Improving Customer Churn Prediction Using Domain-Driven Feature Engineering, Resampling, and CatBoost with Explainability Extensions

Muhammad Jamaludin Nur
Faculty of Computer Science,
Universitas Dian Nuswantoro
Semarang 50131, Indonesia
111202214616@mhs.dinus.ac.id

De Rosal Ignatius Moses Setiadi *
Research Center for Quantum
Computing and Materials Informatics,
Universitas Dian Nuswantoro
Semarang 50131, Indonesia
moses@dsn.dinus.ac.id

Arnold Adimabua Ojugo
Department of Computer Science,
Federal University ofPetroleum
Resources
Effurun, Nigeria
ojugo.arnold@fupre.edu.ng

Minh T. Nguyen
Thai Nguyen University of
Technology, Thai Nguyen University,
Thai Nguyen, Vietnam
nguyentuanminh@tnut.edu.vn

***Abstract*— This study proposes an innovative pipeline for customer churn prediction in the telecommunications sector, integrating resampling, ensemble learning, and interpretability with fairness analysis to enhance prediction accuracy and provide actionable insights. Using the IBM Telco Customer Churn dataset (7,043 samples, 21 features, 26.54% churn), the study addresses class imbalance with SMOTE, introduces domain-specific features such as Engagement_Score, and optimizes ensemble models (CatBoost, XGBoost) via Bayesian tuning with Optuna and 5-fold cross-validation. The CatBoost model under the FE_SMOTE scenario achieved a ROC-AUC of 0.937 ± 0.010, accuracy of 0.862, and F1-score of 0.863, significantly improving churn detection performance. SHAP analysis identified key features, including InternetService_FiberOptic and tenure, while DICE-ML provided counterfactual explanations to guide retention strategies. Fairness evaluation revealed a 22.55% bias toward senior citizens, indicating a need for demographic adjustment. Overall, this pipeline provides a robust, interpretable, and equitable approach to churn prediction, enabling telecommunications companies to design data-driven interventions with potential scalability to other sectors.**

Keywords*— *Customer Churn Prediction; Ensemble Learning; Feature Engineering; Interpretability; Fairness Analysis

I. INTRODUCTION

High customer churn rates in the telecommunications sector can significantly reduce revenue and damage the company's reputation [1]. Effective retention strategies, including service quality improvement, personalized offerings, loyalty programs, proactive communication, and customer data analytics, have proven successful in mitigating churn and increasing customer lifetime value [2]–[4]. By modeling customer behavior using historical data, churn prediction becomes feasible through the analysis of customer attributes. Various machine learning (ML) algorithms have been employed for churn prediction, such as Decision Trees, Gradient Boosting Machines (GBM), XGBoost, LightGBM, CatBoost, Random Forests (RF), Logistic Regression (LR), Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP), and AdaBoost. GBM-based models, such as XGBoost, LightGBM, and CatBoost, exhibit strong performance on complex tabular data but are prone to overfitting, necessitating careful hyperparameter tuning [5]. RF is valued for stability [6], while LR offers interpretability but underperforms on nonlinear or imbalanced data [7]. SVMs are well-suited for high-dimensional data, but they are computationally intensive and sensitive to hyperparameters [8]. MLPs can capture complex patterns, and AdaBoost enhances accuracy through weak learner iteration, although it is sensitive to noise [9], [10].

Class imbalance remains a significant challenge in churn data, resulting in biased models that fail to accurately detect the minority class. Techniques such as SMOTE, SMOTEENN, RandomUnderSampler, and ADASYN have been used to address this [11]. This study adopts SMOTE as the primary resampling technique due to its effectiveness in balancing class distribution [12], [13]. Despite the success of models like XGBoost and LightGBM, they struggle with class imbalance, reducing minority class prediction accuracy [11], [14]. Additionally, they lack intrinsic interpretability, making it hard for stakeholders to understand churn drivers without auxiliary tools [15], [16]. The integration of interpretability tools, such as SHAP or DICE-ML, can be limited by technical constraints. To overcome these issues, this study proposes features such as EngagementScore and ChurnRiskScore to enhance predictive capacity [17], [18]. It applies Bayesian optimization with Optuna for efficient hyperparameter tuning [19], [20], and integrates SHAP for global and local interpretability [21]. Furthermore, counterfactual explanations complement SHAP by suggesting the smallest changes needed to flip a prediction, providing clear and actionable recommendations. Lastly, fairness analysis using demographic parity and equalized odds ensures equitable predictions, particularly for senior citizens [22], [23].

This study aims to develop a churn prediction framework that combines high accuracy, interpretability, and fairness. By integrating feature engineering, optimization, explainability (SHAP and counterfactual explanations), and fairness analysis, the proposed pipeline addresses the gaps in prior studies that have focused solely on accuracy. The primary contribution lies in the end-to-end integration of ensemble model optimization, interpretable machine learning, and fairness evaluation, delivering a more transparent, ethical, and applicable solution for churn prediction in the telecommunications industry.

II. RELATED WORK

Customer churn prediction remains a central problem in machine learning due to its direct impact on customer

retention and business sustainability. Traditional models such as logistic regression and decision trees offer interpretability but often struggle with nonlinear patterns and imbalanced data [24], [25]. Recent studies increasingly adopt ensemble learning methods like XGBoost, LightGBM, and CatBoost, which achieve higher performance on complex tabular data [24], [26]–[28]. However, these models require careful hyperparameter tuning to prevent overfitting and ensure generalization. Studies using the IBM Telco Customer Churn dataset report varied performance. Gradient boosting, combined with SHAP analysis, achieved an AUC of 0.86 [29], while an ensemble-fusion approach reached an AUC of 0.85 and an F1-score of 0.80 [29]. Other research applying XGBoost without resampling or interpretability tools showed a lower AUC of around 0.75 [25]. Additionally, clustering-based segmentation approaches provide descriptive insights but lack predictive performance metrics [30], [31].

Despite these advances, significant research gaps remain. Many studies focus solely on improving accuracy without addressing class imbalance using techniques like SMOTE, limiting minority class detection [25], [29]. Furthermore, interpretability tools such as SHAP and counterfactual explanation frameworks like DICE-ML are rarely integrated, which reduces actionable insights for decision-making [24], [32]. Fairness evaluation is also often neglected, raising ethical concerns regarding demographic bias [29].

To address these gaps, this study proposes a comprehensive pipeline that combines domain-informed feature engineering, SMOTE resampling, Bayesian hyperparameter optimization, and integrated interpretability with fairness analysis. This approach aims to produce robust, explainable, and equitable churn predictions to support effective data-driven retention strategies.

III. Proposed Method

The methodological framework of this study is illustrated in Figure 1, encompassing dataset preparation, feature engineering, resampling, model optimization, evaluation, interpretability, fairness, and counterfactual explanation.

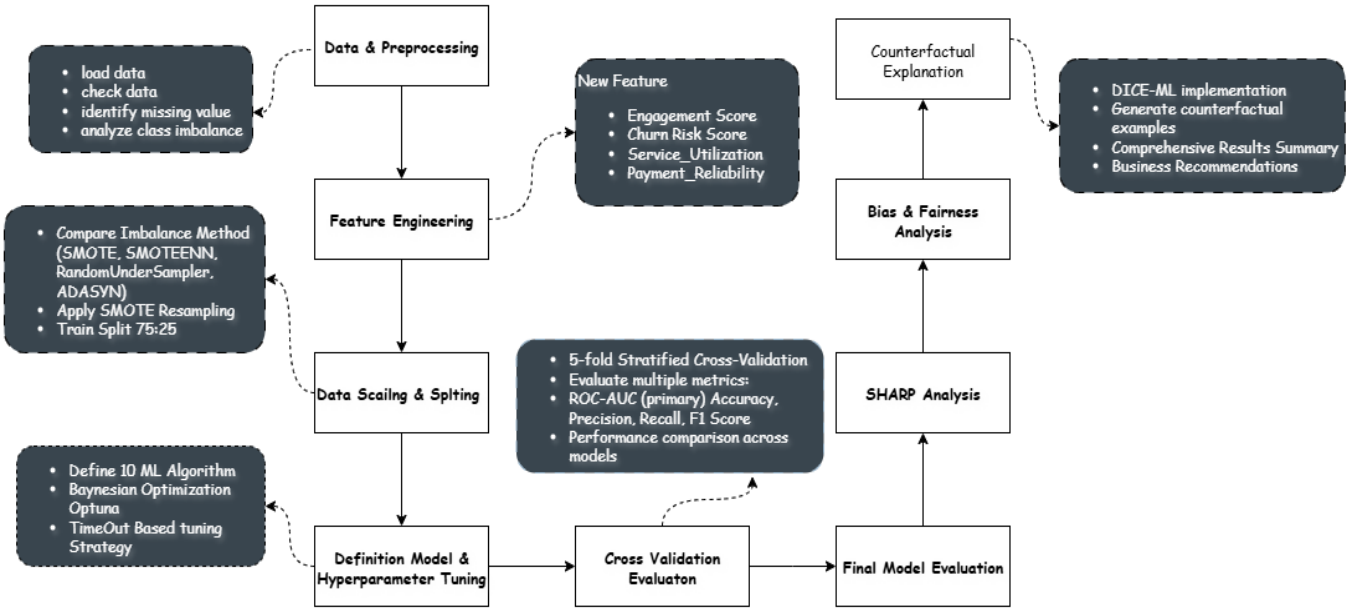

Fig. 1. Top-level methodology of research

A. Dataset and Preprocessing

The dataset used is the IBM Telco Customer Churn dataset from Kaggle, which contains 7,043 samples with 21 features, including the target variable "churn". Preprocessing steps involved normalizing column names, converting TotalCharges from an object to a numeric type by imputing 11 blank entries, and removing customerID as it is non-predictive. Class distribution analysis revealed a significant imbalance, with 26.54% of customers experiencing churn and 73.46% experiencing non-churn (Fig. 2, blue bar), indicating the need for resampling.

B. Feature Engineering

To enhance predictive performance, several domain-informed features were engineered:

- The Engagement Score, representing the intensity of customer interaction, is calculated using Eq. (1).

$$\text{EngagementScore} = \frac{\text{tenture} \times \text{Monthlycharges}}{\text{TotalCharges} + 1} \quad (1)$$

where +1 prevents division by zero.

- Churn Risk Score captures high churn risk for customers with tenure below 12 months:

$$\text{ChurnRiskScore} = \begin{cases} \text{MonthlyCharge}, & \text{if tenture} < 12 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

- Service Utilization sums all binary service indicators marked “Yes” to reflect multi-service adoption.

$$\text{ServiceUtilization} = \sum_{i=1}^{n} 1_{\{x_i=\text{Yes}\}}, x_i \text{ServiceColumns} \quad (3)$$

- Payment Reliability is a binary indicator, where 1 represents the use of automatic payment methods (bank transfer or credit card automatic), and 0 otherwise.

$$PaymentReliability = \begin{cases} 1, if\ PaymentMethod\ is\ automatic \\ 0, Other \end{cases} \quad (4)$$

C. Data Scaling and Splitting

All numerical features were standardized using StandardScaler to ensure consistent distributions. Various resampling methods were compared, including SMOTE, SMOTEENN, ADASYN, and RandomUnderSampler. Based on exploratory results, SMOTE was selected to effectively balance the dataset (Fig. 2, red bar). The dataset was then split into training (75%) and testing (25%) sets using stratified sampling to maintain class proportion.

D. Model Definition and Hyperparameter Tuning

Ten classification and ensemble models were implemented: Decision Tree, Random Forest, Logistic Regression, Support Vector Classifier (SVC), Neural Network, AdaBoost, Gradient Boosting Machine (GBM), XGBoost, LightGBM, and CatBoost. Among them, CatBoost offers a unique advantage by natively handling categorical features and utilizing ordered boosting, which helps reduce overfitting and improve generalization, especially valuable for imbalanced tabular data, such as churn prediction [33]. Bayesian optimization with Optuna was employed for hyperparameter tuning. Table 2 summarizes the parameter settings.

TABLE I. SEARCH RANGES OF HYPERPARAMETERS USED FOR TUNING THE CHURN PREDICTION MODELS

Model	Hyperparameters	Range Value
DecisionTree	max_depth, min_samples_split	3-7, 2-5
GBM	n_estimators, learning_rate, max_depth	100-200, 0.01-0.1, 3-5
XGBoost	n_estimators, learning_rate, max_depth	100-200, 0.01-0.1, 3-5
RandomForest	n_estimators, max_depth, min_samples_split	100-200, 10-None, 2-5
Logistic Regression	C, penalty, solver	0.1-10, l1/l2, liblinear
SVC	C, kernel	0.1-1, rbf/linear
Neural Network	hidden_layer_sizes, alpha	(50,)/(100,)/(50,5) , 0.0001-0.001
AdaBoost	n_estimators, learning_rate	50-100, 0.1-1.0
LightGBM	n_estimators, learning_rate, max_depth	100-200, 0.01-0.1, 3-5
CatBoost	iterations, learning_rate, depth	100-200, 0.01-0.1, 3-5
DecisionTree	max_depth, min_samples_split	3-7, 2-5
GBM	n_estimators, learning_rate, max_depth	100-200, 0.01-0.1, 3-5
XGBoost	n_estimators, learning_rate, max_depth	100-200, 0.01-0.1, 3-5
RandomForest	n_estimators, max_depth, min_samples_split	100-200, 10-None, 2-5

E. Cross-Validation Evaluation

Model performance was evaluated using stratified 5-fold cross-validation to ensure generalizability across data splits. The evaluation employed standard classification metrics, including ROC-AUC as the primary metric, along with accuracy, precision, recall, and F1-score to comprehensively assess classification effectiveness. Additionally, confusion matrices were generated to provide detailed insights into classification outcomes, including true positives, true negatives, false positives, and false negatives, thereby enabling a better understanding of the model's performance on each class.

F. Final Model Evaluation and Interpretability

The best-performing model, CatBoost under FE_SMOTE, was retrained with optimal parameters and evaluated on the test set using confusion matrices, ROC curves, and calibration plots (Figure 4). SHAP analysis (Figure 6) identified InternetService_FiberOptic, tenure, and contract type as key influential features.

G. Fairness Analysis and Counterfactual Explanation

A fairness evaluation was conducted to assess potential demographic bias in model predictions, with a specific focus on the SeniorCitizen attribute. This analysis employed metrics such as demographic parity difference and equal opportunity difference to determine whether the model exhibited unequal treatment or outcomes across customer groups. The fairness evaluation process aimed to ensure that the predictive pipeline aligned with ethical AI standards by identifying and mitigating biases that could negatively impact certain demographics during real-world deployment.

IV. RESULTS AND DISCUSSION

A. Data Balancing

This study utilizes the IBM Telco Customer Churn dataset from Kaggle, which comprises 7,043 samples and 21 features after preprocessing. Analysis revealed a significant class imbalance, with only 26.54% of customers in the "Yes" (churn) class (1,869 instances) and 73.46% in the "No" (non-churn) class (5,174 instances), as shown in Fig. 2 (blue bar). To address this, the SMOTE was applied, resulting in a balanced distribution where both classes contained 5,174 instances each, as illustrated in Fig. 2 (red bar). This balancing ensures improved model performance in detecting minority class churn cases during training.

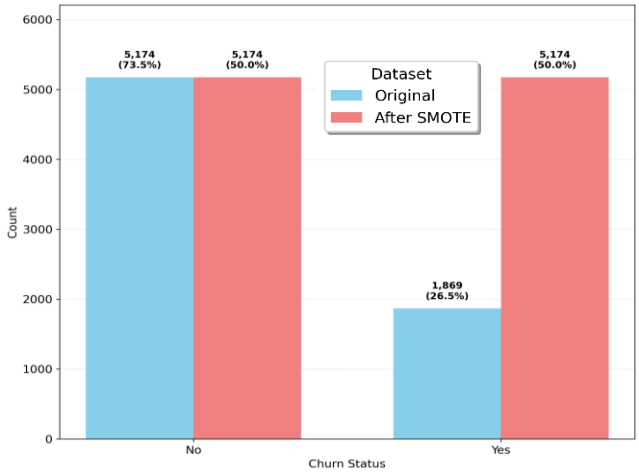

Fig. 2. Target variable distribution before and after SMOTE

B. Evaluated Model Comparison

To assess the predictive performance of the proposed models, this study employed 5-fold cross-validation under four experimental settings: without feature engineering (No_FE), with feature engineering only (FE_Only), with class balancing using SMOTE only (SMOTE_Only), and with both feature engineering and SMOTE (FE_SMOTE). Evaluation was based on five standard classification metrics: ROC-AUC, accuracy, precision, recall, and F1-score. The results are summarized in Table 2.

TABLE II. COMPARATIVE PERFORMANCE OF MODELS ACROSS DIFFERENT EXPERIMENTAL SETTINGS

Case	Model	ROC-AUC	Accuracy	Precision	Recall	F1-Score
No_FE	AdaBoost	0.840±0.022	0.800±0.011	0.656±0.030	0.522±0.027	0.581±0.024
	CatBoost	0.822±0.014	0.783±0.010	0.615±0.027	0.491±0.021	0.545±0.017
	Decision Tree	0.805±0.012	0.779±0.007	0.602±0.021	0.494±0.015	0.542±0.013
	GBM	0.827±0.012	0.784±0.013	0.619±0.031	0.487±0.032	0.545±0.030
	LightGBM	0.836±0.017	0.793±0.008	0.637±0.021	0.512±0.016	0.568±0.017
	Logistic Regression	**0.844±0.016**	**0.806±0.011**	0.661±0.023	0.553±0.042	**0.602±0.030**
	Neural Network	0.836±0.020	0.802±0.015	0.647±0.030	**0.558±0.041**	0.599±0.035
	Random Forest	0.824±0.013	0.787±0.010	0.631±0.024	0.477±0.030	0.543±0.025
	SVC	0.800±0.024	**0.806±0.011**	**0.679±0.030**	0.509±0.030	0.581±0.026
	XGBoost	0.838±0.012	0.795±0.005	0.643±0.017	0.514±0.023	0.571±0.014
FE Only	Ada Boost	0.839±0.019	0.800±0.015	0.656±0.041	0.521±0.027	0.580±0.030
	Cat Boost	0.822±0.010	0.790±0.002	0.631±0.006	0.502±0.011	0.559±0.006
	Decision Tree	0.799±0.015	0.778±0.008	0.600±0.020	0.491±0.012	0.540±0.015
	GBM	0.830±0.013	0.792±0.011	0.639±0.032	0.500±0.015	0.561±0.019
	Light GBM	0.836±0.014	0.796±0.007	0.645±0.024	0.515±0.002	0.572±0.009
	Logistic Regression	**0.845±0.017**	**0.806±0.011**	0.670±0.028	0.533±0.037	0.593±0.028
	Neural Network	0.834±0.018	0.801±0.015	0.654±0.038	**0.535±0.024**	**0.588±0.028**
	Random Forest	0.828±0.009	0.791±0.006	0.641±0.018	0.481±0.015	0.549±0.014
	SVC	0.796±0.027	0.800±0.016	**0.682±0.047**	0.463±0.038	0.551±0.039
	XGBoost	0.838±0.012	0.795±0.007	0.644±0.021	0.510±0.022	0.569±0.015
SMOTE Only	AdaBoost	0.911±0.011	0.823±0.016	0.804±0.015	0.855±0.017	0.829±0.015
	CatBoost	0.932±0.011	**0.856±0.017**	**0.853±0.017**	**0.860±0.022**	**0.857±0.017**
	Decision Tree	0.859±0.014	0.786±0.012	0.772±0.020	0.814±0.011	0.792±0.009
	GBM	0.932±0.009	0.853±0.015	0.852±0.019	0.855±0.019	0.853±0.014
	LightGBM	**0.933±0.009**	0.849±0.014	0.847±0.020	0.854±0.022	0.850±0.014
	Logistic Regression	0.857±0.016	0.777±0.017	0.758±0.023	0.816±0.013	0.785±0.014
	Neural Network	0.867±0.015	0.783±0.015	0.774±0.023	0.800±0.032	0.786±0.015
	Random Forest	0.917±0.013	0.845±0.013	0.840±0.014	0.854±0.020	0.846±0.013
	SVC	0.885±0.015	0.800±0.018	0.781±0.021	0.834±0.016	0.807±0.017
	XGBoost	**0.933±0.009**	0.852±0.012	0.847±0.016	0.859±0.017	0.853±0.012
FE_SMOTE	AdaBoost	0.916±0.012	0.830±0.019	0.807±0.021	**0.868±0.018**	0.836±0.017
	CatBoost	**0.937±0.010**	**0.862±0.012**	**0.858±0.013**	**0.868±0.019**	**0.863±0.013**
	Decision Tree	0.864±0.011	0.792±0.016	0.770±0.027	0.836±0.017	0.801±0.010
	GBM	0.934±0.008	0.856±0.011	0.855±0.011	0.858±0.022	0.856±0.012
	LightGBM	0.936±0.007	0.852±0.011	0.851±0.015	0.854±0.017	0.853±0.011
	Logistic Regression	0.856±0.015	0.774±0.018	0.755±0.023	0.813±0.016	0.783±0.015
	Neura lNetwork	0.870±0.009	0.792±0.012	0.766±0.014	0.840±0.034	0.801±0.014
	Random Forest	0.924±0.012	0.850±0.012	0.842±0.016	0.861±0.021	0.851±0.012
	SVC	0.887±0.013	0.796±0.016	0.778±0.019	0.830±0.019	0.803±0.015
	XGBoost	0.934±0.008	0.854±0.013	0.850±0.013	0.862±0.022	0.855±0.013

In the No_FE and FE_Only scenarios, Logistic Regression achieved the highest ROC-AUC scores of 0.844 and 0.845, respectively, demonstrating moderate discriminative capability. However, its relatively low recall and F1-score indicated limited effectiveness in correctly identifying churned customers. In the SMOTE_Only setting, applying class balancing alone led to substantial performance improvements across models, particularly for ensemble-based classifiers. XGBoost achieved the highest performance in this scenario, with a ROC-AUC of 0.933 ± 0.009, an accuracy of 0.852 ± 0.012, and an F1-score of 0.853 ± 0.012, indicating stable results across validation folds with low standard deviations.

Further improvements were observed under the FE_SMOTE scenario, where the integration of domain-

informed feature engineering with SMOTE resampling yielded the best outcomes. CatBoost outperformed all other classifiers, achieving a ROC-AUC of 0.937 ± 0.010, accuracy of 0.862 ± 0.012, and F1-score of 0.863 ± 0.013, highlighting its superior ability to detect minority class instances while maintaining stability. Other ensemble models, such as XGBoost and LightGBM, also showed enhanced performance under this combined setting. Overall, these results highlight the importance of addressing class imbalance using resampling techniques, such as SMOTE, particularly when combined with engineered features. While accuracy provides a general measure of model performance, metrics such as recall and F1-score offer deeper insights into the effectiveness of models under imbalanced conditions. Fig. 3 summarizes these findings by illustrating the best-performing model for each scenario, confirming that ensemble methods, especially CatBoost under FE_SMOTE, consistently deliver superior predictive accuracy and robustness for churn prediction tasks.

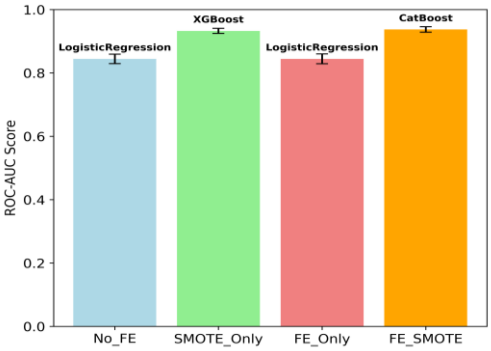

Fig. 3. Best model per scenario by ROC-AUC: Logistic Regression (No_FE, FE_Only), XGBoost (SMOTE_Only), CatBoost (FE_SMOTE).

C. *Final Model Testing*

The final evaluation on the held-out test set used the best-performing model, CatBoost under the FE_SMOTE scenario. It achieved a ROC-AUC of 0.9391 and a precision–recall AUC of 0.9426, demonstrating strong discriminative capability and effective minority class detection. Fig. 4 shows the precision–recall curve, illustrating high precision across various recall thresholds, which confirms the model's robustness and readiness for practical deployment in churn prediction tasks. Additionally, a Confucian matrix is presented in Fig. 5.

D. *SHAP for the CatBoost Final Model*

Figure 6 presents the SHAP summary plot of the final model, specifically the retrained CatBoost model with its optimal hyperparameter configuration. Each point in the plot represents the contribution of a feature to the churn prediction for a single observation, with color encoding the original feature value (red indicating high values and blue indicating low values). Features are sorted by the mean absolute SHAP value, indicating their overall impact on the model's output.

The feature InternetService_Fiber optic emerged as the most influential predictor, with high values significantly influencing the model's prediction of churn. In contrast, features such as tenure, Contract_Two year, and Contract_One year contributed negatively to the churn likelihood, suggesting that long-term customers with extended contracts tend to be more loyal. Other important contributors included Churn_Risk_Score, MonthlyCharges, and the Electronic check payment method, all of which were associated with increased churn probability in various combinations of feature values.

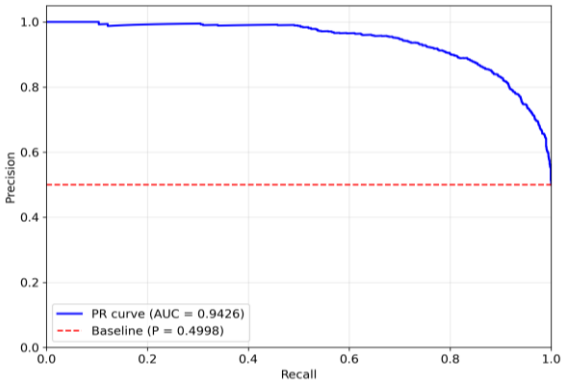

Fig. 4. Precision–Recall curve of CatBoost under FE + SMOTE.

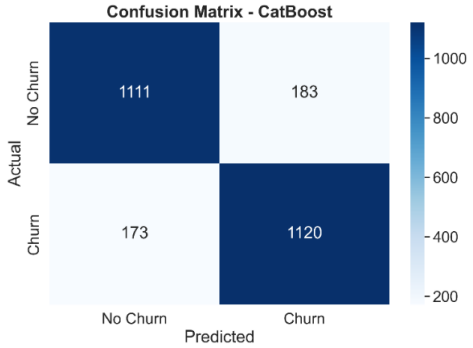

Fig. 5. Confusion matrix of of CatBoost under FE_SMOTE

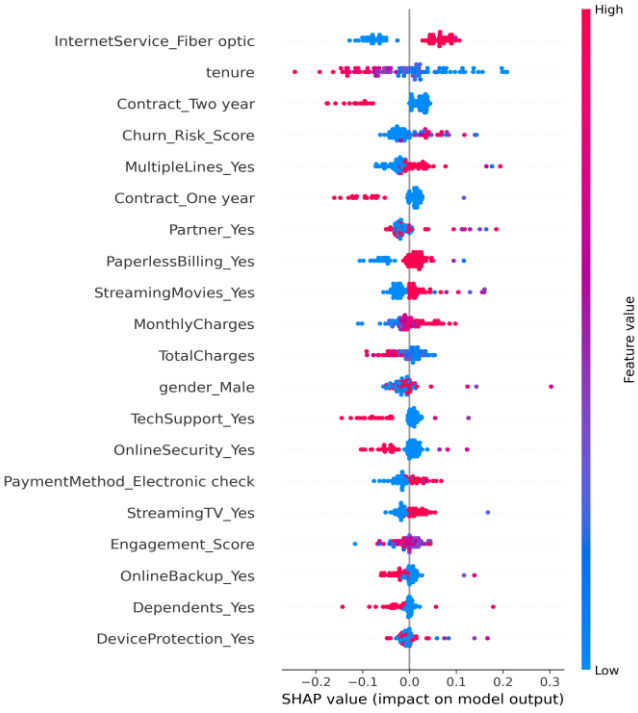

Fig. 6. SHAP summary plot

These findings highlight that the final model captures not only contractual and behavioral dimensions but also risk indicators and payment preferences. This information can be leveraged to design more targeted and data-driven churn intervention strategies, such as offering incentives to high-risk

customers based on their SHAP-based profile. ROC-AUC helps prioritize high-risk customers for retention efforts, while the F1-score ensures a balance between retaining actual churners and avoiding unnecessary interventions [34].

E. Fairness Evaluation Based on Senior Citizen Attribute

Fairness evaluation was conducted with a focus on the SeniorCitizen attribute. The results indicate that the positive prediction rate for senior customers was 46.85%, substantially higher than that of non-senior customers at 24.30%, resulting in a gap of 22.55 percentage points. This discrepancy suggests a potential bias in churn prediction toward the elderly group. In terms of Equal Opportunity, the True Positive Rate (TPR) for senior customers reached 96.01%, compared to 87.94% for non-seniors, yielding a difference of 8.07%. This indicates that the model is more effective at detecting churn among senior customers. Future work will evaluate other demographic features such as gender, partner, and dependents to provide a broader fairness perspective.

F. Counterfactual Explanation Analysis

To enhance the interpretability and practical utility of the final model, this study implemented counterfactual explanations using the Diverse Counterfactual Explanations for ML (DiCE-ML) library. This approach identifies minimal changes to input features needed to alter a customer's prediction from churn to non-churn, thereby providing actionable insights for retention strategies. Counterfactuals were generated for 30 randomly selected test instances, achieving a 100% success rate, meaning that every instance received at least one valid alternative prediction. Fig. 7 summarizes the quality metrics of these counterfactuals. The results showed a mean proximity of 1.838, indicating that the generated counterfactuals remained close to the original instances in feature space, which supports the realism and feasibility of the interventions.

The mean sparsity was 1.578, indicating that, on average, only one to two features required modification, thereby enhancing interpretability. Diversity had an average score of 2.814, reflecting a variety of alternative changes across counterfactuals. Meanwhile, validity recorded a mean of 0.200, suggesting that although counterfactuals were successfully generated for all instances, the model's decision boundaries were relatively rigid, requiring substantial feature changes to flip predictions. This low validity is noted as a limitation, and future work may explore constraint-based or generative approaches to enhance the realism and feasibility of the suggested interventions.

G. Feature-Level Insights from Counterfactual Explanations

Figure 8 shows the features most frequently modified in counterfactual explanations. Engagement_Score had the highest average change, followed by TotalCharges, MultipleLines_Yes, and PaperlessBilling_Yes, indicating their significant influence on model predictions. In contrast, features such as Churn_Risk_Score and various service-related variables were rarely altered, suggesting a lower impact. These insights identify actionable targets for effective intervention strategies to reduce churn.

H. Benchmark Comparison with Prior Studies

To contextualize the performance of the proposed model, we compared it with selected prior studies on the IBM Telco Customer Churn dataset. Table 3 summarizes the key metrics (ROC AUC, F1 Score) from these studies. As shown, our approach outperforms previous methods in both AUC and F1-Score, highlighting the effectiveness of combining feature engineering, SMOTE resampling, and CatBoost optimization.

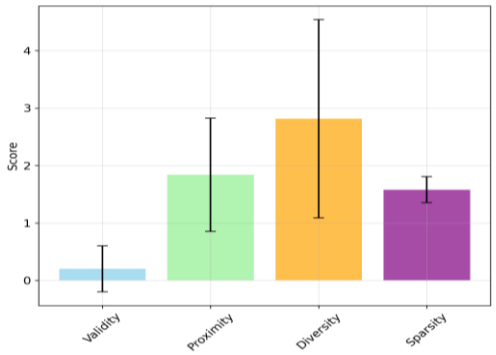

Fig. 7. Summary of counterfactual quality metrics: validity, proximity, diversity, and sparsity.

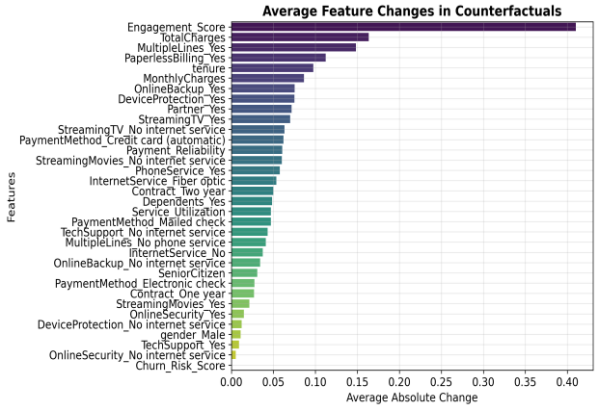

Fig. 8. Ranked Average Feature Changes in Counterfactual Instances

TABLE III. PERFORMANCE COMPARISON ON TELCOM CUSTOMER CHURN

Ref	Model	ROC-AUC	F1-score
[24]	GBM	86.00	60.00
[25]	XGBoost	74.69	74.22
[29]	Ensemble-Fusion	91.00	96.96
Ours	CatBoost	93.70	86.30

V. CONCLUSIONS

This study proposed a comprehensive and reproducible pipeline for customer churn prediction by integrating domain-driven feature engineering, SMOTE-based class balancing, Optuna hyperparameter tuning, and model interpretability via SHAP and counterfactual explanations. The approach outperformed prior benchmarks on the IBM Telco Customer Churn dataset, achieving a ROC-AUC of 0.937 and an F1-score of 0.863 with CatBoost under the FE_SMOTE scenario.

Key contributions include combining interpretability and fairness evaluation within the predictive pipeline. However, limitations remain, such as the relatively small dataset size, a 22.55% fairness disparity for senior customers, low counterfactual validity (0.200), and restricted hyperparameter tuning time budgets. Future work should explore larger and more diverse datasets, incorporate behavioral and contextual features, conduct fairness analyses on multiple sensitive

attributes, integrate the framework into real-time systems, and enhance counterfactual explanations for personalized decision support. Additionally, robustness testing through sensitivity analysis of hyperparameters and noise injection will be explored further to assess model generalization and stability under data uncertainty. These directions aim to enhance the robustness, transparency, and practical applicability of churn prediction in real-world business contexts.

REFERENCES

[1] A. Prashanthan *et al.*, "RetenNet: A Deployable Machine Learning Pipeline with Explainable AI and Prescriptive Optimization for Customer Churn Management," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 2, pp. 182–201, Jun. 2025, doi: 10.62411/faith.3048-3719-110.

[2] M. A. Hajar *et al.*, "The Effect of Value Innovation in the Superior Performance and Sustainable Growth of Telecommunications Sector: Mediation Effect of Customer Satisfaction and Loyalty," *Sustainability*, vol. 14, no. 10, p. 6342, May 2022, doi: 10.3390/su14106342.

[3] R. E. Ako *et al.*, "Effects of Data Resampling on Predicting Customer Churn via a Comparative Tree-based Random Forest and XGBoost," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 86–101, Jun. 2024, doi: 10.62411/jcta.10562.

[4] I. Puspitasari *et al.*, "Investigating the role of utilitarian and hedonic goals in characterizing customer loyalty in E-marketplaces," *Heliyon*, vol. 9, no. 8, p. e19193, 2023, doi: 10.1016/j.heliyon.2023.e19193.

[5] F. Zhu, X. Wu, Y. Lu, and J. Huang, "Strength Estimation and Feature Interaction of Carbon Nanotubes-Modified Concrete Using Artificial Intelligence-Based Boosting Ensembles," *Buildings*, vol. 14, no. 1, p. 134, Jan. 2024, doi: 10.3390/buildings14010134.

[6] S. Han, H. Kim, and Y.-S. Lee, "Double random forest," *Mach. Learn.*, vol. 109, no. 8, pp. 1569–1586, Aug. 2020, doi: 10.1007/s10994-020-05889-1.

[7] S. Han and H. Jung, "NATE: Non-pArameTric approach for Explainable credit scoring on imbalanced class," *PLoS One*, vol. 19, no. 12, p. e0316454, Dec. 2024, doi: 10.1371/journal.pone.0316454.

[8] Y. Rimal, N. Sharma, S. Paudel, A. Alsadoon, M. P. Koirala, and S. Gill, "Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy," *Sci. Rep.*, vol. 15, no. 1, p. 13444, Apr. 2025, doi: 10.1038/s41598-025-93675-1.

[9] H. Xu and H. Yuan, "An SVM-Based AdaBoost Cascade Classifier for Sonar Image," *IEEE Access*, vol. 8, pp. 115857–115864, 2020, doi: 10.1109/ACCESS.2020.3004473.

[10] W. Huang, Y. Li, J. Tang, and L. Qian, "Fault Diagnosis Methods for an Artillery Loading System Driving Motor in Complex Noisy Environments," *Sensors*, vol. 24, no. 3, p. 847, Jan. 2024, doi: 10.3390/s24030847.

[11] M. Imani and H. R. Arabnia, "Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis," *Preprint*. Aug. 21, 2023. doi: 10.20944/preprints202308.1478.v1.

[12] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00943-4.

[13] T. Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, and Y. Lin, "A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare," *BioData Min.*, vol. 16, no. 1, p. 15, Apr. 2023, doi: 10.1186/s13040-023-00330-4.

[14] K. Peng and Y. Peng, "Research on Telecom Customer Churn Prediction Based on GA-XGBoost and SHAP," *J. Comput. Commun.*, vol. 10, no. 11, pp. 107–120, 2022, doi: 10.4236/jcc.2022.1011008.

[15] K. Peng, Y. Peng, and W. Li, "Research on customer churn prediction and model interpretability analysis," *PLoS One*, vol. 18, no. 12, p. e0289724, Dec. 2023, doi: 10.1371/journal.pone.0289724.

[16] X. Zhu, Q. Chu, X. Song, P. Hu, and L. Peng, "Explainable prediction of loan default based on machine learning models," *Data Sci. Manag.*, vol. 6, no. 3, pp. 123–133, Sep. 2023, doi: 10.1016/j.dsm.2023.04.003.

[17] M. Z. Abedin, P. Hajek, T. Sharif, M. S. Satu, and M. I. Khan, "Modelling bank customer behaviour using feature engineering and classification techniques," *Res. Int. Bus. Financ.*, vol. 65, p. 101913, Apr. 2023, doi: 10.1016/j.ribaf.2023.101913.

[18] J. K. Sana, M. Z. Abedin, M. S. Rahman, and M. S. Rahman, "A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection," *PLoS One*, vol. 17, no. 12, 2022, doi: 10.1371/journal.pone.0278095.

[19] M. Wojciuk, Z. Swiderska-Chadaj, K. Siwek, and A. Gertych, "Improving classification accuracy of fine-tuned CNN models: Impact of hyperparameter optimization," *Heliyon*, vol. 10, no. 5, p. e26586, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26586.

[20] F. Abbas *et al.*, "Optimizing Machine Learning Algorithms for Landslide Susceptibility Mapping along the Karakoram Highway, Gilgit Baltistan, Pakistan: A Comparative Study of Baseline, Bayesian, and Metaheuristic Hyperparameter Optimization Techniques," *Sensors*, vol. 23, no. 15, p. 6843, Aug. 2023, doi: 10.3390/s23156843.

[21] T. R. Noviandy, G. M. Idroes, and I. Hardi, "An Interpretable Machine Learning Strategy for Antimalarial Drug Discovery with LightGBM and SHAP," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 84–95, Aug. 2024, doi: 10.62411/faith.2024-16.

[22] M. Pal, S. Pokhriyal, S. Sikdar, and N. Ganguly, "Ensuring generalized fairness in batch classification," *Sci. Rep.*, vol. 13, no. 1, p. 18892, Nov. 2023, doi: 10.1038/s41598-023-45943-1.

[23] C. Castro and M. Loi, "The Fair Chances in Algorithmic Fairness: A Response to Holm," *Res Publica*, vol. 29, no. 2, pp. 331–337, Jun. 2023, doi: 10.1007/s11158-022-09570-3.

[24] S. S. Poudel, S. Pokharel, and M. Timilsina, "Explaining customer churn prediction in telecom industry using tabular machine learning models," *Mach. Learn. with Appl.*, vol. 17, p. 100567, Sep. 2024, doi: 10.1016/j.mlwa.2024.100567.

[25] T. A. R. Akbar and C. Apriono, "Machine Learning Predictive Models Analysis on Telecommunications Service Churn Rate," *Green Intell. Syst. Appl.*, vol. 3, no. 1, pp. 22–34, Jun. 2023, doi: 10.53623/gisa.v3i1.249.

[26] Z. S. Dhahir, "A Hybrid Approach for Efficient DDoS Detection in Network Traffic Using CBLOF-Based Feature Engineering and XGBoost," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 174–190, Sep. 2024, doi: 10.62411/faith.2024-33.

[27] M. A. Shaikhsurab and P. Magadum, "Enhancing Customer Churn Prediction in Telecommunications: An Adaptive Ensemble Learning Approach," *ArXiv*. Aug. 29, 2024.

[28] D. R. I. M. Setiadi, A. R. Muslikh, S. W. Iriananda, W. Warto, J. Gondohanindijo, and A. A. Ojugo, "Outlier Detection Using Gaussian Mixture Model Clustering to Optimize XGBoost for Credit Approval Prediction," *J. Comput. Theor. Appl.*, vol. 2, no. 2, pp. 244–255, Nov. 2024, doi: 10.62411/jcta.11638.

[29] C. He and C. H. Q. Ding, "A novel classification algorithm for customer churn prediction based on hybrid Ensemble-Fusion model," *Sci. Rep.*, vol. 14, no. 1, p. 20179, Aug. 2024, doi: 10.1038/s41598-024-71168-x.

[30] M. Pejić Bach, J. Pivar, and B. Jaković, "Churn Management in Telecommunications: Hybrid Approach Using Cluster Analysis and Decision Trees," *J. Risk Financ. Manag.*, vol. 14, no. 11, p. 544, Nov. 2021, doi: 10.3390/jrfm14110544.

[31] A. Prashanthan, "An Integrated Framework for Optimizing Customer Retention Budget using Clustering, Classification, and Mathematical Optimization," *J. Comput. Theor. Appl.*, vol. 3, no. 1, pp. 45–63, Jul. 2025, doi: 10.62411/jcta.13194.

[32] S. Adamu, A. Iorliam, and Ö. Asilkan, "Exploring Explainability in Multi-Category Electronic Markets: A Comparison of Machine Learning and Deep Learning Approaches," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 4, pp. 440–454, Mar. 2025, doi: 10.62411/faith.3048-3719-58.

[33] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. Big Data*, vol. 7, no. 1, p. 94, Dec. 2020, doi: 10.1186/s40537-020-00369-8.

[34] J. R. Dias and N. Antonio, "Predicting customer churn using machine learning: A case study in the software industry," *J. Mark. Anal.*, vol. 13, no. 1, pp. 111–127, Mar. 2025, doi:10.1057/s41270-023-00269-9.