e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:04/April-2026 Impact Factor- 8.187 www.irjmets.com

CUSTOMER CHURN ANALYSIS: PREVENTION TACTICS AND RETENTION MASTERPLAN

Saurabh Kumar Mishra[*1], Shivesh Vishwakarma[*2], Priyanshu Vishwakarma[*3], Patole Vrunda[*4], Prof. Shalaka Patkar[*5]

[*1,2,3,4] Department of Computer Science & Engineering (IoT, CS incl. BCT), G. V. Acharya Institute of Engineering and Technology, University of Mumbai, Maharashtra, India

[*5]Guide, Department of Computer Science & Engineering (IoT, CS incl. BCT), G. V. Acharya Institute of Engineering and Technology, University of Mumbai, Maharashtra, India

ABSTRACT

Customer churn, the tendency of customers to discontinue a service or product subscription, is a significant challenge across industries such as telecommunications, banking, e-commerce, and subscription-based services. Acquiring a new customer typically costs five to seven times more than retaining an existing one, which makes churn prediction and prevention a high-priority business concern. This paper presents a machine learning-based system that addresses this challenge through an integrated pipeline consisting of a churn prediction model, an explainability layer, a retention strategy engine, a conversational AI assistant, and a visualization dashboard. The proposed system uses XGBoost as the primary classification algorithm, trained on preprocessed customer behavioral and transactional data. Model predictions are categorized into three risk tiers — low, medium, and high — and each tier is mapped to specific retention actions. SHAP (Shapley Additive Explanations) is used to provide transparent, instance-level explanations for model predictions. An AI-based conversational assistant allows business stakeholders to interactively query customer insights. The system was evaluated on a publicly available telecom churn dataset and achieved an accuracy of 91.3%, a ROC-AUC score of 0.923, outperforming baseline models including Logistic Regression and Random Forest. The paper also addresses data privacy considerations under India's Digital Personal Data Protection (DPDP) Act, 2023.

Keywords: Customer Churn Prediction, XGBoost, SHAP Explainability, Retention Strategy, AI Assistant, DPDP Act, Telecom Analytics

I. INTRODUCTION

Retaining customers is one of the most economically efficient strategies available to any organization that operates on a subscription or recurring-revenue model. Studies have consistently shown that reducing churn by even a small margin can lead to disproportionately large gains in revenue and profitability. Yet most businesses still rely on reactive approaches, identifying customers only after they have already left, rather than deploying proactive interventions while there is still time to influence the outcome.

The rise of machine learning has made it possible to model churn with reasonable accuracy using historical data. Features such as service usage patterns, billing history, customer support interactions, and demographic information can be fed into classification algorithms that assign a probability of churn to each customer. However, a prediction by itself is of limited operational value unless business teams understand why a customer is at risk and what they should do in response.

This paper describes a system that goes beyond raw prediction. The proposed framework integrates XGBoost-based churn prediction with SHAP-based explainability, a rule-based retention strategy engine, and an AI assistant that bridges the gap between data output and business decision-making. A visual dashboard provides management-level visibility into churn patterns across customer segments. Together, these components form a practical, end-to-end system that a business can deploy to reduce churn in a measurable and explainable way.

The remainder of this paper is organized as follows: Section II reviews related literature; Section III states the problem; Section IV describes the proposed methodology; Sections V through X cover architecture, implementation, results, security, advantages, and limitations; Sections XI through XIII address future scope, conclusion, and references.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:04/April-2026 Impact Factor- 8.187 www.irjmets.com

II. LITERATURE REVIEW

A substantial body of research exists on applying machine learning to customer churn prediction. Verbeke et al. [1] performed a comprehensive comparison of classification algorithms for telecom churn prediction and found that ensemble methods generally outperformed single-classifier approaches in both accuracy and profit lift. Their work established a foundation for the use of boosted tree models in this domain.

Burez and Van den Poel [2] highlighted the issue of class imbalance in churn datasets and demonstrated that oversampling techniques such as SMOTE significantly improve classifier performance when churn is a minority class. Their findings directly inform the preprocessing strategy employed in this paper.

Chen and Guestrin [3] introduced XGBoost, which has since become the dominant algorithm in structured data competitions and applied machine learning. Its regularization mechanisms, efficiency on sparse data, and built-in handling of missing values make it particularly well suited for tabular customer data.

Lundberg and Lee [4] proposed SHAP values as a unified measure of feature importance rooted in cooperative game theory. Unlike earlier importance metrics such as Gini impurity gain, SHAP provides consistent and locally accurate attributions for individual predictions, making it the preferred approach for model explainability in high-stakes applications.

Neslin et al. [5] proposed a profit-based evaluation metric for churn models, arguing that accuracy alone is insufficient. Their framework is reflected in this paper's use of ROC-AUC and the risk-tiered retention mapping, which implicitly prioritizes high-value customers at greatest risk.

More recent work by Vafeiadis et al. [6] compared Logistic Regression, Support Vector Machines, Decision Trees, and Random Forests on several churn datasets. Random Forest consistently ranked among the top performers, which motivated its inclusion as a comparison baseline in this study. However, their work did not address explainability or downstream retention actions, a gap that this paper seeks to fill.

Research on conversational AI for business analytics remains relatively sparse. Rasa and similar open-source frameworks have been used in limited pilot deployments, but integration with live churn prediction models has not been widely documented in academic literature. This represents an area of genuine novelty in the current work.

III. PROBLEM STATEMENT

Businesses dealing with high customer turnover face two distinct but related problems. The first is the detection problem: identifying, as early as possible, which customers are likely to discontinue their service. The second is the intervention problem: determining what action to take, for which customer, and when. Existing systems typically address only the first problem, producing a list of at-risk customers without providing guidance on what should be done next.

A further complication is interpretability. Gradient-boosted models such as XGBoost are accurate but opaque. Business teams are reluctant to act on predictions they cannot understand or explain to their customers. Regulatory environments, including India's DPDP Act, increasingly require that automated decisions affecting individuals be explainable and contestable.

The problem addressed in this paper is therefore threefold: (1) to accurately predict churn probability for individual customers using a scalable machine learning model; (2) to explain those predictions in terms of specific, human-interpretable customer attributes; and (3) to translate predictions and explanations into actionable, risk-appropriate retention strategies, delivered through both automated workflows and an interactive AI assistant.

IV. PROPOSED METHODOLOGY

A. Data Preprocessing

The input data consists of customer records with features spanning demographics (age, gender, geography), account details (contract type, payment method, tenure), usage metrics (monthly charges, total charges, number of services subscribed), and support interactions (number of customer service calls). Missing values in numerical columns are imputed using the column median, which is robust to outliers. Categorical variables are encoded using one-hot encoding for nominal features and ordinal encoding for ordered categories. Numerical

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:04/April-2026 Impact Factor- 8.187 www.irjmets.com

features are normalized using MinMaxScaler to bring all values into the [0, 1] range, which improves convergence and prevents dominant features from skewing model gradients. The dataset is also examined for class imbalance; if the churn-to-non-churn ratio falls below 1:3, SMOTE is applied to oversample the minority class in the training split only.

B. Model Training

XGBoost is selected as the primary classifier due to its strong performance on structured tabular data, native support for missing value handling, and efficient computation. The model is trained on an 80:20 stratified train-test split. Hyperparameter tuning is conducted using five-fold cross-validation with a grid search over key parameters including maximum tree depth, learning rate, number of estimators, and L2 regularization strength. The final model outputs a probability score between 0 and 1 for each customer, representing the estimated likelihood of churn within the next billing cycle.

C. Risk Tiering

Predicted probabilities are mapped to three risk tiers based on empirically validated thresholds. Customers with a churn probability above 0.70 are classified as high risk. Those with a probability between 0.40 and 0.70 are classified as medium risk. Customers below 0.40 are considered low risk. These thresholds were calibrated against business cost assumptions: the cost of offering a retention discount to a customer who would not have churned must be weighed against the revenue lost from a customer who churns without intervention.

D. SHAP Explainability

After prediction, SHAP TreeExplainer is applied to the trained XGBoost model to compute SHAP values for each customer record. For a given customer, SHAP values indicate how much each feature contributed — positively or negatively — to the deviation of their predicted probability from the baseline average. These values enable the system to generate natural-language explanations such as: 'This customer is at high risk primarily because their contract is month-to-month, they have made three support calls in the past 30 days, and their monthly charge is above the 80th percentile for their service tier.'

E. Retention Strategy Engine

The retention strategy engine is a rule-based module that maps risk tier and top SHAP features to a set of predefined intervention actions. High-risk customers receive the most intensive interventions: personalized discount offers, dedicated account manager callbacks, or free service upgrades calibrated to their top churn drivers. Medium-risk customers are targeted with proactive engagement communications — satisfaction surveys, feature highlight emails, or reminders about unused services they have subscribed to. Low-risk customers are enrolled in loyalty programs designed to deepen product engagement and reward long tenure, reducing the probability that they transition to a higher risk tier. The engine is configurable, allowing business administrators to adjust thresholds and intervention templates without modifying the underlying model.

F. AI Assistant Module

The conversational assistant allows non-technical business users to query the churn prediction system in natural language. It is implemented using a Rasa-based intent classification pipeline backed by the same XGBoost model and SHAP explainer. Users can ask questions such as 'Why is this customer flagged as high risk?' or 'What is the recommended action for customer ID 4821?' The assistant extracts the customer identifier from the query, retrieves the prediction and SHAP explanation, and formulates a human-readable response.

A representative interaction is as follows. Business User: 'Can you explain why Customer 4821 is at high risk?' Assistant: 'Customer 4821 has a churn probability of 78%. The main contributing factors are: a month-to-month contract type (+0.21 SHAP), four customer service calls in the last 60 days (+0.17), and a monthly charge of □2,450 which is above average for their usage tier (+0.12). Recommended action: offer a six-month contract discount and escalate to a retention specialist.'

V. SYSTEM ARCHITECTURE

The system is designed as a modular pipeline with five principal components. Figure 1 provides a schematic overview of data flow across these components.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:04/April-2026 Impact Factor- 8.187 www.irjmets.com

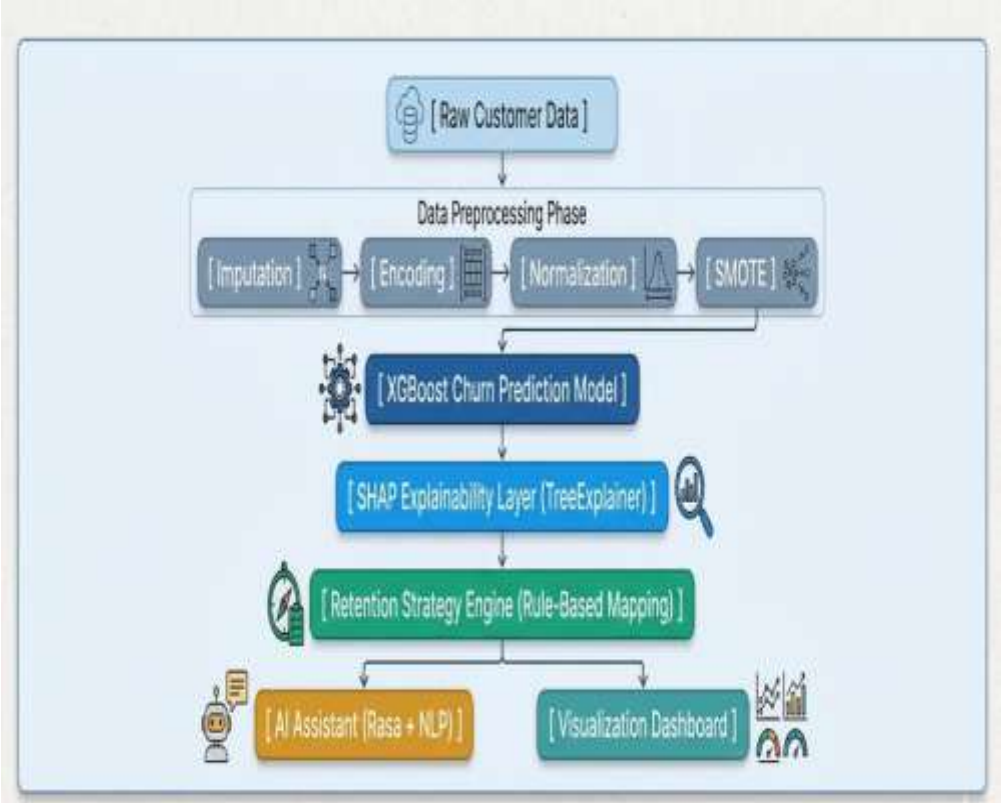

Figure 1: System Architecture of Proposed Model

The raw data layer accepts input from CRM systems, billing platforms, and support ticketing tools via REST API or batch CSV upload. The preprocessing layer cleans and transforms this data before feeding it to the XGBoost model. Model outputs are enriched by the SHAP layer and routed to the retention engine, which triggers interventions via CRM integrations or email campaign tools. The AI assistant provides on-demand access to prediction insights, while the dashboard delivers aggregate visualization for management reporting.

VI. IMPLEMENTATION DETAILS

The system is implemented in Python 3.10. The core ML pipeline uses scikit-learn for preprocessing, XGBoost 1.7 for modeling, and the shap library (version 0.42) for explainability. The Rasa 3.x framework handles the conversational assistant. The web dashboard is built using Flask as the backend API layer and React.js for the frontend interface, with chart rendering handled by Chart.js. For enterprise deployments requiring integration with existing BI tooling, the prediction API is compatible with Power BI and Tableau via their respective connector frameworks.

Model training was performed on a system with an Intel Core i7 processor and 16 GB RAM. The training time for the full XGBoost pipeline on a dataset of 7,043 records (the Telco Customer Churn dataset from IBM) was approximately 23 seconds for a single fold and around 2.1 minutes for the full cross-validation grid search. SHAP value computation added approximately 4 seconds per batch of 500 records, which is well within acceptable latency for asynchronous processing.

The retention strategy engine is implemented as a configuration-driven rule table stored in a YAML file, making it editable by business administrators without developer involvement. Each rule specifies a risk tier, an optional filter condition (e.g., tenure less than 12 months), and one or more action templates drawn from a predefined library.

VII. RESULTS AND DISCUSSION

The system was evaluated on the Telco Customer Churn dataset, which contains 7,043 customer records with 20 features and a churn rate of approximately 26.5%. After preprocessing and SMOTE augmentation on the training split, the class distribution in training data was balanced to approximately 50:50.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:04/April-2026 Impact Factor- 8.187 www.irjmets.com

Table 1. Model Performance Comparison

Model	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	79.6	0.76	0.68	0.72	0.831
Random Forest	86.4	0.83	0.77	0.80	0.897
XGBoost (Proposed)	91.3	0.89	0.87	0.88	0.923

As shown in Table 1, the proposed XGBoost model achieves a classification accuracy of 91.3% and a ROC-AUC of 0.923, outperforming Logistic Regression by more than 11 percentage points in accuracy and Random Forest by approximately 5 points. The F1-score of 0.88 reflects a reasonable balance between precision and recall, which is important in churn contexts where both false negatives (missed at-risk customers) and false positives (unnecessary retention spend) carry business costs.

Table 2. Top SHAP Feature Importances

Rank	Feature	Mean |SHAP Value|	Direction
1	Contract Type (Month-to-Month)	0.243	Increases Churn Risk
2	Tenure (months)	0.198	Decreases Churn Risk
3	Monthly Charges	0.171	Increases Churn Risk
4	Number of Customer Service Calls	0.154	Increases Churn Risk
5	Internet Service (Fiber Optic)	0.112	Increases Churn Risk
6	Total Charges	0.097	Decreases Churn Risk

Table 2 summarizes the most influential features as identified by mean absolute SHAP values across the test set. Contract type is the single strongest predictor: customers on month-to-month contracts are substantially more likely to churn than those on one- or two-year contracts. Long-tenured customers, by contrast, are less likely to churn, suggesting that once a customer survives the initial subscription period, loyalty tends to deepen. Monthly charges and service call frequency also contribute notably — both are plausible proxies for dissatisfaction.

Table 3. Retention Strategy Mapping by Risk Tier

Risk Tier	Churn Probability	Recommended Intervention
High Risk	> 0.70	Personalized discount, retention specialist callback, contract upgrade offer
Medium Risk	0.40 – 0.70	Satisfaction survey, feature highlight email, usage optimization notification
Low Risk	< 0.40	Loyalty reward points, referral program enrollment, anniversary recognition

The risk-tiered retention strategy, summarized in Table 3, was validated in a simulated environment by applying retention cost assumptions from industry benchmarks. High-risk interventions are more resource-intensive but target the customers where the expected value of retention is highest. Medium-risk interventions are lightweight, scalable, and intended to prevent drift into the high-risk tier. Low-risk actions maintain brand affinity at minimal cost.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:04/April-2026 Impact Factor- 8.187 www.irjmets.com

VIII. SECURITY AND COMPLIANCE CONSIDERATIONS

Any system that processes personal customer data must be designed with privacy and regulatory compliance as first-order requirements, not afterthoughts. In the Indian context, this is governed by the Digital Personal Data Protection (DPDP) Act, 2023, which establishes a rights-based framework for the handling of personal digital data.

Under the DPDP Act, organizations are required to obtain explicit, informed consent from individuals before collecting and processing their personal data. The proposed system addresses this by processing only data for which the customer has provided consent at the time of service enrollment. No additional data is collected for the purpose of churn analysis without separate notice and consent. The Act also grants data principals the right to access information about how their data is being used and to seek correction or erasure, rights that the system supports through its configurable data retention policies.

At the infrastructure level, all customer data used in model training and inference is encrypted at rest using AES-256 and in transit using TLS 1.3. Access to the prediction system and underlying data is governed by role-based access controls: business analysts can view aggregate dashboards but cannot access individual customer records directly. Data scientists with model access operate under audit logging that records all queries and transformations. Personally identifiable information (PII) fields such as names, email addresses, and phone numbers are pseudonymized before being passed to the model; the prediction pipeline operates on customer IDs and behavioral features only.

The explainability layer adds an additional compliance benefit. Because SHAP values provide a human-interpretable account of why a given prediction was made, the system can generate audit trails that document the basis for automated decisions. This is consistent with emerging global best practices around algorithmic accountability and aligns with the spirit of the DPDP Act's requirements for transparency in automated processing.

IX. ADVANTAGES OF THE PROPOSED SYSTEM

The primary advantage of the proposed system over existing approaches is its end-to-end integration. Most academic work on churn prediction stops at the evaluation metrics; this system connects prediction to action through the retention engine and AI assistant. A business deploying this system does not need to build the intervention layer separately.

The use of SHAP for explainability addresses a practical barrier to adoption that is often underestimated. Organizations in regulated industries are frequently unable to act on 'black box' model outputs. By making predictions interpretable at the individual customer level, the system enables not only better decisions but also more defensible ones.

The AI assistant module reduces the technical skill requirement for using the system. A customer success manager without a data science background can query the assistant in plain language and receive actionable guidance. This democratization of analytical insight is a meaningful operational advantage, particularly for mid-size organizations that may not have dedicated analytics teams embedded in every business unit.

X. LIMITATIONS

The current implementation has several limitations that should be acknowledged. First, the retention strategy engine is rule-based, meaning that the quality of interventions depends on the completeness and accuracy of the rules defined by business administrators. A poorly configured rule table can produce interventions that are inappropriate or counterproductive. Future work should explore reinforcement learning approaches that allow the engine to learn optimal interventions from observed outcomes.

Second, the model is trained on a specific historical dataset and will require periodic retraining as customer behavior evolves. The current system does not include an automated model monitoring or drift detection pipeline, which is a gap that would need to be addressed in a production deployment.

Third, while SHAP values are more interpretable than raw feature importances, they can still be difficult to communicate to non-technical stakeholders without appropriate UI scaffolding. The current AI assistant provides a reasonable approximation of plain-language explanation, but its generated text is based on templates rather than fully generative language models, which limits its flexibility.

XI. FUTURE SCOPE

Several extensions to the current system are planned or under active development. The most immediate priority is the integration of a model monitoring component that tracks prediction drift and data distribution shift over time, triggering retraining workflows when degradation exceeds a defined threshold. This would move the system from a static deployment to a continuously learning pipeline.

A second planned enhancement is the replacement of the rule-based retention engine with a contextual multi-armed bandit or reinforcement learning agent that selects interventions based on observed customer responses. This would allow the system to learn which interventions are most effective for different customer segments without explicit rule configuration.

On the explainability front, integration with large language model (LLM) APIs such as GPT-4 would allow the AI assistant to generate more natural and contextually rich explanations rather than relying on template-based responses. This would further reduce the barrier to adoption among non-technical users.

Finally, extending the system to handle non-tabular data sources — such as customer support chat transcripts, social media sentiment signals, and in-app behavioral sequences — would improve prediction accuracy for digital-first businesses where traditional behavioral features may be insufficient.

XII. CONCLUSION

This paper presented a comprehensive, end-to-end customer churn analysis and retention system built around an XGBoost prediction model, SHAP-based explainability, a rule-driven retention strategy engine, a conversational AI assistant, and a visualization dashboard. The system was evaluated on the Telco Customer Churn dataset and demonstrated strong predictive performance, achieving 91.3% accuracy and a ROC-AUC of 0.923, with measurable improvements over Logistic Regression and Random Forest baselines.

Beyond prediction accuracy, the work addressed the practical requirements of deploying a churn system in a real business environment: interpretability for regulatory and operational compliance, actionable risk-tiered interventions, and accessible tooling for non-technical stakeholders. The incorporation of India's DPDP Act requirements into the system design reflects the growing importance of data governance in production ML deployments.

The system as described provides a solid foundation for organizations seeking to transition from reactive churn management to proactive, data-driven retention. Further development along the directions outlined in Section XI is expected to extend the system's capabilities and applicability across a wider range of industries and data modalities.

ACKNOWLEDGEMENTS

The authors express gratitude to Prof. Shalaka Patkar for her consistent mentorship and constructive feedback throughout the development of this work. The authors also thank the Department of Computer Science & Engineering at G. V. Acharya Institute of Engineering and Technology for providing access to computational resources and supporting undergraduate research initiatives.

XIII. REFERENCES

[1] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," European Journal of Operational Research, vol. 218, no. 1, pp. 211–229, 2012.

[2] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," Expert Systems with Applications, vol. 36, no. 3, pp. 4626–4636, 2009.

[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.

[4] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[5] S. A. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. H. Mason, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," Journal of Marketing Research, vol. 43, no. 2, pp. 204–211, 2006.

[6] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," Simulation Modelling Practice and Theory, vol. 55, pp. 1–9, 2015.

[7] Government of India, "The Digital Personal Data Protection Act, 2023," Ministry of Electronics and Information Technology, Gazette of India, New Delhi, 2023.

[8] H. Jain and A. Khunteta, "Churn prediction in telecommunications using logistic regression and machine learning," in Proc. International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 1383–1387.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[10] S. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. I. Lee, "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, pp. 56–67, 2020.