Customer Churn Prediction: Machine Learning Classification for Identifying Potential Churn

[1]Mr. Y. P. S. Hari Krishna, [2] K. Srinidhi, [3]D. L. Sravanthi, [4]Ch. Manoj, [5]Ch. Siva
[1]*Assistant Professor, Srinivasa Institute of Engineering and Technology*
[2345]*UG Scholars, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I10-194915-459

Abstract: Customer churn prediction has become a critical research area in the telecommunications industry due to its direct impact on revenue and customer retention strategies. This study proposes a machine learning-based classification model to identify customers who are likely to discontinue services. The dataset used in this research includes customer attributes such as tenure, monthly charges, contract type, and internet service information. Data preprocessing techniques, including handling missing values, categorical feature encoding, and feature scaling, were performed to enhance model performance. Several classification algorithms were evaluated, and Logistic Regression was selected based on its predictive accuracy and interpretability. The developed model was integrated into a web-based interface to facilitate real-time prediction. Experimental results demonstrate that the proposed approach effectively identifies potential churn customers, thereby supporting data-driven decision-making for improving customer retention in the telecom sector.

Keyword: Churn Prediction, Classification, Logistic Regression, Machine Learning, Telecom Industry.

I. INTRODUCTION:

Customer churn is a major concern in the telecommunications industry, as losing customers directly impacts company revenue and growth. Retaining existing customers is more cost-effective than acquiring new ones, making churn prediction an important research problem. Early identification of potential churn customers allows organizations to take preventive actions and improve customer retention strategies.

Telecom companies generate large volumes of customer data, including billing information, service usage details, and contract history. Analyzing such data using traditional methods is challenging. Machine learning techniques provide efficient solutions for identifying patterns and predicting customer behavior based on historical data.

In this study, a supervised machine learning approach is proposed to predict customer churn. Logistic Regression is used as the primary classification algorithm due to its simplicity and effectiveness in binary classification problems. The developed system aims to assist telecom organizations in identifying high-risk customers and supporting data-driven decision-making.

II. LITERATURE SURVEY

Customer churn prediction has been an active area of research for many years, especially in industries such as telecommunications, banking, insurance, and online subscription services. Several researchers have focused on identifying patterns in customer behavior that indicate the possibility of churn. Early studies mainly used statistical methods such as logistic regression and basic decision tree models. These techniques were simple to implement and easy to interpret, but their performance was limited when handling complex datasets.

With the development of machine learning techniques, more advanced classification algorithms have been applied to churn prediction problems. Models such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, and Random Forest have shown improved prediction accuracy compared to traditional methods. Among these, ensemble models like Random Forest are often preferred because they reduce overfitting and handle nonlinear relationships effectively.

Recent research also highlights the importance of data preprocessing and feature engineering. Handling missing values, encoding categorical variables, and scaling numerical features significantly improve model

performance. Some studies also address the issue of class imbalance, as churn datasets often contain fewer churn cases compared to non-churn cases. Techniques such as resampling and synthetic data generation are used to balance the dataset.

Although many studies report high accuracy, some models lack interpretability and practical usability for business applications. Therefore, there is still a need for a structured and efficient churn prediction framework that provides reliable results while remaining understandable for decision-makers.

III. SYSTEM ARCHITECTURE

The system architecture for the proposed customer churn prediction model follows a structured workflow starting from data collection to business decision-making. The overall process is divided into multiple stages to ensure accurate and reliable prediction.

The first stage is Customer Data Collection. This includes customer-related information such as usage details, billing history, and demographic data. These attributes form the input dataset for the prediction model.

The second stage is Data Preprocessing. In this step, the collected data is cleaned and prepared for analysis. Missing values are handled, categorical variables are encoded, and numerical features are normalized where required. Proper preprocessing improves the quality of the dataset and enhances model performance.

The third stage is Feature Selection. Not all attributes contribute equally to churn prediction. Therefore, important customer features are selected based on their relevance and impact. This helps in reducing model complexity and improving prediction accuracy.

The next stage is the Machine Learning Model. Classification algorithms such as Logistic Regression, Decision Tree, and Random Forest are applied to the processed dataset. The model is trained using historical customer data to learn churn patterns.

The final stage is Churn Prediction Output, where the model classifies customers into two categories: Churn or No Churn. Based on the prediction results, businesses can take appropriate actions.

The predicted output supports the Business Decision stage, where retention strategies such as promotional offers, personalized communication, and service improvements are implemented to reduce customer attrition.

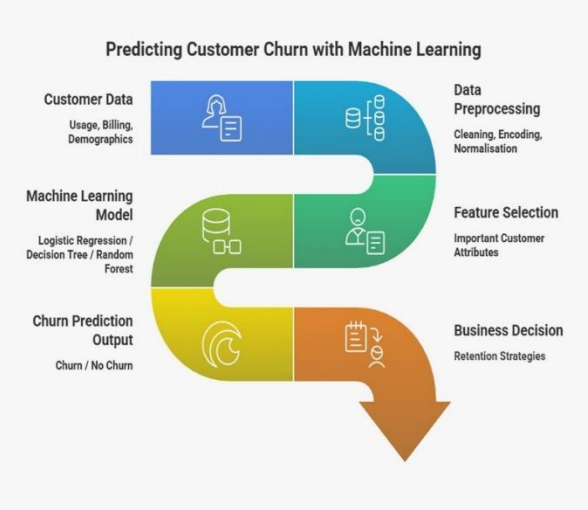

Fig. 1. System Architecture of Customer Churn Prediction Model

IV. METHODOLOGY

The proposed customer churn prediction system follows a structured machine learning approach consisting of data preparation, preprocessing, model training, and evaluation. The overall process ensures accurate identification of potential churn customers.

A. Dataset Description: The dataset used in this study contains historical customer information collected from a telecom service provider. It includes features such as customer tenure, monthly charges, total charges, contract type, payment method, internet service type, and other service-related attributes. The target variable indicates whether the customer has churned or not.

B. Data Preprocessing: Data preprocessing is performed to improve the dataset quality before model training. Missing values are handled appropriately to avoid inconsistencies. Categorical variables are converted into numerical form using encoding techniques. Irrelevant or redundant features are removed to reduce complexity. The dataset is then divided into training and testing sets for model evaluation.

C. Model Implementation: Machine learning classification algorithms are applied to predict customer churn. Logistic Regression is selected as the primary model due to its simplicity, interpretability, and effective

performance for binary classification problems. The model learns patterns from historical customer behavior and builds a predictive framework.

D. Model Evaluation: The performance of the trained model is evaluated using metrics such as accuracy, precision, recall, and F1-score. A confusion matrix is used to analyze correct and incorrect classifications. These evaluation measures help in assessing the effectiveness of the proposed churn prediction system.

V. RESULTS AND DISCUSSION

The performance of different classification models was evaluated using accuracy, precision, recall, and F1-score, which measure overall correctness and the ability to identify churn and non-churn customers. The dataset was split into training and testing sets to ensure reliable evaluation and to prevent overfitting. Logistic Regression, Decision Tree, and Random Forest were implemented and compared to determine the most effective model for predicting customer churn, balancing both accuracy and interpretability for practical use. Evaluating multiple models allows us to understand the strengths and weaknesses of each approach and select the one that performs best on real-world data. The results provide valuable insights for telecom companies to make data-driven decisions in customer retention strategies.

Table I: Performance Comparison of Classification Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	81.97	68.31	59.52	63.61
Decision Tree	70.83	44.95	45.31	45.13
Random Forest	78.99	64.53	45.84	53.61

Table II: Confusion Matrix – Logistic Regression

Confusion Matrix	Predicted No Churn	Predicted Churn
Actual No Churn	933	103
Actual Churn	151	222

VI. OUTPUTS

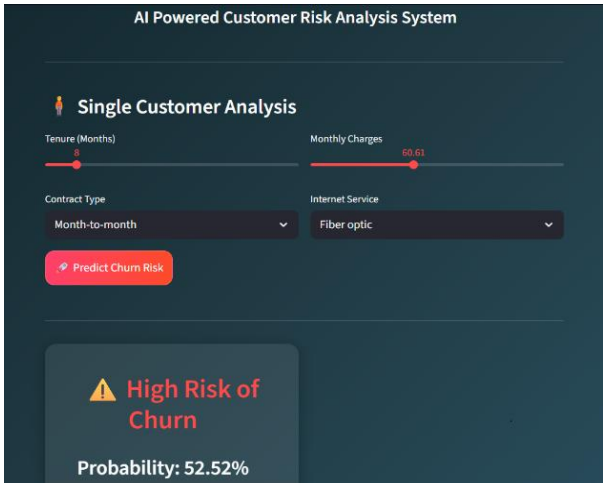

Fig. 2. Single Customer Churn Prediction Output

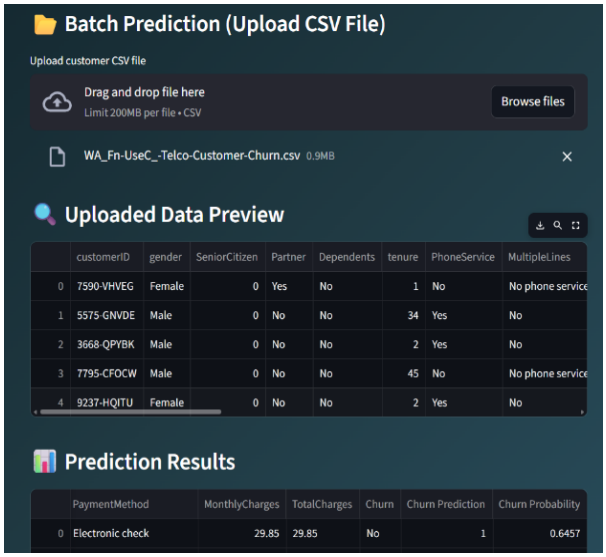

Fig. 3. Batch Customer Churn Prediction Output

VII. DISCUSSION

From Table I, Logistic Regression achieved the highest accuracy of 81.97%, outperforming Decision Tree and Random Forest models. It also showed balanced precision and recall, indicating strong performance in identifying both churn and non-churn customers. Random Forest had competitive accuracy (78.99%) but lower recall, while Decision Tree performed the worst, likely due to overfitting on the training data.

The confusion matrix (Table II) shows that Logistic Regression correctly classified most non-churn

customers and a significant portion of churn customers, demonstrating its effectiveness in practical scenarios. The system output (Figures 1–2) visually presents the predicted churn status, with red highlighting high-risk customers and green indicating low-risk customers, making it easy for decision-makers to identify users requiring attention. Overall, the results confirm that Logistic Regression provides a reliable and interpretable model for customer churn prediction, enabling telecom companies to implement proactive retention strategies and improve revenue by focusing on at-risk customers.

VIII. CONCLUSION

his study presented a machine learning-based approach for predicting customer churn in the telecommunications sector. Logistic Regression, Decision Tree, and Random Forest models were implemented and evaluated using accuracy, precision, recall, and F1-score. Among these, Logistic Regression achieved the highest accuracy of 81.97%, with balanced precision and recall, indicating a strong ability to correctly identify both churn and non-churn customers. The confusion matrix results confirm that the model effectively distinguishes high-risk customers, providing actionable insights for targeted retention strategies.

The results highlight the importance of proper data preprocessing, feature selection, and model evaluation in building reliable churn prediction systems. By implementing this predictive framework, telecom organizations can make data-driven decisions, such as offering personalized promotions, enhancing customer support, and improving service quality, ultimately reducing churn and increasing long-term profitability. The model's simplicity, interpretability, and efficiency make it suitable for real-world deployment, and it serves as a foundation for integrating more sophisticated predictive analytics in future applications.

IX. FUTURE SCOPE

The proposed churn prediction system can be extended and improved in several ways to enhance predictive accuracy and business impact. Future research can incorporate additional features, such as customer interaction logs, call center records, complaint history, social media sentiment, and real-time usage patterns, which may provide deeper insights into customer behavior. Advanced machine learning techniques, including ensemble learning, gradient boosting, or deep neural networks, can be explored to capture complex patterns and improve prediction performance.

Furthermore, integrating the prediction model with an automated decision-support system can allow telecom companies to take immediate retention actions, such as personalized offers or proactive notifications, based on real-time predictions. The system could also be adapted to other service-oriented industries, including banking, insurance, and subscription-based platforms, where customer retention is a critical challenge. Continuous learning and adaptive models can ensure sustained performance under dynamic market conditions, making the system more robust and practically valuable for long-term operational strategies.

REFERENCES

[1] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," IEEE Trans. Neural Networks, vol. 4, pp. 570–578, Jul. 1993.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[3] F. F. Reichheld and P. Schefter, "E-loyalty: Your secret weapon on the web," Harvard Bus. Rev., vol. 78, no. 4, pp. 105–113, 2000.

[4] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in a big data platform," J. Big Data, vol. 6, no. 1, p. 28, 2019.

[5] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, "Improved churn prediction in the telecommunications industry using data mining techniques," Appl. Soft Comput., vol. 24, pp. 994–1012, 2014.

[6] W. Verbeke, D. Martens, and B. Baesens, "Social network analysis for customer churn prediction," Appl. Soft Comput., vol. 14, pp. 431–446, 2014.

[7] 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," J. Mark. Res., vol. 43, no. 2, pp. 204–211, 2006.

[8] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, "Computer-assisted customer churn management: State-of-the-art and future trends," Comput. Oper. Res., vol. 34, no. 10, pp. 2902–2917, 2007.

[9] V. Chang, K. Hall, Q. A. Xu, F. O. Amao, M. A. Ganatra, and V. Benson, "Prediction of customer churn behavior in the telecommunication industry using machine learning models," Algorithms, vol. 17, no. 6, 2024.

[10] A. Sikri et al., "Enhancing customer retention in telecom industry with machine learning driven churn prediction," Sci. Rep., vol. 14, art. 13097, 2024.

[11] R. M. Wahul, A. P. Kale, and P. N. Kota, "An ensemble learning approach to enhance customer churn prediction in the telecom industry," Int. J. Intell. Syst. Appl. Eng., vol. 11, no. 9s, 2023.