Telecommunication Customer Analysis: Application of Machine Learning methods into Customer churn predictions

ABSTRACT

This report investigates customer churn prediction in the telecommunications sector using marketing analytics and behavioural science approaches. The primary objective is to identify key churn predictors and develop accurate classification models to inform data-driven retention strategies. A dataset of 1,409 customer records was analysed using R programming to implement logistic regression, a logistic model with interaction terms, and decision tree algorithms. Exploratory data analysis revealed that contract type, tenure in months, and monthly charges were the most influential variables. The logistic regression model with interaction terms achieved the best performance, with an AUC of 0.881 and an F1 score of 0.825. Results confirm that long-term contracts and customer tenure reduce churn, while higher monthly charges increase it. Findings also support behavioural insights, such as commitment bias and loss aversion, in shaping customer loyalty. These results enable the design of targeted, evidence-based marketing interventions for customer retention.

TABLE OF CONTENTS

LIST OF EQUATIONS

LIST OF FIGURES

LIST OF TABLES

I. INTRODUCTION

1. The Business and the Business Problems

The dataset pertains to the **UK telecommunications sector**, a vital component of the nation's digital infrastructure, supporting both consumer and enterprise connectivity. According to Ofcom (2024), the industry generates over **£30 billion annually**, with more than **90 million mobile connections** and **27 million broadband lines**. The sector is rapidly evolving due to **digitisation**, rising data demand, and increasing consumer expectations for seamless service.

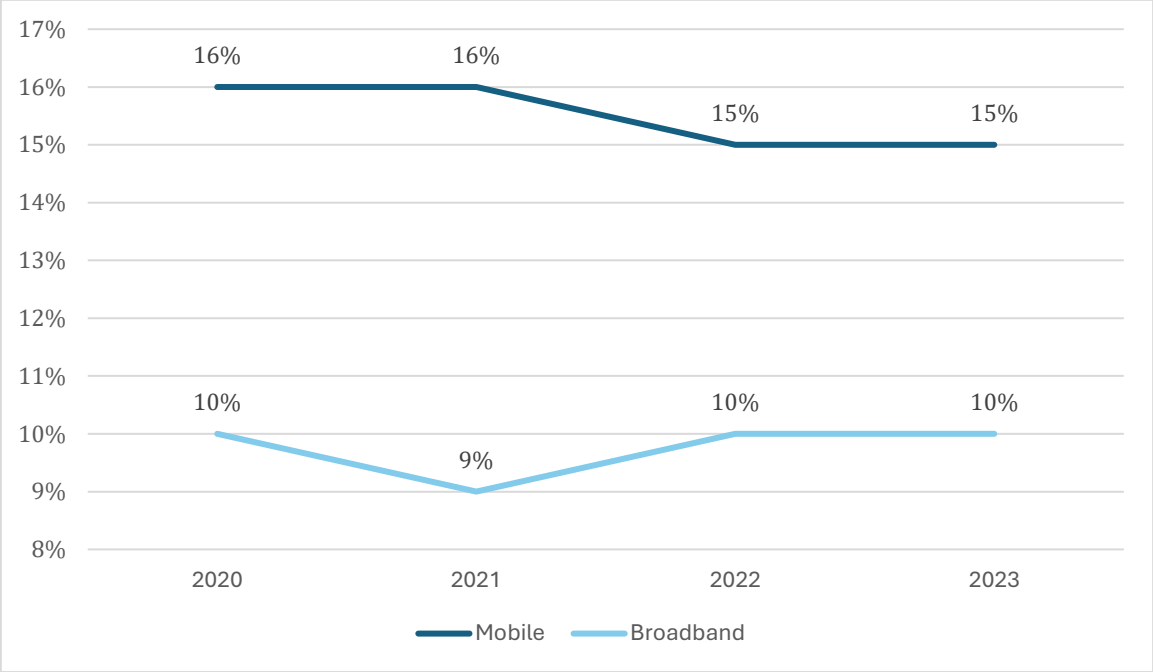

Figure 1: Switching rate across the UK telecommunication industry from 2020 to 2023 (Ofcom, 2023)

A key challenge is **customer churn**, defined as the discontinuation of service, either through cancellation or switching providers. As shown in **Figure 1**, annual churn rates reach **9–10% for broadband** and up to **16% in mobile segments** (Ofcom, 2023). This trend has been amplified by **market liberalisation, new technologies**, and expanding service options, all of which heighten competitiveness (Estrella-Ramón et al., 2013; Wu et al., 2022).

2. Marketing Analytics and Behavioural Science Context

In this dynamic environment, **marketing analytics** enables telecom firms to extract insights from complex customer data to support strategic decisions. Techniques such as **Naïve Bayes** and **Logistic Regression** facilitate **churn prediction**, allowing companies to identify at-risk segments and implement **targeted retention strategies**.

Viewed through **behavioural science**, churn is shaped by cognitive biases such as **loss aversion, perceived fairness**, and **switching costs**. For instance, long-term contract holders are influenced by **commitment bias**, while users of automated payments face less decision friction, improving retention. These insights support the development of psychologically informed marketing interventions. Overall, data-driven churn models not only improve campaign personalisation and customer loyalty but also generate **long-term business value** (Wu et al., 2022; Estrella-Ramón et al., 2013).

3. Research objectives and data analytic framework

This report adopts an integrated marketing analytics and behavioural science approach to investigate customer churn in the telecommunications industry. The primary aim is to derive actionable insights that support retention strategies, pricing decisions, and service design improvements.

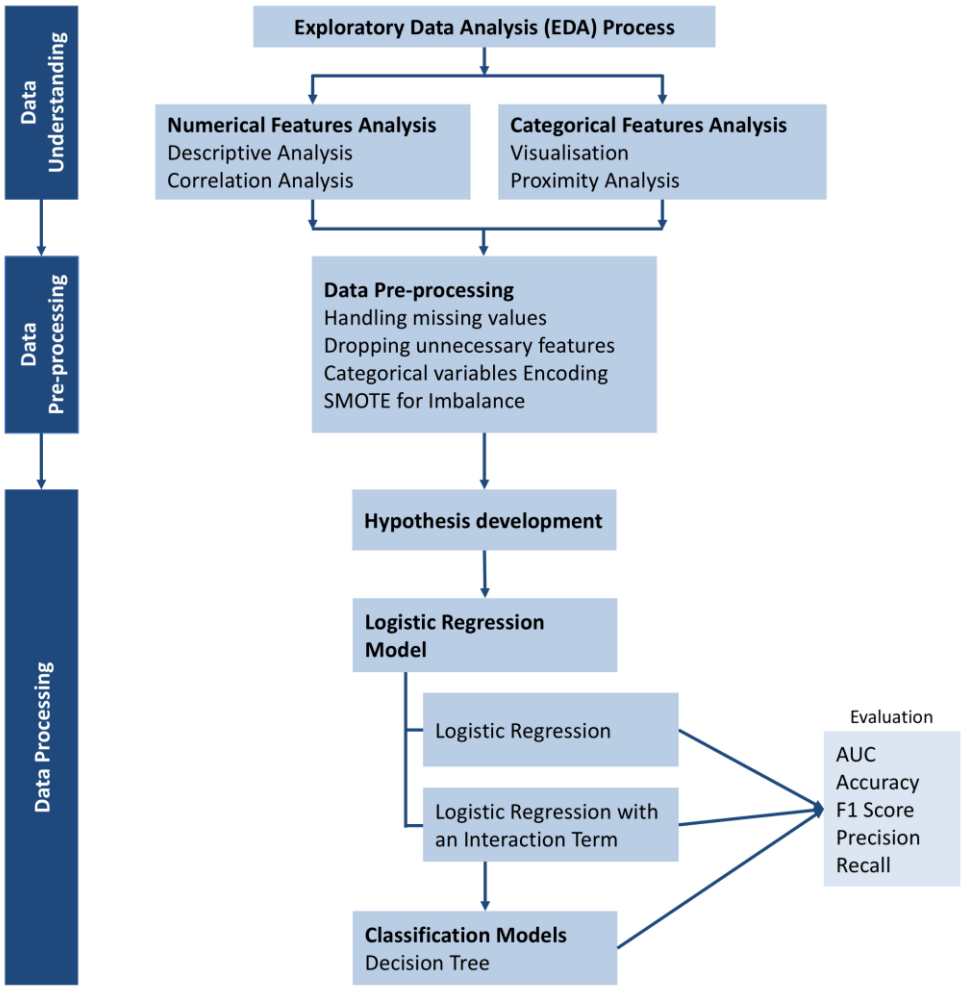

Figure 2: Data Analysis Framework

3.1 Data Analysis Framework Overview

The analytical workflow begins with data understanding, comprising exploratory data analysis (EDA) of both numerical and categorical variables. This step identifies variable distributions, relationships, and outliers, laying the foundation for informed model building.

Subsequently, data pre-processing is conducted to address missing values, remove irrelevant variables, and encode categorical predictors. Crucially, SMOTE (Synthetic Minority Over-sampling Technique) is applied to address class imbalance in churn outcomes, ensuring more reliable model performance.

The next stage involves the development of hypotheses, grounded in theoretical expectations and behavioural science literature. These hypotheses guide the interpretation of results and model evaluation.

The core modelling stage includes the application of:

- Logistic Regression, to quantify the effect of individual predictors on churn probability.
- Logistic Regression with Interaction Terms, to capture how variable effects vary across subgroups (e.g., the impact of pricing across different contract types).
- Decision Tree Algorithm, offering an interpretable, rule-based classification tool for churn prediction.

Model performance is evaluated using metrics such as AUC, accuracy, F1 score, precision, and recall, ensuring robustness and enabling comparison across methods.

3.2 Alignment with Research Objectives

Predict Churn Behaviour Using Classification Models

The models developed successfully estimate churn likelihood and are validated using standard classification performance metrics. This satisfies the objective of providing a predictive foundation for customer retention strategies.

Identify Key Drivers of Churn through Regression Modelling

Logistic regression models quantify the effect of key predictors such as contract type, monthly charges, and tenure, directly addressing this objective. The inclusion of interaction terms further reveals how these effects vary by customer subgroup, enhancing managerial insight.

Apply Behavioural Science to Marketing Interpretation

Model results are interpreted through a behavioural science lens, incorporating theories such as loss aversion, fairness perception, and switching cost. This allows findings to be translated into customer-centric marketing actions, supporting the development of personalised offers, retention communications, and loyalty program designs.

II. Variable selection and data dictionary

The chosen data set has been taken from the website "Kaggle" (Md. Abdur Rahman, n.d.). From the original dataset comprising **52 variables** and **1409 observations**, a total of **10 variables** were selected for further analysis. The basis for this selection was twofold: (1) **theoretical relevance and conceptual clarity** of the variables in capturing key customer demographics, usage patterns, and subscription details, and (2) **consistency with variables frequently cited in peer-reviewed literature** on churn prediction in the telecommunications sector (see the *Reference* column in Table 1).

#	Variable Name	Definition	Variable Group	Data level	Data Type	Data Variable	Data Time	Data Organisation	Reference
1	Churn	Indicates whether the customer churned. 0 = Retained, 1 = Churned	Target variable	Nominal (Categorical, Binary)	Qualitative	Multivariate	Cross-Sectional	Structured	(Dhariya, 2023)
2	Age	The customer's age in years.	Customer Demographics	Ratio (Numerical)	Quantitative				(Huang et al., 2012)
3	Gender	The customer's gender. Binary nominal (Male/Female)		Nominal (Categorical)	Qualitative				(Huang et al., 2012)
4	Monthly Charge	The customer's total monthly charge for all their services, presumably in GBP (£).	Service Usage Patterns	Ratio (Numerical)	Quantitative				(Lalwani et al., 2022)
5	Tenure in Months	Time (in months) since customer joined the service.		Ratio (Numerical)	Quantitative				(Lalwani et al., 2022)
6	Avg Monthly GB Download	Average actual GB data consumed per month by customer.		Ratio (Numerical)	Quantitative				(Lalwani et al., 2022)
7	Avg Monthly Long Distance Charges	Average of charges from calls beyond standard regional rate zones (e.g., international)"		Ratio (Numerical)	Quantitative				(AL-Najjar et al., 2022)
8	Number of Dependents	Number of declared dependents using the same account.		Ratio (Numerical)	Quantitative				(AL-Najjar et al., 2022)
9	Contract	The customer's contract type.	Subscription Details	Nominal (Categorical)	Qualitative				(Qureshi et al., 2013)
10	Payment Method	The customer's payment method.		Nominal (Categorical)	Qualitative				(Qureshi et al., 2013)

Table 1: Descriptive Analysis of Numerical Variables

The selected variables include one binary target variable (Churn) and nine predictor variables grouped into three dimensions: **Customer Demographics** (Age, Gender), **Service Usage Patterns** (Monthly Charge, Tenure in Months, Avg Monthly GB Download, Avg Monthly Long Distance Charges, Number of Dependents), and **Subscription Details** (Contract, Payment Method). Each variable's role and data structure are detailed in Table 1, highlighting their contribution to the multivariate, cross-sectional, and structured nature of the dataset.

III. DATA UNDERSTANDING

1. Exploratory Data Analysis

1.1 Numerical Variables Analysis

1.1.1 Descriptive Analytics

Variable	N	N*	Mean	StDev	Min	Q1	Median	Q3	Max	Range	IQR	Mode	N for Mode	Skewness	Kurtosis
Age	1409	0	46.62	17.002	19	32	47	61	80	61	29	19	36	0.09	-1.07
Monthly Charge	1409	0	64.89	30.264	18.8	35.48	70.45	90.225	118.35	99.55	54.75	20.05	11	-0.21	-1.27
Tenure in Months	1409	0	31.10	24.406	1	9	28	55	72	71	46	1	113	0.28	-1.37
Avg Monthly GB Download	1409	0	20.82	21.011	0	4	17	27	85	85	23	0	308	1.21	0.77
Avg Monthly Long Distance Charge	1409	0	23.09	15.475	0	8.99	23.34	36.58	49.98	49.98	27.595	0	137	0.03	-1.26
Number of Dependents	1409	0	0.44	0.933	0	0	0	0	6	6	0	0	1096	2.12	3.86

Table 2: Descriptive Statistics of Numerical Variables

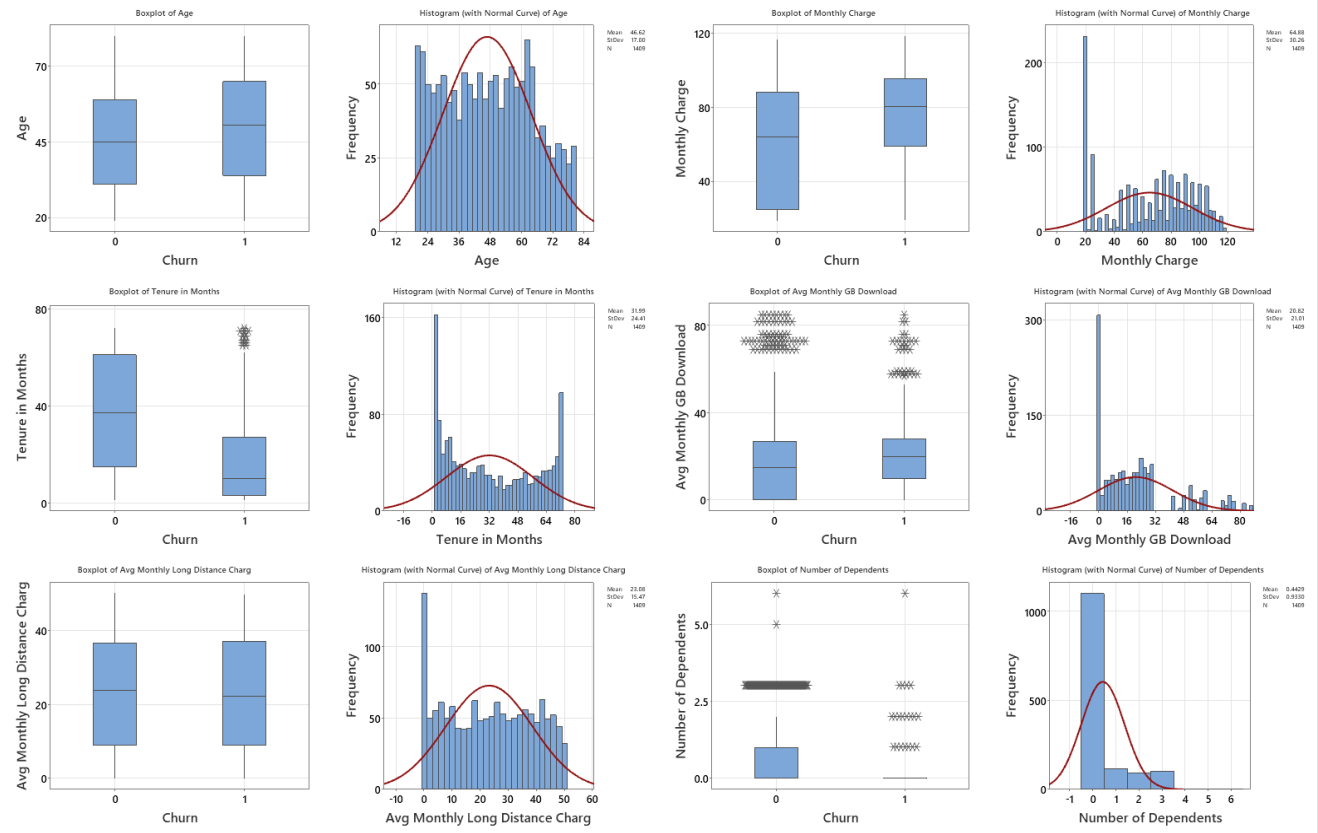

Figure 3: Boxplots (for Retained (Churn = 0) and Churned (Churn = 1)) and Histograms of Numeric Variables

As shown in Table 2 and Figure 3, **Monthly Charge**, **Tenure in Months**, and **Number of Dependents** are particularly informative for distinguishing churned from retained customers.

Monthly Charges are right-skewed, with most customers paying low fees and a smaller segment paying significantly more. Churned customers tend to face **higher charges**, suggesting that **pricing may trigger churn**, possibly due to a mismatch between cost and perceived value. Conversely, **Tenure in Months** shows that customers who churn typically do so early, while retained users have longer service durations. This highlights the **importance of early engagement**, as the risk of attrition is highest in the first year. The **Number of Dependents** is skewed toward zero, especially among churners. Retained customers tend to have more dependents, suggesting that those with families may be more stable due to shared service reliance.

While **Avg Monthly GB Download** and **Long Distance Charges** are also skewed, they show only modest differences between churn groups. Their value may emerge **in combination with other predictors**. Notably, clear **outliers in Tenure and GB Download** warrant preprocessing—such as capping or transformation—to avoid distorting model training.

1.1.2 Correlation Analysis

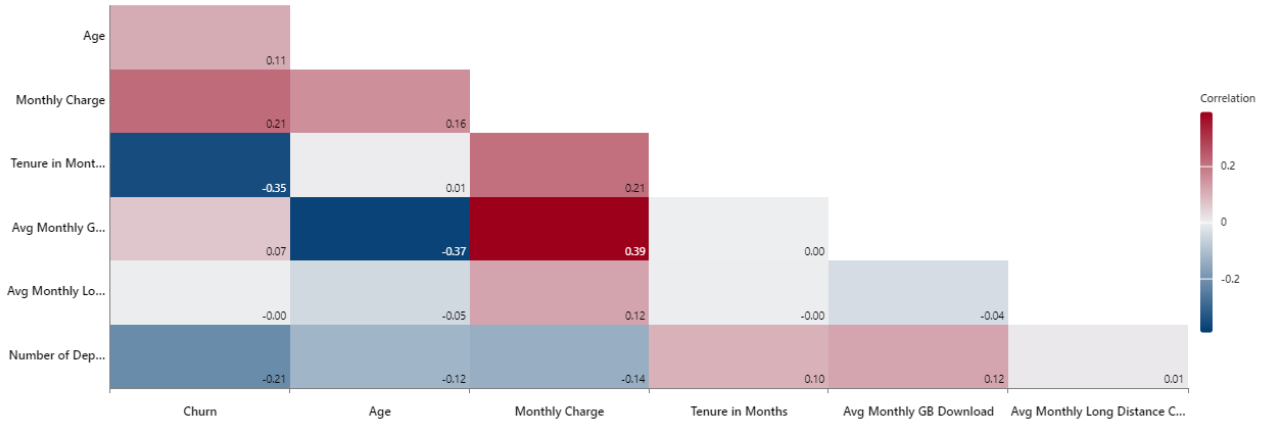

Figure 4: Correlation heatmap of Numerical Variables

All correlations with churn are relatively weak, with the highest being a **negative correlation between tenure and churn (-0.35)**, suggesting that customers with longer service durations are less likely to churn. Similarly, **monthly charge shows a modest positive correlation (0.21)**, aligning with earlier findings that higher bills may be a risk factor for attrition (Lalwani et al., 2022).

The generally **low correlation coefficients across variables indicate a lack of strong multicollinearity**, which is favorable for predictive modeling using **logistic regression**, as this technique assumes that predictor variables are not highly correlated (Midi et al., 2010; Menard, 2002). The independence among features further ensures that each variable contributes uniquely to the model's estimation of churn probability.

1.2 Categorical Variables Analysis

1.2.1 Visualisation

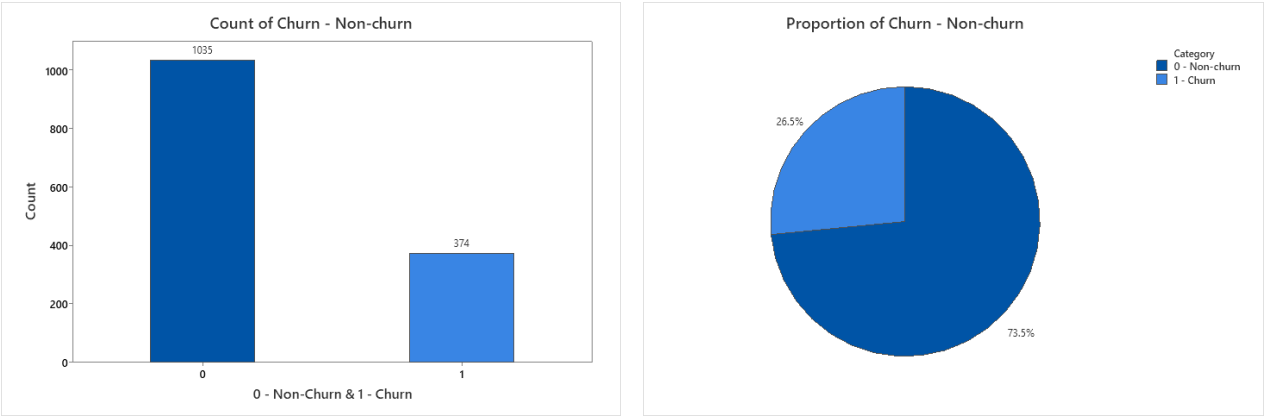

Figure 5: Churn - Non-churn count and distribution

In Figure 5, the distribution of churn (1) and non-churn (0) customers highlights a significant imbalance, with more non-churned customers than churned ones. This imbalance is typical in churn datasets and emphasizes the need for balancing techniques, such as Synthetic Minority Oversampling Technique (SMOTE), during modeling to prevent biased predictions (Field, 2018).

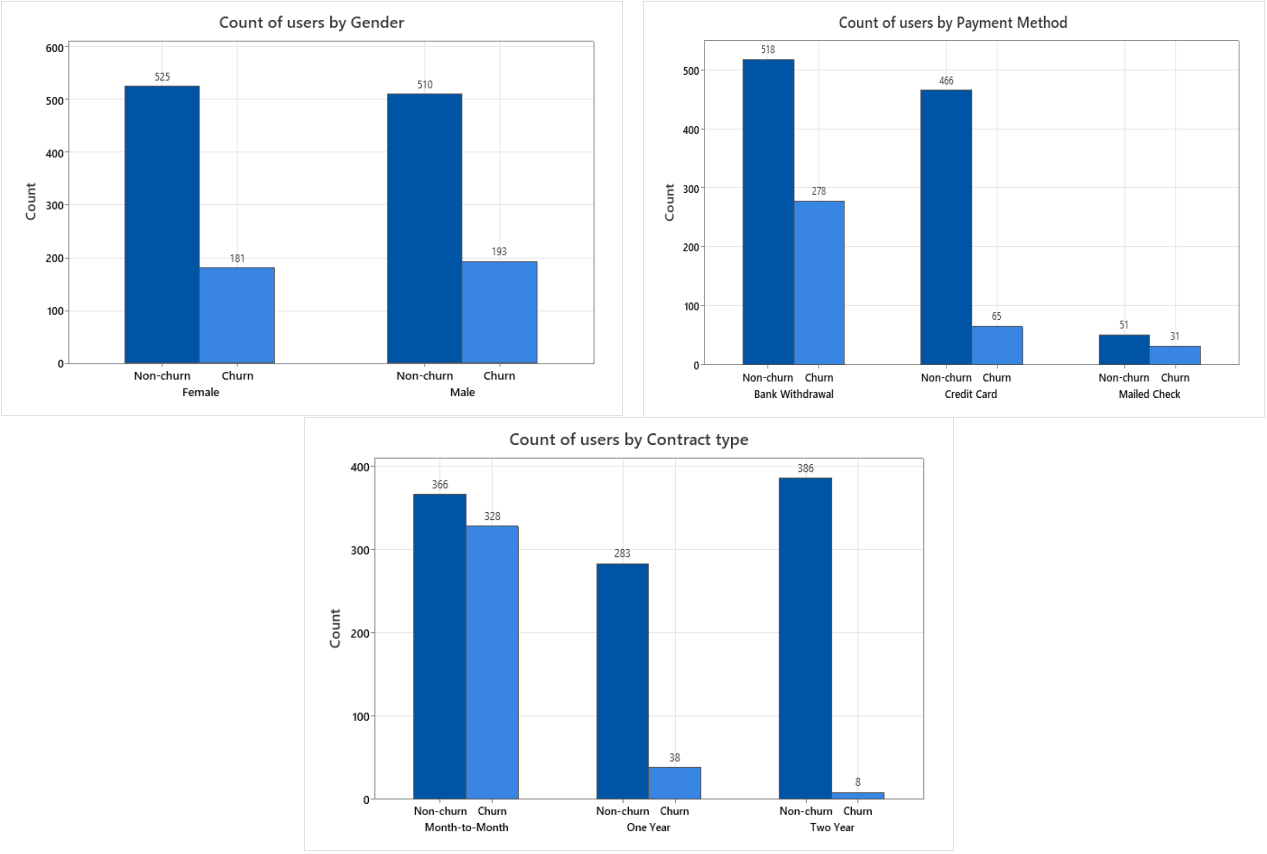

Figure 6: Categorical variables count for Churn - Non-churn

Figure 6 delves deeper by examining churn rates across categorical variables, such as contract type and payment method. For example, customers on month-to-month contracts are likely to exhibit higher churn rates compared to those on longer-term contracts. This reflects the flexibility of shorter contracts, making it easier for customers to switch providers. Similarly, payment methods like mailed checks might show higher churn compared to automated methods like bank withdrawals or credit cards, where ease of payment fosters longer retention.

1.2.2 Proximity Analysis

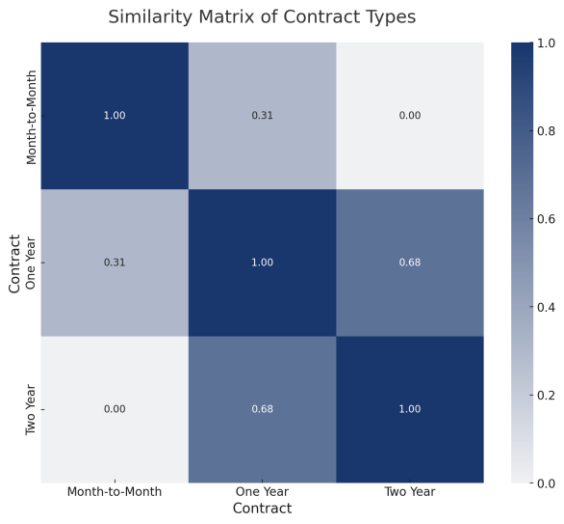

Figure 7: Similarity matrix - Contract types comparison
R Code

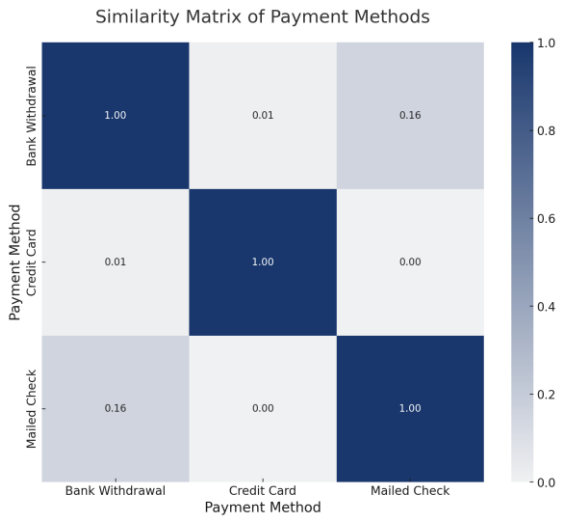

Figure 8: Similarity matrix - Payment Methods comparison
R Code

In Figure 7, the Euclidean distance matrix compares contract types (e.g., month-to-month, one-year, two-year). The results may show that month-to-month contracts are significantly different from longer-term contracts. This indicates that customers on shorter contracts behave differently, likely being less committed and exhibiting higher churn rates compared to those with annual or biennial contracts. The Euclidean distance between categories could

reveal that automated payment methods, such as bank withdrawals or credit cards, are closely related in terms of customer behaviour. Conversely, mailed check payments might stand out as more dissimilar due to their association with higher churn rates.

III. DATA PRE-PROCESSING

1. Elimination of unnecessary features

Unique and irrelevant variables, such as customer IDs and city names, were removed from the dataset. These features do not contribute to predictive modelling and can introduce noise, reducing the model's efficiency.

2. Outlier Capping

R Code

Outliers in numerical features (except *Number of Dependents*) are capped using the interquartile range (IQR) method. Values outside 1.5 × IQR from the first and third quartiles are replaced with the respective boundary values. This step helps reduce the skewing effect of extreme values, ensuring model robustness and better generalization.

3. Categorical variables Encoding

Column name	Variables	Encoding
Contract	Month-to-Month	1
	One Year	2
	Two Year	3
Gender	Female	0
	Male	1
Payment_Method	Mailed Check	1
	Bank Withdrawal	2
	Credit Card	3

Table 3: Details of Digitalisation of Categorical Variables

R Code

Categorical variables were encoded to facilitate their use in machine learning models, which generally require numerical inputs. Table 3 provides a summary of the encoding process which ensure that categorical variables are appropriately represented as numeric data while preserving their categorical relationships. The encoding step also supports the SMOTE process in the following step.

4. SMOTE

R Code

As shown in Figure 5, there is an imbalance between churned (1) and non-churned (0) customers, with a higher proportion of non-churned customers. This imbalance can bias machine learning models toward the majority class, reducing the accuracy of churn predictions. To address this, balancing techniques such as Synthetic Minority Oversampling Technique (SMOTE) were applied to oversample the minority class.

IV. DATA PROCESSING

1. Hypotheses Statements for the Econometric Models

To examine the drivers of customer churn in the UK telecommunications sector, the following hypotheses are developed based on academic literature and industry insights. These will be tested using **logistic regression** and **decision tree** models.

H1: Contract Type Influences Customer Churn

- **H_0:** Contract type has no significant effect on churn.
- **H_1:** Customers on longer-term contracts (e.g., one-year or two-year) are less likely to churn than those on monthly plans.

Justification: Longer contracts increase switching costs and foster customer commitment. **Zheng (2025)** found that contract length significantly lowers churn rates.

H2: Tenure Negatively Affects Customer Churn

- **H_0:** Tenure has no effect on churn.
- **H_1:** Longer tenure is associated with lower churn likelihood.

Justification: Longer customer relationships often lead to higher loyalty. **Chang et al. (2024)** identified tenure as a key retention predictor.

H3: Monthly Charges Positively Influence Customer Churn

- **H_0:** Monthly charges do not affect churn.
- **H_1:** Higher monthly charges increase the likelihood of churn.

Justification: Higher costs can lead to dissatisfaction and switching. **Zheng (2025)** observed that monthly spending impacts churn, especially among younger users.

H4: Number of Dependents Reduces Customer Churn

- **H_0:** Number of dependents has no effect on churn.
- **H_1:** Customers with dependents are less likely to churn.

Justification: Multi-user households rely more on services and are less likely to switch. **Chang et al. (2024)** reported lower churn in such segments.

H5: Interaction Between Contract Type and Monthly Charges Affects Customer Churn

- **H_0:** No interaction exists between contract type and monthly charges.
- **H_1:** The effect of monthly charges on churn varies by contract type.

Justification: Spending sensitivity may differ across contract types. **Zheng (2025)** found that pricing impacts churn differently among contract categories.

2. Logistic Regression Models

2.1 Logistic Regression Model

The logistic regression model was developed to estimate the likelihood of customer churn based on a set of selected predictors.

$$log\left(\frac{P(Churn=1)}{1-P(Churn=1)}\right) = \beta_0 + \beta_1 A + \beta_2 C_1 + \beta_3 C_2 + \beta_4 G + \beta_5 P_1 + \beta_6 P_2 + \beta_7 \text{MC} + \beta_8 \text{T} + \beta_9 \text{GB} + \beta_{10} \text{LD} + \beta_{11} \text{ND}$$

In which:

Predictors	Coefficient	Coefficient Value	p-value
(Intercept)	β_0	0.8723	0.6805
A = Age	β_1	1.0112	**0.0073**
C_1 = Contract_One Year	β_2	0.2630	**7.94E-16**
C_2 = Contract_Two Year	β_3	0.0471	**2E-16**
G = Gender_Male	β_4	1.1975	0.1198
P_1 = Payment Method – Bank Withdrawal	β_5	0.7391	0.2076
P_2 = Payment Method – Credit Card	β_6	0.3080	**2.22E-06**
MC = Monthly Charge	β_7	1.0286	**2E-16**
T = Tenure in Months	β_8	0.9743	**2.03E-12**
GB = Avg Monthly GB Download	β_9	1.0005	0.9052
LD = Avg Monthly Long Distance Charges	β_{10}	0.9874	**0.0014**
ND = Number of Dependents	β_{11}	0.5035	**2.94E-14**

Equation 1: Logistic Regression Model Equation

R Code

The model applies **reference categories**, with "Month-to-Month" as the baseline for contract type, "Female" for gender, and "Mailed Check" for payment method. Coefficients are interpreted relative to these baselines, explaining their absence from the output.

Among all predictors, **contract type** is most influential. Compared to month-to-month plans, customers on **one-year** and **two-year contracts** are significantly less likely to churn, with odds ratios of **0.2630 (p < 0.001)** and **0.0471 (p < 0.001)**, corresponding to **74% and 95% reductions**, respectively. This reinforces the role of contractual commitment as a key retention factor.

Monthly charge is statistically significant (p < 0.001); each unit increase raises churn odds by **2.9%**, suggesting pricing sensitivity, particularly in competitive segments. **Tenure** similarly offers protection, with each additional month reducing churn odds by **2.6% (OR = 0.9743, p < 0.001)**. The **number of dependents** also lowers churn risk (**β = 0.5035, p < 0.001**), likely reflecting household reliance on service.

In contrast, **gender (β = 1.1975, p = 0.1198)**, **bank withdrawal (β = 0.7391, p = 0.2076)**, and **average monthly GB download (β = 1.0005, p = 0.9052)** are not statistically significant. Their odds ratios are close to 1, with high p-values, indicating **limited or random effects** on churn.

2.2 Logistic Regression Model with an Interaction Term

$$log\left(\frac{P(Churn=1)}{1-P(Churn=1)}\right) = \beta_0 + \beta_1 A + \beta_2 C_1 + \beta_3 C_2 + \beta_4 G + \beta_5 P_1 + \beta_6 P_2 + \beta_7 \text{MC} + \beta_8 \text{T} + \beta_9 \text{GB} + \beta_{10} \text{LD} + \beta_{11} \text{ND} + \beta_{12}(C_1 \times MC) + \beta_{13}(C_2 \times MC)$$

In which:

Predictors	Coefficient	Coefficient Value	p-value
(Intercept)	β_0	1.2601	0.5058
A = Age	β_1	1.0130	**0.0020**
C_1 = Contract_One Year	β_2	0.0465	**5.95E-10**
C_2 = Contract_Two Year	β_3	0.0010	**1.38E-06**
G = Gender_Male	β_4	1.1921	0.1314
P_1 = Payment Method – Bank Withdrawal	β_5	0.7454	0.2415
P_2 = Payment Method – Credit Card	β_6	0.2939	**2.35E-06**
MC = Monthly Charge	β_7	1.0213	**8.11E-12**
T = Tenure in Months	β_8	0.9715	**3.50E-14**
GB = Avg Monthly GB Download	β_9	1.0018	0.6644
LD = Avg Monthly Long Distance Charges	β_{10}	0.9890	0.0054
ND = Number of Dependents	β_{11}	0.4944	**7.25E-15**
$C_1 \cdot MC$ = ContractOne Year:Monthly.Charge	β_{12}	1.0238	**6.79E-05**
$C_2 \cdot MC$ = ContractTwo Year:Monthly.Charge	β_{13}	1.0459	**0.0024**

Equation 2: Logistic Regression Model Equation with an Interaction Term

R Code

Building upon the earlier logistic regression analysis, the extended model introduces interaction terms between **contract type** and **monthly charge** to explore whether the relationship between pricing and churn varies depending on the customer's subscription plan. Interestingly, the most influential predictors—contract type, monthly charge, and tenure—remain consistent across models.

Notably, both interaction terms—**Contract One Year × Monthly Charge** and **Contract Two Year × Monthly Charge**—are statistically significant, with odds ratios just above 1. This suggests that **the effect of monthly charges on churn becomes more pronounced among customers on longer-term contracts**. In other words, although one-year and two-year contracts generally serve as a buffer against churn, this protective effect weakens when monthly bills rise. Even customers who have made a long-term commitment to the provider are not immune to pricing pressure.

3. Decision Tree Classification Algorithm

To predict customer churn, a Decision Tree classification model was chosen for its **interpretability and practical value** in translating analytical outputs into actionable marketing strategies. This approach handles categorical and numerical data, captures non-linear relationships, and models variable interactions without complex preprocessing, making it both accessible and robust.

The **effectiveness of decision trees** in **churn prediction** is **well-supported**. Idris et al. (2013) and Amin et al. (2019) highlight their competitive performance and interpretability, particularly in customer relationship management contexts where **transparency is essential**.

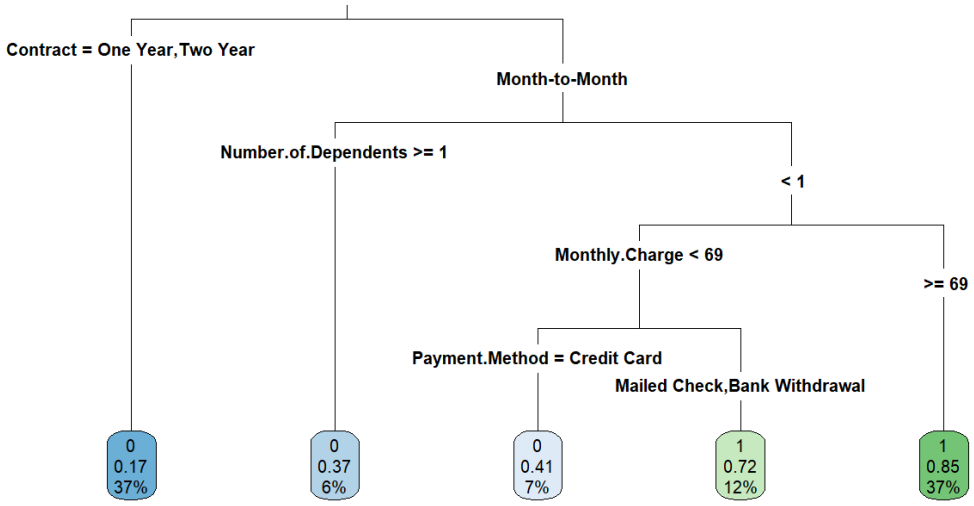

Figure 9: Decision Tree for Predicting Customer Churn Using Full Dataset
R Code

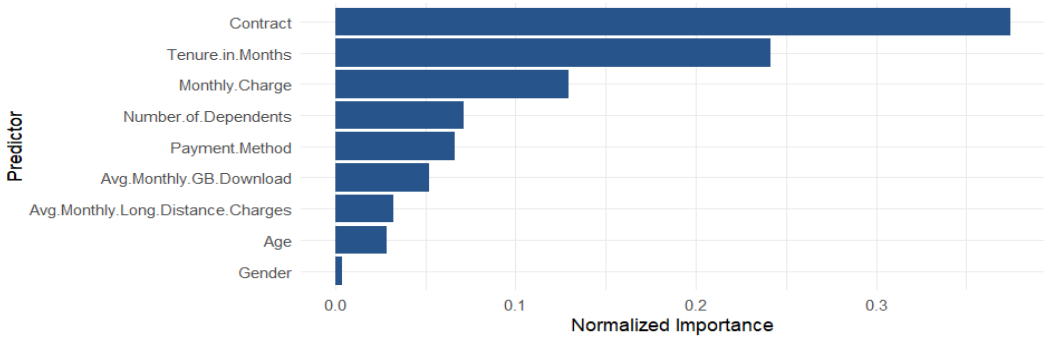

Figure 10: Predictor Importance in Decision Tree Model for Customer Churn
R Code

At the top of the tree, **contract type** is the **primary splitting variable**. Customers on one-year or two-year contracts are largely predicted not to churn, as shown by the leftmost terminal nodes with low churn probabilities **(0.17 and 0.37)**. This supports the notion that contractual commitment is a strong retention mechanism, consistent with the logistic regression results.

For month-to-month customers, the tree further splits based on number of dependents, monthly charges, and payment method. Notably, **churn risk** is highest **(0.85)** among customers with **no dependents**, high monthly charges **(≥ £69)**, and payment via bank withdrawal or mailed check. This segment represents vulnerable customers—short-term users with high bills and minimal household reliance, who may perceive lower service value.

The variable importance plot (Figure 10) identifies **contract, tenure, and monthly charge as the top predictors**—consolidating findings from the logistic model, though with minor ranking differences. Contract dominates the tree's structure, reflecting its central role in early splits. Tenure also ranks high, reinforcing its protective effect against churn. While still important, monthly charge appears slightly less dominant than in regression, where it was among the most statistically significant continuous variables.

Additionally, **payment method and dependents play meaningful roles** in the tree. While dependents were significant in the regression model, the decision tree also captures subtle interactions—particularly between payment type and pricing—that linear models may overlook.

4. Comparison of performance metrics of classification methods

Metric	Logistic Regression	Logistic Regression with an Interaction Term	Decision Tree classification algorithm
AUC	0.878	**0.881**	0.848
Accuracy	0.806	0.807	**0.811**
F1 Score (Churn=1)	0.820	**0.825**	0.798
Precision (Churn=1)	0.790	0.781	**0.820**
Recall (Churn=1)	0.853	**0.873**	0.777

Table 4: Summary of Performance Metrics under Logistic Regression Equations and Decision Tree Classification Algorithms

	Logistic Regression		Logistic Regression with an Interaction Term		Decision Tree classification algorithm	
Reference / Prediction	**0**	**1**	**0**	**1**	**0**	**1**
0	781	165	760	142	844	250
1	254	957	275	980	191	872

Table 5: Summary of Confusion Matrices under Logistic Regression Equations and Decision Tree Classification Algorithms

With the **highest AUC (0.881), F1 Score (0.825)** and **recall (87.3%)**, the logistic regression model with interaction terms demonstrates superior ability to identify churners reliably.

The **standard logistic regression model**, while slightly behind the interaction-enhanced version, also performs admirably. It maintains a strong **AUC (0.878)** and F1 score (0.820), and its **recall of 85.3%** suggests it is still highly effective in flagging churners. However, it lacks the additional nuance provided by interaction effects, which slightly limits its ability to capture more complex churn dynamics.

The **decision tree classification algorithm**, on the other hand, offers a different area of strength. With an **accuracy of 81.1%** and the **highest precision (82%)**, it excels in making confident predictions about churners. In contexts where false positives (incorrectly flagging a loyal customer as a churner) are costly—such as when retention incentives are expensive, this precision could be highly beneficial. However, its relatively **lower AUC (0.848)** and **recall (77.7%)** suggest it may miss a larger portion of actual churners, making it less suitable when the priority is to capture as many true churners as possible.

Predictors	Value	Churn probability by Logistic Regression	Churn probability by Logistic Regression with an interaction term	Churn probability by Decision Tree classification algorithm
Age	24	31.8%	38.2%	41.2%
Gender	Male			
Contract	Month-to-month			
Monthly charge	10			
Tenure in months	7			
Avg monthly GB download	11			
Avg monthly long-distance charge	0			
Number of dependents	0			
Payment Method	Credit Card			

Table 6: Sample churn probabilities under classification algorithms

R Code

The sample pertains to a 24-year-old MSc student in the UK using Giffgaff's £10 month-to-month rolling plan. Based on the classification models, the predicted churn probability ranges from **31.8% to 41.2%**, indicating a moderate likelihood of churn. This estimation is driven by factors such as short tenure, low monthly charge, and absence of dependents—characteristics statistically linked to higher churn risk.

However, in reality, the user has **no intention to churn**, underscoring a key limitation of predictive models: while effective at identifying population-level trends, they may misclassify individuals due to unobservable factors like customer satisfaction or brand loyalty. This example highlights the importance of integrating predictive analytics with qualitative insights to better understand and retain at-risk customers.

V. GENERAL INTERPRETATIONS AND RECOMMENDATIONS OF MARKETING APPROACHES

Based on the predictors identified as most significant across classification models—particularly **contract type**, **tenure**, **monthly charge**, and **number of dependents**—several targeted marketing strategies can be recommended to enhance customer retention.

1. Promote Long-Term Contractual Commitments

The models consistently showed that customers on **one-year or two-year contracts** are significantly less likely to churn than those on monthly rolling plans.

Recommendation: Introduce incentive-based campaigns (e.g., discounts, loyalty points, or data boosts) to motivate monthly plan users to upgrade to longer contracts. Jain and Surana (2017) reports that structured retention bundles tied to contracts can reduce churn by up to **20%** when applied to price-sensitive segments.

2. Tenure-Based Loyalty Strategies

Tenure was found to be inversely associated with churn—longer-term customers tend to stay. However, early-stage customers (e.g., < 12 months) remain highly vulnerable.

Recommendation: Deploy **proactive onboarding and engagement campaigns** within the first year of service—such as welcome emails, usage tips, or milestone rewards—to improve early retention. Elena and Anshu (2023) shows that structured engagement in the first 90 days reduces first-year churn by **over 15%**.

3. Smart Pricing and Billing Transparency

Although the effect size is moderate, higher **monthly charges** are positively associated with churn. This suggests that cost-sensitive customers perceive high charges as a reason to exit.

Recommendation: Offer **tiered service bundles** and **bill shock prevention** alerts to enhance pricing transparency. According to (Ofcom, 2023), pricing concerns account for **58% of customer exits**, highlighting the value of communication and flexible billing options.

4. Leverage Household Dynamics

Customers with **dependents** are significantly more loyal. This is likely due to higher service reliance, household bundling, and increased perceived cost of switching.

Recommendation: Design **family-oriented bundles** or "multi-line discounts" that reward additional line activations or service combinations. Evidence from Elena and Anshu (2023) suggests multi-line packages reduce churn probability by up to **30%** among family households.

5 Personalise Based on Usage and Payment Behaviour

While payment method and data consumption showed weaker individual predictive power, combining them with demographic and behavioural profiles can support micro-segmentation strategies.

Recommendation: Implement **data-driven marketing automation** to personalise offers based on customer payment preferences, usage trends, and churn scores. According to Jain and Surana (2017), such predictive personalisation improves retention rates by **10–15%**.

VI. APPENDIX – R SCRIPTS

Exploratory Data Analysis

```r
#Similarity matrix
library(dplyr)
library(scales)
compute_similarity <- function(df, var) {
  dummies <- model.matrix(~ get(var) - 1, data = df)
  colnames(dummies) <- gsub("get\\(var\\)", "", colnames(dummies))
  dist_mat <- dist(t(dummies), method = "euclidean")
  max_dist <- max(as.matrix(dist_mat))
  sim_mat <- 1 - as.matrix(dist_mat) / max_dist
  diag(sim_mat) <- 1
  return(sim_mat)
}
plot_similarity <- function(sim_mat, title_text, xlab_text, ylab_text) {
  melted_sim <- melt(sim_mat)
  ggplot(melted_sim, aes(x = Var1, y = Var2, fill = value)) +
    geom_tile(color = "white") +
    geom_text(aes(label = sprintf("%.2f", value)), size = 4) +
    scale_fill_gradient(low = "white", high = "#003366", limits = c(0, 1)) +
    theme_minimal() +
    labs(title = title_text, x = xlab_text, y = ylab_text, fill = "") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1),
          panel.grid = element_blank(),
          plot.title = element_text(hjust = 0.5))
}
# Contract similarity
sim_contract <- compute_similarity(Dataset, "Contract")
plot_similarity(sim_contract,
                "Similarity Matrix of Contract Types",
                "Contract", "Contract")

# Payment method similarity
sim_payment <- compute_similarity(Dataset, "Payment.Method")
plot_similarity(sim_payment,
                "Similarity Matrix of Payment Methods",
                "Payment Method", "Payment Method")
```

Figure 11: Similarity matrix

Data Pre-processing

```r
# Capping outliers
cap_outliers <- function(x) {
  q1 <- quantile(x, 0.25, na.rm = TRUE)
  q3 <- quantile(x, 0.75, na.rm = TRUE)
  iqr <- q3 - q1
  lower <- q1 - 1.5 * iqr
  upper <- q3 + 1.5 * iqr
  x[x < lower] <- lower
  x[x > upper] <- upper
  return(x)
}
# Identify numeric columns, excluding "Number.of.Dependents"
numeric_cols <- sapply(Dataset, is.numeric)
target_cols <- setdiff(names(Dataset)[numeric_cols], "Number.of.Dependents")
for (col in target_cols) {
  Dataset[[col]] <- cap_outliers(Dataset[[col]])
}

# Encoding dataset
Dataset$Contract <- recode(Dataset$Contract,
                      "Month-to-Month" = 1,
                      "One Year" = 2,
                      "Two Year" = 3)
Dataset$Gender <- recode(Dataset$Gender,
                      "Female" = 0,
                      "Male" = 1)
Dataset$Payment.Method <- recode(Dataset$Payment.Method,
                      "Mailed Check" = 1,
                      "Bank Withdrawal" = 2,
                      "Credit Card" = 3)
```

Figure 12: Handling outliers

```r
#SMOTE
library(smotefamily)
Dataset$Churn <- as.factor(Dataset$Churn)
X <- Dataset[, setdiff(names(Dataset), "Churn")]
y <- Dataset$Churn

# Apply SMOTE
smote_result <- SMOTE(X, y, K = 5, dup_size = 2)
Dataset_smote <- data.frame(smote_result$data)
Dataset_smote$Churn <- as.factor(Dataset_smote$class)
Dataset_smote$class <- NULL

#Rounding decimals of encoded categorical variables after SMOTE
Dataset_smote$Gender <- round(Dataset_smote$Gender)
Dataset_smote$Contract <- round(Dataset_smote$Contract)
Dataset_smote$Payment.Method <-round(Dataset_smote$Payment.Method)

#Assigning the rounded variables as factors
Dataset_smote$Gender <- factor(Dataset_smote$Gender, levels = c(0, 1),
                               labels = c("Female", "Male"))
Dataset_smote$Contract <- factor(Dataset_smote$Contract, levels = c(1, 2, 3),
                                 labels = c("Month-to-Month", "One Year", "Two Year"))
Dataset_smote$Payment.Method <- factor(Dataset_smote$Payment.Method,levels = c(1, 2, 3),
                                       labels = c("Mailed Check", "Bank Withdrawal", "Credit Card"))
```

Figure 13: SMOTE

Data Processing

Logistic Regression Model

```
# Logistic regression
names(Dataset_smote) <- make.names(names(Dataset_smote))
predictors <- setdiff(names(Dataset_smote), "Churn")
formula <- as.formula(paste("Churn ~", paste(predictors, collapse = " + ")))
logit_model <- glm(formula, data = Dataset_smote, family = binomial)
summary(logit_model)
exp(coef(logit_model))
```

Figure 14: Logistic Regression

```
# Model evaluation
library(caret)
library(pROC)
library(e1071)

# Predict probabilities (for class = 1)
pred_prob <- predict(logit_model, type = "response")

# Classify using 0.5 threshold
pred_class <- ifelse(pred_prob >= 0.5, 1, 0)
pred_class <- as.factor(pred_class)
actual <- as.factor(Dataset_smote$Churn)

# Confusion matrix (positive class = "1")
conf_matrix <- confusionMatrix(pred_class, actual, positive = "1")
print(conf_matrix)

# Extract key metrics
accuracy <- conf_matrix$overall["Accuracy"]
precision <- conf_matrix$byClass["Precision"]
recall <- conf_matrix$byClass["Recall"]
f1 <- conf_matrix$byClass["F1"]

# AUC
roc_obj <- roc(actual, pred_prob)
auc_value <- auc(roc_obj)

# Print metrics
cat("Accuracy: ", round(accuracy, 3), "\n")
cat("Precision (Churn=1): ", round(precision, 3), "\n")
cat("Recall (Churn=1): ", round(recall, 3), "\n")
cat("F1 Score (Churn=1): ", round(f1, 3), "\n")
cat("AUC: ", round(auc_value, 3), "\n")
```

Figure 15: Logistic Regression Evaluation

Logistic Regression Model with an Interaction Term

```r
# Logistic regression model with interaction: Contract * Monthly.Charge
logit_model_interact <- glm(
  Churn ~ Contract * Monthly.Charge +
    Age + Gender + Tenure.in.Months +
    Avg.Monthly.GB.Download + Avg.Monthly.Long.Distance.Charges +
    Number.of.Dependents + Payment.Method,
  data = Dataset_smote,
  family = binomial
)
summary(logit_model_interact)
exp(coef(logit_model_interact))
```

Figure 16: Logistic Regression with an Interaction Term

```r
# Model evaluation

# Predict probabilities
pred_probs <- predict(logit_model_interact, type = "response")

# Classify based on threshold (0.5 default)
pred_class <- ifelse(pred_probs >= 0.5, 1, 0)

# Create confusion matrix
conf_matrix <- confusionMatrix(as.factor(pred_class),
                               as.factor(Dataset_smote$Churn), positive = "1")
print(conf_matrix)

# Extract metrics
accuracy <- conf_matrix$overall["Accuracy"]
precision <- conf_matrix$byClass["Precision"]
recall <- conf_matrix$byClass["Recall"]
f1 <- conf_matrix$byClass["F1"]

# Print
cat("Accuracy: ", accuracy, "\n")
cat("Precision: ", precision, "\n")
cat("Recall: ", recall, "\n")
cat("F1 Score: ", f1, "\n")

# 5. ROC Curve and AUC
roc_obj <- roc(Dataset_smote$Churn, pred_probs)
auc_value <- auc(roc_obj)
cat("AUC: ", auc_value, "\n")
```

Figure 17: Logistic Regression with an Interaction Term Evaluation

Decision Tree Classification Algorithm

```r
# Decision tree

library(rpart)
library(rpart.plot)
library(caret)

Dataset_smote$Churn <- as.factor(Dataset_smote$Churn)
dt_model <- rpart(Churn ~ ., data = Dataset_smote, method = "class")
dt_pred <- predict(dt_model, Dataset_smote, type = "class")
conf_matrix <- confusionMatrix(dt_pred, Dataset_smote$Churn, positive = "1")
print(conf_matrix)
rpart.plot(dt_model, extra = 106, type = 3, fallen.leaves = TRUE)
```

Figure 18: Training Decision Tree Classification Algorithm

```r
# Decision tree evaluation
library(pROC)
dt_prob <- predict(dt_model, Dataset_smote, type = "prob")[, "1"]
conf_matrix <- confusionMatrix(dt_pred, Dataset_smote$Churn, positive = "1")
print(conf_matrix)
cm <- as.matrix(conf_matrix$table)
TP <- cm[2,2]
TN <- cm[1,1]
FP <- cm[2,1]
FN <- cm[1,2]
precision <- TP / (TP + FP)
recall <- TP / (TP + FN)
f1_score <- 2 * precision * recall / (precision + recall)

cat("\nPrecision:", round(precision, 3))
cat("\nRecall:", round(recall, 3))
cat("\nF1 Score:", round(f1_score, 3))

roc_obj <- roc(Dataset_smote$Churn, dt_prob)
auc_value <- auc(roc_obj)
cat("\nAUC:", round(auc_value, 3))
```

Figure 19: Evaluating Decision Tree Classification Algorithm

```r
#Ranking importance
importance <- dt_model$variable.importance
importance_norm <- importance / sum(importance)
importance_df <- data.frame(
  Predictor = names(importance_norm),
  Importance = round(importance_norm, 4)
)
importance_df <- importance_df[order(-importance_df$Importance), ]
print(importance_df)

ggplot(importance_df, aes(x = reorder(Predictor, Importance), y = Importance)) +
  geom_col(fill = "#27548A") +
  coord_flip() +
  labs(title = "Predictor Importance in Decision Tree Model",
       x = "Predictor",
       y = "Normalized Importance") +
  theme_minimal()
```

Figure 20: Plotting importance of predictors under Decision Tree Classification Algorithm

Models' comparisons

```
# Example: Predicting churn probabilities
new_data <- data.frame(
  Age = 24,
  Gender = factor("Male", levels = c("Female", "Male")),
  Contract = factor("Month-to-Month", levels = c("Month-to-Month", "One Year", "Two Year")),
  Monthly.Charge = 10,
  Tenure.in.Months = 7,
  Avg.Monthly.GB.Download = 11,
  Avg.Monthly.Long.Distance.Charges = 0,
  Number.of.Dependents = 0,
  Payment.Method = factor("Credit Card",
                          levels = c("Mailed Check", "Bank Withdrawal", "Credit Card"))
)

# Predict churn probability using logistic regression model
churn_prob <- predict(logit_model, newdata = new_data, type = "response")
print(churn_prob)

# Predict churn probability using logistic regression model with an interaction term
churn_prob <- predict(logit_model_interact, newdata = new_data, type = "response")
print(churn_prob)

# Predict churn probability using Decision Tree
dt_prob <- predict(dt_model, new_data, type = "prob")
print(dt_prob)
```

Figure 21: Sample calculation of churn probabilities under classification algorithms

VI. LIST OF REFERENCES

1. AL-Najjar, D., Al-Rousan, N. and AL-Najjar, H. (2022) Machine Learning to Develop Credit Card Customer Churn Prediction. *Journal of Theoretical and Applied Electronic Commerce Research*, 17 (4): 1529–1542. doi:10.3390/jtaer17040077.
2. Amin, A., Al-Obeidat, F., Shah, B., et al. (2019) Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, 94: 290–301. doi:https://doi.org/10.1016/j.jbusres.2018.03.003.
3. Dhariya, S. (2023) "Customer Churn Prediction in Telecommunication Industry using Machine Learning and Deep Learning Approach." In *3rd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2023 - Proceedings*. 2023. Institute of Electrical and Electronics Engineers Inc. pp. 804–810. doi:10.1109/ICIMIA60377.2023.10426097.
4. Elena, M. and Anshu, G. (2023) *Consumers in Focus: 5G Consumer Behaviour Survey Dashboard 2023 | GSMA Intelligence*. Available at: https://www.gsmaintelligence.com/research/consumers-in-focus-5g-consumer-behaviour-survey-dashboard-2023 (Accessed: 20 April 2025).
5. Estrella-Ramón, A., Sánchez Pérez, M., Swinnen, G., et al. (2013) A marketing view of customer value: Customer lifetime value and customer equity. *SOUTH AFRICAN JOURNAL OF BUSINESS MANAGEMENT*, 44: 47–64. doi:10.4102/sajbm.v44i4.168.
6. Field, A. (2018) *Discovering Statistics Using IBM SPSS Statistics*. SAGE Publications. Available at: https://books.google.co.uk/books?id=JIrutAEACAAJ.
7. Huang, B., Kechadi, M.T. and Buckley, B. (2012) Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39 (1): 1414–1425. doi:10.1016/j.eswa.2011.08.024.
8. Idris, A., Khan, A. and Lee, Y.S. (2013) Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification. *Applied Intelligence*, 39. doi:10.1007/s10489-013-0440-x.
9. Jain, P. and Surana, K. (2017) Reducing churn in telecom through advanced analytics. *McKinsey & Company*. Available at: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/reducing-churn-in-telecom-through-advanced-analytics (Accessed: 20 April 2025).
10. Lalwani, P., Mishra, M.K., Chadha, J.S., et al. (2022) Customer churn prediction system: a machine learning approach. *Computing*, 104 (2): 271–294. doi:10.1007/s00607-021-00908-y.
11. Md. Abdur Rahman (n.d.) *Telco Customer Churn*. Available at: https://www.kaggle.com/datasets/borhanitrash/telco-customer-churn (Accessed: 2 December 2024).
12. Menard, S. (2002) *Applied Logistic Regression Analysis*. Thousand Oaks, California: SAGE Publications, Inc. doi:10.4135/9781412983433.
13. Midi, H., S.K., S. and and Rana, S. (2010) Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13 (3): 253–267. doi:10.1080/09720502.2010.10700699.
14. Ofcom (2023) *Switching Tracker Technical Report*. London. Available at: https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/data/statistics/2023/switching-tracker-2023/switching-tracker-technical-report?v=330323 (Accessed: 19 April 2025).
15. Ofcom (2024) *Telecommunications Market Data Update - Q4 2023*. London. Available at: https://www.ofcom.org.uk/research-and-data/telecoms-research/data-updates/telecommunications-market-data-update-q4-2023 (Accessed: 19 April 2025).
16. Qureshi, S.A., Rehman, A.S., Qamar, A.M., et al. (2013) Telecommunication subscribers' churn prediction model using machine learning. *8th International Conference on Digital Information Management, ICDIM 2013*, pp. 131–136. doi:10.1109/ICDIM.2013.6693977.
17. Wu, X., Li, P., Zhao, M., et al. (2022) Customer churn prediction for web browsers. *Expert Systems with Applications*, 209: 118177. doi:https://doi.org/10.1016/j.eswa.2022.118177.