ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2024.1631.269

MACHINE LEARNING-DRIVEN STRATEGIES FOR CUSTOMER RETENTION AND FINANCIAL IMPROVEMENT

N. Rakesh[1], B.A. Mohan[2], U. Kumaran[3], G.L. Prakash[4], Rajakumar Arul[5], Kalaipriyan Thirugnanasambandam[6]

[1]Associate Professor, Department of Information Science and Engineering, BMS Institute of Technology and Management, Bengaluru, Karnataka, India.
e-mail: n_rakesh@bmsit.in, orcid: https://orcid.org/0000-0001-8966-5831
[2]Associate Professor, Department of Information Science and Engineering, BMS Institute of Technology and Management, Bengaluru, Karnataka, India.
e-mail: ba.mohan@bmsit.in, orcid: https://orcid.org/0000-0002-0711-1550
[3]Assistant Professor, Selection Grade, Department of CSE, Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, India.
e-mail: u_kumaran@blr.amrita.edu, orcid: https://orcid.org/0000-0002-0160-2703
[4]Associate Professor, Department of Information Science and Engineering, BMS Institute of Technology and Management, Bengaluru, Karnataka, India.
e-mail: glprakash@bmsit.in, orcid: https://orcid.org/0000-0002-0724-8469
[5]Assistant Professor, Centre for Smart Grid Technologies (CSGT)/, School of Computer Science and Engineering (SCOPE), VIT Chennai, Vandalur-Kelambakkam Road, Chennai, India. e-mail: rajakumar.arul@vit.ac.in,
orcid: https://orcid.org/0000-0002-1385-7965
[6]Senior Assistant Professor, Centre for Smart Grid Technologies, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai, Tamilnadu, India. e-mail: kalaipriyan.t@vit.ac.in,
orcid: https://orcid.org/0000-0001-9512-4687

SUMMARY

In the telecom sector, which generates vast amounts of data daily due to its extensive client base, retaining present customers is more cost-effective than attaining new ones. Business analysts and CRM specialists must comprehend the reasons behind customer churn and identify behavioral patterns within client data. This study develops a churn forecast model employing clustering and classification algorithms to recognize churn consumers and highlight the factors influencing customer churn in the telecom industry. Feature selection is done using information gain and correlation feature ranking filters. The Random Forest (RF) algorithm achieved superior performance, correctly classifying 88.63% of churned customer data. An essential CRM function is to formulate effective retention strategies to prevent customer departure. Post-classification, the proposed model clusters the churned customer data and provides group-based retention strategies using cosine similarity among the groups. The performance of the model is assessed through metrics like precision, accuracy, recall, f-score, and ROC-AUC. The findings indicate that the RF algorithm enhances churn classification and customer profiling through k-means clustering, and the classification algorithm helps identify the factors driving customer churn through generated rules.

Key Words: *machine learning churn prediction model, random forest, ROC.*

Received: August 17, 2024; Revised: October 22, 2024; Accepted: November 12, 2024; Published: December 24, 2024

INTRODUCTION

In today’s dynamic business landscape, customer shake poses a significant challenge for companies across various industries. The phenomenon of customers discontinuing their relationship with a company, not only impacts revenue but also reflects customer satisfaction and loyalty [1]. Under-standing the underlying factors driving churn and implementing strategies to mitigate it are crucial for sustainable growth and competitiveness. Machine learning techniques offer a powerful approach to churn analysis, enabling organizations to delve deep into their data to uncover patterns, trends, and predictive insights. By leveraging advanced algorithms and predictive modeling, industries can proactively identify customers at risk of churning and implement retention strategies [3]. This work aims to explore and implement machine learning techniques for churn analysis, focusing [1] on predicting customer churn and understanding the key drivers behind it. By utilizing historical customer data, demographic information, transactional records, and behavioral patterns, we seek to build robust predictive models capable of accurately identifying potential churners.

A. Motivation

In today’s hyper-competitive marketplaces, where customer expectations are constantly evolving, retaining existing customers is just as critical as acquiring new ones [22]. Customer churn, the loss of customers over time, not only represents lost revenue but also signifies missed opportunities for long-term relationships and brand advocacy [3]. Therefore, businesses across industries are increasingly recognizing the importance of churn analysis as a strategic imperative for sustainable growth.

B. Objective

A study that sought to address consumer loyalty outcomes in prepaid mobile phone companies expanded on this work. The prediction process in this paper is two-step. First, four clustersof similar attributes are created. Next, the first step’s extracted churn data is assessed using many methods, including DecisionTrees and Neural Networks [18].

C. Problem Statement

Churning of customer refers when clients decide to end their relationships with a corporation. Companies prefer the data analysts to forecast and prevent the loss of customers. Every firm needs to forecast the risk of churning of its client, especially when there is still time to prevent them from dropping services offered by the firm. Besides the direct loss of profits because of this, it is always difficult to gain a new customer than it is to retain client who is currently subscribed to the service [4].

D. Proposed System

In today's competitive market, attracting new customers can be five to ten times more expensive than maintaining the satisfaction and loyalty of existing customers. Businesses typically receive a customer loss rate of 10% to 30% annually. Recognizing this trend, many organizations are now employing customer retention and satisfaction strategies [19]. In sectors that rely heavily on subscription models, such as banking, telecommunications, insurance, and client relationship management, businesses manage large client bases. These companies generate their income through regular payments from their consumers. To maintain profitability, it's crucial for these businesses to retain their consumers' loyalty while minimizing operational costs. This approach helps ensure a steady revenue flow with the least possible cost [9].

LITERATURE SURVEY

From [1], proposes a hybrid churn prediction model for the telecommunication industry. The hybridmodel suggests a comprehensive approach, likely integrating both traditional statistical methods and machine learning techniques. A critical analysis would need to delve into the effectiveness of this hybrid approach compared to existing models, the datasets used, and the evaluation metrics employed. This paper [2], presents a churn prediction model for telecommunication subscribers utilizing machine learning techniques. Given the rapid advancements in machine learning, this paper's methodology and algorithmic choices would be of interest. The critical analysis should examine the model's predictive performance, scalability, and generalizability across different telecommunication datasets. Additionally, discussing the novelty of the approach compared to existing literature would be valuable. A study of data mining methods for customer churn prediction in the telecommunication industry was conducted [3].

This paper likely provides a synthesis of various data mining techniques applied in churn prediction. The critical analysis should evaluate the comprehensiveness of the review, highlighting the strong suit and weak suit of different techniques and identifying gaps in the existing literature that warrant further research [3]. A survey on churn prediction in telecommunication, focusing on classification techniques based on data mining [4]. Like the previous paper, this survey would require critical evaluation regarding the scope, depth, and relevance of the classification techniques discussed. Assessing the adequacy of the surveyed techniques in addressing real-world challenges in churn prediction would be crucial.

A study on churn prediction [5] in mobile telecom systems employing data mining techniques. This paper's contribution likely lies in its application of specific data mining techniques to address churn prediction in mobile telecom systems. A critical analysis should examine the effectiveness of these techniques, their computational efficiency, and their scalability to large-scale telecom datasets [5].

In their research, Kirui et al. explore the use of probabilistic data mining methods to forecast customer attrition in the mobile phone sector. Their study emphasizes the application of these classification methods to forecast when customers might drop their services [6]. The use of probabilistic classifiers suggests a probabilistic approach to churn prediction, which may offer insights into the uncertainty associated with churn predictions. The critical analysis should delve into the performance of these classifiers compared to deterministic approaches and discuss the implications for practical churn management strategies. The paper [7], discuss the application of data mining techniquesin for churn prediction. This paper likely presents specific data mining techniques tailored to the telecom churn prediction domain. A critical analysis should assess the suitability of these techniques for real-world telecom datasets, considering factors such as data sparsity, class imbalance, and feature relevance.

The paper [8], focus on churn estimate in telecommunication using data mining technology. This paper's critical analysis should evaluate the effectiveness of data mining techniques addressing the unique challenges of churn prediction in telecommunication industry. Assessing the scalability, interpretability, and computational efficiency of the proposed approach would be essential. Contribution lies in the empirical analysis of customer churn patterns in the mobile industry [9]. A critical analysis should assess the robustness of the findings, the generalizability of the results to other mobile telecom markets, and the implications for churn management strategies [9]. Propose a churn prediction model for telecommunication. This paper's critical analysis should evaluate the novelty and effectiveness of the proposed churn prediction model [10]. Assessing the model's predictive performance, scalability, and ease of implementation would be crucial for determining its practical utility in real-world telecom domain. These critical analyses provide a framework for evaluating the contributions, methodologies, and implications of each paper within the context of churn prediction in the telecommunication industry. The work [11], discusses the efficiency of ensemble methods for customer retention. Ensemble methods are known for their ability to improve predictive performance by combining multiple models. A critical analysis of this paper should evaluate the specific ensemble techniques proposed, their effectiveness in customer retention compared to individual models, and any practical considerations for implementing ensemble approaches in real-world customer retention scenarios. This paper explores the purpose of artificial intelligence in e-commerce [12]. Artificial intelligence (AI) has significant implications for electronic commerce, ranging from personalized recommendations to fraud detection [2] [8]. A critical analysis should assess the specific AI techniques discussed, their impact on different aspects of electronic commerce, and any challenges or ethical considerations associated with their implementation.

The study [20], discuss the outlook of e-commerce systems in coming years, particularly beyond the year 2030. Critical Analysis: Predicting the outlook of e-commerce systems involves anticipating technological advancements, changes in consumerbehavior, and regulatory developments. A critical analysis should evaluate the authors' predictions considering current trends and emerging technologies, considering factors suchas AI, blockchain, augmented reality, and sustainability, and assess the feasibility and potential implications of these predictions. The paper [14], propose genetic modelling for customer retention. Critical Analysis: Genetic algorithms offer a unique approach to modeling customer retention by mimicking natural selection processes. A critical analysis should evaluate the suitability of genetic algorithms for modeling customer retention compared to traditional statistical or machine learning approaches, considering factors such as interpretability, scalability, and computational efficiency. In present sequential pattern analysis for network banking churn [15]. Analysis can uncover valuable insights into customer behavior leading to churn. A critical analysis should evaluate the effectiveness of goal-oriented sequential pattern analysis in identifying churn indicators compared to other techniques. Additionally, [15] assessing the practical utility of the discovered patterns for preemptive churn management strategies would be essential.

From the paper [16], apply data mining to insurance consumer churn management. Critical Analysis: Insurance client churn presents unique challenges due to the long- term nature of insurance contracts [17]. A critical analysis should evaluate the effectiveness of data mining techniques in identifying churn risk factors specific to the insurance industry. Additionally, assessing the practical implications of churn management strategies informed by data mining insights would be crucial. In propose an efficient classifier for forecasting churn in telecommunication industry [21]. Critical Analysis: This paper likely focuses on the development or evaluation of a specific churn prediction classifier tailoredto the telecommunication industry. A critical analysis should assess the classifier's predictive performance, scalability, and generalizability across different telecom datasets. Additionally,discussing the practical implications of using this classifier forchurn management in telecom companies would be important. These critical analyses provide insights into the contributions, methodologies, and implications of each paper within their respective domains of customer churn prediction and management.

EXISTING SYSTEMS

Existing systems for churn prediction vary across industriesand can be broadly categorized into the following types:

1. **Telecommunication Systems:** Telecommunication companies often utilize churn prediction systems to identify customers at risk of leaving their services. These systems typically analyze usage patterns, customer demographics, billing information, and customer service interactions to identify churnindicators. Examples include:

   - Customer Relationship Management (CRM) systems with built-in churn prediction modules.
   - Proprietary churn prediction software developed in-house by telecommunication companies.
   - Third-party analytics platforms specializing in churn pre- diction for telecommunication providers [4].

2. **Banking and Financial Services Systems:** Banks and financial institutions deploy churn prediction systems to reduce customer attrition, particularly in the context of banking services, credit cards, and investment products [13]. These systems analyze transaction histories, account activities, customer demographics, and market trends to forecast potential churn. Examples include:

- Customer segmentation and retention modules integratedinto banking software suites.
- Predictive analytics platforms tailored for the bankingsector, offering churn prediction as a service.
- In-house data science teams developing custom churn prediction models using machine learning algorithms [4].

3. **E-commerce and Retail Systems:** E-commerce companies and retail chains employ churn prediction systems to retain customers and maximize lifetime value. These systems analyze purchase histories, browsing trends, product preferences, and demographic data to identify consumers likely to churn. Examples include:

   - Customer analytics tools integrated into e-commerce plat- forms, such as Shopify or Magento.
   - Marketing automation platforms with built-in churn prediction features, like HubSpot or Marketo.
   - In-house data analytics teams leveraging machine learning algorithms to develop custom churn prediction models [3].

4. **Insurance Systems:** Insurance companies utilize churn pre- diction systems to minimize policy cancellations and retain policyholders. These systems analyze policyholder behavior, claim histories, premium payments, and demographic factors to forecast churn risk. Examples include:

   - Policy management software with churn prediction mod- ules tailored for the insurance industry.
   - Predictive modeling platforms specializing in insuranceanalytics and customer retention.
   - Custom-built churn prediction solutions developed bydata science teams within insurance companies [4].

5. **Utility and Energy Systems:** Utility providers, including electricity, water, and gas companies, implement churn pre-diction systems to reduce customer attrition and improve customer satisfaction. These systems analyze usage patterns, billing data, customer inquiries, and demographic information to predict churn likelihood. Examples include:

   - Customer engagement platforms with churn prediction capabilities designed for utility companies.
   - Data analytics solutions customized for the energy sector, offering churn prediction as part of their feature set.
   - In-house churn prediction models developed by utilities using data science expertise and industry-specific knowledge.

Overall, existing churn prediction systems leverage a combination of data sources, analytics techniques, and domain-specific knowledge to forecast customer churn and enable proactive retention strategies.

PROPOSED SYSTEM DESIGN

A. System Architecture

The Figure 1 shows the system architecture of proposed work. The following steps are followed during the processingof data. Use Case Diagram shown in Figure 2.

- **Data Collection and Preprocessing:** Data can be collected from various sources such as CRM systems, transaction databases, customer service logs, etc. Preprocess to address missing values, outliers, and inconsistencies in the data. This can include methods like data cleaning, feature engineering and normalization.

- **Feature Engineering:** Extract significant features from the data to forecast churn. Features like demographic information, purchase history, customer interactions, etc. are considered. Further these features may need to be transformed or encoded appropriately for machine learning algorithms.

- **Model Training:** Separate the data into training and validation sets. Select suitable machine learning algorithms for churn prediction, such as logistic regression, random forests, decision trees, or gradient boosting algorithms. Train multiple models using different algorithms and hyperparameters to find the best performing model.

- **Model Evaluation:** Evaluate the trained models using appropriate metrics like accuracy, precision, F1-score, recall, and ROC-AUC. Also perform cross-validation to ensure the model's generalization ability.

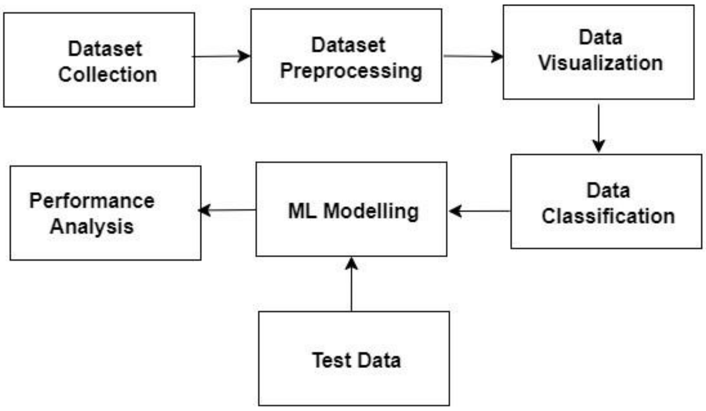


Figure 1. System Architecture

B. Use Case Diagram

1. **Actors: Customer:** Represents the users whose churn behavior is being analyzed.

- **Data Analyst:** Analyzes and preprocesses the data before feeding it into the machine learning model.

- **Machine Learning Model:** The system itself, which predicts customer churn based on the input data.

- **Administrator:** Manages the deployment and maintenanceof the machine learning model.

- **Business Decision Maker:** Uses insights from the churn analysis to make strategic decisions.

2. **Use cases:**

- **Collect Data:** The system collects data from sources like CRM systems, transaction databases, etc.

- **Preprocess Data:** Data Analyst preprocesses the collecteddata to deal with missing values, outliers, etc.

- **Train Model:** The system trains the machine learning model using the data which is preprocessed.
- **Evaluate Model:** Data Analyst evaluates the trained model's performance using validation data.
- **Deploy Model:** Administrator deploys the trained model into production.
- **Monitor Model:** Administrator monitors the deployed model's performance in real-time.
- **Update Model:** Data Analyst periodically updates the model using new data.
- **Generate Insights:** Business Decision Maker uses insightsgenerated by the model to make strategic decisions for customer retention.
- **Visualize Results:** Users interact with the system to visualize churn trends, model performance, and key insights.

3. **Relationships:** Include Relationship: Data preprocessing, model training, evaluation of model, model deployment, monitoring, and model updating are all included in the main use case of "Customer Churn Analysis".

   - **Extend Relationship:** The "Visualize Results" use casemay extend from "Customer Churn Analysis" when usersneed to interactively explore the analysis results.

4. **System Boundary:** The system boundary encompasses all the actors and use cases involved in the customer churn analysis process.

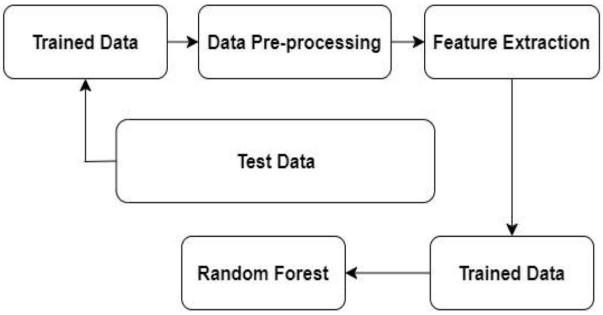


Figure 2. Use Case Diagram

C. Activity Diagram

1. **Data Collection and Preprocessing:** Acquire data from various resources such as CRM systems, transaction databases, etc. Preprocess the data for handling missing values, outliers, etc.
2. **Feature Engineering:** Extract relevant features from the preprocessed data. Transform or encode features as needed formachine learning.
3. **Model Training:** Split the dataset into training and validation sets. Train multiple models using different algorithms. Evaluate models using validation set.
4. **Model Selection:** Choose the best model based on evaluation metrics. Then deploy the selected model into production.
5. **Feedback Loop:** Continuously collect new data for retraining. Use model predictions to improve the model over time.
6. **Visualization and Reporting:** Generate dashboards and reports to visualize churn trends and model performance. Provide actionable insights and recommendations.

D. System Design

It gives the overall working framework of the proposedsystem and the flow of data between the various components shown in Figure 3.

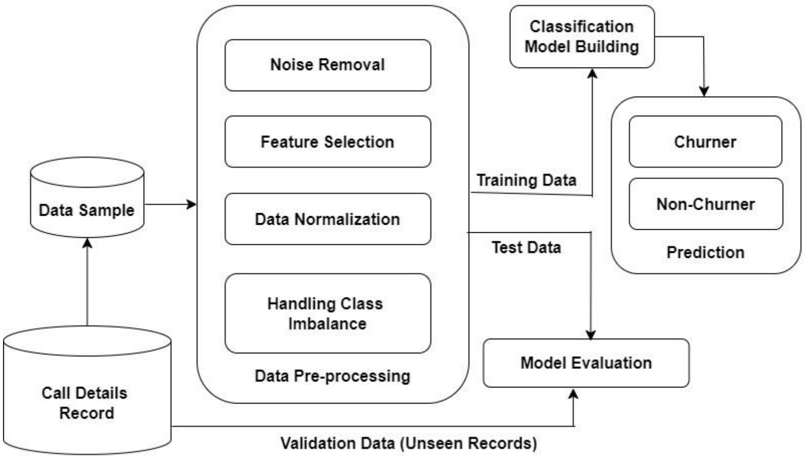


Figure 3. System Design

IMPLEMENTATION

A. Methodology

Data Preprocessing: Examining the variety of values in each column provides valuable intuitions. However, the varied ranges across different columns can lead to certain values overshadowing others. To address this issue, we used Min-Max scaling technique to normalize the data, as illustrated in Figure 4.

From sklearn.preprocessing import MinMaxScaler

$featureslogtransformed = pd.DataFrame(data{=}data[num_{c}ols])$

$featureslogtransformed[num_{c}ols] = data[num_{c}ols].apply(lambdax : np.log(x + 1))$

$scaler = MinMaxScaler()$

featureslogmin$_{m}ax_{t}ransform = pd.DataFrame(data = featureslogtransformed)$

featureslogmin$_{m}ax_{t}ransform[num_{c}ols] \quad = scaler.fittransform(featureslogtransformed[num_{c}ols])$

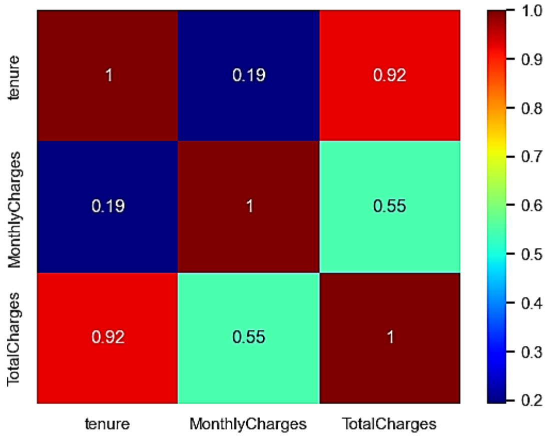


Figure 4. Correlation of different numerical data

Following data preprocessing and cleaning stages, we move on to model implementation. This process is executed in three primary phases.

- Splitting data into two sets - training and testing set.
- Designing classifier () function for applying different algorithms.
- Creating grid search () function for hyper parameter tuning.

For the first part, the data used 30% for testing and 70%for training. In applying the classifier() function we applied different supervised algorithms to the data and then showed the results of metrics and plotted the AUC curve. Then we defined hyperparameter and sent it to the grid search () function to findthe best values. we used AUC For the scoring for the grid search. For the first step, we applied decision tree since it is the base model and considered it as base model for comparing with other models. Then we applied Logistic Regression, Support Vector Machine, Random Forest, and XGboost.

In Table 1 and Table 2, we can see the result of algorithms before and after parameter tuning respectively. The challenge in the coding part was to write one helper function that could run all the algorithms. Since there were two plots for each algorithm, confusion matrix, and RUC curve, it was time-consuming to put everything in the right places usingthe subplot function.

Table 1. Metric for default values of algorithms

Models	Precision	Recall	F1-score	AUC
Decision Tree	0.73	0.73	0.73	0.6621
Logistic Regression	0.79	0.80	0.79	0.7090
SVM	0.78	0.79	0.77	0.6677
Random Forest	0.76	0.78	0.76	0.6649
XGBoost	0.79	0.80	0.79	0.7034

Table 2. Metric after hyper parameters tune with grid search

Models	Precision	Recall	F1-score	AUC
Decision Tree	0.78	0.77	0.78	0.7296
Logistic Regression	0.78	0.80	0.79	0.7014
SVM	0.78	0.80	0.78	0.6903
Random Forest	0.80	0.74	0.75	0.7525
XGBoost	0.78	0.78	0.78	0.7015

B. Algorithms and Techniques

The prediction is done using various supervised machine learning techniques like Logistic Regression, Decision Tree, S.V.M., Random Forest, and XGBoost and then compared the results to check which one predicts better.

- **Decision Tree:** A decision tree is a graph kind of structure in which each node represents a feature and each branch represents a decision made as shown in Figure 5. It is a widespread algorithm for both classification and regression problems.

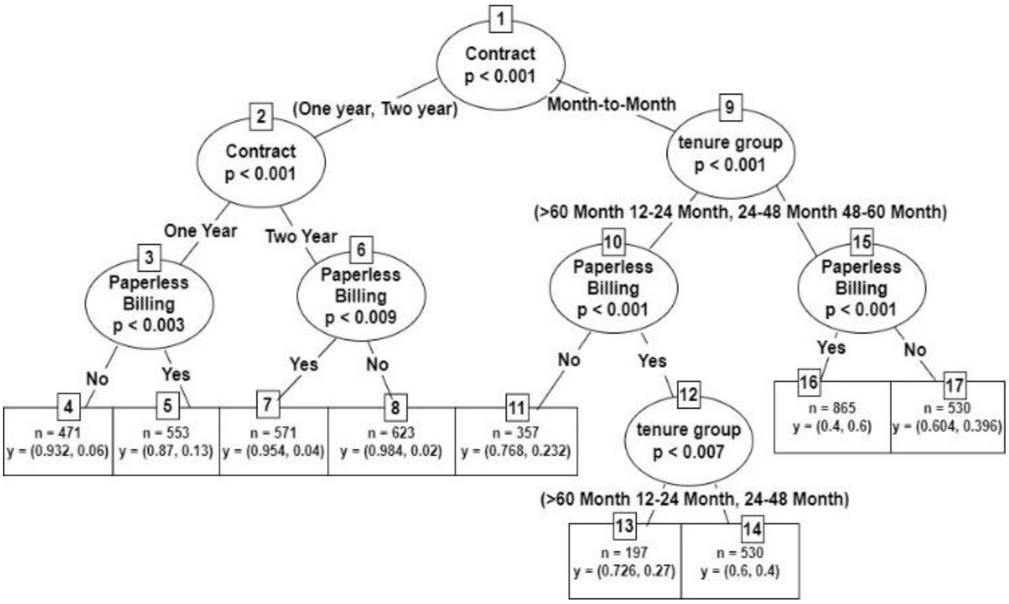


Figure 5. Decision Tree

Logistic Regression: Logistic regression is a statistical technique used to analyze datasets. It requires one or more predictor variables and features a binary outcome variable. In this model, the dependent variable is dichotomous, represented as 1 for a positive result and 0 for a negative result. The sigmoid function as follows shown in Eq. (1).

$$\mathrm{f}(x) = \frac{1}{1 + e - ^{(x)}} \qquad \ldots\ldots (1)$$

The end goal of Logistic Regression is to find the line that separates the data shown in Figure 6.

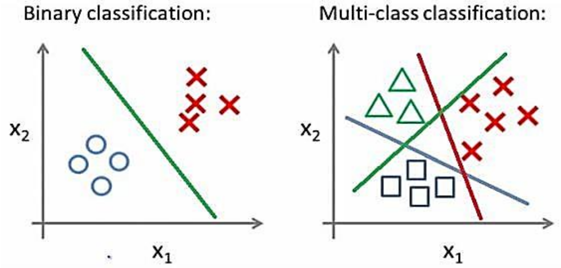


Figure 6. Logistic Regression

The process begins by implementing a linear equation, followed by applying the Sigmoid function to the outcome. This transforms the results into a range between 0 and 1. The next step involves determining the error by reducing the cost function to its min value. Finally, we use the Gradient descent method, as depicted in Eq. (2).

$$\mathrm{J}(\theta) = \frac{1}{m \sum_{n-1}^{m} Cost\big(h_\theta\{(x\}^{i})\big), y(i)} \qquad \ldots. (2)$$

Support Vector Machine: The SVM illustrated in Figure 7, is a supervised learning algorithm in machine learning. It's adaptable, capable of handling both classification and regression tasks. In the SVM approach, each data point is denoted in an n-dimensional coordinate space, where n is the number of attributes. Each feature's value corresponds to a specific coordinate in this space. The algorithm then performs classification by identifying the optimal hyperplane that separates the different classes of data points.

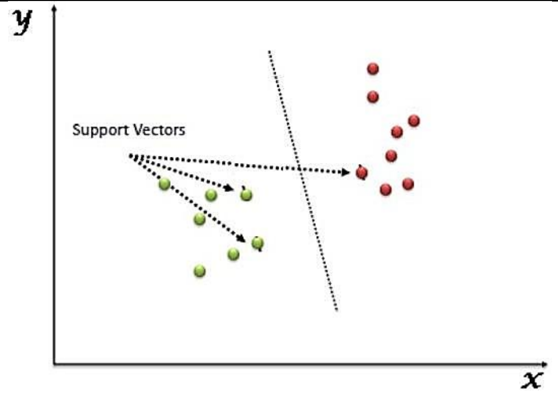


Figure 7. SVM

Random Forest: is a versatile machine learning technique applicable to both classification and regression problems. It activates by generating various decision trees during the training process. The final result is determined by the most common class among the individual trees' predictions. When each $hk(x)$ represents a decision tree, the collection of decision trees forms a random forest. The Kth decision tree results in a classifier, as formulated in Eq. (3).

$$h_k(x) = h(\frac{x}{\theta}) \qquad ………. (3)$$

XGBoost: XGBoost is a gradient-boosted decision tree implementation tuned for rapid learning and enhanced efficiency [10]. The boosting process involves three key steps:

- Initial predictive model $m(0)$ is established for the target variable y.
- A subsequent model $m(1)$ is developed to fit the residuals from the preceding step.
- The models $m(0)$ and $m(1)$ are then merged to create $N(1)$, which represents the boosted iteration of $m(0)$.

RESULTS AND DISCUSSION

A. Model Evaluation and Validation

The best result before up-sampling: Up-sampling in AdaBoost involves assigning higher weights to the misclassified instances in each iteration to focus more on the complex samples. This technique aims to alleviate the imbalance issue by giving more weight to the minority class during the training process. By iteratively adjusting the weights of the misclassified instances, AdaBoost aims to improve the classification performance, particularly for imbalanced datasets presented in Figure 8. Through up-sampling, AdaBoost can effectively handle datasets with unequal class distributions, boosting the performance of the classifier on minority class samples without introducing bias towards the majority class. Classification report before upsampling shown in Table 3.

Table 3. Classification report before upsampling

Class	Precision	Recall	F1-score	AUC
0	0.91	0.71	0.8	1549
1	0.51	0.81	0.62	561
Avg/Total	0.8	0.74	0.75	2110

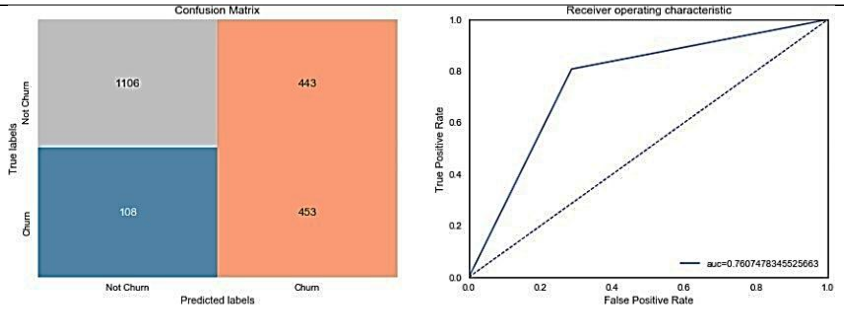


Figure 8. Adaboost method before up-sampling

The best result after up-sampling: In the context of AdaBoost, up-sampling after classification typically involves adjusting the weights of the misclassified instances to emphasize their importance in subsequent iterations presented in Table 4. This post-classification up-sampling technique aims to correct errors made by the classifier by giving more weight to the misclassified samples, allowing AdaBoost to iteratively refine its model to better classify difficult-to-distinguish instances. By adjusting the weights of the misclassified samples and retraining the model, AdaBoost strives to improve its over-all performance and enhance its ability to correctly classify challenging data points as shown in Figure 9.

Table 4. Classification report after upsampling

Class	Precision	Recall	F1-score	AUC
0	0.95	0.85	0.9	1571
1	0.86	0.95	0.9	1527
Avg/Total	0.91	0.9	0.9	3098

In both cases, Adaboost integrated with random forest gives best results. It can be because of Adaboost uses some weak learners and combined the result with a strong learner, which can give a better classification accuracy in this dataset. Also, upsampling made the result better since it prevents the algorithms bias to majority class.

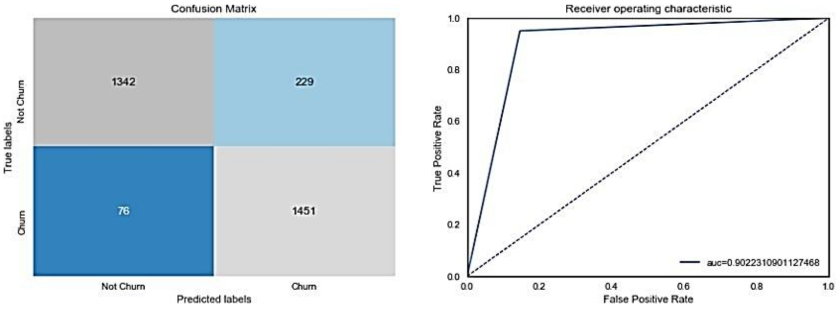


Figure 9. Adaboost method after unsampling

B. Justification

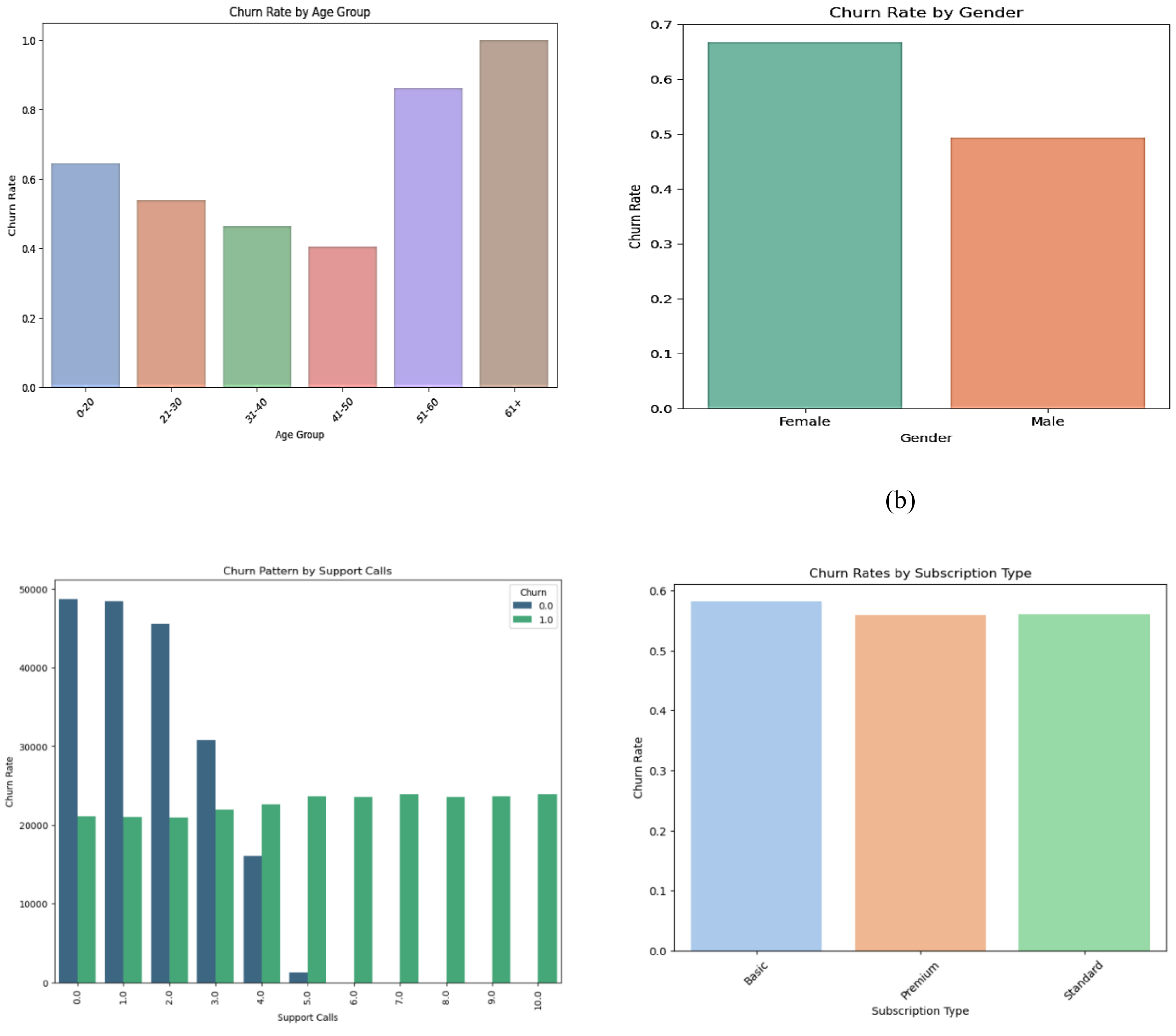
When comparing a benchmark model to a final model using metrics like precision, F1 score, recall, and AUC, we can observe significant improvements. The benchmark model typically serves as a baseline, showcasing the initial performance of the system. In contrast, the final model represents the optimized version after refining and tuning. Precision calculates the fraction of correct positive forecasts among all positive forecasts made. Recall on the other hand decides fraction of actual positive instances that were properly recognized. The F1-score serves as a balanced metric, combining both precision and recall into a single value. Additionally, the AUC (Area Under the ROC Curve) measures the model's ability to differentiate between positive and negative classes across various thresholds. Typically, we expect to see higher precision, F1-score, recall, and AUC values in the final model compared to the benchmark, indicating improved predictive performance and better discrimination between classes. Comparing final model and benchmark model shown in Table 5.

Table 5. Comparing final model and benchmark model

Model	Precision	Recall	F1-score	AUC
Bench Mark Model	0.79	0.80	0.79	0.6621
Final Model	0.91	0.90	0.90	0.9022

C. Visualization Process of Data Analysis

Data visualization of dataset columns offers a graphical representation of the dataset's characteristics, facilitating insights into patterns, trends, and relationships within the data. These visualizations not only aid in understanding the structure of the dataset but also assist in identifying potential data preprocessing steps and informing feature selection strategies for machine learning tasks. Figure 10 shows dataset analysis using graphs based on different parameters like churn rate by age group, gender, subscription type and churn pattern.



(b)

(c)

(d)

Figure 10. Data Visualization of dataset columns (a) churn rate by age group (b) churn rate by gender (c) churn pattern by support calls (d) churn rate by subscription type

CONCLUSION

To sum up, this study offers a thorough customer churn forecasting system for the telecommunications industry. It employs both classification and clustering methods to recognize and analyze the patterns of customers likely to discontinue services. Feature selection is done using information gain and correlation feature ranking filters. With the deployment of Random Forest for classification achieved superior performance, correctly classifying 88.63% of churned customer data and post-classification, k-means clusters the churned customer data and provides group-based retention strategies using cosine similarity among the groups, the model achieved high accuracy in predicting churn. Additionally, the model's ability to generate rules for churn factorsaids in devising effective retention strategies. The proposed approach offers valuable insights for CRM, enabling themto enhance customer retention efforts and optimize marketing campaigns.

REFERENCES

[1] Olle GD, Cai S. A hybrid churn prediction model in mobile telecommunication industry. International Journal of e-Education, e-Business, e-Management and e-Learning. 2014 Feb 1;4(1):55-62. https://doi.org/10.7763/IJEEEE.2014.V4.302

[2] Rami S, Abrar S, Mohammed AA, Tayseer A, Belal MA, Mahmaod A. Assessment of Cybersecurity Risks and threats on Banking and Financial Services. Journal of Internet Services and Information Security. 2024;14(3):167-90. https://doi.org/10.58346/JISIS.2024.I3.010

[3] Dahiya K, Talwar K. Customer churn prediction in telecommunication industries using data mining techniques-a review. International journal of advanced research in computer science and software engineering. 2015;5(4):417-33.

[4] Saini N. Churn Prediction in Telecommunication Using Classification Techniques Based on Data Mining: A Survey. International Journal of Advanced Research in Computer Science and Software Engineering. 2015 Mar;5(3).

[5] Balasubramanian M, Selvarani M. Churn prediction in mobile telecom system using data mining techniques. International Journal of scientific and research publications. 2014 Apr;4(4):1-5.

[6] Kirui C, Hong L, Cheruiyot W, Kirui H. Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining. International Journal of Computer Science Issues (IJCSI). 2013 Mar 1;10(2 Part 1):165-172.

[7] Kamalraj N, Malathi A. Applying data mining techniques in telecom churn prediction. International Journal of Advanced Research in Computer Science and Software Engineering. 2013 Oct;310:363-70.

[8] Srinadi NL, Hermawan D, Jaya AA. Advancement of banking and financial services employing artificial intelligence and the internet of things. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications. 2023;14(1):106-17. https://doi.org/10.58346/JOWUA.2023.I1.009

[9] Churi A, Divekar M, Dashpute S, Kamble P. Analysis of customer churn in mobile industry using data mining. International Journal of Emerging Technology and Advanced Engineering. 2015 Mar;5(3): 225-30.

[10] Shaaban E, Helmy Y, Khedr A, Nasr M. A proposed churn prediction model. International Journal of Engineering Research and Applications. 2012 Jun;2(4):693-7.

[11] Bhujbal NS, Bavdane GP. Leveraging the efficiency of Ensembles for Customer Retention. In 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) 2021 Nov 11 (pp. 1675-1679). IEEE. https://doi.org/10.1109/I-SMAC52330.2021.9640757

[12] Song X, Yang S, Huang Z, Huang T. The application of artificial intelligence in electronic commerce. In Journal of Physics: Conference Series 2019 Aug 1 (Vol. 1302, No. 3, p. 032030). IOP Publishing. https://doi.org/10.1088/1742-6596/1302/3/032030

[13] Kalinin O, Gonchar V, Abliazova N, Filipishyna L, Onofriichuk O, Maltsev M. Enhancing Economic Security through Digital Transformation in Investment Processes: Theoretical Perspectives and Methodological Approaches Integrating Environmental Sustainability. Natural and Engineering Sciences. 2024 May 5;9(1):26-45. https://doi.org/10.28978/nesciences.1469858

[14] Eiben AE, Koudijs AE, Slisser F. Genetic modelling of customer retention. In European Conference on Genetic Programming 1998 Apr 14 (pp. 178-186). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0055937.

[15] Chiang DA, Wang YF, Lee SL, Lin CJ. Goal-oriented sequential pattern for network banking churn analysis. Expert systems with applications. 2003 Oct 1;25(3):293-302. https://doi.org/10.1016/S0957-4174(03)00073-3

[16] Soeini RA, Rodpysh KV. Applying data mining to insurance customer churn management. International Proceedings of Computer Science and Information Technology. 2012 Feb;30(2012):82-92.

[17] Thomas KP, Rajini DG. Evolution of Sustainable Finance and its Opportunities: A Bibliometric Analysis. Indian Journal of Information Sources and Services. 2024 Jun 28;14(2):126-32. https://doi.org/10.51983/ijiss-2024.14.2.18

[18] Qureshi SA, Rehman AS, Qamar AM, Kamal A, Rehman A. Telecommunication subscribers' churn prediction model using machine learning. In Eighth international conference on digital information

management (ICDIM 2013) 2013 Sep 10 (pp. 131-136). IEEE. https://doi.org/10.1109/ICDIM.2013.6693977

[19] Jadhav RJ, Pawar UT. Churn prediction in telecommunication using data mining technology. International Journal of Advanced Computer Science and Applications. 2011;2(2):17-19.

[20] Mohdhar A, Shaalan K. The future of e-commerce systems: 2030 and beyond. Recent Advances in Technology Acceptance Models and Theories. 2021:311-30. https://doi.org/10.1007/978-3-030-64987-6_18

[21] Pamina J, Raja B, SathyaBama S, Sruthi MS, VJ A. An effective classifier for predicting churn in telecommunication. Jour of Adv Research in Dynamical & Control Systems. 2019 Jun 6;11(01sp).

[22] Kapoor S, Sharma V. A Comprehensive Framework for Measuring Brand Success and Key Metrics. In Brand Management Metrics. 2024: 16-30. Periodic Series in Multidisciplinary Studies.