Proceedings of ICFTBA 2026 Symposium: Strategic Management & Business Analytics: Data-Driven Performance Models
DOI: 10.54254/2754-1169/2026.LD37687

Optimization of an Explainable Machine Learning Model for Customer Churn Prediction and Business Retention Strategy

Zixi Zhang

Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, China
v530026292@mail.bnbu.edu.cn

Abstract. Customer churn forecasting is critical for companies that want to identify potential churn customers and develop more targeted customer maintenance strategies. This study explores whether Bayesian optimization can improve the performance of machine learning in customer churn prediction, and whether explainable model output can provide useful business insights. This research used Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) to analyze the IBM Telecom customer churn data set containing 7,043 customer records. We applied Bayesian hyperparameter optimization to XGBoost by using five-fold hierarchical cross-validation, and used SHapley Additive exPlanations (SHAP) to explain the optimized model. The optimized XGBoost model achieves the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.8424 and the Precision-Recall Area Under the Curve (PR-AUC) of 0.6518, and the recall rate also increased from 0.6408 in the baseline XGBoost model to 0.7721. SHAP analysis shows that the type of contract, user duration and Internet services are the most influential predictors. Risk segmentation further shows that there are significant differences in the churn rates of different customer groups. Research results show that combining model optimization with interpretable analysis helps to make more transparent and targeted customer retention decisions.

Keywords: Customer Churn Prediction, XGBoost, Bayesian Optimization, SHAP

1. Introduction

Customer churn is a major challenge for service-oriented enterprises, because losing existing customers will reduce recurring income and increase the demand for new customers. Therefore, machine learning has been increasingly used to identify high-risk customers. However, predictive performance alone is not enough for effective customer retention management. There may be imbalances in the lost data set, the performance of the model may be highly dependent on the selection of hyperparameters, and complex models may not fully explain why some customers are classified as high-risk customers.

This study explores an integrated method that combines model optimization, interpretable machine learning and business-oriented customer analysis. Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) are first compared as baseline models. Bayesian optimization is applied to XGBoost to test whether

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.LD37687

systematic hyperparameter tuning can improve the accuracy of customer churn prediction. Subsequently, SHapley Additive exPlanations (SHAP) analysis is used to identify the characteristics that contribute the most to the forecast, and the predicted customer churn probability is used to divide customers into different risk groups, so as to support customer retention-oriented decision-making.

Accordingly, this study aims to answer three questions: whether model optimization can improve the performance of customer churn forecasting; which factors have the greatest impact on customer churn forecasting according to explainable machine learning; and how to use optimized forecast results and explainable insights to guide customer retention strategies. This study provides a theoretical basis for the formulation of customer retention strategies by evaluating the impact of optimization on multiple performance indicators, using the SHAP method to identify the prediction drivers, and linking risk scoring and explanation with customer retention-oriented customer segmentation. Therefore, this study can integrate predictive optimization, interpretability and business decision-making support into an empirical framework.

2. Literature review

Customer churn prediction is a mature machine learning application, because identifying customers who may be lost helps to formulate more effective customer retention strategies. One of the common challenges is category imbalance, which distorts model training and makes overall accuracy no longer a comprehensive indicator of performance. Burez and Van den Poel pointed out that loss forecasting in the case of class imbalance needs to be handled carefully and evaluated using indicators such as Receiver Operating Characteristic Area Under the Curve (ROC-AUC) [1]. This is especially important when the observation value of the loss category is relatively small, for example, in many customer retention data sets.

Ensemble methods are widely used in complex classification tasks. XGBoost is a scalable tree-eleveing system designed to achieve efficient predictive modeling [2]. However, its performance depends on hyperparametric configuration. Bayesian optimization uses previously evaluated information to guide the search, so as to find promising parameter settings, providing a systematic alternative to manual or exhaustive tuning [3]. These studies provide a methodological basis for testing whether system optimization can improve the performance of XGBoost in customer churn forecasting. In this study, optimization is regarded as a model development process rather than a new algorithm contribution.

Predictive performance alone cannot fully meet the needs of business decision-making. Complex models may be accurate, but difficult to explain. SHAP provides a unified method for assigning the contribution of the feature level to the prediction results of each model [4]. Recent research that can explain loss combines machine learning, SHAP-based feature analysis, customer segmentation, and customer retention-oriented recommendations [5]. These studies also show that contract terms, service terms, fees, payment methods and service-related variables are crucial to explain the predicted risk of loss [5]. Therefore, explainability not only helps to sort variables globally, but also helps to find out why specific customers receive high-risk predictions.

Recent research has also gone beyond classification and turned to commercial value. Jiang et al. combined customer churn forecasting with interpretation and profitability considerations based on the Shapley model [6]. At the same time, Devriendt and others emphasized that identifying customers who may be lost is not equivalent to identifying customers who will respond positively to customer retention interventions [7]. This difference suggests a valuable research direction: predictive optimization and interpretability can support customer retention decision-making, but

DOI: 10.54254/2754-1169/2026.LD37687

prediction-based recommendations should not be regarded as experimentally verified interventions. This study follows this logic and integrates Bayesian optimization based on XGBoost, interpretation based on SHAP model, customer risk segmentation and decision-making support for customer retention, while clearly distinguishing between predictive evidence and causality.

3. Data source and dataset characteristics

This study uses the IBM Telco Customer Churn dataset from Kaggle [8]. The dataset contains 7,043 telecom customers, initially containing 21 variables, covering customer characteristics, service subscriptions, account information, billing conditions and loss status. The reason for choosing this dataset is that it contains behavioral information and contract information related to supervised churn forecasting and customer retention analysis.

The target variable, *Churn*, is a binary variable. Among the 7,043 observations, 5,174 customers were not lost, and 1,869 customers were lost, accounting for about 73.5% and 26.5% of the data set respectively. This distribution shows a moderate degree of class imbalance. No duplicate records were found during the preprocessing process.

Predictive variables cover multiple dimensions of customer relationships. Demographic variables include gender, whether you are elderly, whether you are a partner, and whether you have dependents. Account-related variables include account duration, contract type, whether to use electronic billing, payment method, monthly fee and total cost. Service-related variables describe telephone, Internet, network security, technical support, device protection, streaming media and other telecommunications services. The customer identifier has been excluded from the prediction model because it is only used as a record identifier.

4. Methodology

4.1. Data preprocessing

The dataset was cleaned and transformed in R using RStudio. *TotalCharges* contains 11 missing values, all of which correspond to customers with zero-month service time. Since these customers are new customers and have no cumulative costs, replace the missing values with zero instead of deleting these observations. The customer identifier is removed, the classification variables are converted into factors, and the prediction variables are uniquely coded.

Using the 80:20 layered sampling method, 5,636 training samples and 1,407 test samples were generated while maintaining the distribution of customer churn. The training set contains 4,140 non-lost users and 1,496 lost users. After sampling, training-related preprocessing was carried out to avoid information leakage. Standardization is only applicable to logistic regression models, while tree-based models use unscaled coding prediction variables.

In order to solve the problem of moderate category imbalance, we apply category weighting in the model training process instead of generating synthetic observation data. The positive weight is only calculated according to the training set. The formula is 4140/1496, and the result is about 2.77. The weight of non-loss cases is 1. The test set remains unchanged and is not weighted.

4.2. Baseline model development and evaluation

We have developed four baseline models: Logistic Regression, Random Forest, XGBoost, and LightGBM. XGBoost is a scalable tree-boosting framework [2]. These models are selected to

DOI: 10.54254/2754-1169/2026.LD37687

compare traditional statistical classifiers, bagging-based integration methods, and two gradient enhancement methods. Category weights are introduced during the training process, so as to give a small number of lost user categories a higher weight without changing the test distribution.

All models use the same training set-test set division. The performance evaluation indicators of the test set include accuracy rate, precision rate, recall rate, F1 score, Receiver Operating Characteristic Area Under the Curve (ROC-AUC) and Precision-Recall Area Under the Curve (PR-AUC). The threshold value of the threshold-related indicators is set to 0.50. ROC-AUC is the main optimization indicator because it evaluates the ranking performance under different thresholds; at the same time, the recall rate and PR-AUC are also valued, because identifying customers with loss risk is the key to retention screening.

4.3. Bayesian optimization of XGBoost

XGBoost further adopts Bayesian hyperparameter optimization for tuning, which uses the previous evaluation results to find a better configuration [3]. The optimization is only carried out on the training data, and the five-fold layered cross-verification is adopted, aiming at the average ROC-AUC of five verification folds. The test set is not used in the tuning process.

Six hyperparameters were optimized: boosting rounds, maximum tree depth, learning rate, subsampling ratio, feature subsampling ratio, and minimum child weight. The first search used a wide range of parameters: the number of boost wheels 100-800, tree depth 3-10, learning rate 0.01-0.30, subsampling rate 0.60-1.00, characteristic subsampling rate 0.60-1.00, minimum child weight 1-10. Because several optimal values appeared near the initial range, the second fine search was carried out, using 500-1500 rounds of lifting rounds, tree depth 2-6, learning rate 0.005-0.08, sub-sampling rate 0.50-0.85, and characteristic sub-sampling rate 0.50-0.85; the minimum child weight is 1-6. Each search uses 10 initial points and 20 additional Bayesian iterations.

The final configuration uses 500 rounds of Boosting iteration, with a maximum depth of 2, a learning rate of 0.02312649, a sub-sampling rate of 0.50, a characteristic sub-sampling rate of 0.5733002, and a minimum sub-node weight of 2. It achieves an average ROC-AUC of 50% cross-verification of 0.8514524. Then, the optimized model is retrained on the complete training set and evaluated on the unmodified test set.

4.4. Explainability and business-oriented analysis

In order to quantify the contribution of the feature level to customer churn forecasting, we calculated the SHAP value of the optimized XGBoost model [4]. For classification variables represented by multiple virtual variables, we add the SHAP contribution value of each observation to obtain its corresponding original characteristic value. Then, we calculate the global importance as the absolute average of the group SHAP contribution value. In addition, we also conducted a SHAP analysis of the direction and size of the contribution of different customers.

Customers in the test concentration are divided into three rule-based risk groups according to the predicted loss probability: less than 0.30 is a low-risk group, 0.30 to less than 0.60 is a medium-risk group, and 0.60 or more is a high-risk group. These thresholds are used for descriptive subdivision and are not statistically optimized.

For each customer, we regard the maximum positive group SHAP contribution value as the main predictive risk driver. These drivers are mapped to a wide range of customer retention areas, including new customer guidance and early life cycle interactions, contract flexibility and loyalty incentives, service quality and customer support, billing and payment experiences, and price-related

DOI: 10.54254/2754-1169/2026.LD37687

customer value. The analysis aims to provide a basis for retention recommendations, rather than claiming a causal relationship or experimentally verified intervention results.

5. Results

5.1. Predictive performance and XGBoost optimization

Table 1 compares the performance of five evaluation models on the reserved test set. In the baseline model, the random forest model achieved the highest accuracy rate (0.7761) and accuracy rate (0.5642); the logistic regression model achieved the highest recall rate (0.7748) and the area under the ROC curve (0.8416); and the LightGBM model achieved the highest F1- score (0.6294).

Table 1. Performance comparison of customer churn prediction models

Model	Accuracy	Precision	Recall	F1	ROC_AUC	PR_AUC
Logistic Regression	0.7470	0.5152	0.7748	0.6188	0.8416	0.6513
Random Forest	0.7761	0.5642	0.6836	0.6182	0.8373	0.6316
XGBoost	0.7605	0.5407	0.6408	0.5865	0.8195	0.6016
LightGBM	0.7690	0.5476	0.7399	0.6294	0.8311	0.6270
Optimized XGBoost	0.7541	0.5246	0.7721	0.6247	0.8424	0.6518

Bayesian optimization has improved many indicators of XGBoost. ROC-AUC increased from 0.8195 to 0.8424, PR-AUC increased from 0.6016 to 0.6518, the recall rate increased from 0.6408 to 0.7721, and the F1 score increased from 0.5865 to 0.6247. The accuracy rate decreased slightly, from 0.7605 to 0.7541, while the accuracy rate decreased from 0.5407 to 0.5246. Figure 1 summarizes these changes. Among the five models, the optimized XGBoost model has achieved the highest ROC-AUC and PR-AUC, but its advantage over logistic regression is not significant.

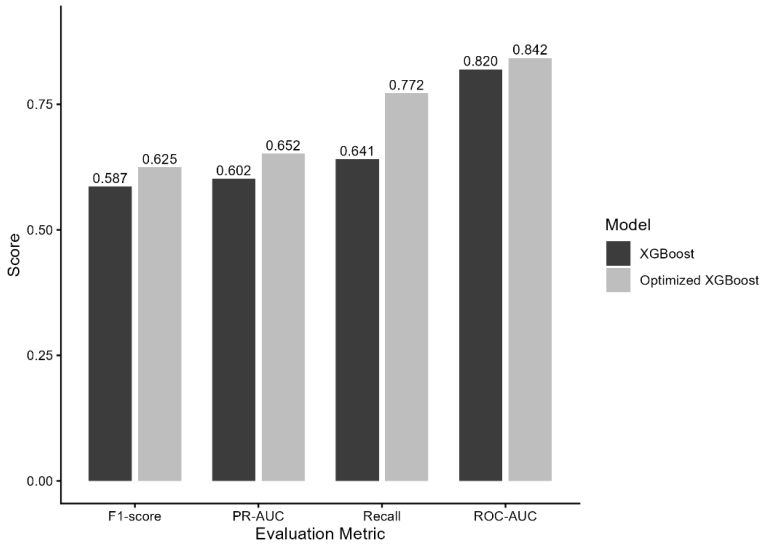

Figure 1. Performance comparison between baseline and optimized XGBoost

DOI: 10.54254/2754-1169/2026.LD37687

5.2. Explainable analysis of churn predictions

SHAP analysis was applied to the optimized XGBoost model. Figure 2 shows the ten most important original features based on mean absolute aggregated SHAP values. Contract ranked first at 0.563, followed by tenure at 0.520 and InternetService at 0.455. PaymentMethod ranked fourth at 0.204, followed by TotalCharges (0.188) and MonthlyCharges (0.184). PaperlessBilling, OnlineSecurity, MultipleLines, and StreamingTV were also among the top ten features.

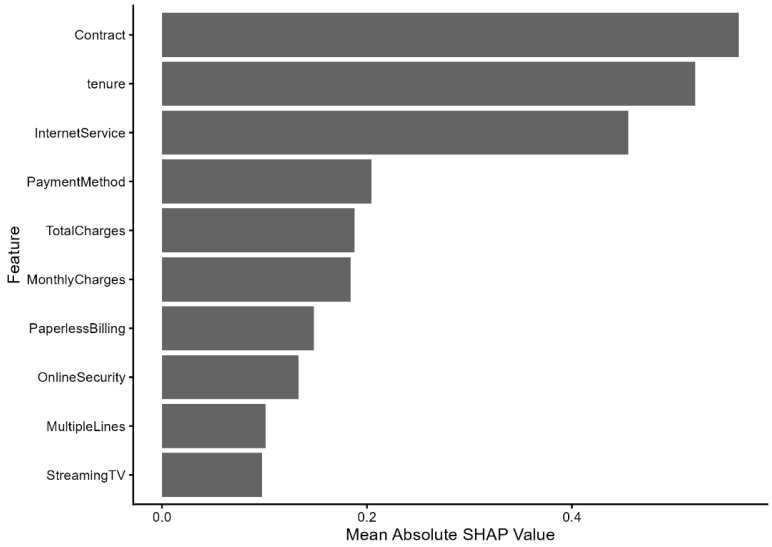

Figure 2. Global feature importance based on mean absolute SHAP values

These results show that contractual conditions, customer relationship duration, service configuration, payment arrangements, and charges contributed strongly to model predictions. Figure 3 provides a complementary SHAP summary view showing the distribution, direction, and magnitude of contributions across the most influential encoded predictors.

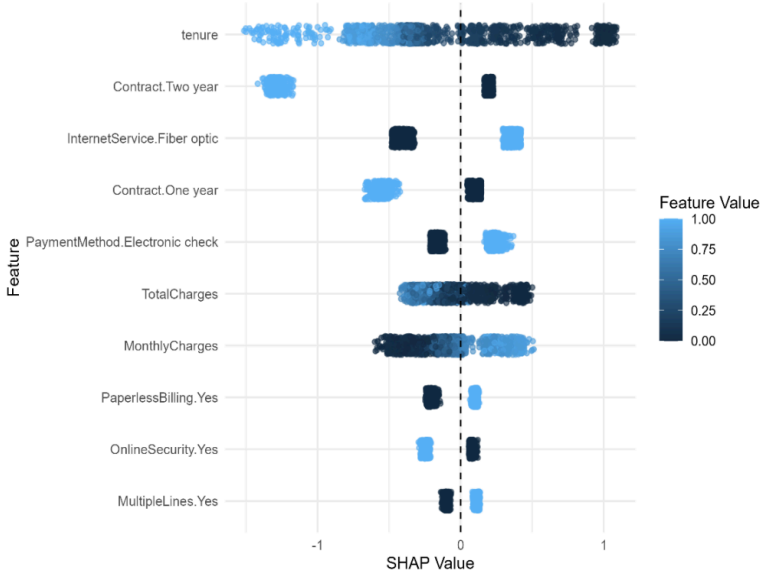

Figure 3. Distribution of SHAP values for the most influential predictors

DOI: 10.54254/2754-1169/2026.LD37687

5.3. Customer risk segmentation

According to the predicted loss probability, the test set customers are divided into low-risk, medium-risk and high-risk groups. The high-risk group contains 430 customers, accounting for about 30.6% of the test sample. The average predicted loss probability is 0.776, and the actual loss rate is 59.3%. The medium-risk group contains 334 customers, accounting for about 23.7%. The average predicted loss probability is 0.456, and the actual loss rate is 23.6%. The low-risk group contains 643 customers, accounting for about 45.7%. The average predicted loss probability is 0.110, and the actual loss rate is 6.1%. Table 2 shows the customer distribution, mean predicted churn probability, and observed churn rate across the three risk groups.

Table 2. Customer risk segmentation and observed churn rates

RiskLevel	Customers	CustomerPercentage	MeanPredictedProbability	ActualChurnPercentage
High Risk	430	30.6	0.7762	59.3
Medium Risk	334	23.7	0.4558	23.6
Low Risk	643	45.7	0.1105	6.1

The observed customer churn rate continued to decline from the high-risk group to the medium-risk group to the low-risk group. In the high-risk group, user duration, contract type and Internet service type are the strongest positive influencing factors for predicting the risk of customer churn. The results of customer segmentation and SHAP analysis together provide an empirical basis for customer retention-oriented discussions.

6. Discussion

6.1. Effect of model optimization

Bayesian optimization mainly improves the performance of XGBoost by improving its ability to identify lost users. The recall rate has increased significantly, and the F1 score, ROC-AUC and PR-AUC have also improved. This result is consistent with the optimization goal, because the average ROC-AUC value is maximized during cross-verification [3]. The simultaneous decline in accuracy and precision shows that optimization has not improved all indicators. However, in the case of unbalanced loss of users, the overall accuracy rate alone may be misleading [1]. A higher recall rate is conducive to retention screening because it can identify more real lost users, but a lower accuracy rate means that more non-lost users may also be marked as potential lost users.

The optimized model is slightly better than logistic regression in terms of ROC-AUC and PR-AUC, and its recall rate is close to the results of logistic regression. Therefore, rather than optimizing and improving all indicators, it improves the overall ranking ability and minority detection ability of XGBoost.

6.2. Interpretation of churn predictions

SHAP analysis results show that contracts, service terms and InternetService are the three most influential characteristics, followed by PaymentMethod, TotalCharges and MonthlyCharges. These findings are basically consistent with the results of a recent explainable loss study, which also emphasizes contract-related variables, service periods, fees and payment arrangements [5].

DOI: 10.54254/2754-1169/2026.LD37687

This similarity supports the practical application value of these prediction models. However, the SHAP value should not be interpreted as evidence of causality. They quantify the contribution of characteristics to model prediction [4], and may reflect the correlation and other patterns in the training data. Therefore, it is better to interpret these results as predictive explanations to guide further business investigations. In fact, global SHAP importance can identify recurring predictive patterns in the customer base, and customer-level explanations help clarify why specific customers get high-risk scores.

6.3. Business implications and research boundaries

Risk segmentation shows an obvious gradient in the observed customer churn rate: the high-risk customer churn rate is 59.3%, the medium-risk customer churn rate is 23.6%, and the low-risk customer churn rate is 6.1%. This difference shows that the prediction probability can support the prioritization of customers for customer retention assessment. SHAP explanations can explain why some customers are predicted to be high-risk customers, helping to match the customer retention focus with the customer's life cycle stage, contract, service experience, payment process, and perceived value. For example, when customer duration is the main risk driver, early life cycle interaction is crucial; while when contract or service is the main factor in forecasting, an assessment centered on these variables may be more appropriate.

This method is consistent with the research results that link customer churn forecasting with explainable and economically significant decision-making support [6]. However, the proposed customer retention strategy is still a model-based suggestion, not a verified intervention. Predicting which customers may be lost is different from identifying which customers will respond positively to specific retention measures. The enhanced modeling study emphasizes this important difference [7]. Similarly, rule-based risk thresholds should be interpreted as descriptive subdivision, not the best business critical point. Future research should verify customer retention interventions and, under appropriate circumstances, use experimental, causal, profit-sensitive or promotion-based methods to optimize decision-making thresholds.

7. Conclusion

This study explores the optimization of machine learning in customer churn forecasting and its potential contribution to enterprise customer maintenance strategy. Before applying Bayesian hyperparameter optimization to the XGBoost model, the study compared four benchmark models. Compared with the benchmark model, the recall rate, F1 score, ROC-AUC and PR-AUC of the optimized XGBoost model have improved, but the accuracy rate and precision rate have decreased slightly. These results show that optimization can improve the ability to identify and sort customers at risk of loss, but it needs to be weighed against various evaluation indicators.

SHAP analysis shows that contracts, service duration and Internet services are the most important factors affecting customer churn forecasts, followed by payment and cost-related variables. The prediction probability also clearly shows the difference in the actual turnover rate of high-, medium-, and low- risk groups, indicating that model-based risk segmentation has practical value in customer maintenance management.

There are some limitations in this research. It only uses a single telecommunications data set, and the risk threshold is based on rules rather than statistically optimized. In addition, the proposed retention strategy is not based on intervention evidence, but on prediction and explanatory output.

DOI: 10.54254/2754-1169/2026.LD37687

Future research can test more data sets, optimize risk thresholds, and use methods such as causality, experiments or enhancement effects to evaluate retention measures.

References

[1] Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3), 4626–4636. https://doi.org/10.1016/j.eswa.2008.05.027

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

[3] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems (Volume 25)* (pp. 2951–2959).

[4] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (Volume 30)* (pp. 4765–4774).

[5] El Attar, A., & El-Hajj, M. (2026). Explainable AI-driven customer churn prediction: A multi-model ensemble approach with SHAP-based feature analysis. *Frontiers in Artificial Intelligence, 9*, Art. no. 1748799. doi: 10.3389/frai.2026.1748799.

[6] Jiang, P., Liu, Z., Abedin, M. Z., Wang, J., Yang, W., & Dong, Q. (2024). Profit-driven weighted classifier with interpretable ability for customer churn prediction. *Omega, 125*, Art. no. 103034. doi: 10.1016/j.omega.2024.103034.

[7] Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. *Information Sciences, 548*, 497–515. doi: 10.1016/j.ins.2019.12.075.

[8] BlastChar. (2026, August 20). Telco customer churn. *Kaggle*. dataset, based on IBM Sample Data Sets.