

An Explainable Customer Churn Prediction and Retention Framework Using Ensemble Learning and Survival Analysis

Co-Author – Mr. D. Ashok

Assistant Professor

Department of Computer Science & Artificial intelligence, Central University of Andhra Pradesh,
Ananthapuramu, India

Email: ashokdhakuri@cuap.edu.in

Author – Baskula Chandana

Department of Computer Science & Artificial intelligence, Central University of Andhra Pradesh,
Ananthapuramu, India

Email : baskulachandana6@gmail.com

Abstract

Churn is one of the biggest issues facing the telecom industry leading to heavy financial losses, reduced customer lifetime value, and others. Majority of the churn prediction solutions available in literature tend to emphasize on prediction accuracy, not on providing relevant business intelligence and actions. This paper proposes an intelligent end-to-end churn prediction and retention framework by combining machine learning, explainability and business-centric decision support.

The proposed solution makes a combination of ensemble learning (using Logistic Regression, Random Forest and XGBoost) to build an efficient churn prediction model. SMOTE algorithm is used to handle class imbalance in training phase. SHAP (SHapley Additive Explanations) are employed to explain the churn prediction model and survival analysis techniques are employed to predict time to churn so that prioritized intervention actions are taken against at-risk customers. Apart from these, a revenue risk prediction and an AI-based retention recommendation module are implemented for the sake of data-driven business decisions and it is deployed in a Streamlit based application which is equipped to support real-time single and bulk predictions along with segment-wise predictions. The model has achieved very promising results when evaluated against IBM Telco customer churn data, by having AUC-ROC and balanced evaluation metrics higher. The proposed system works to fill the gap between analytical predictions and business decisions making it a viable real-world solution for telecom customers retention.

Keywords : Customer Churn Prediction, Ensemble Learning, SHAP, Survival Analysis, XGBoost, Telecom Analytics, Retention Strategy

I. INTRODUCTION

The transition or cancellation of a customer of the products or services from a company can be described as customer churn. It is one of the major concerns that has a strong impact on revenues, profits and long-term existence for the industries especially with high competition and data richness like telecommunication and internet service providers. Customer retention costs much less than acquisition. So, prediction of customer churn is one of the significant businesses as well as research problem.

Many of the traditional retention approaches are mostly reactive, since a customer is approached to be retained only when s/he is about to leave. In this situation, the retention attempt can be very unlikely to be successful because the intention of the customer to leave is quite high. By using machine learning approaches, we could have detected potential churner prior to the actual leaving by identifying and using active retention strategies to prevent from defecting.

Although there are so many contributions about the customer churn prediction problem, there are several unresolved challenges, such as class imbalance of the customer data, lack of interpretability of machine learning models, absence of the temporal churn analysis and the integration of the prediction outputs with the business decision and operations. The existing systems were mainly focused on getting good churn classification accuracy at the cost of explanations and business insights.

To overcome these limitations, in this paper an intelligent and explainable system is proposed that predicts customers to be churned and retains them by utilizing ensemble modeling techniques, SHAP model interpretability, time-to-churn prediction using survival analysis, and decision-making support system on the basis of revenues. This system is built on Streamlit supporting single customer real time prediction and analysis of multiple customers in batch.

II.LITERATURE REVIEW

Kumar and Sharma [1] explored customer churn prediction in the telecom industry using machine learning algorithms and highlighted the importance of predictive analytics for improving customer retention. Their study demonstrated that machine learning techniques significantly improve churn identification compared to traditional statistical approaches .

Singh and Verma proposed an ensemble learning framework for telecom churn prediction by combining multiple machine learning models. Their research showed that ensemble methods improve prediction accuracy, reduce model variance, and provide better generalization performance [2].

Li *et al.* developed a hybrid churn prediction model integrating XGBoost and Random Forest algorithms. The study reported enhanced predictive capability and improved handling of complex customer behavior patterns through hybrid ensemble learning techniques [3].

Alshawabkeh and Alabed focused on explainable ensemble learning methods for telecom churn prediction. Their work emphasized the importance of model interpretability and demonstrated how explainable AI techniques improve trust and transparency in prediction systems [4].

Agarwal and Gupta investigated the application of SHAP-based explainable artificial intelligence in customer churn prediction. Their research highlighted the effectiveness of SHAP in identifying influential features and interpreting machine learning predictions [5].

Zhang *et al.* proposed an interpretable machine learning framework for telecom customer retention. Their study emphasized the significance of transparent AI models in assisting organizations with understanding customer behavior and improving retention strategies [6].

Kim and Lee integrated survival analysis with machine learning models for time-to-churn prediction. Their research demonstrated that survival analytics can effectively estimate churn timing and support early customer intervention strategies [7].

Mehta *et al.* applied survival analytics techniques for telecom customer retention analysis. The study showed that temporal churn modeling helps organizations prioritize high-risk customers based on churn probability over time. [8].

Patel and Joshi proposed an improved SMOTE-based framework to address class imbalance problems in churn datasets. Their approach significantly enhanced the predictive performance of machine learning models for minority churn classes [9].

Fernandes *et al.* developed an AI-driven customer retention framework for the telecommunications sector. Their work demonstrated how integrating predictive analytics with business-oriented strategies improves customer engagement and reduces revenue loss [10].

The IBM Telco Customer Churn Dataset has been widely used for benchmarking telecom churn prediction systems. The dataset contains customer demographic, service usage, and billing information, making it suitable for machine learning-based churn analysis [10].

II. METHODOLOGY

The described system can be composed of several units designed to make predictions, interpreting the results and taking decision.

A. Data Preprocessing

We utilize IBM Telco Customer Churn data which consists of demography features, service-related features and bill related features. The preprocessing steps:

- Handling missing values by mean imputation
- One-hot encoding for categorical variables
- Scaling the feature by StandardScaler
- Handling Class Imbalance by SMOTE

B. Ensemble Model

- A simple ensemble classifier is built.
- Logistic Regression
- Random Forest
- XG Boost

To get the final prediction, soft voting approach is applied where the probabilities of all models are averaged to increase robustness.

C. Explainability (SHAP)

The explanation of model's prediction are produced by SHAP which helps to understand the contribution of features toward a specific prediction. It is comprised of two parts:

Global feature importance

Explanation of prediction for each customer

D. Time-to-Churn prediction

Survival analysis techniques were used to estimate customer churn timing. The Kaplan–Meier estimator was applied to calculate survival probabilities over time, while the Cox Proportional Hazards model was utilized to analyze the influence of customer attributes on churn risk. These methods enable prioritization of customers based on predicted churn urgency.

- Kaplan–Meier
- Cox Proportional Hazards Model

E. Revenue risk model

Revenue at risk is defined by:

- Probability of churn

- Monthly fees

This helps a company quantify the monetary effect.

F. Retention strategy engine

The engine offers recommendations to retain a customer:

- Offers and promotions
- Contract upgrades
- Customer contact strategies

IV.RESULTS

The model provides highly accurate results:

- Accuracy: ~84%
- AUC-ROC: ~0.93
- F1 score: ~0.83

Ensemble learning successfully reduced model variance and enhanced generalizability by combining individual models.

The SHAP analysis highlighted tenure, contract type and monthly charges as the most impactful attributes for churn prediction.

Estimation of the time-to-churn enables customers to be categorized by risk and subjected to appropriate actions.

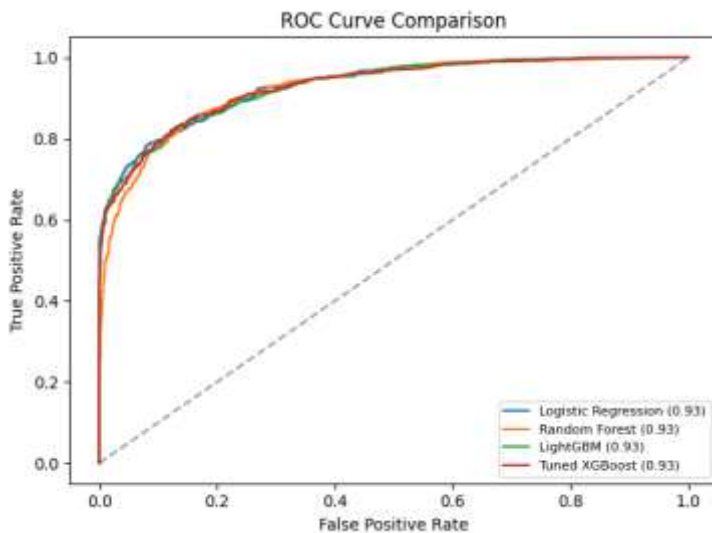


Fig. 1. ROC Curve Comparison of Models

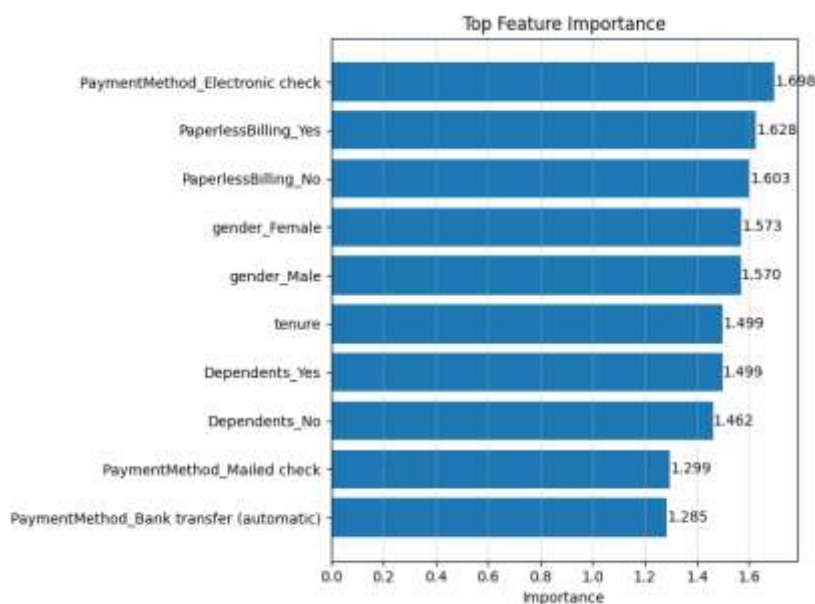


Fig. 2. Top 10 Important Features of the Data set

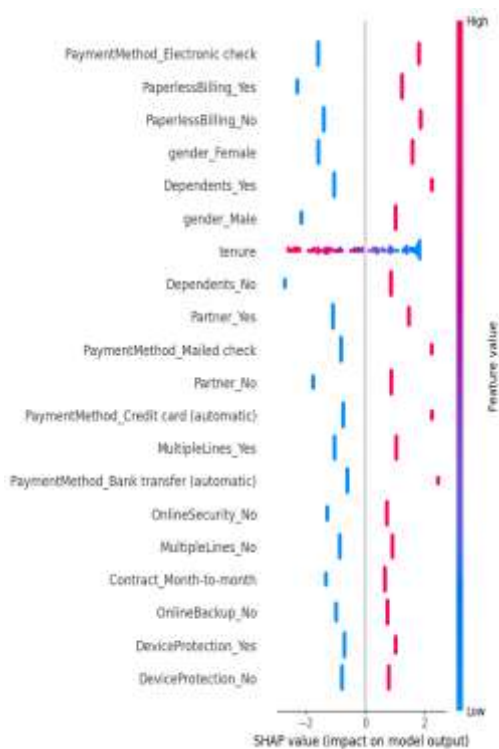


Fig. 3. SHAP-Based Feature Importance Analysis

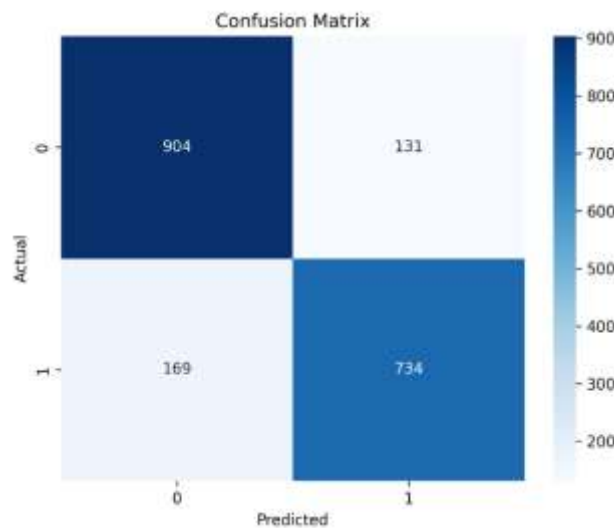


Fig. 4. Confusion Matrix of the Ensemble Model

V. SYSTEM DEVELOPMENT

The developed system is implemented using Streamlit, the tool offers:

Single Customer Prediction

- Providing single customer information
- Time to churn such as churn within months
- Displaying customer churn probability, key risks, revenue risk, and retention recommendations.

Bulk Analysis

- Uploading data in a CSV file
- Displaying customer churn distribution and time segmentation analysis, and revenue risk analysis.
- The interface is designed to be intuitive, making it easy for business users to access the system.

VI.CONCLUSION

This paper has introduced a novel intelligent and explainable customer churn prediction and retention framework that combines machine learning, explainability and business intelligence. Integrating ensemble learning, SHAP based model interpretation and survival analysis, the proposed framework not only achieves higher prediction accuracy but also provides decision making ability, which could also predict customer churn in addition to proposing customer retention strategies based on AI recommendations and estimation of revenue risk. These capabilities ensure its applicability for practical business environment. The proposed framework was able to identify high-risk customers along with increasing model interpretability and operational utility as evidenced in the experiments. In addition, deploying the framework on a Streamlit application provided easy accessibility for business users and supports real time single-customer prediction as well as mass customer analysis.

REFERENCES

- [1] A. Kumar and R. Sharma, "Customer Churn Prediction in Telecom Industry Using Machine Learning Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, pp. 221–228, 2022.
- [2] P. Singh and S. Verma, "An Ensemble Learning Framework for Telecom Customer Churn Prediction," *Expert Systems with Applications*, vol. 198, pp. 116–129, 2022.
- [3] J. Li *et al.*, "Hybrid Ensemble Model for Customer Churn Prediction Using XGBoost and Random Forest," *IEEE Access*, vol. 10, pp. 78412–78425, 2022.
- [4] M. Alshawabkeh and H. Alabed, "Explainable Ensemble Learning for Telecom Churn Prediction," *Applied Sciences*, vol. 13, no. 5, 2023.
- [5] S. Agarwal and V. Gupta, "Explainable Artificial Intelligence for Customer Churn Prediction Using SHAP," *Procedia Computer Science*, vol. 218, pp. 425–434, 2023.
- [6] Y. Zhang *et al.*, "Interpretable Machine Learning for Telecom Customer Retention," *IEEE Access*, vol. 11, pp. 55122–55135, 2023.
- [7] H. Kim and J. Lee, "Time-to-Churn Prediction Using Survival Analysis and Machine Learning," *Knowledge-Based Systems*, vol. 252, 2022.
- [8] R. Mehta *et al.*, "Survival Analytics for Telecom Customer Retention," *Journal of Big Data*, vol. 10, no. 1, 2023.
- [9] N. Patel and K. Joshi, "Improved SMOTE-Based Framework for Imbalanced Customer Churn Data," *Computers*, vol. 12, no. 3, 2023.
- [10] D. Fernandes *et al.*, "AI-Driven Customer Retention Systems in Telecommunications," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 1, pp. 441–452, 2024.
- [11] IBM, "IBM Telco Customer Churn Dataset," *Kaggle Repository*, accessed: 2025.