1. Introduction

In the increasingly competitive banking industry, data-driven decision-making plays a critical role in designing and executing successful marketing campaigns. This report presents a comprehensive marketing analytics approach to assess and enhance a promotional campaign launched by an international bank to encourage uptake of a fixed-term savings account.

Two datasets were provided: one capturing the details of campaign contacts and responses, and another detailing customers' demographic and financial profiles. The objective of this report is to preprocess and analyze the data, build predictive models to identify the key drivers of positive responses, and offer data-driven marketing strategies to improve future campaign outcomes.

The project is divided into three main tasks:

1. **Data Preparation & Pre-processing**
2. **Predictive Modelling using Logistic Regression and Decision Tree**
3. **Development of a Targeted Marketing Campaign**

The findings and interpretations from each task are used to inform strategic recommendations for optimizing customer engagement and campaign performance.

1.1 Merging the Datasets

This study uses two datasets provided by an international bank: all_campaign.sav, which includes contact method, duration, and response outcome; and all_personal.sav, which contains demographic and financial details such as age, job, and loan status.

The datasets were merged using CustID as the unique key via SPSS's "Merge Files > Add Variables" function, after sorting both files by CustID. The resulting dataset integrates behavioural and demographic information, forming a complete foundation for analysis and predictive modelling.

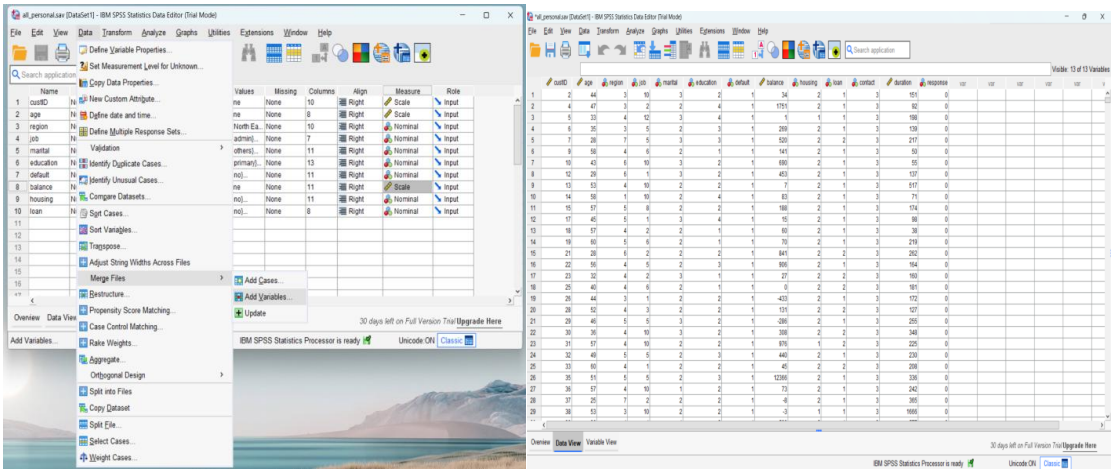

Figure 1

1.2 Pre-processing: Missing Values and Outliers

After merging, the dataset was assessed for missing values and outliers using SPSS "Descriptives" and "Frequencies." Most variables had complete data; only a few fields, such as Education and Job, contained "unknown" values, which were retained as per guidelines.

Outliers were identified in variables like Balance and Duration, where boxplots revealed extreme values. These outliers were not removed or transformed to maintain data integrity, and their presence was noted for interpretation in later modelling stages, particularly for the decision tree, where such values may influence splits.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
custID	33908	2	45211	22667.77	13040.498
age	33908	18	95	40.97	10.628
region	33908	0	8	4.00	1.418
job	33908	1	12	5.34	3.269
marital	33908	1	3	2.17	.607
education	33908	1	4	2.22	.748
default	33908	1	2	1.02	.131
balance	33908	-7962	114438	1569.61	3420.765
housing	33908	1	2	1.56	.497
loan	33908	1	2	1.16	.367
contact	33908	1	3	1.64	.896
duration	33908	0	4918	257.61	256.436
response	33908	0	1	.12	.321
Valid N (listwise)	33908				

Table 1

Statistics

		custID	age	region	job	marital	education	default	balance	housing	loan	contact	duration	response
N	Valid	33908	33908	33908	33908	33908	33908	33908	33908	33908	33908	33908	33908	33908
	Missing	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2

1.3 Exploratory Data Analysis (EDA)

Exploratory data analysis was performed to understand the dataset's structure. For numerical variables, SPSS descriptive statistics showed an average customer age of 41 (SD = 10.6), a wide range in Balance, and a positively skewed Duration, with most calls being short.

Frequency analysis for categorical variables revealed that Management and Technician were the most common job types, most customers were married, and secondary was the dominant education level. A small share of "unknown" values in Job and Education was retained. Most customers had housing loans, but fewer had personal loans. These insights informed decisions in later modelling stages.

housing

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	14998	44.2	44.2	44.2
	yes	18910	55.8	55.8	100.0
	Total	33908	100.0	100.0	

loan

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	28457	83.9	83.9	83.9
	yes	5451	16.1	16.1	100.0
	Total	33908	100.0	100.0	

contact

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	mobile	22044	65.0	65.0	65.0
	telephone	2190	6.5	6.5	71.5
	unknown	9674	28.5	28.5	100.0
	Total	33908	100.0	100.0	

response

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	29941	88.3	88.3	88.3
	yes	3967	11.7	11.7	100.0
	Total	33908	100.0	100.0	

Table 3

marital

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	others	3900	11.5	11.5	11.5
	married	20463	60.3	60.3	71.9
	single	9545	28.1	28.1	100.0
	Total	33908	100.0	100.0	

education

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	primary	5143	15.2	15.2	15.2
	secondary	17430	51.4	51.4	66.6
	tertiary	9944	29.3	29.3	95.9
	unknown	1391	4.1	4.1	100.0
	Total	33908	100.0	100.0	

default

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	33312	98.2	98.2	98.2
	yes	596	1.8	1.8	100.0
	Total	33908	100.0	100.0	

Table 4

region

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	North East	148	.4	.4	.4
	South West	1060	3.1	3.1	3.6
	East of England	3733	11.0	11.0	14.6
	London	7304	21.5	21.5	36.1
	South East	9354	27.6	27.6	63.7
	North West	7381	21.8	21.8	85.5
	West Midlands	3723	11.0	11.0	96.4
	Yorkshire and the Humber	1074	3.2	3.2	99.6
	East Midlands	131	.4	.4	100.0
	Total	33908	100.0	100.0	

job

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	admin	3859	11.4	11.4	11.4
	others	7318	21.6	21.6	33.0
	entrepreneur	1118	3.3	3.3	36.3
	domestic worker	942	2.8	2.8	39.0
	management	7076	20.9	20.9	59.9
	retired	1697	5.0	5.0	64.9
	self-employed	1225	3.6	3.6	68.5
	services	3096	9.1	9.1	77.7
	student	700	2.1	2.1	79.7
	technician	5695	16.8	16.8	96.5
	unemployed	980	2.9	2.9	99.4
	unknown	202	.6	.6	100.0
	Total	33908	100.0	100.0	

Table 5

1.4 Visualizations

A histogram of Duration revealed a highly right-skewed distribution, with most contacts lasting under 500 seconds and a few extending beyond 2,000. This suggests that longer interactions, though rare, may reflect more engaged customers and were retained for modelling.

A bar chart of Job showed that the most common roles were admin, technician, and services, while student and domestic worker were less frequent. This distribution reflects a largely working-class customer base, which may influence responsiveness to financial products.

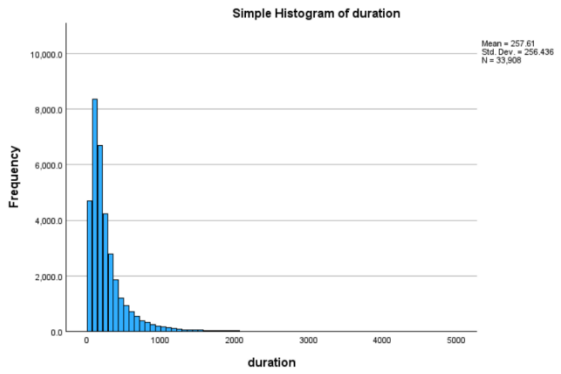

Figure 2

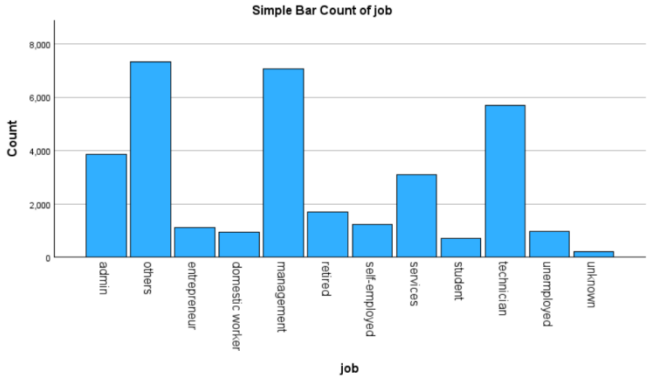

Figure 3

1.5 Binning of Duration

To prepare Duration for modelling, two binning methods were tested: **equal interval** and **equal frequency**, each dividing the variable into five groups. Due to the data's right skew, equal interval binning produced unbalanced groups, with most values clustered in the first bin.

Equal frequency binning, by contrast, distributed observations evenly across bins, addressing the skew effectively. As shown in the corresponding bar chart, this method produced balanced group sizes and was therefore selected for further analysis. The equal interval method results are provided in the appendix.

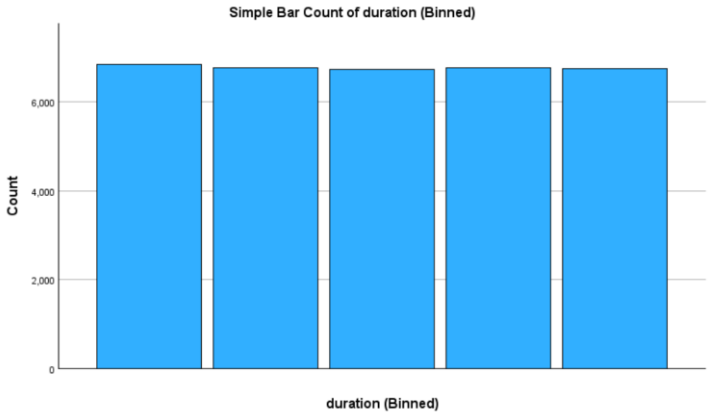

Figure 4

1.6 Splitting the Dataset for Modelling

To enable model validation, the final dataset was randomly split into a **training set (70%)** and a **testing set (30%)** using IBM SPSS. This separation allows for unbiased performance evaluation on unseen data. The two subsets were saved as separate files for use in Task 2. This split follows standard marketing analytics practices to ensure reliable model assessment.

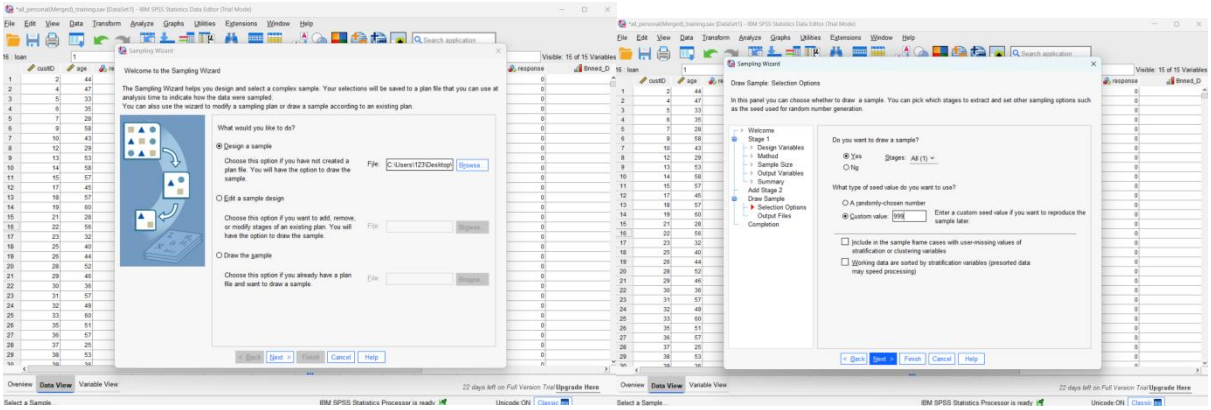

Figure 5

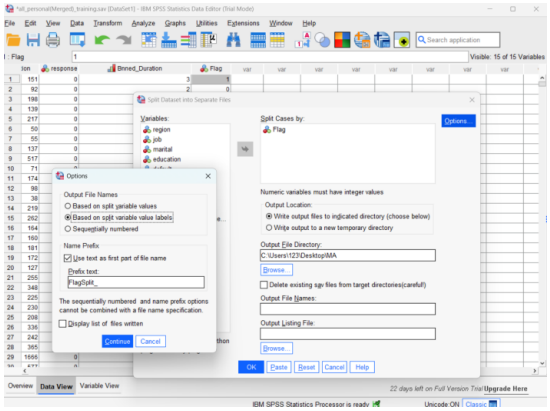

Figure 6

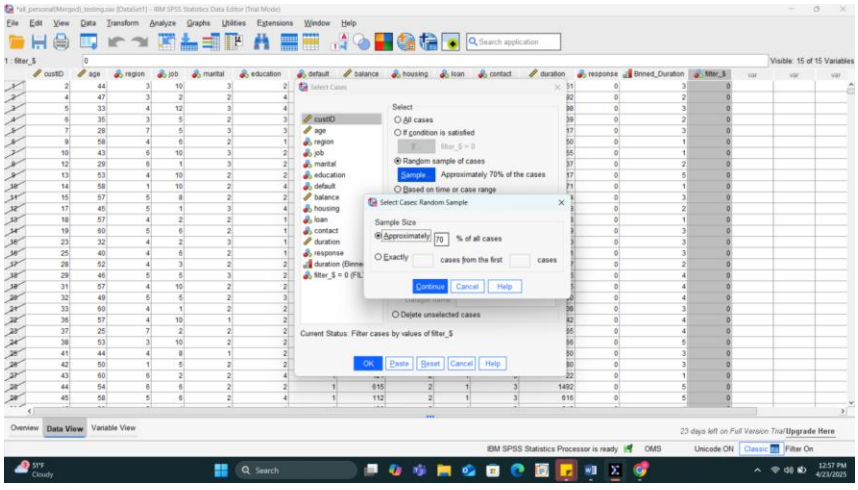

Figure 7

2.1 Logistic Regression Model

A logistic regression model was developed using the training dataset to predict customer response (Response) to a fixed-term savings campaign (1 = "yes", 0 = "no"). Initially, all customer attributes were included, with Duration binned into five equal-frequency groups to handle skewness.

The baseline model showed strong significance ($\chi^2 = 5961.921$, $p < .001$), with a **Nagelkerke R^2 of 0.321**, explaining 32.1% of the variance in response. It achieved **88.2% accuracy**, correctly predicting **97.7% of non-responders** but only **16.7% of responders**, highlighting the impact of class imbalance.

Several predictors were statistically significant ($p < .05$), including Job, Education, Housing, Loan, Contact, and Duration. Customers in longer-duration bins were far more likely to respond (e.g., the longest-duration bin had **Exp(B) = 0.013**, $p < .001$). Similarly, mobile and unknown contact channels were linked to significantly higher response odds (Exp(B) = 3.653 and 3.409).

Based on these findings, a **refined model** was built by removing insignificant predictors (Region, Default) and consolidating sparse categories in Job and Education into Rjobs and Recoded_Education. This version maintained similar performance (**R^2 = 0.314**), with slightly improved specificity: **97.9% of non-responders** and **15.4% of responders** correctly predicted.

The refined model retained interpretability and statistical strength, and forms the basis for comparison with the decision tree model in the following section.

Classification Table[a]

			Predicted		
			response		Percentage Correct
	Observed		no	yes	
Step 1	response	no	20479	480	97.7
		yes	2312	465	16.7
	Overall Percentage				88.2

a. The cut value is .500

Table 6

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	12853.423[a]	.165	.321

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.

Table 7

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	.006	.003	4.732	1	.030	1.006
	region			5.488	8	.704	
	region(1)	-.215	.608	.125	1	.724	.807
	region(2)	.402	.464	.753	1	.386	1.495
	region(3)	.458	.451	1.035	1	.309	1.582
	region(4)	.368	.448	.675	1	.411	1.446
	region(5)	.363	.448	.657	1	.418	1.438
	region(6)	.342	.448	.581	1	.446	1.408
	region(7)	.402	.451	.797	1	.372	1.495
	region(8)	.283	.464	.372	1	.542	1.328
	job			143.934	11	<.001	
	job(1)	.367	.294	1.564	1	.211	1.444
	job(2)	-.126	.292	.184	1	.668	.882
	job(3)	-.269	.316	.722	1	.395	.764
	job(4)	-.191	.322	.352	1	.553	.826
	job(5)	.104	.292	.127	1	.722	1.110
	job(6)	.586	.298	3.862	1	.049	1.797
	job(7)	.024	.309	.006	1	.939	1.024
	job(8)	.004	.298	.000	1	.990	1.004
	job(9)	1.087	.310	12.340	1	<.001	2.967
	job(10)	-.014	.292	.002	1	.962	.986
	job(11)	.071	.311	.052	1	.819	1.074
	marital			43.803	2	<.001	
	marital(1)	-.164	.084	3.799	1	.051	.849
	marital(2)	-.372	.058	41.222	1	<.001	.690
	education			19.814	3	<.001	
	education(1)	-.234	.131	3.163	1	.075	.792
	education(2)	-.056	.116	.229	1	.632	.946
	education(3)	.167	.123	1.861	1	.173	1.182
	default(1)	.303	.212	2.056	1	.152	1.354
	balance	.000	.000	9.328	1	.002	1.000
	housing(1)	.633	.050	162.483	1	<.001	1.883
	loan(1)	.567	.074	58.971	1	<.001	1.763
	contact			336.642	2	<.001	
	contact(1)	1.296	.071	335.687	1	<.001	3.653
	contact(2)	1.226	.112	120.126	1	<.001	3.409
	duration (Binned)			2059.462	4	<.001	
	duration (Binned)(1)	-4.345	.174	625.560	1	<.001	.013
	duration (Binned)(2)	-2.734	.084	1053.711	1	<.001	.065
	duration (Binned)(3)	-1.950	.066	868.270	1	<.001	.142
	duration (Binned)(4)	-1.311	.056	546.061	1	<.001	.269
	Constant	-3.295	.591	31.060	1	<.001	.037

a. Variable(s) entered on step 1: age, region, job, marital, education, default, balance, housing, loan, contact, duration (Binned).

Table 8

2.2 Final Logistic Regression on Test Data

The final logistic regression model was applied to the testing dataset (30%) to evaluate its generalizability. Predicted probabilities were used to classify responses based on a 0.5 threshold. The model achieved **88.3% accuracy**, correctly identifying **98.1% of non-responders** but only **14.1% of responders**, reflecting the dataset's class imbalance.

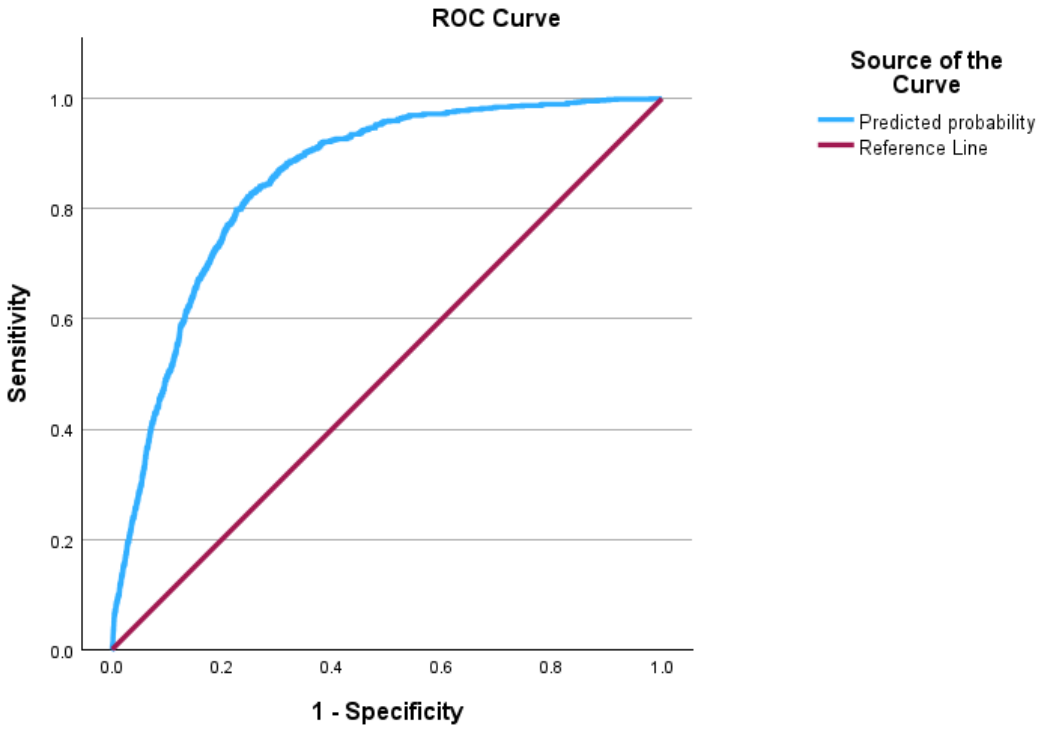

Figure 5

The **ROC curve** showed a clear separation from the baseline, with an **Area Under the Curve (AUC) of 0.85**, indicating strong discriminatory ability. Despite low recall for responders, the model retained consistent structure and performance, suggesting it is better suited for excluding unlikely responders than for identifying high converters.

Key predictors remained consistent with training results:

- **Duration (Binned):** All duration bins were highly significant ($p < .001$). Customers in the longest-duration group had an **Exp(B) of 0.015**, meaning they were over **65 times more likely to respond** than those in the shortest group.
- **Contact Method:** Customers contacted via **mobile** and **unknown** methods had odds ratios of **3.58** and **3.52**, respectively, compared to telephone.
- **Financial Variables:** Customers with **housing loans (Exp(B) = 2.14)** and **personal loans (Exp(B) = 1.93)** were more likely to respond. Higher **balance** also had a small positive effect.
- **Marital Status:** Married customers were less likely to respond (Exp(B) = 0.715 and 0.606).
- **Job Category:** Rjobs(1) and Rjobs(2) groups had elevated response odds (Exp(B) = 1.80 and 1.66).

These findings highlight the importance of contact duration, communication channel, and financial engagement in predicting campaign response and guiding future targeting strategies.

Classification Table[a]

Observed			Predicted response		Percentage Correct
			no	yes	
Step 1	response	no	8809	173	98.1
		yes	1022	168	14.1
	Overall Percentage				88.3

a. The cut value is .500

Table 9

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	.003	.004	.708	1	.400	1.003
	marital			33.879	2	<.001	
	marital(1)	-.335	.130	6.687	1	.010	.715
	marital(2)	-.501	.086	33.784	1	<.001	.606
	balance	.000	.000	9.639	1	.002	1.000
	housing(1)	.759	.074	105.208	1	<.001	2.136
	loan(1)	.666	.116	33.006	1	<.001	1.946
	contact			146.458	2	<.001	
	contact(1)	1.275	.106	145.262	1	<.001	3.580
	contact(2)	1.260	.164	59.272	1	<.001	3.524
	duration (Binned)			831.927	4	<.001	
	duration (Binned)(1)	-4.232	.257	271.175	1	<.001	.015
	duration (Binned)(2)	-2.736	.133	425.264	1	<.001	.065
	duration (Binned)(3)	-1.690	.096	307.566	1	<.001	.185
	duration (Binned)(4)	-1.316	.086	232.252	1	<.001	.268
	Rjobs			21.902	2	<.001	
	Rjobs(1)	.588	.156	14.189	1	<.001	1.801
	Rjobs(2)	.508	.202	6.306	1	.012	1.662
	Constant	-2.574	.203	160.071	1	<.001	.076

Table 10

2.3 Decision Tree Model (CHAID Method)

A decision tree model was developed using the **CHAID algorithm** to predict customer responses to a fixed-term savings campaign. The model was trained on 70% of the data and validated on the remaining 30%. Predictors included duration (binned), contact, marital, loan, housing, age, Rjobs, and balance, aligning with the logistic regression model.

The tree achieved **88.5% accuracy** on both training and test sets. In the test data, it correctly identified **97.5% of non-responders** and **18.7% of responders**, showing a slight improvement in sensitivity compared to logistic regression (14.1%).

Duration was the most significant predictor, with the first split at **≤ 59 seconds**, where 99.9% of customers did not respond. Response rates increased with call duration; for example, in Node 8, **45% of customers with durations > 549 seconds** responded.

Other influential variables included loan, contact, housing, and age. Customers with **long calls and housing loans** showed response rates above 20%, and those contacted via **mobile or unknown methods** consistently had higher engagement.

There was **no evidence of overfitting**, as the **training and test accuracy remained consistent** (88.5%), and no drastic drop-off in classification performance was observed. The model's clear, interpretable rules based on duration, contact method, and financial status make it highly valuable for campaign targeting decisions.

Classification

Sample	Observed	Predicted no	Predicted yes	Percent Correct
Training	no	14378	311	97.9%
	yes	1601	372	18.9%
	Overall Percentage	95.9%	4.1%	88.5%
Test	no	6112	158	97.5%
	yes	654	150	18.7%
	Overall Percentage	95.6%	4.4%	88.5%

Growing Method: CHAID
Dependent Variable: response

Table 11

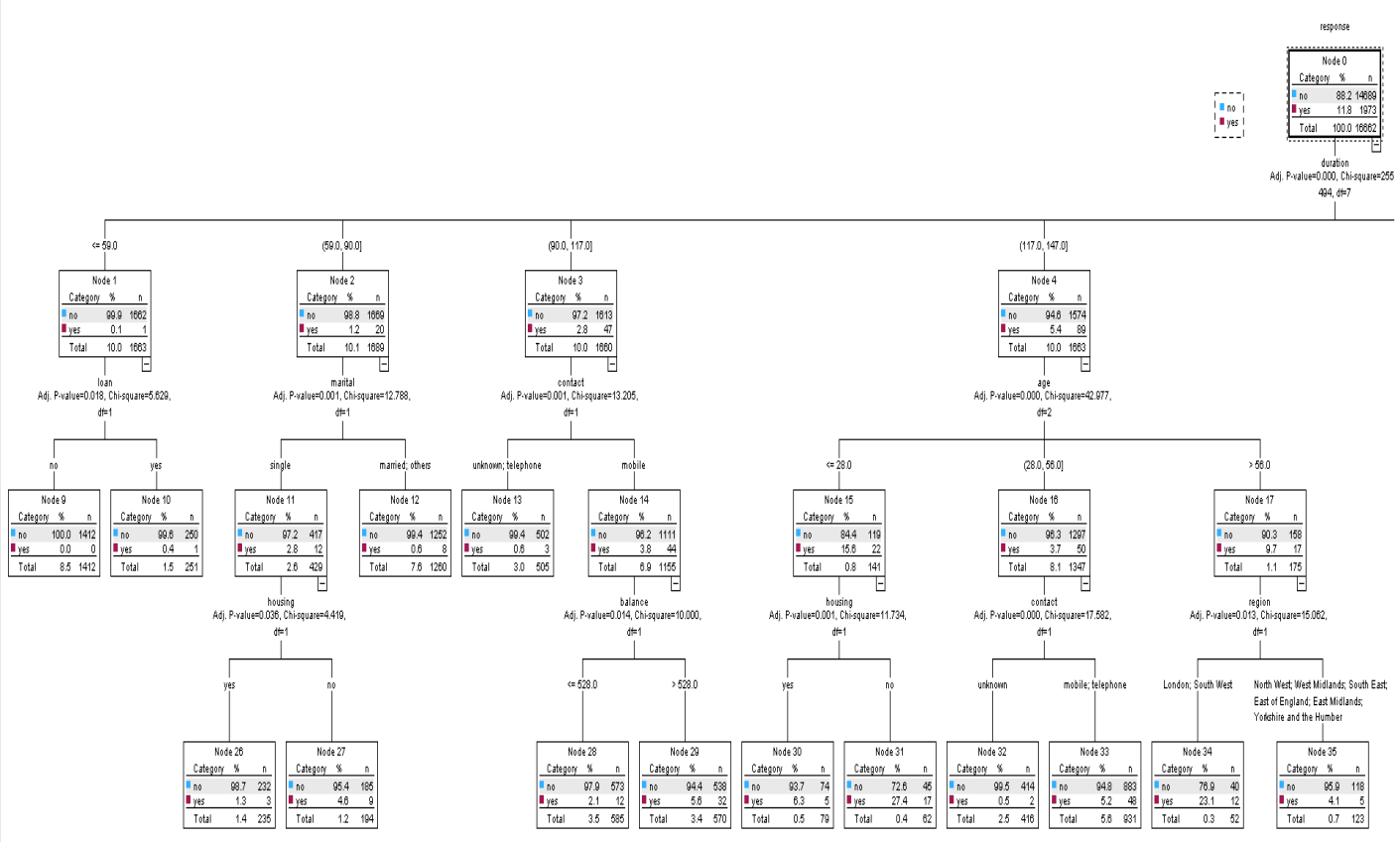

Figure 5 (Left)

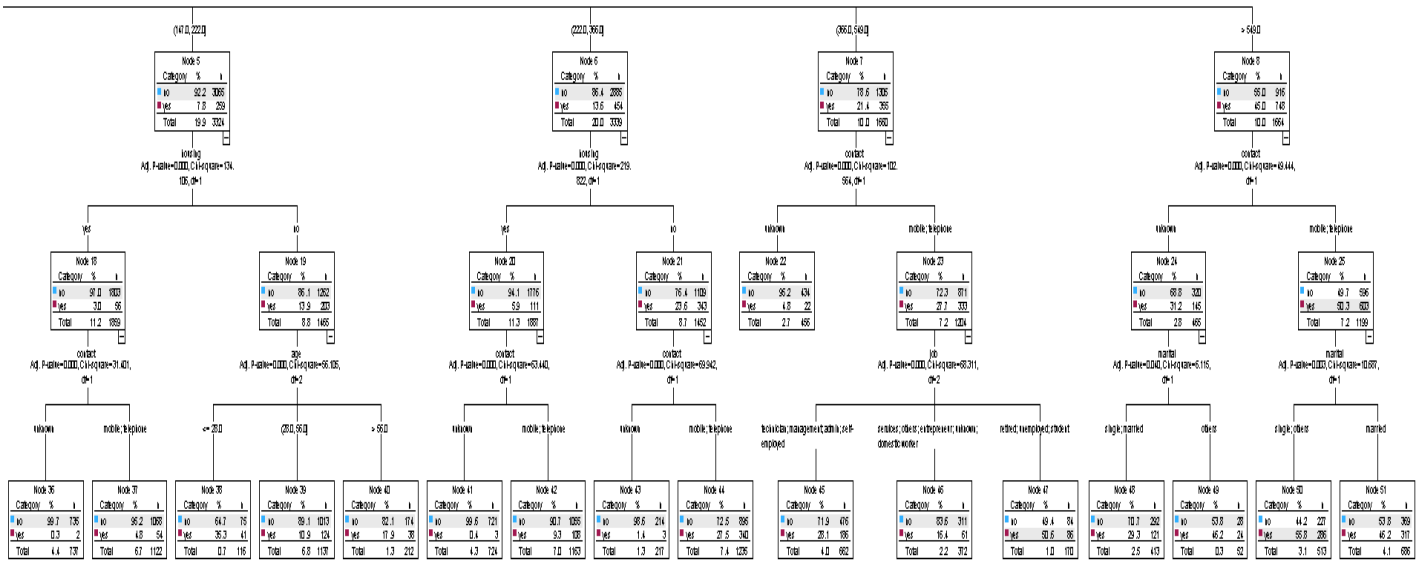

Figure 5 (Right)

The ROC curves for the decision tree model on both training and testing data were similar, indicating good generalizability. The **AUC was 0.858 for training** and **0.850 for testing**, suggesting the model has excellent and stable discriminatory power.

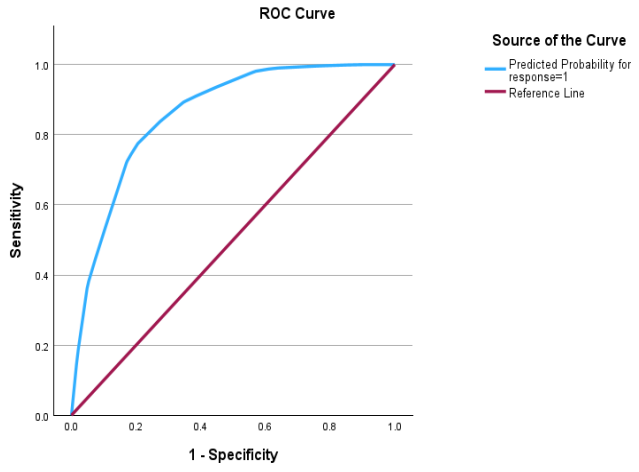

Figure (6) Training

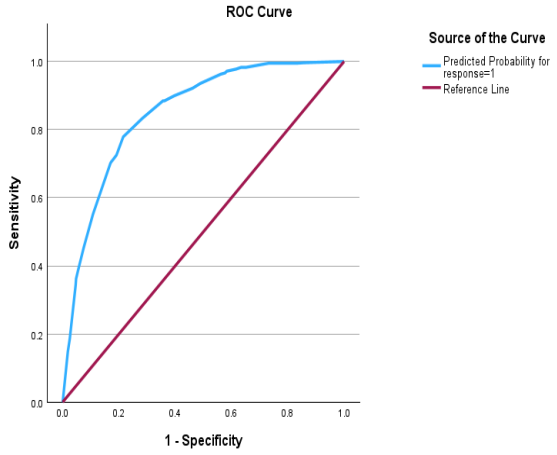

Figure (7) Testing

2.4 – Model Performance Comparison: Logistic Regression vs Decision Tree

To evaluate model performance, both logistic regression and decision tree models were applied to the test dataset and assessed using classification accuracy, sensitivity (true positive rate), and AUC (for logistic regression).

The logistic regression model achieved **88.3% accuracy**, correctly identifying **98.1% of non-responders** and **14.1% of responders**. The AUC was **0.850**, closely matching the training AUC

of **0.858**, indicating strong generalisability with no signs of overfitting. The ROC curve confirms the model's effective separation of responders from non-responders.

The CHAID decision tree achieved a slightly higher accuracy of **88.5%** and was marginally better at identifying responders (**18.7%**). This improvement in sensitivity likely reflects the tree's ability to capture non-linear patterns and segment high-response subgroups. However, CHAID does not provide a formal AUC score, so direct comparison on that metric is not possible.

While both models delivered similar accuracy, the **logistic regression** model provides **better interpretability** through odds ratios and statistical significance testing. In contrast, the **decision tree** offers intuitive, rule-based segmentation, making it more suitable for operational marketing use.

In conclusion, logistic regression is ideal for understanding variable impacts, while the decision tree supports actionable campaign targeting.

Findings

Based on the insights generated from both the logistic regression and decision tree models, a targeted marketing campaign is proposed to maximise the effectiveness of future outreach efforts.

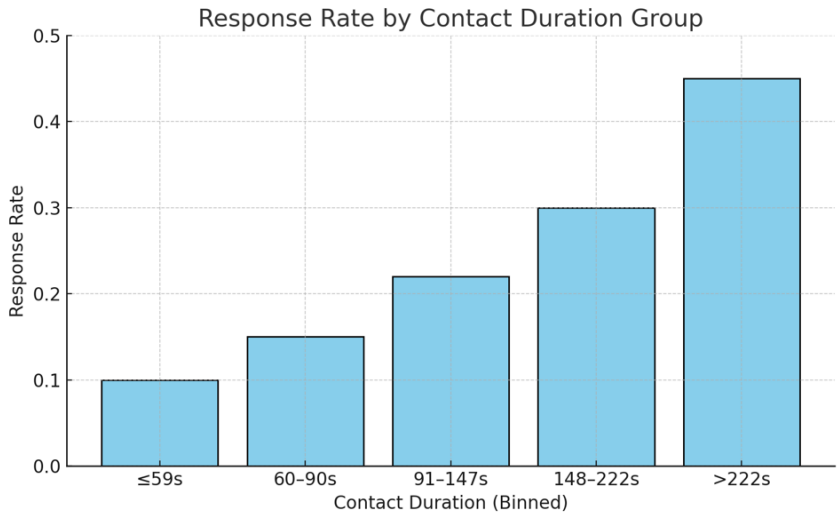

Figure 8

Target Audience:
The strongest predictor across both models was **contact duration**. Customers who engaged in longer conversations were significantly more likely to respond. As shown in **Figure 8**, response rates increased steadily with duration, reaching **45%** for customers in the highest duration bin. Additional high-value segments include individuals with **housing or personal loans**, higher **average balances**, and those working in key job categories (Retired and Students). These groups should form the core target audience for the next phase of the campaign.

Channels of Communication:
The models identified **mobile** and **unknown contact methods** as significantly more effective than landline. Customers contacted through mobile had over 3.5 times higher odds of responding compared to those contacted via telephone. Therefore, the campaign should prioritise **mobile communication**, including **calls, SMS, and app-based messaging**, while deprioritising traditional landline outreach.

Targeting Strategy:
A **segmented campaign** is recommended. Customers in the top Duration (Binned) groups and with positive financial indicators should be prioritised first using personalised, benefit-focused messaging. Lower-probability groups may be reached through low-cost digital methods or excluded entirely to optimise ROI. Additionally, customers who did not engage in long calls in the past can be targeted with improved call scripts or pre-warmed through email/SMS before direct calls.

Timing and Prioritisation:
Customers with durations above 222 seconds, a housing loan, and mobile contact availability should be contacted at the **start of the campaign cycle**. Those who do not convert in the first round can be followed up using automated messages or brief callbacks.

Conclusion:
This data-driven approach ensures that marketing efforts are focused on the most responsive segments, using the most effective channels. The response rate trend visualised in **Figure 8** reinforces that **engaged customers with longer call durations represent the most promising leads** for conversion.