Customer Churn Prediction in Telecom Industry

R. Anisha, Manogna. R, Shravani. G, B. Uma, Mithan Gowda. M.K
*Department of Computer Science and Engineering Malnad College of Engineering*Hassan, India
*Professor, Department of Computer Science and Engineering Malnad College of Engineering*Hassan, India

***Abstract*—Client retention really matters in today's telecom world. It's expensive to win new customers, and honestly, people can switch providers pretty easily now. When customers leave—what we call churn—it hits a telecom company right where it hurts: profits and long-term stability. We set out to build a machine learning model that could spot when customers were ready to leave. We looked at everything—how long they'd been around, what kind of contract they had, their monthly bills, internet details, even how often they called tech support. The Telco Customer Churn Dataset gave us a good starting point, but honestly, it wasn't enough. So we made our own custom dataset, adding features that really captured what we wanted to dig into. Getting the data ready took some work. We coded the categories, filled in missing spots, and normalized the numbers so everything lined up for training. Then tried out several classification methods—Logistic Regression, Random Forest, and XGBoost. We didn't just run the models; we dug into how each one performed and what worked best.**

Keywords—Customer Churn Prediction, Machine Learning, telco dataset, Data preprocessing, Ensemble models, SMOTE, SHAP, Telecom analytics, Predictive Modelling.

I. INTRODUCTION

In today's telecom world, things change fast. Keeping your customers isn't just important—it's crucial if you want your business to grow. With more and more players jumping into the market, telecom companies feel the squeeze to win over new customers while holding tight to the ones they already have. When customers leave—what the industry calls churn—it hits revenue hard and can damage a company's reputation too. And since it costs almost five times more to bring in a new customer than to keep an old one, it just makes sense to focus on retention. That's why companies need to spot the warning signs of churn as early as possible and step in with smart, well-timed engagement. Here, machine learning steps up as a practical tool. With predictive analytics, telecoms can make smarter retention decisions and actually predict who's at risk of leaving.

Churn isn't caused by just one thing. It's a tangled mix—pricing, customer service, product quality, what competitors are offering, and more. Some people bail because they're unhappy with the service. Others get tempted by a better deal from a rival. The truth is, old-school statistical models just can't handle all these messy, non-linear connections. Machine learning, on the other hand, thrives here. It chews through massive, messy datasets and finds the hidden patterns that matter, leading to stronger, more accurate predictions. In this paper, we lay out a predictive machine learning framework designed to predict which customers might leave, using historical data on things like service usage, how long customers have stayed, and billing info. This kind of insight lets telecom operators pinpoint high-risk customers and take action, boosting retention and cutting down on churn.

We worked with a dataset of about 10,000 telecom subscribers. It's got all the basics—how long each person's been with the company, their contract details, monthly bills, which services they use (think internet, phone, streaming), and how they pay. Before diving into modeling, we had to clean things up. That means filling in missing values, converting categories to numbers (using stuff like label or one-hot encoding), and making sure all the features are on the same scale. Once that's done, we rolled up our sleeves for some exploratory data analysis. This is where you really poke around—looking for patterns, checking how different features connect, and pulling out little nuggets that help with feature engineering and picking the right models. EDA and

preprocessing are basically the warm-up before you get to the real action with predictive algorithms.

We tried out a few different machine learning models: Logistic Regression, Random Forest, and Gradient Boosting Classifier. They're all a bit different. Some are easier to explain, others are more accurate or just plain faster. To keep things fair, we measured them using accuracy, precision, recall, F1-score, and ROC-AUC. Honestly, Gradient Boosting usually comes out on top. It's flexible, handles both simple and complex relationships, doesn't overfit easily, and learns step by step—so you end up with solid predictions. In the end, our research makes it clear: machine learning isn't just good at spotting churn, it actually helps reduce it. With these tools, companies can make smarter moves to keep their customers and grow their business.

II. LITERATURE SURVEY

A. Predicting Customer Attrition in the Telecom Sector through Machine Learning on a Big Data Framework

This study zooms in on SyriaTel, a major telecom player in Syria, and explores how machine learning can flag customers who are ready to jump ship. We're talking about nearly 70 terabytes of data—nine months' worth—crunched using big data tech. That's a massive pile, but it's worth it. The idea is straightforward: create a model that pinpoints which subscribers are most likely to leave. With machine learning and big data working together, telecom companies don't just react—they stay a step ahead, hang on to more customers, and keep churn in check.

B. Analyzing and Forecasting Customer Churn Patterns in the Telecom Industry with Machine Learning Techniques

This paper breaks down how machine learning can help predict customer churn in telecom, with a real focus on how well the models perform and how easy they are to understand. The authors tested several supervised learning algorithms—Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and SVM—to see which one does the best job at finding likely churners. They didn't just stop at picking models; they worked through data prep, feature selection, and tuning to make sure the results hold up. The cool part? They used explainable AI tools like LIME and SHAP to actually show what's driving customers to leave. When you look at the numbers—accuracy, precision, recall, F1-score, and ROC-AUC—the ensemble models pulled ahead, delivering sharper predictions and better insights about what makes people jump ship. In the end, the study offers a practical, interpretable framework that helps telecom companies build smarter retention strategies and manage customer relationships better.

C. Improving Customer Retention Strategies in the Telecom Sector through ML-Based Churn Prediction Models

This research goes headfirst into the messy world of telecom data, where predicting who's about to leave is never simple—especially when the data's all over the place. To fix that, the authors came up with a new ratio-based balancing method that helps machine learning models get better at spotting churners. They tried out a bunch of algorithms—Perceptron, Multi-Layer Perceptron, Naive Bayes, Logistic Regression, K-Nearest Neighbor, Decision Tree, Gradient Boosting, and XGBoost. XGBoost took the crown, beating out old-school over-sampling and under-sampling. Bottom line: the right algorithm and balancing method can seriously boost your churn predictions.

D. Integrating Machine Learning and Social Network Analysis for Telecom Churn Prediction on Big Data Platforms

This study shakes up churn prediction by blending machine learning with social network analysis (SNA) features. Instead of just looking at customer details, it pays attention to how customers connect and interact. The result? A model with a 93.3% AUC, leaving traditional models in the dust. By adding SNA, the research shows you can pick up on subtle patterns—relationships and behaviors—that regular models miss, giving telecom companies a smarter way to understand and fight churn.

E. Understanding Telecom Customer Churn through Explainable Artificial Intelligence (XAI) Techniques

This work puts transparency front and center. Using explainable AI tools like LIME and SHAP, the study opens up the black box of machine learning, showing both big-picture trends and individual reasons behind churn predictions. This level of interpretability gives telecom companies a clear view of what's driving customers away—think service quality, satisfaction,

and more—so they can build retention strategies that actually fit what customers want and need.

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

This paper introduces an adaptive ensemble learning setup that pulls together several base models—XGBoost, LightGBM, LSTM, Multi-Layer Perceptron, and SVM—using a stacking ensemble .

III. METHODOLOGY

This Customer Churn Prediction system zeroes in on telecom customers who look likely to leave, so operators can step in early with smart retention moves. The project runs the whole machine learning pipeline, end to end. That means everything from grabbing the raw data, cleaning it up, and engineering features, to training models, tuning them, making sense of the results, and finally, rolling out the finished model. The dataset covers a mix of customer details—stuff like gender, age, how long they've been with the company, contract type, monthly and total charges, whether they stream movies, use online or tech support, and if they have multiple lines. Each of these features matters when it comes to figuring out who might churn. First up, the data needs some love. That involves filling in missing values, encoding categories (label or one-hot encoding, depending), and normalizing numbers so everything lines up. Churn data in telecom is always imbalanced—way more people stay than leave—so the system uses SMOTE to balance things out. This step helps the models actually learn what a churning customer looks like, instead of just guessing "no churn" all the time. For the modeling, it's a three-way race: Logistic Regression, Random Forest, and XGBoost. These algorithms are solid picks for both linear and nonlinear problems. Each one gets a deep dive with hyperparameter tuning—tweaking regularization, tree depth, number of trees, learning rate, all of it—using grid search and cross-validation. To judge performance, the system tracks accuracy, precision, recall, F1-score, and AUC-ROC. Confusion matrices and ROC curves give a more visual, gut-level sense of how each model is actually doing. But it's not just about numbers. SHAP analysis comes in for model interpretability, showing how much each feature—like tenure, contract type, monthly charges—actually pushes the model's predictions. This makes the results easier to trust and gives operators clear direction for customer retention strategies. The whole thing is built in Python, using both built-in and third-party libraries. You've got os , warnings, and argparse for handling files, silencing noise, and managing command-line inputs. NumPy and Pandas handle the heavy lifting with data. Scikit-learn takes care of the basic machine learning, while XGBoost adds the power of gradient boosting. Imbalanced-learn manages class imbalance with SMOTE, and matplotlib is there for plotting. Joblib wraps it up by saving the trained models and all their artifacts, so nothing gets lost. With advanced preprocessing, solid balancing, careful optimization, and SHAP-based explanations, this system delivers a robust, transparent, and scalable pipeline. It doesn't just predict churn accurately—it also shows you why, making it a practical tool for real telecom operations.

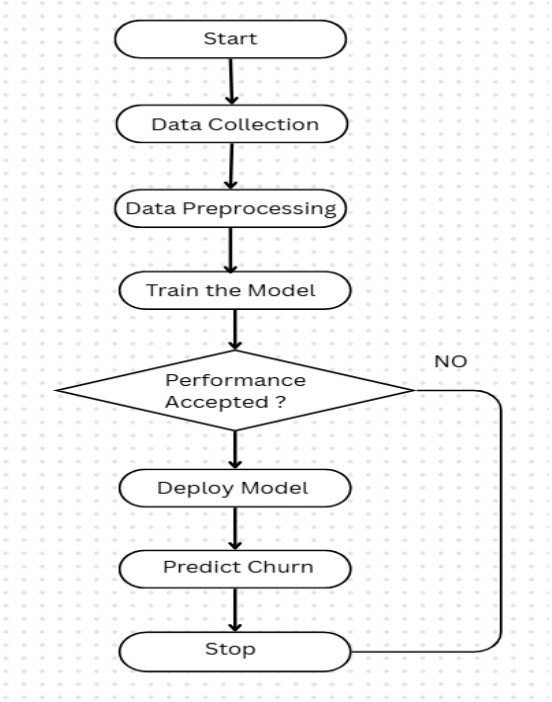

Fig. 1. Flowchart For Customer Churn Prediction

IV. RESULTS AND DISCUSSION

The results from testing our customer churn prediction system are pretty solid. The machine learning framework we built does a great job spotting which customers are likely to leave, especially in the telecom world. We fed the models a wide mix of customer data—everything from demographics and account details to internet and service usage. Before training, we did all the usual prep: normalized the data, encoded features, and balanced things out with SMOTE so both

churners and non-churners were represented fairly. We trained three main models: Logistic Regression, Random Forest, and XGBoost. For each one, we fine-tuned the hyperparameters to make sure we weren't getting tripped up by bias or variance. These models didn't just beat the old-school approaches—they really pushed up the accuracy for predicting churn. To really see how well everything worked, we checked several evaluation metrics: Accuracy, F1-score, and ROC-AUC. The ROC curves told the story—especially for the ensemble models, and XGBoost stood out. It did a great job telling apart customers who leave from those who stay. The ROC curves further validated the higher discriminative ability of the ensemble-based models, particularly XGBoost, in distinguishing churners from non-churners.

The most important features from the feature importance analysis, done through Smart Analysis and supported by SHAP interpretability, revealed that several customer-related attributes played an important role in determining churn behavior. Key influences included contract type, tenure, monthly charges, total charges, internet services, online security, technical support, and the method of payment, among others. Customers with short-term contracts, high monthly charges, and less technical support were found to have a much higher chance of churning. On the other hand, those with long-term contracts, stable tenure, and comprehensive bundled services tend to exhibit better retention and loyalty behavior. Moreover, the findings showed service-related attributes such as preference for streaming services, internet reliability, and billing preference affecting customer satisfaction, which in turn impacts churn decisions. This detailed understanding of the model's predictive behavior, combining Smart Analysis and SHAP-based feature evaluation, guaranteed interpretability and transparency. Results showed that the proposed integration of ensemble learning methods with feature-level interpretability enhanced not only the prediction accuracy but also provided premium insights about the behavioral patterns behind the customer churn. This would be beneficial for the telecom company to proactively identify customers at risk and implement data-driven retention strategies that focus on service enhancement, price optimization, and customized offers.

Overall, the results validate that the proposed model is indeed effective as a reliable decision-support framework that integrates interpretability with predictive precision and business applicability. By integrating optimized machine learning techniques along with feature-driven analysis and explainable AI tools, it would enable telecom providers to build up customer engagement, reduce churn rates, and optimize long-term profitability.

V. CONCLUSION AND FUTURE WORK

This churn prediction framework really shows off what machine learning can do for telecom companies trying to hang on to their customers. It brings together smart data processing with sharp prediction tools, so companies can spot which customers are most at risk of leaving. But it's more than just about getting the numbers right. The whole point is to make sense of messy, complex customer data and turn it into insights that actually help people make decisions. By calling out the biggest reasons customers leave, the system gives telecom companies the info they need to take action, keep people satisfied, and build loyalty that lasts. In the end, it's not just about crunching numbers—it gives companies a solid way to cut down on churn and build a stronger business. And honestly, there's a lot of room to grow. Imagine having a real-time system that watches customer behavior as it happens and reacts right away. Bring in deep learning or mix-and-match ensemble models, and you can catch even the really subtle warning signs. Add more data—stuff like transactions, usage patterns, even what customers say on social media—and you get an even clearer picture of what drives people to leave or stay. If you make the results more transparent with explainable AI, people will actually trust what the models tell them. And when you wrap all this into a cloud service or API, any telecom company can use it, no matter how big or small they are. So, this isn't just about predicting churn—it opens up new ways for telecoms to use analytics, make smarter decisions, and react in real time. Churn prediction isn't just some fancy tech for telecom companies—it's become something they can't really do without. The market's crowded, competition is fierce, and honestly, keeping your current customers is way cheaper than always trying to win over new ones. That's where machine learning comes in. Algorithms like Logistic Regression, Random Forest, Decision Trees, XGBoost, and Neural Networks help companies figure out who's likely to leave next. When you really dive into the details—demographics, how people use

their phones, billing habits, contract types, complaints—you start to see real patterns. With good data and a careful approach to cleaning and organizing it, these models can predict churn with surprising accuracy. And it's not just about the math. The way you prepare the data, pick out the right features, and tune your models makes a huge difference. These insights aren't just for researchers. Decision-makers actually use this info to roll out targeted offers, start loyalty programs, or fix service problems for customers who might leave. When you have real-time prediction systems, you can reach out right away when someone's thinking of jumping ship. And there's even more potential coming up. Deep learning, natural language processing, and sentiment analysis from things like social media or support chats can make predictions even sharper. Big data tools and cloud analytics let companies handle tons of information without breaking a sweat. But as companies lean more on all this data, privacy and ethics become even more important. Data protection laws like GDPR aren't just some box to check—they're about earning trust, and that trust keeps customers around. Bottom line: churn prediction lets telecom companies stop scrambling to keep up and start getting ahead. They keep customers happy, cut down on turnover, and grow the business.

REFERENCES

[1] A. Idris, A. Khan, and Y. S. Lee looked into smarter ways to predict customer churn in telecom. They used mRMR for feature selection and a RotBoost-based ensemble classifier. Their work, published in Applied Intelligence (2013), dives into how these techniques can help flag customers likely to leave. doi: 10.1007/s10489-012-0380-1

[2] In another study, A. Amin and colleagues focused on churn prediction in telecom using a rough set approach. Their research, featured in Neurocomputing (2017), shows how rough sets can help pick out key factors behind customer departures. doi: 10.1016/j.neucom.2017.03.044

[3] R. Ahmed and M. Maheswaran pushed churn prediction further by using an enhanced ensemble learning method. Their 2017 paper in Procedia Computer Science explains how blending different models can bump up prediction accuracy. doi: 10.1016/j.procs.2017.08.162

[4] W. Verbeke and his team wanted models that don't just predict churn but actually make sense to humans. Their work, in Expert Systems with Applications (2011), used advanced rule induction techniques to build clearer, more understandable models. doi: 10.1016/j.eswa.2010.08.088

[5] T. Vafeiadis and co-authors compared different machine learning methods for predicting customer churn. They laid out the strengths and weaknesses of each technique in Simulation Modelling Practice and Theory (2015). doi: 10.1016/j.simpat.2015.03.003

[6] A. Keramati's group improved churn prediction in telecom by tapping into various data mining methods. Their findings, published in Applied Soft Computing (2014), show how these tools can give telecom companies a real edge. doi: 10.1016/j.asoc.2014.08.041

[7] B. Huang, M. T. Kechadi, and B. Buckley also tackled churn prediction in telecommunications. Their research, which appeared in Expert Systems with Applications (2012), breaks down how to spot customers at risk of leaving. doi: 10.1016/j.eswa.2011.08

[8] M. Liu, J. Wang, and J. Su explored explainable AI for churn prediction in telecom. Their 2022 paper in the Journal of Intelligent & Fuzzy Systems shows how using SHAP and ensemble learning can make predictions not just accurate, but also easier to understand. doi: 10.3233/JIFS-211118

[9] H. He and E. A. Garcia tackled the tough problem of learning from imbalanced data—something that comes up a lot in churn prediction. Their work, published in IEEE Transactions on Knowledge and Data Engineering (2009), digs into methods for handling datasets where one class hugely outweighs the other. doi: 10.1109/TKDE.2008.239.