



Machine Learning Methods for Proactive Customer Churn Prediction in Mobile Networks: A Case Study

Eric Michel Deussom Djomadji¹, Mbah Shalom Agwa², Bienvenue Arsene Roger Fouba³,
Jean Gutenbert Kenfack Wamba⁴

Page | 1

^{1,4} Department of Electrical and Telecommunications Engineering, National Advanced School of Engineering of Yaoundé, University of Yaoundé I

² Division of Information and Communication Technology, National Advanced School of Post, Telecommunication and ICT

³ Department of computer Engineering, National Advanced School of Engineering of Yaoundé

Cite as:

Deussom Djomadji, E. M., Agwa, M. S., Fouba, B. A. R., & Kenfack Wamba, J. G. (2026). Machine Learning Methods for Proactive Customers Churn Prediction in Mobile Networks: A Case Study.. *International Journal of Electrical, Electronics and Communication Engineering*, 7(1), 1-23.
<https://doi.org/10.5281/zenodo.23767938>

© 2026 The Author(s). International Journal of Electrical, Electronics & Communication Engineering published by ACADEMIC INK REVIEW.

Abstract

Customer churn is one of the most critical and cost-intensive challenges facing the telecommunications industry today. With the rapid expansion of mobile networks and the saturation of the digital market, retaining existing customers has proven to be significantly more cost-effective than acquiring new ones. The present study proposes a comprehensive, data-driven framework utilizing Machine Learning (ML) techniques to predict customer churn proactively within the Cameroon National Operator (CNO) mobile network. By extracting and preprocessing real-world data from the Evolved Packet Core (EPC), specifically the Home Subscriber Server (HSS), a dataset of 13,937 valid customer records was established from an initial pool of 797,508. To address the inherent class imbalance between active and churned users, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. Four supervised machine learning algorithms were trained and evaluated namely: Logistic Regression, Random Forest, LightGBM, and a heterogeneous Stacking Ensemble classifier. Evaluation metrics demonstrate that the Stacking Ensemble model achieved the highest predictive performance. The model yielded an accuracy of 97.51%, a precision of 97.46%, a recall of 97.54%, and an F1-score of 97.50%. Furthermore, Kaplan-Meier survival analysis and feature importance evaluations revealed that early subscription phases (the first 50 days) and specific financial dynamics (monthly charges and recharge frequencies) are the most critical determinants of customer loyalty. The developed system empowers telecommunication decision-makers to identify at-risk subscribers in real-time, facilitating the deployment of personalized, proactive retention strategies.

Keywords: Customer Churn Prediction; Ensemble Learning; Stacking Classifier; Machine Learning

Introduction

The telecommunications industry has undergone a radical transformation over the last few decades, evolving from basic analog voice services (1G) to ultra-reliable, low-latency, and high-bandwidth digital ecosystems (5G). This technological evolution has exponentially increased global connectivity, fundamentally altering how societies and economies operate. However, this rapid advancement has also dismantled traditional market monopolies, significantly lowering the barriers to entry for competitors and lowering the switching costs for consumers. Today, mobile network subscribers are empowered to seamlessly transition between service providers in search of optimal network coverage, superior customer service, and cost-effective data bundles.

In this fiercely competitive landscape, customer loyalty is highly volatile. This phenomenon of subscribers abandoning a service provider in favor of a competitor is formally known as "customer churn." Industry benchmarks indicate that telecommunication operators globally experience average annual churn rates ranging from 20% to 35%. The economic implications of this attrition are profound. Decades of market research have established that the financial cost of acquiring a new customer (factoring in marketing, onboarding, and promotional subsidies) is five to ten times greater than the cost of retaining an existing

one (Manro et al., 2025). Consequently, telecom operators have shifted their strategic focus from aggressive customer acquisition to proactive customer retention.

The Cameroonian Telecommunications Context

In Cameroon, the telecommunications market is characterized by intense rivalry among primary Mobile Network Operators (MNOs), including MTN, Orange, and the incumbent Cameroon National Operator (CNO), Cameroon Telecommunications (CAMTEL). This competitive environment forces operators to continuously innovate their pricing models and digital service deliveries. Despite these efforts, customer loyalty remains fragile. CNO, which since some years acquired concessions to expand its mobile network services, faces the continuous challenge of voluntary churn where subscribers actively choose to discard their cards or port their activities to competing networks due to perceived inadequacies in service or the lure of better promotional offers.

Objectives of the Study

The primary objective of this research is to transition national operator's churn management strategy from a reactive framework where retention teams intervene only after a customer has already decided to leave to a proactive, predictive model. By analyzing vast amounts of historical customer data generated by the mobile network infrastructure, this study aims to:

1. Analyze and preprocess authentic CNO subscriber data to identify key behavioral, geographical, and financial factors contributing to churn.
2. Develop, compare, and evaluate multiple Machine Learning algorithms (Logistic Regression, Random Forest, LightGBM, and Stacking Ensembles) using appropriate validation metrics.
3. Validate a highly accurate and robust predictive model capable of flagging high-risk customers before they defect.
4. Design and deploy an interactive Churn Prediction System to translate algorithmic outputs into actionable business intelligence for CNO's strategy and marketing departments.

Context and Problem Statement

Churn Concept

a. Definition of Churn

Churn is one of the most challenging problems that affects revenue and customer base in mobile network operators. It refers to the number of customers who stop using a product or service over a given period of time: monthly, quarterly or annually (Lu, 2001).

In a mobile network churn prediction refers to the process of identifying subscribers which are likely to stop using the network's product or services in the near future, either by cancelling their subscription, switching to a competitor (Tridens Technology, 2024), or becoming inactive for a prolonged period. It involves analysing customer behaviour, usage patterns and other relevant data to forecast which customers are at risk of leaving (Grevisse & Yende, 2023).

The churn rate measures the percentage of customers lost during a specific time frame.

$$\text{Churn Rate} = \frac{\text{Number of Customers Lost during a Period}}{\text{Total Customers at the Start of the Period}} \times 100 \quad (1)$$

b. Reasons for Churn

Customer churn is not random; it happens for a reason or rather several reasons. Understanding these reasons is a key to building a model that works (Stripe, n.d.). Below are some reasons of churn.

- i. Dissatisfactions with service
- ii. Poor customer experience
- iii. Lack of engagement
- iv. Better offer from competitors
- v. Inadequate customer support

- vi. Product or service quality issues
- vii. Impact of reviews and reputation

c. Types of Churns

There are two major types of churns which are voluntary churn and involuntary churn (Wagh et al., 2024).

Voluntary Churn: The customer actively chooses to leave the service. It is caused by Poor service quality, high prices, better offers from competitors, dissatisfaction with customer service. An example is a mobile subscriber switches from the CNO to MTN for better data offers.

Involuntary Churn: The customer is removed by the service provider due to reasons beyond the customer's control or without the customer's intention to leave. It is caused by Non-payment of bills, fraudulent activities, service suspension, expired contracts not renewed. An example is a customer's account is deactivated due to non-payment of subscription fees. In our case we are dealing with voluntary churn.

Importance of Churn Prediction

Churn prediction allows the mobile network operator to more proactively address potential customer loss. It is important in the following ways (Wagh et al., 2024):

- i. Proactive retention
- ii. Cost savings
- iii. Enhanced customer experience
- iv. Optimized marketing
- v. Business insights
- vi. Offer Competitive Pricing
- vii. Revenue growth
- viii. Data-driven decisions
- ix. Long term value

Currently, the CNO employs traditional, rule-based methodologies to identify churn. The system primarily monitors usage activity, subscription status, and billing consistency, flagging a customer only after an extended period of total inactivity (typically 90 days). This approach is fundamentally reactive. It assesses the symptoms of churn rather than predicting the intent to churn, thereby failing to leverage predictive analytics to intervene when the customer's decision might still be reversed.

Research Problem and Objectives

a. Research Problem

The existing approach to customer churn management have relied on reactive, last-minute support and actions after the customer has made up his mind to churn. Such are inefficient, expensive, and do not treat the root causes of churn. With the increase in competition and customer demands, the need for a proactive, data-driven support capable of anticipating churn before it happens is crucial.

b. Research Question

From the research problem we asked ourselves how can machine learning techniques be applied to CNO's customer data to accurately predict churn?

c. Objective

The main objective of this research is to design and implement a predictive model based on ML which will serve as a decision-support tool that will allow the company to take timely actions aimed at reducing churn rates, enhancing customer satisfaction, and ultimately supporting CNO's digital transformation strategy. This section presented the general context and framework of our study, which focuses on the contribution of Machine Learning techniques to predicting customer churn and evaluating performance in mobile networks.

Literature Review

Several authors have explored the use of machine learning to solve certain problems in the field of mobile networks on various topics, including the following points:

Fraud Detection & Security

- i. SIM Box Fraud Identification (2023): Leverages Random Forest, SVM, and XGBoost on Call Detail Records (CDRs) to detect bypass fraud (Deussom Djomadji et al., 2023a).
- ii. Collaborative Fraud Detection Model (2022): Uses K-Means clustering and differential privacy to identify fraudulent anomalies across networks (Deussom Djomadji et al., 2022).

Network Maintenance and Management

- i. Alarm Classification: Applies K-Means and DBSCAN for fault analysis and root cause identification in optical networks (Deussom Djomadji et al., 2023b).
- ii. LTE Network Maintenance: Compares K-NN, Decision Trees, and Random Forest for managing and predicting incidents (Bernabe et al., 2022).
- iii. Resource Allocation: Evaluates Linear Regression, CNN, and LSTM for radio resource optimization (Deussom Djomadji et al., 2023c).

Traffic Forecasting: 4G Traffic Prediction (2025): An analysis focused on time-series forecasting using LSTM, CNN, and Linear Regression (Deussom Djomadji et al., 2025).

The application of Machine Learning (ML) to predict customer churn has garnered significant attention in academic research, moving from simple statistical models to highly complex ensemble networks.

Traditional Machine Learning Approaches

Early research primarily relied on standalone ML models. Varudandi (2020) demonstrated the efficacy of Logistic Regression in identifying churn trends, noting an accuracy of 81.5% on Kaggle datasets (Hutsady et al., 2024). The study successfully correlated specific service types (e.g., fiber optic usage) and payment methods with high churn probabilities. Similarly, Wagh et al. (2023) explored the use of decision trees and Boosted Trees in the telecom sector, identifying Random Forest as an optimal standalone model with a predictive accuracy of 91.66%, precision of 82.2%, and recall of 81.8%.

More recently, Manro et al. (2025) expanded on this by building a complete ML pipeline comparing classical algorithms with advanced boosting models like XGBoost and LightGBM (Varudandi, 2020). Their findings highlighted how data preprocessing and hyperparameter tuning significantly influence the detection of customer behavior patterns.

Advanced Ensemble and Stacking Methods

Recent literature has heavily pivoted toward Ensemble Learning by combining multiple weak learners to form a highly accurate predictive meta-model. A 2024 study by Hutsady et al (2024) analyzing Lao Telecom customer data achieved a remarkable 93% accuracy using Decision Trees and Random Forests, proving that aggregating decision pathways mitigates overfitting.

The most cutting-edge research emphasizes "Stacking Classifiers." A 2025 study on heterogeneous multi-stacking ensemble techniques integrated Random Forest, XGBoost, and K-Nearest Neighbors (KNN) with a Logistic Regression meta-model (Bhushan, 2025). This approach effectively balanced recall and precision, outperforming all base models by capturing complex, non-linear relationships in telecommunication data. Furthermore, research published in the *Bulletin of Electrical Engineering and Informatics* (2025) utilized a complex stacking model (combining XGBoost, Random Forest, LightGBM, SVM, and Neural Networks) paired with SMOTE and stratified k-fold cross-validation (Next-Gen Customer Retention, 2025). This sophisticated framework achieved an unprecedented 98% accuracy and 98.38% F1-score, definitively proving that optimizing stacking ensembles with robust class-balancing techniques is the optimal path for telecom churn prediction.

Research Gaps and Thesis Contribution

Despite extensive literature, several gaps persist:

- Dataset Authenticity:** A vast majority of studies utilize heavily recycled, publicly available datasets (such as the IBM Telco dataset on Kaggle). These datasets fail to capture current, localized market dynamics or the specific behavioral nuances of Sub-Saharan African consumers.
- Algorithmic Stagnation:** Many operational telecom deployments still rely on single-model architectures rather than adopting high-performance Stacking algorithms.
- Class Imbalance:** Real-world telecom data is inherently imbalanced (non-churners drastically outnumber churners). Many studies fail to rigorously apply techniques like SMOTE, leading to models that report high accuracy but possess poor sensitivity (recall) for actual churners.

This study directly addresses these gaps by utilizing raw, authentic, and localized subscriber data extracted directly from CNO's HSS. It seamlessly integrates SMOTE to resolve class imbalance and deploys a highly tuned Stacking Ensemble algorithm to outpace traditional predictive methodologies.

Methodology

The proposed methodology encompasses a robust machine learning pipeline designed to extract, clean, engineer, and model customer data. This pipeline is presented in figure 1.

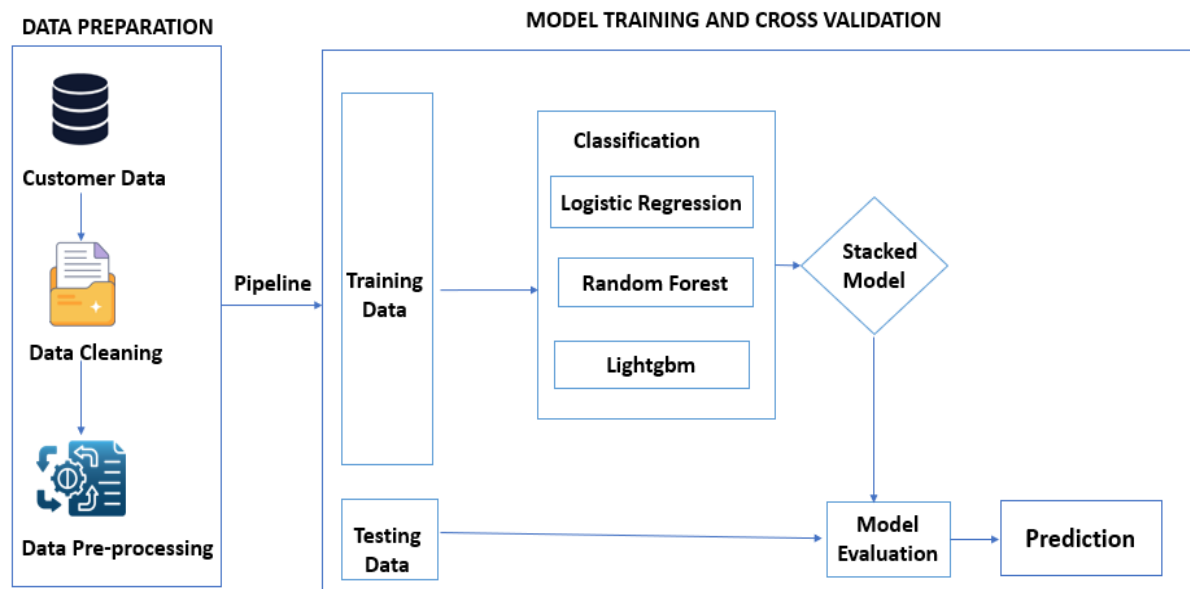


Figure 1: Machine learning pipeline for churn prediction

Data Collection and Preprocessing

Data was securely extracted from the CNO database encompassing a target timeline from January 1, 2025, to August 25, 2025. The raw dataset consisted of 797,508 subscribers.

The raw data is cleaned up, filled and transformed to ensure quality and consistency before we use it to develop the ML system. It's necessary to preprocess the data after it has been collected to make it ready for an ML task. Figure 2 and figure 3 are illustrations of raw data.

Figure 2a: Raw Data Set (1)

Figure 2b: Raw Data Set (1)

Figure 3a: Raw Data Set (2)

AN	AO
SUB_CREATE_TIME	SUPPSERVICE
1643365437	0
1643982965	0
1644500017	0
1645194710	0
1676020923	0
1674476979	0
1644482541	0
1644321675	0
1644834062	0
1645458395	0
1660116609	0
1657553047	0
1645706975	0
1645536362	0
1645629558	0
1645097295	0
1645004476	0
1642691090	0
1668183586	0
1646141546	0
1652707884	0
1653205187	0
1653990585	0
1655194868	0
1655719054	0
1656936622	0
1662990355	0

Figure 3b: Raw Data Set (2)

Data Cleaning:

The initial dataset was saturated with noise, including invalid entries (-1 values), null fields, and redundant structural errors. Through automated filtering using Python's Pandas library, duplicate records and rows lacking vital predictive attributes were eliminated. By filtering for accounts created strictly within the 2025 window and applying the cleaning protocols, the dataset was refined to 13,937 high-quality customer records.

Target Variable Definition (Labeling)

a. Derivation of Churn Status

The churn status in our project was derived based on customer activity timelines using the features SUB_CREATE_TIME and PURGE_TIME_ATSGSN. We began by sorting all customers whose accounts were created as from January 1st 2025. From this group, a customer was classified as churned if, between their subscription creation date and the reference date of 25th August 2025 when we started analysing the data, they remained inactive on the platform for a period exceeding three months. This approach provided a clear distinction between active customers and churners, allowing us to build a reliable target variable (CHURN_STATUS) for our predictive modelling. Figure 4 and figure 5 present a capture of the source codes used for the purpose.

```
# Ensure columns are datetime before filtering
df['PURGE_TIME_ATSGSN'] = pd.to_datetime(df['PURGE_TIME_ATSGSN'], errors='coerce')
df['SUB_CREATE_TIME'] = pd.to_datetime(df['SUB_CREATE_TIME'], errors='coerce')

df = df[df['PURGE_TIME_ATSGSN'] >= pd.Timestamp('2025-01-01')]
df = df[df['SUB_CREATE_TIME'] >= pd.Timestamp('2025-01-01')]

✓ 0.2s
```

Figure 4: Code source for converting PURGE_TIME_ATSGSN and SUB_CREATE_TIME to datetime

```

today = pd.Timestamp.today()
churn_cutoff = today - timedelta(days=90)

df['TENURE_DAYS'] = (today - df['SUB_CREATE_TIME']).dt.days
df['PURGE_TIME_ATSGSN'] = pd.to_datetime(df['PURGE_TIME_ATSGSN'], errors='coerce')
df['DAYS_SINCE_PURGE'] = (today - df['PURGE_TIME_ATSGSN']).dt.days

def label_status(row):
    if row['PURGEDONMME'] == 'YES' or row['PURGEDONSGSN'] == 'YES':
        return 'Churned'
    elif pd.notnull(row['PURGE_TIME_ATSGSN']) and row['PURGE_TIME_ATSGSN'] < churn_cutoff:
        return 'Churned'
    else:
        return 'Active'

df['CUSTOMER_STATUS'] = df.apply(label_status, axis=1)

```

Figure 5: Code source for finding Churn Status

Feature Engineering and Selection

a. Final Dataset and Feature Engineering for Model Training

Before processing, our dataset contained a sample of 797,508 customers. We first filtered out customers created from January 1st 2025 and we obtained 61,245 customers. Through a series of cleaning and preprocessing steps such as removing duplicates, handling missing values, and filtering out incomplete or irrelevant records the dataset was refined to 13,937 customers with valid entries in all the selected features. This cleaned and well-structured dataset then served as the foundation for training and evaluating our churn prediction models. To enhance the accuracy and reliability of our churn prediction model, we applied a thorough feature engineering process. The raw dataset provided contained sensitive customer details and lacked certain predictive attributes that are essential for robust training. Since some customers information could be disclosed due to data privacy restrictions, we carefully simulated some features to enrich the dataset without compromising the integrity of the dataset. For instance, we generated variables such as regional customer distributions to provide the model with more informative inputs including the longitude and latitude coordinates to better capture customer behaviour geographically and have a view of the active and churned customers on a map. The dataset is illustrated by figure 6.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
CUSTOMER_ID	VISION	PURGE_TIME_4TSGSN	TOPPHARMACYTOPPHARMACOL	PURGEDONMME	PURGEDONSGSN	P5_UTRAH	P5_GERAN	P5_5-UTRAH	EPSCORPOS	SUB_CREATE_TIME	TENURE_MONTHS	MONTHS_SINCE_PURGE	PLAN_NAME	MONTHLY_CHARGE	RECHARGE/TOTAL_CHARGE	REGION	LATITUDE	LONGITUDE	GENDER	AGE	CHURN_STATUS			
1	2.37625-41	24/07/2025 09:49	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	22/07/2025 08:16	1	1 Blue One S	2000	2	4000	Adamawa	7.25	13.583	Male	15	0		
2	2.37625-41	24/07/2025 09:49	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	02/07/2025 09:17	2	1 Blue One S	2000	2	4000	Far North	12	14.25	Male	16	0		
3	2.37625-41	27/02/2025 09:43	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	10/02/2025 09:52	8	3 Blue One L	5000	0	0	Adamawa	7.25	13.583	Female	32	1		
4	2.37625-41	13/07/2025 09:50	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	06/07/2025 12:19	8	2 Blue One L	5000	3	15000	South	2.917	11.217	Male	26	0		
5	2.37625-41	06/06/2025 12:15	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	20/02/2025 09:12	7	3 Blue One L	5000	3	15000	Far North	12	14.25	Male	33	0		
6	2.37625-41	06/02/2025 21:07	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	05/02/2025 19:48	7	7 Blue One L	5000	2	10000	North	8.75	13.5	Female	24	1		
7	2.37625-41	22/06/2025 06:15	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	16/04/2025 15:00	4	2 Blue One M	3000	3	9000	South West	4.95	9.3	Male	21	0		
8	2.37625-41	08/07/2025 20:06	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	12/07/2025 14:38	7	2 Blue One L	5000	3	15000	South West	4.95	9.3	Male	29	0		
9	2.37625-41	24/06/2025 14:35	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	17/02/2025 11:59	6	2 Blue One M	3000	3	9000	North West	6.333	10.667	Male	23	0		
10	2.37625-41	12/06/2025 08:04	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	07/04/2025 19:14	5	3 Blue One M	3000	2	6000	Adamawa	7.25	13.583	Female	22	0		
11	2.37625-41	25/07/2025 12:40	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	02/05/2025 06:04	4	1 Blue One M	3000	3	9000	Littoral	4.05	9.7	Female	19	0		
12	2.37625-41	26/07/2025 15:41	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	16/07/2025 09:20	1	1 Blue One S	2000	2	4000	West	5.5	10.417	Female	23	0		
13	2.37625-41	26/07/2025 09:01	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	15/07/2025 15:18	1	1 Blue One S	2000	1	2000	Adamawa	7.25	13.583	Female	19	0		
14	2.37625-41	27/07/2025 08:54	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	18/07/2025 09:20	1	1 Blue One S	2000	2	4000	Far North	12	14.25	Male	24	0		
15	2.37625-41	15/06/2025 17:12	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	17/07/2025 09:13	4	2 Blue One M	3000	3	9000	Adamawa	7.25	13.583	Male	24	0		
16	2.37625-41	10/07/2025 14:12	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	21/02/2025 15:46	7	2 Blue One L	5000	3	15000	Centre	3.848	11.502	Male	27	0		
17	2.37625-41	06/02/2025 08:12	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	06/02/2025 09:59	8	7 Blue One L	5000	2	10000	South West	4.95	9.3	Male	25	1		
18	2.37625-41	02/07/2025 13:15	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	22/07/2025 09:47	7	2 Blue One L	5000	3	15000	Far North	12	14.25	Female	26	0		
19	2.37625-41	27/07/2025 05:13	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	19/02/2025 08:43	7	1 Blue One L	5000	3	15000	South West	4.95	9.3	Male	29	0		
20	2.37625-41	30/06/2025 18:40	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	11/09/2025 11:14	6	2 Blue One M	3000	3	9000	North West	6.333	10.667	Female	25	0		
21	2.37625-41	05/06/2025 09:57	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	15/03/2025 07:38	6	4 Blue One M	3000	0	0	East	4.583	13.667	Male	30	1		
22	2.37625-41	02/04/2025 06:05	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	20/04/2025 11:53	5	5 Blue One M	3000	2	6000	Littoral	4.05	9.7	Female	26	1		
23	2.37625-41	13/06/2025 05:39	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	20/04/2025 13:45	4	3 Blue One M	3000	2	6000	Centre	3.848	11.502	Female	22	0		
24	2.37625-41	16/06/2025 15:25	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	14/06/2025 07:03	3	2 Blue One S	2000	2	4000	Far North	12	14.25	Male	26	0		
25	2.37625-41	05/06/2025 16:05	200000000	100000000	YES	NO	NO	NO	NO	NORPIS	17/02/2025 09:43	3	2 Blue One S	2000	2	4000	Far North	12	14.25	Male	26	0		

Figure 6: Final Data set after filtering, cleaning and preprocessing

b. Description of data features

In this section, we will describe the various features that make up the dataset used for our churn prediction model.

Table 1: Description of Data Feature

Feature Name	Description
CUSTOMER_ID	Unique identifier for each customer in the dataset. Used for internal tracking.
MSISDN	MSISDN
PURGE_TIME_ATSGSN	The timestamp when the subscriber's session or profile was purged from the Serving GPRS Support Node (SGSN), indicating inactivity or disconnection events.
TGPPAMBRMAXUL	Maximum u+plink bit rate (UL) allowed for the subscriber as defined in 3GPP standards. Represents the maximum upload speed allocated.
TGPPAMBRMAXDL	Maximum downlink bit rate (DL) allowed for the subscriber as defined in 3GPP standards. Represents the maximum download speed allocated.
PURGEDONMME	Binary indicator feature showing if the subscriber was purged on the MME (Mobility Management Entity), reflecting deregistration events
PURGEDONSGSN	Binary indicator feature showing if the subscriber was purged on the SGSN, reflecting inactivity in the 2G/3G packet-switched domain
PS_UTRAN	Packet-switched usage on UTRAN (3G network). Reflects customer activity in 3G data sessions
PS_GERAN	Packet-switched usage on GERAN (2G network). Reflects customer activity in 2G data sessions.
PS_E-UTRAN	Packet-switched evolved universal terrestrial radio access network which represents the data traffic or activity level of a subscriber within the 4G data sessions.
EPSOBBPOS	Subscription option parameter related to EPS (Evolved Packet System) bearer configuration. Indicates data service provisioning.
SUB_CREATE_TIME	Date when the subscriber account was created in the HSS, marking the start of the customer relationship.
TENURE_MONTHS	Number of months since the last purge event, measuring recent inactivity duration.
MONTHS_SINCE_PURGE	Number of months since the last purge event, measuring recent inactivity duration.
PLAN_NAME	Name of the customer's tariff or subscription plan indicating service preferences
MONTHLY_CHARGE	Monthly fee associated with the subscription plan.
RECHARGE_FREQUENCY	Number of times the customer recharged their account in a month. Indicates engagement and spending behaviour.
RECHARGE_AMOUNT	This is the amount that is recharge in the account of the subscriber.
TOTAL_CHARGE	Cumulative amount spent by the customer during their subscription period
REGION	Geographical region where the customer is registered active, important for churn analysis across territories.
LATITUDE / LONGITUDE	Coordinates of the customer's registered location. Used for mapping and clustering customers spatially.
GENDER	Customer's gender.
AGE	Customer's age.
CHURN_STATUS	Indicator showing if the customer has churned (1) or is still active (0). Target variable for prediction.

c. Description of Feature Statistics

To develop a better understanding of the attributes, it is required to construct summary statistics for all the features contained in the dataset used in this study. The figure7 below shows the outcomes.

	TGPPAMBRMAXUL	TGPPAMBRMAXDL	TENURE_MONTHS	MONTHS_SINCE_PURGE	MONTHLY_CHARGE	RECHARGE_FREQUENCY	TOTAL_CHARGE	AGE	CHURN_STATUS
count	13937	13937	13937	13937	13937	13937	13937	13937	13937
mean	1996362201.3346	99906723.1111	5.5086	2.9916	3470.6178	1.7381	5999.7847	24.9599	0.4402
std	84148570.3336	2157655.6496	1.5304	1.7172	1010.2168	1.0423	4197.778	3.3649	0.4964
min	50000000	50000000	1	1	2000	0	0	15	0
25%	2000000000	100000000	5	2	3000	1	2000	23	0
50%	2000000000	100000000	6	3	3000	2	6000	25	0
75%	2000000000	100000000	7	4	5000	2	9000	27	1
max	2000000000	100000000	8	8	5000	3	15000	38	1

Figure 7: Feature Statistics

d. Correlation between variables

We use a correlation matrix to understand the relationships between each variable as presented in figure 8. It shows the strength and direction of relationships between all the variables in the dataset and helps identify which features are strongly related to churn. Each cell in the matrix displays a correlation coefficient a value between -1 and +1 (*Customer Churn Analysis in Telecom Organization, 2024*):

- A value close to +1 means a strong positive correlation
- A value close to -1 means a strong negative correlation
- A value near 0 means no significant relationship between the variables

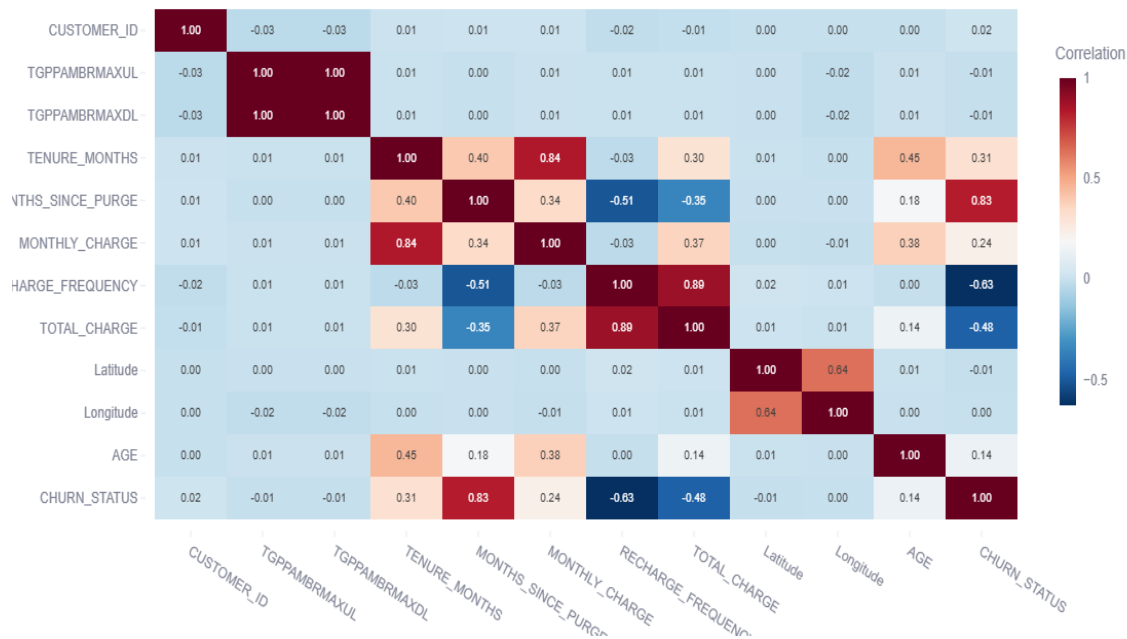


Figure 8: Heatmap Correlation matrix between the features

To maximize model accuracy, raw data points were engineered into predictive features.

- Financial Features:** Variables such as MONTHLY_CHARGE, RECHARGE_AMOUNT, and TOTAL_CHARGE were aggregated to track spending behavior.
- Network Features:** PS_E-UTRAN, PS_GERAN, PURGEDONMME, and bandwidth limits (TGPPAMBRMAXUL) were encoded to reflect technical engagement with 2G, 3G, and 4G infrastructures.

- iii. **Geographical Simulation:** To account for privacy constraints, synthetic baseline features such as REGION, LATITUDE, and LONGITUDE were logically mapped based on cellular tower distributions to facilitate spatial churn mapping.

e. Data Visualisation

Data visualization is a key aspect in exploratory data analysis as it helps to understand the underlying patterns, relationships and distribution of variables within the dataset. In this study, visualisations were used to analyze both numerical and categorical features related to customer churn.

Distribution of numerical features

The numerical features were visualized using histograms graphs in figure 9. These visualisation helps to observe the spread and central tendency of data.

- **TENURE_MONTHS:** This represents the number of months a customer has stayed with the company. The distribution is right skewed showing that most customers have stayed for five to seven months while fewer customers have very short tenure.
- **MONTHS_SINCE_PURGE:** This variable measures the time spent since a customer's last interaction in the platform. The distribution is left skewed with the majority of customers having one to three months since their last purge. Fewer customers have more than six months since their last purge.
- **MONTHLY_CHARGE:** This shows the amount billed monthly to customers.
- **RECHARGE_FREQUENCY:** This variable represents how often customers recharge their accounts. There are two distinct customer behaviors low frequency and high frequency rechargers. Low frequency rechargers might be less engaged and more likely to churn.
- **AGE:** this represents the age distribution of customers.

Distribution of Numeric Features

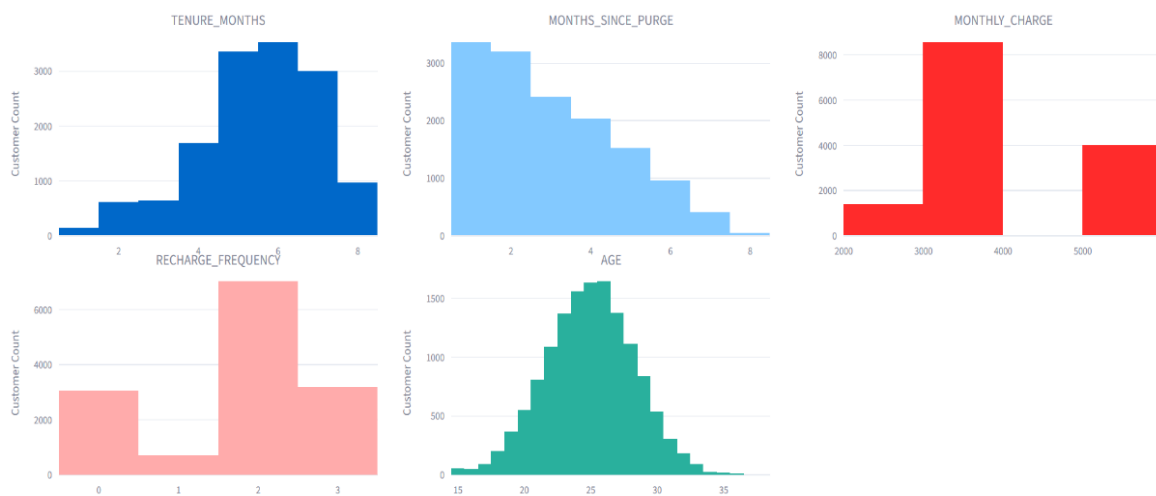


Figure 9: Distribution of Numeric Features

Distribution of Categorical features

Categorical variables were analyzed using bar charts and count plots to explore their relationship with customer churn (See figure 10). The visualisations below show how churn and non-churn customers are distributed across several features.

- **PS_E-UTRAN:** This feature reflects whether a customer has accessed the E-UTRAN network, commonly associated with LTE(4G) services. The chart shows a higher count of non-churned customers compared to churned ones. This suggests that customers with access to LTE service are more likely to remain active. The availability of high speed connectivity may enhance user satisfaction and reduce churn risk.
- **PS_GERAN:** This feature represents access to the GSM/EDGE radio access network (GERAN), commonly associated with 2G services. The chart shows a higher count of non-churned customers compared to churned ones.

- **PURGEONMME:** This feature indicates whether a customer's session was purged from the mobility management entity (MME). It is usually done on customers who has not used mobile data or voice services for a long period of time.
- **PURGEONSGSN:** This feature indicates whether a customer's session was purged from the SGSN (serving GPRS support node). It is usually done on customers who has not used mobile data or voice services for a long period of time.
- **PLAN_NAME:** This feature shows the distribution of different plan type the customers uses.
- **GENDER:** This feature shows the gender distribution of customers in CNO

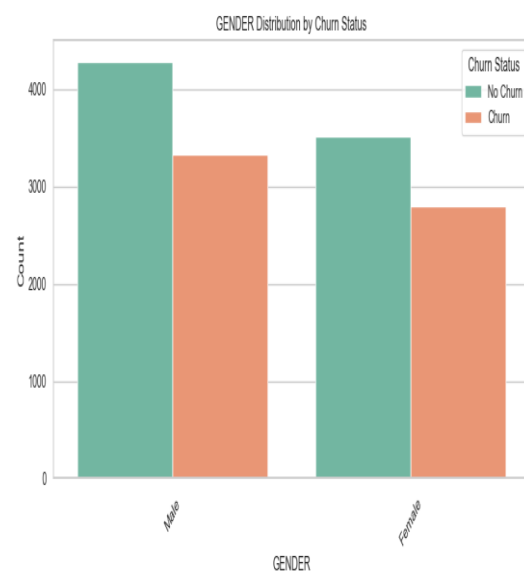
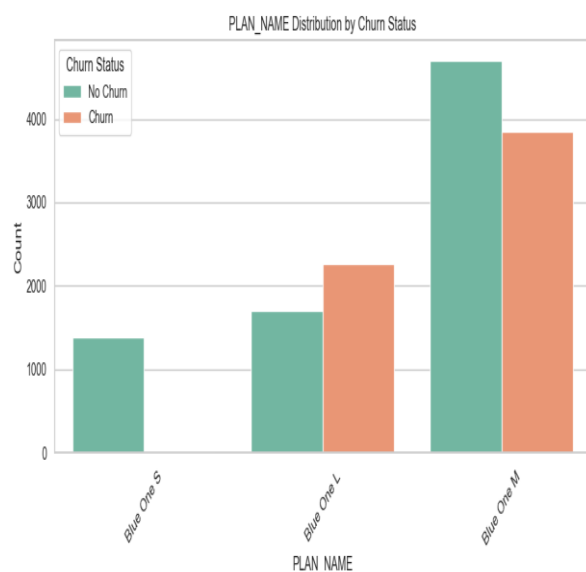
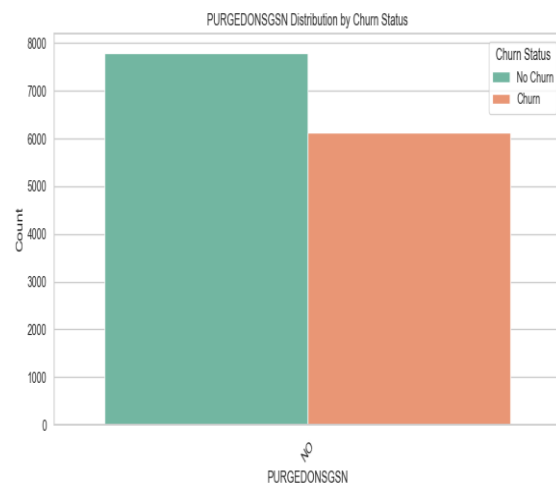
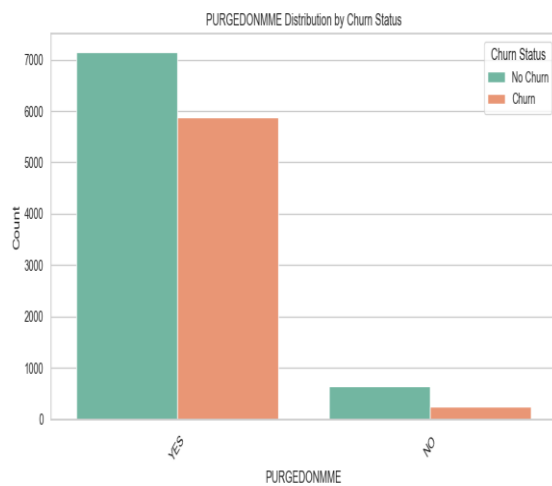
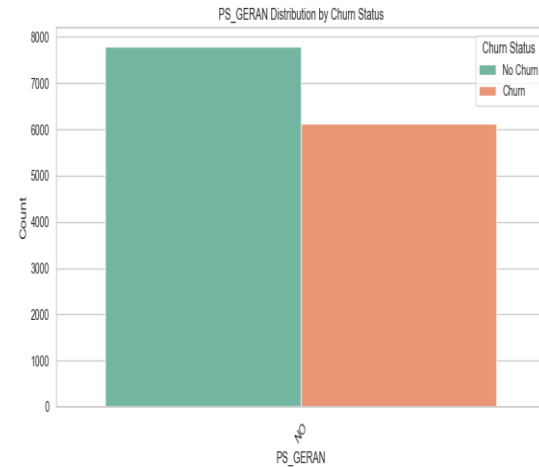
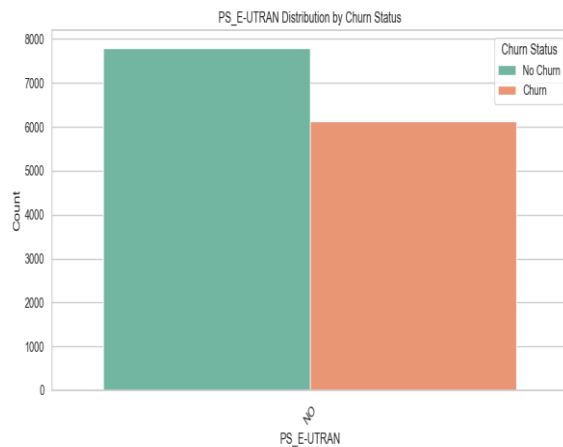


Figure 10: Distribution of Categorical features with respect to the churn status

f. Data splitting Process

The data is split in the ratio 70/10/20, where 70% of the data is used for training, 10% is used for validating the model and 20% for testing. This ensures that the model learns effectively while still having enough data for reliable evaluation.

g. Cross Validation Method

We make use of K-fold cross validation which divides the data into k-folds, train the model on k-1 folds and test it on the remaining folds. It helps to prevent the model from overfitting and improve robustness.

h. Handling Class Imbalance

Imbalance data is a common challenge in churn prediction, where the number of non-churners outweighs churners. This imbalance can lead to inaccurate models that fail to identify customers at risk of leaving. For our case we have 7801 subscriber who have not churned and 6136 subscribers who have churned. In order to balance both classes, Synthetic Minority Oversampling Technique (SMOTE) was applied to the training dataset, so the size of the minority class, in this case churner, was increased to the similar size of the majority class, non-churner, through random sampling of churner.

i. Model Evaluation Metrics

An evaluation metric is a quantifiable measure to evaluate the performance of a ML model. It is a way to check how well a model performs meaning how close are the predicted values to the actual values for unseen data. Our aim is to have perfect predictions for our ML model but getting 100% is correct predictions is not possible. So, evaluation metrics helps us to quantify that error so we know how accurate our model is. For our evaluation we will be using the metric accuracy, f1-score, recall, precision, roc-auc and confusion matrix.

Machine Learning Algorithms

Four specific algorithms were selected based on their distinct mathematical strengths namely logistic regression, random Forest, Light Gradient Boosting Machine and Stacking Ensemble Algorithm.

Logistic Regression

Used as a baseline and later as a meta-learner, Logistic Regression is a statistical model used for binary classification. It models the probability that an instance belongs to a particular class using the Sigmoid function. It operates under the assumption of a linear relationship between the independent variables and the log-odds of the dependent variable.

Random Forest

An ensemble learning method that constructs a multitude of decision trees during training. Random Forest applies "bagging" (Bootstrap Aggregating) and feature randomness to ensure that individual trees are uncorrelated. For classification tasks, the final output is determined by majority voting among the trees. This structure drastically reduces the variance and overfitting typically associated with deep single decision trees.

Light Gradient Boosting Machine (LightGBM)

A highly efficient framework based on decision tree algorithms that utilizes gradient boosting. Unlike traditional boosting tools that grow trees level-wise, LightGBM grows trees leaf-wise, choosing the leaf with the maximum delta loss to grow. It employs Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), making it exceptionally fast and highly accurate for large-scale telecommunication datasets.

Stacking Ensemble Algorithm

Stacking (Stacked Generalization) is an advanced meta-learning strategy designed to combine the predictive strengths of heterogeneous base models.

- i. **Level-0 (Base Learners):** The dataset is passed through Logistic Regression, Random Forest, and LightGBM models.

- ii. **Level-1 (Meta-Learner):** The predictions (probabilities) outputted by the Level-0 models are compiled into a new dataset. A meta-learner (in this study, a secondary Logistic Regression model) is trained on this new dataset to make the final prediction. This allows the system to intelligently weigh which base model is most reliable for specific types of data patterns.

System Design and Software Environment

The ML pipeline was integrated into a comprehensive, interactive web application designed for CNO's strategic analysts. The platform was built utilizing Python as the core programming language.

- i. **Backend & ML:** Scikit-learn, Imblearn (SMOTE), Pandas, and NumPy.
- ii. **Frontend & Interactivity:** Streamlit framework, HTML, CSS.
- iii. **Visualizations & Mapping:** Plotly for interactive charting and Folium for geographic network mapping.
- iv. **Reporting:** ReportLab for automated PDF generation of churn insights.

Unified Modeling Language (UML) diagrams (Use Case, Class, and Activity) were utilized to map the architectural flow, ensuring seamless interaction between data upload, model training, geographical visualization, and prediction downloads. Figure 11 presents the use case diagram follow by figure 12 for the class diagram and figure 13 for the activity diagram.

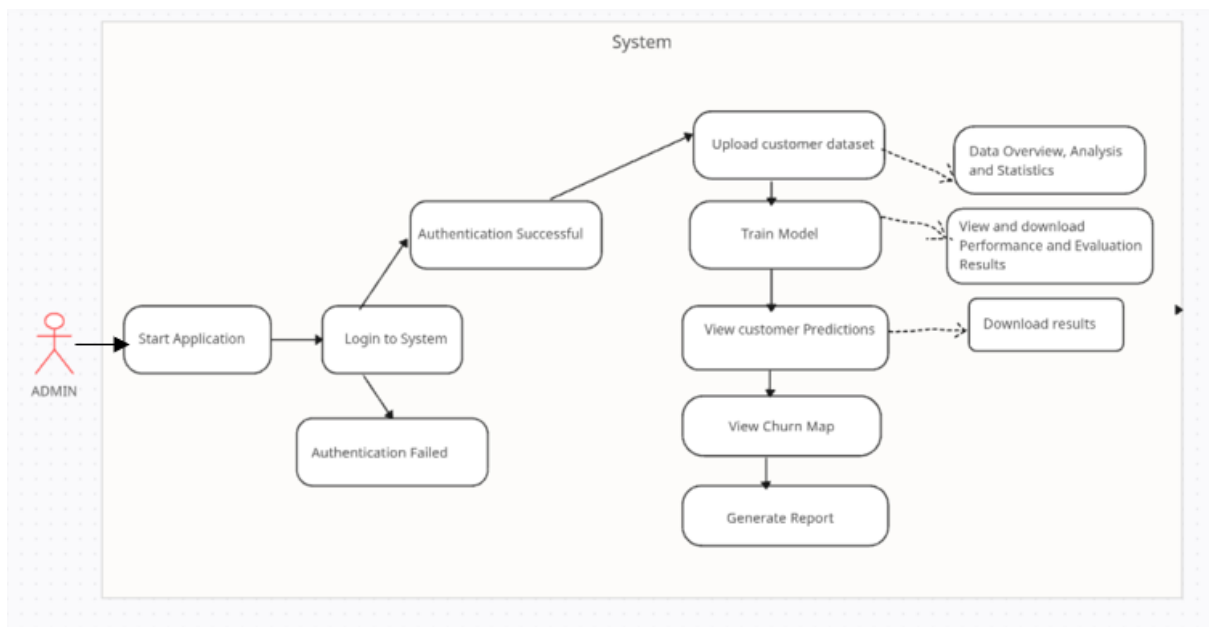


Figure 11: Use case Diagram

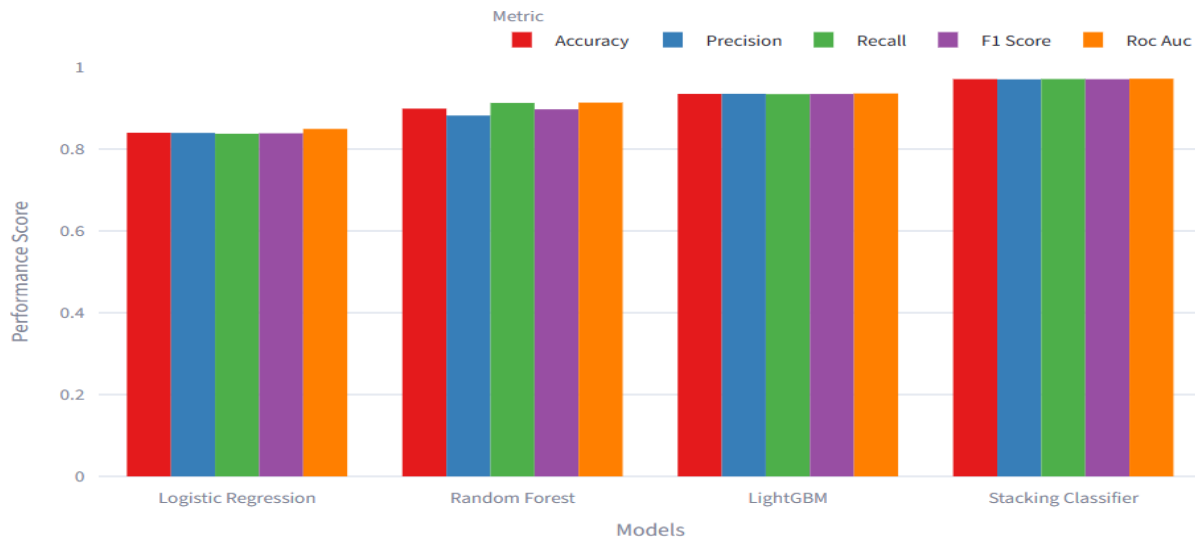


Figure 14: Models Performance Result

Training Results Summary

	accuracy	precision	recall	f1_score	roc_auc
Logistic Regression	84.753	84.7171	84.51	84.6134	85.6931
Random Forest	89.0406	87.3659	90.4276	88.8704	90.494
LightGBM	92.5753	92.5995	92.524	92.5617	92.6749
Stacking Classifier	97.5153	97.4614	97.5408	97.5011	97.6199

Best Model: Stacking Classifier with 97.5% accuracy

Figure 15: Model Performance Table

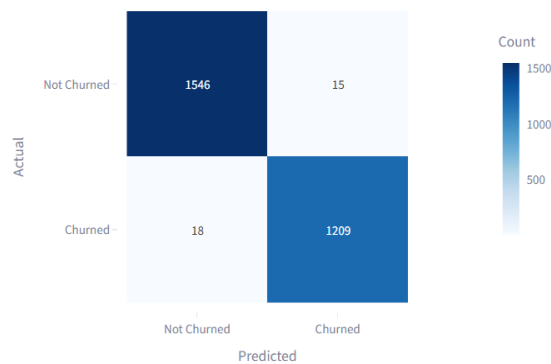
Results Summary:

- Logistic Regression: Achieved a baseline accuracy of 84.75%, precision of 84.71%, and recall of 84.51%. While computationally light, it struggled to capture complex, non-linear subscriber patterns.
- Random Forest: Improved performance to 89.04% accuracy and a 90.42% recall, proving its robustness in handling diverse feature interactions.
- LightGBM: Achieved highly competitive results with 92.57% accuracy and a 92.52% recall, showcasing the power of gradient boosting on tabular data.
- Stacking Classifier: Outperformed all individual models across every metric. The meta-learner achieved a remarkable 97.51% Accuracy, 97.46% Precision, and 97.54% Recall, with an F1-Score of 97.50%.

Confusion Matrix Analysis

The confusion matrix provides a granular view of the model's predictive precision (True Positives vs. False Positives).

Confusion Matrix - Logistic Regression



Confusion Matrix - Random Forest

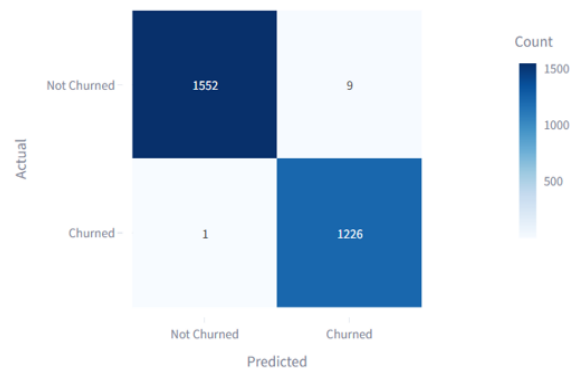
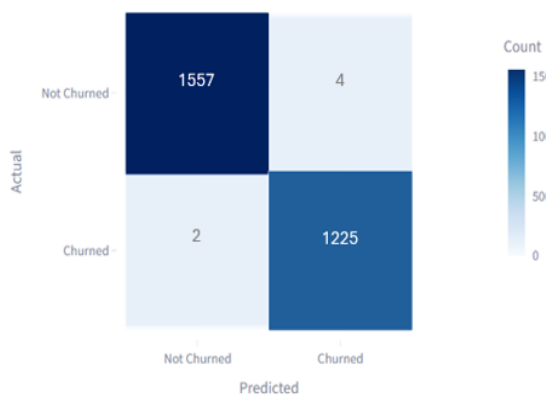


Figure 16: Confusion Matrix for Logistic Regression (left) and Random Forest (right)

Confusion Matrix - LightGBM



Confusion Matrix - Stacking Classifier

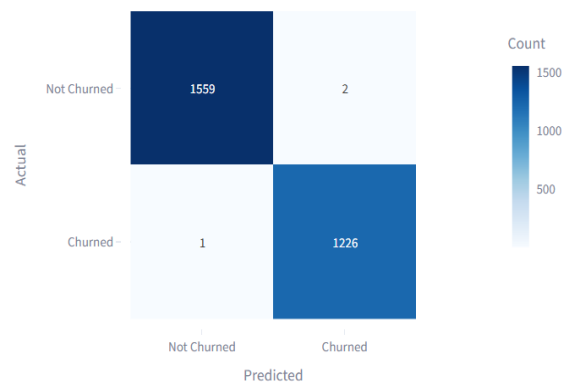


Figure 17: Confusion Matrix for LightGBM (left) and Stacking (right)

For the premier Stacking Classifier, out of 2788 observations, the model achieved near-perfect classification:

- i. **True Negatives (TN):** 1,559 active customers correctly identified.
- ii. **True Positives (TP):** 1,226 churned customers accurately flagged.
- iii. **False Positives (FP):** Only 2 active customers misclassified as churners.
- iv. **False Negatives (FN):** Only 1 actual churner missed by the system.

This exceptionally low rate of False Negatives is crucial for telecom operators; missing a churner represents a direct loss of revenue, whereas a False Positive merely results in an unnecessary retention promotion being sent to a loyal customer.

Feature Importance and Behavioral Insights

The Random Forest and LightGBM models allowed for the extraction of feature importance scores show by figure 30 and figure 31, revealing the primary drivers of customer attrition.

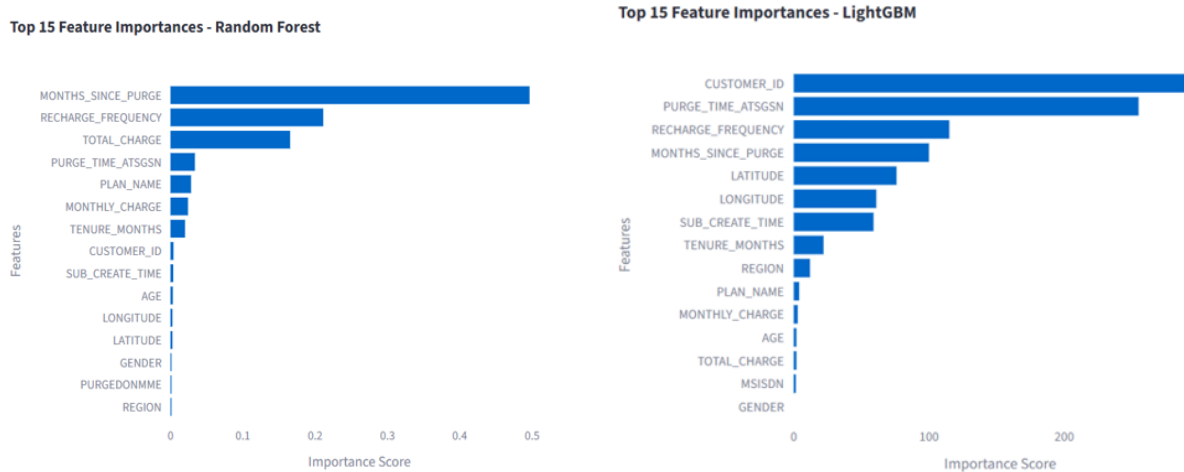


Figure 18: Feature importance for Random Forest (left) and LightGBM (right)

- Time-Based Engagement (MONTHS_SINCE_PURGE & TENURE_MONTHS):** The recency of a user's connection to the EPC was the most dominant predictor. Customers demonstrating initial signs of reduced network interaction were exponentially more likely to churn entirely.
- Financial Expenditure (TOTAL_CHARGE & MONTHLY_CHARGE):** Financial data analysis revealed distinct risk clusters. Density distribution graphs (Figure 19) indicated that churners heavily concentrated in the higher monthly charge bracket (4,500 – 5,500 FCFA). Conversely, active, loyal customers dominated the mid-tier spending bracket (3,000 – 3,500 FCFA). This counter-intuitive finding suggests that high-paying customers possess higher expectations regarding network quality and customer service; when these expectations are unmet, they actively seek competing providers.

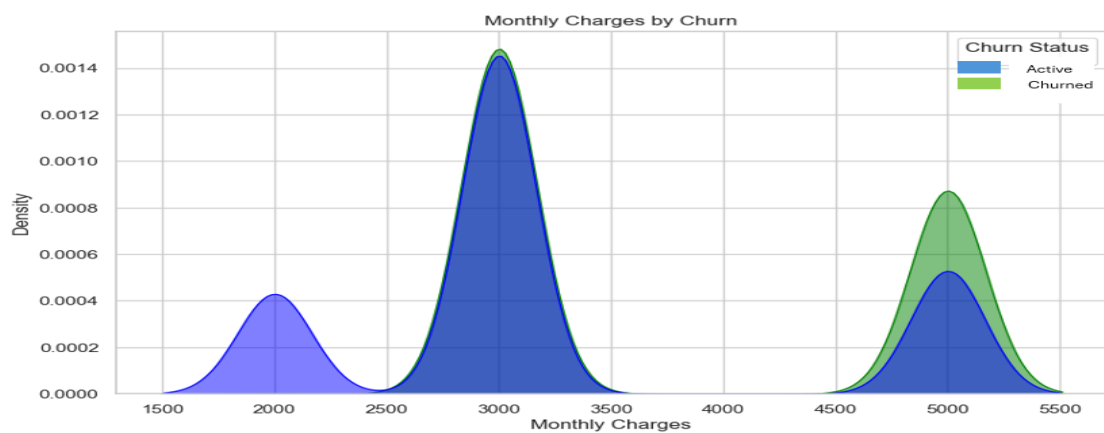


Figure 19: Monthly Charges by Churn

- Recharge Habits (RECHARGE_FREQUENCY):** Analysis of recharge volumes (see figure 20) showed that users making infrequent, low-volume recharges (clustering below 2,500 FCFA) were highly prone to churn, indicating a gradual disengagement from the CNO ecosystem.

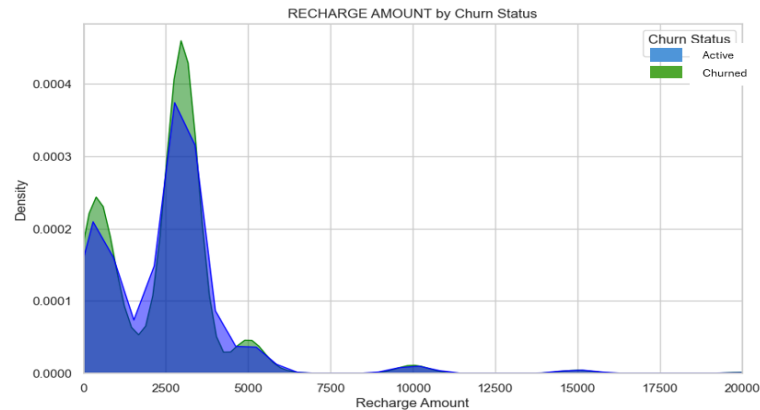


Figure 20: Recharge amount by churn

Survival Analysis (Kaplan-Meier Curve)

To predict the lifespan of a customer's loyalty, the Kaplan-Meier survival estimator was applied (See Figure 21). The generated survival curve illustrated a precipitous decline in customer retention probability within the first 50 days of account creation.

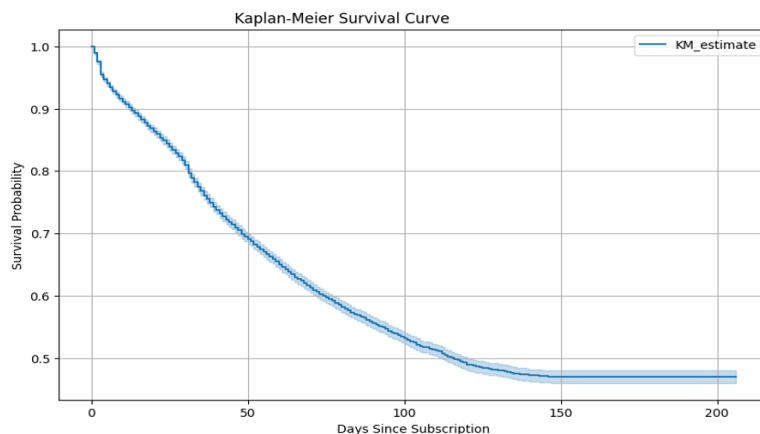


Figure 21: Kaplan - Meier Survival Curve

Interpretation: This steep initial drop indicates widespread "early-life churn," potentially driven by poor onboarding experiences, immediate dissatisfaction with network coverage, or the conclusion of introductory promotional data bundles. Beyond 150 days, the curve flattens significantly. Therefore, CNO must aggressively focus retention efforts and loyalty campaigns strictly within the first two months of a subscriber's lifecycle to yield maximum return on investment.

The Churn Prediction System Interface

The practical output of this research was the deployment of the Streamlit-based Churn Prediction Dashboard. The system features:

- i. Data Overview: Automated analysis of missing values and dataset shape.
- ii. Geographic Insights (Folium): An interactive map plotting subscriber distribution based on latitude/longitude variables, allowing regional managers (e.g., in the Littoral or Centre regions) to visually pinpoint clusters of high churn risk (figure 22).
- iii. Automated PDF Reporting: A one-click generation tool using ReportLab to export predictive summaries and strategic recommendations for executive review (Figure 23 and figure 24).

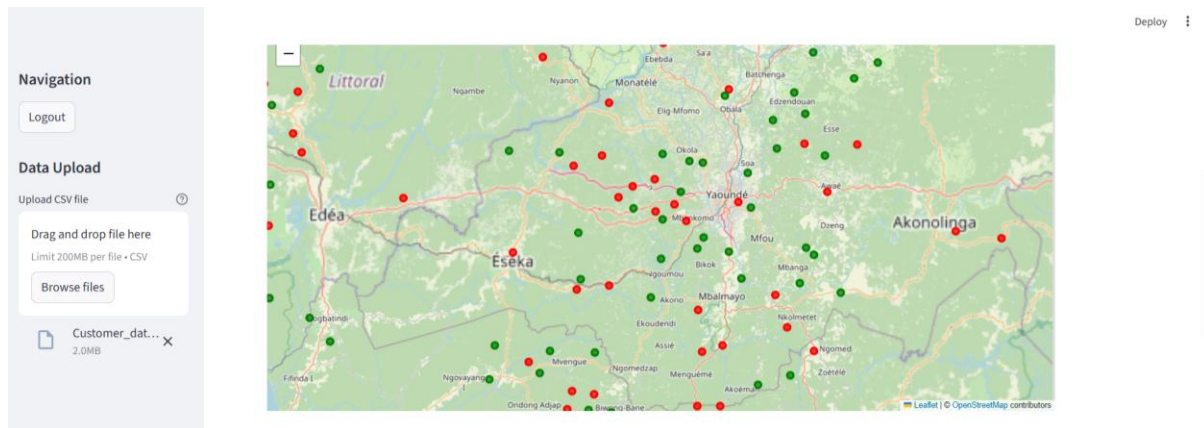


Figure 22: Map Visualization of churn status

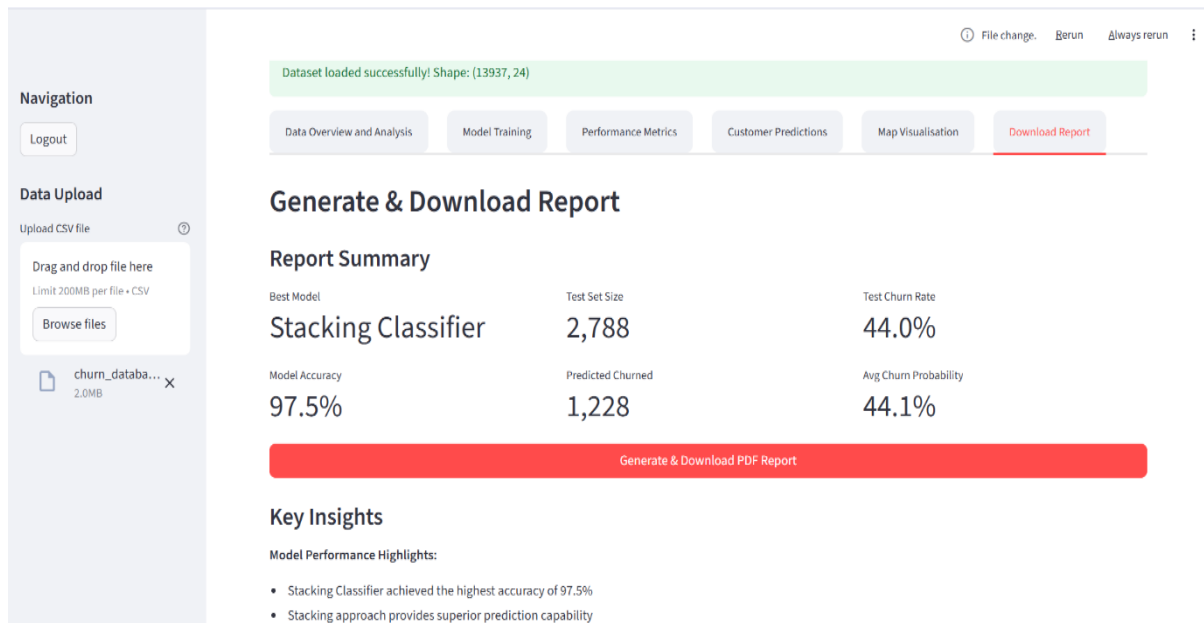


Figure 23: Download Report Page

Camtel Customer Churn Analysis Report

Executive Summary

This report presents a comprehensive analysis of customer churn prediction for Camtel mobile network. The test dataset contains 2,788 customer records. Multiple machine learning algorithms were evaluated, with the Stacking Classifier achieving the highest performance. The analysis identified 1,228 customers at high risk of churning with an overall churn rate of 44.0%.

Model Performance Results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
Logistic Regression	84.8	84.7	84.5	84.6	85.7
Random Forest	89.0	87.4	90.4	88.9	90.5
LightGBM	92.6	92.6	92.5	92.6	92.7
Stacking Classifier	97.5	97.5	97.5	97.5	97.6

Customer Risk Analysis

Risk Distribution:

- High Risk Customers: 1,226 (44.0%)
- Low Risk Customers: 1,562 (56.0%)

Strategic Recommendations

Immediate Actions:

- Focus retention efforts by implementing targeted campaigns on 1,226 high-risk customers
- Monitor churn probability trends monthly

Recommendations:

- Send SMS to 1,226 high-risk customers immediately to remind them to subscribe to various bundle offers
- Retention campaigns with personalized SMS, app notifications, or call center follow-ups.
- Offer tailored bundles based on usage patterns.
- Introduce loyalty rewards for long-tenured customers to reinforce retention.
- Monitor recharge frequency and trigger reactivation offers for dormant users.
- Investigate infrastructure quality in areas with high churn percentage
- Focus marketing efforts on underserved regions

Figure 24: Summary Report of Prediction and Insights

Conclusion

Customer churn remains one of the most persistent threats to profitability in the highly competitive telecommunications sector. This article successfully developed, evaluated, and deployed an advanced Machine Learning framework dedicated to predicting subscriber attrition within CNO's mobile network. By moving beyond reactive monitoring and leveraging raw behavioral data from the Evolved Packet Core, this study established a robust, predictive approach to customer lifecycle management.

Among the algorithms tested, the Stacking Ensemble classifier combining the strengths of Random Forest, LightGBM, and Logistic Regression yielded very good results, achieving a 97.51% predictive accuracy. Furthermore, deep analytical exploration via feature importance and Kaplan-Meier survival curves uncovered critical business insights: the first 50 days of a subscription are critical to long-term loyalty, and high-spending customers exhibit a disproportionately high risk of churning, likely due to unmet

service expectations. The integration of these models into a user-friendly, web-based prediction system provides CNO's decision-makers with a highly effective, data-driven tool to execute targeted, cost-efficient retention strategies.

References

- Manro, E., Malhotra, D., & Kamthania, D. (2025, April). Customer churn prediction using machine learning. *BCIIT eJournal*, 2(1).
https://bciit.ac.in/pdf/eJournal/Issue1_2025/eshanmanro_V2I1_FullText.pdf
- Lu, J. (2001). *Predicting customer churn in the telecommunications industry: An application of survival analysis modeling using SAS*.
- Tridens Technology. (2024). *Why telecom customers churn and how to measure it?*
<https://tridenttechnology.com/telecom-churn/>
- Grevisse, R., & Yende. (2023). Analysis of the churn in mobile telecommunications companies and customer loyalty through targeted marketing: Study of IT discriminant. *International Journal of Academic and Applied Research (IJAAR)*.
- Stripe. (n.d.). *What causes churn and how businesses can minimize it*.
<https://stripe.com/resources/more/what-causes-churn-and-how-businesses-can-minimize-it>
- Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, 14, Article 100342. <https://doi.org/10.1016/j.rico.2023.100342>
- Deussom Djomadji, E. M., Basile, K. I., Christian, T. T., Djoko, F. V. K., & Michael, S. E. (2023). Machine learning-based approach for identification of SIM box bypass fraud in a telecom network based on CDR analysis: Case of a fixed and mobile operator in Cameroon. *Journal of Computer and Communications*, 11(2), 142–157. <https://doi.org/10.4236/jcc.2023.112010>
- Deussom Djomadji, E. M., Matemtsap Mbou, B., Tchagna Kouanou, A., Ekonde Sone, M., & Bayonbog, P. (2022). Machine learning-based approach for designing and implementing a collaborative fraud detection model through CDR and traffic analysis. *Transactions on Engineering and Computing Sciences*, 10(4), 46–58. <https://doi.org/10.14738/tmlai.104.12854>
- Deussom Djomadji, E. M., Clovis, T. N., Christian, T. T., Mamadou, A., & Sone, M. E. (2023). Machine learning-based alarms classification and correlation in an SDH/WDM optical network to improve network maintenance. *Journal of Computer and Communications*, 11, 122–141.
<https://doi.org/10.4236/jcc.2023.112009>
- Bernabe, B., Deussom Djomadji, E. M., Marie, C., & Fabrice, M. (2022). Comparing machine learning algorithms for improving the maintenance of LTE networks based on alarms analysis. *Journal of Computer and Communications*, 10, 125–137. <https://doi.org/10.4236/jcc.2022.1012010>
- Deussom Djomadji, E. M., Kabiena, I., Nkemeni, V., Belinga À Njere, A., & Sone, M. (2023). Dynamic resource allocation in LTE radio access network using machine learning techniques. *Journal of Computer and Communications*, 11, 73–93. <https://doi.org/10.4236/jcc.2023.116005>
- Deussom Djomadji, E. M., Nkemeni, V., Aurelle, T. K., & Issikamle, M. (2025). Machine learning-based traffic prediction in 4G LTE networks: Case study of a mobile operator in Cameroon. *American Journal of Engineering and Applied Sciences*, 18(1), 47–54. <https://doi.org/10.3844/ajeassp.2025.47.54>
- Hutsady, O., Vongpasith, P., Sonevilay, P., Phonekhamma, C., & Khammanivong, O. (2024, August). *Lao Telecom customer churn prediction: A case study in Borikhamxay branch*.
- Varudandi, S. (2020, November). *Predicting customer churn using logistic regression*.
- Enhancing customer churn prediction with stacking ensemble and stratified k-fold. (2025, February). *Bulletin of Electrical Engineering and Informatics*, 14.
<https://beei.org/index.php/EEI/article/download/8112/4047>

Bhushan, S. (2025, January). *Enhancing customer churn prediction in the telecom sector using advance machine learning techniques and explainable AI*.

<https://norma.ncirl.ie/8683/1/shreyasbhargavbhushan.pdf>

Next-gen customer retention: A stacked ensemble model for churn prediction. (2025, June).

<https://www.thesesjournal.com/index.php/1/article/download/503/449>

Customer churn analysis in telecom organization. (2024, February). *International Journal for Multidisciplinary Research (IJFMR)*. <https://www.ijfmr.com/papers/2024/2/18886.pdf>