Customer Churn Prediction for Telecommunication Industry using Extreme Gradient Boosting and Minimum Redundancy and Maximum Relevance

Yi Yang Lim2, a), Thian Song Ong1, 2, b), Kian Ming Lim3, c) and Chin Poo Lee3, d)

*1Centre for Advanced Analytics, COE for Artificial Intelligence , Multimedia University, Jalan Ayer Keroh Lama, Bukit Beruang, 75450 Melaka, Malaysia*

2Faculty of Information Science & Technology, Multimedia University, Jalan Ayer Keroh Lama, Bukit Beruang, 75450 Melaka, Malaysia

3School of Computer Science, University of Nottingham Ningbo China, Zhejiang Province 315100, China

b) Corresponding author: tsong@mmu.edu.my  
a)1181103468@student.mmu.edu.my  
c)Kian-Ming.Lim@nottingham.edu.cn  
d)Chin-Poo.Lee@nottingham.edu.cn

**Abstract.** Customer churn poses significant challenges for the telecommunications industry. To address this issue, a comprehensive churn prediction framework is needed to accurately predict and analyze the reasons behind customer churn. In this study, we propose a customer churn prediction framework that leverages eXtreme Gradient Boosting (XGBoost) in conjunction with the SHapley Additive exPlanations (SHAP). The SHAP model is applied to XGBoost to interpret its prediction output and explain feature contributions. The study utilizes three publicly available Telco datasets, namely Cell2cell, IBM, and BigML for performance evaluation. First, the data preprocessing phase involves categorical value conversion and missing value imputation is carried out. Feature scaling is then performed using RobustScaler to standardize features to a specific range. Dimensionality reduction is achieved using the minimum Redundancy and Maximum Relevance (mRMR) feature selection method. The class imbalance problem is then addressed using the Synthetic Minority Over-sampling Technique (SMOTE). The proposed model predicts customer churn with the hyperparameter tuning performed using Optuna. XGBoost achieved the highest accuracy across all three datasets: 72.90% on Cell2cell dataset, 82.04% on IBM dataset, and 96.10% on BigML dataset. Besides, we also compared the performance with and without feature selection. The results indicates that the mRMR had only a minor impact on the prediction accuracy but effectively reduced the model complexity, while still achieving results comparable to those obtained using the full feature sets.

# Introduction

The telecommunications Industry (TCI) is currently experiencing intense competition and an oversaturated market, resulting in a highly competitive landscape. Consequently, customer churn, defined as the loss of customers due to their departure from a company, has emerged as a significant challenge for telecommunication companies. Customer churn negatively impacts revenue and necessitates prompt attention and effective countermeasures. According to [1], fostering long-term relationships with existing clients is important, as loyal customers serve as both valuable assets and market ambassadors, while [2] stated that customers are prone to purchase additional services and share their positive experiences, thus attracting more customers. Customer churn can be caused by various factors such as high costs, privacy concerns, or dissatisfaction with the service provided.

To address the issues, telecommunications companies are investing in solutions that can predict customer churn and understand the reason for churn, which allows them to invest in targeted retention strategies. The Customer Churn Prediction (CCP) method has proven to be an effective solution. Research shows that adopting CCP techniques can result in significant improvements in earnings and brand recognition. All of this is made possible by the vast amounts of customer data that were gathered through centralized information systems in different industries. Now, researchers are able to build predictive models using more diverse methodologies.

Researchers have utilized various approaches to tackle churn prediction, including the use of traditional models such as Random Forest, Decision Tree, and XGBoost [3-6], and often in combination with ensemble methods to enhance performance [6], [7]. Additionally, deep learning methods have been shown to effectively capture complex, non-linear relationships and extract features more effectively without the need for manual feature selection [8].

# materials and methods

The proposed solution is evaluated on three publicly available datasets, which are the Orange, IBM, and Cell2cell datasets. Table 1 shows the characteristics of each dataset. All three datasets have class imbalance issues, and the Cell2cell dataset contains missing values that require further processing.

**TABLE 1.** Information of the datasets

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Cell2Cell** | **IBM** | **BigML** |
| Total of features | 57 | 20 | 19 |
| Total of customers | 51047 | 7043 | 3333 |
| Missing value | Yes | No | No |
| Churn | 28.8% | 26.5% | 14.5% |
| Not churn | 71.2% | 73.5% | 85.5% |
| Categorical features | 23 | 18 | 3 |
| Numerical features | 35 | 3 | 16 |

Figure 1 presents the overall methodology adopted in this study. The work begins with the data preprocessing as the initial step. Categorical data are transformed into numerical data using an ordinal encoder for further processing. The missing values of each dataset are imputed with the median value to preserve data integrity. The features of the dataset are then scaled using RobustScaler to handle outliers. Next, the mRMR feature selection method is employed to select the most relevant features. Next, the SMOTE is then used to address class imbalance and minimize prediction bias. For each dataset, the training and testing data are split into an 80:20 ratio. The performance of XGBoost, Random Forest, and Logistic Regression is then evaluated and compared. The Optuna hyperparameter tuning method is used on the classifiers to ensure the best hyperparameter values are selected. Lastly, the SHAP model is applied to explain the output of XGBoost, which shows how each feature influences the outcome. Besides, the performance of the mRMR is evaluated by testing it with a different number of selected features.

*A screenshot of a computer

AI-generated content may be incorrect.*

**Figure 1.** The process of the proposed method

# results and discussion

## Comparison of Results with and Without mRMR Feature Selection

To determine the optimal number of features that balance performance and complexity, the number of features in each dataset is gradually reduced. The predictions for this part were carried out using XGBoost, combined with automatic feature selection using Optuna. For this part, only accuracy and F1 score are used to give an overall evaluation. Figures 2, 3, and 4 show the impact of the number of features on accuracy and F1 score.

For the Cell2cell dataset, the accuracy remains consistent with different numbers of features, as shown in Figure 2(a). The highest accuracy score obtained is 72.89%. An attempt was made to use only a single feature, which is CurrentEquipmentDays to predict customer churn. The only feature is selected by the mRMR feature selection, where the K value is set to only one. The accuracy obtained is 69.91%. This could mean that other features could introduce noise or do not add value to the prediction. The F1 score fluctuates more than accuracy, where the scores lower significantly between 36 and 32 features, as shown in Figure 2(b).

|  |  |
| --- | --- |
| *A graph with numbers and a number of features  Description automatically generated* | *A graph of a number of features  Description automatically generated* |
| (a) | (b) |

**Figure 2.** (a) Accuracy based on number of features (b) F1 score based on number of features

For the IBM dataset, the accuracy obtained using all 20 features is 81.68%, and it remained relatively stable down to 6 features, as shown in Figure 3(a). However, it drops to 77.99% when using 4 features and further declines to 65.57% with only one feature. The F1 score, shown in Figure 3(b), follows a similar trend.

|  |  |
| --- | --- |
| *A graph of a number of features  Description automatically generated* | *A green bar graph with white text  Description automatically generated* |
| (a) | (b) |

**Figure 3.** (a) Accuracy based on number of features (b) F1 score based on number of features

As for the BigML dataset, when using all 19 features to predict customer churn, the accuracy obtained is 93.55%. The best results for both metrics were obtained by using 11 features, resulting in 96.10% for the accuracy and 84.94% for the F1 score, as shown in Figure 4(a) and Figure 4(b). There is a significant decline in performance when the number of features is reduced to 5, suggesting that at least 7 features are necessary to maintain optimal performance for this dataset.

|  |  |
| --- | --- |
| *A graph of a number of features  Description automatically generated* | *A graph of a number of features  Description automatically generated* |
| (a) | (b) |

**Figure 4.** (a) Accuracy based on number of features (b) F1 score based on number of features

Overall, across the three datasets used, the mRMR feature selection method does not improve model performance. However, it effectively reduces the model’s complexity and training time while still getting similar results when using the full feature sets.

## Comparison of Results Using Different Classifiers

Based on the feature selection test, the most suitable number of features for prediction was selected. The performance of Logistic Regression and Random Forest was also evaluated and compared alongside XGBoost. The previous process remains the same, and the hyperparameter tuning for each method is conducted using Optuna. Tables 2, 3, and 4 present the prediction output of the models. Overall, XGBoost obtained better results across different metrics most of the time, followed by Random Forest and Logistic Regression.

**TABLE 2.** Cell2cell dataset (44 features)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **F1 score** |
| Logistic Regression | 70.01 | 42.37 | 15.23 | 22.40 |
| Random Forest | 72.27 | 56.19 | 11.09 | 18.53 |
| XGBoost | **72.90** | **58.01** | **16.95** | **26.24** |

**TABLE 3.** IBM dataset (18 features)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **F1 score** |
| Logistic Regression | 76.65 | **67.74** | 22.52 | 33.80 |
| Random Forest | 78.63 | 60.52 | 55.49 | 57.90 |
| XGBoost | **82.04** | 66.94 | **63.53** | **65.19** |

**TABLE 4.** BigML Dataset (11 features)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **F1 score** |
| Logistic Regression | 80.95 | 40.00 | 67.36 | 50.19 |
| Random Forest | 94.75 | 80.61 | **83.15** | 81.86 |
| XGBoost | **96.10** | **90.36** | 78.94 | **84.26** |

## SHAP Summary Plot

The SHAP summary plot for the IBM dataset is presented in Figure 5. The y-axis shows the most important features from the top to the bottom. The x-axis of the plot shows how much the impact of each feature on the prediction. The contract is the most important feature. Customers who subscribe to a two-year contract (represented by red color) and a one-year contract (purple color) are more likely to continue using the services. While customers with a month-to-month contract are more likely to churn. The second most important feature is tenure. Customers with a longer tenure (red) are more likely to continue using the services, whereas those with shorter tenures are more prone to churn. Customers who subscribe to additional services such as TechSupport, OnlineSecurity, and OnlineBackup also tend to be more loyal to the company. Conversely, we notice that customers with higher monthly charges are positively correlated to customer churn, as well as customers who have multiple lines.

A graph of different colored lines

AI-generated content may be incorrect.

**Figure 5.** SHAP summary plot of IBM dataset

Figure 6 presents the SHAP summary plot of the Cell2cell dataset. The most influential feature is CurrentEquipmentDays, which represents the number of days the handset is used. The analysis suggests that a longer usage period is positively associated with a high likelihood of churn. This could be attributed to customer considering device upgrades or switching service providers after extended use. The second important feature, MonthsInService, appears to have a mixed impact on the prediction. The feature PercChangeMinutes refers to the percentage change in the customer's usage. The high positive values, which show the increase in minutes usage, suggest that the customer is more engaged with the service and less likely to churn. Besides, customers with higher monthly minutes, credit ratings, and total recurring charges are also more loyal to the company. On the contrary, customers who have had their handsets refurbished and have contacted the retention team exhibit a higher likelihood of churning.

A graph of a number of different colored lines

Description automatically generated with medium confidence

**Figure 6.** SHAP summary plot of Cell2cell dataset

Figure 7 presents the SHAP summary plot of the BigML dataset. Features such as ‘Total intl calls’ and ‘Voice mail plan’ are negatively correlated with churn, which means that as the usage of these features increases, the likelihood of churn decreases. ‘Total day minutes’ and ‘Customer service calls’ are the most important features, but they seem to have a mixed impact on the model’s output, where higher values can contribute positively or negatively to churn. Features such as ‘international plan’, ‘Total day charge’, and ‘total eve minutes’ have a positive correlation with customer churn.

# conclusion

This research enables telecommunications companies to identify potential customer churn and its underlying causes. With this information, they can implement personalized retention strategies specifically for customers who are at a higher risk of churning. Moving on to the future research direction, prediction of customer churn by using deep learning methods would be the most potential direction. As the data is too specific for each dataset, deep learning models can capture all the complex relationships and patterns in the data without manual feature engineering. Besides, they are also better at handling high-dimensional data with numerous features, making them a promising approach for improving the performance of customer churn predictions.

A graph of a number of people

Description automatically generated with medium confidence

**Figure 7.** SHAP Summary Plot of BigML Dataset

# references

1. 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 **94**, 290–301 (2019).
2. M. Óskarsdóttir, C. Bravo, W. Verbeke, C. Sarraute, B. Baesens, and J. Vanthienen, “Social network analytics for churn prediction in telco: Model building, evaluation and network architecture,” Expert Systems with Applications **85**, 204–220 (2017).
3. R. Nkolele, and H. Wang, “Explainable Machine Learning: A Manuscript on the Customer Churn in the Telecommunications Industry,” in *2021 Ethics and Explainability for Responsible Data Science (EE-RDS)*, (Institute of Electrical and Electronics Engineers, Johannesburg, South Africa, 2021), pp. 1–7.
4. S.M. Sina Mirabdolbaghi, and B. Amiri, “Model Optimization Analysis of Customer Churn Prediction Using Machine Learning Algorithms with Focus on Feature Reductions,” Discrete Dynamics in Nature and Society **2022**(1), 5134356 (2022).
5. A. Idris, A. Khan, and Y.S. Lee, “Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification,” Applied Intelligence **39**(3), 659–672 (2013).
6. I. Ullah, B. Raza, A.K. Malik, M. Imran, S.U. Islam, and S.W. Kim, “A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector,” Institute of Electrical and Electronics Engineers Access **7**, 60134–60149 (2019).
7. S. Fakhar Bilal, A. Ali Almazroi, S. Bashir, F. Hassan Khan, and A. Ali Almazroi, “An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry,” PeerJ Computer Science **8**, e854 (2022).
8. L. Elvitaria, E.F. Ahmad Shaubari, N.A. Samsudin, S.K. Ahmad Khalid, S. Salamun, I.P. Sari, Z. Indra, and R. Rudiansyah, “A Data-Driven Approach for Batik Pattern Classification Using Convolutional Neural Networks (CNN),” Semarak International Journal of Electronic System Engineering **4**(1), 22–30 (2024).