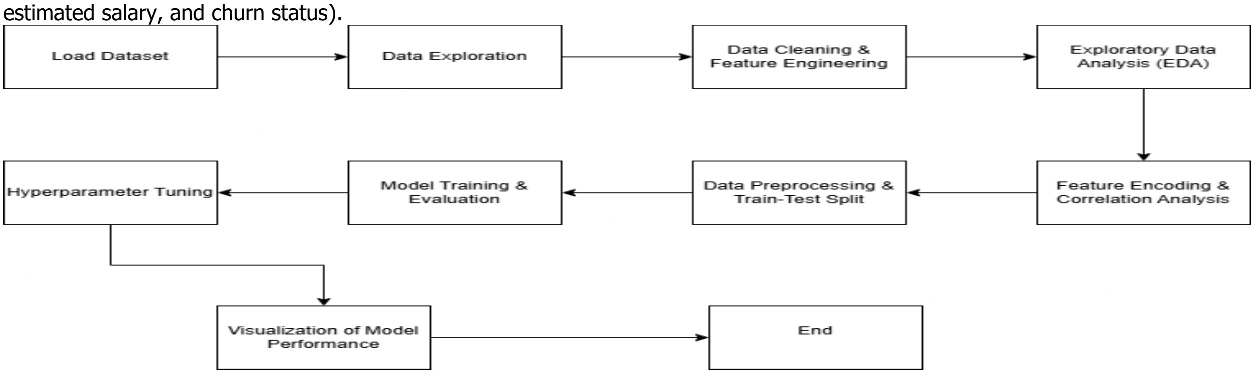


Figure 1 illustrates these elements.

In contrast, the subscription dataset offers insights into customer engagement with digital services, featuring variables such as account age, monthly and total charges, subscription type, payment method, viewing habits, and support interactions [18]. This dual-dataset approach facilitates the capture of a broader spectrum of customer behaviors and identification of nuanced patterns that may contribute to churn.

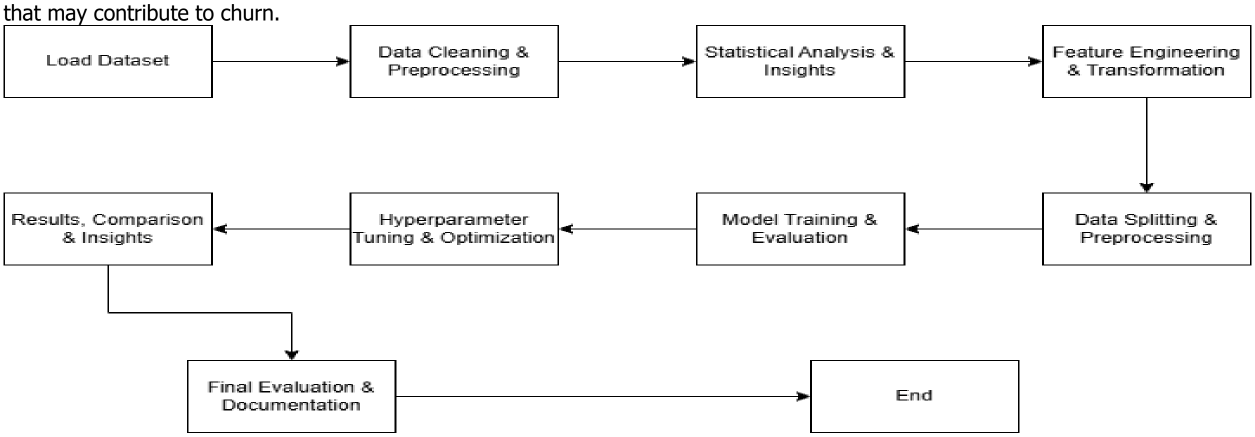


Figure 2 illustrates the data flow.

Recent advancements in data mining and machine learning have introduced new opportunities for addressing churn [9] [18] [20]. By applying a diverse range of predictive models—including ensemble methods, linear classifiers, probabilistic approaches, tree-based methods, hybrid techniques, and artificial neural networks—this study aims to evaluate and compare the effectiveness of these models in predicting customer churn [6] [10] [13]. Importantly, the analysis examined model performance both before and after hyperparameter tuning to assess improvements in accuracy and efficiency. Integrating advanced feature engineering and ensemble strategies has been shown to significantly improve the churn prediction accuracy, making these approaches increasingly relevant for industrial applications [8] [10] [18]. The remainder of this paper is organized as follows. Section 3 reviews the relevant literature and outlines the current challenges in churn prediction in both sectors. Section 4 details the methodology, including data collection, preprocessing, feature extraction, and modeling framework. Section 5 presents the experimental results and discusses the implications of the customer retention strategies. Finally, Section 6 concludes the study with insights into future research directions and potential industrial applications.

3) Literature Survey

Predicting customer churn presents a challenge across industries, notably in telecommunications and subscription services, where retention is more cost-effective than acquisition. Machine learning (ML) advancements have enabled accurate churn predictions using historical data. This study examines datasets from subscription-based and telecommunications sectors. For subscription data, missing values, outliers, and variable categorization were addressed through encoding and scaling. In telecommunications data, usage metrics were rectified, temporal features extracted, and age groups categorized. The modeling approach evaluated Random Forest (RF), Gradient Boosting (GB), extreme Gradient Boosting (XGBoost), LightGBM (LGBM), CatBoost, Logistic Regression (LR), Ridge, Passive-Aggressive classifiers, and neural networks. RF and Gaussian Naive Bayes (GNB) performed well on subscription data, while telecom models achieved 80% accuracy and 0.44 F1-Score, revealing trade-offs between complexity and performance. Studies have shown ML algorithms' effectiveness in churn prediction. One study [3] showed Stochastic Gradient Boosting (SGB) outperformed RF, LR, and K-Nearest Neighbors (KNN) with 83.9% accuracy. Another study [2] achieved 96.3% accuracy using RF with stratified k-fold cross-validation, while a hybrid SMOTE-ENN approach achieved 95% accuracy with RF [1].

Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) have been applied for hyperparameter tuning and feature selection [15], efficiently navigating complex search spaces to enhance model performance. Preprocessing steps like imputation and cleaning significantly improve model robustness [5]. Deep learning methods have shown promise, with Recurrent Neural Networks (RNNs) achieving 94.8% accuracy on customer churn data [6]. Composite models combining Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM) networks achieved 98.36% accuracy on the Telco Customer Churn dataset [10]. Gradient boosting methods remain central to churn prediction, with XGBoost achieving 82.35% accuracy using correlation-based feature selection [7]. Feature importance ranking and advanced feature engineering optimize model strategies [11]. Studies have integrated SMOTE-ENN with ensemble methods like RF to enhance accuracy [8][9], while RNNs outperform Feedforward Neural Networks (FNNs) in capturing sequential customer behavior [13][14]. In banking, SMOTE with GA achieved 96% accuracy using KNN [15], while a hybrid approach using RF achieved 85% accuracy [19]. GA-XGBoost achieved an AUC of 99% and F1-score of 90% with SHAP value interpretation [20]. In telecom, a study [17] applied machine learning to customer data, integrating Social Network Analysis (SNA) features to improve AUC to 93.3%. Another approach [18] used data transformation methods and feature selection, boosting AUC by 26.2% using logistic regression and neural networks.

Table 1 provides a comparison of the original and additional references between different existing algorithms.

Table 1: Summary of Research Papers on Customer Churn Prediction

Ref.	Paper Title	Algorithms/Approach	Dataset	Key Contribution	Best Accuracy	Drawbacks
[1]	Customer Churn Prediction Using Machine Learning Approaches	RF, DT, LR	Telecom (7,043 samples)	Combined SMOTE-ENN with RF to address imbalance	95% (RF)	Lower performance compared to deep models
[2]	CUSTOMER CHURN PREDICTION	RF, LR, KNN	Telecom (10,000+ samples)	Addressed class imbalance with SMOTE-ENN	96.3% (RF)	Model explainability not emphasized
[3]	Customer Churning Analysis Using Machine Learning Algorithms	SGB, RF, LR, KNN	Telecom (7,044 samples)	Highlighted SGB's superiority with hyperparameter tuning	83.9% (SGB)	Limited feature engineering; small dataset
[4]	Comparative Analogy Among Different Machine Intelligence Based Techniques for Customer Churn Prediction	RF, XGBoost, ANN	Multi-industry	Emphasized ensemble methods and hybrid optimization	91.66% (RF)	Lack of domain-specific feature customization
[5]	Customer Churn Prediction in Telecom Sector Using Machine Learning Techniques	Decision Tree, Random Forest, KNN, Survival Analysis, Cox PH Model	Telecom (Telco-Customer-Churn, 7,043 samples)	Integrates robust preprocessing, data balancing (SMOTE-ENN), and temporal analysis to enhance churn prediction; demonstrates ensemble	99% (Random Forest)	High overfitting risk due to SMOTE on small data

				methods achieving up to 99% accuracy		
[6]	AI-Based Customer Churn Prediction Model for Business Markets in the USA	CNN, RNN, LSTM, KNN, SVM, DT, LR, NB	Business market (10,000 records)	Deep learning (RNN, LSTM) for sequential behavioral modeling; high accuracy from RNN; focus on CRM integration	94.8% (RNN)	Higher computational cost
[7]	Bank Customer Churn Prediction Using Machine Learning	LR, DT, RF, KNN, XGBoost	Bank dataset (10,000 records)	Demonstrated superior performance of XGBoost with correlation-based feature selection	82.35% (XGBoost)	No hybrid sampling for imbalance issues
[8]	Customer Churn Prediction Using Machine Learning Approaches	RF, DT, LR	Telecom (7,043 samples)	Combined SMOTE-ENN with RF to address imbalance	95% (RF)	Model interpretability not addressed
[9]	CUSTOMER CHURN PREDICTION	RF, LR, KNN	Telecom (10,000+ samples)	Addressed class imbalance with SMOTE-ENN	96.3% (RF)	Imbalance handling may lead to noisy samples
[10]	Customer churn prediction using composite deep learning technique	CNN, GRU, LSTM (Composite DL)	Telco Customer Churn	Developed a fused CNN-GRU-LSTM model capturing deep sequential features	98.36% (Composite DL)	High training time and hardware dependency
[11]	Customer Churn Prediction using Machine	KNN, SVM, Decision Tree, RF, LR	Telecom dataset	Highlights feature importance and model interpretability; RF performed best	89.9% (RF)	Basic models; no deep learning exploration
[12]	Customer churn prediction using composite deep learning technique	CNN + LSTM	Telecom dataset	Leverages spatial and temporal features via deep learning; outperforms ML models	95.2% (CNN+LSTM)	Model complexity; risk of overfitting
[13]	Neural Network Based a Comparative Analysis for Customer Churn Prediction	FNN, RNN	Real-world telecom	RNNs shown to be superior due to sequential learning; uses dropout and early stopping to improve generalization	93.45% (RNN)	Overfitting with small datasets
[14]	Customer churn prediction using composite deep learning technique	CNN + LSTM	Real-world telecom	Proposes a CNN-LSTM hybrid for advanced sequential and spatial feature learning; utilizes regularization techniques	95.02% (CNN+LSTM)	Lack of interpretability
[15]	Bank Customer Churn Prediction Using SMOTE: A Comparative Analysis	LR, RF, XGBoost, ANN	Bank dataset (Kaggle - Credit Card Customers)	Applied SMOTE for balancing; compared traditional ML vs. ANN; ANN slightly outperformed others	89.3% (ANN)	Risk of minority class oversampling bias
[16]	Customer churning analysis using machine learning algorithms	RF, DT, SVM, NB, KNN	Telecom (International Journal of Intelligent Networks)	Compared multiple classifiers; RF and DT performed the best among all, reinforcing ensemble effectiveness	92.1% (RF)	Lack of ensemble or hybrid techniques
[17]	Customer churn prediction in telecom using machine learning in big data platform	DT, RF, GBM, XGBoost		Integrated big data techniques and SNA features; XGBoost achieved top AUC	93.3% (XGBoost)	Big data pipeline setup not detailed
[18]	A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection	LR, RF, KNN, FNN, RNN, NB, GB, DT	Telecom (4 datasets)	Applied data transformations (WOE, Rank, Z-score) and feature selection; WOE+LR/FNN achieved best results	+26.2% AUC gain	Only evaluated on publicly available datasets
[19]	Customer Churn Prediction in Banking Sector- A Hybrid Approach	LR, RF, SVM, DT	Bank dataset	Clustered dataset and applied best model per cluster; Random Forest emerged as strongest performer	85% (Random Forest)	Limited model interpretability
[20]	Research on customer churn prediction and model interpretability analysis	GA-XGBoost, SHAP analysis	Bank dataset (Kaggle - Credit Card Customers)	Optimized XGBoost with Genetic Algorithm; SHAP explained model behavior; highest AUC and F1 achieved	99% (AUC)	Computationally intensive model tuning

6) Key Findings

1. In the evaluation of model performance and efficiency, ensemble models such as Random Forest and Gradient Boosting demonstrated approximately 80% accuracy. In contrast, models such as XGBoost, LGBM, and CatBoost exhibited an accuracy of approximately 79.95% when applied to telecommunications datasets. Notably, simpler models, including Gaussian Naive Bayes, also yielded satisfactory results, underscoring the significance of feature selection and preprocessing in model performance.

2. To address the class imbalance, the application of SMOTE-ENN effectively balanced the dataset, thereby enhancing the recall rate for churners. Additionally, the use of stratified sampling ensured equal representation, which mitigated bias and improved precision-recall trade-offs, particularly within telecom datasets.

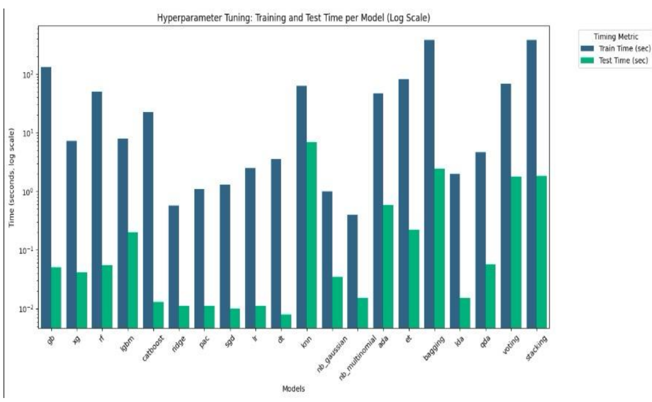
3. Optimization strategies, such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO), have been employed to optimize hyperparameters and facilitate feature selection, thereby enhancing model efficiency. The feature selection process effectively reduces dimensionality, which accelerates the training and mitigates the risk of overfitting.

4. In examining the algorithm performance trends, it was observed that Gradient Boosting, XGBoost, and Random Forest consistently demonstrated robust performance. The integration of hybrid models, such as SMOTE-ENN, combined with ensemble methods, has been shown to enhance the accuracy of churn prediction. It is imperative that model selection be tailored to the complexity of the dataset and the specific characteristics of the churn behavior.

5. Practical Applications and Future Research: Preprocessing techniques, including feature engineering, outlier treatment, and encoding, have enhanced accuracy. Logistic Regression offers interpretability and emphasizes the key indicators of churn. Future research should concentrate on hybrid models and real-time churn-prediction systems.

6. Behavioral and temporal analyses have identified the key predictors of customer churn by examining customer interactions, service usage, and complaints. Temporal factors, such as contract renewals and periods of inactivity, were also found to be significant. Notably, behavioral characteristics have been shown to enhance the accuracy of churn prediction, particularly within subscription-based datasets.

7. Algorithm Performance on Telecom Dataset



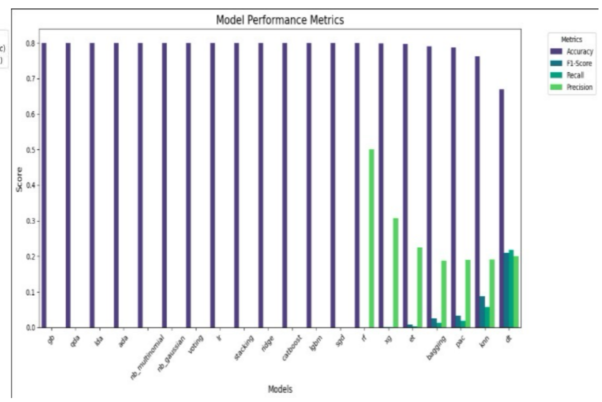
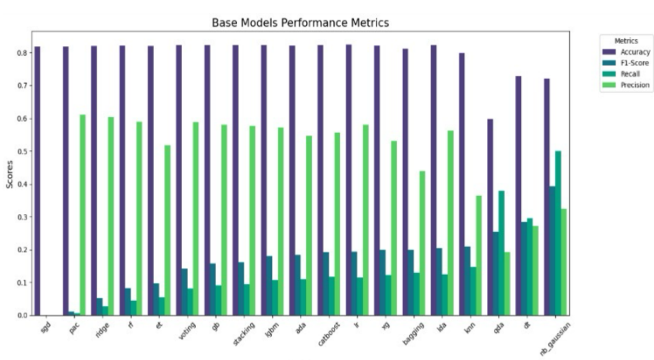


Figure 3 illustrates the train and test time.

Figure 4 illustrates the performance of different algorithms.

8. Algorithm Performance on Subscription Dataset



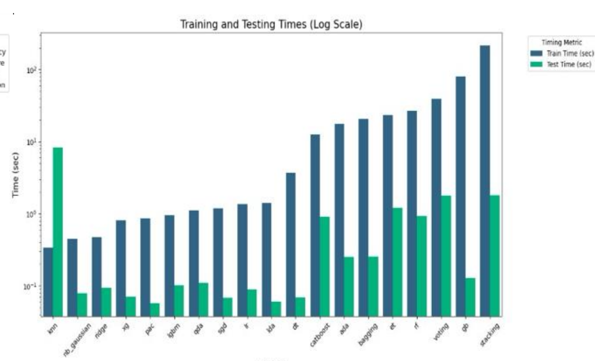


Figure 5 illustrates the train and test time.

Figure 6 illustrates the performance of different algorithms.

7)Conclusion

This study explores customer churn prediction using advanced machine learning, focusing on Artificial Neural Networks (ANN) and traditional ensemble methods. By analyzing datasets from telecommunications and subscription platforms, this study emphasizes the importance of feature engineering, class imbalance management, and hyperparameter tuning. Ensemble methods, such as

Random Forest, Gradient Boosting, and XGBoost, showed stable accuracy (~80%) across both datasets, with simpler models, such as Gaussian Naive, also performing well. Metaheuristic optimization techniques, including Genetic Algorithms and Particle Swarm Optimization, are valuable for hyperparameter tuning. Integrating behavioral and temporal data further improved predictions, particularly for subscription services. While some advanced models are computationally expensive, future research should focus on hybrid models that combine deep learning and optimization, as well as real-time churn prediction systems. This study also highlights the potential of customer segmentation and personalized interventions for better retention strategies.

8) References

[1] Srinivasan Rajendran and Rajeshwari Devrajan, "Customer Churn Prediction Using Machine Learning Approaches," International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), 2023. DOI: 10.1109/ICECONF57129.2023.10083813

[2] Senthilnayaki B, Swetha M, and Nivedha D, "CUSTOMER CHURN PREDICTION", International Advanced Research Journal in Science, Engineering and Technology, Vol. 8, Issue 6, June 2021. DOI: 10.17148/IARJSET.2021.8692

[3] B. Prabadevi, R. Shalini, and B.R. Kavitha, "Customer Churning Analysis Using Machine Learning Algorithms," International Journal of Intelligent Networks, Vol. 4, 2023, pp. 145–154. DOI: 10.1016/j.ijin.2023.05.005SciSpace

[4] Prachi Dahiya and Umang Kant, "Comparative Analogy Among Different Machine Intelligence Based Techniques for Customer Churn Prediction", 2023.

[5] Sharmila K. Wagh, Jayshree R. Pansare, Aishwarya A. Andhale, Sarita P. Ambadekar, Kishor S. Wagh, and S.H. Gawande, "Customer Churn Prediction in Telecom Sector Using Machine Learning Techniques", Results in Control and Optimization, Volume 14, March 2024. DOI: 10.1016/j.rico.2023.100144

[6] Nisha Gurung, Md Rokibul Hasan, Md Sumon Gazi, and Faiaz Rahat Chowdhury, "AI-Based Customer Churn Prediction Model for Business Markets in the USA: Exploring the Use of AI and Machine Learning Technologies in Preventing Customer Churn", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol. 9, Issue 2, 2023. DOI: 10.32628/CSEIT239211

[7] Jeetesh Tiwari and Arpita Shukla, "Bank Customer Churn Prediction Using Machine Learning," 2023 International Conference on Innovative Computing (ICIC). DOI: 10.1109/ICIC60129.2023.10383749

[8] Srinivasan Rajendran and Rajeshwari Devrajan, "Customer Churn Prediction Using Machine Learning Approaches", ICECONF 2023. DOI: 10.1109/ICECONF57129.2023.10083813Scribd

[9] Senthilnayaki B, Swetha M, and Nivedha D, "CUSTOMER CHURN PREDICTION", IARJSET, Vol. 8, Issue 6, June 2021. DOI: 10.17148/IARJSET.2021.8692Academia+2ResearchGate+2Scribd+2

[10] A. Senthilkumar and B. Subash Chandra Bose, "Customer Churn Prediction Using Composite Deep Learning Technique", Materials Today: Proceedings, 2023. DOI: 10.1016/j.matpr.2023.09.368

[11] Tauseef Ahmad and Abid Ali Minhas, "Customer Churn Prediction Using Machine Learning Techniques" 2023 4th International Conference on Computing and Big Data (ICCBD). DOI: 10.1109/ICCBD57450.2023.00025

[12] Amrit Pal et al., "Customer Churn Prediction Using Composite Deep Learning Technique," 2023 International Conference on Computational Performance Evaluation (ComPE). DOI: : 10.1109/ComPE57084.2023.00037

[13] Swati Mahajan, Shruti Patil, and Sushma Agarwal, "Neural Network Based a Comparative Analysis for Customer Churn Prediction", 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). DOI: 10.1109/ICACITE53722.2022.9823726

[14] Debashish Senapati et al., "Customer Churn Prediction Using Composite Deep Learning Technique," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC). DOI: 10.1109/ICCMC57318.2023.10070712

[15] Moshood, A., Hambali., Ishaku, Andrew. (2024). Bank Customer Churn Prediction Using SMOTE: A Comparative Analysis. Qeios, doi: 10.32388/h82xtw

[16] B. Prabadevi, R. Shalini and B.R. Kavitha, "Customer churning analysis using machine learning algorithms" in International Journal of Intelligent Networks, Volume 4, pp. 145-154, 2023. DOI: https://doi.org/10.1016/j.ijin.2023.05.005.

[17] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1-24.

[18] Joydeb, Kumar, Sana., Mohammad, Zoynul, Abedin., M., Sohel, Rahman., Md., Saidur, Rahman. (2022). A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection. PLOS ONE, doi: 10.1371/journal.pone.0278095

[19] Sreenitya, Mandava., Anvita, Gupta., Komanduri, Srikar. (2022). Customer Churn Prediction in Banking Sector- A Hybrid Approach. International Journal For Science Technology And Engineering, doi: 10.22214/ijraset.2022.47541

[20] Ke, Peng., Yan, Peng., Wenguang, Li. (2023). Research on customer churn prediction and model interpretability analysis. PLOS ONE, doi: 10.1371/journal.pone.0289724