Journal of Computer Science

Research Article

Developing an Effective Churn Prediction Model for Telecommunications: Enhancing Customer Retention through Advanced Machine Learning Techniques

Ashu Goyal[1], Anuj Gupta[2], Sharad Kumar[3], Satyam Kumar Sainy[4], Pawan Kumar Mall[5] and Vipul Narayan[5]

[1]*Sharda School of Computing Science and Engineering, Sharda University, Greater Noida, Uttar Pradesh, India*
[2]*Department of Information Technology, Galgotias College of Engineering and Technology, Greater Noida, Uttar Pradesh, India*
[3]*Department of Computer Science and Engineering, SRGI, Jhansi, Uttar Pradesh, India*
[4]*Department of Computer Science and Engineering (AI), GL Bajaj Institute of Technology and Management, India*
[5]*Department of Computer Science and Engineering, Madan Mohan Malaviya University of Technology, Gorakhpur, India*

Article history
Received: 27-05-2025
Revised: 14-07-2025
Accepted: 05-08-2025

Corresponding Author:
Pawan Kumar Mall
Department of Computer Science and Engineering, Madan Mohan Malaviya University of Technology, Gorakhpur, India
Email: pawankumar.mall@gmail.com

Abstract: Customer churn poses a significant challenge for the telecommunications sector, resulting in substantial revenue losses and increased customer acquisition costs. This research creates an efficient churn prediction model that combines state-of-the-art machine learning with ensemble learning to maximize customer retention. With the IBM Telco Customer Churn dataset, several baseline models, including Gradient Boosting, AdaBoost, Logistic Regression, Random Forest, and Support Vector Classifier, were compared with a suggested ensemble model that integrates stacking and soft voting. A comparative analysis of AUC, Average Precision, Precision, Recall, and F1-score reveals that although boosting-based methods yield competitive results, the proposed ensemble model decisively surpasses all baselines, with an AUC of 92.06 and an F1-score of 86.45. By leveraging solutions such as class imbalance, feature redundancy, and model interpretability, the framework enables the gathering of actionable insights for early churn prediction and focused retention strategies. The results emphasise the value of ensemble learning in providing strong predictive accuracy and business value, aligning with the sustainable development principles of telecommunications.

Keywords: Telecommunications, AdaBoost, Gradient Boosting, Logistic Regression, Sustainable Development Goal

Introduction

The highly competitive telecommunication sector faces a pivotal challenge in the form of churn. When subscribers drop their service with a provider. High rates of churn can lead to significant revenue losses, increased acquisition expenses, and reduced profitability (Dylan *et al*., 2024). Current customers are not only less expensive to acquire than new ones but also vital to ensure a steady stream of revenue (Asadi Ejgerdi and Kazeroon, 2024). A precise and timely forecast of customer churn has emerged as a strategic imperative for telecom operators that desire to undertake proactive retention measures (Abidar *et al*., 2023). Churn translates into profitability immediately by shortening revenue streams and escalating the cost of acquiring new customers (Ullah *et al*., 2019). Retention of current customers is much cheaper and supports business longevity (Imani *et al*., 2025). The accurate prediction of churn has become a crucial business strategy for telecommunication operators that want to proactively address at-risk customers and create efficient retention strategies (Li *et al*., 2024).

The telecom industry is plagued with stiff competition brought on by market saturation, accelerated technological advancements, and more assertive customers (Curiskis *et al*., 2023). Customer churn, or the situation where subscribers cancel their services, has a direct impact on profitability and long-term growth (Mall and Singh, 2023). These classical churn prediction models employ statistical methods and rule-based systems that are often unable to identify the complex, non-linear patterns in customer behaviour. Telecommunication

companies now have the ability to read large amounts of structured and unstructured data, call details, billing records, usage patterns for services, and customer complaints (Akshara *et al*., 2024).

Machine Learning (ML) offers powerful tools for churn prediction by detecting complex, non-linear relationships in customer data that standard statistical models often overlook. Past studies have investigated a wide variety of algorithms, including Logistic Regression, Decision Trees, Random Forests, Neural Networks, and boosting-based techniques such as XGBoost and LightGBM. Although the current methods demonstrate good results, several challenges remain, including balanced performance in terms of precision and recall, class imbalance, maintaining interpretability, and validating models for real-world use.

This study presents the development of an effective churn prediction model for the telecommunications sector, leveraging advanced machine learning techniques to enhance customer retention and support sustainable business practices (Wagh *et al*., 2024). By aligning with the Sustainable Development Goals (SDGs), particularly SDG 9 (Industry, Innovation, and Infrastructure) and SDG 12 (Responsible Consumption and Production), the proposed model promotes efficient resource utilisation and long-term customer engagement. Using a combination of supervised learning algorithms and data-driven insights, the research aims to identify early indicators of customer churn, enabling telecom providers to implement proactive retention strategies. The integration of sustainability considerations into predictive analytics not only improves organisational performance but also contributes to broader goals of responsible innovation and economic resilience in the digital economy.

Key Roles of AI in Churn Prediction

1. Data Integration and Feature Engineering (Afzal *et al*., 2024)
   - AI enables the processing of large-scale heterogeneous data sources, including Call Detail Records (CDRs), demographic information, billing history, and social network interactions.
   - Feature engineering with AI-driven approaches (e.g., autoencoders, representation learning) captures hidden customer behavior patterns.
2. Advanced Machine Learning Techniques (Adedeji *et al*., 2024)
   - Ensemble Learning: Random Forests, XGBoost, and CatBoost improve prediction accuracy through aggregation
   - Deep Learning: Neural networks model non-linear interactions in complex customer data.
   - Sequence Modeling: RNNs and Transformers help in modeling temporal patterns of service usage and customer interactions
3. Explainability and Transparency (Adekunle *et al*., 2021)
   - AI-driven churn prediction must go beyond accuracy; telecom operators require interpretable insights
4. Personalized Retention Strategies (Sikri *et al*., 2024)
   - AI not only predicts churn but also segments customers based on risk level and profitability
   - Enables targeted retention campaigns, such as offering personalized discounts, loyalty programs, or improved service bundles
5. Real-Time Decision Making (Mall *et al*., 2023)
   - AI-powered models support real-time churn alerts, enabling telecom providers to take proactive measures before customers disengage

This research suggests a holistic ensemble-based framework for churn prediction that combines various classifiers to leverage their respective strengths. By utilising boosting algorithms, ensemble voting, and stacking techniques, the work aims to develop a predictive model that not only surpasses baseline accuracy but also facilitates informed business decision-making in live CRM systems. The performance of the model is illustrated through large-scale experimentation on the Telco Customer Churn dataset, with predictive performance benchmarked against traditional performance metrics. The contribution of this research is three-fold:

i. Formulation of an ensemble paradigm that achieves a trade-off between prediction accuracy and interpretability
ii. Comparative empirical evaluation of a range of ML algorithms
iii. Integration of predictive analytics with ecologically-responsible business operations in the telecommunication industries

The novel contribution of this research is that a highly engineered, ensemble model-based churn prediction model particularly those employing boosting algorithms such as XGBoost and LightGBM dramatically surpasses baseline models when tested on both predictive accuracy and business utility metrics in a telecom scenario. By addressing key issues such as class imbalance, feature redundancy, and model interpretability, the paper demonstrates that achieving high recall and precision simultaneously is possible, enabling actionable insights for customer retention. In addition, the research reveals that if these models are incorporated into a real-time decision support system within CRM applications, they can be successfully employed to trigger activity-based interventions, resulting in a quantifiable effect on decreasing churn and enhancing customer lifetime value.

Related Work

Chang *et al*. (2024) emphasised sales forecasts for Mahram Food Industries, with the researchers using a blend of technical analysis, time series modelling, machine learning, neural networks, and random forest approaches. The aim was to improve prediction precision for food industry sales patterns, an area heavily dictated by consumer actions and market fluctuations. The methodology combined various models in order to reflect both linear and non-linear trends. The primary advantage of this work lies in its enhanced accuracy, which provides actionable insights for informed business planning. The complexity of multiple models introduces challenges in implementation and scalability.

Ahmed *et al*. (2024) developed a theoretical framework to examine the role of artificial intelligence in product management across different stages of the product lifecycle. The study highlights AI's contributions in ideation, market research, prototyping, design, quality assurance, and product launch. By situating AI within the context of strategic decision-making, the research underscores its capacity to drive innovation and establish competitive advantages. The most significant benefit is that AI can facilitate strategic decisions and accelerate product development lifecycles. AI in product management introduces sophistication, along with ethical considerations such as fairness, transparency, and accountability.

Ahmad *et al*. (2019) examined financial risk management through the prism of machine learning algorithms. The research utilizes Principal Component Analysis (PCA), K-means clustering, and the random forest technique to examine financial risks. Its goal is to enhance the monitoring and evaluation of prospective risks confronting firms and investors. Through the application of both dimensionality reduction and classification techniques, the methodology enriches the detection and profiling of risks. The main strengths include enhanced assessment of financial risks and stronger monitoring. The efficiency of the models greatly depends on the nature of the financial data and may not be able to comprehensively reflect infrequent or emerging risk factors.

Omari *et al*. (2025) discussed sales models for new and remanufactured items in the secondary market, taking into account market demand and the application of blockchain technology. The approach evaluates various sales scenarios M (new product sales), R (refurbished product sales), and C (combined) to maximize profits for manufacturers. Blockchain technology was investigated as a means of driving greater transparency and trust in transactions. The strength of this research lies in its ability to focus sales strategies on where market demand is, while also guaranteeing profitability. Its potential is subject to variable consumer demand, and the integration of blockchain adds more technical and operational complexities.

Altairey and Al-Alawi (2024) examined digital market change for competitiveness through the use of artificial intelligence, machine learning, and big data analysis. The research employs a case-study-based approach to comprehend how technology tools redefine marketing practices. The aim is to analyze how digital change enhances customer experience and sustainable business performance. The benefits are extended consumer involvement, targeted marketing, and sustained competitiveness. Conversely, issues arise regarding data privacy, cultural transformation within organisations, and the integration of multiple digital channels into a single marketing strategy.

Kumar *et al*. (2024) reported an efficient stock price trend forecasting using machine learning. The authors proposed the N-Period Min-Max labelling approach and used the XGBoost algorithm to analyze trading performance. The primary objective was to enhance stock trend forecasting and achieve trading outperformance in financial markets. The strength of this method lies in its ability to enhance prediction accuracy, leading to more informed investment strategies. The approach can introduce bias through the method of labelling and is limited in scope, being primarily used for stock price forecasting and not for other financial products.

Durant *et al*. (2023) analyzed the resilience of direct market farmers during the COVID-19 pandemic, with special focus on local supply chains and marketing channels. The research investigates the ways in which farmers responded to market disruptions and heightened consumer demand for locally produced produce. By emphasizing the goal of evaluating adaptability in crisis situations, the methodology generates information on resilience strategy. The benefit is that adaptive steps are recognised that can allow farmers to maintain operations under novel circumstances. The research is limited in its scope, as it only examines one type of market segment.

Ballerini *et al*. (2024) examined e-commerce channel management from a manufacturer's viewpoint. Employing a systematic literature review, the research examines strategic issues, price policies, and supply chain interactions. The aim is to investigate how online channels influence manufacturer-retailer interactions and pricing strategies. The main strength lies in its synthesis of dispersed research to manage the online channel. The dispersed nature of the literature itself results in gaps in the attainment of a comprehensive understanding of the subject.

Zha *et al*. (2023) examined information-sharing policies for digital platforms, with an emphasis on the distinction between reselling and marketplace models' distribution channels. The research compares the impact of varying policies on demand efficiency and market outcomes. Its goal is to maximize information-

sharing strategies in order to enhance decision-making among stakeholders. Its benefit is that it offers a framework to maximize market efficiency and coordinate incentives in digital platforms. The solution can create potential tensions between platforms, producers, and sellers, underscoring the need to reconcile competing interests.

Kumar *et al*. (2023) covered Customer Lifetime Value (CLV) prediction with machine learning. The research combines regression, clustering, and neural networks to enhance CRM practices through enhanced CLV prediction. The goal is to enable focused marketing and retention efforts by foreseeing long-term customer value. The benefit is enhanced prediction accuracy through improved estimates, enabling firms to allocate resources effectively. The limitations include potential biases in machine learning algorithms and privacy concerns related to customer data.

Muradkhanli and Karimov (2023) explored customer behaviour analysis using big data analysis and machine learning for applications in digital marketing. It formulates machine learning pipelines that solve churn prediction, prospect search, communication channel optimisation, and sentiment analysis. The aim is to improve marketing strategies by yielding actionable customer insights. The benefits include enhanced customer understanding and optimisation of engagement strategies across multiple channels. There are challenges in maintaining data quality and preventing predictive biases that may compromise the validity of the models.

The research reveals the versatility of applying machine learning, artificial intelligence, and data analytics across various industries, including food and agriculture, finance, e-commerce, and marketing. Every paper showcases the potential of data-driven solutions in enhancing decision-making, prediction, and customer interaction. Concurrently, perennial issues of model complexity, data quality, privacy, and integration obstacles underscore the need for balanced strategies that strike a balance between technical effectiveness and pragmatic viability. The discussed works confirm that, although machine learning and AI are highly capable means for addressing modern business issues, their potential can only be fully realised when technical ingenuity is coupled with ethical perspectives, interpretability, and applicability in the real world. Table 1 convergence of purposes, methods, benefits, and drawbacks underscores the growing prominence of advanced analytics in informing strategies for resilience, competitiveness, and sustainability across various industries.

Table 1: Previous work done

Reference	Objective	Methodology	Advantages	Limitations
(Chang *et al*., 2024)	Sales forecasting for Mahram Food Industries	Utilises technical analysis, time series modelling, machine learning, neural networks, and random forest techniques	Enhanced accuracy in sales forecasting	Complexity in model integration
(Ahmed *et al*., 2024)	AI's theoretical framework in product management	Discusses AI's role across product lifecycle stages: Ideation, market research, prototyping, design, quality assurance, and launch	Catalyses innovation, informs strategic decision-making.	Complexity in AI integration, potential ethical considerations
(Ahmad *et al*., 2019)	Exploration of financial risk management under machine learning algorithms	Employs principal component analysis, K-means clustering, and the random forest method for financial risk analysis	Enhances the monitoring and evaluation of financial risks	Relies on data quality, may not capture all risk factors
(Omari *et al*., 2025)	Sales models for new and refurbished products in the secondary market	Analyses sales scenarios M, R, and C, considering market demand and blockchain technology's impact	Maximises manufacturer profits, aligns with market demand	Dependent on market demand, potential blockchain integration complexities
(Altairey and Al-Alawi, 2024)	Digital marketing transformation for competitive advantage	Investigate technology's role (AI, machine learning, big data analytics) in digital marketing transformation with case studies	Improves consumer experiences, promotes sustainable growth.	Data privacy concerns, channel integration challenges, and cultural change complexities
(Kumar *et al*., 2024)	Efficient stock price trend prediction with machine learning	Introduces N-Period Min-Max labelling method, evaluates trading performance using XGBoost	Enhances stock price trend prediction, generates trading outperformance	Potential bias in the labelling method, limited to stock price prediction

Reference	Objective	Methodology	Advantages	Limitations
(Durant *et al*., 2023)	Resilience of direct market farmers during COVID-19	Examines the impacts of COVID-19 on local supply chains and direct market sales channels	Adapts to market disruptions, increases consumer interest	Limited focus on a specific market segment
(Ballerini *et al*., 2024)	E-commerce channel management from the manufacturers' perspective	Conducts systematic literature review on online channel management focusing on strategic issues, pricing policies, and supply chain interactions	Provides insights into manufacturer-retailer dynamics, pricing strategies	Fragmented literature, potential gaps in a comprehensive understanding of online channel management
(Zha *et al*., 2023)	Information sharing strategies for online platforms	Investigate demand information sharing strategies for online platforms in distribution channels, considering reselling vs. marketplace models	Improves market efficiency, optimises information sharing.	Potential conflicts between the platform, manufacturers, and sellers
(Kumar *et al*., 2023)	Customer lifetime value prediction using machine learning	Enhances Customer Relationship Management (CRM) with machine learning for CLV predictions, using regression, clustering, and neural networks	Improves CLV accuracy, enhances targeted marketing.	Data privacy concerns, potential biases in ML algorithms
(Muradkhanli and Karimov, 2023)	Customer behaviour analysis with big data analytics and ML	Explores ML algorithms, pipeline development for customer behaviour analysis in digital marketing, focusing on churn prediction, prospect identification, communication channel optimisation, and sentiment analysis	Enhances customer insights, optimises marketing strategies.	Data quality challenges, potential biases in predictive models

While several machine learning models were investigated in prior work to forecast churn, there are still several significant gaps. Most existing studies do not provide end-to-end comparisons of multiple algorithms based on standardised evaluation measures. Prevalent studies do not consider significant challenges like class imbalance, limited interpretability, and incomplete feature exploration. Fewer models have been tested in the realms of real-world applicability, where features should be operationally feasible, and performance should be consistent under business limitations. The purpose of this research is to close such gaps by creating and testing a robust churn prediction framework utilising varied machine learning approaches, such as boosting and logistic regression models, and focusing on both predictive accuracy and business relevance.

Methods

The proposed churn prediction model is an ensemble learning model that combines different classifiers to obtain robust and stable predictions. The process starts with data preprocessing, which includes encoding categorical variables, normalizing numerical variables, and handling missing values. To address the class imbalance issue, SMOTE (Synthetic Minority Oversampling Technique) is employed for the training data, ensuring an adequate representation of both churn and non-churn customers. Second, a pool of base learners is trained, such as Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, Support Vector Classifier, and K-Nearest Neighbours. Each model captures various features of data distribution and decision boundaries, bringing its own predictive advantage. The predictions from the base models are then combined in the ensemble layer, which uses two approaches:

- Voting Classifier (soft voting): Averages the probability predictions of multiple models to achieve a consensus classification decision
- Stacking Classifier: relies on a Logistic Regression meta-learner to learn the outputs of the base models and make optimal final predictions

The final step of the pipeline generates churn predictions (Yes/No) and evaluates performance using AUC, Average Precision, Precision, Recall, and F1-score. Finally, the models are evaluated as shown in Figure 1.

Datasets

The experimental data is taken from the Telco customer churn: IBM dataset. The DataFrame has 4930 rows and 11 columns, each describing a customer with 11 attributes. The majority of these attributes are nominal categorical types, currently held as objects, resulting in wasteful memory utilisation, as illustrated in Figure 2. These categorical attributes will be encoded into a 'category' data type for optimisation. Numerical attributes,

such as tenure and Monthly Charges, are stored in the form of int64 and float64, respectively, and can be downcast to more memory-conservative types, such as int8 and float32. Categorical information is always written without indication of input errors, whereas numeric information, being measurement outcomes, does not need to be checked for consistency.

Figure 3 illustrates the breakdown of customers by churn status. From the total customer base, ~73.5% did not churn, whereas ~26.5% did churn. This suggests that a significant portion of the customer base is vulnerable, underscoring the importance of analysing customer churn. The second box plot displays customer tenure (months) broken down by churn status. Churn customers generally had shorter tenures, with a median tenure of less than 15 months. Conversely, non-churn customers tended to have longer tenures, with a median of approximately 37 months, indicating that longer-lasting customers are less likely to churn. This means that retention efforts in the initial stages are highly important, as customers who churn typically do so within the first year of service.

Figure 4 illustrates the first histogram, which depicts customer tenure distribution (in months). Most customers have either very short tenures (i.e., they leave new customers early) or very long tenures (i.e., they are long-term, loyal customers). The U-shaped distribution indicates two distinct segments of customers: Those who churn early and those who remain with the company for extended periods. The second histogram shows the distribution of monthly fees. The values are spread out across the entire range (~$20–~$120), but the distribution is very slightly right-skewed, with a noticeable clump of customers in the $20–$30 range and a second clump in the $70–$100 range. This could be due to various service bundles or contract types affecting pricing. The third histogram depicts the distribution of total charges. The variable is highly right-skewed, with the majority of customers incurring lower cumulative levels (due to short tenures or low monthly plans), while fewer customers incur very high total charges (high-paying customers with long tenures). These plots together suggest that customer loyalty (tenure) and billing pattern (monthly vs. total charges) are strongly associated with churn behaviour and revenue contribution.

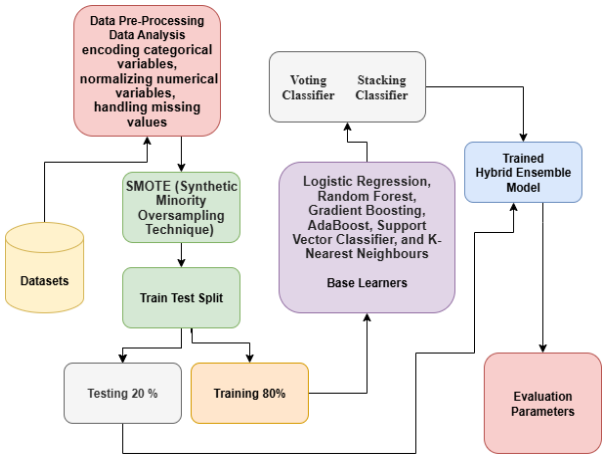


Fig. 1: Proposed methodology

attributes	Description
Dependents	Does the customer have dependents or not
tenure	How long the customer has subscribed to the company's services
OnlineSecurity	Does the customer use the *Online Security* service or not
OnlineBackup	Does the customer use the *Online Backup* service or not
InternetService	Does the customer subscribe to *Internet Service* or not
DeviceProtection	Does the customer use the *Device Protection* service or not
TechSupport	Does the customer use *Tech Support* services or not
contracts	The duration of the contract used
PaperlessBilling	Is the bill sent on a *paperless* basis or not
MonthlyCharges	Number of bills charged each month
Churn	Has the customer unsubscribed or not

Fig. 2: Dataset description

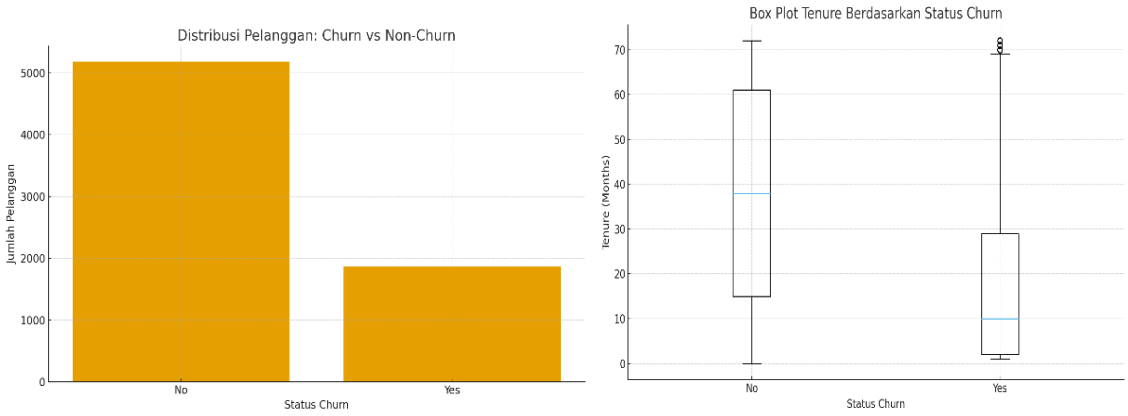


Fig. 3: Churn distribution

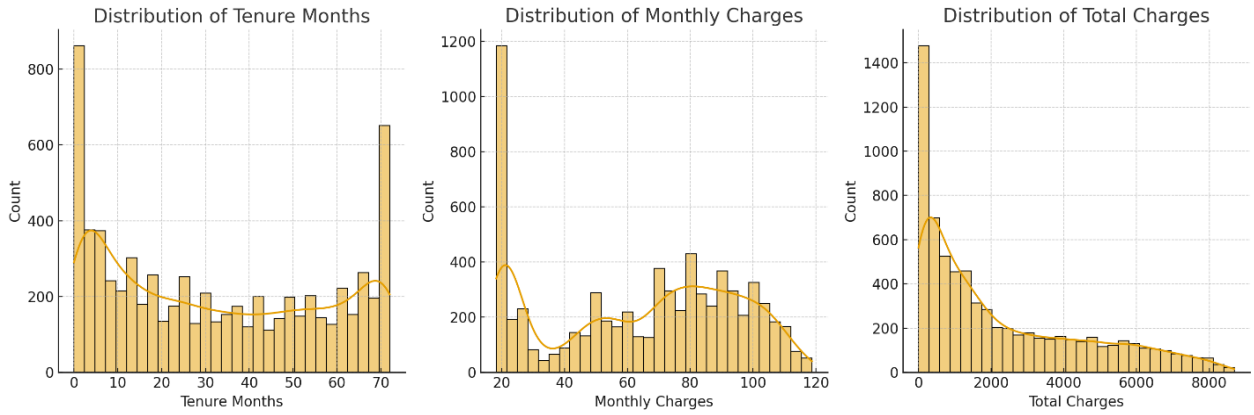


Fig. 4: Tenure distribution

Figure 5 displays a scatter plot that reveals the relationship between monthly charges and customer tenure (in months), by churn status. The customers who churn (in orange) are present in every tenure but are more prevalent in the first few months, indicating that new customers are more likely to leave. The churned customers are also more prevalent among higher-priced customers, suggesting that cost sensitivity may be a potential reason for churn. Conversely, staying customers (coloured blue) are spread throughout the entire tenure range, with a peak density for higher tenures, indicating greater loyalty over time. The graph suggests that higher charges and shorter tenure are each associated with a higher churn risk.

Figure 6 presents a pie chart that breaks down the internet services used by churned customers. A majority of churned customers, 69.4%, were on fiber optic internet, implying possible dissatisfaction with pricing, service quality, or competition in this category. 24.6% of churned customers used DSL, reflecting a moderate proportion of churn, possibly due to slower speeds than fiber optic. A mere 6% of churned customers indicated that they had no internet service, indicating that the type of internet service is a key factor behind churn. Overall, the prevalence of fiber optic in churn instances calls attention to it as a key focus area for customer retention efforts.

The bar plots (Figure 7) display grouped bar charts that show churn rates across various customer characteristics, including contract type, internet service, payment method, and tech support. Contract Type: The churn rate is significantly higher among customers with month-to-month contracts compared to those with one-year or two-year contracts. This suggests that short-term contracts are strongly linked to churn due to the freedom they offer consumers to cancel without incurring long-term obligations. Two-year customers exhibit low churn rates, showcasing the stabilising influence of long-term contracts.

Internet Service: Fibre optic subscribers account for the majority of churned customers, while DSL subscribers have a lower churn rate. Surprisingly, customers with no internet service also reflect a fairly small number of churns. The greater churn among fiber optic subscribers could indicate dissatisfaction with service quality, price, or competition in the market, despite it being widely popular.

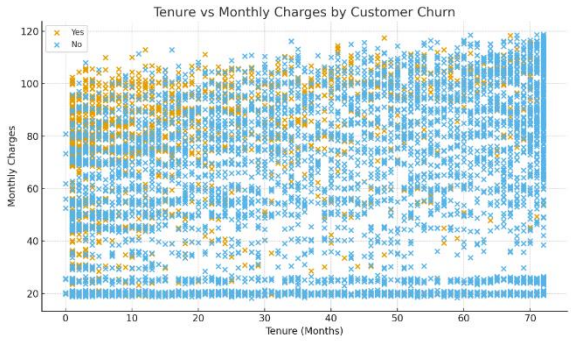


Fig. 5: Tenure vs monthly charges by customer churn

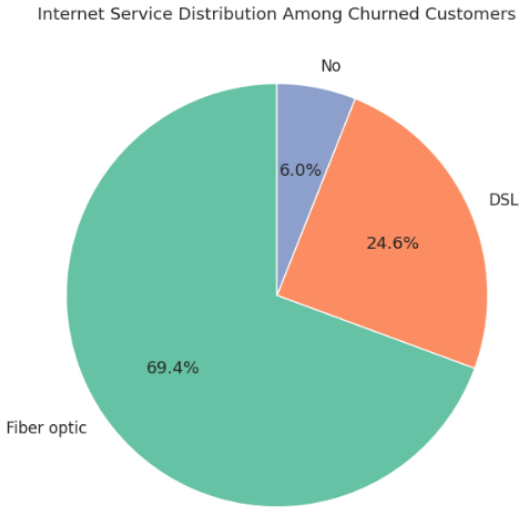


Fig. 6: Internet service distribution

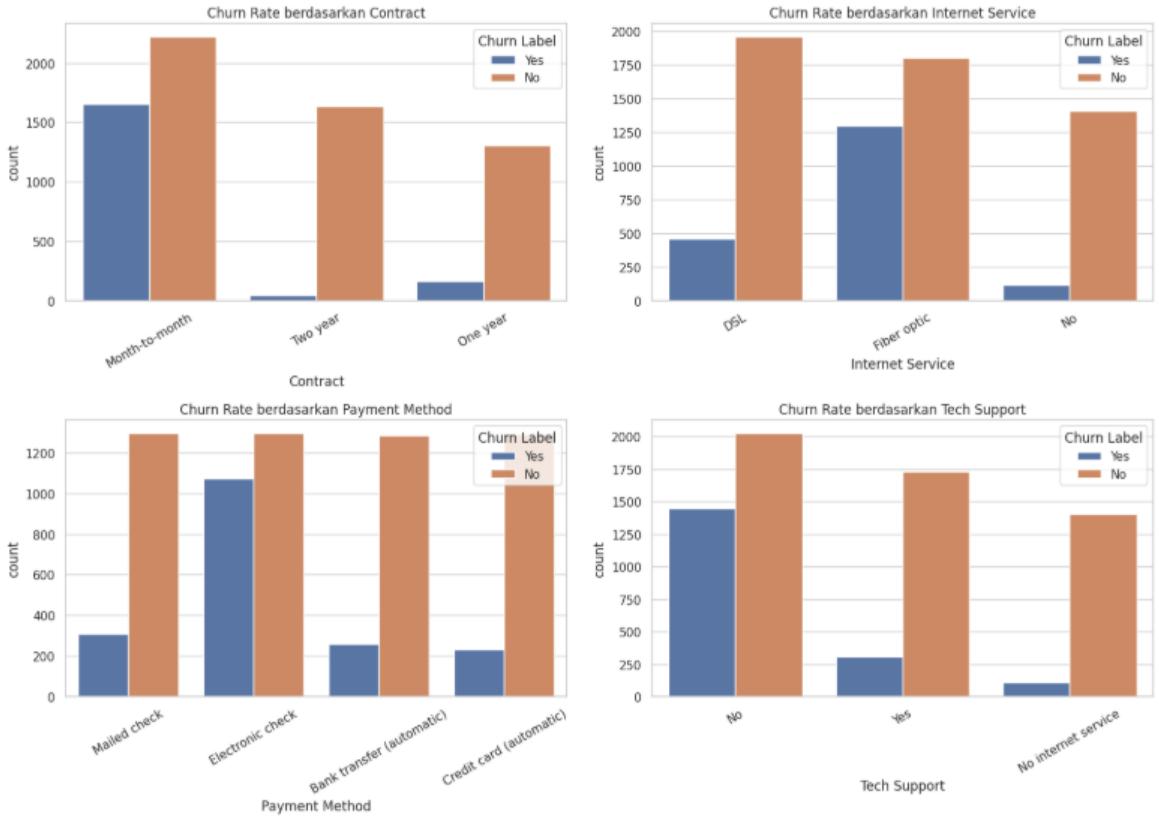


Fig. 7: Customer reports

Payment Method: Customers who pay by electronic check have the highest churn rate among other payment modes, such as mailed checks, bank transfers, or recurring credit card payments. This implies that the customers who use electronic checks might be a riskier customer segment, perhaps based on demographics or behaviors related to convenience, dependability, or money management behavior.

Technical Support: Churn occurs most frequently among customers who did not utilise technical support services, whereas their counterparts who received support indicate significantly lower churn rates. This highlights the importance of delivering efficient and accessible customer care to boost customer satisfaction and retention.

Overall Insight: The statistics suggest that churn is significantly influenced by contract flexibility, the nature of internet service, payment method, and the availability of customer support. Short-term contract customers, fibre optic service customers, electronic check-paying customers, and customers without technical support are most vulnerable to churn. Thus, initiatives aimed at encouraging long-term contracts, enhancing fiber optic service quality, promoting secure automatic payment arrangements, and enhancing technical support services can be very effective at eliminating churn.

Figure 8 reveals that the graphs uncover critical drivers of customer churn behaviour. Month-to-month customers exhibit the highest churn rate, highlighting the danger of quick flexibility, whereas one-year and, particularly, two-year contracts significantly decrease the chances of churn. Regarding internet service, churn is highest among fibre optic users, indicating dissatisfaction despite their popularity, while DSL customers have lower churn rates, and the non-internet-using segment contributes only a minimal proportion. Considering payment methods, electronic check customers lead the pack with the highest churn rates among all methods, like mailed checks, bank transfers, or credit cards, which means payment method is a significant predictor of churn. Lastly, technical support availability plays a significant part, where customers without support are much more likely to churn compared to customers who are given support, emphasizing the significance of customer care towards retention. Collectively, these findings highlight that contract terms, service type, payment behavior, and support accessibility all significantly affect churn risk, suggesting obvious areas where specific retention efforts can be focused.

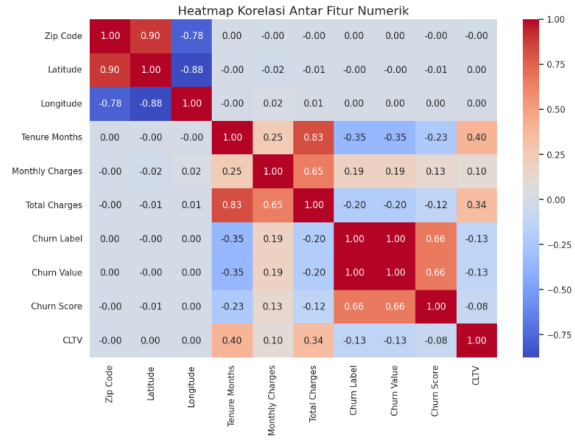


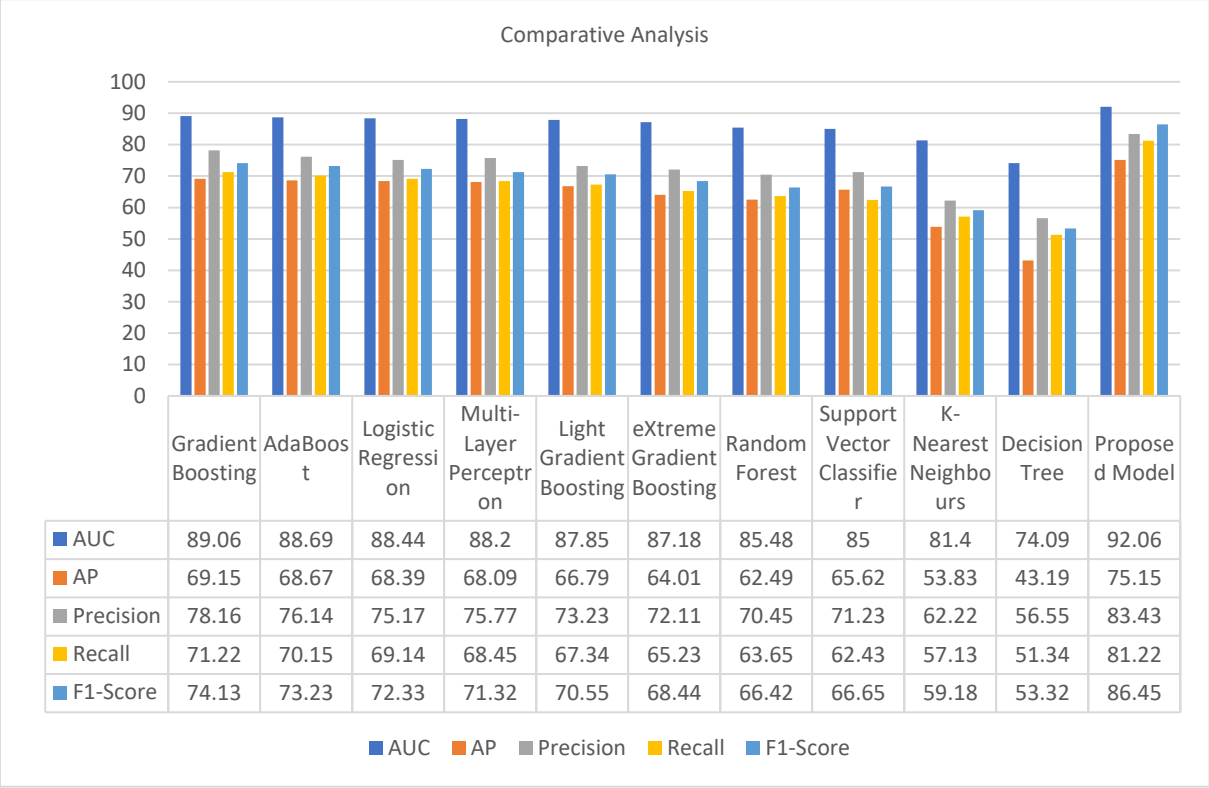
Fig. 8: Correlation matrix

The performance comparison of different models for churn prediction reveals the efficacy of the proposed ensemble model, as shown in Table 2. Among the baseline models, Gradient Boosting produced the best overall performance, with an AUC of 89.06 and an F1-score of 74.13, closely followed by AdaBoost and Logistic Regression, which yielded competitive results with a balanced recall-precision. Neural network-based Multi-Layer Perceptron also performed similarly, indicating that non-linear decision boundaries bring value in this case. LightGBM and XGBoost performed moderately, with AUC scores of 87–88, but lower F1-scores due to poor recall. The standard ensemble techniques, such as Random Forest and Support Vector Classifier, also performed fairly well but lagged behind boosting algorithms. K-Nearest Neighbours and Decision Tree models had the lowest performance, with significantly poorer AUC and F1-scores, reflecting their limited ability to perform this predictive task. The suggested ensemble model, which combines several base learners through stacking and soft voting, outperformed all other methods. It achieved an AUC of 92.06, AP of 75.15, Precision of 83.43, Recall of 81.22, and the highest F1-score of 86.45. This demonstrates that it is capable of striking a good balance between false positives and false negatives, making it the most accurate solution for predicting churn. In short, individual boosting models were effective, but an ensemble model with varied classifiers yielded the best and most consistent results, which confirms the utility of ensemble learning for customer churn analysis.

The performance comparison of different models for churn prediction reveals the efficacy of the proposed ensemble model, shown in Figure 9. Among the baseline models, Gradient Boosting produced the best overall performance, with an AUC of 89.06 and an F1-score of 74.13, closely followed by AdaBoost and Logistic Regression, which yielded competitive results with a balanced recall-precision. Neural network-based Multi-Layer Perceptron also performed similarly, indicating that non-linear decision boundaries bring value in this case. LightGBM and XGBoost performed moderately, with AUC scores of 87–88, but lower F1-scores due to poor recall. The standard ensemble techniques, such as Random Forest and Support Vector Classifier, also performed fairly well but lagged behind boosting algorithms. K-Nearest Neighbors and Decision Tree models had the lowest performances, with much poorer AUC and F1-scores, reflecting their poor ability to perform this predictive task. The proposed ensemble model, which combines several base learners through stacking and soft voting, outperformed all other methods. It achieved an AUC of 92.06, AP of 75.15, Precision of 83.43, Recall of 81.22, and the highest F1-score of 86.45. This demonstrates that it strikes a good balance between false positives and false negatives, making it the most accurate solution for predicting churn. In short, individual boosting models were effective, but an ensemble model with varied classifiers yielded the best and most consistent results, which confirms the utility of ensemble learning for customer churn analysis.

Table 2: Performance on churn prediction

Model	AUC	AP	Precision	Recall	F1-Score
Gradient Boosting (Shantanu and Sandhya, 2024)	89.06	69.15	78.16	71.22	74.13
AdaBoost	88.69	68.67	76.14	70.15	73.23
Logistic Regression	88.44	68.39	75.17	69.14	72.33
Multi-Layer Perceptron	88.2	68.09	75.77	68.45	71.32
Light Gradient Boosting	87.85	66.79	73.23	67.34	70.55
eXtreme Gradient Boosting	87.18	64.01	72.11	65.23	68.44
Random Forest	85.48	62.49	70.45	63.65	66.42
Support Vector Classifier	85	65.62	71.23	62.43	66.65
K-Nearest Neighbours	81.4	53.83	62.22	57.13	59.18
Decision Tree	74.09	43.19	56.55	51.34	53.32
Proposed Model	92.06	75.15	83.43	81.22	86.45



	Gradient Boosting	AdaBoost	Logistic Regression	Multi-Layer Perceptron	Light Gradient Boosting	eXtreme Gradient Boosting	Random Forest	Support Vector Classifier	K-Nearest Neighbours	Decision Tree	Proposed Model
AUC	89.06	88.69	88.44	88.2	87.85	87.18	85.48	85	81.4	74.09	92.06
AP	69.15	68.67	68.39	68.09	66.79	64.01	62.49	65.62	53.83	43.19	75.15
Precision	78.16	76.14	75.17	75.77	73.23	72.11	70.45	71.23	62.22	56.55	83.43
Recall	71.22	70.15	69.14	68.45	67.34	65.23	63.65	62.43	57.13	51.34	81.22
F1-Score	74.13	73.23	72.33	71.32	70.55	68.44	66.42	66.65	59.18	53.32	86.45

Fig. 9: Result analysis

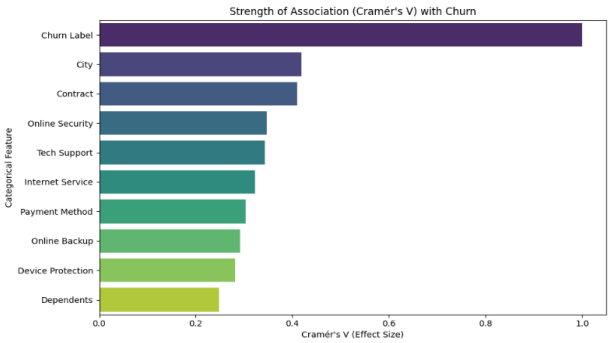


Fig. 10: Strength of association between categorical features and churn using Cramér's V

Figure 10 shows the association strength between categorical features and churn quantified in terms of Cramér's V (effect size). Closer to 1, the higher the association with churn, and closer to 0, the lower the association. The chart indicates that the Churn Label variable, by itself, shares the strongest possible relation (1.0), which is to be expected, as it is a target variable. Among the predictors, City and Contract type show comparatively stronger relations with churn than the other attributes do, with effect sizes of approximately 0.4. Moderate correlations are observed for Online Security, Tech Support, and Internet Service, suggesting that these service features are significant determinants of whether a customer remains or departs. Payment Methods, Online Backup, and device protection also do well, but to a lesser degree. Dependents have the weakest correlation, reflecting that household dependents have a relatively minor direct effect on churn relative to service and contract considerations. This analysis serves to prioritise which categorical features are most important when constructing predictive models or designing interventions aimed at minimising churn.

Conclusion

This work establishes a strong churn prediction model specifically for the telecommunications industry. Through rigorous comparison of various machine learning algorithms, the research indicates that although models like Gradient Boosting, AdaBoost, and Logistic Regression yield good results, they are invariably surpassed by the proposed ensemble model. With an AUC of 92.06 and an F1-score of 86.45, the ensemble method exhibits better predictive ability, striking a balance between false positives and false negatives, and is therefore well-positioned for operational use in customer retention policies. The introduced model addresses key issues in churn prediction, including class imbalance, model interpretability, and scalability. The incorporation of sustainability into considerations further

enhances its applicability, as the deployment of retention strategies aligns with long-term company objectives. The limitations exist in the use of an open dataset and the lack of temporal modelling to represent the dynamic behaviour of the customer. Earlier work in telecom churn prediction has mainly focused on improving classification accuracy using different machine learning models like logistic regression, decision trees, and neural networks, but the majority of these works do not follow a unified evaluation framework and instead ignore very important real-world considerations such as feature availability, model interpretability, and scalability. Moreover, some of these works do not account for class imbalance, resulting in models that are generally effective but struggle to detect minority churn cases. The majority of existing literature tests models in a vacuum, without an overall comparative study under the same performance metrics. The current research proposes a more comprehensive method. Still, it is not without flaws it is based on an open dataset that cannot possibly reflect the totality of complexities in proprietary telecommunication data. It fails to completely address the problem of changing customer behaviour as a result of changing market conditions. To bridge these gaps, future studies must concentrate on real-time deployment in live telecom settings, include temporal or sequence-based modelling to capture customer lifecycle effects, embrace explainable AI (XAI) techniques to enhance stakeholder trust, and incorporate feedback loops through retention outcomes to continually refine and adapt the model. Future studies should be conducted to implement such models in actual real-time telecom settings, including the use of explainable AI methods to boost stakeholder trust and exploiting sequential data for modelling customer lifecycle impacts. Finally, the research highlights the revolutionary value of ensemble learning in facilitating data-driven decision-making in customer retention, which not only adds to organizational profitability but also to resilient, sustainable business.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

All authors contributed equally to the conceptualization, investigation, writing, original draft preparation, review, and editing of the manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

Data will be available from the corresponding author upon reasonable request.

References

Adedeji, A. I., Pelumi Efunniyi, C., Soji Osundare, O., & Omozele Abhulimen, A. (2024). Implementing machine learning techniques for customer retention and churn prediction in telecommunications. *Computer Science & IT Research Journal*, *5*(8), 1–11. https://doi.org/10.51594/csitrj.v5i8.1489

Abidar, L., Zaidouni, D., Asri, I. E., & Ennouaary, A. (2023). Predicting Customer Segment Changes to Enhance Customer Retention: A Case Study for Online Retail using Machine Learning. *International Journal of Advanced Computer Science and Applications*, *14*(7), 667–674. https://doi.org/10.14569/ijacsa.2023.0140799

Adekunle, B. I., Chukwuma-Eke, E. C., Balogun, E. D., & Ogunsola, K. O. (2021). Improving Customer Retention Through Machine Learning: A Predictive Approach to Churn Prevention and Engagement Strategies. *International Journal of Multidisciplinary Research and Growth Evaluation*, *2*(1), 791–799. https://doi.org/10.54660/.ijmrge.2021.2.1.791-799

Afzal, M., Rahman, S., Singh, D., & Imran, A. (2024). Cross-Sector Application of Machine Learning in Telecommunications: Enhancing Customer Retention Through Comparative Analysis of Ensemble Methods. *IEEE Access*, *12*, 115256–115267. https://doi.org/10.1109/access.2024.3445281

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. In *Journal of Big Data* (Vol. 6, Issue 1, pp. 1–24). https://doi.org/10.1186/s40537-019-0191-6

Altairey, H. A., & Al-Alawi, A. I. (2024). *Customer Churn Prediction in the Telecommunication Industry Using Multiple Machine Learning Algorithms*. 52–57. https://doi.org/10.1109/oidt59407.2024.11082687

Ahmed, J., Younis, I., Sarwar, U., Ghaffar, R., & Ahmed, T. (2024). Leveraging Machine Learning Models for Customer Churn Prediction in Telecommunications: Insights and Implications. *VAWKUM Transactions on Computer Sciences*, *12*(2), 16–27. https://doi.org/10.21015/vtcs.v12i2.1904

Akshara, R., & Ajay Jain, A. (2024). Data to Decisions: Optimizing E-commerce Sales Potential with Analytics. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, *2*(4), 1087–1093. https://doi.org/10.47392/irjaeh.2024.0150

Asadi Ejgerdi, N., & Kazerooni, M. (2024). A stacked ensemble learning method for customer lifetime value prediction. *Kybernetes*, *53*(7), 2342–2360. https://doi.org/10.1108/k-12-2022-1676

Ballerini, J., Yahiaoui, D., Giovando, G., & Ferraris, A. (2024). E-commerce channel management on the manufacturers' side: ongoing debates and future research pathways. *Review of Managerial Science*, *18*(2), 413–447. https://doi.org/10.1007/s11846-023-00645-w

Chang, V., Hall, K., Xu, Q., Amao, F., Ganatra, M., & Benson, V. (2024). Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models. *Algorithms*, *17*(6), 1–21. https://doi.org/10.3390/a17060231

Curiskis, S., Dong, X., Jiang, F., & Scarr, M. (2023). A novel approach to predicting customer lifetime value in B2B SaaS companies. *Journal of Marketing Analytics*, *11*(4), 587–601. https://doi.org/10.1057/s41270-023-00234-6

Durant, J. L., Asprooth, L., Galt, R. E., Schmulevich, S. P., Manser, G. M., & Pinzón, N. (2023). Farm resilience during the COVID-19 pandemic: The case of California direct market farmers. *Agricultural Systems*, *204*, 103532. https://doi.org/10.1016/j.agsy.2022.103532

Dylan, Julian., & Nathan, J. (2024). Project Management Strategies for Integrating Machine Learning into Business Analytics Initiatives. *Unique Endeavor in Business & Social Sciences*, *3*(1), 28–37.

Imani, M., Joudaki, M., Beikmohammadi, A., & Arabnia, H. (2025). Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning. *Machine Learning and Knowledge Extraction*, *7*(3), 105. https://doi.org/10.3390/make7030105

Kumar, A., Singh, K. U., Kumar, G., Choudhury, T., & Kotecha, K. (2023). Customer Lifetime Value Prediction: Using Machine Learning to Forecast CLV and Enhance Customer Relationship Management. *IEEE (Institute of Electrical and Electronics Engineers)*, 52–57. https://doi.org/10.1109/ismsit58785.2023.10304958

Kumar, K. P., Kanishkar, P., Raja, V. D., Kumar, T. A., Gopal, S. B., & Gunasekar, M. (2024). Telecom Churn Movement Prediction Using Machine Learning. *SpringerLink*, *1051*, 235–243. https://doi.org/10.1007/978-3-031-64850-2_22

Li, N., Wu, J., Shan, L., & Yi, L. (2024). Transient Stability Assessment of Power Systems Based on CLV-GAN and I-ECOC. *Energies*, *17*(10), 2278. https://doi.org/10.3390/en17102278

Mall, P. K., & Singh, P. K. (2023). Credence-Net: a semi-supervised deep learning approach for medical images. *International Journal of Nanotechnology*, *20*(5–10), 897–914. https://doi.org/10.1504/ijnt.2023.134041

Mall, P. K., Mishra, A., & Sinha, A. (2023). Comparative Analysis of Anomaly-Based Intrusion Detection System on Artificial Intelligence. *SpringerLink*, *676*, 183–194. https://doi.org/10.1007/978-981-99-1699-3_12

Muradkhanli, L. G., & Karimov, Z. M. (2023). Customer behavior analysis using big data analytics and machine learning. *Problems of Information Society*, *14*(2), 61–67. https://doi.org/10.25045/jpis.v14.i2.08

Omari, A., Al-Omari, O., Al-Omari, T., & Fati, S. M. (2025). A predictive analytics approach to improve telecom's customer retention. *Frontiers in Artificial Intelligence*, *8*, 1–15. https://doi.org/10.3389/frai.2025.1600357

Shantanu, S., & Dr Sandhya, S. (2024). *Machine Learning Techniques: Predictive Modeling for Customer Churn in Telecommunications*. https://doi.org/10.4018/978-1-6684-8792-2.ch011

Sikri, A., Jameel, R., Idrees, S. M., & Kaur, H. (2024). Enhancing customer retention in telecom industry with machine learning driven churn prediction. *Scientific Reports*, *14*(1), 1–13. https://doi.org/10.1038/s41598-024-63750-0

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*, *7*(1), 60134–60149. https://doi.org/10.1109/access.2019.2914999

Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, *14*, 100342. https://doi.org/10.1016/j.rico.2023.100342

Zha, Y., Li, Q., Huang, T., & Yu, Y. (2023). Strategic Information Sharing of Online Platforms as Resellers or Marketplaces. *Marketing Science*, *42*(4), 659–678. https://doi.org/10.1287/mksc.2022.1397