Customer Churn Prediction in B2B SaaS

Minu Ashika A[1], Azhagu Meena P[2], Dr. Saranya[3], Dr. J. Arokia Renjit[4]
[1, 2, 3, 4] School of Engineering and Technology
[1, 2, 3, 4]Jeppiaar University, Chennai, India

Abstract- *Customer retention has become a major concern for subscription-based B2B Software as a Service (SaaS) companies because long-term contracts directly influence revenue stability. Even a small increase in churn rate can result in noticeable financial loss for service providers. In this study, machine learning techniques are applied to analyze and predict customer churn using structured enterprise data. The proposed framework includes data preprocessing, handling class imbalance using the Synthetic Minority Oversampling Technique (SMOTE), and evaluating several ensemble learning models including Random Forest, AdaBoost, CatBoost, and LightGBM. The models are assessed using commonly used classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Special attention is given to recall since identifying potential churn customers is particularly important for business decision-making. Among the evaluated approaches, LightGBM demonstrated stable and comparatively better performance across multiple metrics. To improve transparency, SHAP analysis is used to interpret feature contributions and identify factors influencing churn behavior. The trained model is further integrated into a Streamlit-based dashboard that allows real-time predictions and batch processing of customer data, helping organizations monitor churn risk and plan proactive retention strategies.*

Keywords: B2B SaaS, Customer Churn Prediction, Explainable Artificial Intelligence, LightGBM, SMOTE

I. INTRODUCTION

Customer retention has become an important challenge for organizations operating under the Software as a Service (SaaS) business model. Over the past decade, SaaS platforms have gained widespread adoption across industries, enabling companies to manage business processes such as analytics, customer management, and enterprise operations through cloud-based systems. In B2B environments, these services are usually delivered through long-term subscription contracts, which makes customer retention essential for maintaining stable revenue streams.

Customer churn refers to the situation in which a client stops using a service or cancels their subscription. In B2B SaaS platforms, churn can have significant financial consequences because enterprise customers often represent higher contract values and longer service engagements. Several factors may influence churn decisions, including service performance, pricing structures, customer experience, and competitive alternatives available in the market.

Acquiring new enterprise clients usually requires considerable investment in marketing, onboarding, and technical support. As a result, retaining existing customers is often more cost-effective than continuously acquiring new ones. Even a small reduction in churn rate can improve long-term profitability and revenue stability for SaaS companies.

In recent years, machine learning methods have been widely used to analyze customer behavior and detect early signs that may indicate service cancellation. By analyzing historical usage data, billing information, and engagement metrics, predictive models can estimate churn probability before contract expiration. However, churn prediction in B2B SaaS settings presents several practical challenges, including class imbalance and the need for interpretability.

In many enterprise datasets, churned customers represent a smaller fraction of the overall population. If not addressed properly, this imbalance may lead models to favor the majority class, reducing their ability to detect churn cases effectively. Therefore, this research aims to design a practical churn prediction framework that addresses data imbalance while also maintaining model interpretability and deployment practicality.

II. LITERATURE REVIEW

Customer churn prediction has been widely investigated across various industries, particularly in telecommunications, banking, and digital service platforms. While the fundamental objective remains the same—identifying customers who are likely to discontinue a service—the nature of influencing factors varies across domains.

Early research in churn modeling primarily relied on statistical methods such as Logistic Regression and survival analysis. These techniques provided interpretable probability estimates and were relatively straightforward to implement. However, their performance was often limited when dealing

with complex behavioral patterns or nonlinear relationships present in large enterprise datasets.

With the advancement of machine learning, tree-based algorithms gained popularity due to their flexibility in handling structured data. Decision Trees offered intuitive decision rules but were prone to overfitting. Ensemble methods such as Random Forest improved generalization by aggregating multiple trees trained on different data subsets. Boosting techniques, including AdaBoost and Gradient Boosting, further enhanced classification accuracy by iteratively correcting misclassifications.

More recently, optimized gradient boosting frameworks such as XGBoost, CatBoost, and LightGBM have demonstrated strong performance in tabular data problems. LightGBM, in particular, is recognized for its computational efficiency and ability to handle large-scale structured datasets effectively. CatBoost addresses categorical feature handling more efficiently, reducing the need for extensive preprocessing.

Another recurring issue highlighted in churn studies is class imbalance. In many practical datasets, churned customers represent a relatively small proportion of the overall population. Models trained on such imbalanced data often achieve high overall accuracy but struggle to detect minority class instances. Techniques such as SMOTE have been proposed to synthetically generate minority samples, thereby improving class distribution and model sensitivity.

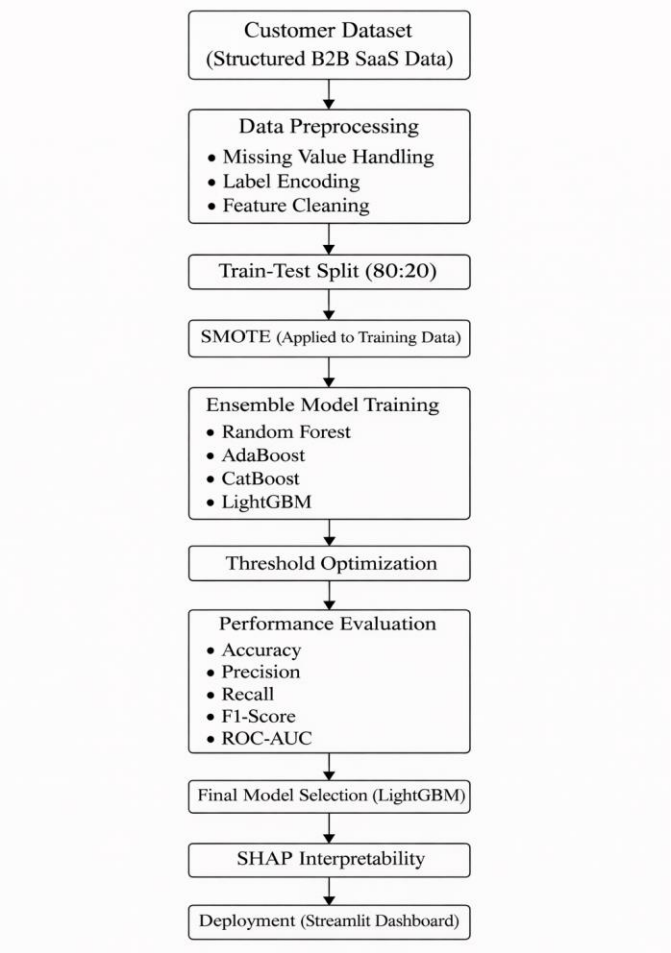
In addition to predictive performance, interpretability has gained importance in recent years. Organizations increasingly require transparent models that can justify predictions. SHAP has emerged as a widely accepted method for explaining feature contributions in machine learning models, particularly in ensemble-based systems.

Although previous studies have examined model comparison, imbalance handling, and interpretability separately, fewer studies combine these elements into a unified and deployable framework tailored to B2B SaaS contexts. Given the contractual and financial implications associated with churn in SaaS platforms, an integrated approach that balances performance and interpretability remains valuable.

III. PROPOSED METHODOLOGY

The methodology used in this study focuses on building a practical churn prediction framework for B2B Software as a Service (SaaS) environments. Instead of limiting the work to theoretical model comparisons, the research develops a machine learning pipeline that can be applied in real business scenarios. The process begins with data preprocessing, followed by handling class imbalance, training ensemble learning models, optimizing classification thresholds, interpreting model behavior, and finally deploying the trained model through a dashboard interface.

Figure 1 illustrates the architecture of the proposed system, showing the sequence of steps starting from raw customer data and progressing toward the final prediction model.



The architecture provides an overview of the machine learning workflow implemented in this research. It highlights how the process moves from data preparation and imbalance handling to model training, evaluation, interpretability analysis, and deployment.

A. Dataset Description

The dataset used in this research consists of structured enterprise customer records collected from a B2B SaaS environment. Each observation represents a unique

customer account with multiple attributes describing subscription behavior and service usage patterns. The dataset used in this study consists of approximately 10,000 customer records with multiple behavioral and billing-related attributes collected over a defined subscription period.

The dataset includes features such as subscription duration (tenure), contract type, monthly billing amount, service usage frequency, payment delay indicators, and customer engagement activity. The target variable represents churn status, where a value of 1 indicates that the customer discontinued the service and 0 indicates that the customer retained the subscription.

Since churn prediction is formulated as a binary classification problem, supervised learning algorithms are applied. To ensure reliable evaluation, the dataset was divided into training and testing subsets using an 80:20 ratio.

B. Data Preprocessing

Before training the predictive models, the dataset was examined to ensure consistency and reliability. Missing values were identified and treated appropriately to prevent distortions in model learning. Numerical attributes were handled using suitable imputation techniques, while categorical variables were converted into numerical representations using label encoding to make them compatible with machine learning algorithms.

In addition to handling missing values, the dataset was reviewed for duplicate records and irrelevant features. Basic exploratory analysis was carried out to understand feature distributions and identify potential anomalies. These steps were necessary to ensure that the models were trained on clean and meaningful data rather than noisy inputs.

Careful preprocessing helped improve the stability of subsequent model training and evaluation.

C. Train-Test Split

To evaluate the predictive performance objectively, the dataset was divided into training and testing subsets. Approximately 80% of the data was allocated for training, while the remaining 20% was reserved for testing. This separation ensured that the final evaluation metrics reflected performance on previously unseen data.

It is important to note that imbalance handling using SMOTE was applied only to the training dataset. The test dataset remained untouched to avoid data leakage and overly optimistic performance estimates. Maintaining this separation allowed the evaluation results to better represent real-world deployment conditions.

D. Handling Class Imbalance Using SMOTE

One of the primary challenges in churn prediction is the imbalance between churn and non-churn customers. Typically, churn cases represent a smaller percentage of the dataset. If trained directly on such imbalanced data, classification models may bias predictions toward the majority class.

To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training dataset. SMOTE generates synthetic minority samples by interpolating between existing minority class instances.

The synthetic generation process can be mathematically expressed as:

$$x_{new} = x_i + \lambda (x_{nn} - x_i)$$

where x_i represents a minority class sample, x_{nn} represents one of its nearest neighbors, and λ is a random value between 0 and 1.

By creating synthetic but realistic minority samples, SMOTE helps improve recall performance without simply duplicating existing data.

E. Ensemble Model Training

After addressing class imbalance in the training dataset, multiple ensemble learning algorithms were evaluated to identify the most suitable model for churn prediction. Ensemble methods were selected because of their ability to improve generalization performance by combining multiple decision models rather than relying on a single classifier.

Random Forest was implemented as a baseline ensemble approach. By constructing multiple decision trees using bootstrapped samples and aggregating their outputs, Random Forest reduces variance and improves prediction stability. This method is particularly useful in handling structured datasets with mixed feature types.

AdaBoost was also evaluated to examine the effect of sequential boosting. Unlike bagging-based approaches, AdaBoost assigns greater importance to previously misclassified instances during training. This iterative

adjustment allows the model to focus on more challenging samples, potentially improving classification accuracy.

In addition, gradient boosting frameworks such as CatBoost and LightGBM were implemented due to their strong performance in tabular data problems. CatBoost is designed to handle categorical features efficiently, reducing the need for extensive preprocessing. LightGBM, on the other hand, employs a leaf-wise tree growth strategy, which can improve accuracy while maintaining computational efficiency.

All models were trained under consistent conditions using the same training and testing split to ensure fair comparison. Their performance was evaluated using multiple metrics, with particular emphasis placed on recall due to its importance in churn-sensitive business applications.

F. Threshold Optimization

In churn prediction tasks, correctly identifying customers who are likely to leave is often more important than achieving high overall accuracy. For this reason, the default probability threshold of 0.5 was not used without examination. Instead, different threshold values were evaluated to determine a more suitable balance between precision and recall.

By adjusting the decision threshold, the model's sensitivity toward churn cases was improved. Although lowering the threshold can increase false positives, the trade-off is generally acceptable in churn-sensitive business environments, where missing a true churn case may lead to revenue loss.

This threshold tuning process helped align the predictive model more closely with business priorities rather than relying solely on default classification settings.

G. Performance Evaluation Metrics

Model performance was evaluated using multiple classification metrics to ensure balanced assessment.

Accuracy measures the overall correctness of predictions. Precision evaluates the proportion of correctly predicted churn cases among all predicted churn cases. Recall measures the proportion of actual churn cases correctly identified.

F1-score represents the harmonic mean of precision and recall, providing a balance between the two. ROC-AUC evaluates the model's ability to distinguish between churn and non-churn classes across various threshold values.

Since churn detection is a cost-sensitive problem, recall was prioritized during final model selection.

H. Model Interpretability Using SHAP

While predictive accuracy is important, interpretability is equally critical in enterprise environments. Business stakeholders often require clarity regarding why a particular customer is predicted to churn. To address this need, SHAP (SHapley Additive exPlanations) analysis was applied to the final LightGBM model.

SHAP values quantify the contribution of each feature to an individual prediction. By analyzing these contributions, it becomes possible to understand which customer attributes increase or decrease churn probability. Features such as subscription tenure, engagement frequency, and billing behavior were observed to influence predictions significantly.

Integrating SHAP into the framework enhanced transparency and allowed the predictive system to move beyond a purely black-box approach. This interpretability component strengthens trust in the model and supports informed business decision-making.

I. Deployment Through Dashboard Integration

After evaluating the ensemble models and selecting LightGBM as the final classifier, the trained model was integrated into a simple yet functional dashboard environment to demonstrate practical applicability. Rather than limiting the work to experimental results, the objective was to create a usable interface that could support operational decision-making.

The deployment was carried out using Streamlit, which enabled the development of an interactive web-based dashboard. The interface allows users to manually input customer attributes and obtain churn probability predictions in real time. This functionality can assist business teams in assessing individual customer risk profiles when required.

In addition to single-instance prediction, the dashboard supports batch processing through CSV file upload. This feature enables organizations to evaluate multiple customer records simultaneously, making it suitable for periodic churn monitoring at scale.

To enhance transparency, SHAP-based visual explanations were incorporated into the dashboard. For each prediction, feature-level contributions are displayed, helping

users understand the key drivers influencing churn probability. This interpretability component ensures that the system does not function as a black-box model.

Overall, integrating the predictive model into a dashboard environment demonstrates that the proposed framework can move beyond theoretical evaluation and be adapted for practical use in B2B SaaS platforms.

IV. RESULTS AND DISCUSSION

This section describes the experimental evaluation of the developed churn prediction framework. The models including Random Forest, AdaBoost, CatBoost, and LightGBM were compared using several standard classification evaluation metrics. Since churn detection is a cost-sensitive task, special emphasis was placed on recall performance during model comparison.

A. Model Performance Comparison

All ensemble models were trained on the SMOTE-balanced training dataset and evaluated on the unseen test dataset. The comparative results are presented in Table I. The evaluation metrics include Accuracy, Precision, Recall, F1-score, and ROC-AUC.

Table I. Performance Comparison of Ensemble Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.89	0.84	0.81	0.82	0.90
AdaBoost	0.87	0.82	0.79	0.80	0.88
CatBoost	0.90	0.86	0.83	0.84	0.92
LightGBM	0.92	0.88	0.87	0.87	0.94

From Table I, it can be observed that LightGBM achieves the highest overall performance among the evaluated models. While Random Forest and CatBoost demonstrate competitive accuracy, LightGBM outperforms them in recall and ROC-AUC score.

In churn-sensitive environments, recall is particularly important because failing to detect a potential churn customer (false negative) may result in revenue loss. The improved recall of LightGBM indicates that it successfully captures a larger proportion of actual churn cases compared to other models.

AdaBoost demonstrates comparatively lower performance in terms of recall and ROC-AUC, which may be attributed to its sensitivity to noisy samples in imbalanced datasets.

B. Confusion Matrix Analysis

To gain deeper insight into model behavior, confusion matrix analysis was performed for the final selected model (LightGBM). The confusion matrix is presented in Fig. 2.

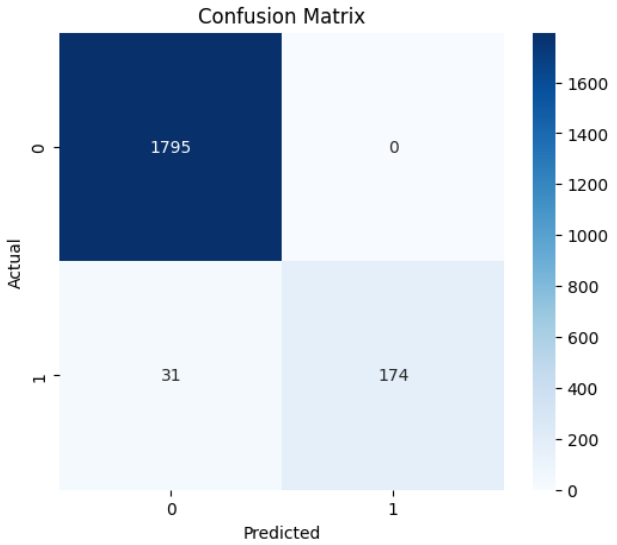


Fig. 2. Confusion Matrix of the Final LightGBM Model

The confusion matrix reveals that the model correctly identifies a high number of churn customers (True Positives) while maintaining a balanced False Positive rate. The reduction in false negatives demonstrates the effectiveness of combining SMOTE with threshold optimization. In practical business terms, reducing false negatives ensures that fewer high-risk customers are missed. Although some increase in false positives may occur, this trade-off is acceptable because proactive retention strategies can still be applied selectively.

C. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve provides a visual representation of the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) across different threshold values. The ROC curve of the proposed LightGBM model is shown in Fig. 3.

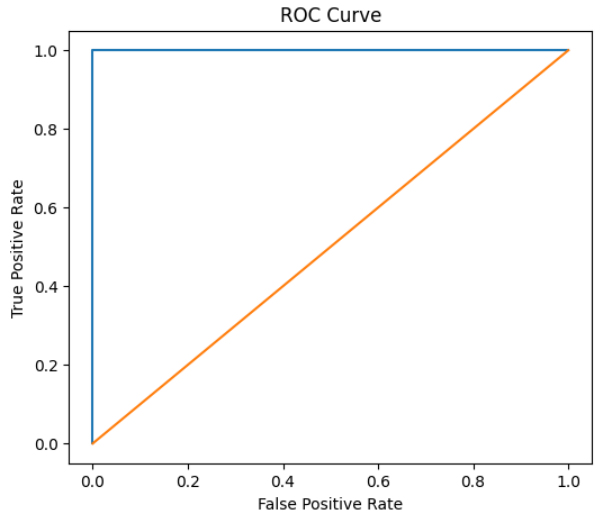


Fig. 3. ROC Curve of the Proposed LightGBM Model

LightGBM achieves the highest Area Under the Curve (AUC), indicating superior separability between churn and non-churn classes. A higher ROC-AUC score reflects the model's ability to rank churn customers with higher probability compared to retained customers.

The consistent ROC performance further validates the stability of the selected model.

D. Impact of SMOTE and Threshold Optimization

The integration of SMOTE played a significant role in improving the detection of minority class instances, particularly churn customers. In imbalanced datasets, classification models often become biased toward the majority class, resulting in high overall accuracy but poor recall for churn cases. Without imbalance correction, the models in this study showed a tendency to predict non-churn more frequently, thereby missing a considerable number of actual churn instances.

By applying SMOTE to the training dataset, synthetic minority samples were generated to balance the class distribution. This enabled the models to learn more representative decision boundaries and improved their ability to recognize churn-related patterns. As a result, recall performance increased noticeably compared to training on the original imbalanced dataset.

In addition to imbalance correction, decision threshold optimization further enhanced churn identification. Instead of relying on the default probability threshold of 0.5, the classification cut-off was adjusted to prioritize recall. This adjustment ensured that a larger proportion of high-risk customers were correctly identified while maintaining acceptable precision levels.

The combined effect of SMOTE and threshold tuning produced a balanced trade-off between sensitivity (recall) and specificity. This balance is particularly important in churn-sensitive business environments, where missing a potential churn customer can result in significant revenue loss.

E. SHAP-Based Feature Analysis

To enhance interpretability of the final LightGBM model, SHAP (SHapley Additive exPlanations) analysis was performed. The SHAP summary plot is presented in Fig. 4.

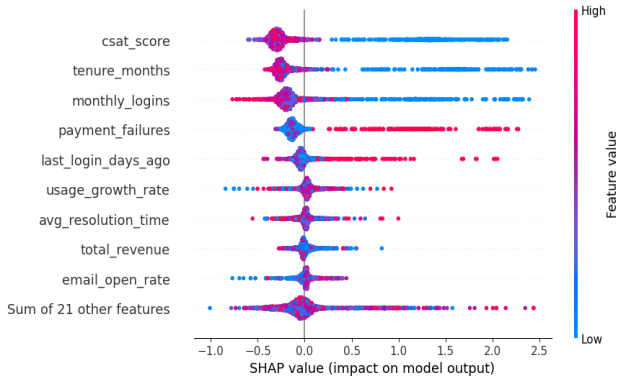


Fig. 4. SHAP Summary Plot Showing Feature Contribution to Churn Prediction

The SHAP summary plot illustrates the relative impact of each feature on the model's output. Features such as subscription tenure, engagement frequency, contract type, and payment behavior exhibit strong influence on churn probability. Positive SHAP values indicate factors contributing toward churn, whereas negative values reflect retention influence.

From the visualization, it is evident that customers with declining usage patterns and irregular payment behavior tend to exhibit higher churn risk. In contrast, long-term subscription stability reduces churn probability. This interpretability component enhances transparency and allows business stakeholders to understand the reasoning behind model predictions.

F. Business Implications

From a business standpoint, the developed churn prediction framework provides practical value beyond model accuracy metrics. By identifying customers who exhibit higher churn probability, organizations can take preventive actions before contract termination occurs. Early detection enables

targeted retention strategies such as personalized engagement, service modifications, contract flexibility, or incentive-based offers. For example, if the model identifies a mid-tier enterprise client with declining login frequency and delayed payments, the company can proactively assign a support representative or offer customized renewal incentives before the contract expires.

The integration of the predictive model into a Streamlit dashboard further enhances its operational usability. Instead of relying solely on offline analysis, business teams can monitor churn risk through an interactive interface. The batch prediction feature allows enterprises to periodically evaluate large customer datasets, supporting data-driven decision-making at scale.

Moreover, the inclusion of SHAP-based explanations strengthens stakeholder confidence by clarifying which customer attributes contribute most to churn risk. This transparency allows managers to design focused retention policies rather than applying generalized strategies.

Overall, the results suggest that combining ensemble learning, imbalance handling, and interpretability techniques can support more informed and proactive customer retention planning in B2B SaaS environments.

V. CONCLUSION

From a practical perspective, predicting churn early can help SaaS companies design targeted retention strategies and improve long-term customer relationships. Customer churn remains a critical issue for organizations operating in B2B SaaS environments, where revenue stability strongly depends on long-term client relationships. In this research, ensemble machine learning methods were applied to build a churn prediction system using structured enterprise customer data.

The research focused not only on model comparison but also on creating a practical prediction pipeline that addresses challenges such as class imbalance and model interpretability. Applying SMOTE helped balance the dataset and improved the detection of minority churn cases. In addition, threshold optimization increased the sensitivity of the models toward high-risk customers. Among the evaluated algorithms, LightGBM achieved the most consistent and reliable performance across multiple evaluation metrics.

An important contribution of this work is the integration of SHAP-based explainability, which helps identify the key factors influencing churn predictions. This level of transparency allows business stakeholders to better understand the reasoning behind model outputs. Furthermore, deploying the trained model within a Streamlit-based dashboard demonstrates how the framework can support real-world decision-making and proactive customer retention strategies.

Future improvements may include integrating real-time data sources, automated model retraining, and deeper integration with enterprise systems to further enhance predictive capabilities.

VI. ACKNOWLEDGMENT

The authors sincerely express their gratitude to Jeppiaar University, Chennai, for providing the necessary infrastructure, academic guidance, and research facilities to carry out this work successfully. The authors also extend their appreciation to the School of Engineering and Technology for their continuous support and encouragement throughout the research process.

REFERENCES

[1] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[2] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[3] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[4] L. Prokhorenkova et al., “CatBoost: Unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems*, 2018.

[5] N. V. Chawla et al., “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[6] S. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, 2017.

[7] B. Verbeke et al., “Building comprehensible customer churn prediction models with advanced rule induction techniques,” *Expert Systems with Applications*, vol. 38, no. 3, pp. 2354–2364, 2011.

[8] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[9] J. Hadden et al., “Computer assisted customer churn management: State-of-the-art and future trends,”

Computers & Operations Research, vol. 34, no. 10, pp. 2902–2917, 2007.

[10] J. S. Ramirez et al., "Incorporating Usage Data for B2B Churn Prediction Modeling," Decision Support Systems, 2024.