

Machine learning analysis of telecom customer churn in California

Tianyu Lan¹, Hanan Aldowah¹, Sohail Ahmed Shahani²

¹Department of Business Analytics, School of Management, Universiti Sains Malaysia (USM), George Town, Malaysia

²Department of Chemical Engineering, Universiti Teknologi PETRONAS (UTP), Seri Iskandar, Malaysia

Article Info

Article history:

Received Feb 15, 2026

Revised Jun 7, 2026

Accepted Jul 19, 2026

Keywords:

Customer churn

Logistic regression

Machine learning

Random forest

SHAP

Support vector machine

ABSTRACT

Customer churn is a significant concern in the telecommunications industry, particularly in highly competitive and price-sensitive markets such as California. Although machine learning methods are widely applied for churn prediction, their practical adoption is often limited by insufficient interpretability. This study develops an interpretable and business-oriented churn analysis framework by comparing multiple machine learning models, including logistic regression (LR), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost), combined with shapley additive explanations (SHAP)-based explanation techniques. Using a real-world California telecom customer dataset, the framework identifies key churn determinants related to contract characteristics, pricing sensitivity, and customer service behaviors. Experimental results demonstrate that XGBoost achieves the best predictive performance among the evaluated models, while SHAP provides transparent explanations of feature contributions, highlighting the importance of contractual and pricing factors for customer retention strategies.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Hanan Aldowah

Department of Business Analytics, School of Management, Universiti Sains Malaysia (USM)

George Town, Penang, Malaysia

Email: hanan@usm.my

1. INTRODUCTION

Customer churn is a persistent and economically significant issue in the telecommunications industry, particularly in highly competitive and saturated markets such as California [1]. The cost of acquiring new customers is substantially higher than that of retaining existing ones, making churn reduction a critical component of a sustainable business strategy.

Customer churn prediction has become a critical task for telecommunications companies because inaccurate churn identification can lead to customer loss and reduced business sustainability. Although machine learning approaches have significantly improved churn prediction performance, many existing studies still focus primarily on predictive accuracy while providing limited explanations of the factors influencing customer decisions. This limitation reduces the practical applicability of machine learning models in real-world telecom management systems. In recent years, increasing competition, service commoditization, and price sensitivity have further intensified churn behavior, particularly among urban and cost-sensitive customer segments. These challenges have made churn prediction an important research topic in both academic and industrial domains [2], [3].

To address this issue, machine learning (ML) and artificial intelligence (AI) techniques have been widely adopted for predictive modeling of customer churn [4]. These approaches significantly improve

prediction accuracy compared to traditional statistical methods and enable large-scale data-driven decision-making in telecom systems [5], [6].

However, despite these advancements, most existing approaches remain limited in terms of interpretability and operational usability, restricting their application in real-world decision-making environments. Therefore, this study develops an interpretable machine learning framework for customer churn analysis by integrating multiple predictive models with explainable artificial intelligence (XAI) techniques. Shapley additive explanations (SHAP) is used to enhance model transparency by identifying the contributions of individual features to churn predictions [7]. In this study, customer characteristics, including contract type, monthly charges, and service-related factors, are analyzed using a California telecommunications dataset to explore the major determinants of customer churn [8], [9].

Unlike studies that focus solely on model performance comparisons, this research emphasizes the integration of predictive capability, model reliability, and practical interpretability. Logistic regression (LR), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost) are compared using systematic model evaluation strategies, while SHAP analysis is further applied to explain the decision-making process of the best-performing model. The objective of this study is therefore to provide an interpretable analytical framework that connects machine learning prediction with actionable customer retention strategies.

2. PROPOSED APPROACH AND CONTRIBUTIONS

Customer churn prediction has been widely investigated in the telecommunications industry due to its significant impact on customer retention and business performance [10]. Previous studies have applied various statistical and machine learning methods, ranging from traditional models such as LR to advanced ensemble algorithms such as XGBoost [11]. While traditional statistical models provide better interpretability, their ability to capture complex nonlinear relationships may be limited. Conversely, machine learning models generally achieve stronger predictive performance but often suffer from reduced transparency, making it difficult for decision-makers to understand the reasons behind predictions [12].

Recent studies have introduced XAI techniques, such as SHAP, to improve the interpretability of machine learning-based churn prediction models [13]-[15]. However, many existing studies primarily emphasize algorithmic performance or methodological development, while the practical interpretation of churn drivers and their implications for customer retention remain underexplored. Furthermore, limited attention has been paid to regional differences in customer churn behavior, particularly in specific markets with unique economic and competitive characteristics such as California [16], [17]. Therefore, a research gap remains in developing churn prediction frameworks that simultaneously achieve reliable predictive performance, model interpretability, and practical applicability. Recent studies have highlighted the importance of integrating explainable machine learning with business decision support; however, further empirical investigation using regional telecom datasets and comprehensive evaluation strategies is still needed [18].

To address these limitations, this study proposes an interpretable machine learning framework for telecom customer churn prediction in the California telecommunications market.

- The proposed framework evaluates multiple machine learning algorithms, including LR, RF, SVM, and XGBoost, to compare their predictive performance and identify the most effective model for churn classification. LR provides a transparent baseline, while ensemble learning approaches such as XGBoost are adopted to capture nonlinear relationships and complex feature interactions [19].
- To enhance model interpretability, SHAP is applied to quantify feature contributions and identify the key factors influencing churn prediction [13]. This approach enables machine learning outputs to be translated into interpretable insights that support managerial decision-making [14].
- The study further contributes by conducting empirical analysis using a real-world telecom dataset, focusing on region-specific behavioral patterns in California, where customer churn is influenced by economic pressure and service heterogeneity [20].
- The proposed framework emphasizes the translation of predictive results into actionable retention strategies, aligning machine learning outcomes with business decision requirements in telecom operations [21].

3. RESEARCH METHOD

Customer churn is a core issue for organizations in highly competitive industries such as telecommunications. Higher churn rates can seriously affect an organization's profitability, customer lifecycle value, and operational planning. This study analyzes California telecom customer churn using machine learning techniques to identify key factors driving churn. The data for the study were obtained from

a California telecommunications company. By building multiple machine learning models, including LR, RF, SVM, and XGBoost, a prediction framework was constructed.

3.1. Data source

The dataset for this study was sourced from the California Telecom Customer Churn Dataset available on Kaggle. The customer churn dataset contains information on all 7,043 customers of a telecommunications company in California during the second quarter of 2022.

3.2. Descriptive statistics

In this study, descriptive statistical analysis of customer characteristics and churn behavior was conducted using the California telecom dataset, and the distribution characteristics of the dataset are summarized in Table 1. The dataset covers 1,129 cities in California, with Los Angeles representing the largest proportion of customers (4.33%), followed by San Diego (2.13%) and San Jose (1.59%), as presented in Table 1. The distribution of contract types shows that month-to-month contracts are the dominant category (55.02%), followed by two-year contracts (24.07%) and one-year contracts (20.91%). The gender distribution is relatively balanced, with male customers accounting for 50.48% and female customers accounting for 49.52%.

Regarding customer churn behavior, Table 1 indicates that 73.46% of customers were classified as non-churned, while 26.54% were classified as churned customers. This indicates a moderate class imbalance between churn and non-churn categories, which should be considered during model development and evaluation. In addition, Table 1 summarizes the numerical characteristics of customer expenditure, where total charges range from 18.80 to 8,684.80 USD, and monthly charges range from 18.25 to 118.75 USD, demonstrating considerable variation in customer spending patterns and service costs.

Table 1. Descriptive statistics of telecom customer churn

Columns	Category/Range	Percentage (%)
City	Los Angeles	4.33
	San Diego	2.13
	San Jose	1.59
	Sacramento	1.53
	San Francisco	1.48
	Fresno	0.91
	Long Beach	0.85
	Oakland	0.74
	Stockton	0.62
	Bakersfield	0.57
Contract	Other (1119 cities)	85.25
	Month-to-month	55.02
	Two years	24.07
	One year	20.91
Gender	Male	50.48
	Female	49.52
Churn label	No	73.46
	Yes	26.54
Total spend	[18.80, 8684.80]	
Monthly payment	[18.25, 118.75]	

3.3. Data preprocessing

Before model training, data preprocessing was performed to improve data quality and ensure compatibility with machine learning algorithms. Six variables were selected from the California telecom churn dataset, including contract, city, gender, monthly charges, total charges, and churn label. The total charges variable was converted into numerical format, and missing values in numerical variables were replaced using the median value, while categorical missing values were imputed using the mode value. The churn label was transformed into a binary variable, where 1 represents churned customers and 0 represents non-churned customers. Gender was also encoded into binary values. Since the city variable contained 1,129 unique categories, the 50 most frequent cities were retained, and the remaining cities were grouped into an “Other” category to reduce feature complexity. Categorical variables were then transformed using one-hot encoding, and numerical variables were standardized using z-score normalization. The processed dataset was subsequently used for machine learning model development and evaluation.

3.4. Machine learning model

The machine learning models in this study were implemented using the Scikit-learn and XGBoost Python libraries.

3.4.1. logistic regression method

LR was adopted as a baseline classification model for customer churn prediction. LR estimates the probability of churn based on the relationship between input features and the binary target variable. Due to its simplicity and interpretability, LR provides a reliable benchmark for comparing more complex machine learning algorithms. In this study, the regularization parameter C was tuned with values of 0.01, 0.1, 1, and 10, while L2 regularization was applied to control model complexity and improve generalization performance.

3.4.2. XGBoost method

XGBoost was selected as the primary predictive model because of its strong capability in handling structured customer data and modeling complex feature interactions. XGBoost constructs an ensemble of decision trees through gradient boosting optimization. In this study, several important hyperparameters were optimized, including the number of estimators (n_estimators: 100, 200, and 300), maximum tree depth (max_depth: 3, 5, and 7), learning rate (0.01, 0.05, and 0.1), subsample ratio (0.8 and 1), and column sampling ratio (colsample_bytree: 0.8 and 1). Furthermore, the optimized XGBoost model was selected for SHAP-based interpretation to analyze the contribution of individual features to customer churn prediction.

3.4.3. Random forest method

RF was employed as an ensemble learning model to capture nonlinear relationships between customer characteristics and churn behavior. RF combines multiple decision trees to improve prediction stability and reduce overfitting. In this study, the number of trees (n_estimators) was evaluated with values of 100 and 200, while the maximum tree depth (max_depth) was tested using values of 5, 8, and 12 to optimize the balance between model complexity and predictive performance.

3.4.4. SVM method

SVM was applied to identify complex decision boundaries between churned and non-churned customers. An RBF kernel was adopted to capture nonlinear patterns within customer behavior features. The regularization parameter C was optimized using values of 0.1, 1, and 10, while the gamma parameter was set to “scale” for kernel coefficient calculation.

3.5. Cross-validation and model evaluation indicators

To optimize model performance and ensure reliable evaluation, hyperparameter optimization and model validation were conducted using grid search and stratified 5-fold cross-validation. Grid search was applied to systematically evaluate predefined combinations of hyperparameters for each machine learning model, and the optimal parameter configuration was selected based on the highest ROC-AUC score. During the stratified 5-fold cross-validation process, the dataset was divided into five subsets while maintaining the original proportion of churned and non-churned customers. Each model was trained on four folds and validated on the remaining fold, and the final performance was calculated by averaging the results across five folds.

The performance of the developed models was evaluated using multiple classification indicators, including accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. Accuracy represents the overall proportion of correctly classified samples, while precision measures the reliability of churn predictions. Recall indicates the ability of the model to identify actual churn customers, and F1-score provides a balanced measurement between precision and recall. ROC-AUC was adopted as the primary evaluation metric because it measures the model's discrimination ability across different classification thresholds. In addition, PR-AUC was included because it provides a more informative evaluation for imbalanced churn prediction tasks by focusing on the minority churn class. Additionally, McNemar's test was applied to assess whether the prediction differences between XGBoost and other baseline models were statistically significant based on paired classification results.

3.6. SHAP visualization

To improve the interpretability of the XGBoost model and identify the key factors influencing customer churn, SHAP analysis was applied. SHAP is an XAI approach based on cooperative game theory, which quantifies the contribution of each feature to individual predictions by assigning a SHAP value to each

feature. A positive SHAP value indicates that a feature increases the probability of customer churn, while a negative SHAP value indicates a decreasing effect on churn prediction.

In this study, TreeSHAP was employed because XGBoost is a tree-based ensemble model. The trained XGBoost model was used to calculate SHAP values for all input features. The global importance of each feature was measured using the mean absolute SHAP value, where larger values indicate greater overall contribution to the model prediction. The top five most influential features were selected and visualized using SHAP summary plots and SHAP importance ranking to provide a comprehensive interpretation of the factors associated with telecom customer churn.

4. RESULTS AND DISCUSSION

4.1. Model performance

Table 2 presents the performance comparison of four machine learning models for telecom customer churn prediction. XGBoost achieved the best overall performance, with the highest accuracy (0.79) and F1-score (0.55), indicating its stronger ability to balance overall classification performance and churn identification capability.

RF achieved the highest precision (0.69), but its lower recall (0.35) led to a lower F1-score (0.47). LR achieved comparable performance with an accuracy of 0.78 and an F1-score of 0.51, demonstrating that linear models can still provide reasonable prediction capability. SVM showed the lowest overall performance, with an accuracy of 0.77 and F1-score of 0.48.

In addition to the performance metrics, McNemar's test was conducted to examine whether the prediction outcomes of XGBoost were statistically different from those of other classifiers. As shown in Table 3, significant differences were observed between XGBoost and LR ($p = 0.0057$) as well as SVM ($p < 0.001$). However, no significant difference was found between XGBoost and RF ($p = 0.0684$), suggesting that both tree-based ensemble models achieved comparable classification behaviors.

Table 2. Performance comparison of machine learning models

Model	Accuracy	Precision	Recall	F1-score
XGBoost	0.79	0.65	0.47	0.55
RF	0.78	0.69	0.35	0.47
LR	0.78	0.64	0.42	0.51
SVM	0.77	0.62	0.39	0.48

Table 3. Statistical comparison between XGBoost and other models using McNemar's test

Model A	Model B	Chi-square	p-value	Significant
LR	XGBoost	7.654	0.0057	Yes
RF	XGBoost	3.321	0.0684	No
SVM	XGBoost	20.679	<0.001	Yes

4.2. ROC curve

Figure 1 compares the receiver operating characteristic (ROC) curves of the four machine learning models, including LR, RF, SVM, and XGBoost. ROC analysis evaluates the capability of classifiers to distinguish churn customers from non-churn customers across different decision thresholds.

Among all evaluated models, XGBoost achieved the highest area under the curve (AUC = 0.831), followed by RF, LR, and SVM. The superior ROC performance of XGBoost indicates its stronger ability to capture nonlinear relationships among customer characteristics, enabling more accurate discrimination between churn and non-churn customers.

Although the remaining three models also demonstrate acceptable predictive capability, their ROC curves consistently remain below that of XGBoost. These findings suggest that ensemble learning methods provide better generalization performance than traditional linear or kernel-based classifiers for customer churn prediction. Overall, the ROC analysis confirms that XGBoost offers the best classification performance and therefore was selected as the final prediction model for subsequent XAI analysis.

4.3. Precision–recall curve

Since customer churn prediction involves an imbalanced classification problem, Precision–Recall (PR) analysis was additionally conducted to evaluate model performance on the minority churn class. Unlike

ROC analysis, PR curves focus specifically on the trade-off between precision and recall, making them particularly suitable for evaluating churn prediction models.

As illustrated in Figure 2, XGBoost consistently achieves the highest Precision–Recall curve among the four evaluated classifiers. The corresponding PR-AUC value of XGBoost was (0.642), which was higher than RF (0.632), LR (0.589), and SVM (0.584). This result indicates that XGBoost maintains relatively high precision while simultaneously identifying a larger proportion of actual churn customers. RF also demonstrates competitive performance, whereas LR and SVM experience a more rapid decline in precision as recall increases. The consistency between the ROC and PR analyses further verifies the robustness of XGBoost. These findings suggest that the proposed model is capable of effectively identifying potential churn customers while minimizing unnecessary retention efforts.

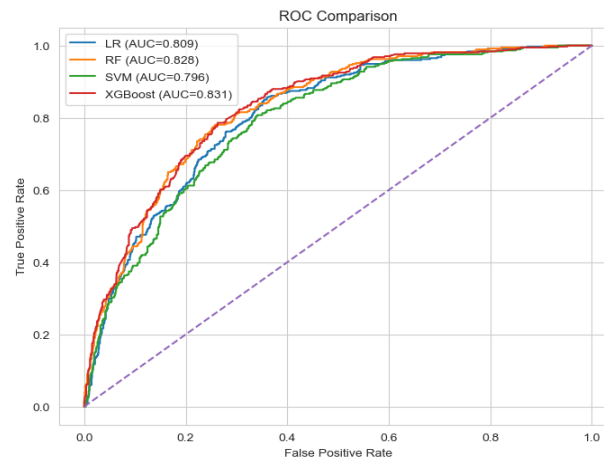


Figure 1. The comparison of the ROC curve

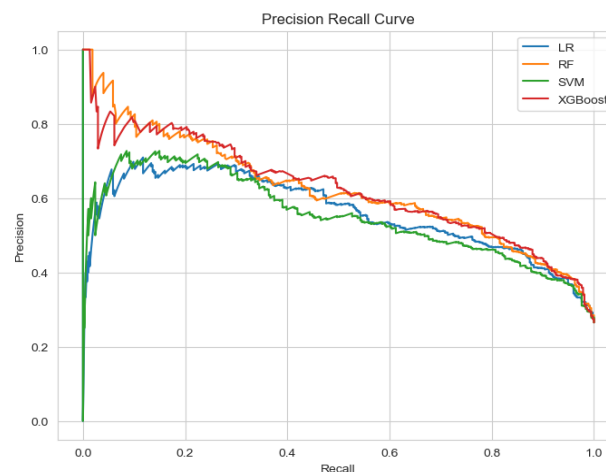


Figure 2. The comparison of the precision–recall curve

4.4. XGBoost confusion matrix

In previous evaluation metrics, XGBoost has demonstrated excellent performance. Figure 3 presents the confusion matrix of the optimized XGBoost classifier on the testing dataset. The model correctly classified 844 non-churn customers (true negatives) and 177 churn customers (true positives), while 49 non-churn customers were incorrectly predicted as churn (false positives) and 107 churn customers were misclassified as non-churn (false negatives).

The relatively small number of false positives indicates that the model effectively avoids unnecessary retention interventions for loyal customers. Meanwhile, although some churn customers remain undetected, the classifier achieves a reasonable balance between sensitivity and specificity. This balance is

particularly important in telecommunications because excessive false alarms may increase operational costs, whereas missed churn customers directly contribute to revenue loss.

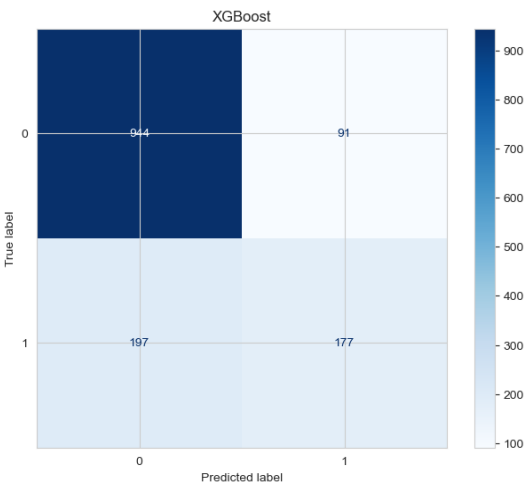


Figure 3. The XGBoost confusion matrix

4.5. SHAP visualization

Figure 4 illustrates the global feature importance derived from SHAP values. Features are ranked according to their mean absolute SHAP values, representing their average contribution to the prediction outcomes across all customers.

Monthly charges is identified as the most influential feature affecting customer churn, indicating that pricing plays a dominant role in customer retention decisions. Contract_two year ranks second, suggesting that long-term contractual commitments substantially reduce churn probability by increasing switching costs. Total charges is the third most important feature, reflecting customers’ historical investment in the service. Contract_One Year also contributes to churn prediction, although its influence is weaker than that of two-year contracts. Finally, City_San Diego appears among the top five variables, implying that regional market characteristics exert a modest influence on customer behavior. The SHAP feature importance analysis demonstrates that financial and contractual variables dominate churn prediction, whereas geographical factors contribute relatively less to the model’s overall performance.

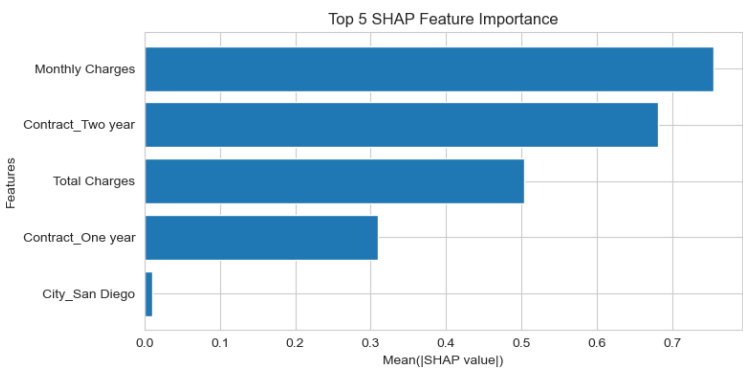


Figure 4. SHAP feature importance

Figure 5 presents the SHAP summary plot, which simultaneously illustrates both the magnitude and direction of feature contributions to customer churn predictions. Each point represents one customer, and the color gradient indicates the corresponding feature value, with red indicating higher values and blue indicating lower values.

Monthly charges exhibit the largest impact on model predictions. Higher monthly charges generally yield positive SHAP values, indicating a higher probability of customer churn. Conversely, customers with

relatively lower monthly fees tend to exhibit negative SHAP values, suggesting a stronger likelihood of remaining with the service provider.

The effect of Contract_Two Year is predominantly negative, demonstrating that customers enrolled in long-term contracts are substantially less likely to churn. Similarly, Contract_One Year also reduces churn probability, although its influence is less pronounced than that of two-year contracts. Total charges display a moderate contribution, indicating that accumulated spending reflects customer loyalty to some extent but is less influential than current pricing policies. In contrast, City_San Diego exhibits comparatively limited SHAP variation, suggesting that geographical location plays only a secondary role in predicting customer churn.

The SHAP summary plot therefore confirms that customer churn is primarily driven by contractual commitment and pricing-related factors rather than demographic or regional characteristics. This finding provides valuable business insights for telecommunications companies seeking to improve customer retention through pricing optimization and contract management strategies.

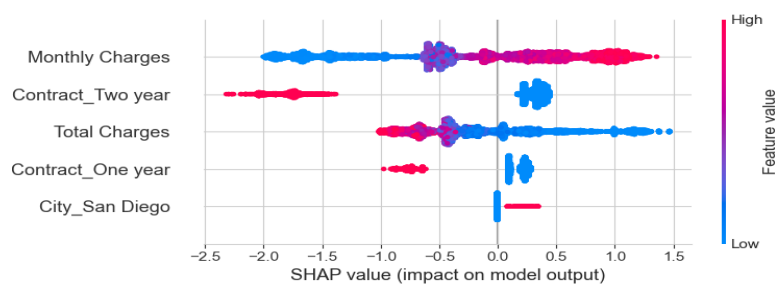


Figure 5. SHAP summary plot

4.6. Discussion and recommendations

This study developed and evaluated four machine learning models, including LR, RF, SVM, and XGBoost, for telecom customer churn prediction in California. The experimental results demonstrate that XGBoost achieved the best predictive performance, while SHAP analysis further provided interpretable insights into the major factors influencing customer churn. The findings highlight the importance of contractual and financial characteristics in churn behavior and demonstrate the potential of combining machine learning and XAI for telecom customer retention management.

4.6.1. Key findings and interpretation

The model comparison results indicate that XGBoost achieved the strongest predictive capability among all evaluated models, obtaining the highest accuracy, ROC-AUC, and PR-AUC values. This superior performance suggests that customer churn behavior involves complex nonlinear relationships that can be effectively captured by gradient boosting algorithms. Although RF also demonstrated competitive performance, XGBoost achieved a better balance between precision and recall, indicating a stronger capability in identifying potential churn customers.

The SHAP analysis reveals that monthly charges and contract-related variables are the most influential factors affecting churn prediction. Higher monthly charges increase the likelihood of churn, suggesting that perceived pricing and service affordability play critical roles in customer retention decisions. In contrast, long-term contracts, particularly two-year contracts, reduce the probability of churn by increasing customer commitment and switching costs. These findings indicate that customers with flexible contracts and higher recurring costs represent higher-risk segments.

Additionally, total charges contributed significantly to churn prediction, indicating that accumulated customer spending provides useful information regarding customer relationship status. However, geographical factors such as City_San Diego showed relatively limited influence compared with contractual and financial variables. This suggests that customer churn is primarily driven by service-related and economic factors rather than demographic or regional characteristics.

4.6.2. Comparison with previous studies

The findings are consistent with previous studies on telecom customer churn prediction [1]. Consistent with prior research, this study confirms that pricing factors and customer commitment mechanisms are important determinants of churn behavior in competitive telecommunications markets. Previous studies have also emphasized that increasing competition and market liberalization intensify customer switching behavior [22], [23].

Previous research has demonstrated the effectiveness of machine learning approaches, including LR and ensemble learning methods, for churn prediction tasks [24]. The current study further supports these findings by demonstrating that XGBoost achieves better predictive performance than traditional models by capturing nonlinear feature interactions. Meanwhile, LR remains valuable as an interpretable baseline model, reflecting the trade-off between prediction accuracy and model transparency discussed in previous studies [18].

Although XAI methods, such as SHAP, have been increasingly applied in predictive modeling research [13], [14], many existing studies focus mainly on algorithm development. By integrating SHAP with XGBoost in a telecom churn prediction framework, this study provides interpretable evidence linking machine learning predictions with practical customer retention strategies.

4.6.3. ICT and managerial implications

From an ICT perspective, this study demonstrates the value of integrating machine learning and explainable AI techniques into telecom decision-support systems [9]. The proposed framework combines predictive capability with model interpretability, enabling telecom operators to identify high-risk customers and understand the reasons behind churn predictions. Such an approach can support automated churn monitoring, customer risk assessment, and intelligent retention management systems [25], [26].

From a managerial perspective, the results suggest that retention strategies should prioritize customers with month-to-month contracts and relatively high monthly charges. Telecom providers can reduce churn risk by introducing contract upgrade incentives, personalized pricing plans, and targeted retention campaigns. Furthermore, since SHAP results indicate that contractual and financial variables contribute more significantly than demographic features, customer segmentation strategies should focus on behavioral and service characteristics rather than demographic information [6], [27].

4.6.4. Limitations and future research

This study has several limitations. First, the dataset was obtained from a single telecom market in California, which may limit the generalizability of the findings to other regions with different market structures. Second, the dataset mainly contains demographic, contractual, and billing information, while behavioral factors such as customer interactions, service complaints, and sentiment information are not included.

Future research could incorporate dynamic customer behavior data, sentiment analysis, and advanced deep learning approaches to capture more complex churn patterns. Furthermore, cross-regional and cross-country studies could be conducted to investigate how different market environments influence customer churn behavior and retention strategies.

5. CONCLUSION

This study investigated telecom customer churn prediction in California by integrating machine learning models and XAI techniques. Four classification models, including LR, RF, SVM, and XGBoost, were developed and evaluated using multiple performance indicators. The results demonstrated that XGBoost achieved the best predictive performance, indicating its effectiveness in capturing complex relationships among customer characteristics. The SHAP-based analysis further revealed that customer churn was mainly influenced by contractual and financial factors. Monthly charges, contract type, and total charges were identified as the most important predictors, whereas demographic and regional factors contributed relatively little. These findings suggest that telecom operators should prioritize pricing strategies, contract management, and targeted retention programs for high-risk customer segments.

From an ICT perspective, integrating machine learning and explainable AI offers a practical framework for developing transparent, data-driven customer retention systems. By combining accurate prediction with interpretable insights, the proposed approach can support more effective decision-making in telecom customer management.

This study is limited by the use of a single-region dataset and the absence of dynamic behavioral information. Future research could incorporate customer interaction data, sentiment analysis, and cross-regional datasets to further improve the generalizability and predictive capability of churn models.

FUNDING INFORMATION

The authors state that no funding was involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Tianyu Lan	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Hanan Aldowah		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		
Sohail Ahmed Shahani		✓	✓	✓		✓				✓				

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The dataset used in this study is publicly available and open-access. It does not contain any sensitive or personally identifiable information. All analyses and experiments conducted in this study are fully reproducible using this dataset. Dataset URL: <https://www.kaggle.com/datasets/shilongzhuang/telecom-customer-churn-by-maven-analytics/data>.

REFERENCES

- [1] B. Ç. Uslu *et al.*, "Customer churn prediction analysis in a telecommunication company with machine learning algorithms," *Journal of Industrial Engineering*, vol. 32, no. 3, pp. 496–512, 2021, [Online]. Available: <http://orcid.org/0000-0001-8214-825Xhttp://orcid.org/0000-0002-0271-5865>
- [2] V. Mahajan, R. Misra, and R. Mahajan, "Review on factors affecting customer churn in telecom sector," *International Journal of Data Analysis Techniques and Strategies*, vol. 9, no. 2, pp. 122–144, 2017, doi: 10.1504/IJDATS.2017.085898.
- [3] A. Eshghi, D. Haughton, and H. Topi, "Determinants of customer loyalty in the wireless telecommunications industry," *Telecommunications Policy*, vol. 31, no. 2, pp. 93–106, Mar. 2007, doi: 10.1016/j.telpol.2006.12.005.
- [4] M. Z. Alotaibi and M. A. Haq, "Customer churn prediction for telecommunication companies using machine learning and ensemble methods," *Engineering, Technology and Applied Science Research*, vol. 14, no. 3, pp. 14572–14578, Jun. 2024, doi: 10.48084/etasr.7480.
- [5] P. K. Dalvi, S. K. Khandge, A. Deomore, A. Bankar, and V. A. Kanade, "Analysis of customer churn prediction in telecom industry using decision trees and logistic regression," in *2016 Symposium on Colossal Data Analysis and Networking, CDAN 2016*, IEEE, Mar. 2016, pp. 1–4. doi: 10.1109/CDAN.2016.7570883.
- [6] H. Aldowah, S. A. Shahani, A. Saif, and A. Alsswey, "Machine learning solutions for talent loss: predictive insights and strategic countermeasures," *International Journal of Academic Research in Business and Social Sciences*, vol. 15, no. 5, May 2025, doi: 10.6007/ijarbs/v15-i5/25064.
- [7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4768–4777, 2017.
- [8] H. Aldowah, H. Al-Samarraie, S. U. Rehman, I. N. Umar, and A. I. Alzahrani, "Mapping online engagement: a DEMATEL-based analysis of instructor perspectives," *International Journal of Information Technology and Decision Making*, vol. 25, no. 06, pp. 2137–2177, Jul. 2025, doi: 10.1142/S0219622025501056.
- [9] N. Nguyen, T. H. Nguyen, Y. N. Nguyen, D. Doan, M. Nguyen, and V. H. Nguyen, "Machine learning-based model for customer emotion detection in hotel booking services," *Journal of Hospitality and Tourism Insights*, vol. 7, no. 3, pp. 1294–1312, Jun. 2024, doi: 10.1108/JHTI-03-2023-0166.
- [10] J. H. Ahn, S. P. Han, and Y. S. Lee, "Customer churn analysis: churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry," *Telecommunications Policy*, vol. 30, no. 10–11, pp. 552–568, Nov. 2006, doi: 10.1016/j.telpol.2006.09.006.
- [11] J. Bhattacharyya and M. K. Dash, "What do we know about customer churn behaviour in the telecommunication industry? A bibliometric analysis of research trends, 1985–2019," *FIIB Business Review*, vol. 11, no. 3, pp. 280–302, Sep. 2022, doi: 10.1177/23197145211062687.
- [12] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," *Journal of Business Research*, vol. 94, pp. 290–301, Jan. 2019, doi: 10.1016/j.jbusres.2018.03.003.
- [13] A. M. Salih *et al.*, "A perspective on explainable artificial intelligence methods: SHAP and LIME," *Advanced Intelligent Systems*, vol. 7, no. 1, Jan. 2025, doi: 10.1002/aisy.202400304.
- [14] A. El Attar and M. El-Hajj, "Explainable AI-driven customer churn prediction: a multi-model ensemble approach with SHAP-based feature analysis," *Frontiers in Artificial Intelligence*, vol. 9, Feb. 2026, doi: 10.3389/frai.2026.1748799.
- [15] H. Aldowah, S. U. Rehman, and S. A. Shahani, "Addressing data privacy concerns in digital emerging technologies: Strategies and best practices," in *Challenges and Solutions for Cybersecurity and Adversarial Machine Learning*, IGI Global Scientific Publishing, 2025, pp. 449–475. doi: 10.4018/979-8-3373-2200-1.ch014.
- [16] D. H. Zhu, Z. Z. Deng, and Y. P. Chang, "Understanding the influence of submission devices on online consumer reviews: a comparison between smartphones and PCs," *Journal of Retailing and Consumer Services*, vol. 54, p. 102028, May 2020, doi: 10.1016/j.jretconser.2019.102028.
- [17] S. Ul Rehman and H. Aldowah, "Integration of artificial intelligence and IoT for environmental sustainability," in *Emerging Technologies in the Healthcare System for Sustainable Development*, Elsevier, 2025, pp. 253–264. doi: 10.1016/B978-0-443-26492-4.00014-1.

- [18] M. A. Shaikhsurab and P. Magadum, "Enhancing customer churn prediction in telecommunications: an adaptive ensemble learning approach," *arXiv preprint arXiv:2408.16284*, 2024, [Online]. Available: <http://arxiv.org/abs/2408.16284>
- [19] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, "Customer churn prediction in telecom sector using machine learning techniques," *Results in Control and Optimization*, vol. 14, p. 100342, Mar. 2024, doi: 10.1016/j.rico.2023.100342.
- [20] J. Y. V. Quek, Y. H. Pang, Z. Y. Lim, S. Y. Ooi, and W. H. Khoh, "Customer churn prediction through attribute selection analysis and support vector machine," *Journal of Telecommunications and the Digital Economy*, vol. 11, no. 3, pp. 180–194, Sep. 2023, doi: 10.18080/jtde.v11n3.777.
- [21] L. S. Ling, N. Mustafa, and S. F. A. Razak, "Customer churn prediction for telecommunication industry: a malaysian case study," *F1000Research*, vol. 10, p. 1274, Dec. 2021, doi: 10.12688/f1000research.73597.1.
- [22] X. Lin, T.-J. Lu, and X. Chen, "Technological Innovation, market competition, and regulatory reform in telecommunications," *Wireless Personal Communications*, vol. 102, no. 2, pp. 997–1007, Sep. 2018, doi: 10.1007/s11277-017-5128-9.
- [23] M. E. Meena and J. Geng, "Dynamic competition in telecommunications: a systematic literature review," *SAGE Open*, vol. 12, no. 2, Apr. 2022, doi: 10.1177/21582440221094609.
- [24] A. Idris, A. Khan, and Y. S. Lee, "Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification," *Applied Intelligence*, vol. 39, no. 3, pp. 659–672, Oct. 2013, doi: 10.1007/s10489-013-0440-x.
- [25] D. Asif, M. S. Arif, and A. Mukheimer, "A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry," *Results in Engineering*, vol. 26, p. 104629, Jun. 2025, doi: 10.1016/j.rineng.2025.104629.
- [26] P. D. Joshi and R. Shende, "Predicting customer churn in telecom industry using machine learning algorithms," in *3rd IEEE International Conference on Networks, Multimedia and Information Technology, NMITCON 2025*, IEEE, Aug. 2025, pp. 1–6, doi: 10.1109/NMITCON65824.2025.11187974.
- [27] T. C. Johnston and M. A. Hewa, "Fixing service failures," *Industrial Marketing Management*, vol. 26, no. 5, pp. 467–473, Sep. 1997, doi: 10.1016/S0019-8501(96)00158-7.

BIOGRAPHIES OF AUTHORS



Tianyu Lan    received his bachelor's degree in Statistics from Datong University, China, in 2023. He obtained his master's degree in Business Analytics from Universiti Sains Malaysia (USM), Malaysia, in 2025. He is currently engaged in research in the areas of machine learning and data analytics. His research interests include machine learning, statistical modeling, and customer churn prediction. He can be contacted at email: lanta@ldy.edu.rs.



Hanan Aldowah    is a senior lecturer at Universiti Sains Malaysia (USM) and holds a Ph.D. and a master's in IT and Multimedia from the same university, along with a bachelor's in computer science. Her extensive expertise spans diverse domains, including data analytics, data mining, ML, artificial intelligence, IoT, Cognitive Maps, VR and augmented reality, and instructional technology. Aldowah holds a keen interest in applying emerging technologies to enhance sectors such as management, education, and computer education, as well as in technology adoption, digital transformation, and organizational behavior. She can be contacted at email: hanan@usm.my.



Sohail Ahmed Shahani    serves as a teaching assistant at Universiti Teknologi PETRONAS (UTP) in Malaysia. His doctoral research, culminating in a Ph.D. in chemical engineering with a focus on pharmaceuticals and life sciences, has equipped him with strong computational simulation and modeling skills. His current work applies these techniques to challenges within the food and pharmaceutical industries. His specific areas of expertise include the development of drug-delivery formulations, environmental toxicology, advanced biochemical characterization, CO₂ capture and separation, pharmaceutical formulations, COSMO-RS, machine learning, and DFT techniques. He can be contacted at email: sohail_21000836@utp.edu.my.