Executive Summary

This report applies marketing analytics to address customer churn and value optimisation in the telecommunications industry. Using a real-world dataset, the analysis explores which engagement behaviours drive customer value and identifies high-risk churn profiles to inform targeted strategies.

A multivariate linear regression revealed that among usage metrics, frequency of SMS was the strongest and most consistent driver of customer value. Seconds of call usage also contributed positively, while subscription length was not a significant predictor. A second linear regression model introduced tariff plan as an interaction term, revealing that premium users derive greater value from engagement behaviours, while long-tenured premium users may yield lower returns due to diminishing usage over time.

To classify churn risk, a decision tree (C5.0 algorithm) was applied. The pruned decision tree identified complaint behaviour as the most powerful churn indicator, with 80% of complainers at risk. Additionally, inactive users with long tenure and low call diversity were flagged as high-risk segments.

Based on these findings, four marketing strategies are recommended: prioritising SMS-engaged users, fast-tracking complaint recovery, reactivating inactive long-tenured users, and optimising plan-based targeting to deepen value across segments. These strategies translate predictive insights into action to drive retention and profitability.

Table of Contents

List of Figures

List of Tables

1. Introduction

In the highly competitive telecommunications industry, marketing analytics plays a pivotal role in driving customer retention, optimising lifetime value, and informing targeted engagement strategies (Kobi and Otieno, 2024). As telecom providers face intense pressure from commoditised services and low switching costs, retaining customers becomes both a financial necessity and a strategic priority (Thoumy and Abdallah, 2017).

Analytics enables firms to harness behavioural data, such as usage patterns, service preferences, and support interactions to not only understand customer behaviour but to predict and influence it (Theodorakopoulos and Theodoropoulou, 2024).

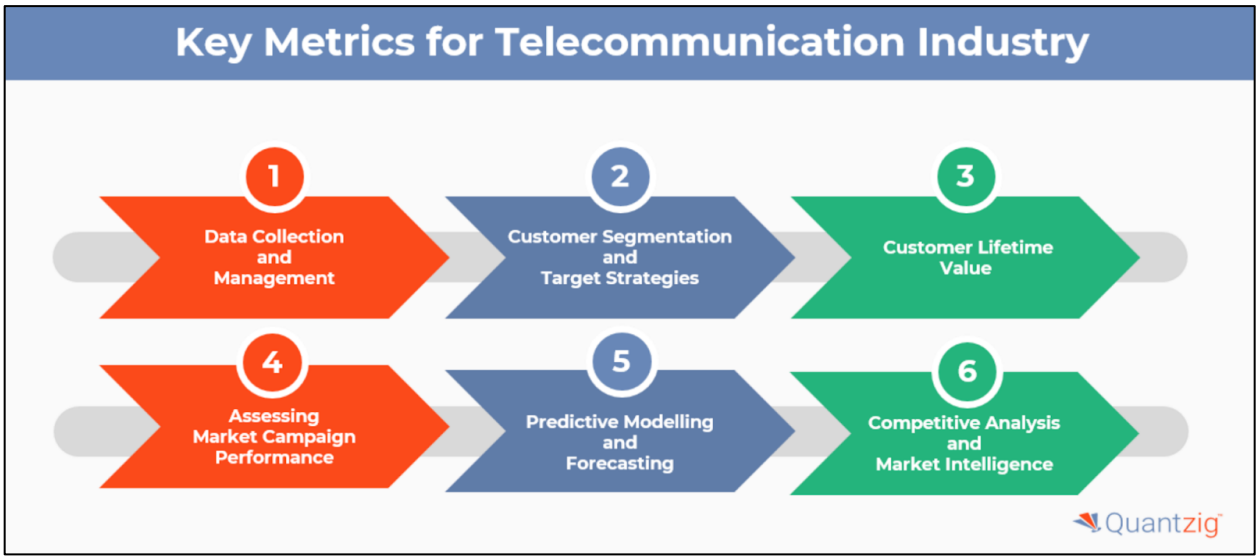


Figure 1: Key Metrics for Telecommunication Industry (Quantzig, 2024)

Instead of relying solely on descriptive reports or reactive churn analysis, Figure 1 shows how forward-thinking telecom companies are embedding predictive and prescriptive models into their marketing operations (Quantzig, 2024). These models help identify high-value customers, tailor interventions based on risk and value segments and optimise resource allocation (Bose *et al.*, 2023). By turning the model outputs into targeted campaigns and prioritised outreach, marketing analytics directly supports strategic decision-making and ROI-driven execution. Viasat, a US-based telecom company, exemplifies this shift, having transformed its marketing strategy by leveraging analytics to attribute sales to digital channels and significantly improve conversion rates (Santy, 2025).

1.1. Primary Challenge

In the telecom industry, customer churn poses a significant threat to revenue and long-term growth (Adeniran *et al.*, 2024). While many churn prediction models aim to explain

why customers churn, the **primary challenge** of this company is to proactively identify high-value customers at risk of churning and take targeted marketing actions to retain them. The company has access to rich behavioural and service-related data including usage patterns, complaints, membership plans, and customer demographics. The following **research questions** will guide the analysis:

- What factor(s) best drives customer value, so the company can focus on it to drive marketing strategies and promote?
- Does the relationship between customer engagement and value differ by tariff plan?
- Can the company classify which customers are likely to churn, so they can proactively target high-value customers at risk?

Multivariate regression and classification techniques will be utilized to address these questions, and the answers will shed light on customer behaviour and churn patterns.

2. Methodology

2.1. Dataset Overview

A real-world telecommunications dataset was obtained from Kaggle (Link), comprising 3,150 rows and 13 columns related to customer behaviour, service usage, and churn. The dataset is complete, with no missing values, and largely consistent in format. It is considered believable, as it was created following extensive research by the original author (Jafari, 2020). Some issues were noted with uniqueness, as 300 rows were duplicates, highlighting the need for further cleaning during data pre-processing.

2.2. Data Dictionary

Table 1 displays the data dictionary that outlines the variables included in the dataset to provide clarity and context.

Variable	Description	Measurement Units	Data Level	Data Type
Customer Demographic				
age_group	Age group category of the customer	5 categories	Ordinal	Qualitative
Service Issues				
call_failure	Number of failed call attempts by the customer	Count (Integer)	Ratio	Quantitative
complains	Indicator of whether the customer has submitted any complaints	Yes/No	Nominal	Qualitative
Plan & Subscription Information				
tariff_plan	If the customer is subscribed to the basic plan or the premium plan	Basic/Premium	Nominal	Qualitative
subscription_length	Number of months the customer has been subscribed to the service	Months	Ratio	Quantitative
charge_amount	Charge paid by the customer last month	Currency (GBP)	Ratio	Quantitative
Customer Engagement				
seconds_of_use	Duration of outgoing calls of the customer last month	Seconds	Ratio	Quantitative
frequency_of_use	Number of calls made by the customer last month	Count (Integer)	Ratio	Quantitative
frequency_of_sms	Number of sms sent by the customer last month	Count (Integer)	Ratio	Quantitative
distinct_called_numbers	Number of unique phone numbers the customer has called last month	Count (Integer)	Ratio	Quantitative
status	Indicator of whether the customer is actively using the service or not	Active/Inactive	Nominal	Qualitative
Customer Value & Outcome				
customer_value	The monetary value of the customer to the company	N/A (0 - 2170)	Ratio	Quantitative
churn	Indicator of whether the customer has churned	Yes/No	Nominal	Qualitative

Table 1: Data Dictionary (Source: Author)

2.3. Feature Selection

Purpose	Variable	References
Multivariate Regression (with and without Interaction)		
Dependent Variable	customer_value	(Homburg and Wielgos, 2022), (Hossain, Akter and Yanamandram, 2021), (Jerab, 2023)
Independent Variables	seconds_of_use	(Sim *et al.*, 2022), (Agbonasevbaefe, 2023)
	subscription_length	(Hajar *et al.*, 2022), (Gazi *et al.*, 2024)
	frequncy_of_sms	(Awuku, Agyei and Gonu, 2023), (Awwad, 2024)
Interaction	tariff_plan	(Saha *et al.*, 2022), (Tripathy *et al.*, 2021)
Classification		
Target Variable	churn	(Singh *et al.*, 2024), (Routh, Roy and Meyer, 2021)

Table 2: Feature Selection (Source: Author)

Table 2 shows for the multivariate regression analysis, the dependent variable selected will help to understand the drivers of profit contribution from each individual. Independent variables represent the duration and diversity of customer engagement, helping to identify which behavioural patterns most influence customer value and should therefore be prioritised in marketing strategies. Introducing an interaction with tariff_plan will enable the business to refine its service offerings, ensuring they align with targeted promotions based on the membership plans. Lastly, a classification model is applied using churn as the target variable to help identify high-risk customers and guide proactive retention strategies based on all other variables.

3. Data Exploration and Preparation

3.1. Numerical Feature Analysis

Numerical Feature Analysis uses statistical summaries and visuals to understand distributions, detect outliers, and identify trends that may require transformation for better model performance (Choi *et al.*, 2021).

	call._failure	subscription _length	charge_amount	seconds_of_use	frequency_of_use	frequency_of_sms	distinct_called_numbers	customer_value
Mean	7.80	32.45	0.97	4534.24	70.48	73.79	23.87	474.99
Standard Error	0.14	0.16	0.03	78.67	1.08	2.10	0.32	9.64
Median	6	35	0	3041	54.5	22	21	232.52
Mode	0	36	0	0	0	0	0	0
Standard Deviation	7.33	8.72	1.55	4199.71	57.40	112.06	17.19	514.44
Sample Variance	53.67	76.09	2.40	17637583.42	3294.93	12557.98	295.63	264650.77
Kurtosis	0.84	1.07	8.45	0.96	0.81	3.26	1.39	1.22
Skewness	1.07	-1.25	2.53	1.31	1.15	1.97	1.04	1.42
Range	36	44	10	17090	255	522	97	2165.28
Minimum	0	3	0	0	0	0	0	0
Maximum	36	47	10	17090	255	522	97	2165.28
Sum	22237	92491	2778	12922593	200882	210301	68031	1353722.545
Count	2850	2850	2850	2850	2850	2850	2850	2850

Table 3: Descriptive Statistics of Numerical Variables (Source: Author)

Table 3 shows the descriptive statistics which reveal key information about the central tendency, spread, and distribution of numerical variables, enabling the identification of patterns in customer behaviour (Alabi *et al.*, 2023). Variables seconds_of_use and frequency_of_use have high mean (4534.24 and 70.48 respectively) and strong right skew (skewness > 1), showing that while most customers engage moderately, some are heavy users. Frequency_of_sms is also highly skewed (1.97) with the largest standard deviation (112.06), suggesting uneven SMS usage across the base.

The feature customer_value shows the widest range (0–2165.28) and high kurtosis (1.22), indicating both variation and extreme outliers. Its positive skew (1.42) and high standard deviation (514.44) point that while most customers are moderately profitable, a few contribute significantly more to total revenue. In contrast, subscription_length is left-skewed (-1.25), with a median of 35 and a narrow spread, suggesting customer tenures are more stable. charge_amount has low overall values but exhibits high kurtosis (8.45), showing that rare high charges inflate the upper tail. call_failure shows a moderate spread (SD = 7.33), low median (6), and a skewness of 1.07, suggesting that failures are infrequent but vary significantly across users. These patterns highlight diverse behavioural profiles that can inform segmentation and prediction in subsequent analysis.

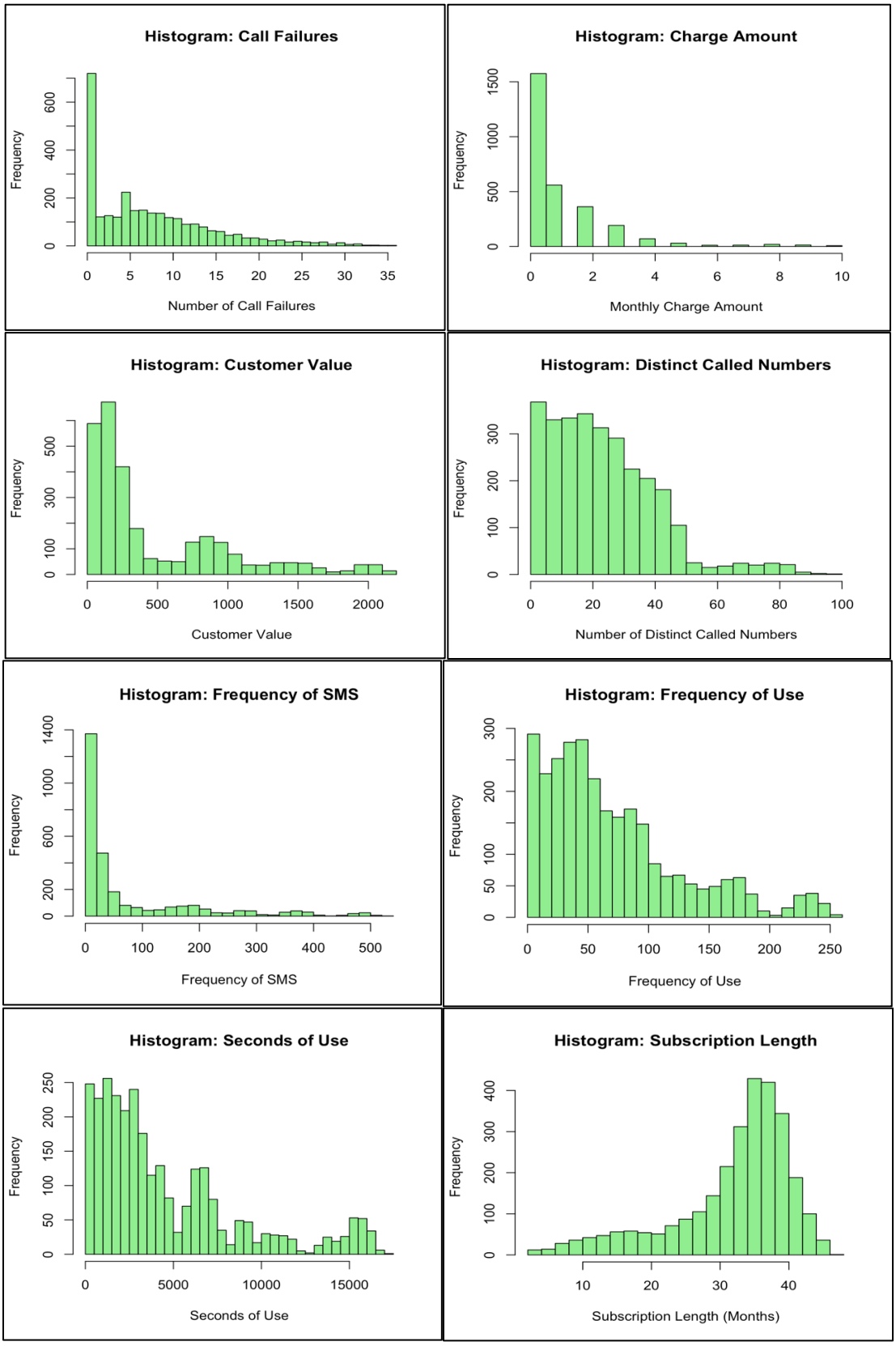
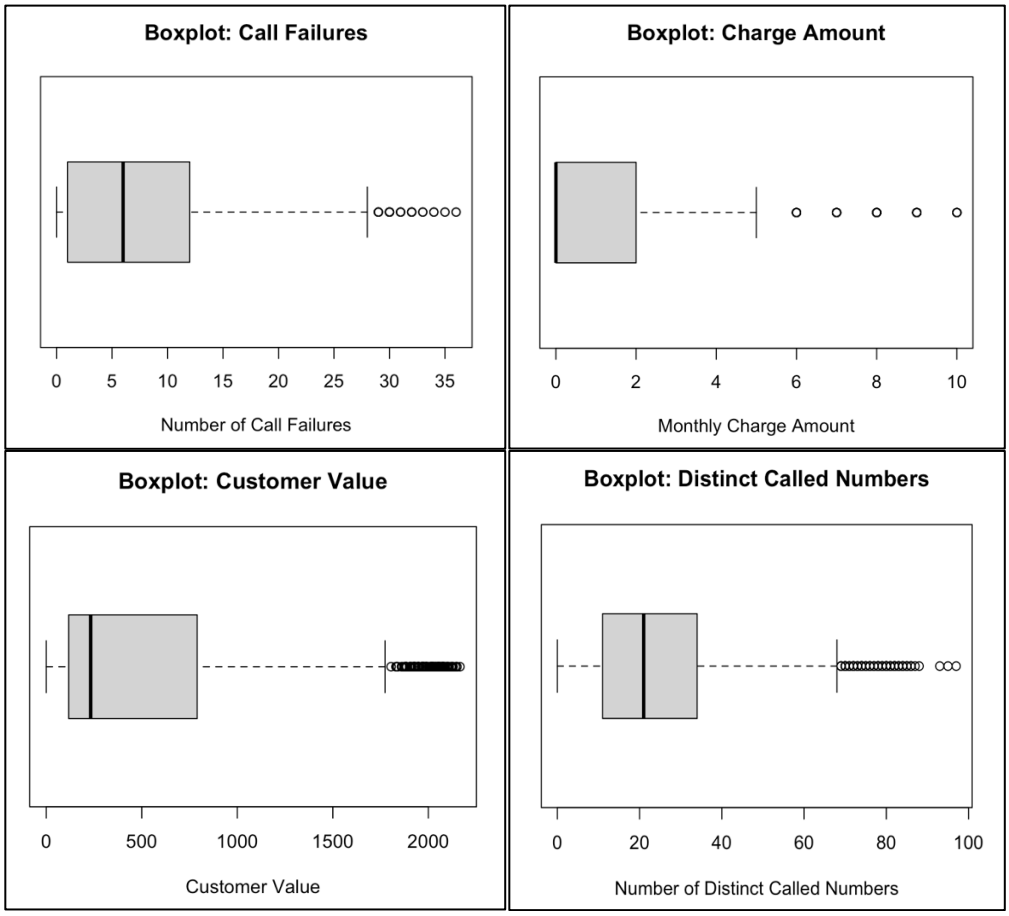


Figure 2: Histograms of Numerical Variables (Source: Author)

Figure 2 displays the histograms which show the distribution and spread of numerical variables, helping to detect skewness, concentration points, and potential outliers (Madhavan, 2025). Most variables, including call_failure (skew = 1.07) and charge_amount (skew = 2.53), show a strong right-skew with high frequencies near zero, indicating that customers rarely experience call failures or incur high charges.

Customer_value is also right-skewed, suggesting that a small subset of customers contributes disproportionately to overall profitability. Frequency_of_sms (SD = 112.06) shows extreme variation, indicating that only a small group uses SMS heavily. Seconds_of_use and frequency_of_use are more dispersed, reflecting varied engagement levels. Subscription_length follows a near-normal distribution, peaking around 35 months, indicating a stable core of long-term subscribers who may be more loyal. These patterns support the need for variable transformation and help uncover customer segments with differing engagement and value profiles.



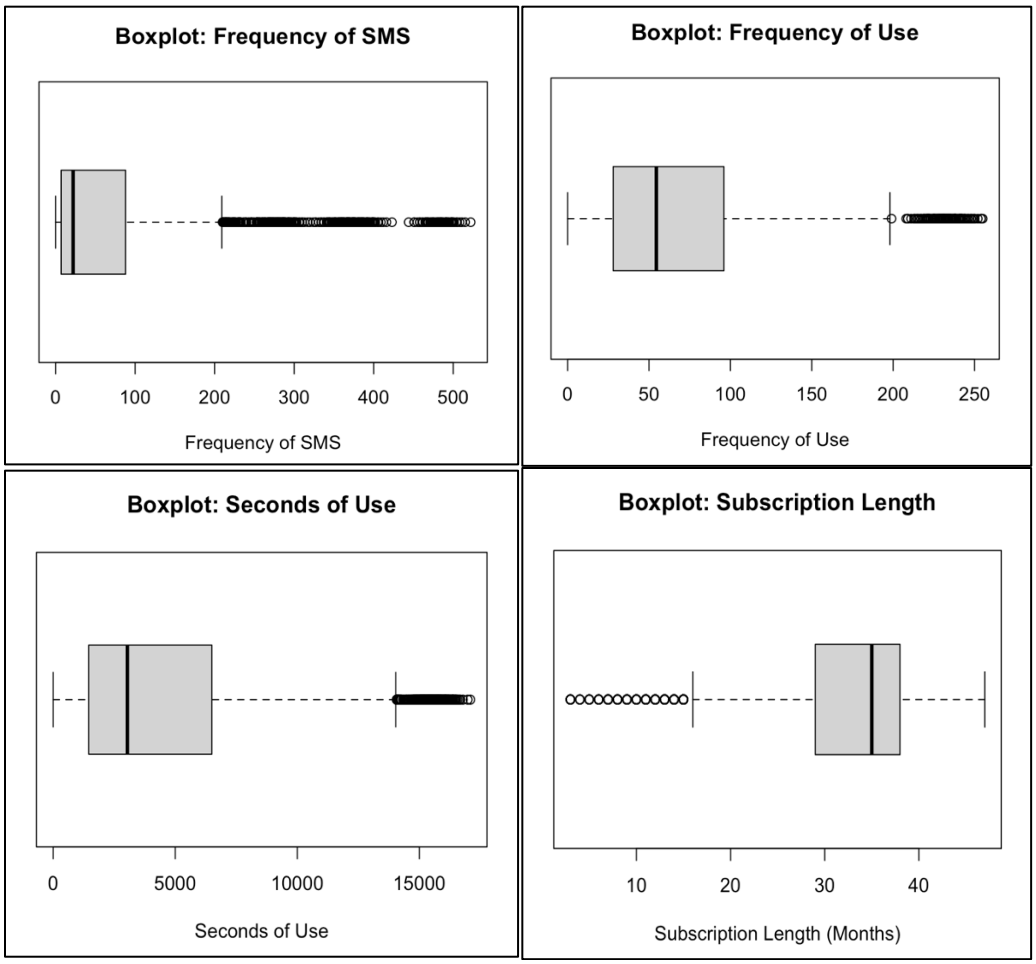


Figure 3: Boxplots of Numerical Variables (Source: Author)

Figure 3 shows the boxplots which reveal the spread and outliers across numerical variables (Mazarei *et al.*, 2025). Customer_value shows a wide interquartile range and many high-end outliers, highlighting a small group of highly profitable customers. Charge_amount and Frequency_of_sms have tight interquartile ranges but numerous outliers, indicating uneven costs and usage. Seconds_of_use and frequency_of_use display greater variability, reflecting diverse engagement levels. Subscription_length appears balanced with minimal outliers, suggesting consistent tenure among users. Outliers in call_failure and distinct_called_numbers indicate differences in service experience and calling patterns. No outliers will be removed as they are real but rare cases. These patterns support segmentation and reveal edge cases worth deeper investigation.

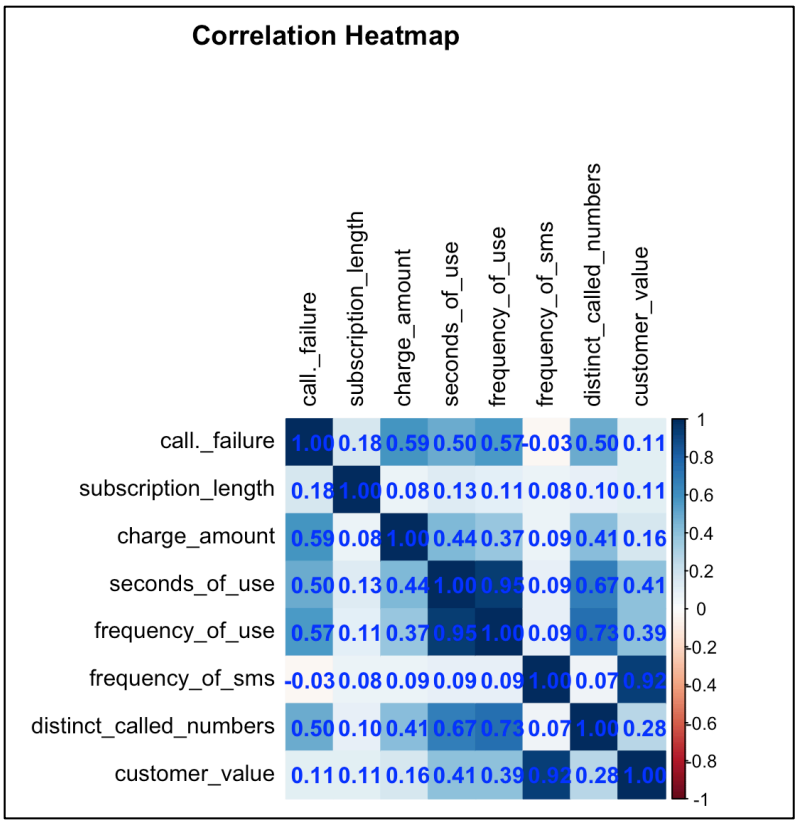


Figure 4: Correlation Heatmap of Numerical Variables (Source: Author)

The correlation heatmap (Figure 4) displays the strength and direction of linear relationships among numerical variables, helping to detect associations and potential multicollinearity (Rais *et al.*, 2025). The strongest positive correlation with customer_value is seen in seconds_of_use (0.41), followed by frequency_of_use (0.39) and distinct_called_numbers (0.28), suggesting that usage intensity and call diversity are key drivers of customer profitability.

Notably, seconds_of_use and frequency_of_use are highly correlated (0.85), which may indicate potential multicollinearity if both are used in the same model. Charge_amount is moderately correlated with several variables, but weakly associated with customer_value (0.16), implying it is less useful for predicting profitability.

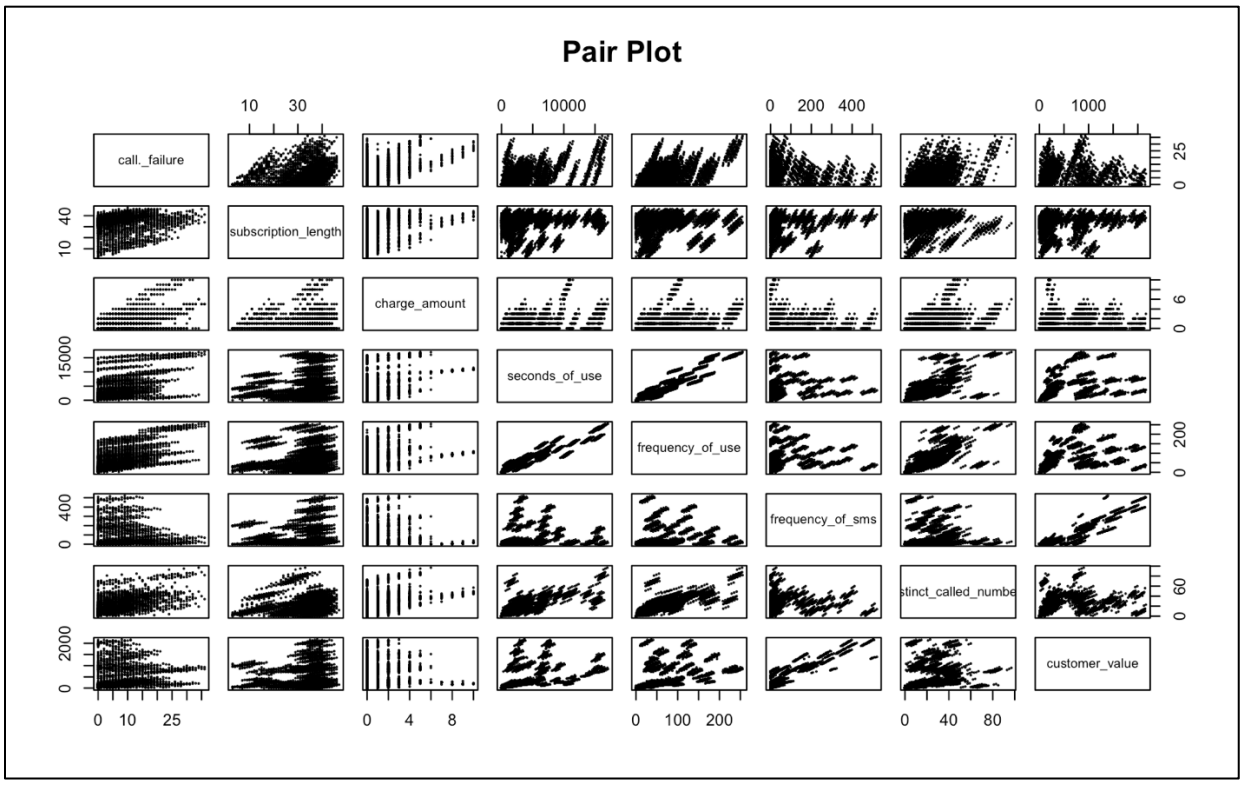


Figure 5: Pair Plot of Numerical Variables (Source: Author)

The pair plot (Figure 5) helps visualise relationships between numerical variables (Kaushik *et al.*, 2022). A strong linear trend appears between seconds_of_use and frequency_of_use, indicating potential multicollinearity. customer_value is positively associated with usage-related variables, supporting their predictive relevance. In contrast, charge_amount and frequency_of_sms show weaker, scattered patterns. These visual cues complement earlier findings and support more targeted variable selection for modelling.

3.2. Categorical Feature Analysis

Categorical Feature Analysis helps examine the distribution of non-numerical variables, enabling comparisons across groups and identifying dominant customer segments (Bilder and Loughin, 2024).

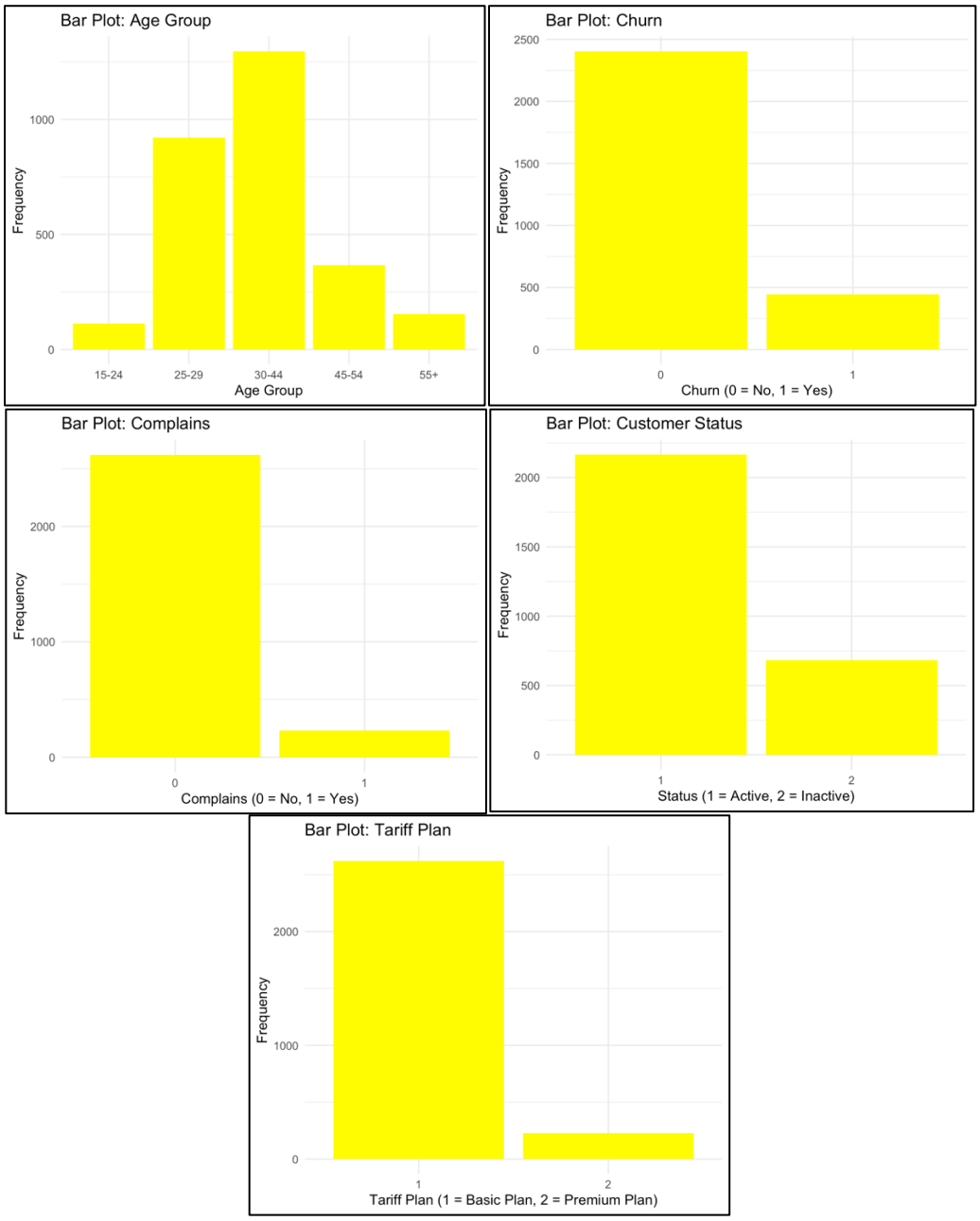


Figure 6: Bar Plots of Categorical Features (Source: Author)

In Figure 6, the bar plots highlight the distribution of categorical variables. Most customers fall into the 30–44 age group and use the basic tariff plan. A large majority are active users with no complaints and have not churned. The churn rate appears low overall, though

notable enough to explore. These distributions inform segmentation, as younger, active users on basic plans form the core customer base worth targeting.

3.3. Data Pre-processing

Table 4 highlights the pre-processing steps followed to prepare the dataset for further analysis.

Processes	Actions
Data Integation	No integration was done
Data Cleaning	300 duplicate rows were removed prior to EDA
Data Transformation	A new column named 'customer_id' was introduced prior to Data Analysis to use it as an unique identifier for each customer, making the final column count 14
	The column named 'tariff_plan' was a binary categorical with values of either 1 & 2. It has been transformed to binary where 1 = 0 and 2 = 1. (0 = Basic Plan, 1 = Premium Plan)
	The column named 'status' was a binary categorical with values of either 1 & 2. It has been transformed to binary where 1 = 0 and 2 = 1. (0 = Active Member, 1 = Inactive Member)
	Factorised the categorical columns to ensure they are in the correct format for analysis

Table 4: Data Pre-processing steps (Source: Author)

4. Data Analysis

4.1. Multivariate Linear Regression (without Interaction)

4.1.1. Hypothesis Statement

In the telecom industry, customer value often reflects the revenue generated over a subscriber's lifecycle. Understanding which usage behaviours drive higher customer value is essential for designing effective marketing strategies and retention efforts. In this context, variables such as seconds_of_use, subscription_length, and frequency_of_sms serve as behavioural indicators of customer engagement and service utilisation. It is expected that increased usage across these dimensions leads to greater value. Therefore, the following hypothesis statements are proposed to test whether customer engagement significantly predicts customer value:

H_0: *None of the engagement metrics significantly explain the variation in customer value.*
H_1: *At least one of the engagement metrics significantly explains the variation in customer value.*

To test these hypotheses, we apply a multivariate linear regression model. This method is appropriate because it allows us to simultaneously assess the influence of multiple engagement variables on customer value, while controlling for the effect of each predictor (Gligor and Bozkurt, 2020).

$$\text{customer\_value} = \beta_0 + \beta_1(\text{seconds\_of\_use}) + \beta_2(\text{subscription\_length}) + \beta_3(\text{frequency\_of\_sms}) + \epsilon \quad \ldots\ldots\ldots\ldots\ldots (1)$$

In this equation,
β_0 = Intercept
β_1 = Coefficient of seconds_of_use
β_2 = Coefficient of subscription_length
β_3 = Coefficient of frequency_of_sms
ϵ = Random Error Component

4.1.2. Regression Outputs and Insights

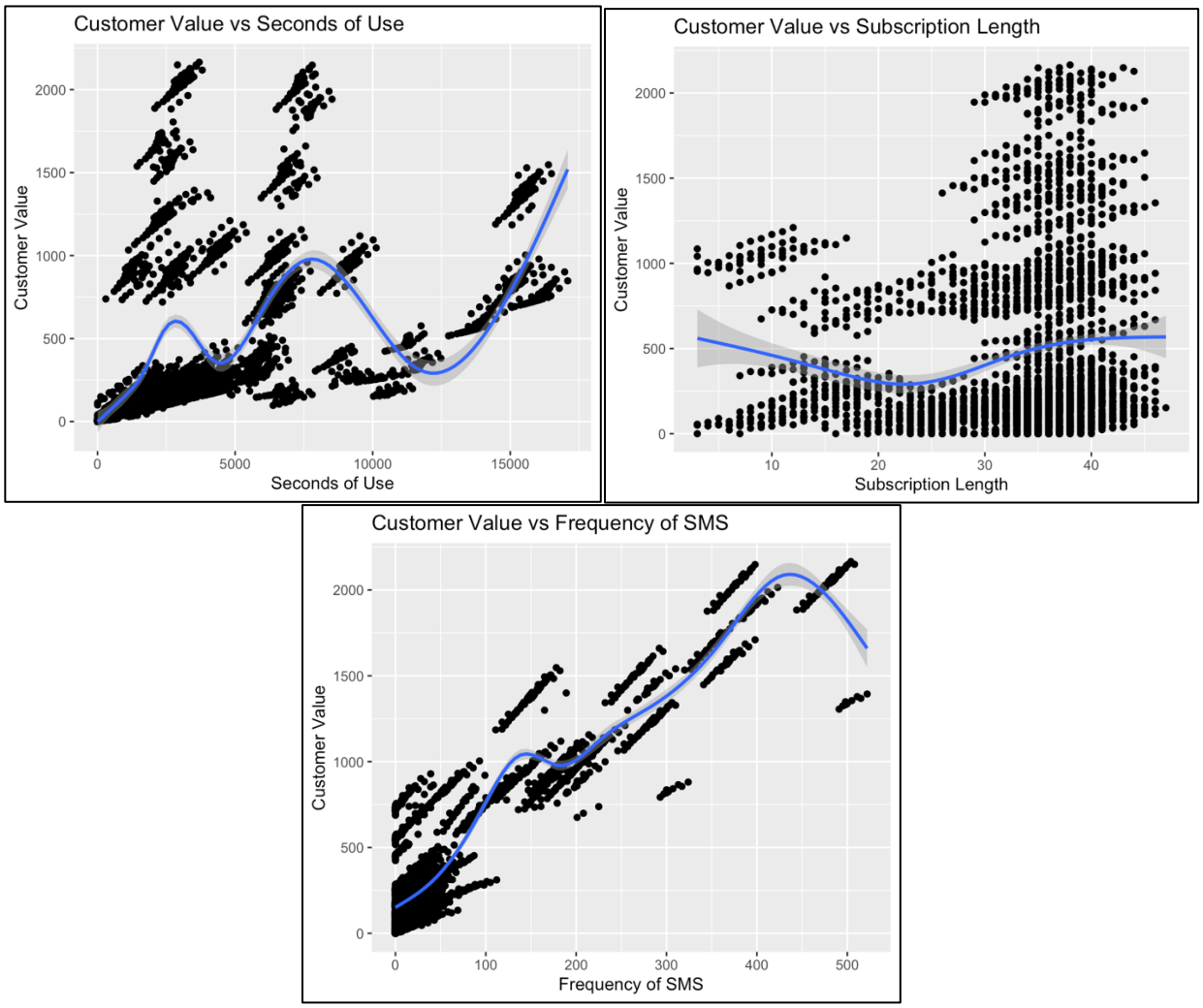


Figure 7: Relationships between the dependent and independent variables (Source: Author)

The visual trends (Figure 7) suggest a strong positive relationship between frequency_of_sms and customer_value, with value peaking at higher SMS levels. Seconds_of_use also shows an upward pattern, particularly beyond 15,000 seconds, though some mid-range dips are present. Subscription_length displays a more subtle effect, with customer value gradually increasing after 25 months. These observations imply that both communication intensity and loyalty duration contribute to value, justifying their inclusion in the multivariate regression (Foroudi, Cuomo and Foroudi, 2020).

```
Call:
lm(formula = customer_value ~ seconds_of_use + subscription_length +
    frequency_of_sms, data = customer)

Residuals:
    Min      1Q  Median      3Q     Max
-876.99   -7.92    9.05   28.04  256.07

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -7.9654322  7.8547634  -1.014    0.311
seconds_of_use       0.0398688  0.0004836  82.441   <2e-16 ***
subscription_length -0.0052995  0.2325740  -0.023    0.982
frequency_of_sms     4.0974862  0.0180351 227.195   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 107.1 on 2846 degrees of freedom
Multiple R-squared:  0.9567,    Adjusted R-squared:  0.9566
F-statistic: 2.094e+04 on 3 and 2846 DF,  p-value: < 2.2e-16
```

		2.5 %	97.5 %
(Intercept)	-7.965432241	-23.36703575	7.43617127
seconds_of_use	0.039868836	0.03892059	0.04081708
subscription_length	-0.005299479	-0.46133014	0.45073118
frequency_of_sms	4.097486176	4.06212296	4.13284939

Figure 8: Multivariate Linear Regression results (Source: Author)

Revised equation:

$$\text{customer\_value} = -7.965 + 0.040(\text{seconds\_of\_use}) - 0.005(\text{subscription\_length}) + 4.097(\text{frequency\_of\_sms}) + \epsilon$$

The results from Figure 8 indicate that seconds_of_use and frequency_of_sms are both highly significant predictors of customer_value, with p-values less than 2e-16, well below the 0.05 threshold. In contrast, subscription_length is not statistically significant (p = 0.982), suggesting that merely retaining customers for longer periods does not necessarily drive higher value in this dataset.

Among the predictors, frequency_of_sms shows the strongest effect, with each additional SMS increasing customer value by 4.10. Seconds_of_use adds 0.04 value per second, while subscription_length has a negligible negative effect (−0.0053). Of these, only frequency_of_sms stands out with a narrow 95% confidence interval [4.06, 4.13], confirming its reliability. This suggests that high messaging activity is the most consistent and impactful driver of customer value.

The model explains a remarkable 95.7% of the variance in customer value ($R^2 = 0.9567$), indicating excellent fit and predictive accuracy. The adjusted R^2 remains virtually unchanged, confirming that all included predictors contribute meaningfully without inflating the model. Residuals are relatively well-behaved with a median error of 9.05,

though the presence of outliers (min = -876.99, max = 256.07) suggests a few high-value customers are not perfectly predicted, possibly due to unmeasured factors such as promotions or plan upgrades.

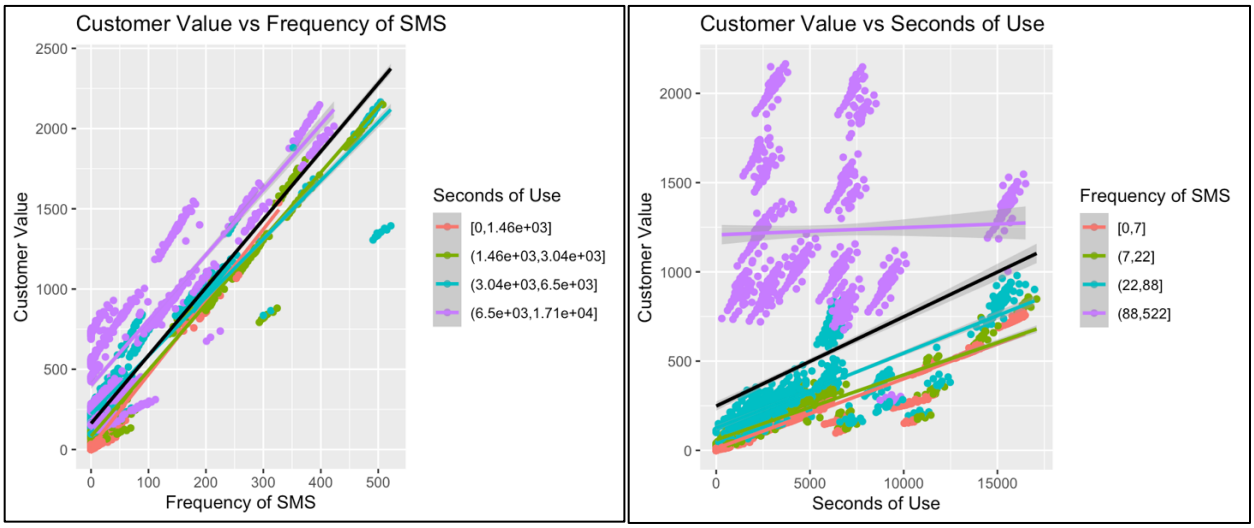


Figure 9: Relationship between dependent and independent variables (Source: Author)

Figure 9 reveal that the positive effect of frequency_of_sms on customer value strengthens as seconds_of_use increases. Similarly, customers with high SMS usage show the steepest value gains with longer call duration. This indicates a reinforcing interaction: when both messaging and call engagement are high, value rises disproportionately. These behaviours are most impactful when combined, not in isolation. Subscription_length is not examined here, as it was found to be statistically insignificant in the initial model.

4.2. Multivariate Linear Regression (with Interaction)

4.2.1. Hypothesis Statement

Customer behaviour does not always translate into equal value across all segments. In telecom, the type of subscription, whether basic or premium, often reflects distinct customer needs, usage habits, and pricing structures. A usage pattern that is profitable in one segment may be less impactful in another. To explore whether value-driving behaviours differ by plan type, tariff_plan is introduced as a moderator. This aligns with marketing practices that tailor offers based on customer tiers. Therefore, the following hypothesis statements are proposed:

H_0: *Tariff plan does not influence the relationship between engagement behaviours and customer value.*
H_1: *Tariff plan influences the relationship between at least one engagement behaviour and customer value.*

To test these hypotheses, we apply a multivariate linear regression with interaction terms, as it enables us to assess whether the effects of engagement behaviours on customer value vary across different tariff plans:

$$\begin{aligned}\text{customer\_value} = {} & \beta_0 + \beta_1(\text{seconds\_of\_use}) \\ & + \beta_2(\text{subscription\_length}) \\ & + \beta_3(\text{frequency\_of\_sms}) \\ & + \beta_4(\text{tariff\_plan}) \\ & + \beta_5(\text{seconds\_of\_use x tariff\_plan}) \\ & + \beta_6(\text{subscription\_length x tariff\_plan}) \\ & + \beta_7(\text{frequency\_of\_sms x tariff\_plan}) + \epsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots\ (2)\end{aligned}$$

In this equation,

β_0 = Intercept
β_1 = Coefficient of seconds_of_use
β_2 = Coefficient of subscription_length
β_3 = Coefficient of frequency_of_sms
β_4 = Coefficient of tariff_plan
β_5 = Coefficient of seconds_of_use and tariff_plan
β_6 = Coefficient of subscription_length and tariff_plan
β_7 = Coefficient of frequency_of_sms and tariff_plan
ϵ = Random Error Component

4.2.2. Regression Outputs and Insights

To examine whether tariff_plan moderates the value relationship, the first step is to test the significance of tariff_plan as a standalone predictor.

```
Call:
lm(formula = customer_value ~ seconds_of_use + subscription_length +
    frequency_of_sms + tariff_plan, data = customer)

Residuals:
    Min      1Q  Median      3Q     Max
-859.82    2.38   11.45   28.05  232.50

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.274e+01  7.881e+00  -2.886  0.00393 **
seconds_of_use       3.923e-02  4.805e-04  81.646  < 2e-16 ***
subscription_length  4.370e-01  2.334e-01   1.872  0.06129 .
frequency_of_sms     4.062e+00  1.812e-02 224.224  < 2e-16 ***
tariff_plan          7.336e+01  7.602e+00   9.650  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 105.5 on 2845 degrees of freedom
Multiple R-squared:  0.958,     Adjusted R-squared:  0.958
F-statistic: 1.624e+04 on 4 and 2845 DF,  p-value: < 2.2e-16
```

Figure 10: Multivariate Linear Regression results (Source: Author)

Figure 10 shows the model's R-squared rises slightly to 0.958, suggesting a marginal improvement in explanatory power as more predictors are included, making the model's ability to explain variation generally improved. Tariff_plan is statistically significant ($p < 0.001$), indicating that basic and premium customers differ meaningfully in value. The median residual also increases slightly to 11.45, hinting at more variation across customer segments.

Revised equation:

$$\begin{aligned} \text{customer\_value} = &- 43.643 + 0.040(\text{seconds\_of\_use}) \\ &+ 1.082(\text{subscription\_length}) \\ &+ 4.034(\text{frequency\_of\_sms}) \\ &+ 220.223(\text{tariff\_plan}) \\ &+ 0.014(\text{seconds\_of\_use} \times \text{tariff\_plan}) \\ &- 13.589(\text{subscription\_length} \times \text{tariff\_plan}) \\ &+ 0.972(\text{frequency\_of\_sms} \times \text{tariff\_plan}) + \epsilon \end{aligned}$$

```
Call:
lm(formula = customer_value ~ seconds_of_use + subscription_length +
    frequency_of_sms + tariff_plan + seconds_of_use:tariff_plan +
    subscription_length:tariff_plan + frequency_of_sms:tariff_plan,
    data = customer)

Residuals:
    Min      1Q  Median      3Q     Max
-851.80   -1.98    9.43   32.02  239.27

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -43.643334   7.948235  -5.491 4.35e-08 ***
seconds_of_use                   0.039602   0.000485  81.651  < 2e-16 ***
subscription_length              1.082125   0.234328   4.618 4.05e-06 ***
frequency_of_sms                 4.033966   0.018103 222.836  < 2e-16 ***
tariff_plan                    220.223487  25.415539   8.665  < 2e-16 ***
seconds_of_use:tariff_plan       0.014415   0.002287   6.304 3.36e-10 ***
subscription_length:tariff_plan -13.588614   1.045990 -12.991  < 2e-16 ***
frequency_of_sms:tariff_plan     0.971893   0.096013  10.123  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 102 on 2842 degrees of freedom
Multiple R-squared:  0.9608,    Adjusted R-squared:  0.9607
F-statistic:  9939 on 7 and 2842 DF,  p-value: < 2.2e-16
```

Figure 11: Multivariate Linear Regression results with Interaction (Source: Author)

In Figure 11, the updated model shows all three interaction terms are statistically significant ($p < 0.001$), indicating that usage behaviours influence value differently for premium versus basic users. For premium users (tariff_plan = 1), the value added by seconds_of_use increases by 0.0144, while frequency_of_sms adds nearly 0.97 more per SMS than for basic users. However, subscription_length has a sharply negative interaction

(−13.59), suggesting that longer tenures may reduce value for premium users, potentially due to front-loaded pricing or lower usage over time.

In comparison to the previous model, seconds_of_use and frequency_of_sms remain highly significant for basic users (tariff_plan = 0), confirming their strong predictive value. For premium users (tariff_plan = 1), the interaction terms for both behaviours are also highly significant ($p < 0.001$), indicating that these effects are even more pronounced. Meanwhile, subscription_length remains insignificant for basic users ($p = 0.982$), though its interaction term is significant and negative for premium users, suggesting a different behavioural pattern in that group.

The R^2 increases marginally to 0.9608, and residual error slightly improves, suggesting modest gains in fit. Median residual remains stable at 9.43. The main effects, representing basic users, remain strong and significant, confirming that usage behaviours reliably drive customer value in this segment.

4.3. Classification

Decision tree was selected as the classification model because it supports categorical outcomes like churn, offers rule-based segmentation, and is highly interpretable (Manzoor *et al.*, 2024). The C5.0 algorithm enables efficient classification while highlighting the most important predictors (Wang, 2024).

4.3.1. Pre-pruning results

```
Classification Tree
Number of samples: 2850
Number of predictors: 12

Tree size: 38
```

```
Evaluation on training data (2850 cases):

           Decision Tree
         ----------------
         Size      Errors

           37   87( 3.1%)   <<


          (a)   (b)    <-classified as
         ----  ----
         2345    59    (a): class 0
           28   418    (b): class 1
```

Figure 12: Pre-pruning results of Classification (Source: Author)

Figure 12 shows the initial model had a low error rate of 3.1% on the training data, with only 87 misclassifications. However, the tree contained 38 terminal nodes, indicating overfitting. Despite high accuracy, the model's complexity limited its usefulness for interpretation and actionable decision-making for marketing strategies.

4.3.2. Post-pruning results

```
Classification Tree
Number of samples: 2850
Number of predictors: 12

Tree size: 5
```

```
Evaluation on training data (2850 cases):

          Decision Tree
        ----------------
        Size      Errors

           5  258( 9.1%)   <<
```

Figure 13: Post-pruning results of Classification (Source: Author)

To improve interpretability, the tree was pruned down to five leaf nodes. As expected, the error increased to 9.1% (Figure 13), but the structure became far more transparent.

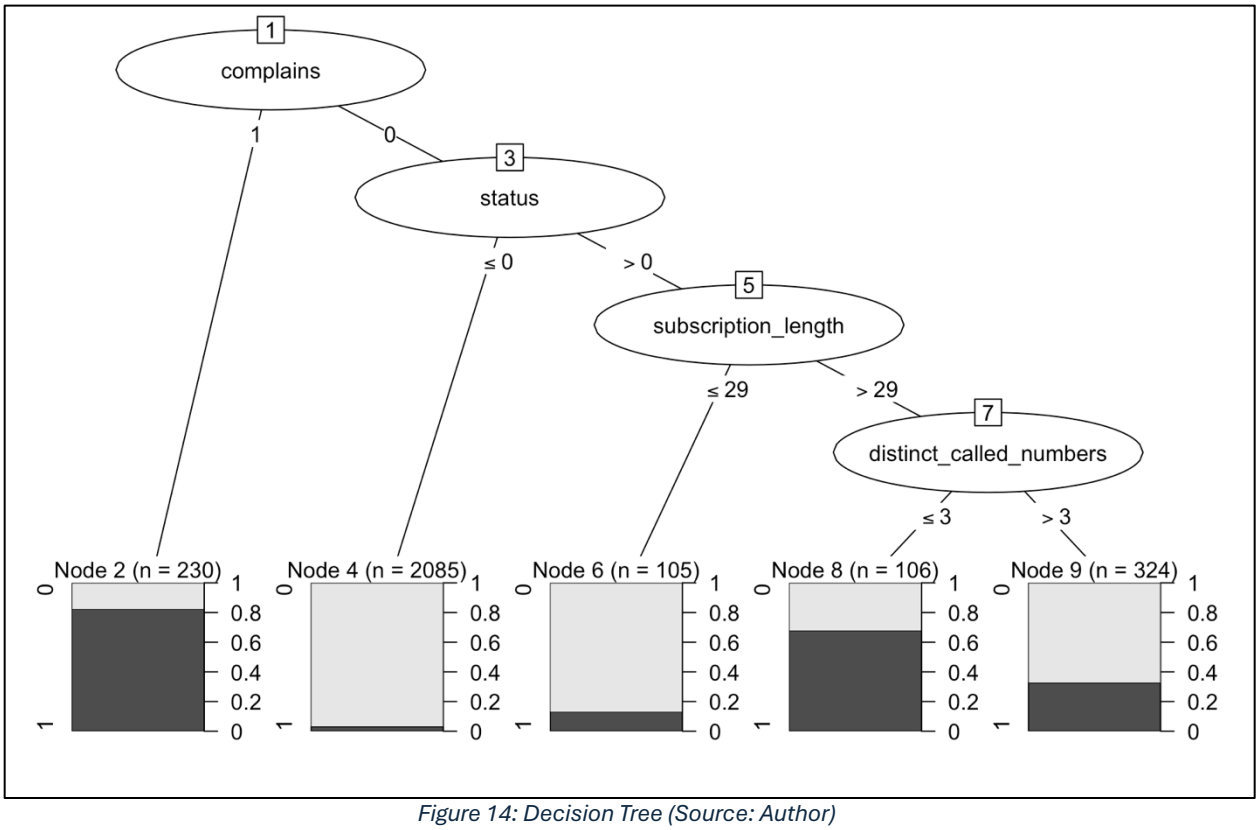


Figure 14: Decision Tree (Source: Author)

Figure 14 shows the root split on complaints highlights complaint behaviour as the strongest indicator of churn. Among the 230 customers who complained, 80% were at risk of churning, which is a clear signal that unresolved complaints require urgent, personalised recovery strategies.

Among the 2,620 customers who did not complain, the model next split on status. Inactive users (status > 0) with longer subscription lengths (> 29 months) formed the highest-risk segment, with over 60% likely to churn. This challenges the assumption that tenure equals loyalty as long-term customers who become inactive may feel neglected or disengaged and should be prioritised for reactivation strategies.

Interestingly, inactive users with shorter tenures (≤ 29 months) showed relatively low churn risk, possibly because they are still exploring the service. These users may benefit from structured onboarding and education.

Inactive users with longer tenures (> 29 months) were further split by distinct_called_numbers. Those with three or fewer contacts had over 60% churn risk, suggesting disengagement despite being subscribed long-term. This low diversity in contact behaviour may reflect limited service use, making them prime targets for usage-based nudges or re-engagement offers. In contrast, those with more than three contacts were less likely to churn, indicating residual value worth retaining.

```
Evaluation on training data (2850 cases):

           Decision Tree
         ----------------
         Size      Errors

            5  258( 9.1%)   <<


          (a)   (b)    <-classified as
         ----  ----
         2330    74    (a): class 0
          184   262    (b): class 1
```

Figure 15: Confusion Matrix and Error results (Source: Author)

Figure 15 highlights that the model achieved 90.9% accuracy, correctly classifying 2,592 of 2,850 cases. It correctly identified 2,330 non-churners (true negatives) and 262 churners (true positives). However, 184 churners were misclassified as non-churners (false negatives), representing silent attrition.

5. Conclusion and Marketing Implications

The findings from the regression and classification models offer clear opportunities to improve retention and revenue through targeted marketing strategies. Below are four evidence-based actions grounded in customer behaviour insights and core marketing principles:

- **Prioritise High-Frequency SMS Users for Value Maximisation**
  The regression model identified frequency_of_sms as the strongest driver of customer value, particularly for premium users. These customers represent a high-value behavioural segment. Using **customer lifetime value (CLV)** as a guiding metric, the company should prioritise these users for loyalty programmes, personalised bundles, and SMS-driven promotions that deepen engagement and extend profitability (Binh, Thy and Phuong, 2021). Airtel, an Indian telecom company, enhanced customer lifetime value by targeting high-SMS users through its "Airtel Thanks" loyalty program, offering personalised rewards and usage-based incentives (Shah, 2024).

- **Proactively Address Complaints to Prevent Churn**
  Complaint behaviour was the top churn predictor, with 80% of complainers at risk. To reduce this, firms should adopt service recovery strategies by implementing Gen-AI to predict contact reasons and route customers efficiently, ensuring rapid, high-touch follow-up for complaints. Verizon, a U.S telecom company, implemented Gen-AI to predict the reason for customer calls and automatically route them to the right agents, enabling faster complaint resolution and reducing churn risk through more effective service recovery (Mukherjee, 2024). Incorporating complaint triggers into churn scoring models also supports **predictive marketing**, enabling automated and timely retention actions before dissatisfaction leads to defection.

- **Reactivate Inactive Long-Term Users with Personalised Engagement**
  Inactive, long-tenured customers with low call diversity were a hidden high-risk group. These users may be experiencing **customer fatigue** or misalignment between needs and services. Reactivation campaigns should apply **personalised lifecycle marketing**, using data to deliver tailored messaging (e.g., plan check-ins or "We Miss You" offers) that rebuild perceived value and re-spark usage.

- **Optimise Tariff Plan Strategies Based on Segment Performance**
  Since the company already segments customers by tariff plan, marketing should now focus on refining tactics within each group. The regression interaction analysis revealed that premium users show stronger value response to engagement behaviours like SMS and call usage. Therefore, this group should be prioritised for **value-based targeting** with personalised usage incentives and upsell offers that deepen engagement. Reliance Jio, an Indian telecommunications company,

optimised segmented tariff plans with ultra-low data pricing for high-usage customers, leveraging usage behaviour to upsell and capture price-sensitive market segments (Canvas Business Model, 2024).

By translating predictive insights into targeted strategies, this analysis equips the company to proactively retain high-risk customers, optimise value within existing segments, and drive more data-informed marketing execution across the customer lifecycle.

6. Bibliography

Adeniran, I.A. *et al.* (2024) 'Implementing machine learning techniques for customer retention and churn prediction in telecommunications', *Computer Science & IT Research Journal*, 5(8), pp. 2011–2025. Available at: https://doi.org/10.51594/csitrj.v5i8.1489.

Agbonasevbaefe, A.O. (2023) *The Impact of Customer Engagement on Performance in Nigeria Mobile Telecoms Industry*. thesis. Anglia Ruskin Research Online (ARRO). Available at: https://aru.figshare.com/articles/thesis/The_Impact_of_Customer_Engagement_on_Performance_in_Nigeria_Mobile_Telecoms_Industry/23765331/1 (Accessed: 5 May 2025).

Alabi, O. *et al.* (2023) 'Introduction to Descriptive Statistics', in *Recent Advances in Biostatistics*. IntechOpen. Available at: https://doi.org/10.5772/intechopen.1002475.

Awuku, E., Agyei, P.M. and Gonu, E. (2023) 'Service innovation practices and customer loyalty in the telecommunication industry', *PLOS ONE*, 18(3), p. e0282588. Available at: https://doi.org/10.1371/journal.pone.0282588.

Awwad, A. (2024) 'The impact of Over The Top service providers on the Global Mobile Telecom Industry: A quantified analysis and recommendations for recovery', *World Journal of Advanced Research and Reviews*, 21(1), pp. 1638–1669. Available at: https://doi.org/10.30574/wjarr.2024.21.1.0113.

Bilder, C.R. and Loughin, T.M. (2024) *Analysis of Categorical Data with R*. CRC Press. Available at: https://books.google.co.uk/books?hl=en&lr=&id=uSUVEQAAQBAJ&oi=fnd&pg=PP1&dq=Categorical+Feature+Analysis&ots=8lqDGPSfr0&sig=KR5DT66ZacZyWY_v5InjksDsIUc&redir_esc=y#v=onepage&q=Categorical%20Feature%20Analysis&f=false.

Binh, T.V., Thy, N.G. and Phuong, H.T.N. (2021) 'Measure of CLV Toward Market Segmentation Approach in the Telecommunication Sector (Vietnam)', *SAGE Open*, 11(2), p. 21582440211021584. Available at: https://doi.org/10.1177/21582440211021584.

Bose, N. *et al.* (2023) 'Leveraging Reinforcement Learning and Predictive Analytics for Enhanced Customer Lifetime Value Optimization', *International Journal of AI Advancements*, 12(8). Available at: https://ijoaia.com/index.php/v1/article/view/20 (Accessed: 5 May 2025).

Canvas Business Model (2024) *Customer Demographics and Target Market of Reliance Jio*, *CANVAS, SWOT, PESTEL & BCG Matrix Editable Templates for Startups*. Available at: https://canvasbusinessmodel.com/blogs/target-market/reliance-jio-target-market (Accessed: 6 May 2025).

Choi, K. *et al.* (2021) 'Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines', *IEEE Access*, 9, pp. 120043–120065. Available at: https://doi.org/10.1109/ACCESS.2021.3107975.

Foroudi, P., Cuomo, M.T. and Foroudi, M.M. (2020) 'Continuance interaction intention in retailing: Relations between customer values, satisfaction, loyalty, and identification', *Information Technology & People*, 33(4), pp. 1303–1326. Available at: https://doi.org/10.1108/ITP-09-2018-0421.

Gazi, Md.A.I. *et al.* (2024) 'The relationship between CRM, knowledge management, organization commitment, customer profitability and customer loyalty in telecommunication industry: The mediating role of customer satisfaction and the moderating role of brand image', *Journal of Open Innovation: Technology, Market, and Complexity*, 10(1), p. 100227. Available at: https://doi.org/10.1016/j.joitmc.2024.100227.

Gligor, D. and Bozkurt, S. (2020) 'FsQCA versus regression: The context of customer engagement', *Journal of Retailing and Consumer Services*, 52, p. 101929. Available at: https://doi.org/10.1016/j.jretconser.2019.101929.

Hajar, M.A. *et al.* (2022) 'The Effect of Value Innovation in the Superior Performance and Sustainable Growth of Telecommunications Sector: Mediation Effect of Customer Satisfaction and Loyalty', *Sustainability*, 14(10), p. 6342. Available at: https://doi.org/10.3390/su14106342.

Homburg, C. and Wielgos, D.M. (2022) 'The value relevance of digital marketing capabilities to firm performance', *Journal of the Academy of Marketing Science*, 50(4), pp. 666–688. Available at: https://doi.org/10.1007/s11747-022-00858-7.

Hossain, M.A., Akter, S. and Yanamandram, V. (2021) 'Why doesn't our value creation payoff: Unpacking customer analytics-driven value creation capability to sustain competitive advantage', *Journal of Business Research*, 131, pp. 287–296. Available at: https://doi.org/10.1016/j.jbusres.2021.03.063.

Jafari, R. (2020) *Customer Churn*, *Kaggle*. Available at: https://www.kaggle.com/datasets/royjafari/customer-churn (Accessed: 5 May 2025).

Jerab, D. (2023) ‘Value, Value Drivers & Valuation: Basics and Principles’. Rochester, NY: Social Science Research Network. Available at: https://doi.org/10.2139/ssrn.4539493.

Kaushik, K. *et al.* (2022) ‘Machine Learning-Based Regression Framework to Predict Health Insurance Premiums’, *International Journal of Environmental Research and Public Health*, 19(13), p. 7898. Available at: https://doi.org/10.3390/ijerph19137898.

Kobi, J. and Otieno, D.B. (2024) ‘Predictive Analytics Applications for Enhanced Customer Retention and Increased Profitability in the Telecommunications Industry’, *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 1762–1774. Available at: https://doi.org/10.38124/ijisrt/IJISRT24MAY1148.

Madhavan, G. (2025) ‘Descriptive Statistics’, in G. Madhavan (ed.) *Mastering Machine Learning: From Basics to Advanced*. Singapore: Springer Nature, pp. 37–52. Available at: https://doi.org/10.1007/978-981-97-9914-5_6.

Manzoor, A. *et al.* (2024) ‘A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners’, *IEEE Access*, 12, pp. 70434–70463. Available at: https://doi.org/10.1109/ACCESS.2024.3402092.

Mazarei, A. *et al.* (2025) ‘Online boxplot derived outlier detection’, *International Journal of Data Science and Analytics*, 19(1), pp. 83–97. Available at: https://doi.org/10.1007/s41060-024-00559-0.

Mukherjee, S. (2024) ‘Verizon uses GenAI to improve customer loyalty’, *Reuters*, 18 June. Available at: https://www.reuters.com/technology/artificial-intelligence/verizon-uses-genai-improve-customer-loyalty-2024-06-18/ (Accessed: 6 May 2025).

Quantzig (2024) *Marketing Mix Strategies in Telecom Industry: A Detailed Case Study*, *quantzig*. Available at: https://www.quantzig.com/case-studies/marketing-mix-models-telecom-industry/ (Accessed: 5 May 2025).

Rais, Z.A. *et al.* (2025) ‘Analysis of Air Pollution in Malaysia: Implications for Environmental Conservation using Granger Causality and Pearson Correlation’, *International Journal of Conservation Science*, 16(1), pp. 149–164. Available at: https://doi.org/10.36868/IJCS.2025.01.09.

Routh, P., Roy, A. and Meyer, J. (2021) ‘Estimating customer churn under competing risks’, *Journal of the Operational Research Society*, 72(5), pp. 1138–1155. Available at: https://doi.org/10.1080/01605682.2020.1776166.

Saha, L. *et al.* (2022) ‘A Machine Learning Model for Personalized Tariff Plan based on Customer’s Behavior in the Telecom Industry’, *International Journal of Advanced Computer Science and Applications*, 13(10). Available at: https://doi.org/10.14569/IJACSA.2022.0131023.

Santy, L. (2025) *Digital Transformation in Telecom: Benefits and 6 Examples*, *Invoca*. Available at: https://www.invoca.com/uk/blog/how-conversation-intelligence-drives-digital-transformation-for-telcos (Accessed: 5 May 2025).

Shah, M. (2024) 'Telecom Loyalty Programs: Rewards And Examples - Guide 2025', 26 December. Available at: https://www.loyaltyxpert.com/blog/telecommunication-loyalty-programs/ (Accessed: 6 May 2025).

Sim, M. *et al.* (2022) 'Customer engagement with service providers: an empirical investigation of customer engagement dispositions', *European Journal of Marketing*, 56(7), pp. 1926–1955. Available at: https://doi.org/10.1108/EJM-12-2020-0879.

Singh, P.P. *et al.* (2024) 'Investigating customer churn in banking: a machine learning approach and visualization app for data science and management', *Data Science and Management*, 7(1), pp. 7–16. Available at: https://doi.org/10.1016/j.dsm.2023.09.002.

Theodorakopoulos, L. and Theodoropoulou, A. (2024) 'Leveraging Big Data Analytics for Understanding Consumer Behavior in Digital Marketing: A Systematic Review', *Human Behavior and Emerging Technologies*, 2024(1), p. 3641502. Available at: https://doi.org/10.1155/2024/3641502.

Thoumy, M. and Abdallah, E. (2017) 'Switching costs impact on customer retention in telecommunication: An exploratory study in Lebanon', *Competition and Regulation in Network Industries*, 18(3–4), pp. 198–216. Available at: https://doi.org/10.1177/1783591718782307.

Tripathy, H.K. *et al.* (2021) 'Amalgamation of Customer Relationship Management and Data Analytics in Different Business Sectors—A Systematic Literature Review', *Sustainability*, 13(9), p. 5279. Available at: https://doi.org/10.3390/su13095279.

Wang, T. (2024) 'Improved random forest classification model combined with C5.0 algorithm for vegetation feature analysis in non-agricultural environments', *Scientific Reports*, 14(1), p. 10367. Available at: https://doi.org/10.1038/s41598-024-60066-x.

7. Appendix

Factorisation

```r
# I will first run the necessary libraries
library(dplyr)
library(readr)
library(tidyverse)

# Since I saved the updated dataset, let's run the updated dataset now
customer <- read_csv("/Users/jawadzaarif7/Desktop/38159/Final/Customer_Churn_Cleaned.csv")

# I will now factorise the categorical features
customer <- customer %>%
  mutate(
    complains = as.factor(complains),
    churn = as.factor(churn),
    age_group = as.factor(age_group),
    tariff_plan = as.factor(tariff_plan),
    status = as.factor(status)
  )

# Done! Let's confirm the changes now
str(customer)

# Summary of factor distributions
summary(customer[, c("complains", "churn", "age_group", "tariff_plan", "status")])
```

Exploratory Data Analysis

```r
# I will now load the psych library for descriptive statistics
library(psych)

# Let's do the descriptive statistics for the numerical features
numeric_summary <- psych::describe(select(customer, where(is.numeric)))
print(numeric_summary)

# Frequency tables for factor variables
cat("\nFrequency tables for categorical variables:\n")

categorical_vars <- c("complains", "churn", "age_group", "tariff_plan", "status")

for (var in categorical_vars) {
  cat("\n", var, ":\n")
  print(table(customer[[var]]))
}

# Time to do the visualisations for numerical and categorical variables!
library(tidyverse)
library(corrplot)
library(ggplot2)
library(dplyr)
library(tidyr)

# Boxplot: Call Failures
boxplot(customer$call._failure,
        xlab = "Number of Call Failures",
        main = "Boxplot: Call Failures",
        horizontal = TRUE)

# Boxplot: Subscription Length
boxplot(customer$subscription_length,
        xlab = "Subscription Length (Months)",
        main = "Boxplot: Subscription Length",
        horizontal = TRUE)

# Boxplot: Charge Amount
boxplot(customer$charge_amount,
        xlab = "Monthly Charge Amount",
        main = "Boxplot: Charge Amount",
        horizontal = TRUE)

# Boxplot: Seconds of Use
boxplot(customer$seconds_of_use,
        xlab = "Seconds of Use",
        main = "Boxplot: Seconds of Use",
        horizontal = TRUE)

# Boxplot: Frequency of Use
boxplot(customer$frequency_of_use,
        xlab = "Frequency of Use",
        main = "Boxplot: Frequency of Use",
        horizontal = TRUE)

# Boxplot: Frequency of SMS
boxplot(customer$frequency_of_sms,
        xlab = "Frequency of SMS",
        main = "Boxplot: Frequency of SMS",
        horizontal = TRUE)
```

```r
# Boxplot: Distinct Called Numbers
boxplot(customer$distinct_called_numbers,
        xlab = "Number of Distinct Called Numbers",
        main = "Boxplot: Distinct Called Numbers",
        horizontal = TRUE)

# Boxplot: Customer Value
boxplot(customer$customer_value,
        xlab = "Customer Value",
        main = "Boxplot: Customer Value",
        horizontal = TRUE)

# Histogram: Call Failures
hist(customer$call._failure,
     main = "Histogram: Call Failures",
     xlab = "Number of Call Failures",
     ylab = "Frequency",
     breaks = 30,
     col = "lightgreen")

# Histogram: Subscription Length
hist(customer$subscription_length,
     main = "Histogram: Subscription Length",
     xlab = "Subscription Length (Months)",
     ylab = "Frequency",
     breaks = 30,
     col = "lightgreen")

# Histogram: Charge Amount
hist(customer$charge_amount,
     main = "Histogram: Charge Amount",
     xlab = "Monthly Charge Amount",
     ylab = "Frequency",
     breaks = 30,
     col = "lightgreen")

# Histogram: Seconds of Use
hist(customer$seconds_of_use,
     main = "Histogram: Seconds of Use",
     xlab = "Seconds of Use",
     ylab = "Frequency",
     breaks = 30,
     col = "lightgreen")

# Histogram: Frequency of Use
hist(customer$frequency_of_use,
     main = "Histogram: Frequency of Use",
     xlab = "Frequency of Use",
     ylab = "Frequency",
     breaks = 30,
     col = "lightgreen")

# Histogram: Frequency of SMS
hist(customer$frequency_of_sms,
     main = "Histogram: Frequency of SMS",
     xlab = "Frequency of SMS",
     ylab = "Frequency",
     breaks = 30,
     col = "lightgreen")
```

```r
# Histogram: Distinct Called Numbers
hist(customer$distinct_called_numbers,
     main = "Histogram: Distinct Called Numbers",
     xlab = "Number of Distinct Called Numbers",
     ylab = "Frequency",
     breaks = 30,
     col = "lightgreen")

# Histogram: Customer Value
hist(customer$customer_value,
     main = "Histogram: Customer Value",
     xlab = "Customer Value",
     ylab = "Frequency",
     breaks = 30,
     col = "lightgreen")

# Pair Plot for Key Numerical Variables
pairs(formula = ~call._failure + subscription_length + charge_amount + seconds_of_use +
        frequency_of_use + frequency_of_sms + distinct_called_numbers + customer_value,
      data = customer,
      cex = 0.15,
      main = "Pair Plot")

# Correlation Matrix and Heatmap
cor_matrix <- cor(customer[, c('call._failure', 'subscription_length', 'charge_amount', 'seconds_of_use',
                               'frequency_of_use', 'frequency_of_sms',
                               'distinct_called_numbers', 'customer_value')],
                  use = "complete.obs")

corrplot(cor_matrix,
         method = 'color',
         addCoef.col = "blue",
         tl.col = "black",
         title = "Correlation Heatmap",
         mar = c(0, 0, 2, 0))  # Add top margin for the title

# Bar Plot: Complains
ggplot(customer, aes(x = complains)) +
  geom_bar(fill = "yellow") +
  labs(title = "Bar Plot: Complains",
       x = "Complains (0 = No, 1 = Yes)",
       y = "Frequency") +
  theme_minimal()

# Bar Plot: Churn
ggplot(customer, aes(x = churn)) +
  geom_bar(fill = "yellow") +
  labs(title = "Bar Plot: Churn",
       x = "Churn (0 = No, 1 = Yes)",
       y = "Frequency") +
  theme_minimal()

# Bar Plot: Age Group
ggplot(customer, aes(x = age_group)) +
  geom_bar(fill = "yellow") +
  labs(title = "Bar Plot: Age Group",
       x = "Age Group",
       y = "Frequency") +
  theme_minimal()
```

```r
# Bar Plot: Tariff Plan
ggplot(customer, aes(x = tariff_plan)) +
  geom_bar(fill = "yellow") +
  labs(title = "Bar Plot: Tariff Plan",
       x = "Tariff Plan (1 = Basic Plan, 2 = Premium Plan)",
       y = "Frequency") +
  theme_minimal()

# Bar Plot: Status
ggplot(customer, aes(x = status)) +
  geom_bar(fill = "yellow") +
  labs(title = "Bar Plot: Customer Status",
       x = "Status (1 = Active, 2 = Inactive)",
       y = "Frequency") +
  theme_minimal()
```

Filtering

```r
# First I will load the necessary libraries
library(dplyr)
library(tidyverse)

# Let's run the dataset now
dataset <- read.csv("/Users/jawadzaarif7/Desktop/38159/Final/Customer Churn.csv", stringsAsFactors = FALSE)

# Now I will check the number of columns and rows
cat("Number of rows:", nrow(dataset), "\n")
cat("Number of columns:", ncol(dataset), "\n")

# Time to check for duplicate rows
duplicate_count <- sum(duplicated(dataset))
cat("Number of duplicate rows:", duplicate_count, "\n")

# I did find duplicate values, time to check for missing values
missing_values <- colSums(is.na(dataset))
cat("Missing values per column:\n")
print(missing_values)

# Since there were no missing values, lets remove the duplicate rows now
dataset_cleaned <- dataset %>% distinct()

# Let's check the updated number of rows and columns now
cat("After removing duplicates:\n")
cat("Number of rows:", nrow(dataset_cleaned), "\n")
cat("Number of columns:", ncol(dataset_cleaned), "\n")

# Time to confirm if the duplicates were removed or not
duplicate_count_cleaned <- sum(duplicated(dataset_cleaned))
cat("Number of duplicate rows:", duplicate_count_cleaned, "\n")

# Just to be sure, I will again check in case any values turned into null
missing_values_cleaned <- colSums(is.na(dataset_cleaned))
cat("Missing values per column:\n")
print(missing_values_cleaned)
```

Data Pre-processing

```
# Let's now convert tariff_plan and status from (1 = 0, 2 = 1)
customer <- customer %>%
  mutate(
    tariff_plan = ifelse(tariff_plan == 1, 0, 1),
    status = ifelse(status == 1, 0, 1)
  )
# Read the updated cells
head(customer[, c("tariff_plan", "status")])
```

Regression without Interaction

```
# Load Packages
install.packages('gridExtra')
library(gridExtra)
install.packages('emmeans')
library(emmeans)

# Plot the regression chart for Seconds of Use VS Customer Value
ggplot(customer, aes(y = customer_value, x = seconds_of_use)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Seconds of Use", y = "Customer Value") +
  ggtitle("Customer Value vs Seconds of Use")

# Plot the regression chart for Subscription Length VS Customer Value
ggplot(customer, aes(y = customer_value, x = subscription_length)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Subscription Length", y = "Customer Value") +
  ggtitle("Customer Value vs Subscription Length")

# Plot the regression chart for Frequency of SMS VS Customer Value
ggplot(customer, aes(y = customer_value, x = frequency_of_sms)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Frequency of SMS", y = "Customer Value") +
  ggtitle("Customer Value vs Frequency of SMS")
```

```r
# Model the regression line
model1.reg <- lm(customer_value ~ seconds_of_use, data = customer)
summary(model1.reg)

model1.reg <- lm(customer_value ~ subscription_length, data = customer)
summary(model1.reg)

model1.reg <- lm(customer_value ~ frequency_of_sms, data = customer)
summary(model1.reg)

model1.reg <- lm(customer_value ~ seconds_of_use + subscription_length + frequency_of_sms, data = customer)
summary(model1.reg)

cbind(coef(model1.reg), confint(model1.reg))

# Create quantile-based bins for the two variables used for coloring
seconds_use.bins <- quantile(pull(customer, seconds_of_use))
customer <- customer %>%
  mutate(seconds_use.bin = cut(seconds_of_use, seconds_use.bins, include.lowest = TRUE))

sms_freq.bins <- quantile(pull(customer, frequency_of_sms))
customer <- customer %>%
  mutate(sms_freq.bin = cut(frequency_of_sms, sms_freq.bins, include.lowest = TRUE))

# Plot 1: Seconds of Use vs Customer Value
ggplot(customer, aes(x = seconds_of_use, y = customer_value, col = sms_freq.bin)) +
  geom_point() +
  geom_smooth(method = lm) +
  geom_smooth(mapping = aes(col = NULL), method = lm, col = "black") +
  labs(x = "Seconds of Use", y = "Customer Value", col = "Frequency of SMS") +
  ggtitle("Customer Value vs Seconds of Use")

# Plot 2: Frequency of SMS vs Customer Value
ggplot(customer, aes(x = frequency_of_sms, y = customer_value, col = seconds_use.bin)) +
  geom_point() +
  geom_smooth(method = lm) +
  geom_smooth(mapping = aes(col = NULL), method = lm, col = "black") +
  labs(x = "Frequency of SMS", y = "Customer Value", col = "Seconds of Use") +
  ggtitle("Customer Value vs Frequency of SMS")
```

Regression with Interaction

```
# Model the regression line with tariff_plan as independent variable
model2.reg <- lm(customer_value~seconds_of_use +
                   subscription_length +
                   frequency_of_sms +
                   tariff_plan, data = customer)
summary(model2.reg)

# Model the regression with the interaction
interaction.reg <- lm(customer_value~seconds_of_use +
                        subscription_length +
                        frequency_of_sms +
                        tariff_plan +
                        seconds_of_use:tariff_plan +
                        subscription_length:tariff_plan +
                        frequency_of_sms:tariff_plan, data = customer)
summary(interaction.reg)
```

Classification

```
customer <- data.frame(customer)  # Ensure it's a dataframe

# Description and summary
str(customer)
summary(customer)

# Look at the class variable
table(customer$churn)

# Load C50 package
install.packages("C50")
library(C50)


# Build the simplest decision tree
# Exclude customer_id (assumed to be the 1st column)
churn_model <- C5.0(customer[,-c(1,14,15,16)], customer$churn)  # Exclude customer_id and churn from predictors

# Display simple facts about the tree
churn_model

# Display detailed information
summary(churn_model)

# Plot the results
plot(churn_model)

# Limit the tree size and try again
churn_model2 <- C5.0(customer[,-c(1,14,15,16)], customer$churn,
                     control = C5.0Control(minCases = 100))  # Exclude customer_id and churn from predictors

# Display tree summary
churn_model2
summary(churn_model2)

# Plot the simplified tree
plot(churn_model2)

### END ###
```