WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Explainable Customer Churn Prediction Using Machine Learning and Intelligent Retention Recommendations

P. Jasmine Lois Ebenezer[a], P. Abinisha[b]

[a]**Associate Professor**, Department of Computer Application and Research Centre, Sarah Tucker College (Autonomous), Tirunelveli.
[b]**Student, Reg.No: 25081205300112001**, Department of Computer Application and Research Centre, Sarah Tucker College (Autonomous), Tirunelveli.

ABSTRACT:

Customer churn presents a significant obstacle for the telecommunications sector, as the loss of current customers can impact income and sustainable business development. This initiative suggests an interpretable machine learning framework for forecasting customer churn and offering tailored retention strategies. The IBM Telco Customer Churn dataset is used to analyze customer characteristics such as tenure, contract type, monthly charges, internet service and payment method. The data is preprocessed and encoded before training different machine learning models. Support Vector Machine (SVM) is selected as the final prediction model based on its overall performance, achieving an accuracy of 81.41%. To improve the interpretability of the predictions, SHAP (SHapley Additive exPlanations) is used to identify the customer features that contribute to churn predictions. Based on the predicted risk and influential customer characteristics, the system generates personalized retention recommendations. A Streamlit-based application provides customer prediction, churn probability, risk assessment, explanations, recommendations, prediction history and an administrative dashboard. The proposed system aims to support telecommunications businesses in identifying customers at risk of churn and taking suitable preventive actions.

Keywords: *Customer Churn, Machine Learning, SVM, Explainable AI, SHAP, Telecommunications, Retention Recommendation.*

I. INTRODUCTION

Customer retention is an important concern for telecommunications companies because customers can easily switch to competing service providers. Customer churn refers to the situation in which an existing customer stops using a company's services. High churn rates can result in revenue loss and increased costs because acquiring new customers is generally more difficult than retaining existing ones.

Machine learning can be used to analyze historical customer information and identify patterns associated with customer churn. Customer attributes such as tenure, contract type, monthly charges, internet service and payment method can provide useful information for predicting whether a customer is likely to leave. However, a prediction alone may not be sufficient for business decision-making. A company also needs to understand why a customer is predicted to churn and what action could be taken to retain that customer.

This project proposes an Explainable Customer Churn Prediction System that combines machine learning prediction, SHAP-based explanation and intelligent retention recommendations. The system uses the IBM Telco Customer Churn dataset and evaluates multiple machine learning approaches. is used as the final deployed model and achieves an accuracy of 81.41%.

The system further uses SHAP to identify the features that have the greatest influence on individual predictions. These explanations are then used along with customer characteristics to generate personalized retention recommendations. The complete system is implemented using a Streamlit interface with database-supported prediction history and an administrative dashboard.

II. LITERATURE REVIEW

Customer churn prediction has become an important application of machine learning in the telecommunications industry, where identifying customers who are likely to discontinue a service can support proactive retention strategies. Earlier churn prediction approaches primarily relied on statistical and traditional machine learning techniques such as Logistic Regression, Decision Trees, Random Forest, and Naïve Bayes. However, recent research has increasingly focused on ensemble learning, advanced machine learning models, and explainable artificial intelligence to improve both predictive performance and the interpretability of churn predictions.

A study published in *Algorithms* in 2024 investigated customer churn prediction in the telecommunications sector using several machine learning algorithms, including Logistic Regression, K-Nearest Neighbors, Naïve Bayes, Decision Tree, and Random Forest. The study reported that Random Forest achieved the best performance among the evaluated models, with an accuracy of 86.94% and an AUC of 0.95. The authors also emphasized that churn prediction models should not focus solely on predictive accuracy but should provide understandable reasons for their predictions [1]. To address this issue, LIME and SHAP were incorporated to explain both overall feature importance and individual customer predictions. The study demonstrates the importance of combining predictive modelling with explainability, particularly when churn predictions are intended to support business decisions.

Asif et al. (2025) proposed an explainable data-driven approach called XAI-Churn TriBoost for customer churn prediction in the telecommunications industry. The proposed framework combines XGBoost, CatBoost, and LightGBM using a soft-voting ensemble [2]. The study used a dataset containing more than two million records and applied extensive preprocessing, feature selection using Boruta with Random Forest, and SMOTE to address class imbalance. The proposed model achieved an accuracy of 96.44%, precision of 92.82%, recall of 87.82%, and F1-score of 90.25%. LIME and SHAP were used to interpret individual predictions and identify important factors associated with churn. The study reported that explainability could provide useful insights for targeted retention strategies.

The findings of Asif et al. also demonstrate a shift from conventional churn prediction toward explainable and actionable analytics. SHAP was used to identify important features affecting individual churn predictions, thereby improving the transparency of complex ensemble models. The authors further suggested that explainable churn models can support targeted interventions and strategic decision-making in telecommunications. However, the proposed approach mainly concentrates on improving predictive performance and explaining model outputs, while the transformation of explanations into specific customer-level retention actions is not the primary focus of the framework.

Mamun (2025) conducted a systematic literature review of machine learning approaches for customer retention and churn analysis. The review considered 112 peer-reviewed studies published between 2015 and 2025 and examined supervised, unsupervised, and hybrid machine learning approaches. Algorithms such as Logistic Regression, Decision Trees, Support Vector Machines, Random Forest, gradient boosting methods, and deep neural networks were identified as widely used techniques for churn prediction [3]. The review also highlighted the increasing importance of feature engineering, preprocessing, evaluation metrics, explainable AI, real-time analytics, and privacy-preserving approaches.

The systematic review further indicates that traditional models remain useful because of their simplicity and interpretability, while more advanced machine learning approaches can better capture nonlinear relationships and complex customer behaviour. In particular, Support Vector Machines, Random Forest, and gradient boosting methods have demonstrated robustness for churn prediction. The review also identifies SHAP and LIME as important post-hoc explainability techniques for addressing the opacity of complex machine learning models.

From the reviewed studies, it can be observed that existing research has successfully addressed customer churn prediction and model explainability, with ensemble models generally receiving considerable attention because of their predictive performance. However, a practical gap remains between identifying the reasons for churn and converting those explanations into personalized retention actions. The literature also shows increasing interest in operational deployment, real-time monitoring, and actionable decision support, indicating that churn prediction systems should move beyond simply producing a churn label or probability.

Therefore, the proposed research focuses on an integrated explainable churn-management framework. Instead of using SHAP only for feature interpretation, the proposed system combines SVM-based churn prediction, churn probability estimation, risk assessment, SHAP-based individual explanations, and personalized retention recommendations. The system is additionally implemented as a Streamlit-based application with database storage, prediction history, and an administrative dashboard. This integration is intended to connect machine learning prediction and explainability with practical retention decision-making.

III. PROBLEM STATEMENT AND OBJECTIVES

A. Problem Statement

Telecommunication companies need to identify customers who are likely to discontinue their services. Traditional analysis may identify general churn patterns but may not provide individual-level explanations or specific retention actions. Therefore, there is a need for a system that can predict customer churn, explain the reasons behind each prediction and recommend suitable retention strategies.

B. Objectives

The proposed research aims to:

- develop an SVM-based customer churn prediction model;
- estimate churn probability and assess customer risk;
- provide individual-level explanations using SHAP;
- generate personalized retention recommendations based on customer characteristics and model insights;
- maintain prediction records through database storage; and
- provide a dashboard for monitoring prediction history and customer-risk information.

IV. PROPOSED METHODOLOGY

The proposed system follows an integrated machine learning and explainable artificial intelligence framework for customer churn prediction and retention decision support. The overall workflow consists of customer data collection, data preprocessing, feature encoding and scaling, SVM-based churn prediction, churn probability estimation, risk assessment, SHAP-based explanation, personalized retention recommendation, database storage, and dashboard-based monitoring.

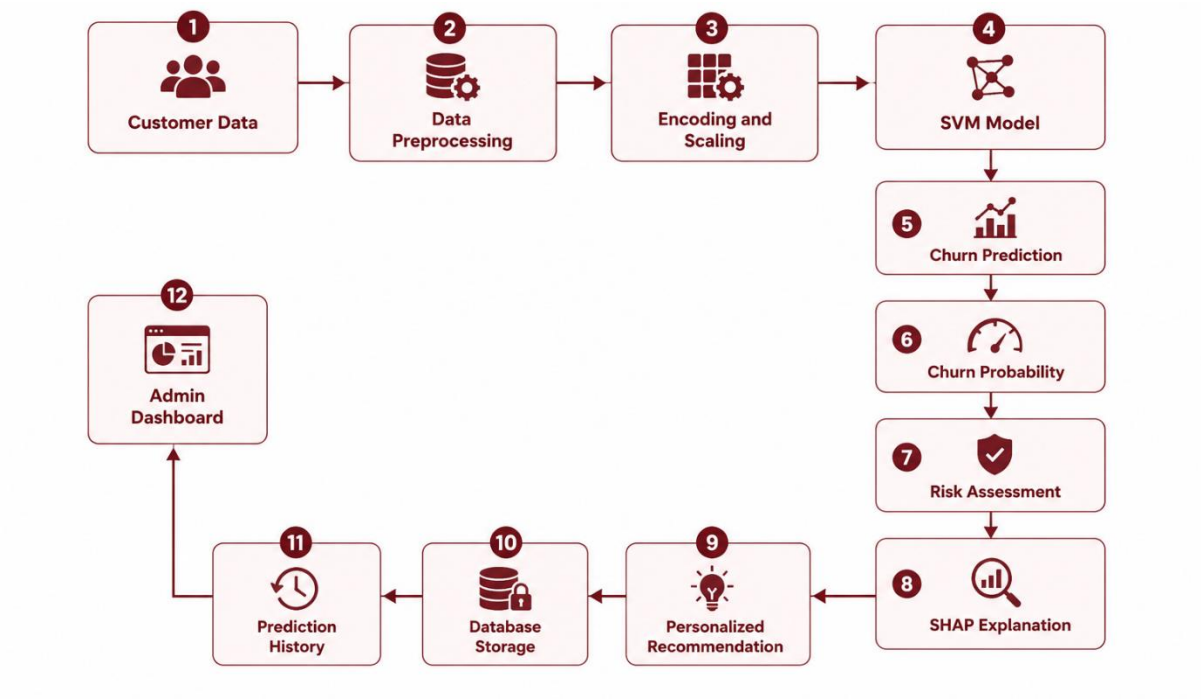


Fig. 1. Proposed System Workflow

A. Dataset Collection

The proposed system uses the IBM Telco Customer Churn dataset, which contains 7,043 customer records with demographic, service-related, contractual, and billing information. The target variable is Churn, which indicates whether a customer has discontinued the service. Features such as tenure, contract type, monthly charges, total charges, internet service, payment method, and subscribed services are used to identify patterns associated with customer churn.

B. Data Preprocessing

The collected dataset is preprocessed to improve its suitability for machine learning. The *customerID* attribute is removed because it serves only as an identifier and does not contribute meaningful predictive information. The *TotalCharges* attribute is converted into a numerical format, and missing values are handled using median imputation.

Categorical attributes are transformed into numerical representations. Binary categorical variables are encoded into binary values, while multi-category attributes are converted using one-hot encoding. This preprocessing produces a numerical feature matrix suitable for model training.

C. Feature Scaling

Since the proposed Support Vector Machine model is sensitive to differences in feature magnitudes, feature scaling is applied after encoding. Standardization is used to transform the input features to a comparable scale. The dataset is then divided into training and testing subsets using an 80:20 ratio with a fixed random state to ensure reproducibility.

D. SVM-Based Churn Prediction

A Support Vector Machine (SVM) classifier with a Radial Basis Function (RBF) kernel is employed as the primary prediction model. SVM identifies a decision boundary that separates churn and non-churn customers while the RBF kernel enables the model to capture nonlinear relationships among customer attributes.

Several machine learning models are evaluated during experimentation, and SVM is selected as the final deployed model based on its overall predictive performance. The final SVM model achieves an accuracy of 81.41% on the test dataset.

E. Churn Probability and Risk Assessment

In addition to the binary churn prediction, the system estimates the probability that an individual customer will churn. This probability is used to determine the customer's risk level. A corresponding customer health score is also calculated from the churn probability, where a lower churn probability indicates better customer health.

The risk assessment enables customers to be categorized according to their likelihood of churn, allowing potentially high-risk customers to be prioritized for retention activities.

F. SHAP-Based Explainability

To improve the interpretability of the SVM predictions, SHAP (SHapley Additive exPlanations) is incorporated into the proposed framework. SHAP assigns contribution values to individual features to indicate how they influence a particular model prediction. For each customer, positive SHAP values indicate features that contribute toward the churn prediction, whereas negative SHAP values indicate features that contribute toward the non-churn prediction. This provides individual-level explanations rather than relying only on overall feature importance.

G. Personalized Retention Recommendation

The explanations generated by the prediction and SHAP modules are used along with customer characteristics to generate personalized retention recommendations. Rule-based actions are associated with important customer conditions such as month-to-month contracts, high monthly charges, fiber-

optic services, electronic check payments, paperless billing, and short customer tenure.

For example, customers with month-to-month contracts may be recommended a long-term contract with suitable benefits, while customers with high monthly charges may be offered an appropriate discount or value-based plan. This module transforms prediction results into actionable customer-retention strategies.

H. Database Storage and Prediction History

The prediction results are stored in a MySQL database to maintain historical records. The stored information includes customer attributes, churn probability, prediction outcome, health score, customer lifetime value, revenue at risk, priority score, and prediction time. The prediction history enables previously generated predictions to be retrieved and analyzed.

I. Streamlit-Based Application and Dashboard

The complete framework is implemented as an interactive Streamlit application. The application provides customer input, churn prediction, probability estimation, risk assessment, SHAP explanations, and personalized recommendations. An administrative dashboard is also provided to visualize prediction statistics, churn distribution, customer health scores, contract information, and high-priority customers.

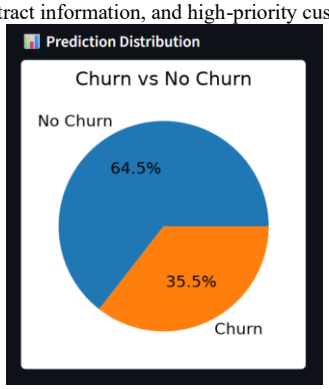


Fig. 2. Dashboard

V. RESULTS AND DISCUSSION

The final SVM model achieved an accuracy of 81.41% on the test dataset. It correctly classified a large proportion of customers who were not expected to churn, with a recall of 92% for the No Churn class.

Model	Accuracy
Random Forest	77.64%
Logistic Regression	79.06%
XGBoost	77.86%
CatBoost	80.34%
LightGBM	80.55%
SVM	81.41%

For the Churn class, the model achieved a precision of 69%, recall of 53% and F1-score of 60%. The results indicate that the model can identify a substantial number of churn-prone customers while maintaining good overall classification performance.

The SHAP module provides an explanation of individual predictions by identifying influential customer characteristics. This improves transparency and allows users to understand the factors associated with a prediction.

The retention recommendation module further converts these explanations into practical actions. Testing with different customer profiles produced different churn probabilities, risk levels and recommendations. For example, customers with high monthly charges, Fiber optic service and month-to-month contracts received recommendations related to discounts, technical support and contract upgrades.

The system also successfully stores predictions in the database. During testing, the Prediction History contained 66 predictions, including 25 predicted churn cases and 41 predicted no-churn cases, with an average predicted churn probability of 47.7%. These records are also reflected in the administrative dashboard.

VI. CONCLUSION

The proposed system provides an explainable machine learning approach for telecommunications customer churn prediction. SVM was selected as the final prediction model and achieved an accuracy of 81.41%. The integration of SHAP improves the interpretability of the prediction by identifying important customer-specific factors. In addition, the personalized retention recommendation module provides actionable strategies based on customer characteristics and churn-related factors. The Streamlit application integrates prediction, probability estimation, risk assessment, explanation,

recommendations, database storage, prediction history and administrative monitoring into a single system. Therefore, the proposed solution can assist businesses in identifying potential churn customers and supporting proactive customer retention decisions.

VII. FUTURE ENHANCEMENT

Future versions of the system can be enhanced by incorporating larger and more diverse telecommunications datasets. Advanced ensemble and deep learning models can also be investigated to improve churn detection, particularly for the churn class. The recommendation module can be extended using customer segmentation, historical response data and optimization techniques to determine the most effective retention strategy for each customer. Real-time integration with customer relationship management systems could further allow companies to automatically identify and respond to high-risk customers.

VIII. REFERENCES

[1] *Algorithms* paper, 2024, "Customer churn prediction" using ensemble learning and XAI techniques, DOI: 10.3390/a17060231. The paper reports Random Forest as the strongest model among its evaluated approaches and uses LIME and SHAP for explainability.

[2] D. Asif, M. S. Arif, and A. Mukheimer, "A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry," *Results in Engineering*, vol. 26, 104629, 2025.

[3] M. N. H. Mamun, "Advancements in Machine Learning for Customer Retention: A Systematic Literature Review of Predictive Models and Churn Analysis," *Journal of Sustainable Development and Policy*, vol. 1, no. 1, pp. 250–284, 2025, DOI: 10.63125/9b316w70.