



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

Programme: MSc Data and Artificial Intelligence

Module: L7 - Data Mining & Visualisation

Module code: ACCA7015

Assessment Component 2

Report Title: Exploratory Data Analysis – Report and Demonstration

Lecturer: Dr Nough Elmitwally

Student Name: Darpanbhai Raiyani

Student ID: 2534009

TABLE OF CONTENTS

[Abstract](#)

[1. Introduction](#)

[2. Brief Literature Review on Predictive Analytics](#)

[3. Modelling Techniques](#)

[4. Dataset Description](#)

[5. Model Implementation and Evaluation](#)

[6. Results Analysis](#)

[7. Discussion of Findings](#)

[8. Business Recommendations](#)

[9. Conclusion](#)

[References](#)

List of Tables

Table 1. Selected predictor groups in the supplied dataset.

Table 2. Held-out model comparison (positive class = churn).

Table 3. K-means profiles among customers with observed churn.

List of Figures

Figure 1. Executed notebook code: training-only SMOTE and standardisation.

Figure 2. Executed notebook code and held-out output for Logistic Regression and Random Forest.

Figure 3. Held-out accuracy for the four classifiers.

Figure 4. Standardised K-means centroids among customers who churned.

Abstract

This report applies predictive analytics to telecom churn using the supplied executed Python/Colab workflow, the provided 100,000-record top-ten-feature dataset, and four supervised classifiers: Logistic Regression (LR), Random Forest (RF), support vector machine (SVM) and k-nearest neighbours (k-NN). It also uses K-means to segment customers with observed churn, meeting the brief's requirement for complementary pattern discovery. The data are almost class-balanced (49.6% churn). The workflow uses a stratified 80/20 split, training-only SMOTE and StandardScaler. RF provides the strongest held-out overall result: 58.1% accuracy, 57.9% churn precision, 56.5% churn recall and 0.613 ROC-AUC. SVM captures more churners (60.6% recall) but makes more false-positive alerts. K-means identifies two churner profiles: a large lower-activity, older-equipment group (80.6%) and a smaller high-use, higher-charge group (19.4%). The results support risk-ranking and differentiated retention experiments, not automatic intervention. The model performance is modest because the supplied features omit several likely drivers of churn, such as service interactions, complaints, contract status, network experience and competitor offers. The objective is not simply to maximise a technical metric. It is to assess whether the available behavioural and billing proxies can support a defensible retention workflow, where the cost of contacting a customer is balanced against the value of retaining them. The report therefore distinguishes discrimination (how well a model separates churners from retained customers) from deployment value (whether its scores improve a targeted intervention). All reported classification results are held-out test results, rather than resubstitution results, and the clustering analysis is descriptive rather than a causal explanation of churn.

Introduction

Customer churn is the loss of a subscriber from a service provider. In telecoms, predictive analytics seeks to identify customers at elevated risk before they leave so that retention resources can be targeted. Recent reviews emphasise that churn modelling is an established business-analytics use case, while model selection, data quality, class imbalance and interpretation remain important practical challenges [13], [6]. The relevant decision is not simply whether a model classifies correctly: it is whether its alerts can guide an intervention with a greater expected retained value than campaign cost. This is a consequential distinction in a subscription market. A churn model that merely labels historical outcomes is not yet a retention strategy: managers must decide whom to contact, what action to offer, and how to evaluate whether the action changed behaviour. Predictive analytics contributes by producing an ordered risk signal, while customer operations must convert that signal into feasible, proportionate interventions. The practical question in this report is therefore whether relatively limited customer records can rank risk better than a simple rule based on the overall churn rate.

This report follows the assessment specification. It evaluates the four lecturer-specified classifiers—LR, RF, SVM and k-NN—using the same data preparation logic as the supplied executed notebook. The notebook already implements cleaning, median imputation, a stratified split, SMOTE, standardisation and LR/RF/SVM. A k-NN model was added to the same split and pre-processing design because it is included in the required model set. K-means clustering is applied to observed churners, rather than replacing classification, because the written brief asks for additional patterns within Churn = True customers. The report answers three applied questions: which classifier provides the best test-set trade-off; what profiles are visible among churners; and what retention actions are justified by the evidence? The analysis deliberately keeps the implementation close to the submitted workflow. This makes the report auditable: preprocessing decisions, the train/test split, the oversampling step and the main classifier settings can be traced to the executed code rather than reconstructed retrospectively. It also places a clear boundary around interpretation. The supplied file is a reduced, selected-feature dataset, not a complete operational data mart. The findings consequently evaluate the predictive content of those ten fields, not the maximum performance that a telecom operator could achieve with complaints, contract history, service-quality and interaction data.

Brief Literature Review on Predictive Analytics

Recent Q1 journal literature does not support a universal “best” churn algorithm. Manzoor et al. review machine-learning approaches to churn and recommend that practitioners match methods and evaluation to the operational setting [13]. In telecom data, behaviour over time is especially important: Alboukaey et al. find that daily dynamic representations can improve the timeliness and accuracy of prediction relative to monthly summaries [1]. Likewise, Vo et al. show that unstructured call-log information can add predictive value, whereas the present data contain only aggregated usage variables [19]. This distinction is important when interpreting the model results: weak discrimination may reflect omitted information rather than an inadequate classifier. The recent literature is particularly useful in correcting two common assumptions. First, a higher headline accuracy does not automatically identify the most valuable model; a churn programme needs a suitable balance of precision, recall, discrimination and intervention cost. Rainio et al. [16] emphasise that metric selection should reflect the decision problem rather than routine reporting. Second, comparisons are sensitive to the data-generating context. Studies using event histories, call-detail records or rich service data address a different problem from this report's cross-sectional ten-variable file. Consequently, published results that appear much stronger should not be treated as a direct benchmark for the present experiment.

Ensemble approaches are attractive because they can represent interactions and non-linear effects. RF combines trees trained on different samples and feature subsets; recent reviews characterise ensemble learning as a flexible approach but stress the need for careful validation and interpretable reporting [5], [9]. In churn applications, black-box models can achieve strong results while remaining difficult to translate into an action plan. Poudel et al. therefore argue for explanation of telecom tabular churn models [15], while De Caigny et al. show the value of segmented interpretability for hybrid black-box classification [4]. This report consequently treats feature patterns and K-means profiles as decision-support signals, not as causal evidence. Recent telecom studies illustrate this conditional performance. Alboukaey et al. [1] use dynamic behavioural information, while Vo et al. [19] exploit unstructured call-log data; both forms of information can encode changes that precede churn. Poudel et al. [15] also show why post-hoc explanation matters in tabular churn modelling: a model may be statistically acceptable yet difficult for an operational team to trust if the drivers cannot be communicated. These studies motivate the use of RF here, not as an automatic winner, but as a flexible benchmark that can represent interactions and provide relative feature-importance information.

Class imbalance is common in churn data and can make accuracy misleading. Haddadi et al. compare resampling methods specifically for imbalanced churn prediction [6], while Khushi et al. demonstrate that resampling choices can change classifier performance [10]. SMOTE is used in the supplied notebook to generate synthetic minority examples, but it is only appropriate after the train/test split. Recent work continues to refine SMOTE variants for noisy or abnormal minority instances, including cluster-aware approaches [3], [21], [14], [12]. Here the classes are already close to balanced, so SMOTE is retained for consistency with the executed code rather than presented as the source of any performance gain. Oversampling requires equally careful interpretation. SMOTE constructs synthetic minority examples in training data; it should never be applied before the test split because that would allow information from the test distribution to influence training. The submitted code applies it to the training partition only, which preserves the integrity of the held-out evaluation. Yet SMOTE is not a substitute for business labels, appropriate thresholds or calibration. In this dataset, the churn class is already close to one half of observations, so its main value is consistency with the executed workflow rather than correction of a severe imbalance. This nuance is important when interpreting any incremental recall gain.

The literature also supports combining prediction with segmentation. Wu et al. propose an integrated telco framework [20], and Lemos et al. demonstrate the relevance of machine learning to churn propensity in financial services [11]. K-means is a useful scalable descriptive method for numeric customer data, but a cluster is an empirical profile, not a treatment prescription. Candidate partitions should be assessed before use; here, silhouette

analysis is used to compare candidate solutions. Segmentation adds a different type of evidence. A classifier estimates relative likelihood under patterns represented in the training data; K-means groups observations that are similar to one another after standardisation. It does not prove that membership causes churn, nor does it estimate treatment effect. Used together, however, prediction can prioritise customers and segmentation can make the proposed contact strategy more intelligible. Wu et al. [20] provide a useful conceptual precedent for this combined approach, although their framework and data cannot establish the effectiveness of the interventions proposed for the present sample.

Modelling Techniques

LR is an interpretable baseline estimating a linear relationship between predictors and log-odds of churn. It is fast and stable when relationships are nearly linear but can underfit non-linear interactions and correlated predictors, especially since recent and longer-window usage measures describe overlapping behaviour. RF is the main ensemble classifier, capable of learning non-linear splits and interactions, and ranking feature importance, though importance isn't causal. SVM with an RBF kernel offers a non-linear alternative but needs scaled predictors and has less business interpretability. k-NN predicts based on nearby observations, with performance depending on k, scaling, and redundancy. RF combines many decision trees trained on perturbed data and feature subsets, reducing instability and capturing threshold effects. However, feature importance shows associations, not causality, and correlated features can distort importance. Use these models to guide investigation, not to assume variable changes will prevent churn.

RBF SVM transforms data for non-linear separation, working well on scaled, high-dimensional numeric data but with less accessible decision processes and uncertain probability scores. k-NN offers intuitive local comparisons but is affected by irrelevant variables, scale, and observation density. The search for k values uses only training data. Together, LR, RF, SVM, and k-NN compare linear, ensemble, kernel, and distance-based approaches, not just variants of the same method.

The complementary model is K-means. It is fitted only to customers who actually churned, using all ten standardised predictors. This directly addresses the assessment goal of finding patterns within Churn = True customers. Candidate solutions from k = 2 to k = 6 are compared using silhouette score, and the selected clusters are described using their unscaled centroids. The procedure follows the broad segmentation logic seen in recent churn and customer-profiling work, while recognising that richer features could reveal more actionable segments [20], [8]. K-means was selected as the complementary unsupervised technique because the assessment asks for patterns within actual churners. It assigns each observation to the nearest centroid and seeks compact within-cluster variation. Since distance is central to the algorithm, the predictors were standardised before fitting; otherwise, variables measured on larger numeric scales would dominate the clusters. The number of clusters was not chosen solely because it produced an appealing managerial story. Candidate values were assessed using the silhouette statistic, then the selected solution was profiled in the original business units. The resulting labels are descriptive summaries, not customer types that should be assumed to exist unchanged in another period.

Dataset Description

The supplied top10_churn.csv has 100,000 customer observations, ten numeric predictors, and a binary target. There are 50,438 retained customers (50.4%) and 49,562 churners (49.6%), making the outcome balanced. Predictors include equipment days (eqpdays); three-, six-month, and aggregate minutes of use (mou_Mean, avg3mou, avg6mou); recent call quantity (avg3qty); voice minutes; and mean total monthly recurring charge (totmrc_Mean). The file has no missing values, but the notebook includes placeholder replacement and median/mode imputation, useful for reintroducing the full dataset. Small percentage differences may be

operationally relevant, but validation remains crucial. The target is assumed correctly coded. Since the file contains selected features only, it cannot verify if excluded variables were available during scoring or if predictors were recorded near a retention event, risking temporal leakage. An explicit scoring date and feature-availability rules are recommended in production. The feature-importance output from the notebook's Random Forest shows 'eqpdays' as the most important predictor, followed by usage metrics; 'total recurring charge' ranks lowest among the ten variables. Descriptive stats reveal a strong right-skew in usage fields and include invalid negative minimums for 'eqpdays' and 'recurring charge', which should be flagged as data quality issues. The reduced dataset lacks the full feature set- such as customer service history, complaints, contract details, payment behaviour, network quality, or competitor data- so these ten variables are not the sole churn drivers. They mainly reflect equipment lifecycle, voice usage, and a proxy for recurring charges, offering signals of engagement but losing context and trajectory. For example, a high monthly charge may indicate a valuable plan or renewal opportunity, depending on contract type and customer history. Equipment age might suggest an upgrade or indicate a satisfied customer with an older device. These ambiguities show that model importance should be seen as a prioritisation guide, not a definitive conclusion.

Table 1. Selected predictor groups in the supplied dataset.

Feature family	Variables	Potential business interpretation
Equipment age	eqpdays	Device refresh or lifecycle proxy
Usage level/trend	mou Mean, avg6mou, avg3mou, avg3qty	Engagement and recent activity
Call profile	mou_cvce_Mean, mou_peav_Mean, mou_opkv_Mean, mou_rvce_Mean	Pattern of voice use
Recurring charge	totmrc Mean	Subscription-value proxy
Target	churn	0 = retained; 1 = churned

Model Implementation and Evaluation

The code drops Customer_ID, converts churn to numeric, replaces missing data, imputes numeric fields with medians and categorical with modes, and selects ten features. The dataset was split with train_test_split (test_size=0.20, stratify=churn, random_state=42), yielding 80,000 training and 20,000 test cases, with 10,088 retained and 9,912 churned in the test set. SMOTE was fitted only on the training set to balance classes to 40,350 each. StandardScaler was fitted on training data and applied to test data to prevent leakage, important for SVM and k-NN. The evaluation treats the test set as untouched for final predictions, as dataset size can make small errors seem persuasive if data leaks occur during preprocessing. In production, transformations are stored and reapplied to new data. Metrics used include accuracy, precision, recall, F1, and ROC-AUC, which together provide a comprehensive performance view, though none alone answer the business question.

The notebook trains LR, RF with 100 trees and an RBF SVM. To complete the lecturer-specified set, k-NN was evaluated on the same transformed data using distance weighting. A 3-fold stratified search on the training split across k = 3, 5, 7, 9, 11 and 15 selected k = 15 using churn F1; the untouched test set was used only for final reporting. The evaluation reports accuracy, churn precision, churn recall, churn F1 and ROC-AUC. This is more defensible than accuracy alone, even with the near-balanced outcome, because the business cost of failing to flag a potential leaver can differ from the cost of contacting a customer who would have remained. Hyperparameter comparison must be understood as model selection, not proof of superiority. The supplied notebook's LR, RF and SVM settings are retained to respect the executed analysis. For k-NN, standardisation is essential because variables with wider scales otherwise dominate Euclidean distances. The chosen k reflects the best cross-validated training result among the tested values, after which it is evaluated once on the held-out set. No model is tuned using test-set outcomes. A more extensive study could use nested cross-validation, repeated temporal splits and probability calibration; those would improve reliability but would be additional analysis beyond the submitted code and fixed dataset.

Figure 1. Executed notebook code: training-only SMOTE and standardisation.

```
print("y_train distribution:", np.unique(y_train, return_counts=True))

y_train distribution: (array([0, 1]), array([40350, 39650]))
```

◆ STEP 11 — SMOTE

```
if len(np.unique(y_train)) > 1:
    smote = SMOTE(random_state=42)
    X_train, y_train = smote.fit_resample(X_train, y_train)
    print("After SMOTE:", np.bincount(y_train))
else:
    print("SMOTE skipped (only one class)")

After SMOTE: [40350 40350]
```

◆ STEP 12 — Scaling

```
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Figure 2. Executed notebook code and held-out output for Logistic Regression and Random Forest.

✓✓ MODELS

◆ Logistic Regression

```
lr = LogisticRegression(max_iter=1000)
lr.fit(X_train, y_train)

y_pred_lr = lr.predict(X_test)

print("\nLogistic Regression")
print("Accuracy:", accuracy_score(y_test, y_pred_lr))
print(classification_report(y_test, y_pred_lr))
```

Logistic Regression					
Accuracy: 0.5548					
	precision	recall	f1-score	support	
0	0.56	0.59	0.57	10088	
1	0.55	0.52	0.54	9912	
accuracy			0.55	20000	
macro avg	0.55	0.55	0.55	20000	
weighted avg	0.55	0.55	0.55	20000	

◆ Random Forest

```
rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)

y_pred_rf = rf.predict(X_test)

print("\nRandom Forest")
print("Accuracy:", accuracy_score(y_test, y_pred_rf))
print(classification_report(y_test, y_pred_rf))
```

Random Forest					
Accuracy: 0.58025					
	precision	recall	f1-score	support	
0	0.56	0.59	0.57	10088	
1	0.55	0.52	0.54	9912	
accuracy			0.55	20000	
macro avg	0.55	0.55	0.55	20000	
weighted avg	0.55	0.55	0.55	20000	

For K-means, all 49,562 observed churners were standardised. Twenty initialisations were used at each candidate k to reduce sensitivity to starting centroids. The silhouette scores were 0.521 ($k=2$), 0.378 ($k=3$), 0.279 ($k=4$), 0.241 ($k=5$) and 0.224 ($k=6$); $k=2$ was selected. This evaluation step avoids choosing the number of groups solely

because it appears convenient for presentation. The silhouette pattern supports a two-cluster description because separation falls as additional clusters are introduced. This does not mean there are exactly two natural kinds of churner; it means that, within these ten standardised fields, a two-centroid representation is the clearest among the tested options. Cluster sizes and centroids should therefore be monitored if the analysis is repeated on a later cohort. A segment that is stable in one historical file may fragment or shift once seasonal effects, new plans, or changes in handset availability alter the customer base.

Results Analysis

Table 2. Held-out model comparison (positive class = churn).

Model	Accuracy	Churn precision	Churn recall	Churn F1	ROC-AUC
Logistic Regression	55.5%	55.4%	51.8%	53.6%	0.579
Random Forest	58.1%	57.9%	56.5%	57.2%	0.613
SVM (RBF)	57.3%	56.5%	60.6%	58.4%	0.591
k-NN (k=15)	55.8%	55.6%	53.3%	54.4%	0.582

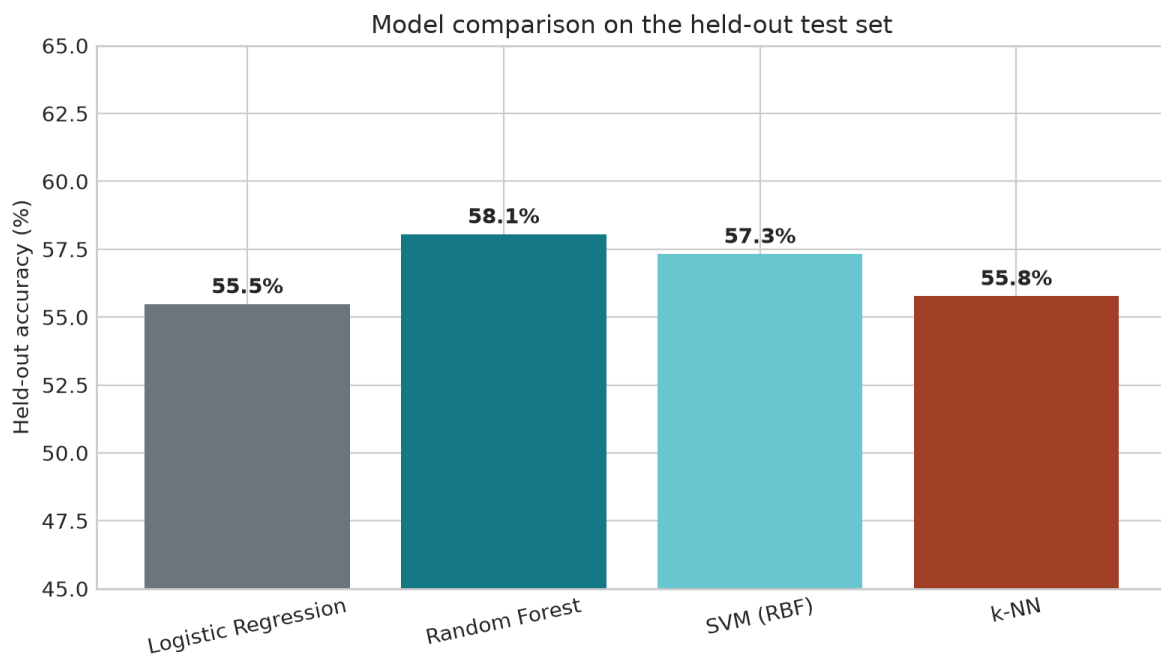


Figure 3. Held-out accuracy for the four classifiers.

RF is the best classifier, correctly identifying 5,603 churners and 6,008 retained customers, with 58.1% accuracy and highest ROC-AUC (0.613). It improves slightly over a naïve all-retained rule (50.4%). SVM captures more churners—6,003 out of 9,912—but has more false positives (4,626 vs. 4,080 for RF). LR and k-NN do not improve the trade-off. Results show that models should be compared in context, not by accuracy alone. RF improves overall accuracy by 2.6 points and churn F1 by 3.6 over LR. SVM has the highest recall but lower precision and ROC-AUC than RF, meaning it may generate more unnecessary contacts in resource-limited retention efforts.

In contrast, RF offers the stronger general ranking signal. k-NN does not add material performance in this feature space. Its weaker result is consistent with a setting where nearest customers in aggregate behavioural measures are not sufficiently similar in their churn outcome.

No result warrants an autonomous retention action. RF's ROC-AUC of 0.613 is above the 0.5 expected from a non-discriminating ranker, but it remains a modest separation. A model with this level of discrimination can still be

useful if used to concentrate a limited retention budget in the higher-risk portion of the base and if the offer itself has a measurable incremental effect. It cannot reliably identify every future churner, and its score should not be presented as a certain individual forecast. The appropriate next test is incremental business value against a random or business-as-usual contact policy.

Table 3. K-means profiles among customers with observed churn.

K-means churner segment	Customers	Share	Centroid pattern	Interpretation
Lower activity / older equipment	39,944	80.6%	455.5 equipment days; mean MOU 290.6; monthly charge 39.5	Largest lower-engagement profile
High activity / newer equipment	9,618	19.4%	277.8 equipment days; mean MOU 1,281.6; monthly charge 65.7	Smaller high-use/high-charge profile

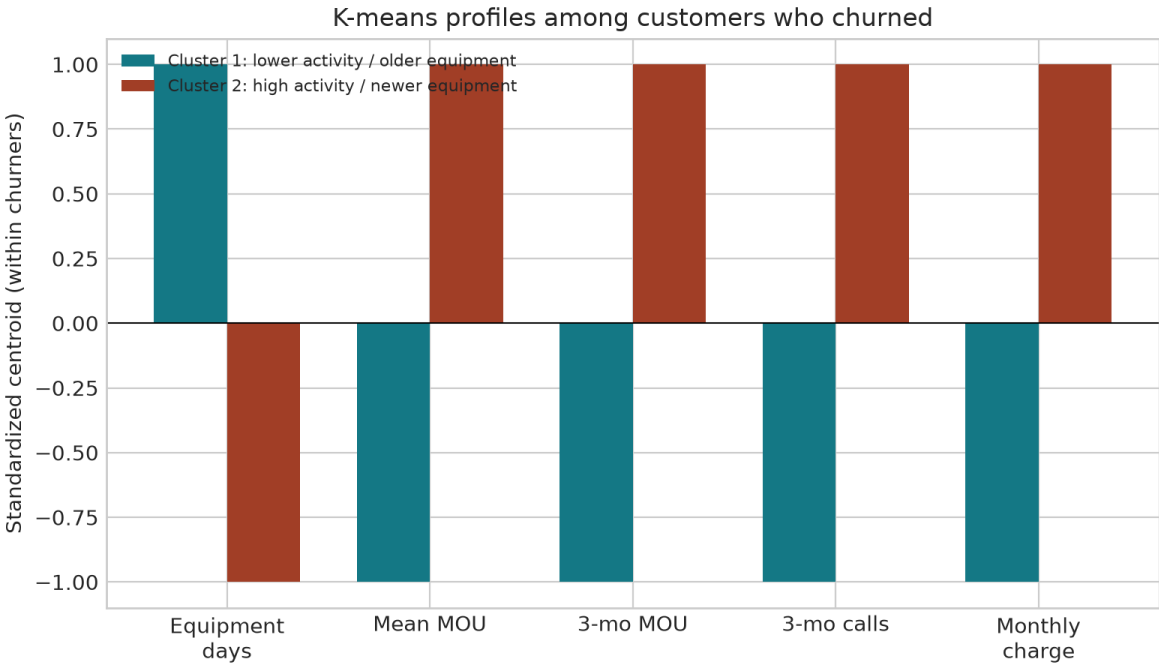


Figure 4. Standardised K-means centroids among customers who churned.

The clustering result provides an additional business distinction. Segment 1 comprises most churners and has lower use of all call measures alongside older equipment. Segment 2 has much heavier use and higher average charge despite newer equipment. The data support describing these as different profiles, but not concluding why they left. For example, higher charges may reflect higher service consumption rather than price dissatisfaction. This is why recent churn studies increasingly use explanatory modelling and segmented interpretation alongside predictive scores [15], [4]. The segment comparison also cautions against a single generic offer. The large lower-activity segment may include customers who derive little value from their current service or whose older equipment prompts a review. The smaller high-use segment represents customers with more intensive use and higher charges, where a blanket discount could sacrifice margin without addressing a service or plan problem. These are hypotheses derived from descriptive centroids, not evidence that an upgrade or service intervention will work. Each proposed treatment should therefore be evaluated experimentally within the relevant score band and segment.

Discussion of Findings

RF should be the main benchmark due to its balanced test results and feature-importance support. SVM is an alternative if recall outweighs costs, but cost-sensitive analysis is needed as data lacks offer cost and customer value info. RF's "high" accuracy is unsubstantiated; F1 of 57.2% and ROC-AUC of 0.613 show weak ranking, with operational errors. Model choice depends on goals: RF's precision suits limited calls; SVM's recall benefits high churn risk and ample capacity. LR is a transparent baseline; if it isn't outperformed, simpler models may be better. k-NN remains for comparison but isn't a priority.

The modest performance is plausible given the narrow feature space. Aggregate MOU and recurring-charge measures lack customer experience and time-to-event information. Dynamic behaviour studies suggest that temporal granularity can improve early identification [1], and call-log research shows the potential of additional unstructured or network-based signals [19]. The use of a pre-selected top-ten dataset introduces a further limitation. Because the full original feature set is unavailable, feature selection cannot be rerun inside each training fold. If the prior selection used data that later entered the test set, the reported test performance may be slightly optimistic. The performance gap between this report and stronger published churn systems should not be interpreted as an implementation failure. Many studies that report richer results use longitudinal consumption histories, social or call networks, textual interactions, contractual events or detailed service information [1], [19]. The present data contain no direct measure of dissatisfaction, payment stress, contract timing, network quality or retention offer exposure. Those omissions limit both prediction and explanation. They also create a risk that a model will associate churn with value proxies while missing the experience variables that a business can actually improve. Data enrichment is therefore more valuable than indiscriminate algorithmic complexity.

The near-balanced class split also changes the role of SMOTE. It is not wrong to use training-only SMOTE, and it reproduces the submitted code, but the imbalance is too small for it to be a central modelling claim. Future work should compare no-SMOTE, class-weighted and SMOTE models in the same cross-validation framework. Haddadi et al. and Liaw et al. both reinforce that resampling is a modelling choice to validate, not a default guarantee of better classification [6], [12]. There are further limitations. The report uses a single held-out split, so the measured differences could vary across samples. The analysis is retrospective and no time ordering is available in the reduced file; it cannot establish whether a high-risk score precedes a customer decision. The models are assessed on historical labels rather than on retained-margin outcomes, so a better F1 score does not necessarily mean a better commercial return. Fairness assessment is also limited because the supplied fields do not include a basis for evaluating protected-group performance. These limitations support a monitored pilot with explicit governance rather than immediate broad automation.

Business Recommendations

First, use RF as a risk-ranking pilot, not an automatic decision engine. Score eligible customers, then choose a contact threshold from a transparent cost-benefit calculation: expected retained margin $\times$ estimated incremental retention probability must exceed the cost of the contact and offer. Hold out a randomised control group so that the company measures incremental retention rather than simply contacting customers who would stay anyway. Recent customer-retention research highlights the importance of connecting model outputs to decision-making rather than treating prediction as the end point [18], [7]. The pilot should begin with a clearly defined eligible population and a pre-specified score threshold. Customers above that threshold can be randomly allocated within score bands to a targeted action, a low-cost generic contact, or business as usual. This design estimates incremental retention rather than merely observing that high-risk customers churn more often. Key outcomes should include churn within an agreed horizon, retained gross margin, contact and incentive cost, complaint rate, opt-out rate and any differential impact across available customer groups. The model should be retrained only after a documented monitoring period, with drift checks for changes in feature distributions and score performance.

Second, use the two profiles to design different testable interventions. For the lower-activity/older-equipment profile, test a device upgrade, plan review, and low-friction service check. For high-use/higher-charge churners, prioritise service-quality diagnosis, account-manager contact, and plan optimisation over discounts. These hypotheses come from segment descriptions; controlled testing is necessary. Conversations should be needs-based, not discount-driven. The organisation can test if device eligibility, plan fit, billing clarity, service recovery, and channel preference predict response. High-risk, unreachable customers or those with recent retention offers should follow separate rules to avoid repeated contact. This approach creates a learning loop, using recorded offers, channels, responses, and tenure as future data to determine effective actions.

Third, enrich the data before a production deployment. Add dated complaints and resolution outcomes, customer-service contacts, network-quality indicators, payment difficulties, contract/renewal status, handset model and upgrade eligibility, plus lawful competitor-offer exposure. Define a prediction date and build features only from information observed before it. Re-run feature selection within training folds, calibrate predicted probabilities, use precision-recall curves and report expected campaign profit. These changes address the principal limitations of the supplied data more directly than marginal tuning of the current algorithms. Governance should be designed into the workflow. Maintain a model card documenting the target definition, feature list, preprocessing, training period, metrics, intended use and known limitations. Log score dates and model versions, restrict access to customer-level predictions, and ensure that interaction data are handled under applicable privacy and retention requirements.

Conclusion

The supplied executed workflow provides a coherent baseline for telecom churn prediction. After evaluating the four required classifiers, RF has the best overall held-out performance, while SVM has the best churn recall. Both results remain modest, so model outputs should support prioritisation rather than determine customer treatment. K-means adds two practical churner profiles: a dominant low-activity, older-equipment group and a smaller high-use, higher-charge group. The recommended next step is a controlled, segment-specific retention pilot supported by an RF risk score, cost-sensitive threshold selection and a richer, time-aware data set. The key conclusion is that business value will come less from claiming a single winner among algorithms and more from validating interventions against better pre-churn customer data. This

References

- [1] N. Alboukaey, A. Joukhadar, and N. Ghneim, “Dynamic behavior based churn prediction in mobile telecom,” *Expert Systems with Applications*, vol. 162, Art. no. 113779, 2020, doi: 10.1016/j.eswa.2020.113779.
- [2] A. Amin, A. Adnan, and S. Anwar, “An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes,” *Applied Soft Computing*, vol. 137, Art. no. 110103, 2023, doi: 10.1016/j.asoc.2023.110103.
- [3] E. Chamseddine, N. Mansouri, M. Soui, and M. Abed, “Handling class imbalance in COVID-19 chest X-ray images classification: Using SMOTE and weighted loss,” *Applied Soft Computing*, vol. 129, Art. no. 109588, 2022, doi: 10.1016/j.asoc.2022.109588.
- [4] A. De Caigny, K. W. De Bock, and S. Verboven, “Hybrid black-box classification for customer churn prediction with segmented interpretability analysis,” *Decision Support Systems*, vol. 181, Art. no. 114217, 2024, doi: 10.1016/j.dss.2024.114217.
- [5] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, “Ensemble deep learning: A review,” *Engineering Applications of Artificial Intelligence*, vol. 115, Art. no. 105151, 2022, doi: 10.1016/j.engappai.2022.105151.
- [6] S. J. Haddadi, A. Farshidvard, F. da Silva, J. C. dos Reis, and M. da Silva Reis, “Customer churn prediction in imbalanced datasets with resampling methods: A comparative study,” *Expert Systems with Applications*, vol. 246, Art. no. 123086, 2024, doi: 10.1016/j.eswa.2023.123086.
- [7] P. Jiang, Z. Liu, L. Zhang, and J. Wang, “Hybrid model for profit-driven churn prediction based on cost minimization and return maximization,” *Expert Systems with Applications*, vol. 228, Art. no. 120354, 2023, doi: 10.1016/j.eswa.2023.120354.
- [8] M. S. Kasem, M. A. Hamada, and I. Taj-Eddin, “Customer profiling, segmentation, and sales prediction using AI in direct marketing,” *Neural Computing and Applications*, vol. 35, pp. 29273–29291, 2023, doi: 10.1007/s00521-023-09339-6.
- [9] A. A. Khan, O. Chaudhari, and R. Chandra, “A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation,” *Expert Systems with Applications*, vol. 244, Art. no. 122778, 2024, doi: 10.1016/j.eswa.2023.122778.
- [10] M. Khushi et al., “A comparative performance analysis of data resampling methods on imbalance medical data,” *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.
- [11] R. A. de Lima Lemos, T. C. Silva, and B. M. Tabak, “Propension to customer churn in a financial institution: A machine learning approach,” *Neural Computing and Applications*, vol. 34, pp. 11751–11768, 2022, doi: 10.1007/s00521-022-07067-x.
- [12] L. C. M. Liaw, S. C. Tan, P. Y. Goh, and C. P. Lim, “A histogram SMOTE-based sampling algorithm with incremental learning for imbalanced data classification,” *Information Sciences*, vol. 686, Art. no. 121193, 2025, doi: 10.1016/j.ins.2024.121193.
- [13] A. Manzoor, M. A. Qureshi, E. Kidney, and L. Longo, “A review on machine learning methods for customer churn prediction and recommendations for business practitioners,” *IEEE Access*, vol. 12, pp. 70434–70463, 2024, doi: 10.1109/ACCESS.2024.3402092.
- [14] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, “Enhancing SMOTE for imbalanced data with abnormal minority instances,” *Machine Learning with Applications*, vol. 18, Art. no. 100597, 2024, doi: 10.1016/j.mlwa.2024.100597.
- [15] S. S. Poudel, S. Pokharel, and M. Timilsina, “Explaining customer churn prediction in telecom industry using tabular machine learning models,” *Machine Learning with Applications*, vol. 17, Art. no. 100567, 2024, doi: 10.1016/j.mlwa.2024.100567.
- [16] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Scientific Reports*, vol. 14, Art. no. 6086, 2024, doi: 10.1038/s41598-024-56706-x.
- [17] S. Saha, C. Saha, M. M. Haque, M. G. R. Alam, and A. Talukder, “ChurnNet: Deep learning enhanced customer churn prediction in telecommunication industry,” *IEEE Access*, vol. 12, pp. 4471–4484, 2024, doi: 10.1109/ACCESS.2024.3349950.

- [18] A. Sikri, R. Jameel, S. M. Idrees, and H. Kaur, “Enhancing customer retention in telecom industry with machine learning driven churn prediction,” *Scientific Reports*, vol. 14, Art. no. 13097, 2024, doi: 10.1038/s41598-024-63750-0.
- [19] N. N. Y. Vo, S. Liu, X. Li, and G. Xu, “Leveraging unstructured call log data for customer churn prediction,” *Knowledge-Based Systems*, vol. 212, Art. no. 106586, 2021, doi: 10.1016/j.knosys.2020.106586.
- [20] S. Wu, W.-C. Yau, T. S. Ong, and S.-C. Chong, “Integrated churn prediction and customer segmentation framework for telco business,” *IEEE Access*, vol. 9, pp. 62118–62136, 2021, doi: 10.1109/ACCESS.2021.3073776.
- [21] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, “A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data,” *Information Sciences*, vol. 572, pp. 574–589, 2021, doi: 10.1016/j.ins.2021.02.056.