Customer Churn Prediction in Telecommunications Using Machine Learning Models

Jiaqi Cao

{ jcao0866@uni.sydney.edu.au }

Business School, The University of Sydney, Sydney, New South Wales 2006, Australia

Abstract. Customer churn prediction has a great impact on the long-term profitability and competitiveness of telecommunications companies. This study uses telecommunications customer data, employs machine learning methods to build a customer churn prediction model, and identifies key influencing factors. By using majority class under-sampling and threshold optimization to alleviate the problem of class imbalance, the overall model performance is improved. Experimental results show that logistic regression, radial basis function kernel support vector machine, random forest, and gradient boosting all have strong predictive ability. The random forest model performs the most evenly, with a recall rate of 0.84 and an F1 score of 0.79 and is selected as the optimal model. The feature importance analysis shows that service availability, Internet service type and cumulative billing amount are the core factors affecting customer churn. These findings provide effective forecasting models and insights for the telecommunications industry to improve customer retention strategies.

Keywords: Customer churn; Telecommunications; Machine learning; Random Forest; Feature importance analysis

1 Introduction

Customer churn is regarded as one of the important challenges in the telecommunications industry. Fierce competition and low conversion costs make it easy for customers to change service providers. Customer churn will directly affect revenue stability and long-term profitability. Therefore, customer retention is the focus of telecommunications operators. Research focusing on deep learning for churn prediction emphasizes that the cost of retaining existing customers is much lower than acquiring new ones, which makes early identification of customer churn very significant for telecommunications operators [1]. The article on Customer churn prediction in the telecom sector further points out that with the increasing growth of customer data, the prediction model can be used to support customer retention strategies [2]. The empirical evidence of the telecommunications case study shows that the phenomenon of customer churn is closely related to customer satisfaction indicators and service experience, which reinforces the necessity of systematic loss analysis [3]. In addition, because operational

ICIAAI 2026, May 22-24, Singapore, Singapore
Copyright © 2026 EAI
DOI 10.4108/eai.22-5-2026.2365122

data is closely related to decision-making, churn prediction has become a core task in the customer relationship management framework [4].

A large number of studies have used a variety of machine learning methods to predict customer churn. Data mining research applies traditional classifiers, including decision trees, simple Bayes and SVM, to prove that preprocessing and feature selection can improve prediction accuracy [5]. Comparative studies based on the Kaggle telecommunications dataset show that ensemble methods such as random forests and gradient boosting consistently perform better than single classifiers, with accuracy rates typically ranging from 75% to 82%, but due to class imbalance, the recall rate for churned users often falls below 60% [6]. The research focusing on interpretability introduces model- agnostic interpretation techniques, such as SHAP, to solve the black box characteristics of high-performance models. And identified the key drivers of churn, including customer duration, contract type and monthly fee standard [7]. More advanced research employed ensemble optimization and imbalance handling techniques, and when the boosted model was combined with sampling strategies, the accuracy rate could exceed 90%, but such results were mostly obtained in controlled experimental environments [8]. Table machine learning research further proves that the explainable gradient boosting model can achieve a balance between predictive performance and explainability, thereby more clearly identifying the loss effect at the feature level [9]. Research based on neural networks and using Relief-F feature selection shows that dimension reduction can significantly improve model performance, and the accuracy rate is increased to more than 90% compared with the decision tree benchmark model [10]. The large-scale experiment in 2024 adopts random forest and boosting tree models. When applying explainable AI-assisted decision-making, the accuracy rate exceeded 90%, and the recall rate reached more than 80% [11]. Research based on ensemble models shows that combining multiple classifiers can enhance robustness, but performance improvement between different data sets is not inevitable [12].

Although a lot of progress has been made on the prediction of customer churn, there are still some limitations. Many studies rely on highly unbalanced datasets, among which the number of non- churning customers far exceeds that of churning customers. As a result, although the model can achieve a high overall accuracy rate, it cannot effectively identify actual churn cases. In addition, most of the existing studies only evaluate churn factors at the overall level, and do not explore the differences in churn drivers under the characteristics of different customer groups. This reduces the value of churn prediction in the formulation of accurate retention strategies. To solve these problems, this study uses majority undersampling techniques to alleviate class imbalances before model training and evaluates various machine learning models under a consistent preprocessing and cross-validation framework. Evaluate logistic regression, random forest, SVM and gradient boosting models through accuracy, precision rate, recall rate, F1 score and ROC-AUC indicators. In addition, analyse the churn behaviour of different customer groups, and use permutation feature importance to identify key churn drivers to provide interpretable insights for predictive modelling and decision-making.

2 Dataset

2.1 Descriptive overview

This study adopts Kaggle's telecommunications customer dataset, which includes basic customer information, service subscription records, billing data and churn status. The dataset contains 7,043 customer records and 21 variables. The target variable is used to identify whether the customer is lost ("Yes" or "No").

In the original dataset, 5,174 customers were marked as not churning ("No") and 1,869 customers were marked as churning ("Yes"), with an overall churn rate of 26.54%. This indicates that most churning customers were from the non-churning category. To mitigate the impact of class imbalance on model training, the majority class undersampling technique is adopted before data splitting. After balanced processing, the dataset contains 3,738 observation records, including 1,869 churning customers and 1,869 non-churning customers, achieving a balanced churn rate of 50%. This method only reduces the amount of most types of observation data, does not generate synthetic data, and retains the original feature structure.

To capture the differences of customer groups, customers are divided into four consumption levels according to monthly spending: low, med-low, med-high, and high. The churn rate at each level of consumption (Figure 1) shows an increasing trend. Low-spending customers have the lowest churn rate, while med-high and high spending customers have a higher churn rate. This pattern shows that high spending customers have higher expectations for service quality, which makes them more sensitive to dissatisfaction and more prone to churn.

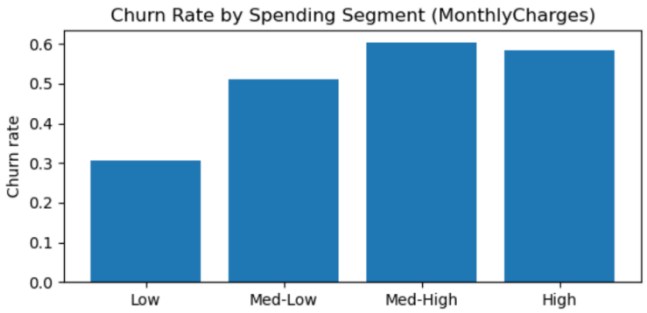

Fig. 1. Churn Rate by Segment (Picture credit: Original)

2.2 Data analysis and preprocessing

Numerical feature analysis (Figure 2) reveals that the retention time of churned customers is shorter, and the churn is concentrated in the early stage of service. In contrast, non-churn customers have a longer retention time and a more even distribution. The monthly fee shows the opposite trend. Churned customers usually incur higher monthly costs than retained customers, but due to the short service cycle, the cumulative cost is generally lower. In addition, customers with fewer subscriptions to additional services are more likely to churn, while customers with a large number of subscriptions tend to show higher retention rates.

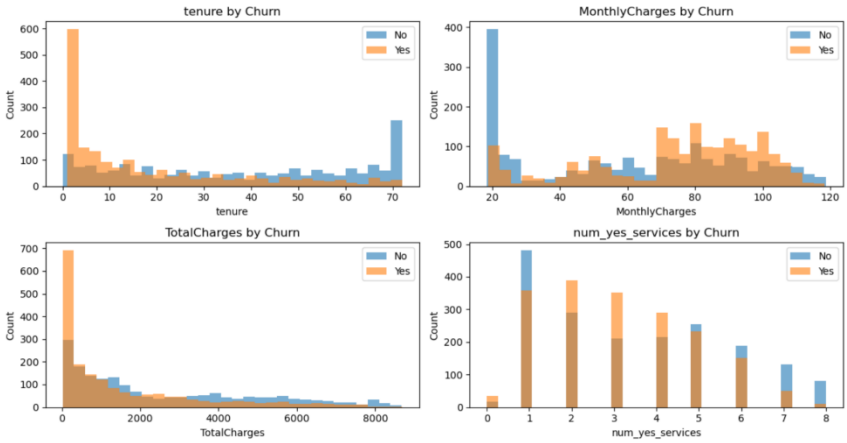

Fig. 2. Numerical Feature Analysis (Picture credit: Original)

Category feature analysis (Figure 3) reveals that the churn rate of monthly payment customers is significantly higher than that of one-year or two-year contract customers. The churn rate of customers who pay by electronic check is higher than that of customers who pay by automatic bank transfer or credit card. The churn rate of optical fibre users is higher than that of DSL users and users without internet. In addition, the lack of online security or technical support services is also associated with the increase in the churn rate.

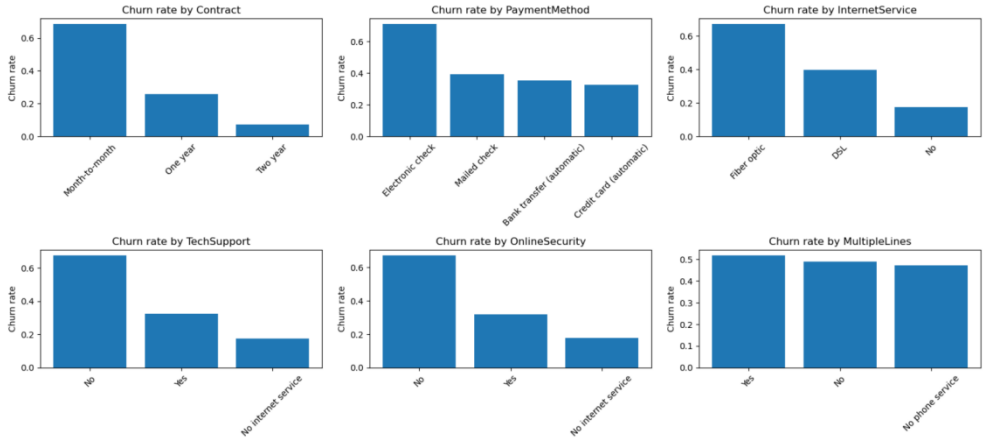

Fig. 3. Categorical Feature Analysis (Picture credit: Original)

Before the model, the data set is preprocessed, including processing missing values, encoding categorical variables, and appropriate scaling of numerical features. Some additional features have been built. Divide customers into early, medium and long-term customer groups through tenure_group. Aggregate the number of subscription services marked as "yes" into num_yes_services characteristics. Network usage is summarised by the binary has_internet indicator, which comes from InternetService. In addition, in the model input stage, the numerical features are standardised and processed, and the categorical features are one-hot encoded.

3 Methodology

3.1 Model training and evaluation strategy

This study uses four supervised classification models: logistic regression, random forest, support vector machine based on radial basis function (RBF) kernel, and gradient boosting. All models are trained using the same preprocessing process.

After preprocessing and class balance, the dataset is divided into training set and test set by stratified sampling. 80% of the observations are training sets and 20% are test sets. Stratified sampling ensures that the two subsets maintain a balanced distribution of lost and non-lost categories. The test set is completely retained and not used in the process of model tuning or threshold selection.

Model training and hyperparameter optimisation are only carried out on the training set. The construction process of each model is consistent to ensure that the preprocessing steps are consistent in the fitting process. Hyperparameters are optimised by RandomisedSearchCV. The method combines sample parameters from predefined intervals instead of exhausting all possibilities and can still explore a wide range of hyperparameters while improving computing efficiency. The main optimisation metric selects the F1 score to balance the precision and recall in the class imbalance scenario.

Hyperparameter tuning adopts 5-fold stratified cross-validation. The training set is divided into five groups according to the proportion of categories. In each iteration, four groups are used for training, one group is used for validation, and finally the average performance of each group is taken. This process reduces the dependence on single training- validation division, making hyperparameter selection more stable. Cross- validation is only used in the tuning stage, not in the final model evaluation.

After selecting the optimal hyperparameters, all models are re-fitted on the complete training set. To improve the classification balance, threshold optimisation is used to replace the fixed 0.5 decision threshold. Evaluate the candidate threshold in the range of 0.10 to 0.90 through the cross-validated out-of-fold prediction probability of the training set. The threshold selection prioritises the F1 score, while setting the minimum precision rate and the recall rate threshold to avoid extreme trade-offs. The selected threshold is fixed in the test stage. The final performance metrics are calculated based on the undisturbed test set to ensure the unbiased evaluation of the generalisation ability.

3.2 Model interpretation

Logistic regression is used as a benchmark model due to its transparency and interpretability. It models the probability of customer churn by applying the sigmoid function to the linear combination of input features. Its formula is:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}} \tag{1}$$

where Y represents the outcome, X denotes the feature vector, and β are the model coefficients. However, when the churn behaviour is nonlinear interaction, the linear decision boundary may limit its performance.

Gradient boosting is an ensemble learning method. It constructs a series of weak learners, usually decision trees. Each new model will correct the errors of the previous one. Its formula is:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{2}$$

where $h_m(x)$ is the weak learner at iteration m, and γ_m is the learning rate. This structure enables the model to capture the complex nonlinear relationship between customer attributes, service usage and billing behaviour.

Random forests construct multiple decision trees through bootstrapped sampling of training data and randomly selected feature subsets. The final prediction will be obtained by a majority vote. Its formula is:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_B(x)\} \tag{3}$$

where $T_b(x)$ is the prediction of the b-th tree. This integrated design can reduce the variance and enhance robustness, but its interpretability will be reduced compared with simple models.

SVM determines the decision boundary by maximising the interval between classes. When the data cannot be separated linearly, the radial basis function kernel is used. Its formula is:

$$K(x_i, x_j) = exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right) \tag{4}$$

where γ controls the influence of individual data points. The radial basis function kernel has flexible nonlinear decision boundaries, making SVM suitable for complex customer churn patterns. However, its computational complexity is higher than that of tree-based methods.

4 Results

4.1 Model performance comparison and selection

The performance of the model is evaluated by a variety of indicators, including accuracy, precision, recall, F1 score and area under the receiver operating characteristic curve (ROC-AUC).

As Table 1 shows, the logistic regression model achieves the highest overall accuracy and F1 score. The random forest and gradient boosting models also performed well, with slightly lower accuracy but higher recall rate. This shows that these two models are more sensitive to churn customers and better at identifying positive churn cases. SVM (RBF) is slightly inferior in terms of recall rate and ROC-AUC.

Table 1. Model Performance Evaluation.

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.7794	0.7555	0.8262	0.7892	0.8549
Random Forest	0.7727	0.7394	0.8422	0.7875	0.8545
Gradient Boosting	0.7513	0.7117	0.8449	0.7726	0.8502
SVM (RBF)	0.7701	0.7715	0.7673	0.7694	0.8362

The ROC curve in Figure 4 confirms that all models are significantly better than random classification. The dense distribution of curves shows that the overall discrimination ability of the model is comparable. Logistic regression has a slight advantage in the ROC-AUC metric, but the difference between models is small.

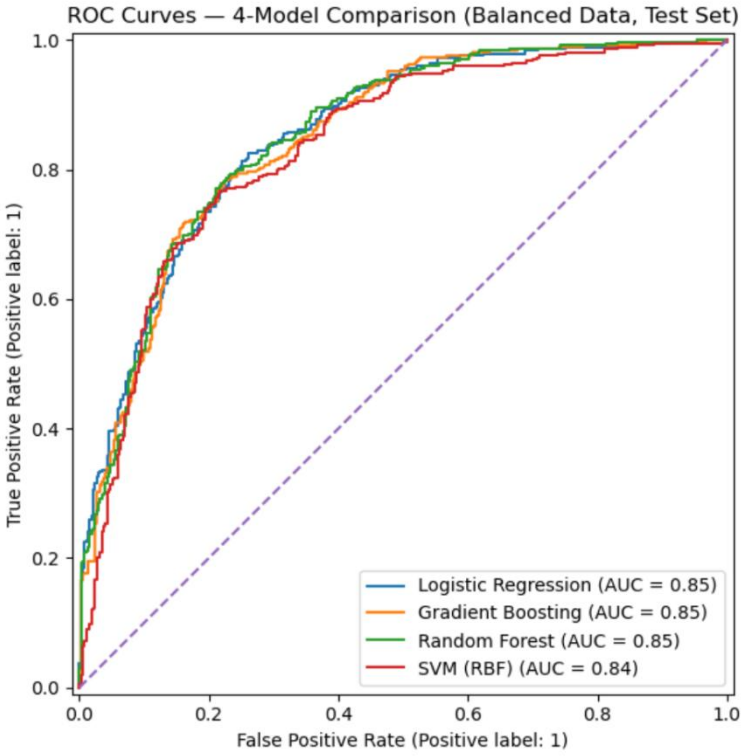

Fig. 4. ROC Curves (Picture credit: Original)

Due to the small performance gap, the model selection needs to comprehensively consider the accuracy, recall rate and interpretability. Considering a higher recall rate, the ability to capture nonlinear feature interactions without strong parameter assumptions, the random forest model was finally selected as the optimal model for subsequent analysis.

4.2 Overall feature importance analysis

Figure 5 shows the analysis of the feature importance of the overall permutation method based on the random forest model. The results show that the lack of multi-line telephone service is the most critical factor affecting customer loss. Whether the customer uses the Internet service and the type of connection are also of high importance. The total cost is also another strong predictor.

In contrast, the monthly fee itself is less important, indicating that short-term pricing cannot fully explain customer churn behaviour. Although marital status also has an impact on loss, its impact is less than service-related variables.

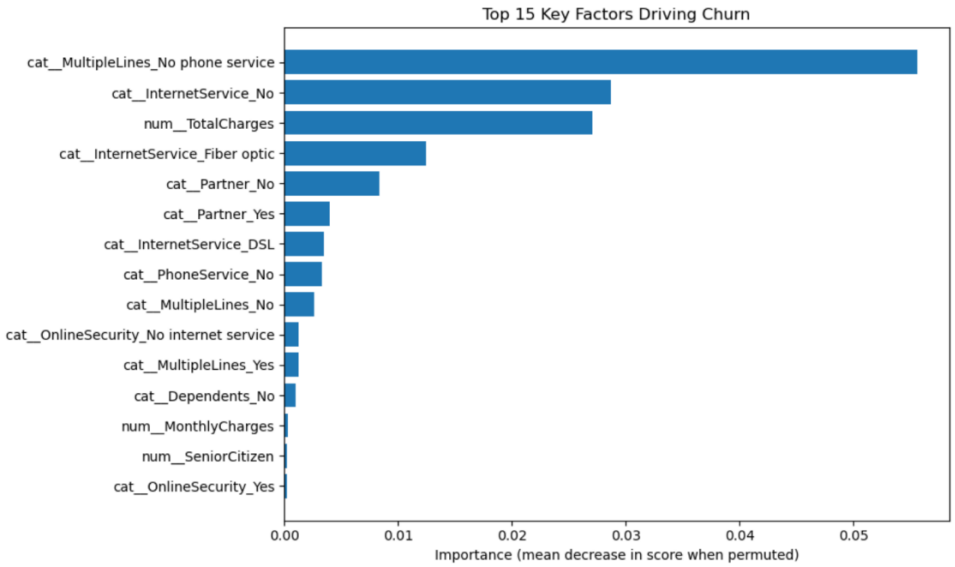

Fig. 5. Top 15 Key Factors Driving Churn (Picture credit: Original)

4.3 Functional importance across different consumer segments

To find out whether the driver of churn varies according to the type of customer, the customers in the test set are divided into four consumption segments according to the monthly cost. Four groups of low, medium-low, medium-high, and high are formed by quartile thresholds (Table 2). This segmentation method ensures that the number of customers in each group is equal.

Table 2. Customer Segmentation Criteria.

Segment	Monthly Charges Range
Low	Bottom 25%
Med-Low	25–50%
Med-High	50–75%
High	Top 25%

The results are shown in Figure 6. The results show that not having the multi-line telephone service and not having Internet access have always been the most important factors in all segments. However, there is a difference between low-consumption and high-consumption customer groups. For low-consumption and medium-low consumption groups, the churn rate is more closely related to the availability of basic services and total cost. In contrast, the churn of medium-high consumption and high-consumption customers is more closely related to service complexity, network configuration and perceived value.

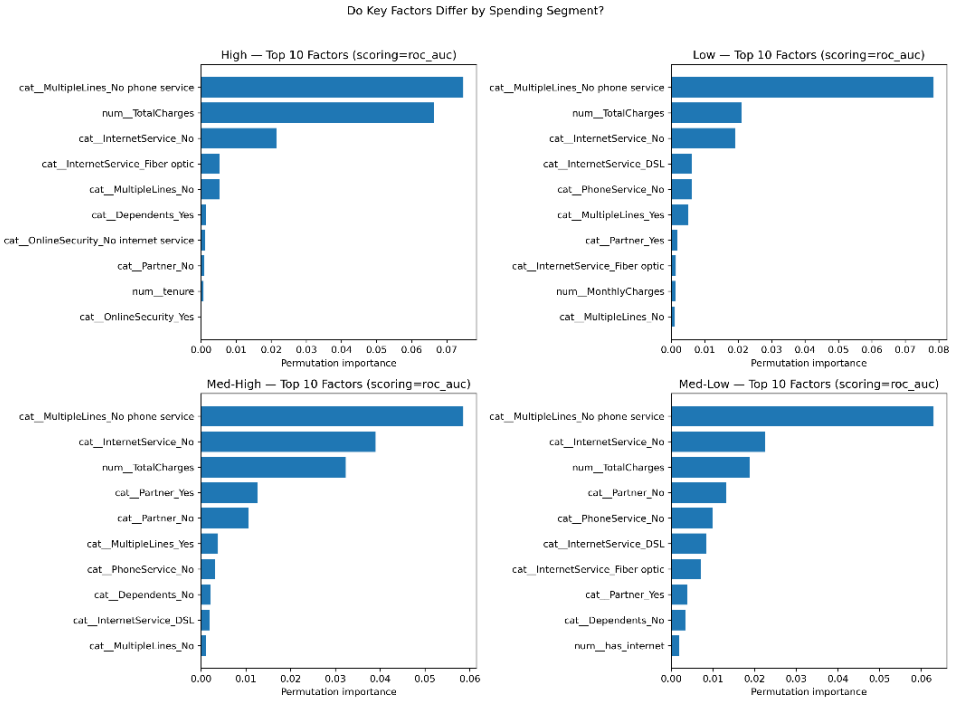

Fig. 6. Key Factors Differ by Spending Segment (Picture credit: Original)

5 Conclusion

Customer churn has always been a major challenge for telecommunications companies. Accurately identifying customers at risk of churn is crucial to improving customer retention strategies and maintaining long-term profitability. This study evaluates the predictive performance and identifies important factors related to the churn behaviour by applying four machine learning models to the telecommunications customer dataset. The results show that all models can achieve reasonable predictive performance under balanced data training. However, integrated models, especially random forests, show stronger recall capabilities and the advantages of identifying lost customers. The feature importance analysis shows that customer churn is mainly driven by service-related factors, such as the lack of multi-line telephone services and Internet connections, rather than simply determined by pricing factors. Hierarchical analysis further reveals that there are differences in the drivers of churn in different consumer groups, indicating that customer behaviour is not homogenised.

However, this study still has limitations: the dataset only covers a single telecommunications scenario and fails to capture the chronological changes of customer behaviour. In addition, although majority class undersampling can effectively solve the problem of category imbalance, it may lose potential information in most classes. These factors may limit the generalizability of the research results. Future research can expand this research by introducing time series data, alternative sampling strategies or more advanced deep learning models.

References

[1] Fujo, S. W., Subramanian, S., Khder, M. A.: Customer churn prediction in telecommunication industry using deep learning. Information Sciences Letters 11(1). pp. 24-24 (2022)

[2] Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., Gawande, S. H.: Customer churn prediction in telecom sector using machine learning techniques. Results in Control and Optimization 14. pp. 100342 (2024)

[3] Mustafa, N., Ling, L. S., Razak, S. F. A.: Customer churn prediction for telecommunication industry: A Malaysian case study. F1000Research 10. pp. 1274 (2021)

[4] Melian, D. M., Dumitrache, A., Stancu, S., Nastu, A.: Customer churn prediction in telecommunication industry: A data analysis techniques approach. Postmodern Openings 13(1 Sup1). pp. 78-104 (2022)

[5] Khalid, L. F., Abdulazeez, A. M., Zeebaree, D. Q., Ahmed, F. Y., Zebari, D. A.: Customer churn prediction in telecommunications industry based on data mining. In: 2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA). pp. 1-6. IEEE (2021)

[6] Chong, A. Y. W., Khaw, K. W., Yeong, W. C., Chuah, W. X.: Customer churn prediction of telecom company using machine learning algorithms. Journal of Soft Computing and Data Mining 4(2). pp. 1-22 (2023)

[7] Noviandy, T. R., Idroes, G. M., Hardi, I., Afjal, M., Ray, S.: A model-agnostic interpretability approach to predicting customer churn in the telecommunications industry. Infolitika Journal of Data Science 2(1). pp. 34-44 (2024)

[8] Senthilselvi, A., Kanishk, V., Vineesh, K., Raj, A. P.: A novel approach to customer churn prediction in telecom. In: 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). pp. 1-7. IEEE (2024)

[9] Poudel, S. S., Pokharel, S., Timilsina, M.: Explaining customer churn prediction in telecom industry using tabular machine learning models. Machine Learning with Applications 17. pp. 100567 (2024)

[10] Abdulsalam, S. O., Arowolo, M. O., Saheed, Y. K., Afolayan, J. O.: Customer churn prediction in telecommunication industry using classification and regression trees and artificial neural network algorithms. Indonesian Journal of Electrical Engineering and Informatics (IJEEI) 10(2). pp. 431-440 (2022)

[11] Chang, V., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., Benson, V.: Prediction of customer churn behavior in the telecommunication industry using machine learning models. Algorithms 17(6). pp. 231 (2024)

[12] Alotaibi, M. Z., Haq, M. A.: Customer churn prediction for telecommunication companies using machine learning and ensemble methods. Engineering, Technology & Applied Science Research 14(3). pp. 14572-14578 (2024).