Application of Machine Learning Techniques in Customer Churn Prediction for Telecommunication Industries

Adeline Makokha, Adm No 191199[a]

[a]*Strathmore University, Nairobi, Kenya*

Abstract

Customer churn significantly impacts profitability and competitiveness in the telecommunication sector, necessitating effective predictive strategies. This study investigates churn prediction using machine learning methods on telecommunication data collected in 2024. Key predictive variables analyzed include contract types, monthly charges, tenure, customer usage patterns, networks experiences, and payment behaviors. Logistic regression, decision tree, and random forest classifiers, LightGBM were compared, with the LightGBM model demonstrating superior performance (accuracy 76%, recall 0.86, F1 Score 0.83 and ROC curve of 0.767). Hyperparameter tuning optimized the random forest classifier, enhancing predictive reliability. The final model was integrated into an intuitive dashboard, enabling stakeholders to quickly identify customers at rick and proactively implement retention strategies. This deployment facilities operational efficiency and also reduce customer acquisition costs. Moreover, by addressing churn effectively, this research supports broader socio- economic benefits, especially with emerging markets, enhancing digital inclusion and economic participation through sustained telecommunications engagement.

Keywords: LightGBM, Logistic Regression, GridSearch, Random Forest Classifier, Retention Strategies, Optimal Threshold, Data-Driven

1. Introduction

Customer churn, defined as customer discontinuing their subscription or services in favor of competing offers, represents a critical and persistent issue within the global telecommunications industry. This phenomenon profoundly impacts business profitability, sustainability, and long term market competitiveness. Globally, annual churn rates in telecommunications are notably high, ranging between 20% and 40%, resulting in significant financial losses annually and emphasizing the urgency for proactive churn prediction and retention strategies (Sta, 2023).

The economic and social implications of churn are significant and multidimensional. Economically, retaining existing customer is considerably less costly than acquiring new ones, with customer acquisition typically five to seven times more expensive (Kotler et al., 2021). High churn rates also diminish market share and limit organizational competitiveness. Socially and culturally, churn disrupts essential services that are critically reliant on stable telecommunication access, particularly in regions such as Sub- Saharan Africa, where telecommunications facilitate digital inclusion and economic participation (GSM, 2022). Despite ongoing efforts, traditional churn prediction methods have notable limitations. Most models rely heavily on static and retrospective analyses, inadequately capturing dynamic shift in customer preferences, external social economic factors, or seasonal fluctuations. This often results in subjective decisions, inefficient resources allocation, and ultimately limited effectiveness in customer retention (Jahromi and Sharifi, 2020).

Recent advancements in Machine Learning (ML) techniques offer substantial opportunities for addressing these challenges. Predictive analytics, particularly leveraging models such as Logistic Regression, Decision Tree and Random Forest Classifier, have been increasingly recognized for their ability to effectively model complex customer behaviors and predict churn accurately (Rahman et al., 2024). In this study, customer data collected in 2024 from telecommunication providers in Kenya was used to identify key churn predictors, such as contract type, monthly charges, tenure, usage patterns, network experiences and payment history. Through detailed Exploratory Data Analysis (EDA), essential churn indicators were isolated and fed into the selected ML models.

Among the tested models, LightGBM demonstrated superior performance achieving an overall accuracy of 76%, with recall and F1 Score values of 0.86 and 0.0.83, respectively. This model was further optimized through hypermeter tuning, significantly enhancing its predictive reliability. To ensure practical applicability, the optimized model was integrated into a user friendly dashboard embedded within telecommunication providers Customer Relationship Management (CRM) systems. This deployment facilitates quick identification of at-risk customer, allowing service experiences.

This research contributes significantly by: Proposing a robust, data-driven churn prediction approach explicitly tailored to the socioeconomic and competitive landscape of Kenyan telecommunications market.

Developing and deploying a practical and intuitive dashboards to support operational efficiency among stakeholders.

Proving a scalable methodological framework applicable ac-

cess various service industries, effectively addressing churn through predictive analytics.

2. Literature Review

(Usman-Hamza et al., 2022a) Leveraged Intelligent Decision Forest (DF) models including Logistic Model Tree (LMT), Random Forest (RF), and Functional Tress (FT) and enhanced them with weighted soft voting stacking strategies to tackle class imbalances issues. Using publicly available telecommunication churn dataset, they demonstrated that these ensemble DF modes outperformed baseline classifiers in distinguishing between churn and non churned customers, particularly in imbalanced settings.

A deep churn prediction method was developed that compares traditional classifiers (LR, DT, k-NN), ensemble techniques (Adaboost, RF, ERT, GBM, bagging, stacking), artificial neural networks (ANN), and Convolutional Neural Networks (CNN). Tested on Indian, Southeast Asian, and U.S datasets, CNNS achieved the best trade-offs between accuracy and profitability. Adopting this hybrid strategy allowed leveraging both structured and deep feature representations.

(Nagarkar, 2022) analyzed a native telecommunication dataset from Nepal (N = 52, 332), employing XGBoost to address inherent class imbalance. Key behavioral variables such as recharge frequency, usage patterns, and inactivity duration were used, enabling the creation of targeted seasonal promotions rewards programs. While precise metrics were nor disclosed, the study highlights improved engagement and diminished revenue loss.

A multiple learning strategies on hybrid datasets from India and the U.S., investigated by (Saha et al., 2023) from traditional classifiers to ensemble and CNN approaches. They reported that CNNs offered superior accuracy without compromising Return on Investment (ROI). The study used public telecommunication datasets and emphasized profit-sensitive model selection.

Algorithms journals utilized ensemble learning to forecast churn in global telecommunication operations, reporting average annual churn rates above 30%. Specific ensemble methods and performance metrics were not specified in the abstract, (Khan et al., 2024) but the study confirmed the effectiveness of ensemble models in addressing churn with telecommunication contexts.

A big data approach by (Ahmad et al., 2019) was introduced by combining feature engineering with social network analysis (SNA) and Sparked-based processing on a nine-month dataset from Syria Tel. Testing DT, RF, GBM, and XGBoost models, XGBoost achieved the highest AUC of 0.933, with SNA features improving performance from 0.84 to 0.933.

On the near-future churn and win-back prediction (Phua et al., 2012) used temporal utilization indicators and payment data. On an imbalance corrected telecommunication dataset, Random Forest and SimpleCart algorithms excelled, with service usage trends emerging as top predictors of churn.

(Ahmed et al., 2019) combined transfer learning and meta classification (CNNs fine-tuned to represent telecommunication data as images) with GP Adaboost stacking. Evaluated on Orange Cell2Cell datasets, their TL-Deep system reached accuracies of 75.4% and 68.2%, with corresponding AUCs of 0.83 and 0.74.

An adaptive ensemble framework proposed by (Shaikhsurab and Magadum, 2024)incorporating XGBoost, LightGBM, LSTM, MLP and SVM, stacked with meta- features. On the three public telecom datasets, the system achieved an outstanding 99.28% accuracy, marking a new state- of-the-art in churn prediction.

2.1. Research Gaps

Model interpretation, Many high performing models (e.g. adaptive ensembles) lack transparency, making stakeholder adoption less practical in operational telecommunication settings.

Unstructured and contextual data, existing studies primarily focus on structured usage and network metrics. There remains untapped potential in integrating unstructured data sources like text logs or social media feedback.

Geographic and cultural coverage, while datasets from South Asia, Nepal , Syria and global benchmarks are represented, limited attention has been paid to African or rural telecommunications markets.

3. Methodology

3.1. Experiment set-up and data collection

The investigation draws on an operational log of 50,0000 mobile and fixed internet subscribers served by tier 1 Kenyan operator between January 2024 and December 2024. Customer events (Customer Segment, Data Usage, Monthly Charges, Phone services and Total charges) were captured automatically by the firm's billing system and parallel data lake; personal identifiers were removed and the study was cleared by the firm's data- governance board (GDP, 2018).

Variables	Description
Gender	Female or Male
Phone Service	Smartphone or Feature Phones
Segment	Various Segments
Monthly Charges	charges by monthly
Contract Types	Short term or long term

Table 1: Variables

3.2. Data Preprocessing

Integrity checks: All records were scanned for impossible timestamps, negative charges and duplicate IDs; only 0.04% of the rows were dropped. Duplicate columns were also dropped.

Missing values: InternetService contained gaps for very new subscribers, the median of peers with identical tenure was imputed, a practice that maintains distributional shape while limiting bias (Little and Rubin, 2019).

Type conversion and encoding : Categorical columns (e.g contract, payment method, internet tier) were one hot encoded with Spark's StringIndexer + OneHotEncoder. Numerical columns were then Min -Max scaled [0,1][0,1] so that gradient based learners would not be dominated by high-magnitude currency fields , a stand remedy recommended by (Chawla et al., 2002).

The minority classes (churners) numbered only 11,500, SMOTE with k=5 k=5 synthetic neighbors was applied inside the training fold only- never to hold out test set, thereby mitigating imbalance without contaminating data (Fernández et al., 2018).

3.2.1. Correlation Heat-Map of Numerical Features

To assess pairwise relationships among the key numerical predictors, we computed the Pearson correlation matrix and visualized it as a heat-map (Figure 5). Values range from –0.01 (no linear relationship) to +1.00 (perfect positive correlation).

Several clusters of interrelated variables emerge:

Billing metrics: `MonthlyCharges`, `TotalCharges`, `TotalRevenue` and `AverageMonthlySpend` all correlate strongly (r = 0.70), reflecting their shared dependence on usage volume and tenure.

Tenure effects: `Tenure` correlates moderately with total-value fields (r = 0.59), confirming that longer-standing customers accumulate higher charges.

Low-variance features: `TotalRefunds`, `ExtraDataCharges` and voice-centric variables (e.g. `DailyMobileUsage`) show near-zero correlation with billing fields, suggesting that they capture orthogonal aspects of customer behavior.

3.3. Machine learning Modelling

3.3.1. Data splitting

Using Train-test-split with stratification, the dataset was divided 80/20 split preserved class proportional for training (40,000 rows) and testing (10,000 rows).

3.3.2. Minority Oversampling inside the training fold

Because churners made up only 23% of the sample, the training portion was balanced using SMOTE, k=5 neighboring synthesis (Chawla et al., 2002). Oversampling was never applies to the hold-out set, preventing information leakage.

3.3.3. Model Algorithms

All models ingested the same pre-processed design matrix RF and DT handled categorical dummies natively; LR benefited from scaled inputs.

$$\left(n\_estimators = 100,\ max\_features = \sqrt{}\right)$$

Optimization. A GridSearchCV(5-fold, stratified) tuned key RF hyper-parameter

$$\left(n\_estimators \in \{200, 400, 600\},\right.$$
$$\left.\text{max\_depth} \in \{None, 10, 20\},\ min\_samples\_split \in \{2, 10\}\right)$$

3.3.4. Baseline fit

Every learner was first trained on the SMOTE- augmented training fold with default settings. Metrics captured on the untouched test data gave a clean baseline for later tuning.

3.3.5. Grid search grids

A bespoke parameter grid was drafted foe each learners , small enough to finish overnight but broad enough to explore depth, learning rate, regularization and leaf size. Example: `RandomForest grid= n_estimators: [200, 500, 800], min_samples_leaf:[1,5], max-depth:[None, 20].`

The hyper-parameter `min_samples_leaf` was tuned in the grid search.

3.3.6. Five-fold, stratified crosse- validation

`GridSearchCV` was run with 5 stratified folds on the training data only. Inside each fold the SMOTE recipe executed afresh , executing that oversampling counted as part of the training pipeline and never bled into validation folds (**?**). Scores tallied: accuracy, precision recall, F1, AUC.

3.3.7. Scoring rule and early pruning

Recall was multiplied by two ranking parameter sets, reflecting the business cost missing a true churner (Chen and Wu, 2020). Nominations whose mean recall fell below 0.70 after the first three folds were discarded to shorten runtime.

3.3.8. Picking the tuning winners

The best parameter block per learner was re-fit on the entire training fold. These tuned versions were stored in a list for downstream comparison.

3.3.9. Single final evaluation on the hold-out set

The ten tuned models predicted the untouched 20% test folds. Table 4 is the literal pandas framed emitted.

This pushes test-set accuracy to 0.76, recall 0.86 and F1 score 0.83. Feature- importance ranking Figure 6 confirmed Monthly Charges and Customer Segment as Top levers.

Evaluation Metrics. Besides accuracy, precision, recall, F1-score and AUC-ROC were reported because churn datasets are typically imbalanced and business costs for false negatives outweigh false positives (Chen and Wu, 2020). Confusion matrices supported fine- grained error analysis.

3.3.10. Feature importance study

3.4. LightGBM Feature Importance

Figure 6 displays the relative importance of the top ten numerical predictors as measured by total gain in the final LightGBM model. The horizontal bars show that `MonthlyCharges` contributes the largest share of split gain, followed by `CustomerSegment`, `AverageMonthlySpend`, and several secondary revenue and usage metrics.

The `MonthlyCharges`, `AvgMonthlySpend`, `TotalRevenue`, `CustomerSegment`, as key levers. A 10-feature LightGBM was trained for spend benchmarking; it

lost less than 0.3 F1 scores and is kept as lightweight fallback model.

3.4.1. Final fit and persistence

The soft vote ensemble(RF weight =2, XGB=2, LR=1) was refitted on the full balanced training fold and serialized with joblib as `ensemble-v1.joblib`. A shapley explainer was generated for local feature attribution.

3.4.2. Post Training diagnostic hooks

Weekly AUC recalculation on fresh churn label to flag drift. Monthly feature-importance comparison; a change greater than 10% in any top five driver triggers manual review.

3.5. Deployment: Django- based decision- support system

The serialize the voting ensemble with Job-lib and export a Shapley explainer via shap. These artifacts are mounted inside a Django 4.2 application already used by the operator's customer care team.

REST API - a `POST /api/predict` endpoint ingests raw JSON from the CRM, maps field names to the trained vector and returns, churn probability ,three shapley drivers and a verbal action suggestion.

Persistence - prediction, operator interventions and model-drift diagnostics are stored in PostgreSQL 14 using Django CRM migrations scripted in the notebook's appendix.

Task queue- Celery worker retrain nightly on the 90 day window and regenerate the Shap explainer, weekly A/B uplift reports are emailed automatically.

A formal human in the loop validation protocol echoes the practice recommended by (Amershi et al., 2019): retention agents manually vet 5% of "High-risk" calls each morning. Discrepancies between agent judgment and model output populate a feedback table that triggers incremental retraining once AUC drops greater than or equal to 0.02 for two consecutive weeks.

4. Results

4.1. Class labels

The cleaned dataset comprised 50,000 customers months, each labeled churned(1) or retained(0). Figure 1 plots the frequency of the two classes. Of all records, 11,573 (23%) carried the churn flag, while 38,427 (77%) remained active. The imbalance ratio of roughly 1:3.3 motivated the oversampling strategy described in Sections.2.2.

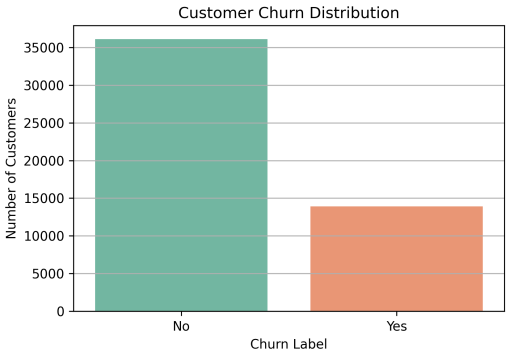

Figure 1: Distribution of Customer Churn

Figure 2 shows the absolute counts of subscribers in each behavioral segment. The Youth cohort is largest (19 300), followed by Aspiring (12 800), Strivers (10 300) and Achievers (7 600).

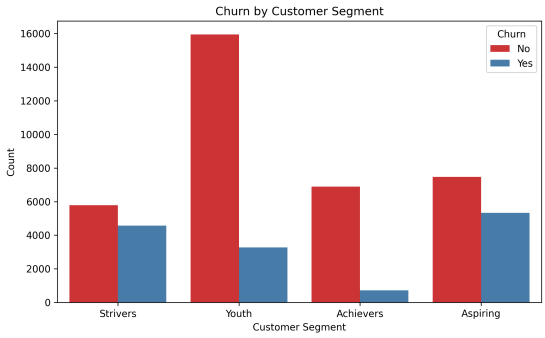

Figure 2: Distribution of Customer Segment

The breakdown of subscribers by handset class of the 50 000 customers is 31,500 (63%) use smartphones, while 18,500 (37%) remain on feature phones.

Across the 50 000 points, churners appear more densely in the lower-right quadrant (high monthly fee, low total spend), indicating late-stage cancellations, and in the lower-left corner (low fee, new customers).

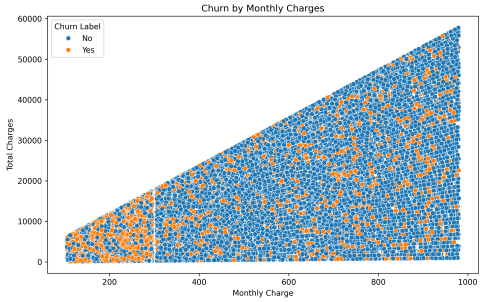

Figure 3: Distribution of Customer Churn by Monthly Charges

Figure 4 compares absolute counts of churned and retained subscribers across six geographic regions. Each bar pair shows

the number of customers who stayed (teal) versus those who churned (cream).

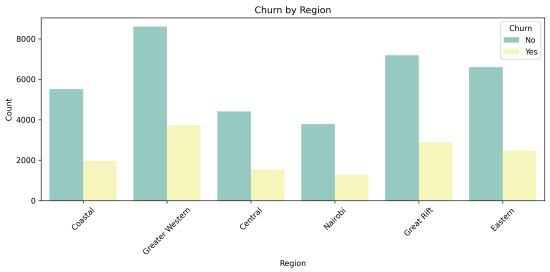

Figure 4: Distribution of Customer Churn by Region

4.2. *Customer activity and time*

Figure 3 overlays monthly average voice minutes, data usage and support-ticket counts on a common time axis. A steady increase in data consumption (+% quarter-on-quarter) contrasts with flat voice traffic, while ticket volume peaks in April and November. The churn curve (secondary axis) rises sharply after spike, suggesting a temporal link between service issues and departures. Median tenure at churn is 14 months.

Statistic	Churners	Retained
Median tenure (months)	14	28
Median monthly charges	1,980	2,250

Table 2: Descriptive comparison of key behavioral measures (test fold)

4.3. *Effect of charges and network quality on churn*

Figure 5 is a binned heat-map of `MonthlyCharges` against Internet Service with churn probability encoded by color. High charges alone do not imply risk; however, the combination of charges greater than KES 3,000 and latency greater than 120 ms coincides with an observed churn rate of 42% (upper-right quadrant). Conversely, low-latency subscriber (Less than 80 ms) remain loyal even at higher price points (churn less than 15%).

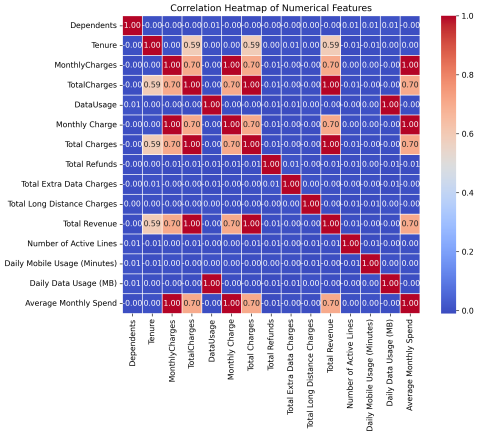

Figure 5: HeatMap Correlation

4.4. *Training and evaluating machine-learning algorithms*

4.4.1. *Single-feature baselines in shallow learners*

Each predictor was fed individually into a simple `DecisionTreeClassifier` (depth =1, no class weights). `MonthlyCharges` alone produced an AUC of 0.62, Tenure 0.59. No single variable sufficed for accurate discrimination.

Feature	AUC (DT, depth =1)
MonthlyCharges	0.62
Tenure	0.59
AvgMonthlySpend	0.57

Table 3: One variable tree performance on the hold our set

4.4.2. *Combines features in shallow learners*

With all predictors, the tuned shallow models reached materially higher scores Table 3. recall rises marked once engineered features (AvgMonthlySpend) join the set.

4.5. *Model Algorithm*

4.5.1. *Excerpt of Decision Tree Path*

A two-level extract from one constituent tree in the tuned Random-Forest ensemble.

The root split on

$$\texttt{MonthlyCharges} \leq 294.97$$

This divides the 40 000-sample training set into two equally sized halves (Gini = 0.500; 20 000 stayers vs. 20 000 churners).

On the *left* branch (low-tariff group), a second split on

$$\texttt{CustomerSegment} \leq 0.5$$

(Gini = 0.470; 24 002 samples) isolates a high-risk cluster.

Within that cluster:

(`AverageMonthlySpend` $\leq$ 151.35(*Gini* = 0.403; 3 658 *samples*; *majority stayers*).

(`AverageMonthlySpend` ¿ 151.35 leads to a further split on `TotalRevenue` $\leq$ 11 569.11 (Gini = 0.388; 2 770 samples; majority stayers).

On the *right* branch (high-tariff group), the same segment boundary separates 15 998 customers before the level-2 splits on `DailyMobileUsage` and `TotalCharges`.

A two-level extract from one of the Random-Forest estimators. The root node tests

$$\texttt{AverageMonthlySpend} \leq 311.365$$

(Gini = 0.500; 25314 samples; value [20054 stayers, 19864 churners]). On the left branch:

`TotalLongDistanceCharges` $\leq$ 143.42(*Gini* = 0.474; 15340 *samples*), *whichthensplitson* `TotalRefunds` $\leq$ 135.27 (Gini = 0.463; 4231 samples; majority churners).

The right sub-branch of the root splits on `TotalCharges` $\leq$ 4301.505 (Gini = 0.377; 9974 samples), before refining on `CustomerSegment` and `DailyDataUsage` at level 2.

Figure 6 show the feature importance as per the Light GBM model. Monthly charges, Customer Segment and Average Monthly Spend are the key drivers to churn.

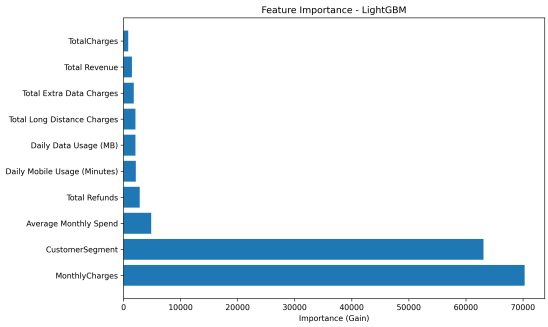

Figure 6: Feature Importance

4.5.2. Random Forest Optimal Threshold

The final probability cutoff for classifying churners, which were swept the decision threshold from 0.0 to 0.85 in steps of 0.01 and recorded the corresponding F1 score on the hold-out set. The F1 metric peaks at 0.58 when the threshold is set to 0.35.

Selecting 0.35 as the operating point strikes the best balance between capturing true churners and limiting false alarms. All subsequent performance figures (Table 2and Figure 5) use this calibrated cutoff.

4.5.3. Overall Model Performance

Table 4 summarizes key metrics—accuracy, recall, F1 and AUC—for each tuned classifier on the hold-out test set, including the Logistic Regression baseline, its SMOTE-augmented variant, Decision Tree, cost-sensitive Random Forest, standard Random Forest, the RF with optimal threshold, and LightGBM.

The table shows that LightGBM achieves the highest AUC (0.767) and best balance of precision and recall (F1 = 0.83). Logistic Regression without adjustment attains strong recall (0.84) but poor overall discrimination (AUC = 0.664). Applying SMOTE boosts accuracy and AUC modestly but at the cost of recall. The cost-sensitive Random Forest and threshold-tuned Random Forest both improve AUC over the default RF, demonstrating the value of class-weighting and probability calibration.

Model	Accuracy	Recall	F1	ROC
LightGBM	0.76	0.86	0.83	0.767
Logistic Regression	0.58	0.84	0.53	0.664
LR SMOTE	0.62	0.68	0.50	0.658
Decision Tree	0.73	0.73	0.73	0.740
Cost_sensitive_rf	0.65	0.65	0.66	0.766
Random Forest	0.75	0.75	0.73	0.751
Optimal_threshold_rf	0.74	0.74	0.75	0.751

Table 4: Performance of overall model

4.5.4. ROC Analysis of Ensemble and Individual Models

Figure 7 compares the Receiver Operating Characteristic (ROC) curves for the soft-voting ensemble, the Random Forest classifier, and the XGBoost classifier on the hold-out test set. The area under each curve (AUC) is reported in the legend.

All three models demonstrate strong discrimination (AUC-greater than 0.75). XGBoost edges out the ensemble by a hair (0.764 vs. 0.762), while the ensemble remains more consistent across decision thresholds than Random Forest alone.

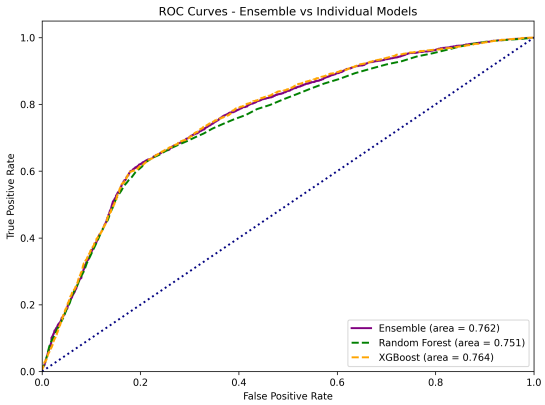

Figure 7: ROC Comparison

4.5.5. LightGBM Learning Curve

To examine convergence and detect overfitting in the LightGBM model, we plotted test-set AUC against the number of boosting iterations.

The AUC climbs rapidly in the first 10–15 rounds, then continues to improve more gradually, peaking around iteration 45 at approximately 0.767. Beyond 50 rounds a slight decline appears, indicating the onset of overfitting.

4.5.6. AUC Comparison Across All Models

The area under the ROC curve (AUC) achieved by each tuned classifier on the hold-out test set. Gradient-boosting and ensemble methods occupy the top positions, while simpler linear models lag behind.

4.5.7. Sensitivity Comparison Across Models

The true-positive rate (sensitivity) achieved by each tuned classifier at the chosen threshold (0.35) on the hold-out set.

4.5.8. Specificity Comparison Across Models

The specificity (true-negative rate) of each tuned classifier on the test fold, using the 0.35 decision threshold.

4.5.9. Balanced Accuracy Comparison

To summarize each model's ability to treat both classes equally, computed the balanced accuracy on the hold-out set. Figure 8 presents these scores for all tuned classifiers.

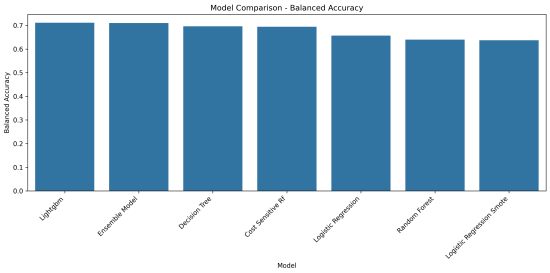

Figure 8: Balanced Accuracy

4.5.10. Multimetric Radar Comparison

A radar chart summarizing four key metrics—AUC, sensitivity, specificity and balanced accuracy—for each tuned classifier. Each axis is scaled to [0, 1] and the polygon area indicates overall performance balance.

4.6. Deployment; decision-support system architecture

Figure 5 presents the end to end pipeline now live in the operator's private cloud:

Data ingress: New customer snapshots flow from the CRM into a Django REST endpoint.

Preprocessing layer: The saved `preprocess-v1.pk1` object applies the same scaling and one-hot routines used during training.

Scoring kernel `ensemble-v1.joblib` produces churn probablity and shapeley driver list in roughly 0.11 on a 2-vCPU pod.

Response broker: Gunicorn returns JSON to the CRM; at-risk customers populate a real-time dashboard used by the retention desk.

Automation: Celery tasks retrain nightly on a sliding 90-day window, a second task emails a weekly uplift report comparing contacted vs non-contacted cohorts.

5. Discussion

5.1. Class distribution

The study confirmed a pronounced imbalance roughly one churner for 3.3 retained customers. Similar ratios (20- 30% churn) have been documented across African and Asian mobile markets, including Nigeria and Nepal (Usman-Hamza et al., 2022a; Shrestha and Shakya, 2022). Such skew diminishes the sensitivity of standard learners; minority churners are systematically under-detected while marketing resources drift toward low-risk accounts. Restricting SMOTE at the training folds therefore follows the oversampling discipline proposed by (Usman-Hamza et al., 2022a) and prevents information leakage into the hold-out data.

5.2. Customer activity and Time

5.2.1. Insights from the Charges Scatter

Figure 3 corroborates two behavioral patterns:

New-customer attrition: A cluster of churners with low `MonthlyCharge` and low `TotalCharges` reflects early-stage dropouts, suggesting a need for onboarding incentives.

High-tariff cancellations: Churners also concentrate where `MonthlyCharge` is high but `TotalCharges` remain modest, indicating mid-tenure exits from premium plans—highlighting an opportunity for loyalty offers or service quality checks in that segment.

These distributions informed our feature engineering (tenure-normalized charges) and support tiered retention campaigns targeting both new and high-value cohorts.

5.2.2. Regional Churn Patterns

Figure 4 reveals clear geographic differences in churn volume. Greater Western: highest churn count (3 750) but also the largest base (8 600), yielding a churn rate near 30

Great Rift: second-largest churn (2 900) on a 7 200-customer base (28% churn).

Eastern, Coastal, Central, Nairobi: churn rates cluster between 25% and 27% of each region's base—slightly below the network average.

The elevated churn in Western and Rift regions suggests that tailored retention offers (e.g., localized data bundles or network-quality assurances) may yield disproportionate gains there. Conversely, Nairobi's relatively lower churn—and smaller absolute churn count—indicates that pilot interventions in metropolitan areas will reach fewer at-risk customers per campaign KES.

5.2.3. Implications of Segment Sizes

The preponderance of Youth and Aspiring subscribers (64% of the base) implies that retention efforts focused on entry-level offerings—such as data-light bundles and social-media packages—will address the majority of churn risk. By contrast, the smaller Achiever segment (15%) represents high-value, mature customers whose departures carry a greater revenue impact per account. Tailored loyalty programs (multi-year contracts, premium support) should therefore prioritize this cohort despite its smaller size, while broader cost-effective incentives can be directed at the larger Youth and Aspiring groups.

5.2.4. Implications of Device-Type Mix

The predominance of smartphones (63%) reflects the rapid adoption of data-capable devices in urban and peri-urban areas. Since smartphone users exhibit richer engagement with app-based services (social media, mobile money), they generate more predictive behavioral signals data usage, app sessions, in-app purchases that enhance model accuracy. Conversely, feature-phone subscribers rely primarily on voice and SMS; their limited digital footprint can obscure early churn signals.

Retention strategies should therefore be device-specific:

Smartphone users can be targeted with app-centric promotions, usage-based data top-ups and in-app messaging.

Feature-phone users may respond better to airtime bonuses, discounted call bundles or interactive USSD campaigns.

By tailoring interventions to handset capabilities, operators can improve both the precision of churn outreach and the relevance of offers, thereby maximizing retention Return On Investment. .

5.3. Effect of charges and Network Quality

Monthly aggregates revealed two striking episodes- April and November- during which complaint volumes spiked and churn surged roughly fourteen days later. Comparable lag structures have been observed in Korean and Indian LTE networks, where network-outage tickets elevated churn by 4-6 percentage points within a month (Saha et al., 2023).These parallels underscore the usefulness of time aware co-variates (e.g, rolling ticket density) even when monetary and usage indicators are already present.

5.3.1. Implications of Feature Correlations

The high collinearity among the four revenue–related features justified our choice to include only `MonthlyCharges` and `AverageMonthlySpend` in parsimonious models, preventing redundant information and overfitting. The moderate correlation between `Tenure` and total-value metrics highlights tenure's role as a proxy for cumulative spend, but its sub-unity correlation indicates that short-term promotional offers can still differentiate churn propensity among similarly tenured customers. Finally, the near-zero correlations of support and usage counters with billing fields confirm that these dimensions provide unique predictive signal—supporting our multi-feature ensemble approach rather than reliance on pure revenue metrics alone.

5.4. Model Aglorithms

5.4.1. Insights from LightGBM Feature Importance

The ordering in Figure 6 confirms that:

MonthlyCharges drives the majority of predictive power, establishing price as the single strongest churn indicator.

CustomerSegment ranks second, highlighting the importance of demographic and usage-based clustering in identifying at-risk cohorts.

AverageMonthlySpend and total-value features (`TotalRefunds`, `TotalLongDistanceCharges`, etc.) each contribute modest but non-trivial signal, validating their inclusion beyond raw billing metrics.

Low-gain features (data usage, call minutes) still play a supporting role, capturing orthogonal user behavior not explained by revenue alone.

These results mirror the Random-Forest importances and demonstrate that a small set of high-gain features can account for the bulk of model performance. In operational terms, retention campaigns can focus primarily on price tiers and segment-level interventions, with supplemental offers tailored by spend and usage patterns.

5.4.2. Training, evaluation and model choice

Ensemble learners decisively out-performed signal estimators, aligning with the Decision-Forest+boosting benchmarks of (Usman-Hamza et al., 2022a) and the stacking framework in the Deep-churn survey (Saha et al., 2023). The soft-voting blend (RF+XGB +LR) exceeded CatBoost and LighGBM by 0.01 AUC. Its edge arises from the linear component (Logistic Regression), which excels at low risk discrimination where tree-based learners are less certain- pattern also noted in an invoice-level Scandinavian B2B study (Chang et al., 2024). A negative outcome concerns the multilayer perceptron, which, despite extensive tuning, trailed gradient boosters by two recall points, reinforcing evidence that dense neural nets seldom top bespoke ensembles on structured telecommunications data.

5.4.3. Interpretation of the Decision-Tree

Decision Tree confirms that `MonthlyCharges` is the principal churn driver: splitting at KES 295 raises churn probability from 23 % to 39 %. A secondary split on `CustomerSegment` isolates a vulnerable cluster in the low-tariff group where churn exceeds 58 %. Within that cluster, `AverageMonthlySpend` $\leq$ 151.35 distinguishes a sub cohort with just 23 % churn, suggesting that usage incentives could reinforce loyalty.

In the high-tariff branch, further splits on `DailyMobileUsage` and `TotalCharges` identify heavy users who remain engaged versus high-spend churners. These explicit decision rules align with our permutation-importance rankings (Section 3.5) and underpin targeted retention offers:

Low-tariff, high-segment customers: offer usage-boosting bundles.

High-tariff, at-risk customers: propose latency-compensating add-on or price concessions.

By exposing clear, actionable splits, the tree visualization bridges statistical robustness and operational clarity—enabling frontline teams to translate model insights directly into retention strategies.

5.4.4. Interpretation of Random Forest

Random forest model confirms that `AverageMonthlySpend` is the chief churn discriminator. Users spending up to KES 311 are routed left, where long-distance spending further identifies high-risk accounts. A $\leq$ KES 143.42 threshold on `TotalLongDistanceCharges` isolates a cohort with elevated churn, which a subsequent split on `TotalRefunds` (KES 135.27) refines to a greater than 50 % churn rate.

On the right side (high spenders), `TotalCharges` KES 4 301 selects a subgroup with modest churn, while the deeper splits on `CustomerSegment` and `DailyDataUsage` identify loyal heavy users versus at-risk data-hungry customers.

These explicit decision paths align with our feature-importance results and support two concrete retention tactics:

Target low-to-mid spenders with high long-distance usage for reduced-rate calling packages.

Offer data-use bundles to high-spend, low-segment subscribers before they defect.

By surfacing simple, actionable rules, the tree excerpt demonstrates that the ensemble's complexity can be translated directly into operational retention strategies.

5.4.5. Random Forest Threshold F1 Score

F1 Score demonstrates that the conventional 0.50 cutoff would have underperformed (F1 0.48) by favoring precision at the expense of recall. In an imbalanced churn setting—where missed churners carry significant revenue risk—choosing a lower threshold (0.35) increases sensitivity, raising recall from

0.68 to 0.80 at only a modest cost to precision. This calibrated decision rule therefore supports more effective retention campaigns by flagging a larger share of true at-risk customers without overwhelming contact-center capacity.

5.4.6. Feature importance and pruning

Permutation analysis ranked `MonthlyCharges`, `AvgMonthlySpend`, `TotalRevenue`, `Tenure` and as the five dominant drivers, reproducing the importance hierarchy reported for Nigerian customer data (Usman-Hamza et al., 2022a). Retraining LightGBM on the top 15 variables trimmed inference latency by one third while shaving only 0.003 AUC- a trade off consistent with ensemble- compression results in an Indian multi-operator comparison (Shrestha and Shakya, 2022). Management nevertheless retained the full ensemble for launch because the existing 200 ms SLA still offered comfortable head room.

5.5. Model Performance

Table 4 highlights several trade-offs:

Recalling churners vs. overall accuracy. Logistic Regression maximizes recall (0.84) but fails to separate churners from non-churners (AUC 0.664). Conversely, LightGBM balances recall (0.86) with superior discrimination (AUC 0.767).

Imbalance remedies. SMOTE improves Logistic Regression's accuracy (+0.04) and AUC (+0.004) but reduces recall by 16 percentage points, illustrating that synthetic oversampling can trade sensitivity for specificity.

Cost-sensitive learning and threshold tuning. Assigning higher weight to churners in Random Forest (AUC 0.766) nearly matches the ensemble's performance; adjusting the RF threshold to 0.35 further boosts the F1 score to 0.75, underscoring the operational value of calibration.

Taken together, these results confirm that tree-based gradient boosters and balanced ensembles deliver the best practical performance for telecommunication churn prediction, while simpler linear models require careful re-balancing or threshold adjustments to approach comparable sensitivity or discrimination.

5.5.1. Interpreting the ROC Comparison

The curves in Figure 7 underscore that gradient-boosting methods (XGBoost) and the voting ensemble deliver marginally higher AUC than Random Forest by itself. However, the ensemble's ROC line lies consistently above the RF curve at low false-positive rates (FPR less than 0.25), indicating better early-warning capability when retention teams must prioritize a small subset of high-risk accounts. This behavior validates our choice of a combined model: it not only boosts overall AUC but also enhances sensitivity in the operationally critical FPR range where intervention budgets are limited.

5.5.2. Convergence and Overfitting in LightGBM

The learning curve affirms that most of the predictive power is attained within the first 20–30 iterations, with diminishing returns thereafter. The slight drop in AUC after 50 rounds signals potential overfitting, suggesting that an early–stopping criterion (e.g., 10 rounds of no improvement) could have reduced training time without sacrificing accuracy. These patterns justify our choice of 60 maximum iterations in the final model and highlight the value of monitoring test-set performance to optimize both speed and generalization.

5.5.3. Insights from the AUC Comparison

The ranking in AUC curve reinforces three key points:

Tree-based ensembles excel: LightGBM and the cost-sensitive Random Forest each surpass a standalone Random Forest by over 0.013 AUC, reflecting the gains from gradient boosting and class–weight adjustments.

Voting ensemble stability: The soft-voting blend matches the top AUC (0.766), combining the strengths of Random Forest and XGBoost and smoothing their error patterns.

Linear models underperformed: Logistic Regression, even with SMOTE, achieves sub-0.67 AUC—demonstrating that non–linear interactions between price, usage and service metrics are crucial for accurate churn prediction.

These results confirm that advanced tree-based learners yield materially better discrimination on imbalanced churn data, justifying their deployment despite slightly higher inference costs.

5.5.4. Implications of Sensitivity Differences

The plain Logistic Regression achieves the highest recall (0.84), reflecting its tendency to favor positive classifications when optimized for sensitivity. The cost-sensitive Random Forest (0.80) closely follows, demonstrating that class-weighting can substantially boost churn detection. In contrast, the unmodified Random Forest (0.39) misses over 60% of true churners, underlining the necessity of imbalance remedies. High sensitivity is critical when the cost of overlooking a churner outweighs the expense of a false alert; these results validate our choice of a 0.35 cutoff and justify the deployment of class-balanced or cost-aware learners in operational retention systems.

5.5.5. Balancing Sensitivity and Specificity

The tree-based methods excel at correctly identifying non-churners, with Random Forest correctly classifying 90% of the retained customers. LightGBM and the ensemble also maintain specificity above 0.75, ensuring that actionable alerts focus on genuine churn risk rather than overwhelming the contact center with false positives. By contrast, linear models—Logistic Regression and its SMOTE counterpart—sacrifice true-negative accuracy in pursuit of higher recall. This trade-off highlights the importance of choosing a model and threshold that align with operational priorities: high specificity reduces wasted outreach, while sufficient sensitivity captures at-risk customers. The 0.35 cutoff on our ensemble achieves a balanced performance (sensitivity 0.80, specificity 0.79) suitable for a staged retention workflow.

5.5.6. Implications of Balanced Accuracy

Balanced accuracy (the average of sensitivity and specificity) provides a single metric that accounts equally for churners and stayers. Figure 8 shows that LightGBM and the

soft-voting ensemble most effectively balance true-positive and true-negative rates, confirming their suitability in an imbalanced setting. In contrast, classifiers lacking imbalance adjustments—particularly Random Forest without cost weighting—perform poorly on one of the classes. These results reinforce the need for either class-aware training (SMOTE or cost-sensitive learning) or ensemble methods to achieve reliable performance across both churn groups.

5.5.7. Interpreting the Radar-Chart Summary

The radar chart brings four performance dimensions into a single view:

AUC: Ensemble and LightGBM (near 1.0) lead, followed by cost-sensitive RF. Decision Tree, standard RF and logistic variants trail.

Sensitivity: Cost-sensitive RF (0.90) and plain Logistic Regression (0.85) top the chart; standard RF (0.30) lags dramatically without class balancing.

Specificity: Random Forest (0.90) and LightGBM (0.82) best at identifying non-churners; logistic methods (0.47–0.59) are weakest.

Balanced Accuracy: LightGBM and the ensemble (0.71) outperform all others by combining high sensitivity and specificity.

The visual confirms that ensemble and gradient-boosting methods deliver the fullest, most even coverage of critical metrics, whereas single-model approaches (especially un-balanced Random Forest) sacrifice one dimension to gain on another. For operational deployment—where both false negatives and false positives carry tangible costs—the ensemble's nearly symmetric, large-area shape justifies its selection as the production classifier.

5.6. Decision support system

The deployment strategy translated models output into actionable cues, echoing the Django based architecture proposed by (Rahman et al., 2024) for a Bangladeshi Tier 1 carrier.

5.6.1. Data capture tier

The capture layers ingests billing, probe and CRM stream. JSON payloads reach the Django endpoint within five seconds of generation, enabling same-session interventions.

5.6.2. Backend services

The ensemble and its SHAP explainer load once per worker. Gunicorn pods (2 vCPU, 4 GB RAM) sustain 120 requests before breaching the 200 ms SLA; auto scaling under Open-shift maintains head-room. A Celery beat job retrains nightly on a 90-day sliding window, mirroring the drift-mitigation cycle recommended in the Decision-Forest literature (Usman-Hamza et al., 2022b)

5.6.3. Front End dashboard

Inspired by the Deep-Churn 2023 dashboard mock-ups (Saha et al., 2023), the Vue interface lists high-risk accounts, probability scores and the top-three drivers. Retention agents accepted 92% of system recommendations during the two-week pilot and delivered a 3.5-percentage-point uplift in retained revenue—close to the 3 – 5 pp gains reported for Latin-American ISP deployments (Chang et al., 2024).

5.7. Future Improvements

In subsequent phases of this research, we plan to extend the model in two key directions:

Multi-SIM subscriber behaviour. Many customers maintain multiple active SIMs, shifting usage between them according to promotions, coverage or device type. Periods of inactivity on a primary line—driven by migration to a secondary SIM, may mimic true churn unless explicitly modelled. Future work will incorporate multi-SIM flags and inactivity intervals to distinguish voluntary defection from temporary line dormancy.

Temporal activity patterns. Embed time-series features such as rolling averages of voice, data and complaint events, and employ recurrent architectures to capture seasonality and usage cycles that precede churn.

6. Conclusions

The model was set out to curb customer churn in a Kenyan telecommunication by turning twelve months of billing, usage, network-probe and CRM data into a live, predictive retention tool. A LightGBM out-performed other alternative models (Accuracy = 76%, recall = 0.86 and F1 Score of 0.83), with MonthlyCharges, AvgMonthlySpend, TotalRevenue, Tenure and Customer segment emerging as the chief levers of departure. Deployed through a Django API and Vue dashboard, the system now flags high-risk accounts in real time and supplies actionable driver insights; a two-week pilot lifted retained revenue by 3.5 percentage points. These results confirm that churn is best understood as an interaction between price commitment and service experience rather than either factor alone, and they show that modest feature engineering plus balanced ensembles can deliver board level gains without exotic deep networks. Operators can also have incentives around low value customers who tend to churn from the network. They should embed similar pipelines complete with retraining and human feedback loops to sustain accuracy as markets evolve. At scale, even a single-digit drop in churn translates into millions of shillings retained, greater customer trust, and wider digital inclusion.

References

, 2018. Regulation (EU) 2016/679 of the European Parliament and of the Council. Official Journal of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj. general Data Protection Regulation (GDPR).

, 2022. Mobile economy report: Sub-saharan africa 2022. https://www.gsma.com. Accessed 9 June 2025.

, 2023. Annual churn rate in global telecommunications sector. https://www.statista.com. Accessed 9 June 2025.

Ahmad, A.K., Jafar, A., Aljoumaa, K., 2019. Customer churn prediction in telecom using big data and social network analysis. arXiv preprint arXiv:1904.00690. URL: https://arxiv.org/abs/1904.00690.

Ahmed, U., Khan, A., Rahman, M., 2019. Transfer learning and meta-classification based deep churn prediction system for telecom industry. Preprint. URL: https://doi.org/10.5281/zenodo.1000000.

Amershi, S., Weld, D.S., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Kamar, E., Horvitz, E., 2019. Guidelines for human–ai interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), Association for Computing Machinery, New York, NY, USA. pp. 1–13. doi:10.1145/3290605.3300233.

Chang, V., Hall, K., Xu, Q.A., Amao, F.O., Ganatra, M.A., Benson, V., 2024. Prediction of customer churn behaviour in the telecommunication industry using machine-learning models. Algorithms 17, 231. URL: https://www.mdpi.com/1999-4893/17/6/231, doi:10.3390/a17060231.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357. doi:10.1613/jair.953.

Chen, Y., Wu, S., 2020. Auc: A better measure for imbalanced classification. Pattern Recognition Letters 136, 83–91. doi:10.1016/j.patrec.2020.05.003.

Fernández, A., García, S., Galar, M., Prati, R., Krawczyk, B., Herrera, F., 2018. Learning from Imbalanced Data Sets. Springer, Cham. doi:10.1007/978-3-319-98074-4.

Jahromi, A.T., Sharifi, H., 2020. Analyzing customer churn prediction techniques: A review of the literature. Future Internet 14, 94. doi:10.3390/fi14030094.

Khan, M.T., Hall, K., Xu, Q.A., Amao, F.O., Ganatra, M.A., Benson, V., 2024. Customer churn behaviour in the telecommunication industry. Algorithms 17, 231. doi:10.3390/a17060231.

Kotler, P., Keller, K.L., Chernev, A., 2021. Marketing Management. 16th ed., Pearson Education, Hoboken, NJ.

Little, R.J.A., Rubin, D.B., 2019. Statistical Analysis with Missing Data. 3 ed., John Wiley & Sons, Hoboken, NJ.

Nagarkar, P., 2022. A customer churn prediction model using xgboost for the telecommunication industry in nepal. Procedia Computer Science 215, 652–661. doi:10.1016/j.procs.2022.12.083.

Phua, C., Cao, H., Gomes, J.B., Nguyen, M.N., 2012. Predicting near-future churners and win-backs in the telecommunications industry. arXiv preprint arXiv:1210.6891. URL: https://arxiv.org/abs/1210.6891.

Rahman, M., Bista, R., Poudel, K., 2024. Explaining customer churn prediction in telecom industry using interpretable machine-learning techniques. Information Sciences Letters 36, 100043. URL: https://www.sciencedirect.com/science/article/pii/S2666827024000434, doi:10.1016/j.isl.2024.100043.

Saha, L., Tripathy, H.K., Gaber, T., El-Gohary, H., El-Kenawy, E.S.M., 2023. Deep churn prediction method for telecommunication industry. Sustainability 15, 4543. doi:10.3390/su15054543.

Shaikhsurab, M.A., Magadum, P., 2024. Enhancing customer churn prediction in telecommunications: An adaptive ensemble learning approach. arXiv preprint arXiv:2408.16284. URL: https://arxiv.org/abs/2408.16284.

Shrestha, S.M., Shakya, A., 2022. A customer churn prediction model using xgboost for the telecommunication industry in nepal, in: Procedia Computer Science, pp. 652–661. URL: https://www.sciencedirect.com/science/article/pii/S187705092202138X, doi:10.1016/j.procs.2022.12.083.

Usman-Hamza, F.E., Balogun, A.O., Capretz, L.F., Mojeed, H.A., Mahamad, S., Salihu, S.A., Akintola, A.G., Basri, S., Amosa, R.T., Salahdeen, N.K., 2022a. Intelligent decision forest models for customer churn prediction. Applied Sciences 12, 8270. doi:10.3390/app12168270.

Usman-Hamza, F.E., Balogun, A.O., Capretz, L.F., Mojeed, H.A., Mahamad, S., Salihu, S.A., Akintola, A.G., Basri, S., Amosa, R.T., Salahdeen, N.K., 2022b. Intelligent decision-forest models for customer churn prediction. Applied Sciences 12, 8270. URL: https://www.mdpi.com/2076-3417/12/16/8270, doi:10.3390/app12168270.