Customer Churn Prediction Model

R Anisha, Shravani G, Mithan Gowda M K, Manogna R, B Uma
Computer Science and Engineering, Malnad College of Engineering, Hassan-573202, India

Customer churn prediction is the practice of identifying which clients are most likely to stop utilizing a company's services in the near future. It helps businesses take proactive steps to retain vulnerable customers and reduce revenue loss.

***Abstract*—Given the high cost of gaining new customers and the comparatively low switching costs for consumers, client retention has become more and more crucial in today's fiercely competitive telecom sector. The profitability and long-term sustainability of a telecom provider are directly impacted by churn, the phenomenon when customers discontinue utilizing a company's service. The goal of this project is to use machine learning techniques to create a reliable forecast model for customer attrition. The study examines customer behavior and service- related characteristics such as contract type, duration, monthly rates, internet service, and technical support utilization using the well-known Telco Customer Churn dataset. To get the dataset ready for model training, a thorough data pretreatment pipeline was put in place. This pipeline involved encoding category variables, managing missing values, and normalizing numerical characteristics. Several classification techniques, such as Random Forest, Support Vector Machine (SVM), Logistic Regression, and Gradient Boosting Classifier, were analyzed and assessed.**

***Index Terms*—AI in the Classroom Speech-to-Text, Natural Language Processing, GPT, Whisper, Mock Interview, HR Inter- view Simulation, Interview Feedback System, and Educational Technology**

I. INTRODUCTION

Retaining clients has become one of the most important aspects of ensuring business success and growth in the quickly changing telecom sector. Telecom businesses are under pres- sure to draw in new clients while keeping their current clientele as the market gets increasingly congested. A major problem that impacts a brand's revenue and reputation is customer churn, which is the phenomenon when customers stop their subscriptions or go to competitors. According to research, it can cost five times as much to acquire a new customer as it does to keep an existing one. Understanding and minimizing client attrition should therefore be a major focus for service providers. Because of the fierce competition, businesses need to come up with creative ways to spot churn early on and respond with proactive, tailored engagement initiatives. In this sense, predictive analytics driven by machine learning offers a practical means of making retention-focused business decisions and accurately forecasting attrition.

Customer churn can be influenced by a wide range of factors, including pricing strategies, customer satisfaction, service quality, customer service experiences, and even promotions from competitors. Some customers may leave because they are dissatisfied with the service, but others may be persuaded by more enticing offers from rival companies or the belief that their current subscription isn't worth much. It might be challenging for conventional statistical models to adequately capture and analyze churn behavior since these elements frequently interact in intricate and non-linear ways. However, because machine learning algorithms can handle massive amounts of heterogeneous data, uncover hidden patterns, and generate data-driven predictions, they are ideal for this purpose. This research intends to construct a powerful machine learning model that can properly forecast whether a telecom client is likely to depart by analyzing historical data on customer demographics, service consumption patterns, account information, and billing details. Assisting telecom operators in identifying high-risk clients and acting promptly to increase client retention is the aim.

About 7,000 customers' personal details, including gender, age, tenure, service plans (such as internet, phone, and streaming), contract type, payment method, and monthly prices, are included in the publicly accessible Telco Customer Churn dataset. In order to

convert raw input into a format appropriate for machine learning, the data preparation phase is essential. This entails addressing missing values, utilizing encoding techniques (e.g., label encoding or one-hot encoding) to convert categorical variables, scaling numerical features to guarantee uniformity, and conducting exploratory data analysis (EDA) to comprehend the distribution and the relationship between the characteristics. In addition to offering visual insights into the data, EDA assists in identifying trends that could guide feature engineering and model selection. This preparation process optimizes the dataset for training and evaluating predictive models.

This study compares and analyzes a number of machine learning methods, such as Random Forest, Logistic Regression, Decision Trees, Support Vector Machine (SVM), and Gradient Boosting Classifier. Every model has advantages and disadvantages with regard to forecast accuracy, computing efficiency, and interpretability. To ensure a thorough evaluation of the models' efficacy, performance measurements like accuracy, precision, recall, F1-score, and the ROC-AUC curve are employed. Given its ability to manage both linear and non- linear interactions and minimize overfitting through sequential learning, gradient boosting in particular is anticipated to per- form well.

Ultimately, this research makes a practical and technical contribution to the field of customer analytics.

II. RELATED WORKS

A. Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform

Description: This study, which focuses on SyriaTel, a major telecom provider in Syria, comprehensively examines the application of machine learning (ML) techniques to anticipate customer attrition in the telecom industry. Using big data technology, the study processed and examined a vast dataset of almost 70 terabytes (TB) gathered over a nine-month period, acknowledging the enormous difficulty of client retention in the fiercely competitive telecom sector. Creating an accurate churn prediction algorithm that could identify customers who were most likely to leave the network was the main objective.

B. Enhancing Customer Churn Prediction in Telecommunications: An Adaptive Ensemble Learning Approach

Description: Using an adaptive ensemble learning frame- work, this study suggests a novel way to improve customer churn prediction in the telecom sector. The necessity for precise and reliable predictive models has increased since telecom service providers continue to face a serious problem with customer attrition, endangering both profitability and client lifetime value. The authors suggest a method for enhancing prediction performance that combines several machine learning algorithms into a coherent group, taking advantage of each one's distinct benefits. The methodology builds a high- performing meta-model that can capture intricate, nonlinear patterns in consumer behavior by carefully combining the outputs of multiple base learners rather than depending on a single model.

C. Application of Machine Learning Techniques for Churn Prediction in the Telecom Business

Description: This study evaluates many machine learning algorithms to predict customer attrition in the telecom sector, with a focus on evaluating the Random Forest (RF) algorithm's effectiveness. Churn prediction is a crucial component of strategic CRM since the telecom industry is always having trouble retaining clients. By accurately identifying people who are most likely to discontinue using their services, telecom providers can take preemptive measures to retain consumers and minimize revenue loss. The study looks at how well different machine learning models perform when applied to real consumer data and finds Random Forest to be a very helpful tool for this.

D. Telecom Customer Churn Forecasting Using Machine Learning: A Data-Driven Predictive Framework

Description: The study" Telecom Customer Churn Fore- casting Using Machine Learning: A Data-Driven Predictive Framework" looks at the application of several machine learning algorithms to forecast customer churn in the telecom industry. In addition to being extremely precise, the authors hope to create a predictive model that is both comprehensible and suitable for real-world application by telecom carriers. The telecom sector is under a lot of pressure to retain consumers due to intense competition and expensive acquisition costs, which makes churn prediction a

helpful strategy. The primary goal of this study is to develop and evaluate a comprehensive churn prediction methodology using structured customer data.

E. Deep Churn Prediction Method for Telecommunication Industry

Description: The research paper" Deep Churn Prediction Method for Telecommunication Industry" looks at a variety of machine learning and deep learning approaches to develop a dependable and incredibly accurate model for predicting customer attrition in the telecom sector. Due to fierce competition and high acquisition costs, telecom service providers continue to have serious concerns about client retention. The study focuses on applying advanced predictive analytics to address this issue. The study contrasts cutting-edge deep learning architectures, ensemble learning methods, and traditional classification models in order to determine the optimal approach for churn prediction.

F. A Hybrid Deep Learning Approach for Customer Churn Prediction in Telecom Sector

Description: This paper proposes a hybrid deep learn- ing framework that combines long- and short-term memory (LSTM) and neural networks (CNN) to improve the accuracy of customer churn prediction in the telecom industry. Recognizing that consumer behavior exhibits both temporal and spatial patterns, the study uses CNN to extract spatial features and LSTM to capture sequential relationships in customer data. The methodology was evaluated using a large telecom dataset that contained attributes such as call detail records, service consumption, and payment history. The hybrid model performed better than both solo machine learning algorithms and conventional deep learning models with high precision and F1 scores.

G. Customer Churn Prediction in Telecom Industry Using Ensemble Learning

Description: This study investigates the effectiveness of ensemble learning techniques to raise the accuracy of churn prediction models in the telecom sector. The researchers used the Telco Customer Churn dataset and preprocessed it using label encoding, feature scaling, and managing null values. Trained models such as Bagging, AdaBoost, and Gradient Boosting were contrasted with conventional classifiers. Gradient Boosting achieved the maximum accuracy of 83.7 percent, demonstrating that the ensemble models outperformed the individual techniques. The feature importance study revealed that contract type, tenure, and monthly costs were important churn indicators. However, the applicability of the work was restricted by its lack of focus on model deployment and interpretability.

TABLE I SUMMARY OF RELATED WORK IN CUSTOMER CHURN PREDICTION

Paper Name	Description	Methodology Used	Models Used	Accuracy/Results
1.Telecom Customer Churn Prediction Applying machine learning (ML) to big data platforms	concentrating on SyriaTel data to predict churn using ML and bid data tools.	Predictive modeling, feature engineering, big data processing	Random Forest, Decision Tree, GBM, and RF	RF accuracy: 78.2%
2. Enhancing customer Chrun Prediction in Telecom	To improve customer churn prediction in telecom, an adaptive ensemble framework is proposed.	Combines base learners using meta-model to capture nonlinear-patterns	AdaBoost, Stacking Ensemble, SVM	Stacking ensemble accuracy: 84.3%
3. Application of Machine Learning Techniques for Telecom Churn Prediction	Examines ML algorithms on actual telecom data with an emphasis on RF. Preprocessing data.	Analyzing the significance of features, and comparing models.	Random Forest, SVM, KNN	Random Forest accuracy: 80.4%
4. Telecom Customer Churn Forecasting Using ML Framework	Develops High-Performing, Interpretable Churn Prediction Model.	structured data modeling that is tested on actual datasets.	XGBoost, Decision Tree, Logistic Regression, and XG-Boost	XGBoost accuracy: 81.7%
5.Deep Churn Prediction Method for Telecom Industry	Examines Deep Learning, Ensemble, and Machine Learning for Robust Churn Prediction.	A comparison between deep learning and conventional models.	CNN, LSTM, RF, Gradient Boosting, and LSTM	LSTM accuracy: 86.1%
6. Hybrid Deep Learning for Telecom Churn Pre-	Temporal spatial patterns by combining CNN and	hybrid model for feature extraction from CDRs,	CNN + LSTM	F1-Score accuracy : 88.5%

diction Models	LSTM.	usage, and payments.		
7. Telecom Industry Customer Churn Prediction Using Ensemble Learning	Improves churn detection by applying ensemble models to Telco datasets.	Model evaluation, feature scaling, and label encoding.	Gradient, AdaBoost, and Bagging, Boosting	Gradient Boosting accuracy: 83.7%

PROPOSED METHODS

Collecting and understanding customer data from the Telco Customer Churn dataset, including demographics, usage pat- terns, and contract details, is the initial stage in creating the proposed customer churn prediction model. After initial research using Python libraries, data preprocessing is done by handling missing values, encoding category features, and scaling numerical ones. Feature engineering is used to se- lect important features and generate new variables. Several machine learning models, such as Random Forest, XGBoost, LightGBM, and Logistic Regression, are trained with class imbalance addressed by SMOTE and class weighting. The model's performance is evaluated using ROC-AUC, F1 score, recall, accuracy, and precision. Interpretability is provided by SHAP values, which highlight significant churn indications. The model is deployed using Flask as a REST API with Docker on cloud, and risk categories and forecasts are dis- played on a dashboard built with React. Continuous monitoring after deployment guarantees model reliability, while future plans call for integrating time series and natural language processing to improve prediction.

The solution assists telecom operators in filtering customer segments, interactively analyzing churn risks, and making educated retention decisions using an intuitive interface. Adapt- ability to changing client behaviors is ensured via real-time feedback loops and continual model upgrades. By incorporating LSTM for sequential behavior analysis and NLP for customer support interaction analysis, predictive accuracy and business value can be further enhanced. Decision-making is more transparent since stakeholders can understand the under- lying reasons for churn predictions according to the model's advanced explainability approaches. Extensive cross-validation and hyperparameter adjustment ensure robust generalization across different client profiles.

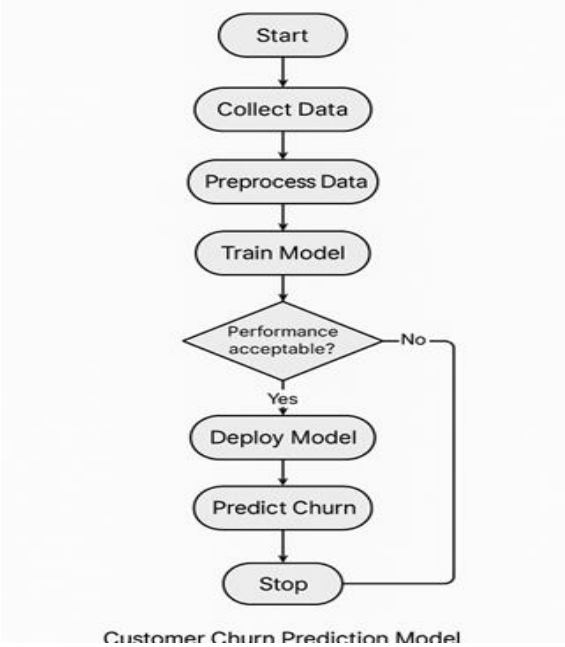

Fig. 1. Flowchart for Customer Churn Prediction

III. CONCLUSION

client attrition, which affects both profitability and client retention, is one of the largest problems facing the telecom industry. In order to proactively predict and control customer attrition, this project aims to propose a machine learning- based framework that integrates data analysis, model building, and deployment strategies. The system is designed to analyze customer demographics, service consumption, and billing trends using the Telcom Customer Churn dataset. Our pipeline combines extensive data preprocessing techniques including scaling, categorical encoding, and missing value imputation with purposeful feature engineering to improve the quality of inputs for modeling. To guarantee equitable model learning, class imbalance problems—a prevalent issue in churn datasets—are addressed using SMOTE and class weighting techniques.

We'll try both more traditional classifiers like Logistic Regression and more potent ensemble methods like Random Forest, XGBoost, and LightGBM. Evaluation criteria including precision, recall, F1-score, and ROC-AUC will be used to gauge the model's performance.

Hyperparameter adjustment and cross-validation will be employed to improve generaliz- ability. To identify the primary reasons for churn and make it easier to generate insightful information, SHAP values will be utilized for interpretability. Telecom operators can use these insights to spot patterns, including how contract type or monthly prices affect the chance of turnover. Real-time predictions and visual analytics of customer groups and churn probability will be made possible by the Flask API coupled into the React-based frontend dashboard that will serve as the user interface.

The solution will be deployed using Docker for containerization and hosted on cloud platforms like AWS or Heroku to offer scalable and reliable access. Telecom businesses will be able to perform real-time churn risk assessments once the system is live, which will help them implement retention tactics on time. It is also meant to include a monitoring system that tracks data drift and performance over time to ensure model accuracy and relevance after deployment. In later phases, NLP-based modules will be added to handle customer support interactions for a more comprehensive churn analysis, and time-series models, like LSTM, will be added to assess sequential usage trends. Consequently, our proposed framework establishes the foundation for a practical, data- driven approach to customer retention in the telecom industry.

REFERENCE

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova," BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Human Language Technologies: Proceedings of the 2019 Conference of the Association for Computational Linguistics' North American Chapter*, vol. 1, pp. 4171–4186, 2019. 10.18653/v1/N19-1423 is the DOI.

[2] D. Y. Dissanayake, V. Amalya, R. Dissanayaka, L. Lakshan, P. Samarasinghe, M. Nadeeshani, and P. Samarasinghe," AI-based Behavioural Analyser for Inter- views/Viva," in *IEEE 16th International Industrial and Information Systems Conference (ICIIS) 2021*, 2021, pp. 1–6. 10.1109/ICIIS53135.2021.9660757 is the DOI.

[3] M. Srivastava, A. Sharma, and R. Singh," AI Chatbot for Job Interview and Candidate Recruitment Based on Automatic Answer Evaluation Using WordNet," *IJCA*, vol. 182, no. 14, pp. 34–39, 2021.

[4] D. Roy and P. Kumar," Virtual Simulation of Technical Interviews Using Machine Learning and NLP," in *Proceedings of the 5th (ICCMC)*, 2023, pp. 512–517.

[5] P. Jain and S. Verma," AI-based Behavioural Analyser for Interviews/Viva using Speech and Facial Emotion Recognition," *v*ol. 12, no. 8, pp. 230–237, 2021.

[6] S. Banerjee and A. Chakrabarti," The Tech-Talk Balance: What Technical Interviewers Expect from Technical Candidates,", vol. 65, no. 2, pp. 112–119, 2022.

[7] H. Li, J. Zhang, and L. Gao," Automatic Scoring of Spoken Responses Using Deep Learning," *IEEE Trans- actions on Learning Technologies*, vol. 14, no. 4, pp. 476–488, 2021.

[8] E. Fernandez and A. Gupta," Affective Computing in Virtual Interviews: Emotion Recognition for Candidate Evaluation," in *Proceedings of the IEEE Conference on Human-Centered AI*, 2022, pp. 233–240.

[9] S. Jadhav, R. Pawar, and A. Kamble," AI-Based Multi- modal Emotion and Behavior Analysis of Interviewees," *International Journal of Scientific Research in Engineering and Management*, vol. 5, no. 3, pp. 101–107, 2023.

[10] X. Lian, Y. Qiu, and L. Deng," Explainable Multimodal Emotion Recognition in Human-Computer Interaction," *arXiv preprint arXiv:2306.15401*, 2023.