International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 9, Issue 3
DOI: https://doi.org/10.15157/IJITIS.2026.9.3.1709-1739
Received: 23.04.2026; Revised: 01.07.2026; Accepted: 03.08.2026

Research Article

From Prediction to Action: Explainable Churn Modelling with Calibration and Threshold-Aware Decision Support

Frida Zisko[1]*, Erarda Vuka[2]

[1]Department of Business Informatics, Mediterranean University of Albania, Tirana, Albania
[2]Department of Information Technology, Mediterranean University of Albania, Tirana, Albania
***frida.zisko@umsh.edu.al**

Abstract

The problem of predicting customer churn plays an important role in the field of telecommunication business decisions. However, the existing literature mainly focuses on achieving better classification accuracy without paying enough attention to calibration, interpretability, and the business value of threshold choice. The goal of this study is to introduce a churn prediction framework that incorporates optimization of the classifier, its calibration, explainability, and business-oriented choice of the threshold. The open-source Telco Customer Churn dataset was used to compare four models like Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). The process of hyperparameter tuning is carried out by means of randomized search and stratified cross-validation. Further experiments include repeated stratified cross-validation and hold-out validation of the models' results. Random Forest showed the best performance in terms of mean cross-validation, while LightGBM demonstrated the best F1-score and Matthew's correlation coefficient on the hold-out set. Due to the lack of statistical significance of the difference between these algorithms, LightGBM was chosen for further use. Isotonic calibration decreases the Brier score from 0.1609 to 0.1371. The threshold optimization shows that choosing the decision boundary equal to 0.30 improves the balance between precision and recall compared to the usual 0.50. Furthermore, the cost-sensitive analysis demonstrates that the optimal value of thresholds decreases as the cost of missing out on a true churner rises. SHapley Additive exPlanations (SHAP) values and permutation-based feature importance were calculated to identify the four most important predictors of churn, including contract type, customer tenure, internet service type, and billing variables. In this research, a decision support framework is developed by calibrating probabilities and tuning decision thresholds.

Keywords: Customer Churn Prediction; Telecom Analytics; LightGBM; SHAP; Cost-Sensitive Thresholding; Decision Support.

INTRODUCTION

Customer churn prediction is one of the main goals in the telecommunication sector because keeping existing customers is less expensive than gaining new ones. In

competitive markets, small gains in churners' identification could have a significant impact on revenue, customer lifetime value, and efficiency of churn prevention activities. Machine learning methods are now irreplaceable tools for solving churn problem by finding potential churners among customers with different features: demographics, contracts, billing and service usage [1-5]. However, two aspects are still missing in practice. Firstly, some papers focus only on ranking of models or their overall features importance without paying much attention to the quality of predicted probabilities that are very important for taking the decision [6, 7]. Second, threshold selection is rarely linked explicitly to business costs and is not always separated from final hold-out evaluation, which can blur the distinction between predictive assessment and deployment policy [8-14]. This analysis considers the following research questions:

- *RQ1:* What is the performance comparison between Logistic Regression, Random Forest, XGBoost and LightGBM regarding their F1-score, recall, Area Under the Receiver Operating Characteristic Curve (ROC-AUC) and Area Under the Precision–Recall Curve (PR-AUC) using repeated stratified cross-validation?
- *RQ2:* How much do Platt (sigmoid) and isotonic calibration improve the reliability of predicted probabilities for the selected model?
- *RQ3:* How do calibrated thresholds and asymmetric false-negative vs. false-positive cost assumptions impact churn detection rates, false alarms, and expected decision costs?
- *RQ4:* To what extent do global SHAP values and permutation importance agree in identifying the top feature groups linked to churn?

These lead to four testable hypotheses:

- *H1:* The benchmarked models differ in their balance between churn capture and false-positive control, with ensemble methods expected to provide competitive F1-score and MCC performance relative to Logistic Regression.
- *H2:* Consistent with probability-calibration theory, post hoc calibration should lower the Brier score by at least 10% without a negative impact on ROC-AUC of more than 0.01.
- *H3:* Consistent with cost-sensitive decision theory, applying a calibrated cut-off point that is less than 0.50 is expected to produce a more desirable trade-off between churn detection and decision cost than the default cut-off point for the asymmetrical cost conditions studied.
- *H4:* Consistent with explanation-consistency theory, SHAP and permutation importance methods should be able to identify at least three shared groups of features in the top five churn predictors.

Each hypothesis corresponds to a distinct stage of the empirical workflow: H1 concerns predictive benchmarking, H2 probability reliability, H3 operational threshold choice and H4 explanation consistency. The hypotheses are revisited explicitly in the Results and Discussion sections.

The hypotheses are based on different theoretical predictions [14-20]. Hypothesis 1 is founded on the ability of ensemble models to account for non-linear patterns and inter-feature dependencies that might not be sufficiently captured by linear decision boundaries [14–17, 21-25]. Hypothesis 2 relies on the calibration theory that views the reliability of probabilities as an attribute independent of ranking accuracy and, therefore, predicts improvement of Brier score and calibration without significant reduction in ROC-AUC values [6, 7]. Hypothesis 3 is grounded in cost-sensitive decision theory, which predicts that the preferred threshold changes when false negatives and false positives have unequal operational consequences [8, 10]. Finally, Hypothesis 4 reflects the findings of explainability studies indicating that consistency across different attribution techniques increases confidence in the importance of features, albeit maintaining association without causation [12, 13, 19–21].

The research gap addressed in this study is not explained by the fact that there are no accurate churn classifiers. There is a substantial body of literature that already successfully applies ensembles of models to solve churn prediction problems on structured telecom datasets. The key practical challenge in building a predictive model that would directly inform business decisions is differentiating between predictions that represent a reliable signal for decision makers and those that do not. Namely, three aspects are often conflated when evaluating models for their ability to inform intervention decisions: whether the model's predicted probability reflects the reality of churn risk; whether the model's explanation is not reduced to a single feature importance plot; and whether the models predicted probability threshold appropriately accounts for the business's different costs associated with missed and retained churners. A model that performs well on predictive accuracy and interpretability benchmarks but fails to account for any of these issues may be misleading or uninformative to a human analyst that seeks to use it in practice. The current study addresses this important gap by evaluating discrimination, probability calibration, and explanation consistency within the same leakage-aware analytical framework.

The study does not aim to propose any algorithm novelty. The authors' contribution is a decision-oriented framework for churn-model evaluation that untangles four questions that are often conflated in practice: what model performs best in terms of discriminatory power, if its outputs are calibrated, if it is amenable to consistent interpretation, and finally, what response actions should be taken on its basis given different error priorities. Their framework, therefore, emphasizes evaluation logic rather than an innovative model. A combination of optimized benchmarking practices, calibration checks, complementary explanation approaches, uncertainty analyses, and cost-sensitivity assessments allows the authors to demonstrate how these questions should ideally be interrelated to evaluate if a churn model is fit for purpose in customer-priority allocation rather than just being generally accurate in predictions.

RELATED WORK

Machine Learning Approaches for Telecom Churn Prediction

Predictive modelling of churn in telecommunication services has received extensive attention, whereby earlier works relied on explainable statistical and machine learning algorithms such as Logistic Regression, Decision Trees, and Survival Analysis [3–5]. As consumer behaviour patterns have become increasingly complex and nonlinear, ensemble learning methods, such as Random Forest and Gradient Boosting, have gained popularity because of their ability to capture complex feature interactions, model nonlinear relationships, and achieve strong predictive performance without relying on restrictive statistical assumptions. Recent studies have also looked into XGBoost, LightGBM, ensemble learning algorithms, as well as class imbalance issues in structured telecom data [14–17]. Additionally, neural network models have also been investigated; however, tree-based models remain highly competitive on low-dimensional structured churn datasets [14–15]. Despite an increase in accuracy of predictions, limitations in implementation of complex models can arise when the output lacks explainability and is not evaluated based on deployability metrics [12, 13]. In order to fill this void, the current study increases the benchmarking framework to include a modern LightGBM algorithm together with the Logistic Regression, Random Forest, and XGBoost [14–17].

Cost-sensitive Decision-Making and Threshold Selection in Churn Management

In the area of churn management operations, classification errors have the following asymmetry such as missing a true churner means losing customer value, while contacting a false churner involves the expenses of the campaign [8, 9]. Thus, in assessing the effectiveness of the churn model, one must not only consider discrimination measures but also the translation of predictions into decisions. Thresholding based on costs implies setting a decision threshold by taking into account the unequal costs of different errors rather than using the usual threshold of 0.50 [10]. The research on churn and uplift modelling motivated by profit maximization also points out that the choice of model and intervention policy needs to match business value and limitations of the campaign [8, 9, 18].

Explainable Artificial Intelligence and Model Interpretability

The concept of Explainable Artificial Intelligence (XAI) is introduced to solve interpretability problems arising from complicated machine-learning algorithms. One of the most popular classification methods divides explainable models into intrinsically interpretable models and post-hoc approaches that allow explaining previously created black-box predictors [12, 19]. Among such post-hoc methods, LIME (Local Interpretable Model-agnostic Explanations) and SHAP have received great attention recently. LIME provides local approximation of a predictor by an interpretable model for each data instance [20]. SHAP makes use of Shapley-value concepts to create additive feature attributions. SHAP allows performing both global and local interpretation of tree-based models [13]. However, stability problem, correlated features and dependencies between

features stay relevant when explaining SHAP values [21]. The analysis of explanations of the factors contributing to telecom churn often shows that tenure, contract type, monthly charges, internet service and payment method are important churn drivers [1, 2]. Yet global importance cannot capture individual differences of customers, and explanation outputs are not directly connected to specific measures of retention [1, 2].

Theoretical Logic of Decision-Oriented Churn Modelling

The theoretical foundation of this research uses the approach which distinguishes between prediction, probability estimation, explanation and decision action. Discriminatory metrics estimate how well a model ranks those customers who will leave from those who won't; however, they cannot answer the question of how a specific probability, such as 0.70, is related to empirical probability of 70% that the customer churns. Calibration answers the latter question by estimating how well the estimated probability is reliable for stratifying risks. On the other hand, the explainability concerns a totally different challenge: it reveals what specific customer properties affect the model predictions and whether they can be explained in terms of meaningful customer retention policy. Finally, threshold optimization translates calibrated probability estimates into decision actions taking into account the cost of incorrect classification of the true churner and non-churner. These four dimensions are analytically distinct but operationally interdependent. One model can have very good discriminatory power but poor calibration, good calibration but poor explainability or good accuracy but poor economic efficiency at the selected threshold. A decision-oriented churn system should be evaluated across all four dimensions rather than judged through a single predictive metric.

The proposed framework encompasses four interrelated theoretical strands identified in the literature. First, statistical learning theory differentiates between ranking capability and probability estimation: ROC-AUC and PR-AUC measure the order of customers according to their risk, while calibration measures the agreement between probabilities predicted and outcome rates observed [6, 7]. Second, decision-theory-based classification shows that the optimal rule for making decisions is conditioned not on the universally adopted threshold of 0.50 but on the ratio of misclassification costs, i.e., the costs of false negatives versus false positives [8–10]. Third, explainable artificial intelligence provides the tools to investigate what drives models' predictions; nonetheless, the credibility of such explanations is to be evaluated according to criteria of consistency, use of complementary methods, and plausibility in the domain, rather than according to a single graphical technique [12, 13, 19–21]. Fourth, churn management literature emphasizes the practical nature of predictive power, which means that model predictions need to be translated into actions of prioritizing and retaining customers [8, 9, 18].

Responsible AI, Transparency and Regulatory Considerations

Besides issues pertaining to the technical interpretability of AI models, there are certain ethical and legal challenges in the field of telecom churn prediction because, through automated scoring of risks, decisions about discounts, deals, or preferences may be made. The GDPR is the legislation that lays out the rules associated with transparency and

automated decision-making in Europe [22]. Explainable AI may help with transparency and accountability by providing explanations for what causes model decisions; nevertheless, these explanations do not guarantee fairness or legality [12, 22]. Models may still reproduce bias when sensitive attributes or correlated proxies influence predictions. Consequently, responsible churn analytics requires both explainability and governance mechanisms that constrain how predictions are operationalized.

To provide a more rigorous positioning against previous research, Table 1 summarizes the dataset scale, validation design, modelling approach, verified predictive results and coverage of calibration, explainability and decision-oriented evaluation. Numerical results are included only when they are explicitly reported in the original study.

Table 1. Quantitative and Methodological Comparison of Selected Telecom Churn Studies

Study	Dataset and validation design	Main models and methodological approach	Verified predictive result	Calibration, XAI and decision-support coverage
[14]	Four public telecom datasets; the largest contained 100,000 observations and 101 features. Grid-search optimization, univariate feature selection and 10-fold cross-validation were applied. Friedman and Holm tests were also reported.	Eight classifiers combined with six data-transformation methods, including logarithmic, rank, Box–Cox, Z-score, discretization and weight-of-evidence transformations.	On Dataset 3, Random Forest with logarithmic transformation achieved AUC = 0.858 and F-measure = 0.820.	Strong multi-dataset benchmarking and statistical comparison; probability calibration, post-hoc XAI and explicit cost-sensitive threshold analysis were not included.
[3]	Public telecom churn dataset; 80/20 train–test split, with K-fold cross-validation on the training data for hyperparameter tuning.	Logistic Regression, Naïve Bayes, SVM, Decision Tree, Random Forest, KNN, Extra Trees, AdaBoost, XGBoost and CatBoost; gravitational search algorithm was used for feature selection.	AdaBoost achieved Accuracy = 81.71%, F-measure = 80.28% and AUC = 84%. XGBoost also achieved AUC = 84%.	Extensive model and feature-selection comparison; calibration, post-hoc explainability and business-cost threshold optimization were not evaluated.
[16]	Three telecom datasets containing 7,043, 3,333 and 3,150 observations. The data were divided into training and testing sets, and hyperparameters were evaluated using 10-fold cross-validation.	XGBoost integrated with feature selection, standardization and SMOTE–ENN hybrid resampling.	On Dataset 3 after SMOTE–ENN and standard scaling: Accuracy = 99.92%, Precision = 99.87%, Recall = 100%, F1-score = 99.93% and AUC = 99%.	Strong imbalance-handling and preprocessing framework; no probability calibration, SHAP analysis or explicit FN:FP cost-sensitive threshold evaluation. The results are not directly comparable because synthetic resampling was applied.

[17]	Telecom churn dataset; 75/25 train–test split. Oversampling, undersampling and five ratio-based balancing configurations—90:10, 80:20, 75:25, 65:35 and 50:50—were evaluated.	Ten standalone and ensemble models, including Perceptron, MLP, Naïve Bayes, Logistic Regression, KNN, Decision Tree, Gradient Boosting and XGBoost.	XGBoost with the 75:25 ratio-based balancing configuration was identified as the most promising overall setting. Exact aggregate F1-score and AUC values were not summarized in the article's main narrative.	Focuses on class balancing and ensemble performance; probability calibration, SHAP and explicit cost-sensitive threshold optimization were not evaluated.
[1]	Large customer-level telecommunications dataset; the exact sample size and detailed validation split were not clearly reported in the abstract.	Decision Trees, Boosted Trees and Random Forest, supported by LIME and SHAP explanations.	Random Forest achieved Accuracy = 91.66%, Precision = 82.2% and Recall = 81.8%.	Provides global and local explainability through LIME and SHAP; probability calibration and explicit cost-sensitive threshold analysis were not examined.
[4]	Real SyriaTel data covering nine months; 5 million labelled customers. A 70/30 train–test split and 10-fold cross-validation were used. A later operational evaluation involved 7.5 million customers.	Decision Tree, Random Forest, Gradient Boosting and XGBoost with statistical and social-network-analysis features on a big-data platform.	XGBoost achieved AUC = 93.301% on the development sample and AUC = 89% in the later operational evaluation.	Strong real-world scale, temporal coverage and operational validation; no post-hoc calibration, SHAP-based explanation or explicit FN:FP threshold scenarios.
Our study	Public Telco Customer Churn dataset with 7,043 customers and 20 predictors. Repeated stratified 5-fold cross-validation with three repeats and an exploratory 70/30 hold-out assessment were used.	Optimized Logistic Regression, Random Forest, XGBoost and LightGBM; Platt and isotonic calibration; SHAP; permutation importance; bootstrap confidence intervals; Friedman and McNemar tests; threshold and cost analysis.	Random Forest achieved CV F1-score = 0.6339 and ROC-AUC = 0.8477. LightGBM achieved hold-out F1-score = 0.6377. Isotonic calibration reduced the Brier score to 0.1371.	Integrates optimized benchmarking, calibration, statistical uncertainty, complementary explanation methods and explicit FN:FP cost-sensitive threshold analysis in one decision-oriented workflow.

Note: Numerical results are provided for methodological context and should not be interpreted as a direct performance ranking because the studies differ in dataset size, class distribution, preprocessing, resampling and validation procedures.

The comparison shows that previous studies provide important contributions in multi-dataset benchmarking, imbalance handling, explainability or large-scale operational validation. However, the selected studies do not jointly evaluate probability calibration,

statistical uncertainty, complementary explanation methods and explicit cost-sensitive threshold selection.

Synthesis and Research Gap Positioning

The existing research demonstrates the evolution of customer churn prediction from the use of classical classification algorithms to the adoption of ensemble approaches, learning to deal with imbalances, uplift modeling, and explainable AI [1–5, 14–18]. Recent studies focus more on XGBoost, LightGBM, hybrid ensembles, neural network-based approaches and use SHAP and other interpretability tools to detect the major predictors of churn [1, 2, 13–17]. However, there are some deployment-oriented components of churn analysis that still are not consistently incorporated. The majority of researches rank models in terms of their accuracy, F1-score, or ROC-AUC, paying less attention to the issue of the reliability of the predicted probabilities [6, 7, 14–17]. At the same time, such metrics as Brier score, log loss, and Expected Calibration Error become especially significant for models used for probability-based retention campaigns [6, 7]. Similarly, explainability findings are often presented in form of feature importance results without any connections with local errors, probability thresholds, and retention activities [1, 2, 12, 13].

A second gap concerns the separation between predictive evaluation and deployment policy. A model may perform well in terms of discrimination but still require calibration and threshold adjustment before it can support cost-sensitive churn management [6–11]. In churn contexts, false negatives and false positives have different business implications: missing a true churner may lead to lost customer value, while targeting a non-churner may generate unnecessary campaign costs [8–10]. For this reason, threshold selection should not rely only on the default 0.50 cut-off, but should be examined under explicit cost scenarios.

The present study addresses these gaps by proposing an integrated and reproducible decision-support workflow for telecom churn prediction. The study combines optimized benchmarking of Logistic Regression, Random Forest, XGBoost and LightGBM with probability calibration, SHAP-based explanation, permutation importance, bootstrap confidence intervals, error analysis and cost-sensitive threshold optimization. By connecting these elements in a single leakage-aware workflow, the study aims to move churn modelling from model comparison toward more reliable, interpretable and action-oriented decision support.

METHODOLOGY

Figure 1 presents an overview of the complete analytical workflow. This study follows a five-stage methodology:

- data acquisition and preprocessing,
- model development and hyperparameter optimization,
- model evaluation and selection,
- probability calibration and threshold analysis, and

- explainability and decision-support interpretation.

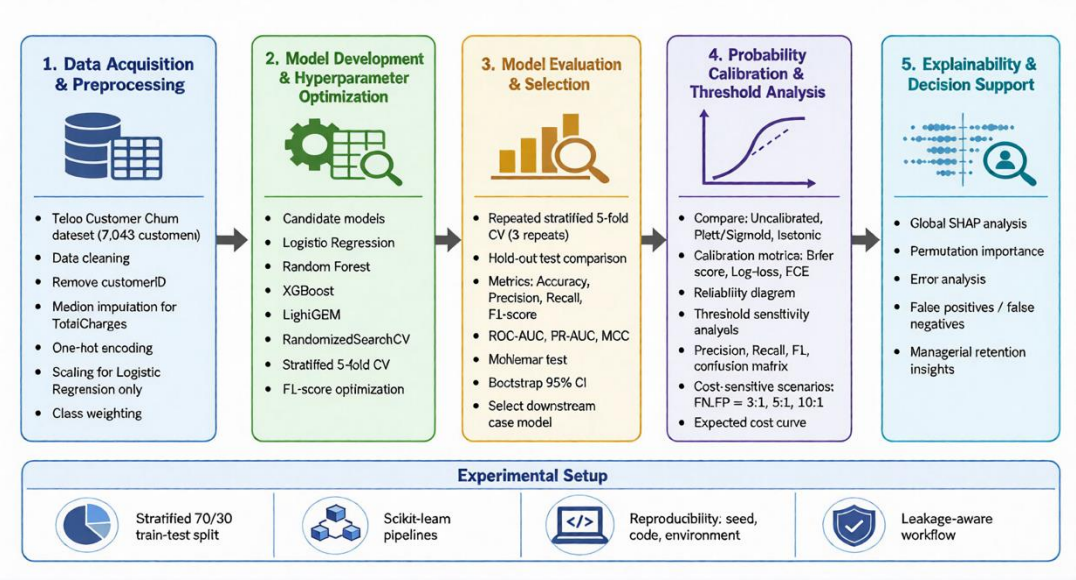

Figure 1. Overview of the Proposed Explainable, Calibrated, And Threshold-Aware Churn Modelling Framework.

Dataset and Prediction Task

The analysis is based on the publicly available Telco Customer Churn dataset, comprising 7043 customer records with demographic attributes, contract and billing information, subscribed services, and a binary churn label [23]. The target variable Churn indicates whether a customer discontinued the service. In light of the business importance of finding the churners, the assessment focuses on both discriminative ability and churn capturing capability of the models apart from their accuracy.

The public dataset is used to make the assessment more transparent, replicable, and comparable to other churn prediction researches. Missing data, imbalanced distribution, and churn rates in different subgroups are assessed prior to model building. The churn rate of the data is about 26.54%, making the classification task moderately imbalanced. Thus, not only accuracy, but other types of metrics are included into assessment.

The dataset represents a moderately imbalanced binary classification problem, with churners accounting for 26.54% of all customers. For this reason, evaluation was not based on Accuracy alone. Class-sensitive metrics such as Precision, Recall, F1-score and PR-AUC were included, together with probability-quality metrics such as Brier score, log-loss and Expected Calibration Error. The latter metrics were evaluated because the study uses predicted probabilities for decision support [6, 7]. The stratified train-test split preserved the churn distribution across training and testing subsets.

Table 2 depicts the summary of the telco customer churn dataset and class distribution use in our study.

Table 2. Summary of the Telco Customer Churn Dataset and Class Distribution

Element	Description
Dataset	Telco Customer Churn (Kaggle)
Total customers	7043
Target variable	Churn (Yes/No)
Total features (excluding target)	20
Non-churned customers	5,174
Churn rate (%)	26.54
Feature groups	Demographic; Contract & billing; Service subscriptions
Churn customers	1,869
Class distribution	73.46% non-churn, 26.54% churn
Class imbalance	Moderate imbalance
Train-test split	Stratified 70/30 split
Training set	4,930 customers
Test set	2,113 customers
Missing value issue	TotalCharges contained non-numeric/blank values after conversion
Missing value handling	Median imputation inside the preprocessing pipeline

Exploratory Data Analysis and Preprocessing

Before model training, an exploratory data analysis was conducted to inspect missing values, class distribution and churn patterns across selected customer groups. The subgroup-level inspection focused on contract type, internet service, payment method and senior citizen status, as these variables represent important customer, service and billing dimensions in telecom churn analysis. The variable TotalCharges contained non-numeric or blank entries after conversion to numeric format. These values were handled using median imputation inside the preprocessing pipeline. The non-informative identifier customerID was removed because it does not carry predictive information and could introduce noise into the modelling process. Categorical variables were transformed using one-hot encoding with handle_unknown="ignore", which prevents errors if unseen categories appear in validation or test folds. A drop-first strategy was applied to reduce redundant dummy variables in the encoded feature space. Continuous numerical variables were standardized only for Logistic Regression, because this model is sensitive to feature scale. Tree-based models, including Random Forest, XGBoost and LightGBM, were trained using the unscaled numerical representation.

All preprocessing operations were implemented inside scikit-learn pipelines. This means that imputation, encoding and scaling were fitted only on the training portion of each cross-validation split and then applied to the corresponding validation fold. The hold-out set was not used during preprocessing fitting or hyperparameter optimization. It was subsequently used for comparative model evaluation and for exploratory calibration, explainability and threshold analyses. This design reduces the risk of information leakage and ensures that performance differences reflect model behaviour rather than preprocessing inconsistencies.

Given the moderate class imbalance in the churn class, class weighting was applied during model training to reduce bias toward the majority non-churn class, see Table 3. No resampling was applied to validation or test folds, so the original class proportions were preserved during evaluation. This supports a realistic assessment of churn detection performance.

Table 3. Churn Rate by Selected Customer Groups

Variable	Category	N	Churn rate (%)
Contract	Month-to-month	3,875	42.71
Contract	One year	1,473	11.27
Contract	Two years	1,695	2.83
InternetService	DSL	2,421	18.96
InternetService	Fiber optic	3,096	41.89
InternetService	No	1,526	7.40
PaymentMethod	Bank transfer (automatic)	1,544	16.71
PaymentMethod	Credit card (automatic)	1,522	15.24
PaymentMethod	Electronic check	2,365	45.29
PaymentMethod	Mailed check	1,612	19.11
SeniorCitizen	No	5,901	23.61
SeniorCitizen	Yes	1,142	41.68

The subgroup-level inspection shows that churn is not uniformly distributed across customer groups. Customers with month-to-month contracts show a much higher churn rate than customers with one-year or two-year contracts. Similarly, customers using optic fibre internet service and electronic check payment show higher churn rates than the corresponding alternative groups. Senior customers also show a higher churn rate than non-senior customers. These descriptive patterns are not interpreted as causal evidence, but they support the need for nonlinear models, class-sensitive evaluation and threshold-aware decision analysis.

Model Selection and Rationale

Four classifiers were benchmarked to represent different levels of interpretability and predictive complexity. Logistic Regression was used as an interpretable linear baseline. Random Forest was included as a bagging-based ensemble capable of capturing nonlinear relationships and feature interactions [24]. XGBoost was included as a widely used gradient-boosting algorithm for structured tabular classification tasks [25]. LightGBM was added as an additional contemporary boosting comparator because recent churn prediction studies increasingly evaluate efficient boosting algorithms on customer-level tabular data [14–17]. The purpose of this benchmark was not only to identify the model

with the highest predictive score, but also to select a suitable model for probability calibration, explainability and threshold-aware decision support. Therefore, model comparison considered Precision, Recall, F1-score, ROC-AUC and PR-AUC rather than Accuracy alone [26].

Hyperparameter Optimization and Training

Hyperparameter optimization was performed on the training data only using RandomizedSearchCV with stratified five-fold cross-validation, see Table 4. For each classifier, a predefined search space was specified based on commonly used settings for binary classification with structured tabular data. The optimization criterion was F1-score, reflecting the need to balance churn detection and false-positive control in a moderately imbalanced classification setting. The hold-out set was excluded from hyperparameter optimization. After the best hyperparameter configuration was selected for each model, the optimized estimators were refitted on the full training set and compared on the hold-out set. The same hold-out predictions were subsequently used for exploratory calibration, explainability and threshold analyses; therefore, these downstream results are not presented as independent external validation. This procedure reduces the risk of optimistic bias and provides a fairer comparison across classifiers than relying on fixed model configurations. Class imbalance was addressed within model training through class-weighting strategies or equivalent imbalance-aware parameters where applicable. This allowed the models to give greater attention to the minority churn class without altering the original class distribution of the validation or test data.

Table 4. Hyperparameter Optimization Design

Model	Main hyperparameters tuned	Search method	Selection metric	Cross-validation design
Logistic Regression	Regularization strength, penalty and solver	RandomizedSearchCV	F1-score	Stratified 5-fold CV
Random Forest	Number of trees, maximum depth, minimum samples split, minimum samples leaf and feature sampling	RandomizedSearchCV	F1-score	Stratified 5-fold CV
XGBoost	Number of estimators, learning rate, maximum depth, subsampling, column sampling, minimum child weight and regularization	RandomizedSearchCV	F1-score	Stratified 5-fold CV
LightGBM	Number of estimators, learning rate, number of leaves, maximum depth, minimum child samples, subsampling and column sampling	RandomizedSearchCV	F1-score	Stratified 5-fold CV

Evaluation Protocol and Statistical Assessment

The optimized models were evaluated through repeated stratified 5-fold cross-validation with three repeats on the training data, followed by comparative assessment on the hold-out set. The hold-out set was excluded from preprocessing fitting and hyperparameter optimization, but it was subsequently used for model comparison and exploratory calibration, explainability and threshold analyses. Accuracy was reported for general comparability, but it was not treated as the primary criterion, because the dataset is moderately imbalanced. Precision, Recall and F1-score were used to assess the trade-off between correctly identifying churners and limiting false positives. ROC-AUC measured ranking performance across thresholds, while PR-AUC was included because it is particularly informative for imbalanced binary classification [26]. Brier score, log-loss and Expected Calibration Error were used to assess probability quality [6, 7]. Matthews Correlation Coefficient was reported as a balanced measure that uses all four elements of the confusion matrix [27]. McNemar's test was used to compare the two strongest models on paired hold-out predictions [28]. For the selected model, bootstrap 95% confidence intervals were estimated on the hold-out set [29]. The repeated cross-validation F1-scores of the four optimized models were compared using the Friedman test. Kendall's W was reported as an effect-size measure. Pairwise post-hoc comparisons were planned only if the omnibus test was statistically significant at $\alpha = 0.05$.

Probability Calibration Assessment

Considering that this study frames churn prediction as a decision-support problem, the reliability of predicted probabilities was evaluated in addition to classification performance. A model with strong ranking performance may still be unsuitable for probability-based decisions if its risk estimates are poorly calibrated [6, 7]. For the selected optimized model, three probability versions were compared: uncalibrated probabilities, Platt/sigmoid calibration and isotonic calibration. Calibration was performed using the training data and evaluated on the hold-out set. Calibration quality was assessed using Brier score, log-loss and Expected Calibration Error, with lower values indicating more reliable probability estimates [6, 7]. ROC-AUC and PR-AUC were also reported after calibration to verify whether ranking performance changed. Reliability diagrams were used to compare predicted churn probabilities with observed churn frequencies.

Threshold-Aware and Cost-Sensitive Decision Analysis

After performing the calibration process, an analysis was performed taking into consideration the threshold to transform probabilities of churners into decisions for retention. The threshold 0.50 might not be the best choice when the costs of mistakes vary between the cases of false negatives and false positives [8-11]. False negatives represent true churners that are missed, and false positives represent non-churners who could receive a needless offer for retention. In the first step, a common threshold analysis was conducted with several probability thresholds. With each threshold, Precision, Recall, F1-score, and the elements of the confusion matrix were calculated. This helped to analyze the impact of changes in threshold conservatism on the process of churn identification. In the

second step, a cost-sensitive threshold analysis was performed using specific cost ratio values for false negatives and false positives.

For each threshold t, the expected decision cost was computed using equation (1), following cost-sensitive and profit-oriented decision principles [8–10]:

$$EC(t) = C_{FN} \times FN(t) + C_{FP} \times FP(t) \quad (1)$$

where C_{FN} is the cost of missing a true churner, FN(t) is the number of false negatives at threshold (t), C_{FP} is the cost of incorrectly targeting a non-churner and FP(t) is the number of false positives at threshold (t).

Three illustrative business scenarios were evaluated: FN: FP = 3:1, 5:1 and 10:1. For each scenario, the threshold with the lowest expected cost was selected. This analysis operationalizes the principle that churn prediction should consider the business consequences of alternative error types rather than relying only on model accuracy [8–10].

Explainability, Permutation Importance and Error Analysis

To interpret the selected optimized model, SHAP analysis was applied to examine the contribution of individual features to churn predictions [13]. Global SHAP analysis was used to identify influential churn drivers and examine the direction and distribution of feature effects across hold-out observations. To avoid relying on a single explanation method, permutation importance was also computed as a complementary model-agnostic assessment of feature relevance [24]. The agreement and differences between the two methods were examined cautiously because feature dependence can affect attribution-based explanations [21]. Finally, an error analysis separated correct predictions, false positives and false negatives to provide a more practical understanding of model limitations.

RESULTS

This section reports the empirical findings of the revised churn prediction workflow. The results have been arranged based on the main stages of the analysis: hyperparameter tuning, cross-validation, hold-out testing, uncertainty quantification, calibration of probabilities, threshold-based cost-sensitivity analysis, explanation, and error analysis. Such organization is useful when switching the focus of the paper from model performance comparison to reliability, interpretation, and decision-making.

Hyperparameter Optimization Results

Table 5 highlights the best cross-validated F1-score and the full hyperparameters settings chosen for each model. The LightGBM model achieved the highest F1-score of 0.6358 while the Random Forest model obtained 0.6353. However, the difference was marginal and should not be interpreted as evidence of statistical superiority.

Table 5. RandomizedSearchCV Optimization Summary

Model	Best CV F1	Best Parameters
Logistic Regression	0.6279	C = 0.1; penalty =l2; solver = lbfgs
Random Forest	0.6353	n_estimators = 200; max_depth = 10; min_samples_split = 5; min_samples_leaf = 6; max_features = sqrt
XGBoost	0.6332	n_estimators = 300; max_depth = 5; learning_rate = 0.03; subsample = 0.8; colsample_bytree = 1.0; min_child_weight = 3; reg_lambda = 0.5
LightGBM	0.6358	n_estimators = 500; learning_rate = 0.01; num_leaves = 31; max_depth = 8; min_child_samples = 30; subsample = 1.0; colsample_bytree = 0.8

Note: Hyperparameter optimization was performed using RandomizedSearchCV with stratified five-fold cross-validation and F1-score as the selection metric.

Repeated Cross-Validation Results

Table 6 reports the repeated stratified cross-validation results for the optimized models. Each model was evaluated using stratified 5-fold cross-validation repeated three times on the training data. This evaluation design was used to obtain a more stable estimate of model performance across different data splits and to reduce dependence on a single train-validation partition. Because the dataset is moderately imbalanced, the interpretation does not rely on Accuracy alone. Recall is important for measuring the ability to identify potential churners, while Precision reflects the extent to which predicted churners are correctly classified. F1-score provides a balance between Precision and Recall, and PR-AUC is particularly relevant because the churn class represents the minority class. ROC-AUC is also reported to assess ranking performance across classification thresholds.

The Friedman test found no statistically significant differences among the four models repeated cross-validation F1-scores, $\chi^2(3) = 2.600$, $p = 0.4575$. Kendall's $W = 0.0578$ indicated a very small effect size. Therefore, the models were broadly comparable in terms of F1-score. The repeated cross-validation results show that the optimized models achieved broadly comparable performance across the evaluated metrics. Random Forest obtained the highest mean Accuracy, ROC-AUC, PR-AUC, Precision and F1-score, while Logistic Regression achieved the highest mean Recall. LightGBM and XGBoost also showed competitive and stable performance, with small standard deviations across repeated folds. These results indicate that no single model clearly dominated all evaluation dimensions during cross-validation.

Therefore, the choice of the downstream case model was not based solely on repeated cross-validation results but also on the exploratory hold-out comparison and its suitability for calibration and threshold analysis.

Table 6. Repeated Stratified Cross-Validation Performance of Optimized Models

Model	Accuracy	ROC-AUC	PR-AUC	Precision	Recall	F1-score
Logistic Regression	0.7499 ± 0.0128	0.8437 ± 0.0089	0.6589 ± 0.0199	0.5189 ± 0.0156	0.7989 ± 0.0195	0.6291 ± 0.0154
Random Forest	0.7721 ± 0.0084	0.8477 ± 0.0075	0.6692 ± 0.0191	0.5525 ± 0.0122	0.7439 ± 0.0203	0.6339 ± 0.0127
XGBoost	0.7624 ± 0.0145	0.8442 ± 0.0099	0.6685 ± 0.0204	0.5373 ± 0.0198	0.7607 ± 0.0221	0.6296 ± 0.0189
LightGBM	0.7632 ± 0.0132	0.8447 ± 0.0096	0.6656 ± 0.0197	0.5385 ± 0.0181	0.7612 ± 0.0196	0.6306 ± 0.0153

Exploratory Hold-out Assessment

After repeated cross-validation, the optimized models were compared on the hold-out set. This set was not used during preprocessing fitting or hyperparameter optimization. However, because the hold-out results also informed the selection of LightGBM for the downstream workflow, they should be interpreted as comparative out-of-sample evidence rather than as independent external validation. Table 7 shows that the optimized models achieved competitive hold-out performance, but no model dominated all metrics.

Table 7. Hold-out comparison across benchmarked models using uncalibrated probabilities at the default threshold of 0.50

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC	Log-loss	Brier score	MCC
Logistic Regression	0.7487	0.5174	0.7950	0.6268	0.8439	0.6425	0.4949	0.1669	0.4735
Random Forest	0.7676	0.5443	0.7665	0.6366	0.8426	0.6524	0.4688	0.1553	0.4879
XGBoost	0.7601	0.5333	0.7718	0.6307	0.8387	0.6481	0.4786	0.1597	0.4790
LightGBM	0.7634	0.5372	0.7843	0.6377	0.8396	0.6507	0.4809	0.1609	0.4896

Logistic Regression achieved the highest Recall, indicating strong churn-capture ability, while Random Forest obtained the highest Accuracy, Precision and PR-AUC among the compared models. LightGBM achieved the highest F1-score and MCC, suggesting the strongest overall balance between churn detection and false-positive control on hold-out sample. Based on this balance, LightGBM was retained as the downstream case model for the subsequent calibration, explainability and threshold-aware decision analyses. At the default 0.50 threshold, the selected LightGBM model produced 1,173 true negatives, 379

false positives, 121 false negatives and 440 true positives. Such error numbers serve as the foundation for subsequent threshold and cost sensitivity analysis.

Confidence Intervals and Statistical Comparison

In order to estimate uncertainty in the performance of selected LightGBM model, the bootstrap 95% confidence intervals for holdout set were derived. Such approach provides interval estimation of main performance characteristics instead of point estimation only. Confidence intervals were estimated for F1-score, Recall, Precision, ROC-AUC, PR-AUC, and Brier Score.

Table 8 shows the bootstrap-based 95% confidence intervals for the main performance metrics of the uncalibrated LightGBM model on the hold-out set. The selected LightGBM model demonstrates stability of hold-out performance in terms of key metrics used for model evaluation. Confidence interval of F1-score for 95% confidence level varies between 0.6069 and 0.6667, while confidence interval for Recall is relatively high and equal to {0.7491, 0.8171}. The relatively narrow confidence interval suggests that the model's churn-capture performance is reasonably stable within the evaluated hold-out sample. The confidence intervals for ROC-AUC and PR-AUC are consistent with the ranking performance of the chosen model. Confidence interval for the Brier score gives additional information for probability quality estimation before calibration. Besides using bootstrap confidence intervals, McNemar's test was applied to comparing hold-out predictions of the two top-performing models. The test evaluates whether two classifiers make significantly different paired errors on the same observations [28]. Table 9 shows that the McNemar p-value is 0.4018, which is above the conventional 0.05 significance level.

Table 8. Bootstrap 95% Confidence Intervals for the Uncalibrated LightGBM Model on the Hold-out Set

Metric	Mean	Lower 95% CI	Upper 95% CI
F1	0.6370	0.6069	0.6667
Recall	0.7840	0.7491	0.8171
Precision	0.5367	0.5036	0.5711
ROC-AUC	0.8396	0.8220	0.8568
PR-AUC	0.6519	0.6095	0.6905
Brier	0.1610	0.1525	0.1702

This indicates that the difference in classification errors between LightGBM and Random Forest is not statistically significant on the hold-out set. Therefore, the selection of LightGBM as the model retained for downstream exploratory analysis is not based on statistical superiority alone, but on its slightly stronger F1-score and MCC, together with

its suitability for the subsequent calibration, explainability and threshold-aware decision analyses.

Table 9. McNemar Test between the Two Best Hold-out Models

Model A	Model B	A correct / B wrong	A wrong / B correct	McNemar p-value
LightGBM	Random Forest	41	50	0.4018

Probability Calibration Results

Because LightGBM was retained as the downstream case model, calibration analysis was conducted on its predicted churn probabilities. Table 10 compares the uncalibrated LightGBM probabilities with two post-hoc calibration methods: Platt/sigmoid calibration and isotonic calibration. The comparison includes Brier score, log-loss and Expected Calibration Error, together with ROC-AUC and PR-AUC to verify whether calibration affected ranking performance. Table 10 shows that both calibration methods substantially improved probability reliability compared with the uncalibrated LightGBM model.

Table 10. Calibration Comparison for the Selected LightGBM Model

Calibration method	Brier score	Log-loss	ECE	ROC-AUC	PR-AUC	F1 at 0.50	Recall at 0.50	Precision at 0.50
Uncalibrated	0.1609	0.4809	0.1267	0.8396	0.6507	0.6377	0.7843	0.5372
Platt / Sigmoid	0.1373	0.4239	0.0218	0.8414	0.6502	0.5879	0.5544	0.6258
Isotonic	0.1371	0.4240	0.0236	0.8424	0.6486	0.5711	0.4973	0.6707

The Brier score decreased from 0.1609 to 0.1373 with Platt/sigmoid calibration and to 0.1371 with isotonic calibration. Expected Calibration Error also decreased sharply, from 0.1267 in the uncalibrated model to 0.0218 and 0.0236 after Platt/sigmoid and isotonic calibration, respectively. Both calibration methods improved the agreement between predicted probabilities and observed churn outcomes, as indicated by the lower Brier scores and Expected Calibration Error values. For isotonic calibration, the minimum Brier score and maximum ROC-AUC values were obtained, while the minimum log-loss and ECE were achieved by Platt/sigmoid calibration. Since the methodology used Brier score as the key factor in making the choice, probabilities from isotonic-calibrated LightGBM classifier were selected for further analysis. The decrease in F1-score and Recall values at the default threshold value of 0.50 after calibration does not mean that calibration was a failure. The main objective of calibration is the improvement of probability estimates, not the classification metrics at a particular threshold. Calibration changes the distribution of probabilities, so the default 0.50 threshold can become inappropriate. Therefore, the next step involves the evaluation of different thresholds and cost-sensitive settings to find the most relevant decision threshold.

Figure 2 shows the reliability curves of the uncalibrated and calibrated LightGBM probabilities compared with the ideal calibration line.

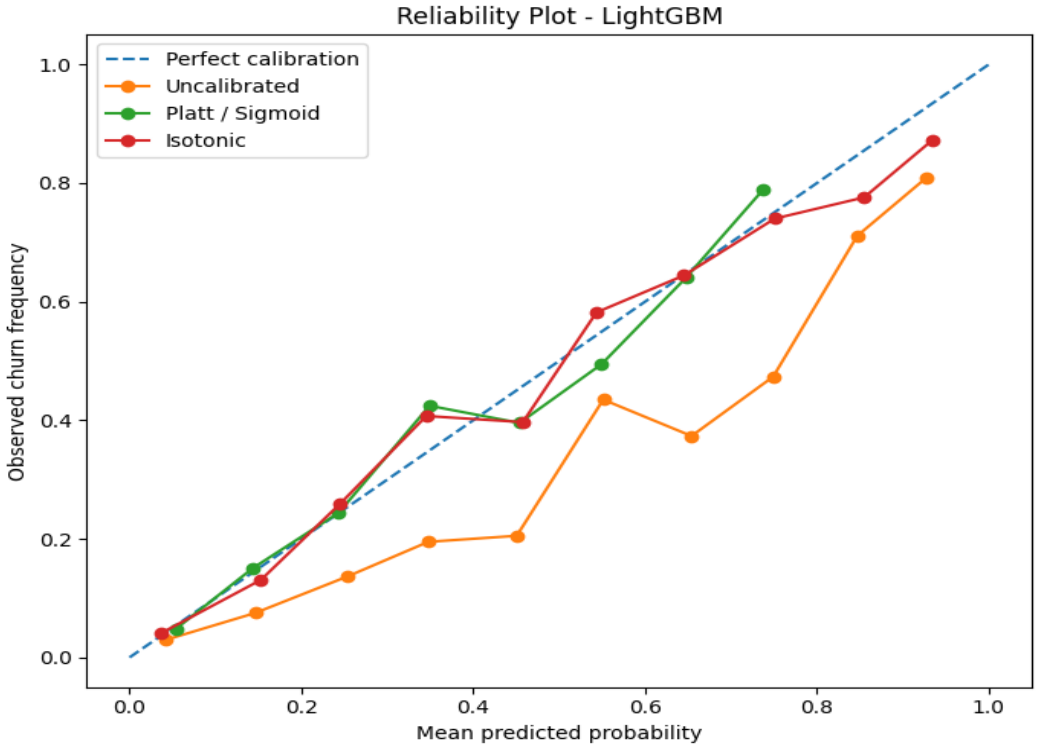

Figure 2. Reliability Diagram of LightGBM Uncalibrated vs. Calibrated Probabilities

The closer the curve is to the diagonal line, the better the calibration of that probability curve. Visual proof adds value to the calibration measures numerically calculated and confirms the quality of probability estimates before optimizing decision thresholds.

Threshold-Aware and Cost-Sensitive Decision Results

After calibration, the threshold-aware method was applied to the selected calibrated output of probability from the LightGBM. The purpose was to analyze the effects of different probability cut-off values on the accuracy of churn identification, false positive results, and false negative outcomes. This procedure is necessary since the default cut-off value of 0.50 may not work best for managing the churn.

Table 11 shows the performance of the isotonic-calibrated LightGBM model across different decision thresholds, highlighting the trade-off between Precision, Recall, F1-score, and classification errors. At the threshold of 0.20, the model achieved the highest Recall, capturing 86.99% of actual churners, but this improvement was accompanied by 538 false positives. At a default threshold value of 0.50, both Precision and Accuracy are improved, while the Recall reduces to 49.73% and thus means that 282 churners are not captured by the model. With the threshold value of 0.30, the maximum F1-Score is observed. Thus, it is clear that the default threshold of 0.50 is not always the best threshold for churn management.

Table 11. Performance of the Isotonic-Calibrated LightGBM Model across Alternative Decision Thresholds

Threshold	Accuracy	Precision	Recall	F1	TN	FP	FN	TP
0.2	0.7108	0.4756	0.8699	0.6150	1014	538	73	488
0.3	0.7686	0.5466	0.7522	0.6332	1202	350	139	422
0.4	0.7832	0.5851	0.6310	0.6072	1301	251	207	354
0.5	0.8017	0.6707	0.4973	0.5711	1415	137	282	279
0.6	0.7889	0.7291	0.3262	0.4507	1484	68	378	183

Cost sensitive threshold optimization beyond conventional threshold sensitivity analysis was done under different cost scenarios in order to compare false negative cost against the false positive cost. It is recognized that in churn management, the cost of missing a churner may be higher than the cost of treating a non-churner as a churner. In each of the cost scenarios, the decision expected cost was calculated for different threshold values.

The cost ratios used in this analysis are expressed in normalized units rather than in euros because the public dataset does not contain campaign cost, customer lifetime value, offer value or retention-success information. A false-positive cost of one unit represents the relative cost of contacting or incentivizing a customer who would not have churned, whereas a false-negative cost of three, five or ten units represents progressively stronger assumptions about the loss associated with failing to identify a genuine churner. These scenarios should therefore be interpreted as sensitivity analyses. They show how the preferred threshold changes as the organization places greater value on avoiding missed churners, rather than estimating the actual financial return of a specific telecom campaign.

For operational use, these relative costs could be replaced with operator-specific estimates. A telecom provider could define C_{FP} from contact and incentive expenditure, and C_{FN} from expected customer-margin loss adjusted by the probability that a retention intervention would have succeeded. The resulting threshold would then depend on customer value, campaign capacity and expected treatment effectiveness. The present analysis demonstrates the decision logic, while the final financial parameters must be estimated from real campaign data.

Table 12 shows that the cost-minimizing threshold selected in the hold-out sample becomes lower when the cost of false negatives (FN) becomes higher. With FN:FP = 3:1, the cost-minimizing threshold is 0.24, which results in the Recall equal to 82.89% and 96 false negatives. If the cost of false negative becomes higher, i.e., FN:FP = 5:1, then the optimal threshold becomes 0.19 and leads to the Recall of 89.13% and 61 false negatives. In the most conservative variant FN:FP = 10:1, the optimal threshold becomes even lower and equals 0.13; this threshold enables the model to recognize 93.23% of churners and produce only 38 false negatives. Thus, it can be stated that the choice of threshold cannot be viewed as a

technical procedure but should be considered as a managerial decision. The higher the cost of missing a churner, the lower the preferred threshold becomes, since it will recognize more customers at the risk of churn at the expense of additional errors and expenses. Speaking about the number of errors, false negatives decrease from 96 in the case of FN:FP = 3:1 to 61 in the case of FN:FP = 5:1 and to 38 in the case of FN:FP = 10:1. Figure 3 illustrates how the expected cost changes across decision thresholds under the three FN:FP cost scenarios.

Table 12. Cost-Sensitive Threshold Optimization under Alternative FN:FP Cost Scenarios

Scenario	C_FN	C_FP	Best threshold	Expected cost	Accuracy	Precision	Recall	F1
FN:FP = 3:1	3	1	0.24	738	0.7416	0.5082	0.8289	0.6301
FN:FP = 5:1	5	1	0.19	879	0.6995	0.4655	0.8913	0.6116
FN:FP = 10:1	10	1	0.13	1120	0.6318	0.4141	0.9323	0.5735

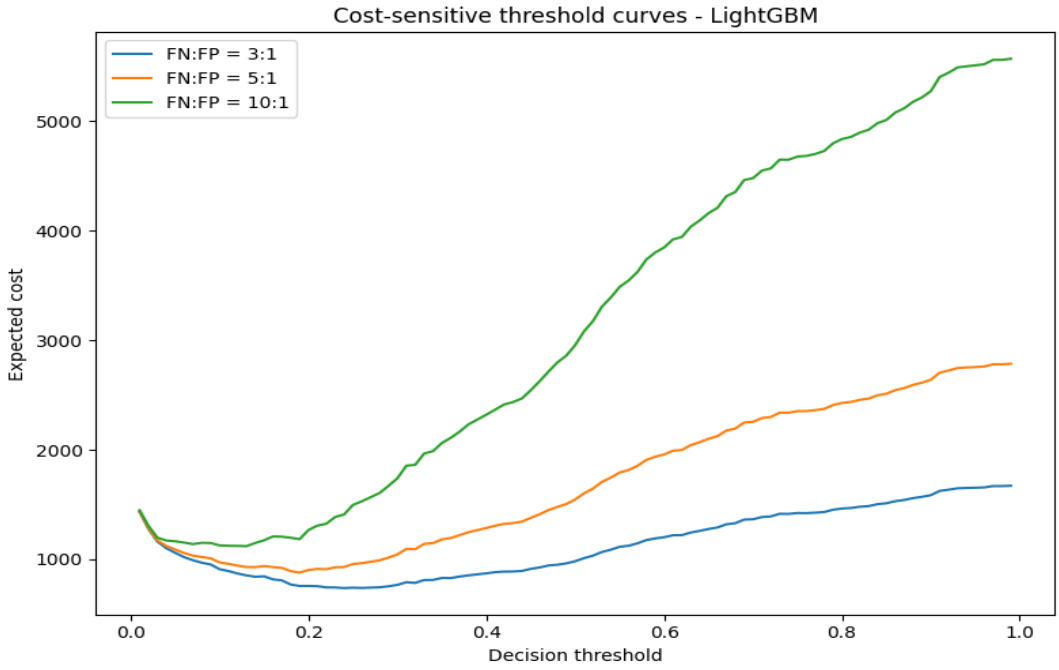

Figure 3. Expected decision cost across probability thresholds under alternative FN:FP cost scenarios.

Figure 3 further corroborates the findings presented in Table 12. With an increase in the relative cost of false negatives, the minimum expected cost is achieved at progressively lower probability thresholds. This indicates that when missing a true churner is considered more costly, the model should operate with a lower decision threshold in order to identify

the "more at-risk" customers. However, this also increases the number of false positives, showing that threshold selection must balance churn capture against campaign cost.

Explainability and Error Analysis

After selecting the optimized LightGBM model and evaluating its calibrated probabilities and decision thresholds, explainability analysis was conducted to identify the core factors of churn prediction. SHAP analysis was used to interpret the contribution of individual features to the model output. In addition, permutation importance was computed as a complementary validation method, and error analysis was performed to examine false-positive and false-negative prediction patterns.

SHAP analysis was used to identify the most influential variables contributing to the selected LightGBM model's churn predictions. The highest mean absolute SHAP values were observed for contract-related variables and tenure, followed by service and billing-related variables. In particular, the most influential features included Contract_Two year, Tenure, Contract_One year, InternetService_Fiber optic, MonthlyCharges, PaymentMethod_Electronic check, TotalCharges, InternetService_No, PaperlessBilling_Yes and OnlineSecurity_Yes. These variables indicate that churn predictions are mainly driven by customer commitment, length of relationship, service type and billing behaviour.

Figure 4 shows the SHAP summary plot for the selected LightGBM model.

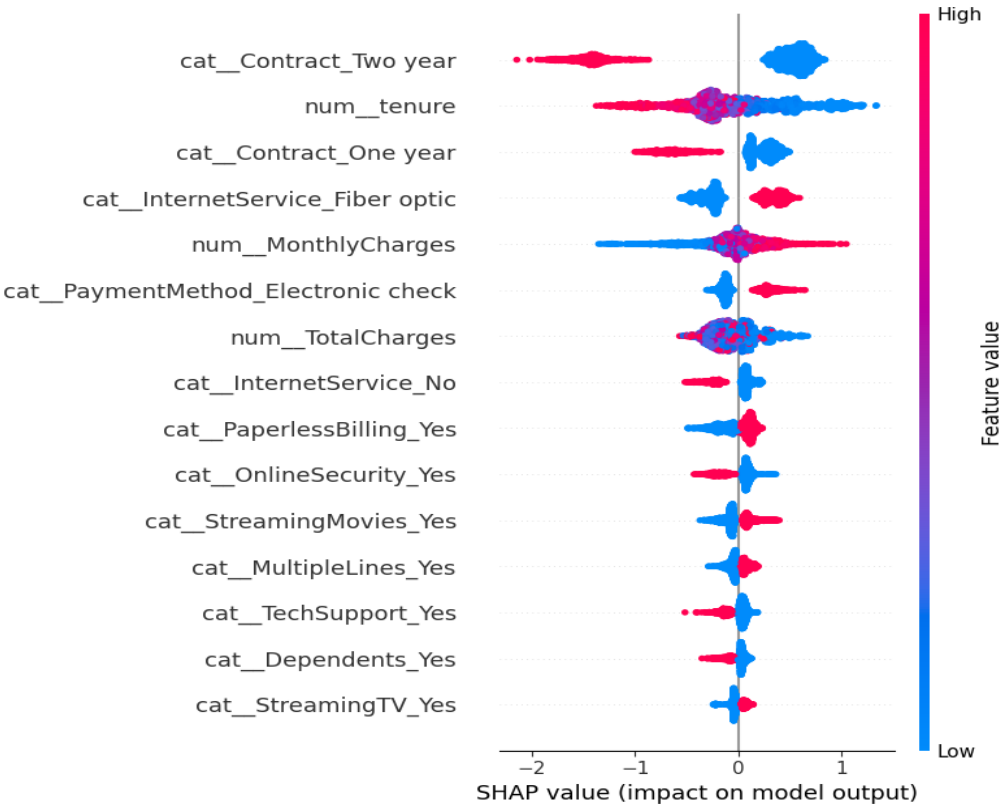

Figure 4. SHAP Summary Plot for the Selected LightGBM Model

Features with positive SHAP values increase the predicted probability of churn, while features with negative SHAP values reduce it. Overall, the SHAP results suggest that the model relies on meaningful customer, contract, service and billing patterns rather than arbitrary variables. The above observations contribute to the interpretability element of the paper and help to translate the model prediction results into actionable retention insights.

These findings suggest that model explanations can support differentiated retention strategies: contract-related risk may motivate commitment incentives, short-tenure risk may call for stronger onboarding, and service-related risk may justify proactive technical support.

Permutation Importance

To complement the SHAP analysis, permutation importance was computed on the hold-out set. This additional analysis was used to verify whether the most influential variables identified by SHAP remained important when feature relevance was evaluated through a different model-agnostic method. While SHAP explains how features contribute to model predictions, permutation importance evaluates how much model performance decreases when the values of a feature are randomly shuffled. Table 13 presents the permutation importance results for the ten most influential features in the selected LightGBM model.

Table 13. Permutation Importance for the Selected LightGBM Model

Rank	Feature	Permutation importance mean	Permutation importance std
1	Contract	0.1136	0.0124
2	InternetService	0.0779	0.0072
3	Tenure	0.0594	0.0072
4	PaymentMethod	0.0159	0.0026
5	TotalCharges	0.0146	0.0030
6	MonthlyCharges	0.0118	0.0059
7	TechSupport	0.0062	0.0035
8	PhoneService	0.0052	0.0017
9	StreamingMovies	0.0044	0.0017
10	Gender	0.0043	0.0025

The above results confirm that contract type, internet service and tenure are the most important predictors under permutation importance. The consistency between SHAP and permutation importance strengthens the interpretability findings. SHAP explains how individual feature values contribute to churn-risk predictions, while permutation importance shows how much predictive performance decreases when a feature is disrupted. Since both approaches highlight contract type, internet service and tenure, the

results suggest that the selected LightGBM model relies on recurring and domain-relevant churn-related patterns

Error Analysis

An additional error analysis was conducted for the isotonic-calibrated LightGBM model at the threshold of 0.50. At this operating point, the model correctly classified 1,694 of the 2,113 hold-out observations, corresponding to 80.17% of the hold-out set. It produced 282 false negatives and 137 false positives. These error counts refer to isotonic-calibrated probabilities and therefore differ from the uncalibrated LightGBM confusion matrix reported in Table 7.

False negatives are particularly important in churn management, since they represent customers who actually churned but were not identified by the model. In practical terms, these customers would not receive retention intervention despite being at risk. False positives, on the other hand, represent non-churners who may receive unnecessary retention offers, potentially increasing campaign costs.

This error distribution reinforces the importance of threshold-aware and cost-sensitive analysis. Depending on business priorities, a telecom operator may prefer a lower threshold to reduce missed churners, even if this increases the number of false positives. Therefore, error analysis supports the study's argument that churn prediction should be evaluated not only through aggregate metrics, but also through the practical consequences of different error types.

Evaluation of the Research Hypotheses

The empirical results provide differentiated support for the four hypotheses.

- H1 is partially supported. Random Forest achieved the highest mean F1-score and ROC-AUC, but the Friedman's test did not identify a statistically significant difference among the models' F1-scores, χ^2 (3) = 2.600, p = 0.4575, W= 0.0578.
- H2 was supported. Isotonic calibration reduced the Brier score from 0.1609 to 0.1371, representing a relative improvement of 14.8%, while ROC-AUC changed from 0.8396 to 0.8424, an absolute difference of only 0.0028.
- H3 was supported by the exploratory hold-out analysis. The calibrated threshold of 0.30 increased F1-score from 0.5711 to 0.6332 compared with the 0.50 threshold. Furthermore, cost-optimized thresholds reduced expected decision cost by 24.9%, 43.2% and 62.1% under the 3:1, 5:1 and 10:1 FN: FP scenarios, respectively.
- H4 was supported at the feature-group level. After aggregating encoded SHAP features to their original variables, Contract, InternetService and Tenure were identified among the five leading groups by both SHAP and permutation importance, corresponding to a top-five overlap of 60%. These relationships remain associative rather than causal.

DISCUSSION

The results indicate that the updated workflow offers a more comprehensive approach to telecom churn prediction than simply comparing the models' performance. LightGBM was selected for downstream analysis because it achieved a marginally higher hold-out F1-score and MCC, although McNemar's test did not indicate a statistically significant advantage over Random Forest. Moreover, the results highlight how the evaluation of predictive models based only on default-threshold metrics can be misleading. Calibration, robustness to changes in threshold settings, cost-based optimizations, and error analysis are very important aspects to consider, especially in case where churn predictions are employed for customer retention decisions.

In contrast to the previous studies of telecom churn that focus on predictive accuracy or class balancing, the current findings suggest that when cost factors are considered in addition to calibration and decision thresholds, similar discriminatory capabilities will lead to different conclusions in practice. Several studies, including [3, 14], have provided detailed comparisons of predictive models, while [1] has focused primarily on model interpretability. The present study extends this line of research by additionally examining probability estimates and decision thresholds.

The results corroborate earlier literature regarding telecom churn, showing that ensembles perform well on structured customer data [3, 14–17]. Yet, the results here also show that small differences in F1-Score or ROC-AUC do not necessarily determine deployment worth. In contrast to other literature which focuses on prediction or explanation, our approach shows that calibrating and thresholding may have a significant impact on the number of people selected and those missed.

The main contribution of the study is in its more comprehensive approach to model evaluation as opposed to its improved prediction performance per se. Since the evaluation is conducted on a single public dataset and not all models previously analyzed are replicated, the results cannot be taken as proof of overall predictive superiority of certain algorithms.

Why the Models Behaved Differently

The comparison between Random Forest and LightGBM is particularly informative because the two models did not produce the same ranking across evaluation settings. Random Forest achieved the strongest average cross-validation performance on several metrics, whereas LightGBM obtained a marginally higher hold-out F1-score and MCC. This difference should not be interpreted as evidence that LightGBM is universally superior. Random Forest reduces variance by averaging many decorrelated trees and may therefore provide stable performance across repeated samples. LightGBM, by contrast, builds trees sequentially and concentrates more directly on residual prediction errors, which may improve the final balance between true-positive detection and false-positive control on a particular hold-out sample. The small magnitude of the differences also indicates that the data contain a relatively stable predictive signal that can be captured by

several model families. From a deployment perspective, this means that the final choice should not depend on a single metric. Probability calibration, threshold behaviour, explanation quality and operational constraints are equally relevant when the leading models perform similarly.

The calibration findings also clarify why discrimination and decision readiness should not be treated as equivalent. A model can rank high-risk customers effectively while assigning probabilities that are systematically too high or too low. This distinction matters in retention campaigns because predicted probabilities may be used to rank customers, estimate campaign size or define eligibility rules. The improvement in Brier score and Expected Calibration Error therefore has a practical interpretation: the calibrated probabilities provide a more credible basis for comparing risk levels across customers. At the same time, calibration altered the distribution of predicted probabilities and reduced classification performance at the unchanged 0.50 cut-off. This is not contradictory. It shows that calibration and threshold selection must be considered jointly, since a threshold chosen for uncalibrated scores may no longer be appropriate after the probability scale has been corrected.

Theoretical Interpretation of the Findings

These results support the theoretical separation of predictive performance from probability reliability, interpretability, and decision utility. Lack of any statistically significant differences among the best classifiers suggests that model discrimination by itself is not enough in choosing the deployment candidate. This result is consistent with the view that ensemble learning techniques can effectively capture the dominant nonlinear patterns in structured telecommunications data, with no substantial differences in the trade-offs among Recall, Precision, and MCC.

Calibration results confirm the framework's second theoretical assumption. Improvement in terms of both Brier score and Expected Calibration Error, along with almost constant ROC-AUC scores, suggest that ranking and probability reliability are two different yet related aspects of a model. From practical perspective, a model could correctly classify high-risk cases but still generate probabilities that need adjustment before using them for prioritization or allocating campaign resources.

Threshold analysis supports the cost-sensitive decision theory. Change of the optimal threshold below 0.50 with the increase of the cost ratio of false negative to false positive suggests that the default classification threshold is not neutral with respect to business goals. In other words, threshold setting represents the organization's policy concerning tolerance for churn, false positives, and campaigns' capacity limitations.

Finally, consistency of SHAP and permutation feature importance supports the explanation consistency hypothesis. Both approaches emphasized the significance of contract type, internet service and tenure, and thus confirmed that these variables are not an artifact of one particular method. However, the findings should be seen as predictions and not as causal relations because the former does not necessarily imply the latter.

Interpretations and Comparison to Previous Research

The empirical evidence supports a number of hypotheses proposed at the beginning of the research. First, benchmarking indicates that the proposed optimized ensemble models show satisfactory performance and adequately capture the nonlinear patterns of churn behavior which could be overlooked by simpler linear approaches [1-5, 14-17].

Second, the results of the calibration analysis show that besides the classification accuracy, the quality of the predicted probabilities needs to be assessed as well. In this particular case, both Platt scaling (sigmoid) and isotonic calibration lower the Brier score and ECE in comparison with the probabilities predicted by uncalibrated LightGBM, but do not impact the metrics such as ROC-AUC and PR-AUC. Such results confirm the difference between the abilities of the models to discriminate and predict the probabilities mentioned earlier [6, 7].

Third, threshold and cost-sensitive evaluations prove that a constant 0.50 cut-off is often inappropriate for managing churn. The threshold which minimizes the cost goes down from 0.24 in case of 3:1 cost scenario to 0.13 in case of 10:1 cost scenario. Thus, previous results which suggested that decision thresholds should be adjusted according to misclassification costs and business needs are confirmed [8, 11].

Fourth, both SHAP values and permutation importance reveal consistent predictors of churn: type of the contract, tenure of the customer, internet service, and way of payment. These factors coincide with those revealed by recent studies on explainable AI approaches to telecom churn prediction [1,2]. Also, the results demonstrate how the global importance of features could be used in combination with other tools of interpretation [13, 21].

Managerial and Decision-Support Implications

These results hold managerial significance for handling customer churn in the telecommunications industry. First of all, although the accurate model performance is vital, churn prediction cannot limit itself to technical classifications since retention strategies rely on calibrated probabilities, the recognition of the significant variables, and the appropriate choice of decision thresholds. The LightGBM model shows high predictive capabilities; nevertheless, its practical importance is maximized by combining it with the calibrated probabilities and thresholds that consider the cost of misclassifications.

Secondly, according to the results of the calibration analysis, it is necessary to pay attention to the model output probabilities since they might not be appropriate to make the customer prioritization decisions. Incorrectly calibrated probabilities may lead to the incorrect assessment of risks and to the over/underestimate of the probabilities of churn occurrence. Therefore, the calibration assessment allows receiving valuable information which might be missed when using discriminative statistics such as the AUC [6, 7].

Thirdly, the analysis of the decision thresholds makes it obvious that the use of the default 0.50 threshold is not always the best choice for making plans for the churn interventions. If the cost of not detecting the churner is high, the lower thresholds allow

identifying more customers at risk though they increase the number of false positives and the marketing costs [8–11].

Finally, the model interpretability allowed us to identify specific factors like the type of contract, tenure, whether the customer uses the internet service, the way of payments, and billing characteristics as predictors of the churn which corresponds with the previous literature on the interpretable models of churn [1, 2].

The explanation results also have different implications depending on whether a variable is potentially actionable. Contract type and payment method can support the design of retention offers, such as incentives for moving from month-to-month arrangements to longer commitments or reducing friction associated with particular payment channels. Tenure is not directly modifiable, but it can identify stages of the customer relationship in which intervention may be most valuable. Service-related variables, including internet service and technical-support access, can guide further investigation into whether dissatisfaction, service complexity or unmet support needs are concentrated within specific customer segments. These interpretations should be treated as decision cues rather than causal prescriptions. SHAP and permutation importance identify predictive associations; they do not prove that changing a feature will prevent churn.

Limitations and Future Work

This study has five main limitations. First, the analysis is based on one public cross-sectional telecom dataset, limiting generalization to other operators, countries and customer populations. Second, the absence of timestamps and longitudinal interaction data prevents temporal validation and concept-drift assessment. Third, the same hold-out observations informed model comparison and exploratory downstream calibration and threshold analyses; consequently, these downstream results should not be interpreted as independent external-validation estimates. Fourth, the evaluated FN:FP scenarios use illustrative relative cost units because actual contact cost, customer lifetime value, campaign response and retention-success data were unavailable. Fifth, SHAP and permutation importance describe predictive associations and do not establish causal drivers of churn. Future work should validate the workflow on longitudinal and operator-specific data, use separate calibration and final-test partitions, evaluate explanation stability across bootstrap samples and integrate campaign-response, uplift and customer-value information.

Future research could extend this work by validating the proposed workflow on real operator data, including Albanian or regional telecom datasets if available. Additional research could incorporate temporal modelling, uplift modelling, customer lifetime value and campaign-response data to test whether predicted churn risk translates into effective retention interventions [18]. Future studies may also compare additional tabular deep-learning models or causal approaches to strengthen the decision-support value of churn prediction systems.

SUMMARY AND CONCLUSION

This study examined telecom churn prediction as a sequence of connected decision problems rather than as a model-ranking exercise alone. The results showed that the leading models were close in predictive performance but differed in their operational profiles. Random Forest produced the strongest mean cross-validation results across several metrics, LightGBM achieved the highest hold-out F1-score and MCC, and Logistic Regression retained the highest Recall. This pattern indicates that no single model should be selected without considering the intended retention objective.

The calibration analysis showed that ranking performance alone was insufficient to judge deployment readiness. Platt and isotonic calibration improved probability reliability, while the subsequent threshold analysis demonstrated that the conventional 0.50 cut-off was not stable across business assumptions. As the relative cost of missing a churner increased, the cost-minimizing threshold moved from 0.24 to 0.13 and Recall increased accordingly. SHAP and permutation importance also produced consistent global evidence that contract type, tenure and service-related variables were central to the model's predictions.

The main contribution is therefore not a new algorithm or a claim of universal predictive superiority. It is a reproducible evaluation framework that links model comparison, probability reliability, explanation consistency and cost-sensitive action. For telecom practice, the results suggest that churn scores should be calibrated, interpreted and aligned with campaign economics before they are used to prioritize customers. The findings remain limited by the use of one public cross-sectional dataset and normalized cost assumptions. Validation with longitudinal operator data, customer lifetime value, campaign response and retention-treatment outcomes is required before the framework can support real financial decisions.

AUTHOR CONTRIBUTIONS

Conceptualization, F.Z.; methodology, F.Z.; software, F.Z.; validation, F.Z. and E.V.; formal analysis, F.Z.; investigation, E.V.; resources, E.V.; data curation, F.Z.; writing, original draft preparation, F.Z.; writing, review and editing, F.Z.; visualization, F.Z. and E.V.; supervision, F.Z. All authors have read and agreed to the published version of the manuscript.

DATA AND CODE AVAILABILITY

The data set utilized in this study is publicly available from the repository for the Telco Customer Churn data set. Analysis scripts and results data set will be made available by the corresponding author upon request.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

1. 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,* **2024**, *17*(6), 231.
2. Poudel, S.S.P., Pokharel, S., Timilsina, M. Explaining Customer Churn Prediction in the Telecom Industry Using Tabular Machine Learning Models. *Machine Learning with Applications,* **2024**, *17*, 100567.
3. Lalwani, P., Mishra, M.K., Chadha, J.S., Sethi, P. Customer Churn Prediction System: A Machine Learning Approach. *Computing,* **2022**, 104, 271–294.
4. Ahmad, A.K., Jafar, A., Aljoumaa, K. Customer Churn Prediction in Telecom Using Machine Learning and Social Network Analysis in Big Data Platform. *Journal of Big Data*, **2019**, *6*, 28.
5. De Caigny, A., Coussement, K., De Bock, K.W. A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees. *European Journal of Operational Research,* **2018**, *269*(2), 760–772.
6. Niculescu-Mizil, A., Caruana, R. Predicting Good Probabilities with Supervised Learning. *In Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, **2005**, pp. 625–632.
7. Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W. Calibration: The Achilles Heel of Predictive Analytics. *BMC Medicine,* **2019**, *17*, 230.
8. Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B. New Insights into Churn Prediction in the Telecommunication Sector: A Profit-Driven Data Mining Approach. *European Journal of Operational Research,* **2012**, *218*(1), 211–229.
9. Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., Snoeck, M. Profit-Maximizing Logistic Model for Customer Churn Prediction Using Genetic Algorithms. *Swarm and Evolutionary Computation*, **2018**, *40*, 116–130.
10. Sheng, V.S., Ling, C.X. Thresholding for Making Classifiers Cost-Sensitive. In Proceedings of the Twenty-First National Conference on Artificial Intelligence, Boston, MA, USA, **2006**, pp. 476–481.
11. Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognition Letters,* **2006**, *27*(8), 861–874.
12. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, **2020**, *58*, 82–115.
13. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, **2020**, *2*, 56–67.

14. Sana, J.K., Abedin, M.Z., Rahman, M.S., Rahman, M.S. A Novel Customer Churn Prediction Model for the Telecommunication Industry Using Data Transformation Methods and Feature Selection. *PLOS ONE*, **2022**, *17*(12), e0278095.
15. Soleiman-garmabaki, O., Rezvani, M.H. Ensemble Classification Using Balanced Data to Predict Customer Churn: A Case Study on the Telecom Industry. *Multimedia Tools and Applications* **2024**, *83*, 44799–44831.
16. Ouf, S., Mahmoud, K.T., Abdel-Fattah, M.A. A Proposed Hybrid Framework to Improve the Accuracy of Customer Churn Prediction in Telecom Industry. *Journal of Big Data* **2024**, *11*, 70.
17. Sikri, A., Jameel, R., Idrees, S.M., Kaur, H. Enhancing Customer Retention in Telecom Industry with Machine Learning Driven Churn Prediction. *Scientific Reports* **2024**, *14*, 13097.
18. Devriendt, F., Moldovan, D., Verbeke, W. A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone toward the Development of Prescriptive Analytics. *Big Data*, **2018**, *6*(1), 13–41.
19. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R., Eds. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer: Cham, Switzerland, **2019**.
20. Ribeiro, M.T., Singh, S., Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, **2016**, pp. 1135–1144.
21. Aas, K., Jullum, M., Løland, A. Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values. *Artificial Intelligence* **2021**, *298*, 103502.
22. European Parliament; Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016: General Data Protection Regulation. *Off. J. Eur. Union* **2016**, *L119*, 1–88. Available online https://eur-lex.europa.eu/eli/reg/2016/679/oj (Access date 12 March 2026)
23. BlastChar. Telco Customer Churn [Data set]. *Kaggle*, **2018**. Available online https://www.kaggle.com/datasets/blastchar/telco-customer-churn (Access date 12 March 2026).
24. Breiman, L. Random Forests. *Machine Learning*, **2001**, *45*, 5–32.
25. Chen, T., Guestrin, C. XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, **2016**, pp. 785–794.
26. Davis, J., Goadrich, M. The Relationship between Precision-Recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, **2006**, pp. 233–240.
27. Chicco, D., Jurman, G. The Advantages of the Matthews Correlation Coefficient over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, **2020**, *21*, 6.
28. McNemar, Q. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika*, **1947**, *12*(2), 153–157.
29. Efron, B., Tibshirani, R.J. *An Introduction to the Bootstrap*. Chapman & Hall/CRC: New York, NY, USA, **1993**.