Bank Customer Prediction with Machine Learning

Fenoniaina Tolojanahary Andrianary

January 2026

Abstract

In today's competitive banking environment, effective customer retention depends on accurately identifying customers at risk of churn. This study applies supervised machine learning techniques to bank customer churn prediction, comparing Logistic Regression, k-Nearest Neighbors, and Random Forest models while addressing the challenge of class imbalance, which can lead to misleading conclusions when accuracy alone is used for evaluation. To overcome this limitation, decision-threshold optimisation using Youden's J statistic was employed to balance recall and precision. The results indicate that Random Forest achieves the highest overall predictive accuracy but fails to detect a substantial proportion of churners, whereas threshold optimisation significantly improves churn detection for Logistic Regression and k-NN by increasing recall. Although this improvement comes at the cost of reduced accuracy, the optimised models are more appropriate in churn-management contexts where the cost of missing a churner exceeds that of incorrectly targeting a loyal customer, highlighting the importance of aligning model evaluation with business objectives. Among the features considered, active membership and age emerge as key drivers of churn, with active customers significantly less likely to leave, while older customers exhibit a higher risk of churn.

1 Introduction

Customer churn poses a significant challenge for retail banks, as losing existing customers directly affects profitability and long-term sustainability. To address this, banks increasingly rely on data-driven decision-support tools that assist human decision-makers by providing objective, consistent estimates of churn risk based on customer characteristics. In this project, churn prediction is treated as a binary classification problem whose purpose is not to guess customer behaviour, but to support targeted and proactive retention actions through reliable probability estimates.

The objective of this study is to develop and evaluate supervised machine learning models that can meaningfully support churn related decision making. The analysis begins with an exploratory data analysis that investigates the structure of the dataset and highlights its key characteristics. Building on these insights, two classification models, Logistic Regression and k Nearest Neighbours, are developed and compared in terms of their predictive performance and practical relevance. Particular attention is given to model comparison and threshold optimisation, demonstrating how predicted probabilities can be transformed into actionable insights rather than simple binary outcomes.

2 Dataset and Exploratory Data Analysis

2.1 Dataset Overview

TThe dataset consists of information on 10,000 bank customers and primarily indicates whether the customer is leaving the company or not. The customer information includes a combination of categorical and numerical features, as summarised in the following table:

Table 1: Feature summary of loan applications.

Feature	Description	Data Type
credit-score	The credit score of the customer	Numerical
country	Country of residence	Categorical
gender	Gender of the customer	Binary
age	Age of the customer	Numerical
tenure	Number of year within the bank	Numerical
balance	The account balance of the customer	Numerical
products-number	Number of products from bank	Numerical
active-member	Is the customer an active member of the bank	Binary
estimated salary	Estimated salary of the customer	Numerical
credit-card	Denotes whether or not a customer has a credit card	Binary
churn	Client has left the bank during some period or not.	Binary

3 Exploratory Data Analysis (EDA)

3.1 Target variable: churn distribution

The target variable is moderately imbalanced, with about 20% churners and 80% non-churners, as shown in Figure 1. As a result, accuracy alone can be misleading, since predicting non-churn for all customers would already yield 80% accuracy. Therefore, confusion matrices are also used to evaluate model performance.

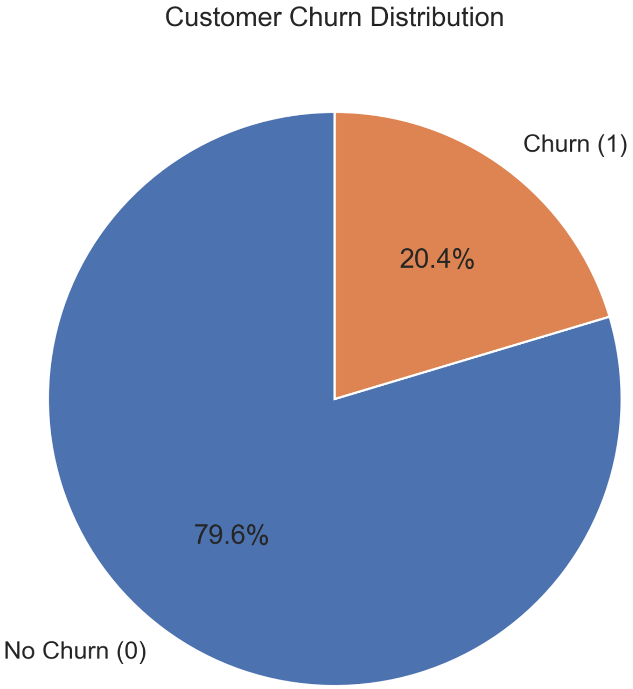


Figure 1: Churn distribution in the dataset.

3.2 Customer demographics

The customer base is mainly middle-aged, around 30–45 years, slightly male-dominated, and largely concentrated in France, with smaller and similarly sized customer groups from Spain and Germany. These demographic differences may influence churn behaviour and are therefore relevant for the classification analysis.

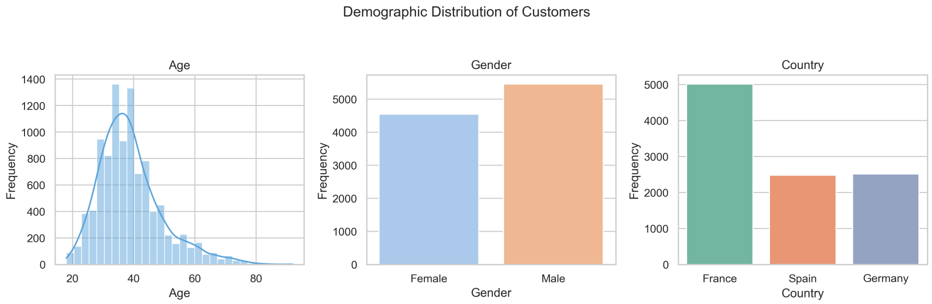


Figure 2: Customer demographic characteristics.

3.3 Financial profile

Credit scores are broadly distributed, while account balances are strongly right-skewed, with many customers holding low or zero balances. Estimated salaries are relatively uniform. This suggests that balance and credit score are likely to be more informative for churn prediction than salary.

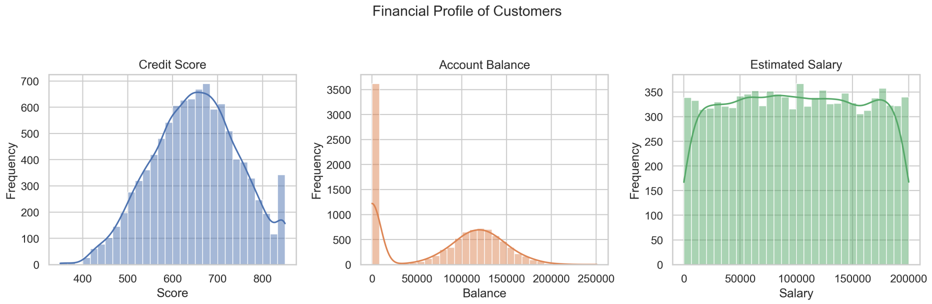


Figure 3: Distribution of financial variables.

3.4 Customer engagement

Customer tenure is fairly evenly distributed, indicating a mix of new and long-standing clients. Most customers hold one or two products, while only a small fraction hold more. Although the active member variable is balanced in terms of counts, it reflects differences in customer engagement, which may result in different churn rates between active and inactive customers and makes it a potentially important predictor.

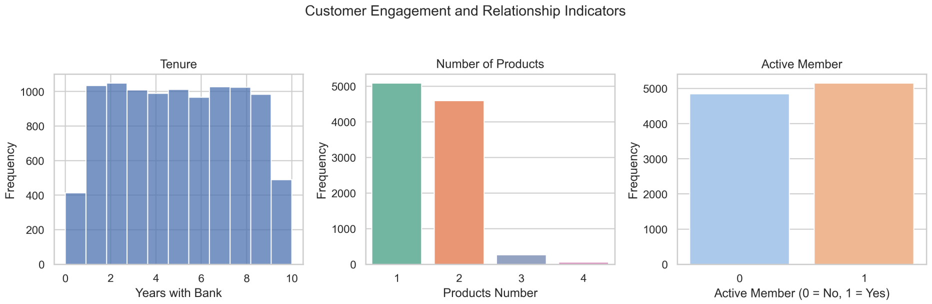
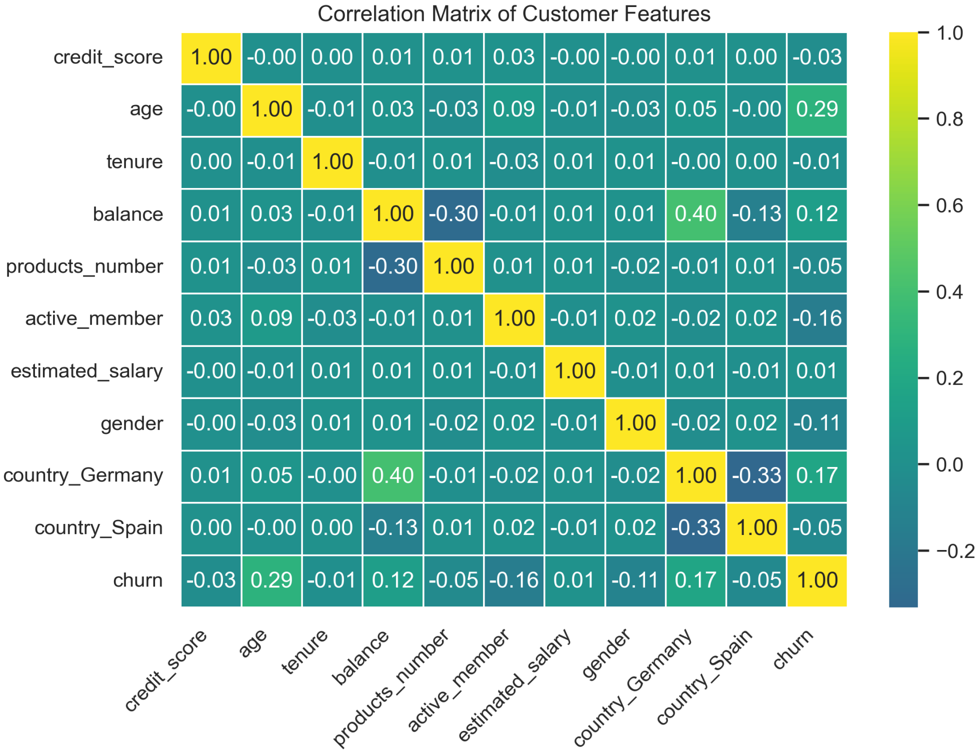


Figure 4: Customer engagement and relationship indicators.

3.5 Feature relationships and correlations

The correlation analysis shown in Figure 5 indicates that no single feature is strongly linearly correlated with churn, suggesting that churn behaviour is driven by multiple factors rather than a single dominant variable. Age exhibits the strongest positive correlation with churn (approximately 0.29), while active membership shows a moderate negative correlation (approximately -0.16). Balance and country indicators display weaker correlations, whereas estimated salary and credit score show near-zero correlation with churn. Overall, these results motivate the use of multivariate classification models, as churn cannot be explained by any single feature in isolation.



	credit_score	age	tenure	balance	products_number	active_member	estimated_salary	gender	country_Germany	country_Spain	churn
credit_score	1.00	-0.00	0.00	0.01	0.01	0.03	-0.00	-0.00	0.01	0.00	-0.03
age	-0.00	1.00	-0.01	0.03	-0.03	0.09	-0.01	-0.03	0.05	-0.00	0.29
tenure	0.00	-0.01	1.00	-0.01	0.01	-0.03	0.01	0.01	-0.00	0.00	-0.01
balance	0.01	0.03	-0.01	1.00	-0.30	-0.01	0.01	0.01	0.40	-0.13	0.12
products_number	0.01	-0.03	0.01	-0.30	1.00	0.01	0.01	-0.02	-0.01	0.01	-0.05
active_member	0.03	0.09	-0.03	-0.01	0.01	1.00	-0.01	0.02	-0.02	0.02	-0.16
estimated_salary	-0.00	-0.01	0.01	0.01	0.01	-0.01	1.00	-0.01	0.01	-0.01	0.01
gender	-0.00	-0.03	0.01	0.01	-0.02	0.02	-0.01	1.00	-0.02	0.02	-0.11
country_Germany	0.01	0.05	-0.00	0.40	-0.01	-0.02	0.01	-0.02	1.00	-0.33	0.17
country_Spain	0.00	-0.00	0.00	-0.13	0.01	0.02	-0.01	0.02	-0.33	1.00	-0.05
churn	-0.03	0.29	-0.01	0.12	-0.05	-0.16	0.01	-0.11	0.17	-0.05	1.00

Figure 5: Correlation heatmap between features and churn.

Figure 6 provides a descriptive overview of customer activity and age patterns in the dataset. *Churn Rate by Active Member Status* shows that inactive customers exhibit higher observed churn rates than active customers. *Age Distribution of Churned and Retained Customers* indicates that churned customers are more concentrated at older ages, while retained customers are more prevalent in younger age groups.

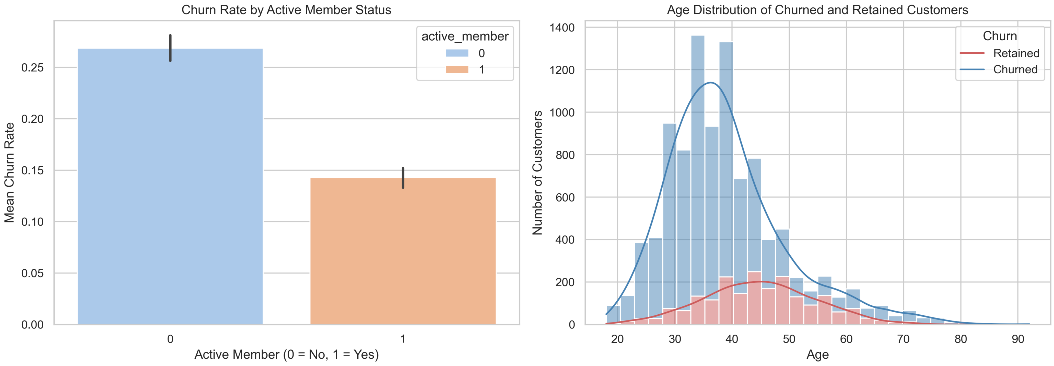


Figure 6: Customer activity status and age distribution by churn outcome.

4 Methodology

4.1 Preprocessing

Categorical variables are encoded and numerical features are standardised to ensure comparability across models, particularly for distance-based methods such as k-Nearest Neighbours. The dataset is split into training and test sets using a stratified strategy within each set to preserve the original class distribution. All preprocessing steps are applied consistently to avoid data leakage

4.2 Models

In the research, two widely applied machine learning methods are constructed: Logistic Regression and k-Nearest Neighbours(kNN).

4.2.1 Logistic Regression

Logistic Regression is a statistical model to estimate the probability of a binary outcome based on one or more predictor (or independent)variables. It is a robust baseline model for this type of prediction task

$$P(y = 1|x) = \sigma(w^\top x + b), \quad \sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

For this research, hyperparameters were selected using a grid search over the learning rate and number of iterations with five-fold cross-validation on the training set. Mean validation accuracy and its standard deviation were used to compare model performance.

To optimize the result of the logistic regression we try to find a threshold that can maximize our recall via the Youden's J statistic

4.2.2 k-Nearest Neighbours (kNN)

It is a non-parametric model that captures local, potentially non-linear decision boundaries by averaging outcomes among neighbouring observations k which is selected based on five-fold cross-validation

To optimize our knn model and to ensure a fair comparison we converts k-NN outputs into probabilities and selects a decision threshold using Youden's J (training-based), then evaluates the resulting classifier on the test set.

4.2.3 Evaluation metrics

Model performance is assessed using accuracy, precision, recall, and ROC–AUC. Given the imbalanced target distribution and our main focus one minimizing False Negative, particular attention is paid to recall and ROC–AUC, as these metrics better reflect a model's ability to identify churners.n

4.2.4 Threshold Optimization

In addition to the standard 0.5 probability threshold, an optimised threshold is derived using Youden's J statistic. This approach selects the threshold that maximises the difference between the true positive rate and the false positive rate. The optimisation is performed using training data only, ensuring a fair evaluation on the test set.

5 Findings and Results

This section constitutes the core of the study and focuses on a detailed evaluation and comparison of model performance. Emphasis is placed on how different evaluation perspectives—fixed thresholds and threshold optimisation—lead to different conclusions about model effectiveness

5.1 Feature Importance

The following table demonstrates the main driver of Churn in bank Customer

Table 2: Logistic Regression Feature Importance

Feature	Coefficient	Abs Coefficient
active_member	-1.1042	1.1042
age	0.7727	0.7727
country_Germany	0.7504	0.7504
gender	-0.5460	0.5460
balance	0.1443	0.1443
products_number	-0.1276	0.1276
credit_score	-0.0892	0.0892
tenure	-0.0514	0.0514
country_Spain	-0.0201	0.0201
estimated_salary	0.0185	0.0185
credit_card	-0.0168	0.0168

The results show that active membership, age, and country (Germany) are the main drivers of churn, with active customers being less likely to leave and older customers facing higher churn risk. In contrast, estimated salary, credit card ownership, and country (Spain) have negligible impact on churn prediction. This is consistent with the earlier correlation analysis, which indicated weak linear relationships for these variables.

5.2 Performance under the standard threshold (0.5)

Using the default probability threshold of 0.5, both Logistic Regression and k-Nearest Neighbours achieve satisfactory overall accuracy. However, class-specific performance reveals important limitations, as recall for the churn class remains relatively low, indicating that many actual churners are misclassified as non-churners. This outcome is consistent with the class imbalance identified during exploratory analysis and confirms that accuracy alone is not an adequate indicator of model quality in this setting. At this threshold, Logistic Regression exhibits more stable predictive behaviour, while kNN shows slightly higher variability in its predictions. Nevertheless, neither model is sufficiently effective at identifying churners when relying solely on the default decision rule.

5.3 Threshold optimisation using Youden's J statistic

To address the limitations observed at the default threshold, an optimised threshold is derived using Youden's J statistic, based exclusively on training data. Applying this optimised threshold to the test set leads to a marked improvement in recall for the churn class for both models.

The results demonstrate that threshold optimisation changes the classification decisions by increasing the number of FP and decreasing the number of FN. To be more specific, more customers at risk of churn are correctly identified, at the cost of an increase in false positives. From a business perspective, this trade-off is often desirable, as failing to identify a churner is typically more costly than incorrectly flagging a loyal customer.

5.4 Comparative interpretation of Logistic Regression and kNN

Under optimised thresholds, Logistic Regression improves recall while maintaining relatively stable precision. Its smoother and more consistent probability estimates across the feature space contribute to its robustness.

In contrast, k-Nearest Neighbours reacts more sharply to threshold adjustments, reflecting its dependence on local neighbourhood structure. While this sensitivity can help capture non-linear churn patterns, it may also introduce instability in sparsely populated regions of the feature space.

Overall, neither model is universally superior. The analysis demonstrates that model selection and threshold choice must be considered jointly, as a model with slightly lower raw accuracy may ultimately be more effective if its decision rule better aligns with churn management objectives.

5.5 Final model comparison using scikit-learn and Random Forest

The final comparison incorporates scikit-learn implementations of Logistic Regression, k-Nearest Neighbours, and Random Forest to evaluate their predictive performance under both standard and optimised thresholds. The following comparison table summarises the key results.

Table 3: Test-set performance comparison across models and thresholds

Model	**Accuracy (TEST)**	**Precision (TEST)**	**Recall (TEST)**
Logistic Regression (standard)	0.8041	0.5615	0.1789
Logistic Regression (Youden)	0.7341	0.4031	0.6324
k-NN (standard)	0.8301	0.6545	0.3529
k-NN (Youden)	0.7976	0.5033	0.5613
Random Forest (standard)	0.8661	0.8182	0.4412
Random Forest (Youden)	0.8571	0.8466	0.3652

Across all models, Random Forest demonstrates the strongest overall discriminative ability, as confirmed by the highest ROC–AUC (0.856) in Figure 7 and the best test accuracy under the standard threshold (0.8661). It also achieves the highest precision (0.8182), indicating that when it predicts churn, it is more likely to be correct. However, its recall remains moderate (0.4412), meaning a substantial proportion of churners are still missed.

k-Nearest Neighbours ($k = 5$) provides a balanced intermediate performance. Under the standard threshold, it attains higher accuracy (0.8301) and recall (0.3529) than Logistic Regression, while Youden optimisation significantly improves recall to 0.5613 at the cost of reduced accuracy (0.7976). This indicates that k-NN responds well to threshold adjustment when churn detection is prioritised.

Logistic Regression shows the lowest overall accuracy among the three models but benefits the most from threshold optimisation. While the standard model has very low recall (0.1789), applying the Youden threshold increases recall substantially to 0.6324 but reduces accuracy to (0.7341). This makes Logistic Regression particularly suitable in churn-sensitive scenarios where identifying churners is more important than minimising false positives. Overall, Random Forest is the preferred choice for general predictive performance, k-Nearest Neighbours offers a strong trade-off between accuracy and recall, and Logistic Regression with threshold optimisation is most effective when churn detection is the primary objective.

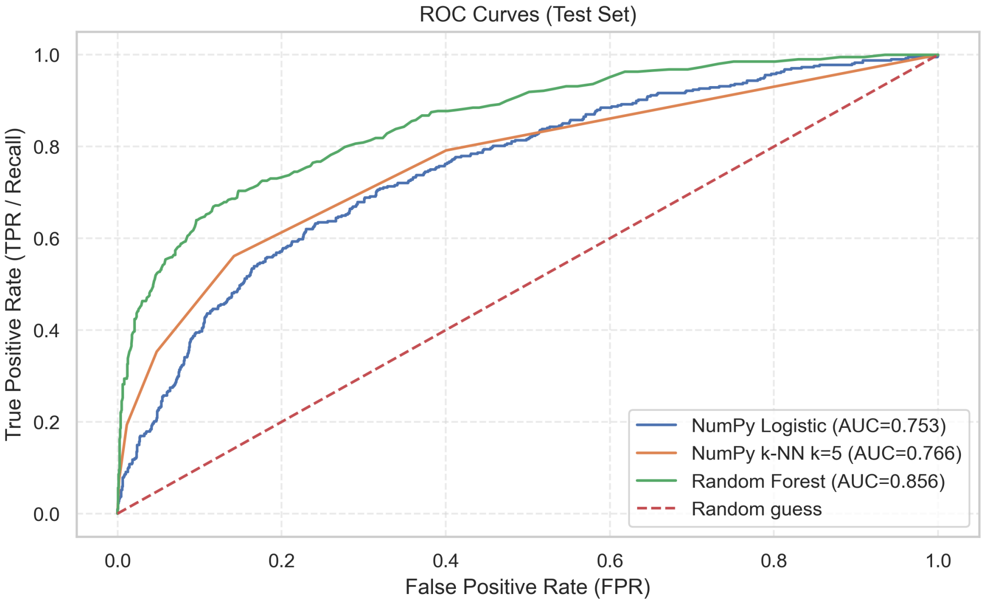


Figure 7: Final comparison of model performance across standard and optimised thresholds.

6 Discussion

The findings of this study highlight several important insights regarding model evaluation and deployment in customer churn prediction. First, reliance on default probability thresholds can significantly underestimate a model's practical value in imbalanced classification problems. Although both Logistic Regression and k-Nearest Neighbours initially appear to perform modestly in identifying churners, threshold optimisation reveals substantially improved performance, particularly in terms of recall for the churn class.

Second, the similarity of the ROC curves underscores the distinction between ranking ability and decision-making performance. While ROC–AUC is an essential metric for comparing the discriminative power of models, it does not fully capture operational effectiveness. This distinction is especially important in applied settings, where decisions must be made at a specific threshold rather than across the full range of possible cut-offs.

Third, the comparison between Logistic Regression and kNN illustrates a broader trade-off between interpretability and flexibility. Logistic Regression provides stable and interpretable probability estimates, making it attractive in regulated environments, whereas kNN offers adaptability to local, non-linear pat-

terns at the expense of consistency and robustness.
Finally, these results align with existing literature emphasising the importance of threshold selection and evaluation beyond accuracy in churn prediction. Previous studies have shown that models with comparable ROC–AUC values may differ substantially in recall once decision thresholds are adjusted to account for class imbalance. Prior work further highlights that optimising decision rules, rather than solely modifying model architectures, can lead to meaningful practical improvements in identifying high-risk customers, particularly in cost-sensitive churn management contexts. The present analysis confirms that significant gains can be achieved not only through model choice, but also through careful translation of probability outputs into decisions.

7 Conclusion

This project presents a complete churn prediction workflow grounded in exploratory data analysis, supervised learning, and rigorous performance evaluation. By comparing Logistic Regression and k-Nearest Neighbours and incorporating threshold optimisation, the study demonstrates the importance of aligning model decisions with practical objectives rather than relying solely on default settings or single performance metrics.

A key limitation of this study is the presence of class imbalance, with churners representing a minority of the dataset. Although threshold optimisation improved recall, the models were still trained on the original imbalanced data, which may limit their ability to fully capture churn patterns. Future work could incorporate resampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to balance the training data and potentially improve churn detection while maintaining better overall performance.