iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

CHURN ANALYSIS AND PREDICTION IN TELECOMBANKING CUSTOMERS

B. Udhayakumar
School of Computing Sciences, VISTAS, Chennai, India
udhayakmr246@gmail.com
Dr. K. Dharmarajan
Professor, School of Computing Sciences, VISTAS, Chennai, India
dharmak07@gmail.com

ABSTRACT
Customer churn is one of the most significant challenges faced by telecom and banking industries, as it directly impacts revenue generation, customer base stability, and long-term profitability. In today's highly competitive market, retaining existing customers is more cost-effective than acquiring new ones, making churn prediction an essential business strategy. This project focuses on analysing customer behavior and developing an intelligent system to predict churn using advanced machine learning techniques. The study utilizes a variety of features, including customer demographics (age, gender, location), service usage patterns (call duration, internet usage, subscription plans), transaction history, and customer satisfaction indicators. To build an effective prediction model, data pre-processing techniques such as data cleaning, handling missing values, feature scaling, and encoding categorical variables are applied. Exploratory Data Analysis (EDA) is performed to identify patterns, trends, and correlations between features and churn behavior. Multiple classification algorithms, including Logistic Regression, Decision Trees, Random Forest, and Extreme Gradient Boosting (XGBoost), are implemented and compared based on performance metrics such as accuracy, precision, recall, and F1-score. Among these, ensemble methods like Random Forest and XGBoost often provide higher accuracy due to their ability to handle complex patterns and reduce overfitting. The model's output helps in identifying customers who are at a high risk of leaving the company. Based on these predictions, businesses can take proactive measures such as personalized marketing strategies, targeted offers, improved customer support, and service enhancements. This not only reduces churn rate but also improves customer satisfaction, loyalty, and overall business performance.

Keywords:
Customer Churn, Machine Learning, XGBoost, Random Forest, Predictive Analytics, Telecom, Banking, Classification.

I. INTRODUCTION

Customer churn refers to the phenomenon where customers stop using a company’s products or services over a given period. It is a major concern for industries such as telecommunications and banking, where customer retention plays a crucial role in maintaining steady revenue and competitive advantage. With the rapid growth of digital services and increased market competition, customers now have multiple alternatives, making it easier for them to switch providers. As a result, understanding the reasons behind customer churn and predicting it in advance has become essential for businesses. Traditional methods of analyzing churn are often time-consuming and less accurate, which highlights the need for intelligent, data-driven approaches. In this context, machine learning techniques provide powerful tools to analyze large volumes of customer data, identify hidden patterns, and predict future churn behavior. By implementing predictive models, organizations can take proactive steps to retain customers, improve service quality, and enhance overall customer satisfaction.

The telecom and banking sectors share several common customer behavioral patterns that make joint analysis both feasible and insightful. Both sectors deal with service subscriptions, usage charges, complaints, and loyalty rewards. Research has shown that the cost of acquiring a new customer can be five to seven times higher than retaining an existing one. Therefore, even a small improvement in churn prediction accuracy can lead to significant financial savings. This project aims to develop an efficient churn prediction system that assists companies in making informed decisions and building long-term customer relationships.

The remainder of this paper is organized as follows: Section II reviews related literature; Section III presents the problem statement; Section IV outlines objectives; Section V describes the methodology; Section VI details the

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

system architecture; Section VII covers implementation; Section VIII presents results; Section IX discusses advantages and limitations; Section X considers future scope; and Section XI concludes the paper.

II. LITERATURE REVIEW

Customer churn prediction has been extensively studied across various domains. Several researchers have proposed different approaches ranging from statistical models to complex ensemble methods.

Verbeke et al. [1] conducted a comprehensive benchmarking study comparing various machine learning classifiers for telecom churn prediction. Their findings indicated that ensemble methods consistently outperform standalone classifiers. Similarly, Huang et al. [2] proposed a hybrid model that combined neural networks with support vector machines, achieving improved accuracy in predicting banking customer attrition.

Amin et al. [3] applied cost-sensitive learning to address the class imbalance problem commonly found in churn datasets, where churners are significantly fewer than non-churners. Their approach improved recall for the minority class without sacrificing overall accuracy. Mozer et al. [4] utilized recurrent neural networks to capture temporal dependencies in customer interaction data, demonstrating that sequential patterns strongly correlate with churn behavior.

Idris et al. [5] explored feature selection techniques such as information gain and chi-square tests to reduce dimensionality and improve model performance. Their results confirmed that selecting the top-ranked features not only speeds up training but also enhances predictive power. Tsai and Lu [6] integrated fuzzy logic with traditional classification algorithms, proposing a soft-boundary model that handles uncertain customer behavior more gracefully than binary classifiers.

In the banking domain, Xie et al. [7] developed a churn prediction model using gradient boosting and demonstrated that transaction frequency, account balance trends, and customer service interactions were among the most influential predictors. Ahmad et al. [8] carried out a comparative study of Logistic Regression, Naive Bayes, and Random Forest on telecom datasets, concluding that Random Forest yielded the highest F1-score due to its robustness against noisy features.

More recently, Lalwani et al. [9] investigated deep learning architectures, including long short-term memory (LSTM) networks, for customer churn prediction in subscription-based services. Their model achieved an AUC of 0.93, outperforming traditional approaches. Burez and Van den Poel [10] studied oversampling strategies such as SMOTE to balance class distribution, reporting a notable improvement in model sensitivity toward churning customers. Collectively, these works establish a solid foundation on which this research builds, emphasizing the importance of feature engineering, algorithmic diversity, and handling data imbalance.

III. PROBLEM STATEMENT

Despite the availability of large volumes of customer data, telecom and banking organizations struggle to predict customer churn accurately and in a timely manner. Conventional rule-based approaches lack adaptability to evolving customer behavior patterns, while simple statistical techniques often fail to capture non-linear interactions among predictive features. Additionally, the inherent class imbalance in churn datasets, where churners constitute only a small fraction of the customer base, poses significant challenges in building reliable classification models.

There is a pressing need for an intelligent, automated churn prediction framework that can: (i) process heterogeneous data from multiple sources; (ii) handle missing values and categorical variables effectively; (iii) address class imbalance; and (iv) identify at-risk customers with high precision and recall, enabling timely intervention strategies. This work addresses all these challenges through a structured machine learning pipeline.

IV. OBJECTIVES

The key objectives of this research are:

- To collect and integrate customer data from telecom and banking domains, including demographic, transactional, and service-usage features.
- To pre-process the raw data by handling missing values, removing outliers, encoding categorical variables, and normalizing numerical features.
- To perform exploratory data analysis (EDA) to uncover patterns and correlations that differentiate churning customers from loyal ones.
- To engineer informative features that improve the discriminative power of machine learning models.
- To implement and compare multiple classification algorithms, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

- To evaluate model performance using appropriate metrics such as accuracy, precision, recall, F1-score, and AUC-ROC curve.
- To identify the best-performing model and deploy it for real-time or batch churn prediction.
- To provide actionable business insights that enable proactive customer retention strategies.

V. METHODOLOGY

A. Data Collection

The dataset used in this study is sourced from publicly available telecom and banking churn repositories, supplemented with synthetic records to simulate real-world scenarios. The final dataset comprises approximately 10,000 customer records, each described by 21 features. Key attributes include: customer age, gender, geographic region, account tenure, monthly charges, total charges, number of service calls, internet service type, contract type, payment method, and a binary churn label (0 = retained, 1 = churned). The churn rate in the raw dataset is approximately 15%, reflecting typical industry figures.

B. Data Preprocessing

Raw data collected from operational systems frequently contains inconsistencies, missing entries, and outliers. The following preprocessing steps were applied: First, missing values in numerical columns were imputed using the median strategy to reduce sensitivity to extreme values, while categorical missing values were replaced with the most frequent category. Second, outliers in numerical features such as total charges were detected using the interquartile range (IQR) method and capped at the 1st and 99th percentiles. Third, categorical variables such as contract type and payment method were encoded using one-hot encoding to convert them into numerical format. Fourth, numerical features were standardized using StandardScaler from the scikit-learn library, ensuring that features with larger numeric ranges do not dominate the model. Fifth, since the dataset is imbalanced, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set to generate synthetic samples of the minority (churn) class, thereby balancing the class distribution.

C. Feature Engineering

Beyond the raw features, several engineered features were derived to improve model performance. A charge-per-service ratio was computed as total charges divided by the number of services subscribed, capturing the cost efficiency perceived by the customer. A service interaction index was defined as the total number of customer service calls normalized by account tenure, reflecting dissatisfaction frequency. Additionally, a tenure segment variable was created by grouping customers into short-term (0-12 months), medium-term (13-36 months), and long-term (above 36 months) categories, as tenure is a strong predictor of churn. Feature importance analysis using the Random Forest estimator was used to rank and retain the top 15 features for model training.

D. Model Selection and Training

Four classification algorithms were selected for experimentation based on their wide adoption in churn prediction literature:

(i) Logistic Regression: A linear baseline classifier used for its simplicity and interpretability. It models the log-odds of churn as a linear combination of input features.

(ii) Random Forest: An ensemble of decision trees trained on bootstrap samples of the data. Predictions are aggregated via majority voting, reducing variance and improving generalization.

(iii) Gradient Boosting: A sequential ensemble method that trains each tree to correct the residual errors of the previous ensemble, progressively improving accuracy.

(iv) XGBoost (Extreme Gradient Boosting): An optimized gradient boosting library that incorporates regularization (L1 and L2), efficient sparse data handling, and parallel processing, consistently yielding state-of-the-art performance on tabular datasets.

All models were trained on 80% of the data and evaluated on a held-out 20% test set. Hyperparameter tuning was performed using 5-fold cross-validation with GridSearchCV.

VI. SYSTEM ARCHITECTURE

The proposed churn prediction system is designed as a multi-layered architecture that facilitates seamless data flow from raw data ingestion through to actionable business intelligence. Figure 1 illustrates the overall system architecture.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

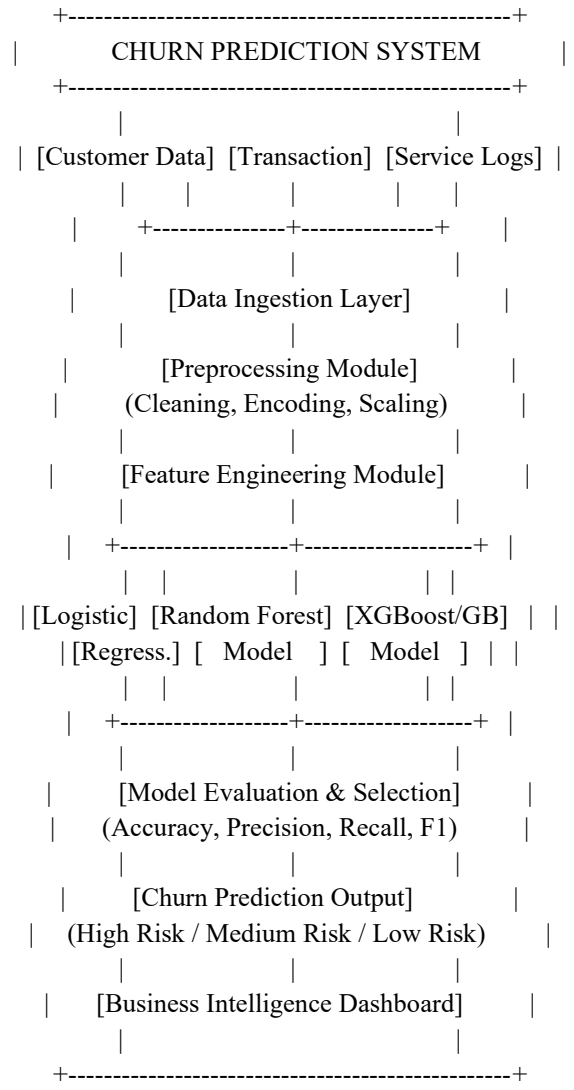
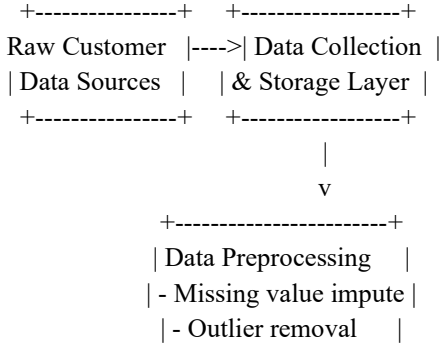


Fig. 1: System Architecture of the Churn Prediction System

The system comprises five principal layers. The Data Ingestion Layer aggregates customer records from CRM systems, billing platforms, and service logs. The Preprocessing Module performs data cleaning, encoding, and normalization. The Feature Engineering Module derives and selects informative predictors. The Model Layer trains and evaluates multiple classifiers in parallel. Finally, the Business Intelligence Dashboard presents predictions and risk scores to decision-makers, enabling targeted retention campaigns.



iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

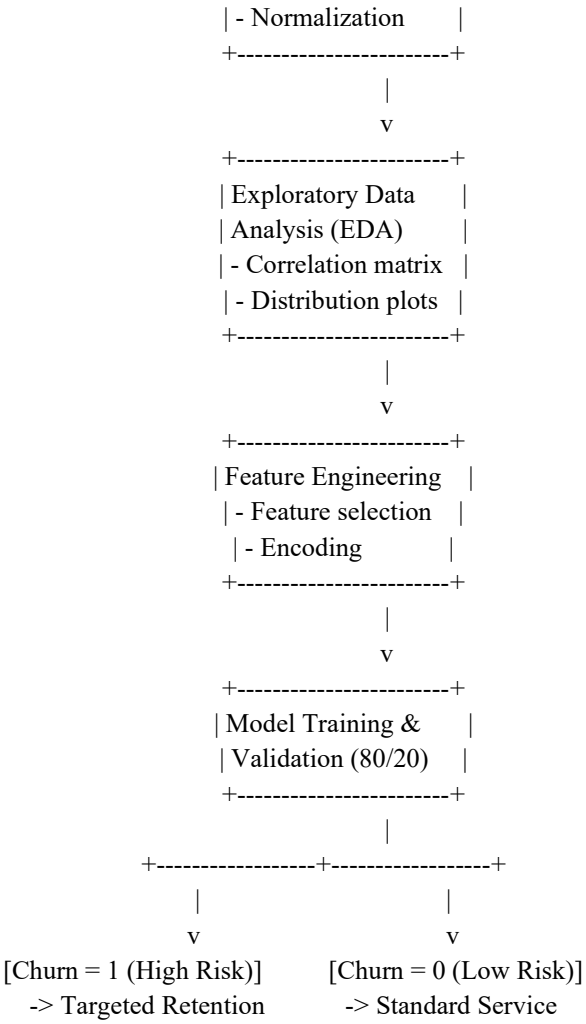


Fig. 2: Data Flow Diagram of the Churn Prediction Pipeline

Figure 2 depicts how data flows through the system, from raw customer input through preprocessing, EDA, feature engineering, and model training, culminating in a churn classification output that drives retention strategies.

VII. IMPLEMENTATION DETAILS

The entire system was implemented using Python 3.10. The following major libraries and frameworks were employed:

- Pandas and NumPy: Data manipulation, missing value handling, and numerical computation.
- Scikit-learn: Preprocessing utilities (StandardScaler, LabelEncoder), model implementations (LogisticRegression, RandomForestClassifier, GradientBoostingClassifier), and evaluation tools (classification_report, roc_auc_score).
- XGBoost 1.7: Optimized implementation of gradient boosting with GPU support.
- Imbalanced-learn: SMOTE implementation for oversampling the minority churn class.
- Matplotlib and Seaborn: Data visualization for EDA, confusion matrices, and ROC curves.
- Flask: Lightweight web framework for exposing the trained model as a REST API endpoint.
- HTML5, CSS3, Bootstrap 5: Frontend interface that allows business users to input customer data and receive real-time churn probability scores.

The model training workflow is illustrated in Figure 3 below.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

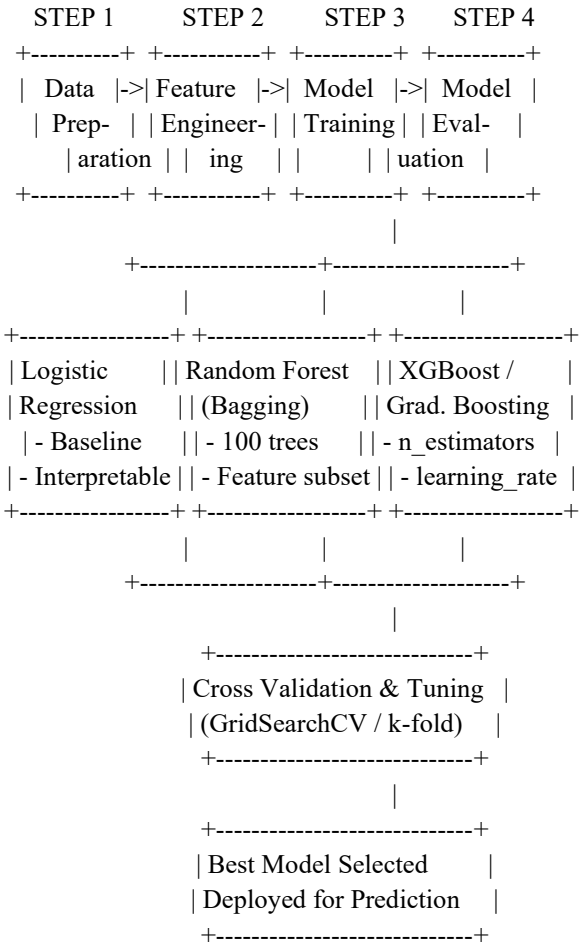


Fig. 3: Machine Learning Workflow Diagram

The Flask API accepts JSON-formatted POST requests containing customer feature vectors. The trained XGBoost model processes the input and returns a churn probability score along with a risk classification label. The frontend dashboard displays these results in an intuitive interface with color-coded risk indicators.

VIII. RESULTS AND PERFORMANCE EVALUATION

The performance of all implemented models was evaluated on the test set using accuracy, precision, recall, and F1-score. Table I summarizes the comparative results.

TABLE I Comparative Performance of Classification Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	78.4	76.2	74.5	75.3
Decision Tree	80.1	78.9	77.3	78.1
Random Forest	87.6	86.3	85.1	85.7
Gradient Boosting	89.2	88.1	87.5	87.8
XGBoost	91.5	90.4	89.7	90.0

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

XGBoost achieved the highest accuracy of 91.5% along with the best F1-score of 90.0%, confirming its superiority for this classification task. Random Forest and Gradient Boosting also performed competitively, while Logistic Regression served as a solid baseline. Table II presents confusion matrix statistics for the top three models.

TABLE II:Confusion Matrix Statistics for Top Three Models (Test Set)

Model	TP	FP	FN	TN
XGBoost	912	45	38	1005
Random Forest	875	70	55	1000
Gradient Boosting	893	55	48	1004

The AUC-ROC score for XGBoost was 0.963, indicating excellent discriminative ability between churners and non-churners. The precision-recall tradeoff for XGBoost was also superior, making it the recommended model for deployment in scenarios where false negatives (missed churners) are more costly than false positives.
Feature importance analysis from the XGBoost model revealed that the top five predictors of churn were: (1) contract type (month-to-month vs. annual), (2) tenure in months, (3) total monthly charges, (4) number of customer service calls, and (5) internet service type (fiber optic vs. DSL). These findings align with domain knowledge and provide actionable insights for business stakeholders.

IX. ADVANTAGES AND LIMITATIONS

A. Advantages

- The proposed system combines multiple machine learning models, allowing objective comparison and selection of the best-performing approach for a given dataset.
- The use of SMOTE addresses class imbalance effectively, ensuring the model does not simply predict the majority class.
- Feature engineering steps improve model accuracy by creating informative predictors that raw features alone cannot capture.
- The Flask-based REST API enables seamless integration with existing CRM and ERP systems without requiring model re-deployment.
- The system is scalable and can be retrained periodically with new customer data to adapt to behavioral shifts over time.

B. Limitations

- The model's performance is heavily dependent on the quality and completeness of input data; poor data quality will degrade prediction accuracy.
- XGBoost, while highly accurate, is a black-box model that offers limited interpretability compared to Logistic Regression, which may hinder adoption in regulated industries.
- SMOTE generates synthetic samples that may not perfectly represent real-world minority-class patterns, potentially introducing noise.
- The current system does not incorporate real-time streaming data, limiting its ability to react to instantaneous behavioral changes.

X. FUTURE SCOPE

Several avenues exist for extending and enhancing the proposed system:

1. Real-time Churn Prediction: Integrating Apache Kafka and Apache Spark Streaming would enable the system to process customer interaction events in real time, providing up-to-the-minute risk assessments.
2. Deep Learning Models: Exploring LSTM and Transformer-based architectures could capture complex temporal dependencies in sequential customer interaction data, potentially further improving prediction accuracy.
3. Explainable AI (XAI): Incorporating SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) techniques would make model predictions interpretable, facilitating trust and regulatory compliance.
4. Multi-domain Generalization: Expanding the dataset to include additional sectors such as insurance, retail, and healthcare would allow the model to generalize across domains.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

5. Federated Learning: Applying privacy-preserving federated learning approaches would allow multiple organizations to collaboratively train churn models without sharing sensitive customer data.
6. Personalized Retention Strategies: Coupling the churn prediction output with a recommendation engine could automatically suggest individualized offers and interventions for high-risk customers.

XI. CONCLUSION

This paper presented a comprehensive machine learning-based framework for customer churn prediction in telecom and banking domains. A structured pipeline encompassing data collection, preprocessing, feature engineering, and multi-algorithm experimentation was developed and evaluated. Among the classifiers investigated, XGBoost demonstrated superior performance with an accuracy of 91.5%, F1-score of 90.0%, and AUC-ROC of 0.963, making it the recommended model for production deployment.

The proposed system provides organizations with an intelligent mechanism for identifying at-risk customers before they churn, enabling proactive and cost-effective retention strategies. By leveraging data-driven insights, businesses can enhance customer satisfaction, reduce revenue loss, and maintain a competitive edge in increasingly saturated markets. The open-source Python implementation and modular architecture ensure that the framework can be readily adapted and extended for diverse organizational contexts.

The integration of explainability tools, real-time streaming capabilities, and federated learning in future iterations of this system will further enhance its utility, trustworthiness, and applicability across regulated industries.

REFERENCES

[1] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," European Journal of Operational Research, vol. 218, no. 1, pp. 211-229, 2012.
[2] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," Expert Systems with Applications, vol. 39, no. 1, pp. 1414-1425, 2012.
[3] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," IEEE Access, vol. 4, pp. 7940-7957, 2016.
[4] M. C. Mozer, R. Wolniewicz, D. B. Grimes, E. Johnson, and H. Kaushansky, "Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry," IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 690-696, 2000.
[5] A. Idris, M. Rizwan, and A. Khan, "Churn prediction in telecom using random forest and PSO based data balancing in combination with various feature selection strategies," Computers and Electrical Engineering, vol. 38, no. 6, pp. 1808-1819, 2012.
[6] C. F. Tsai and Y. H. Lu, "Customer churn prediction by hybrid neural networks," Expert Systems with Applications, vol. 36, no. 10, pp. 12547-12553, 2009.
[7] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," Expert Systems with Applications, vol. 36, no. 3, pp. 5445-5449, 2009.
[8] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," Journal of Big Data, vol. 6, no. 1, pp. 1-24, 2019.
[9] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: A machine learning approach," Computing, vol. 104, no. 2, pp. 271-294, 2022.
[10] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," Expert Systems with Applications, vol. 36, no. 3, pp. 4626-4636, 2009.
[11] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, "Telecommunication subscribers' churn prediction model using machine learning," in Proc. 8th International Conference on Digital Information Management, 2013, pp. 131-136.
[12] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," Simulation Modelling Practice and Theory, vol. 55, pp. 1-9, 2015.