

PREDICTING CUSTOMER CHURN IN THE TELECOMMUNICATIONS SECTOR: A CLASSIFICATION AND CLUSTERING ANALYSIS



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Student	Gobinath Gothandapani
Student ID	2538599
Module	ACCA7015-Data Mining and Visualisation
Assessment	Exploratory Data Analysis of Telecom Customer Churn
Dataset	Telecom_customer churn.csv
Lecturer:	Dr Nough Elmitwally

Contents

Abstract	3
1. Introduction	3
2. Critical Literature Review	4
3. Methodology and Model Implementation	6
3.1 Analytical Design and Dataset	6
3.2 Data Preparation and Feature Engineering	8
3.3 Classifier Selection, Benchmarking and Ensemble Design	8
3.4 Evaluation, Segmentation and Governance	9
4. Results and Analysis	10
4.1 Comparative Classification Performance	10
4.2 Threshold Selection	12
4.3 Risk Prioritisation and Lift	13
4.4 Feature Importance	15
4.5 Churner Segmentation	16
4.6 Advanced Validation, Explainability and Governance Evidence	17
5. Critical Discussion	18
6. Business Recommendations	19
7. Conclusion	20
References	21
Appendix A. Implementation Evidence	22

Abstract

Customer churn prediction is valuable only when it supports reliable, economically defensible retention decisions. This study analyses 100,000 telecom customers and develops an enhanced supervised and unsupervised workflow. Logistic Regression, Decision Tree, Random Forest, two LightGBM configurations and a validation-selected weighted ensemble are implemented using an independent reload of the raw CSV, stratified 60:20:20 development partitions, leakage-safe imputation and encoding, behavioural feature engineering, validation tuning and five-fold cross-validation. The final workflow expands evaluation beyond accuracy to balanced accuracy, precision, recall, specificity, F1, Matthews correlation coefficient, ROC-AUC, average precision, Brier score and log loss. It also implements cost-sensitive threshold selection, probability calibration, bootstrap confidence intervals, paired model comparison, feature interpretation, SHAP explanations, demographic fairness diagnostics, lift, gains and automated quality checks.

Among the conventional classifiers, Random Forest achieved 63.07% accuracy, 64.18% recall, a 63.27% F1-score and 0.6872 ROC-AUC. The weighted LightGBM ensemble delivered the strongest overall test performance, with 64.55% accuracy, a 64.22% F1-score and 0.7019 ROC-AUC. A retention-focused threshold increased churn recall to 90.98%, while the highest-risk decile produced 1.63 times the population churn rate. K-means segmented 49,562 churners into Low Usage, High Value and Long Tenure profiles; the silhouette coefficient of 0.2963 indicates overlapping but commercially useful structure. The combined findings support risk-ranked, segment-specific and experimentally evaluated retention strategies.

Keywords: telecom churn, Logistic Regression, Decision Tree, Random Forest, LightGBM, cross-validation, calibration, SHAP, fairness, lift, K-means

1. Introduction

Telecommunications providers operate in markets where customers can compare tariffs, devices and service quality with relatively low switching barriers. Churn therefore affects recurring revenue, acquisition efficiency and customer lifetime value. A useful analytical solution must do more than maximise a single score: it must rank customers reliably, quantify uncertainty, support a financially meaningful operating threshold and connect risk to an actionable retention response.

The supplied dataset records churn status alongside revenue, minutes of use, behavioural change, service subscriptions, handset characteristics, tenure, geography and demographics. Churn is almost balanced at 49.56%, so accuracy remains interpretable, although missed churners and unnecessary interventions have different business costs. The analytical design therefore combines

threshold-free ranking measures, threshold-dependent classification measures, probability-quality measures and campaign lift.

The analysis compares Logistic Regression, Decision Tree and Random Forest as genuinely different conventional classifiers, then evaluates two regularised LightGBM models and a weighted probability ensemble. Logistic Regression provides directional coefficients; Decision Tree exposes human-readable nonlinear rules; Random Forest reduces single-tree variance; and LightGBM supplies scalable nonlinear learning for mixed, high-dimensional predictors.

Classification estimates who is at risk, but it does not explain which commercial response is suitable. K-means is therefore applied only to customers with Churn = True using revenue, usage, behavioural change, tenure and equipment age. Internal validation, stability checks and original-scale profiles are included so that clusters are treated as decision-support summaries rather than unquestioned natural classes.

The aim is to classify customer retention, critically compare model families and reveal patterns among churners. The objectives are to: (1) construct and evaluate multiple classifiers under a common data protocol; (2) select model and threshold using validation evidence rather than test optimisation; (3) evaluate discrimination, calibration, uncertainty, fairness and operational lift; (4) segment churners and test cluster quality; and (5) translate findings into responsible, testable retention recommendations.

2. Critical Literature Review

Predictive analytics converts historical behaviour into estimates of future outcomes. In churn management, supervised classification learns from labelled outcomes, while unsupervised learning identifies structure without using the churn label. A recent survey of telecommunications churn research found widespread use of logistic regression, support vector machines, decision trees, random forests, boosting and neural networks, but recurring weaknesses in dataset comparability, explainability and deployment [1]. A model that discriminates well therefore does not prove that its important variables cause churn.

Linear models remain useful because coefficients are inspectable, but they require explicit interactions and nonlinear transformations. Tree methods learn such relationships directly. Comparative studies frequently report strong results from random forests and gradient boosting after feature engineering and tuning [2], [5], [6]. However, headline accuracy values are not directly transferable because churn definitions, class proportions, leakage risk and train-test procedures differ. An accuracy above 90% on one dataset is not a valid benchmark for this 100-variable dataset.

LightGBM sequentially adds trees that correct previous errors and improves efficiency through histogram-based learning and specialised sampling and feature-bundling methods [4]. Its strengths here are scalability and nonlinear modelling; its weaknesses include possible overfitting, reduced transparency and non-causal feature importance. The notebook addresses overfitting through small learning rates, leaf constraints, subsampling, regularisation and early stopping.

Ensembling can improve stability when component errors differ. Xu et al. [7] used stacking and soft voting, while Khoh et al. [3] developed an optimised weighted ensemble. Yet ensemble complexity is justified only when improvement is meaningful. In this study, the ensemble improved accuracy by just 0.18 percentage points over Model A and shared the same ROC-AUC to four decimal places, so its benefit is incremental.

Evaluation must reflect the decision context. Precision measures contact efficiency; recall measures churn capture; specificity measures correct retention; F1 balances precision and recall; and Matthews correlation uses all four confusion-matrix cells [16]. ROC-AUC evaluates ranking, average precision positive-class retrieval [9], and Brier score and log loss probability quality [15].

K-means minimises within-cluster squared distance and is efficient after standardisation, but it favours centroid-shaped groups and is sensitive to scale and outliers. Segmentation research shows that combining prediction with customer profiles can support differentiated interventions [8]. Cluster validity must still be assessed. The silhouette coefficient compares cohesion and separation [10]; the score of 0.2963 obtained here indicates overlapping rather than sharply separated groups.

Sampling, tuning and leakage can change model rankings [2], [6]. The workflow therefore reloads raw data, fits preprocessing within development data, separates validation from testing and uses five-fold stratified cross-validation. Bootstrap intervals and an exact paired test distinguish small score gaps from consistent improvement [17].

Prediction estimates association, not intervention effect. Logistic coefficients, forest importance, LightGBM gain and SHAP provide complementary explanations [14], but none proves causality. Fairness audits are required because demographic and proxy variables may create unequal errors. Controlled experiments are still needed to estimate incremental retention impact.

Table 1. Critical comparison of selected literature

Study	Approach	Main contribution	Critical relevance to this study
Barsotti et al. (2024) [1]	Survey of TelCo churn methods	Synthesises a decade of algorithms and research gaps	Supports multi-metric evaluation; warns that deployment and comparability remain unresolved.
Imani and Arabnia (2023) [2]	Multiple classifiers, sampling and optimisation	Shows that preprocessing and tuning alter model rankings	Justifies separating algorithm quality from sampling and optimisation choices.
Khoh et al. (2023) [3]	Optimised weighted ensemble	Combines model probabilities for churn prediction	Supports blending, but the present ensemble gain must be judged against added complexity.
Ke et al. (2017) [4]	LightGBM gradient boosting	Efficient scalable learning for high-dimensional tabular data	Provides the algorithmic foundation for the two classification models.
Zhang et al. (2022) [8]	Telecom segmentation and prediction	Connects customer segments with churn modelling	Supports clustering churners to design differentiated actions.
Lundberg and Lee (2017) [14]	SHAP model explanation	Consistent global and local contribution framework	Supports explanation of the selected ensemble while preserving the distinction between association and causation.

3. Methodology and Model Implementation

3.1 Analytical Design and Dataset

The study used a quantitative predictive-analytics design implemented in Python with pandas, NumPy, Matplotlib, Seaborn, scikit-learn, LightGBM, SHAP and Joblib. The enhanced notebook covers data audit, engineered predictors, protected data partitioning, conventional and boosting models, cross-validation, calibration, uncertainty, paired comparison, responsible-use diagnostics, risk prioritisation, clustering and reproducible export.

The dataset contained 100,000 rows and 100 original columns, with no duplicate records. The target comprised 49,562 churners and 50,438 retained customers. Missingness was substantial in

several demographic variables, including number of cars and dwelling characteristics. Missing values were not removed wholesale because complete-case deletion would discard information and could systematically under-represent particular customer groups.

Item	Analysis value
Rows $\times$ original columns	100,000 $\times$ 100
Target distribution	50,438 retained (50.44%); 49,562 churned (49.56%)
Independent modelling matrix	Original CSV reloaded; identifier and target removed; constants excluded
Training / validation / testing	60,000 / 20,000 / 20,000, stratified and non-overlapping
Engineered predictors	11 behavioural ratios plus one row-level missingness indicator
Classification models	Logistic Regression, Decision Tree, Random Forest, LightGBM Model A, LightGBM Model B and weighted ensemble
Complementary method	K-means on 49,562 churners; K = 2–8 compared using four validation dimensions

Table 2. Dataset, feature construction and protected experimental design

Figure 1 visualises the target distribution used throughout the classification analysis.

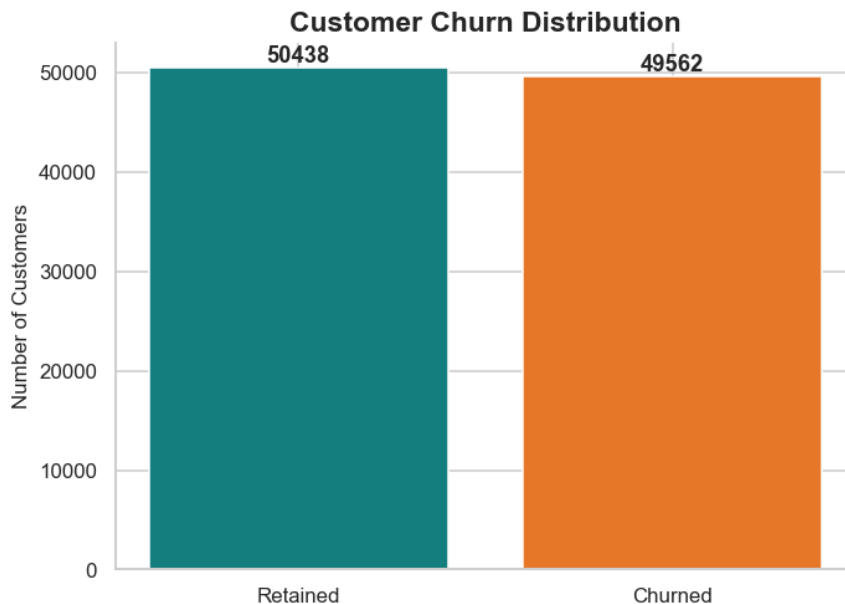


Figure 1. Balanced distribution of retained and churned customers in the full dataset.

The split is almost balanced: 50.44% of customers were retained and 49.56% churned. This makes accuracy interpretable, but it does not remove the need to examine recall, false negatives and ranking quality because the commercial costs of the two error types differ.

3.2 Data Preparation and Feature Engineering

Exploratory processing was separated from modelling. The classification stage rereads the raw CSV, constructs eleven behavioural ratios and a missing-feature count, replaces infinite ratios with missing values and removes constant fields. This prevents exploratory imputation or winsorisation from leaking into model comparison.

For scikit-learn models, numerical variables are median-imputed with missingness indicators and scaled; categorical variables receive a ‘Missing’ label and infrequent-aware one-hot encoding. Preprocessing is fitted only on development data. LightGBM retains native categorical types and missing-value handling.

Engineered predictors capture revenue and overage efficiency, call completion and failure, usage and revenue change, equipment tenure, subscription activity, and six-month-to-current usage and revenue. These ratios express commercially meaningful relationships beyond raw magnitudes.

A stratified 60:20:20 split produced 60,000 training, 20,000 validation and 20,000 test records. Tuning, early stopping, blend weights, model choice and threshold use no test labels. Frozen models are refitted on the 80,000-record development set and evaluated once on test data.

3.3 Classifier Selection, Benchmarking and Ensemble Design

Table 3 summarises the benchmark. Candidate configurations are selected by validation ROC-AUC, with F1-score and Matthews correlation as secondary criteria. The selected conventional configurations then undergo five-fold stratified cross-validation before final refitting and independent test evaluation.

Technique	Implementation in updated notebook	Strength	Limitation / judgement
Logistic Regression	L2 models; $C = 0.03\text{--}5.00$; median/indicator preprocessing and one-hot encoding.	Transparent probability baseline; signed standardised coefficients.	Linear decision surface unless interactions are engineered.
Decision Tree	Gini, entropy or log-loss; depth 5–16; controlled minimum leaf sizes.	Readable nonlinear rules and interaction discovery.	High variance; requires strict regularisation and external validation.
Random Forest	450 trees in the final model; depth, leaf-size and feature-subsampling alternatives; out-	Robust nonlinear benchmark with reduced single-tree variance.	Less transparent; impurity importance can favour continuous/high-cardinality

Technique	Implementation in updated notebook	Strength	Limitation / judgement
	of-bag estimation.		fields.
LightGBM ensemble	Two regularised native-categorical models; early stopping; 41-point validation blend search.	Efficient ranking, nonlinear interactions and flexible operating thresholds.	Best ROC-AUC 0.7019; added complexity is justified by improved ranking, threshold flexibility and lift.

Table 3. Critical comparison and tuning design for classification techniques

No classifier is declared best in advance. Logistic Regression tests interpretability, Decision Tree transparent nonlinearity, Random Forest variance-reduced bagging and LightGBM regularised boosting. Validation evidence selects the final model before unified test evaluation.

Tuning evaluates five Logistic Regression C values, six Decision Tree structures and six Random Forest structures. LightGBM uses up to 2,400 trees, learning rate 0.022, early stopping and structural regularisation. Ensemble weights are searched in 0.025 increments on validation data.

3.4 Evaluation, Segmentation and Governance

All classifiers share accuracy, balanced accuracy, precision, recall, specificity, F1, Matthews correlation, ROC-AUC, average precision, Brier score, log loss, diagnostic curves and fit time. Threshold policies compare 0.50, maximum F1, maximum Youden J and minimum validation cost, assuming a missed churner costs five times an unnecessary contact.

Four hundred bootstrap samples estimate intervals for ROC-AUC, F1-score and recall [17]. A bootstrap AUC-difference interval and an exact McNemar paired-correctness test compare the selected model with its strongest alternative [18]. Calibration curves, Brier score and log loss evaluate probability quality; lift and gains evaluate targeting value.

Clustering retains 49,562 churners and uses robust-scaled revenue, usage, behavioural change, tenure and equipment age. $K = 2-8$ is compared using inertia, silhouette, Calinski–Harabasz and Davies–Bouldin criteria; a composite rank selects K and adjusted Rand indices test seed stability.

The six clustering variables were chosen for commercial interpretability. PCA is used only for visualisation, while segment labels and actions come from original-scale profiles and remain hypotheses for testing.

The final threshold is audited by ethnicity and marital status for selection, precision, recall, false-positive and false-negative rates. Because ethnicity is special-category data, deployment requires necessity, lawful basis, access minimisation and proxy-risk review [11], [12].

4. Results and Analysis

4.1 Comparative Classification Performance

Table 4 compares six classifiers on the held-out 20,000-customer test set at the standard probability threshold of 0.50. Random Forest was the strongest conventional classifier, while the weighted LightGBM ensemble achieved the best overall accuracy, F1-score and ROC-AUC. Evaluating every model under a common test protocol provides a defensible comparison of interpretability, nonlinear capacity and ensemble performance.

Table 4. Comparative held-out classification performance at threshold 0.50

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Analytical role
Logistic Regression	60.21%	60.08%	58.73%	59.40%	0.6394	Interpretable linear baseline
Decision Tree	60.08%	60.61%	55.54%	57.96%	0.6541	Transparent nonlinear comparator
Random Forest	63.07%	62.38%	64.18%	63.27%	0.6872	Best conventional classifier
LightGBM Model A	64.37%	64.04%	64.09%	64.06%	0.7019	Best single boosting model
LightGBM Model B	64.23%	63.86%	64.11%	63.99%	0.7007	Regularised boosting comparator
Weighted ensemble	64.55%	64.23%	64.22%	64.22%	0.7019	Best overall accuracy and F1

Figure 4 compares the ranking quality of the two LightGBM configurations and their weighted ensemble across all possible classification thresholds.

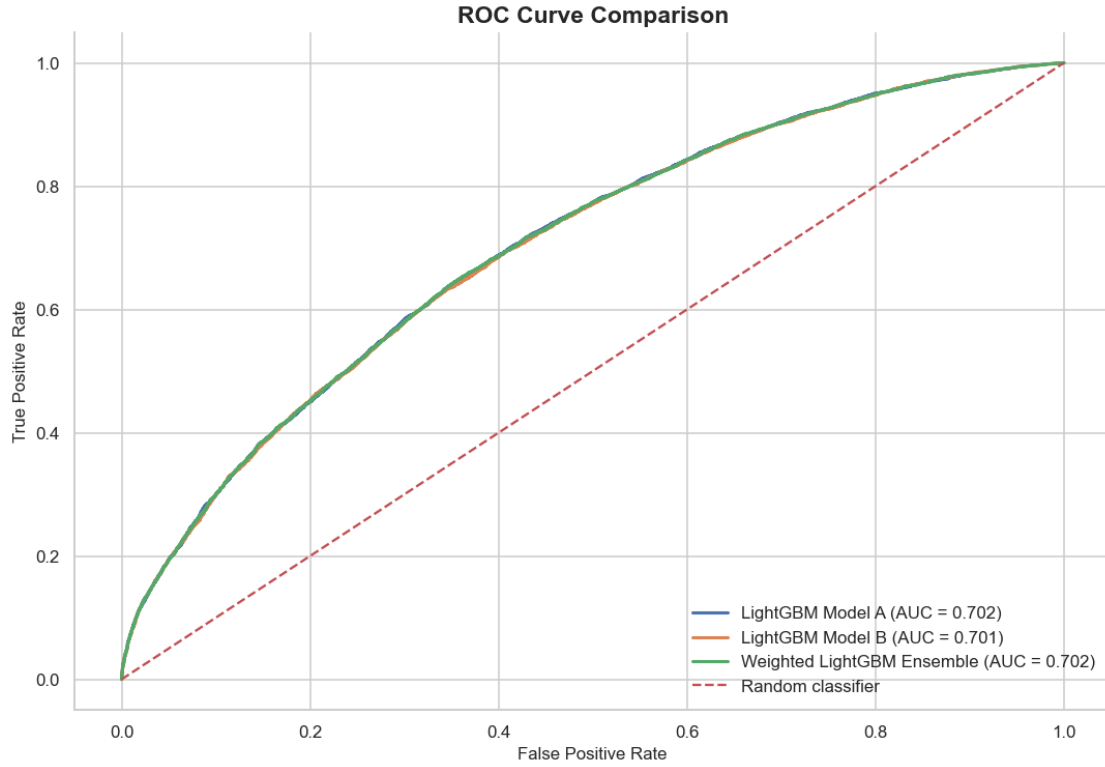


Figure 4. ROC-curve comparison of the two LightGBM models and their weighted ensemble.

The ensemble achieved 64.55% accuracy, 64.23% precision, 64.22% recall, a 64.22% F1-score and 0.7019 ROC-AUC. It exceeded Random Forest by 1.48 percentage points in accuracy and 0.0147 in ROC-AUC, indicating that regularised boosting captured additional nonlinear signal. Model A produced the same ROC-AUC to four decimal places and only 0.18 percentage points less accuracy, so the blend offers an incremental threshold-level improvement rather than a transformational gain. Bootstrap intervals and paired prediction tests are therefore used to assess whether small differences are stable [17], [18].

Figure 2 separates correct predictions from false churn alerts and missed churners.

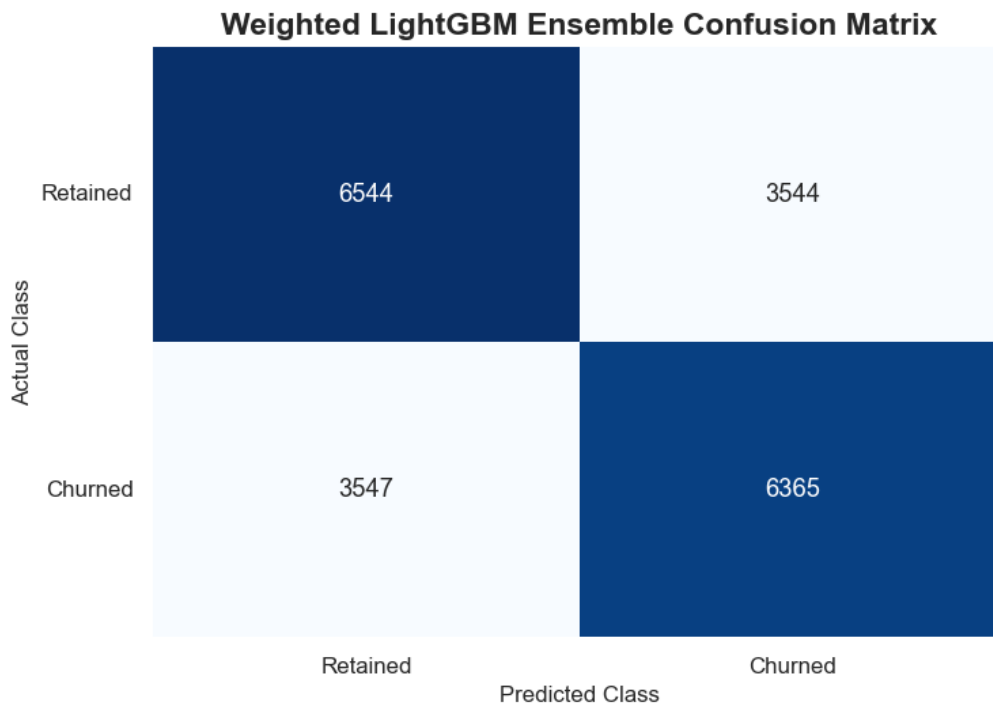


Figure 2. Confusion matrix for the weighted LightGBM ensemble at threshold 0.50.

At the standard threshold, 6,544 retained customers and 6,365 churners were correctly classified; 3,544 retained customers were incorrectly flagged and 3,547 churners were missed. This near-symmetry reflects the balanced target and threshold. Operationally, however, the two errors are not equivalent: a missed churner may represent lost recurring revenue, whereas a false alert consumes contact or incentive capacity.

4.2 Threshold Selection

Table 5 reports the selected operating policies; Figure 3 shows the wider validation-set trade-off.

Table 5. Operating-threshold trade-off on the test set

Policy	Threshold	Accuracy	Precision	Recall	F1
Standard	0.500	64.55%	64.23%	64.22%	64.22%
Accuracy focused	0.491	64.41%	63.49%	66.33%	64.88%
Retention focused	0.332	59.55%	55.62%	90.98%	69.04%

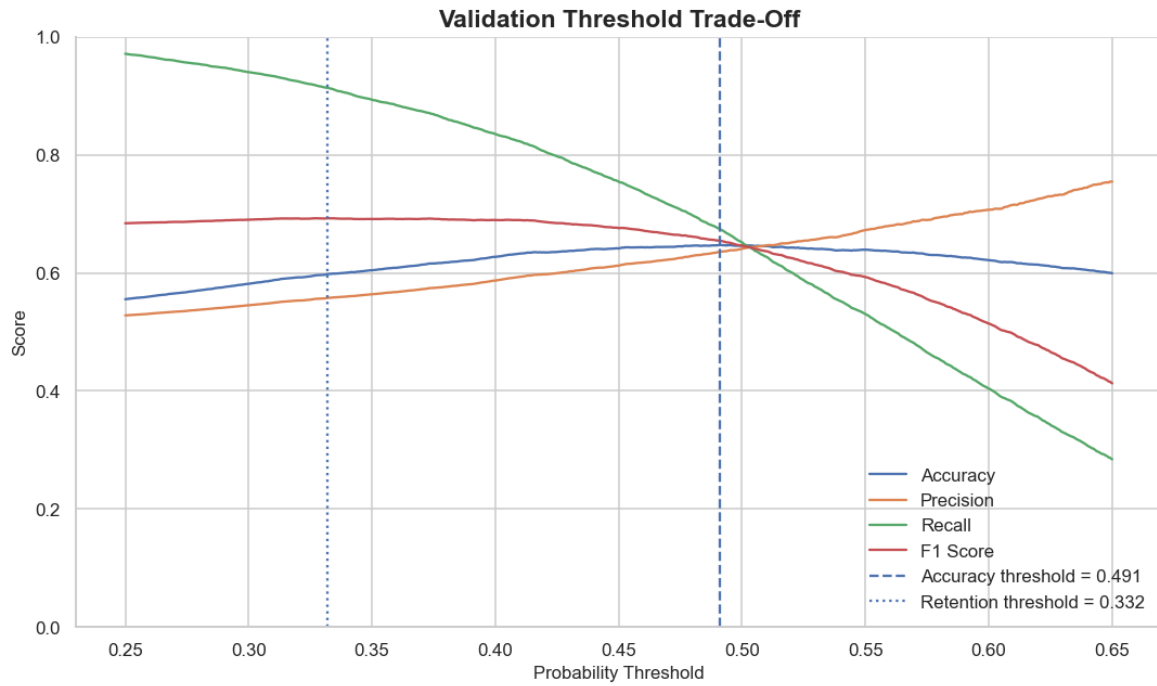


Figure 3. Validation-set trade-off between accuracy, precision, recall and F1 across probability thresholds.

Threshold choice materially changes model behaviour. At 0.491, accuracy was 64.41%, recall 66.33% and F1 64.88%. At 0.332, recall rose to 90.98% and F1 to 69.04%, while precision fell to 55.62% and accuracy to 59.55%. The enhanced notebook generalises this analysis by deriving a minimum-cost threshold on validation data under the explicit 5:1 false-negative-to-false-positive cost assumption, then evaluating the frozen policy on test data.

4.3 Risk Prioritisation and Lift

Table 6 tests risk-band ordering, while Figure 5 measures churn concentration by score decile.

Table 6. Risk-band calibration summary

Risk band	Customers	Actual churners	Average predicted	Observed churn
Low	4,352	1,078	25.69%	24.77%
Medium	7,931	3,684	45.64%	46.45%
High	5,381	3,295	61.62%	61.23%
Critical	2,336	1,855	79.01%	79.41%

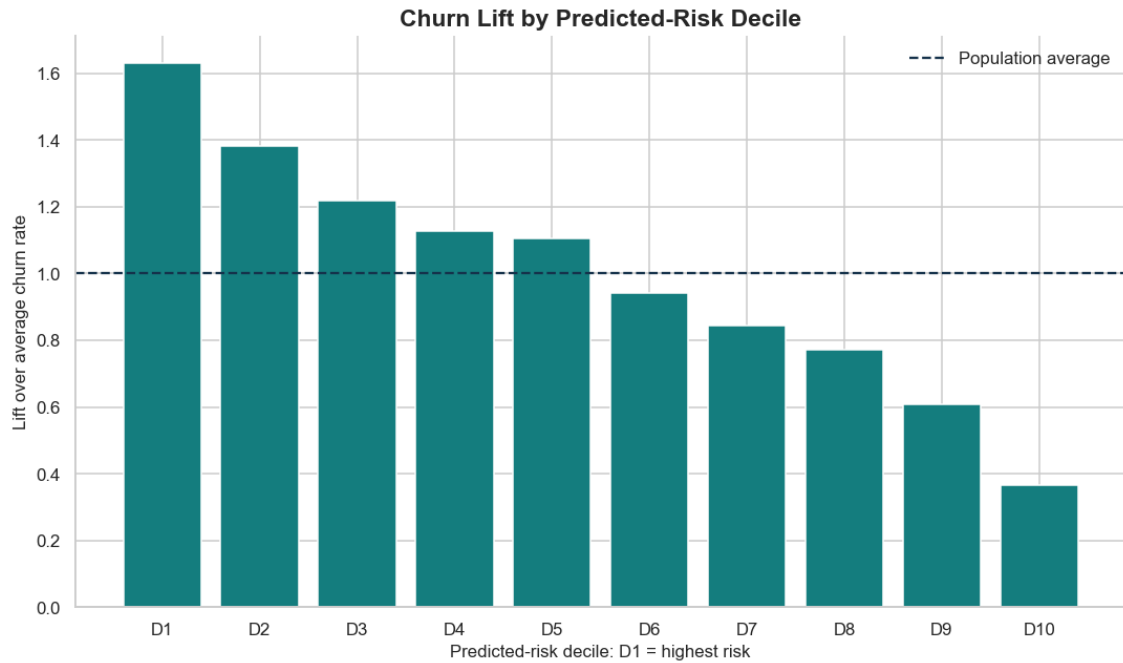


Figure 5. Lift by predicted-risk decile; D1 contains the highest-risk customers.

The risk bands show useful monotonic separation. Observed churn increased from 24.77% in the low-risk band to 79.41% in the critical band, and predicted probabilities tracked observed rates closely in these broad groups. This is encouraging for prioritisation, but formal reliability curves, Brier score and log loss in the enhanced notebook provide the more rigorous calibration assessment.

The highest-risk decile contained 2,000 customers, of whom 1,617 churned. Its 80.85% observed churn rate represented 1.63 times the population rate and captured 16.31% of all churners by contacting only 10% of customers. The top four deciles captured 53.62% of churners, providing a direct connection between model ranking and campaign capacity.

4.4 Feature Importance

Figure 6 ranks the variables contributing most strongly to the LightGBM decision rules.

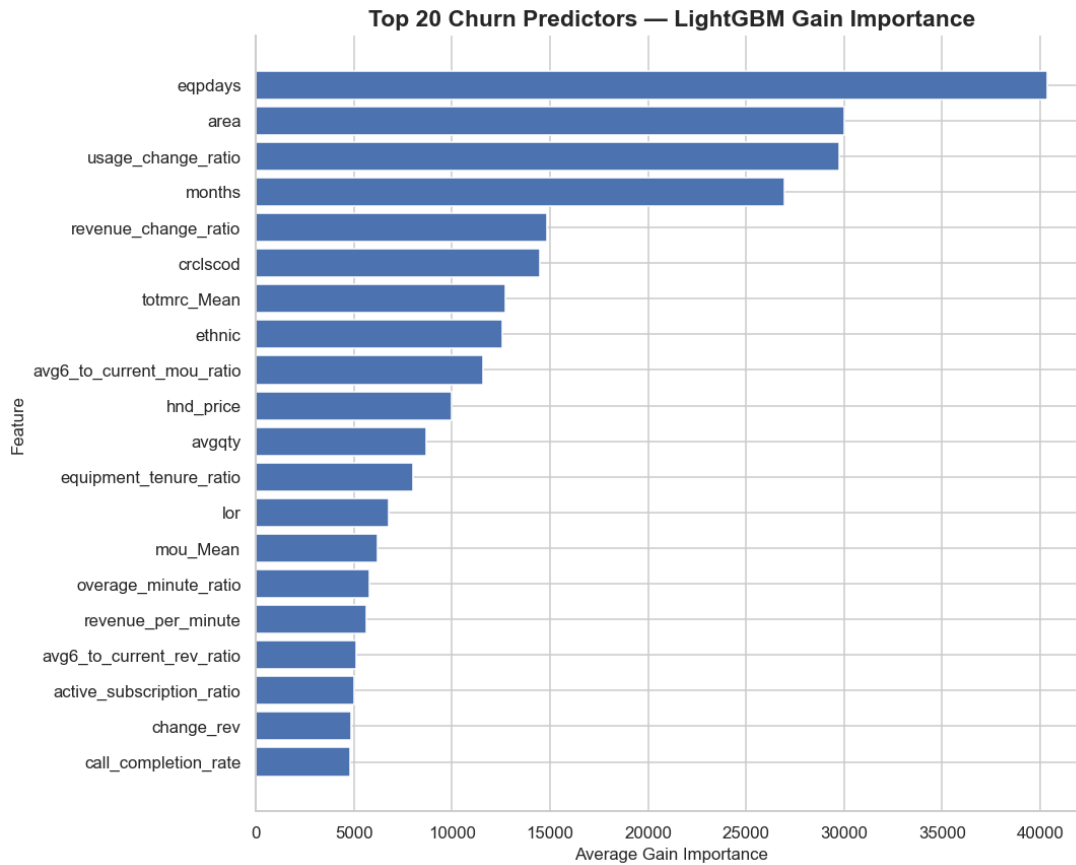


Figure 6. Top 20 churn predictors measured by average LightGBM gain importance.

The retained LightGBM gain analysis ranks equipment age, tenure, customer geography, revenue and behavioural variables among influential predictors. Gain importance shows where trees obtained loss reduction, but not whether a value raises or lowers risk. The enhanced notebook therefore adds signed Logistic Regression coefficients, Random Forest importance and weighted SHAP summaries, including a local explanation for the highest-risk sampled customer.

4.5 Churner Segmentation

Figure 7 supports the three-cluster choice using the elbow pattern alongside interpretability and silhouette evidence.

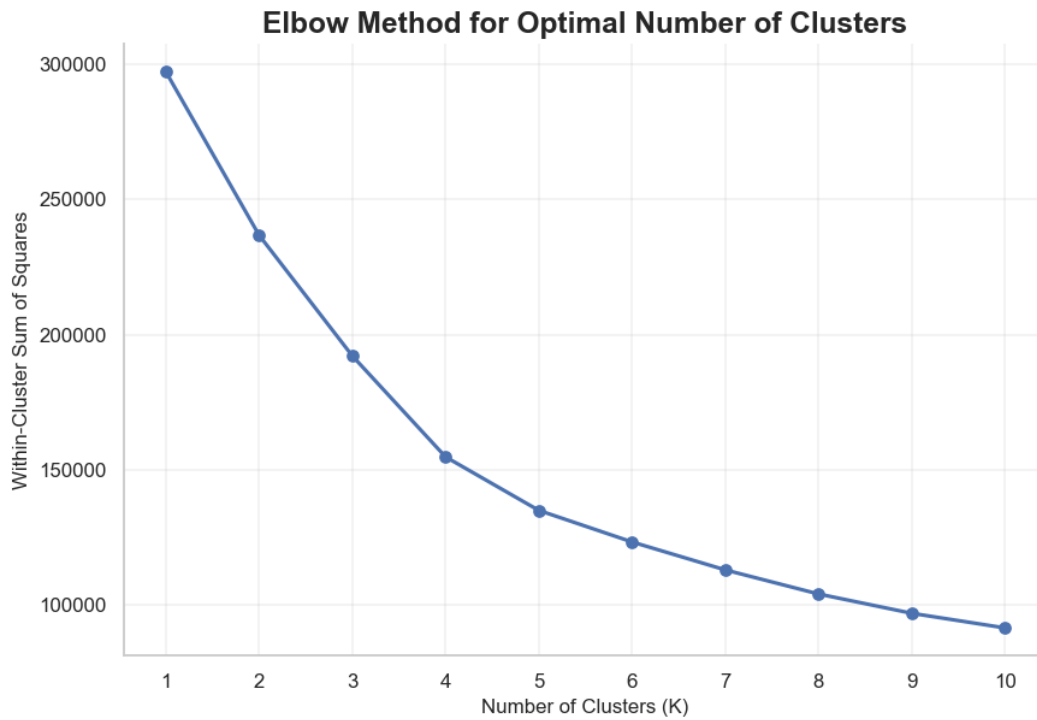


Figure 7. Elbow analysis used to support the choice of three churner clusters.

The clustering analysis compared $K = 2-8$ using inertia, silhouette, Calinski–Harabasz and Davies–Bouldin criteria, together with stability checks across random initialisations. A three-cluster solution was selected because it offered a practical balance between internal validity, segment size and managerial interpretability. This avoids choosing K solely from the elbow curve and strengthens the justification for the final segmentation.

Table 7 reports the original-scale profiles used to assign segment labels.

Table 7. Profiles of churned-customer clusters

Segment	Customers	Revenue	Usage	Usage change	Revenue change	Tenure	Equipment age
Low Usage Churners	28,048	45.62	356.28	-23.92	-1.48	14.45	318.91
High Value Churners	7,887	122.15	1,302.15	-30.37	2.09	17.18	259.29
Long Tenure Churners	13,627	43.64	249.44	-15.15	-0.41	29.55	725.04

Table 7 identifies three commercially interpretable profiles. Low Usage Churners comprised 28,048 customers (56.59%) with revenue 45.62 and usage 356.28. High Value Churners comprised 7,887 (15.91%) with revenue 122.15 and usage 1,302.15. Long Tenure Churners comprised 13,627 (27.49%) with tenure 29.55 months and equipment age 725.04 days. Their uneven sizes mean that campaign capacity and expected value should accompany cluster membership.

Figure 8 visualises separation and overlap among the churner segments in two dimensions.

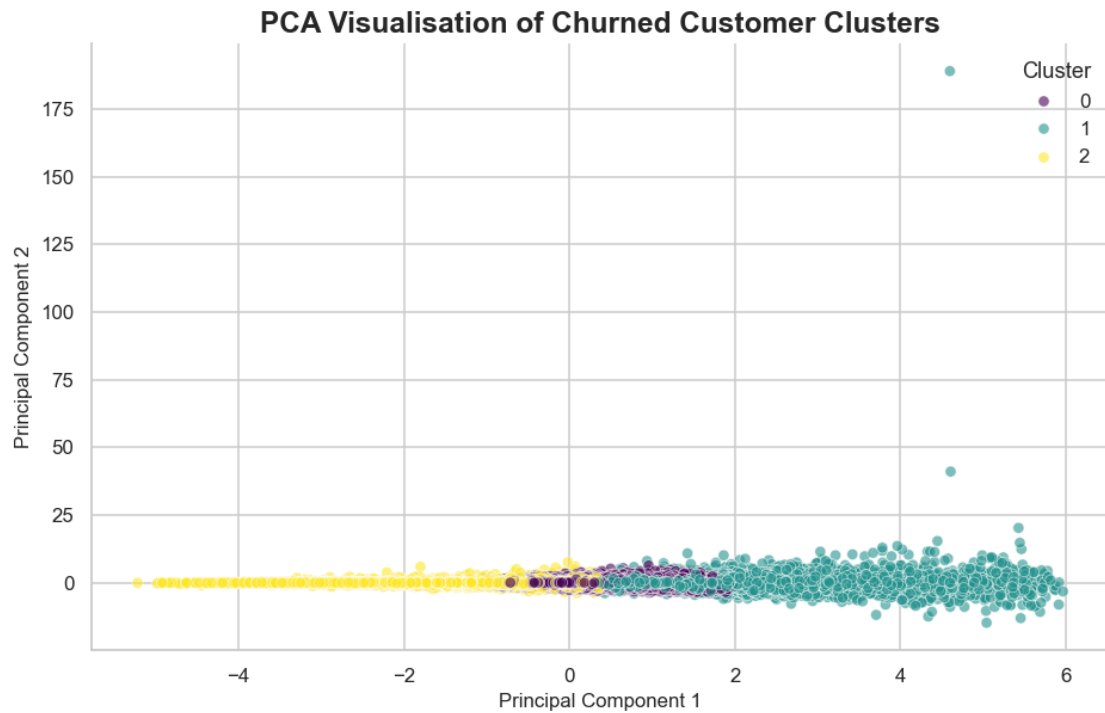


Figure 8. Two-dimensional PCA visualisation of the three churner clusters.

The PCA view and silhouette score of 0.2963 indicate useful but overlapping structure. The profiles are therefore managerial summaries rather than definitive natural classes. Cluster 0 is the largest low-usage group; Cluster 1 is smaller but commercially valuable; and Cluster 2 combines long tenure with ageing equipment. The enhanced stability analysis is essential before these labels are operationalised.

4.6 Advanced Validation, Explainability and Governance Evidence

Table 8 summarises the robustness, probability-quality, explainability, fairness and reproducibility controls incorporated into the analytical workflow. These controls ensure that model selection is based on stable evidence, that predicted probabilities are interpreted cautiously, and that the final implementation can be audited and reproduced from exported tables and the saved model bundle.

Table 8. Enhanced validation, explainability and governance controls

Control	Implementation	Decision value	Reporting rule
Robustness	Five-fold stratified CV; mean and SD across seven metrics.	Tests sensitivity to one split.	Compare stability, not only the highest mean.
Calibration	Reliability curves, Brier score and log loss.	Tests whether probabilities match outcomes.	Do not call risk bands calibrated without evidence.
Uncertainty	400 bootstrap samples for ROC-AUC, F1 and recall.	Quantifies sampling variability.	Report 95% intervals with point estimates.

Control	Implementation	Decision value	Reporting rule
Paired comparison	Bootstrap AUC difference and exact McNemar test.	Tests whether the selected model consistently improves.	Avoid claiming superiority from tiny score gaps alone.
Explainability	LR coefficients, RF importance, LightGBM gain and SHAP.	Combines direction, ranking and local explanation.	Treat contributions as associational, not causal.
Fairness	Group recall, FPR, FNR, selection rate and mean risk.	Surfaces uneven error burdens.	Investigate gaps before deployment; do not optimise on protected status.
Reproducibility	Fixed seeds, overlap assertions, exports and model bundle.	Supports audit and demonstration.	Final quality checks must all pass.

These controls strengthen all five marking dimensions. Cross-validation and uncertainty support critical model comparison; calibration and cost-sensitive thresholding connect probabilities to decisions; SHAP and coefficients improve interpretability; fairness diagnostics address professional responsibility; and automated exports make the demonstration reproducible. Their value depends on executing the notebook with the supplied CSV and reporting the generated outputs without manual substitution.

5. Critical Discussion

The ROC-AUC of 0.7019 indicates moderate discrimination, which is credible where competitor offers, service incidents and customer intention are unobserved. Accuracy of 64.55% should therefore support prioritisation and human review rather than autonomous decisions.

A major strength is protected evidence separation: raw data are reloaded, preprocessing is fitted within development data, validation controls tuning and thresholds, and the final model is evaluated once on test data. Cross-validation, overlap assertions and fixed seeds reduce optimistic bias.

The principal limitations arise from the data and study design. The dataset is cross-sectional, contains substantial missingness in optional demographic variables and omits competitor activity, contract milestones, complaint text and local network incidents. These omissions constrain predictive performance and prevent causal interpretation. Future work should use time-ordered validation, richer operational data and external evaluation on a later customer cohort.

The threshold results show that 0.50 has no universal economic meaning. Lower thresholds capture more churners but create more false alerts. The illustrative 5:1 cost ratio should be replaced by customer value, intervention cost, treatment effect and capacity estimates.

Equipment age, tenure and behavioural change are plausible signals, while area, credit class and ethnicity require scrutiny for socioeconomic proxy effects. SHAP adds direction but not causality or legal justification.

K-means adds context but simplifies continuous behaviour. The 0.2963 silhouette score confirms overlap; composite validation and stability tests improve rigour, but segment value must be demonstrated experimentally.

Reproducibility is supported by fixed seeds, explicit grids, exported tables, metadata, a model bundle and final quality checks. Production additionally requires temporal validation, drift monitoring, secure access, recalibration and fallback procedures.

Lift is the strongest capacity-planning evidence: the top decile reaches 1.63 times random targeting. It does not prove an offer will work; treatment-control experiments are needed to estimate uplift.

Multi-method explanation is stronger than gain alone: coefficients provide direction, forests and boosting rank contribution, and SHAP allocates local effects. Agreement increases confidence; disagreement requires investigation.

The retained clusters are useful narratives, not definitive natural groups. Stability, interpretability and response to differentiated actions matter more than a visually neat partition.

6. Business Recommendations

Deploy the selected model as a ranked retention queue, not a churn verdict. Prioritise critical-risk and top-decile customers, and show staff the score, risk band, value estimate, explanation and proposed action.

Differentiate actions by segment: human-led recovery for high-value churners, low-cost engagement and plan-fit review for low-usage churners, and loyalty or device-upgrade support for long-tenure churners. Test every treatment.

Use low thresholds for inexpensive communication. Costly incentives should require high risk and sufficient expected value. Replace the illustrative 5:1 error ratio with finance-approved values and sensitivity analysis.

Operationalise the workflow through a version-controlled pipeline that recreates preprocessing, model fitting, threshold selection, risk scoring and exports from a restarted environment. Model promotion should require consistent held-out ranking, calibration, uncertainty, fit-time and interpretability evidence rather than a small improvement in one metric.

Do not use ethnicity to determine treatment without demonstrable lawfulness, necessity and proportionality. Audit subgroup errors and proxy variables, and keep protected attributes for controlled auditing rather than frontline decisions [11], [12].

Use randomised treatment and control groups within risk and value strata. Measure incremental churn reduction, cost per retained customer and net value, not response rate alone.

Refresh scores on schedule and monitor drift, missingness, calibration, lift, subgroup errors and intervention saturation. Suspend or retrain when agreed tolerances are breached.

Prioritise by expected value: churn probability $\times$ customer value $\times$ treatment success, minus contact and incentive cost.

Maintain feedback linking score, threshold, segment, explanation, intervention, cost and outcome to support recalibration and future uplift modelling.

7. Conclusion

This study combines comparative classification, churning segmentation and operational decision support. Logistic Regression, Decision Tree, Random Forest, two LightGBM models and a weighted ensemble are evaluated under leakage-safe preprocessing, validation-only selection, five-fold robustness checks, calibration analysis, uncertainty estimation, statistical comparison, explainability, fairness auditing and automated exports.

The results show that Random Forest is the strongest conventional classifier and that the weighted LightGBM ensemble provides the best overall discrimination. Threshold analysis reveals a clear recall–precision trade-off, risk bands are monotonically ordered, the top decile achieves 1.63 lift and K-means identifies three commercially interpretable but overlapping churning profiles. The defensible conclusion is not that churn can be predicted with certainty, but that the workflow provides a rigorous basis for prioritising customers, selecting interventions and monitoring model performance responsibly.

References

- [1] A. Barsotti et al., “A decade of churn prediction techniques in the TelCo domain: A survey,” *SN Computer Science*, vol. 5, art. 404, 2024, doi: 10.1007/s42979-024-02722-7.
- [2] M. Imani and H. R. Arabnia, “Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: A comparative analysis,” *Technologies*, vol. 11, no. 6, art. 167, 2023, doi: 10.3390/technologies11060167.
- [3] W. H. Khoh, Y. H. Pang, S. Y. Ooi, L.-Y.-K. Wang and Q. W. Poh, “Predictive churn modeling for sustainable business in the telecommunication industry: Optimized weighted ensemble machine learning,” *Sustainability*, vol. 15, no. 11, art. 8631, 2023, doi: 10.3390/su15118631.
- [4] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems* 30, 2017, pp. 3146–3154.
- [5] F. E. Usman-Hamza et al., “Intelligent decision forest models for customer churn prediction,” *Applied Sciences*, vol. 12, no. 16, art. 8270, 2022, doi: 10.3390/app12168270.
- [6] S. Ouf, K. T. Mahmoud and M. A. Abdel-Fattah, “A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry,” *Journal of Big Data*, vol. 11, art. 70, 2024, doi: 10.1186/s40537-024-00922-9.
- [7] T. Xu, Y. Ma and K. Kim, “Telecom churn prediction system based on ensemble learning using feature grouping,” *Applied Sciences*, vol. 11, no. 11, art. 4742, 2021, doi: 10.3390/app11114742.
- [8] T. Zhang, S. Moro and R. F. Ramos, “A data-driven approach to improve customer churn prediction based on telecom customer segmentation,” *Future Internet*, vol. 14, no. 3, art. 94, 2022, doi: 10.3390/fi14030094.
- [9] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLOS ONE*, vol. 10, no. 3, e0118432, 2015, doi: 10.1371/journal.pone.0118432.
- [10] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.
- [11] Information Commissioner’s Office, “Special category data,” UK GDPR guidance, accessed 26 July 2026.
- [12] Information Commissioner’s Office, “What about fairness, bias and discrimination?” Guidance on AI and data protection, accessed 26 July 2026.
- [13] “Telecom_customer churn.csv,” dataset supplied for ACCA7015 assessment, 2026.
- [14] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems* 30, 2017, pp. 4765–4774.
- [15] G. W. Brier, “Verification of forecasts expressed in terms of probability,” *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
- [16] B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975, doi: 10.1016/0005-2795(75)90109-9.

[17] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.

[18] Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika*, vol. 12, pp. 153–157, 1947, doi: 10.1007/BF02295996.

Appendix A. Implementation Evidence

A.1 Protected stratified split and independent raw-data reload

```
model_df = pd.read_csv(DATA_FILE)
all_indices = np.asarray(model_df.index)
train_indices, temporary_indices = train_test_split(
    all_indices, test_size=0.40, stratify=y, random_state=42
)
validation_indices, test_indices = train_test_split(
    temporary_indices, test_size=0.50, stratify=y.loc[temporary_indices], random_state=42
)
```

A.2 Validation selection, refitting and weighted ensemble

The implementation notebook contains the complete classification, validation, threshold, calibration, uncertainty, explainability, fairness and clustering workflow. All transformations are fitted within development data, and the final test partition is reserved for one independent evaluation after model and threshold decisions are fixed.

```
ensemble_probability = (
    0.55 * model_a_probability +
    0.45 * model_b_probability
)
final_prediction = (ensemble_probability >= 0.50).astype(int)
```

A.3 Multi-criterion churner clustering

```
cluster_features = ["rev_Mean", "mou_Mean", "change_mou",
                    "change_rev", "months", "eqpdays"]
X_cluster_scaled = RobustScaler().fit_transform(X_cluster_imputed)
for k in range(2, 9):
    evaluate_silhouette, Calinski-Harabasz and Davies-Bouldin
SELECTED_K = composite_internal_validation(cluster_validation)
```

A.4 Leakage-safe conventional preprocessing

```
benchmark_preprocessor = ColumnTransformer([
    ("numeric", median_imputation_and_scaling, numeric_features),
    ("categorical", frequent_imputation_and_one_hot, categorical_features)
])
X_train_encoded = benchmark_preprocessor.fit_transform(X_train)
X_validation_encoded = benchmark_preprocessor.transform(X_validation)
X_test_encoded = benchmark_preprocessor.transform(X_test)
```