MBAS905 (T125) Advanced Business Analytics

MBAS905 T1 2025 Assignment 3

Student name: Huynh Quang Tri Nguyen

Student ID: 8072826

Coordinator and Lecturer: Dr Edmund (James) Brownlow

Table of Contents

1. Introduction

Banks encounter escalating pressure to transform prospects into enduring customers via data-driven marketing strategies in a progressively competitive financial environment. Despite their low risk, term deposit products frequently experience low conversion rates attributable to customer reluctance, economic volatility, and misaligned campaign timing. The primary challenge is ascertaining which customers are most inclined to subscribe and comprehend the behavioural, contextual, and economic factors affecting their decisions. Although customer interaction data is accessible, numerous institutions find it challenging to convert this into actionable insights. This report fills the gap by analysing a real-world banking dataset through descriptive, inferential, and predictive analytics. The objective is to identify essential subscription patterns, evaluate their statistical significance, and develop a classification model to forecast future customer behaviour. The study establishes a basis for refining targeting strategies, optimising campaign efficiency, and synchronising marketing interventions with the appropriate audience at the opportune moment.

2. Methodology

This study employed a systematic analytical approach utilising descriptive, inferential, and predictive methodologies to examine customer behaviour regarding term deposit subscriptions. The dataset from a Portuguese bank comprised 40,188 customer records and encompassed both categorical and numerical variables.

The initial phase utilised descriptive analytics to examine data trends, distributions, and correlations. Visualisations, including bar charts and correlation matrices, provided initial insights, such as elevated subscription rates among administrative positions (3.3%) and maximum campaign responsiveness in May. These observations informed the selection of variables for subsequent statistical analysis.

The second stage utilised inferential statistics to ascertain the statistical significance of observed patterns. A chi-squared test assessed the relationship between job type and term deposit, while an independent samples t-test analysed the variance in campaign contacts between subscriber groups.

The final stage involved conducting predictive modelling utilising SAS Viya's Model Studio. Logistic Regression, Decision Tree, and Random Forest algorithms were evaluated for comparison. The Forest model was chosen due to its superior AUC (0.8025) and F1-score (0.48), and it was employed to assess a holdout dataset for probability-based segmentation. This multi-faceted strategy guaranteed analytical precision and practical insights for campaign enhancement.

3. Descriptive Analysis

Figure 3.1: Term Deposit Subscription Rates by Occupation Group

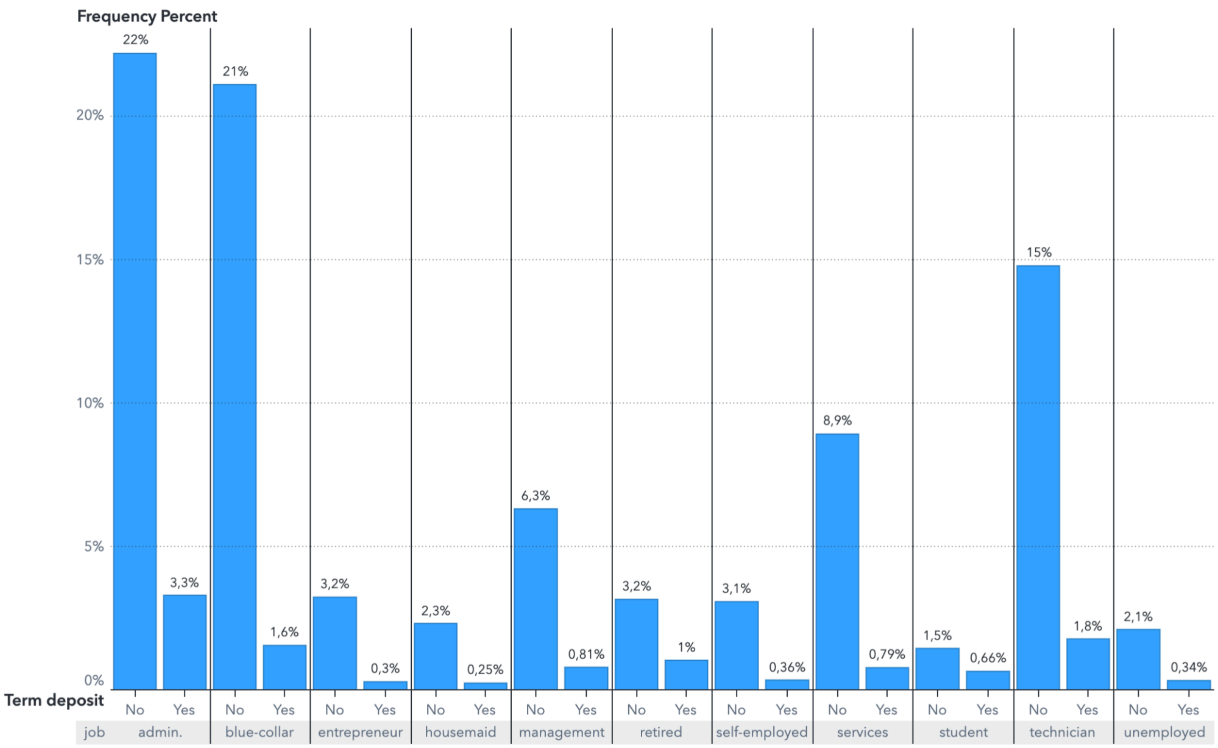


Figure 3.1: Term Deposit Subscription Rates by Occupation Group

While most occupations have low term deposit uptake, administrative (3.3%) and technician (1.8%) roles stand out with higher subscription rates. In contrast, entrepreneurs (0.3%) and housemaids (0.25%) show minimal engagement. Despite a higher percentage of "No" in all groups, these differences suggest that professional stability and information access may drive responsiveness.

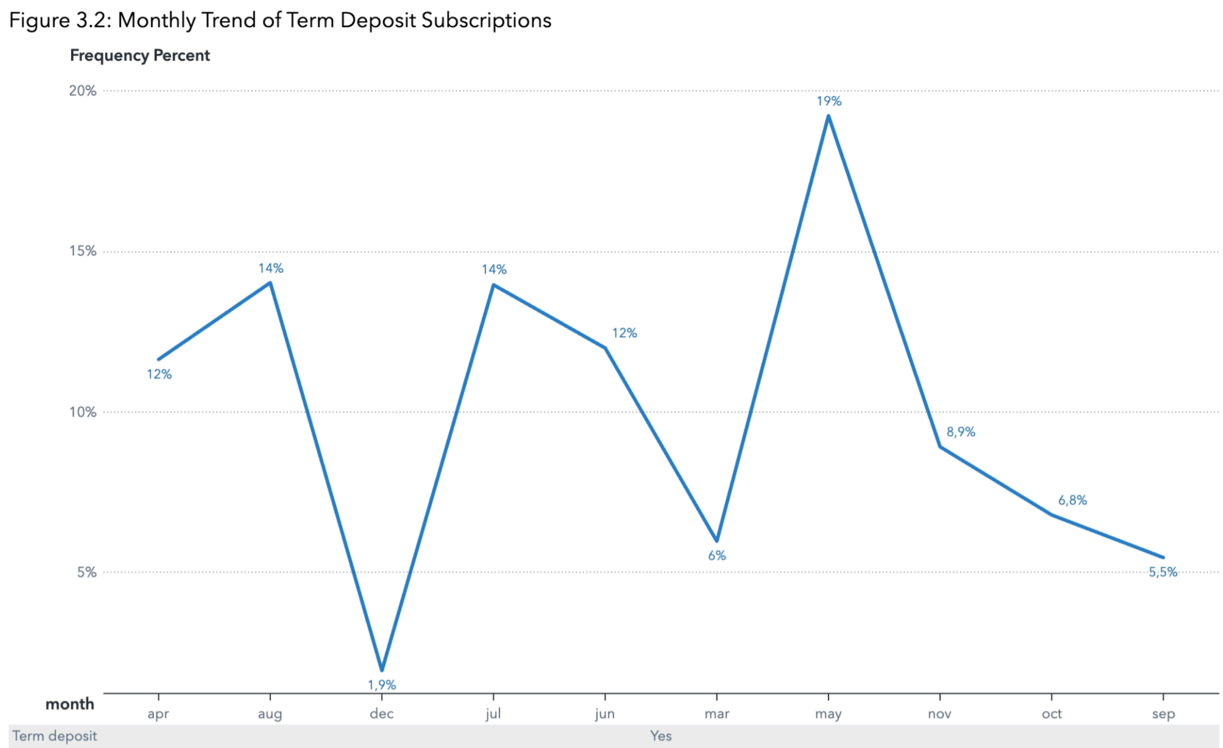


Figure 3.2: Monthly Trend of Term Deposit Subscriptions

Figure 3.2 shows that May demonstrates the highest rate of term deposit subscriptions, followed by July and August. Conversely, December and March exhibit the least engagement. The observed seasonal trends indicate that consumers are more amenable during the mid-year months, likely attributable to financial planning cycles or promotional initiatives, underscoring the significance of timing in enhancing marketing efficacy.

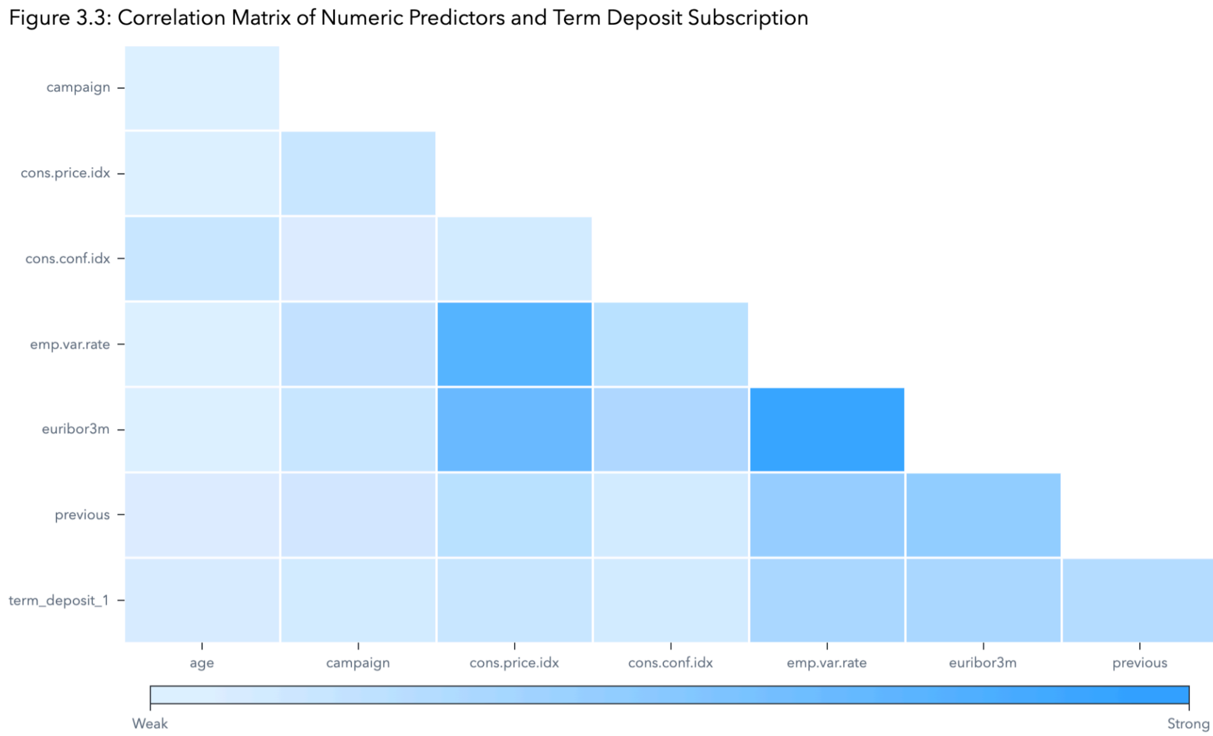


Figure 3.3: Correlation Matrix of Numeric Predictors and Term Deposit Subscription

The correlation analysis revealed that euribor3m (r = -0.31) and emp.var.rate (r = -0.30) exhibited the most pronounced negative correlation with term_deposit_1, indicating that adverse market conditions (e.g., diminished interest rates or employment apprehensions) may heighten customer interest in fixed-term savings. Furthermore, prior research demonstrated a weak-to-moderate positive correlation (r = +0.23), suggesting that historical contact significantly influences customer response.

Overview Insight

The three visualisations reveal significant patterns in customer subscription behaviour. Occupationally, administrative and technical positions exhibit greater involvement with term deposits, indicating a correlation between employment type and financial choices. Seasonally, May, July, and August exhibit peak subscription periods, suggesting that mid-year campaigns may produce superior outcomes. Finally, the correlation matrix indicates that lower interest rates (euribor3m) and deteriorating employment conditions (emp.var rate) are inversely related to subscriptions, whereas prior contact history exerts a slight positive effect. These insights offer critical guidance for campaign scheduling, audience segmentation, and the selection of variables for subsequent statistical analysis and predictive modelling.

4. Inferential Statistics

Chi-Square Test

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
job * term_deposit	40188	100.0%	0	0.0%	40188	100.0%

job * term_deposit Crosstabulation

			term_deposit: no	term_deposit: yes	Total
job	admin.	Count	8846	1321	10167
		Expected Count	9024.3	1142.7	10167.0
	blue-collar	Count	8413	625	9038
		Expected Count	8022.2	1015.8	9038.0
	entrepreneur	Count	1295	120	1415
		Expected Count	1256.0	159.0	1415.0
	housemaid	Count	930	101	1031
		Expected Count	915.1	115.9	1031.0
	management	Count	2521	321	2842
		Expected Count	2522.6	319.4	2842.0
	retired	Count	1264	418	1682
		Expected Count	1492.9	189.1	1682.0
	self-employed	Count	1232	144	1376
		Expected Count	1221.3	154.7	1376.0
	services	Count	3559	316	3875
		Expected Count	3439.5	435.5	3875.0
	student	Count	583	264	847
		Expected Count	751.8	95.2	847.0
	technician	Count	5895	715	6610
		Expected Count	5867.1	742.9	6610.0
	unemployed	Count	846	137	983
		Expected Count	872.5	110.5	983.0
	unknown	Count	287	35	322
		Expected Count	285.8	36.2	322.0
Total		Count	35671	4517	40188
		Expected Count	35671.0	4517.0	40188.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	909.469[a]	11	<.001
Likelihood Ratio	770.976	11	<.001
N of Valid Cases	40188		

a. 0 cells (0.0%) have expected count less than 5. The

Figure 4.1: Chi-Square Test of Job Type and Term Deposit Subscription

Hypotheses:

Null Hypothesis (H_0): There is no statistically significant association between job type and term deposit subscription.

Alternative Hypothesis (H_1): There is a statistically significant association between job type and term deposit subscription.

Interpretation:
A chi-squared test of independence was performed to investigate the association between job type and term deposit subscription. The outcome was statistically significant as per Figure 4.1 (Pearson Chi-Square = 909.469 and $p < .001$), signifying the rejection of the null hypothesis. This establishes a non-random correlation between profession and subscription behaviour. By prior descriptive analyses, administrative and technician positions exhibited subscription rates exceeding expectations, whereas blue-collar and service roles fell short of anticipated levels. These findings correspond with the current literature that associates employment stability and financial literacy with proactive investment behaviour (Struckell et al. 2022). These disparities highlight the significance of occupation-based segmentation in financial marketing. By incorporating job-type insights into predictive modelling and campaign targeting, institutions can more effectively align their messaging and outreach with behavioural patterns identified in the data. Consequently, occupation serves as a demographic designation and a behavioural metric impacting financial decision-making.

Independent Samples t-Test

T-Test

Group Statistics

	term_deposit	N	Mean	Std. Deviation	Std. Error Mean
campaign	yes	4517	2.05	1.668	.025
	no	35671	2.63	2.872	.015

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
campaign	Equal variances assumed	316.228	<.001	-13.293	40186	<.001	-.580	.044	-.666	-.494
	Equal variances not assumed			-19.930	8394.493	<.001	-.580	.029	-.637	-.523

Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval	
				Lower	Upper
campaign	Cohen's d	2.763	-.210	-.241	-.179
	Hedges' correction	2.763	-.210	-.241	-.179
	Glass's delta	2.872	-.202	-.233	-.171

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control group.

Figure 4.2: Independent Samples t-Test on Campaign Frequency Between Subscribed and Non-Subscribed Customers

Hypothesis:

Null Hypothesis (H_0)**:** There is no difference in the number of contact attempts (campaign) between customers who subscribed and those who did not.

Alternative Hypothesis (H_1)**:** There is a significant difference in the number of contact attempts between the two groups.

Interpretation:

An independent samples t-test was performed to evaluate the impact of campaign intensity on term deposit subscriptions. The findings revealed a statistically significant disparity in the mean contact attempts between subscribers and non-subscribers. Subscribers received an average of 2.05 calls (standard deviation = 1.67), whereas non-subscribers received an average of 2.63 calls (standard deviation = 2.87). The computed t-statistic was –13.293 with 40,186 degrees of freedom, and the corresponding p-value was below 0.001. When the p-value decreases below the 0.05 threshold, the null hypothesis is rejected. Although the mean difference of 0.58 seems minimal, the statistical and behavioural ramifications are significant. The descriptive findings, including the substantial rise in subscriptions in May, indicate that excessive communication may result in disengagement. This corresponds with the marketing saturation theory (Deloitte 2019), which promotes fewer, strategically timed interactions. Consequently, a strategy emphasising timing over frequency may produce superior customer engagement results.

5. Prediction Model

A classification pipeline was established using SAS Viya to forecast which customers are likely to subscribe to a term deposit, utilising three algorithms: Logistic Regression, Decision Tree, and Random Forest. Due to the dataset's class imbalance, where merely 11.24% of instances indicate positive responses, meticulous model comparison and evaluation were performed utilising various performance metrics, such as AUC, Gini coefficient, misclassification rate, and F1-score.

The Forest algorithm was chosen as the optimal model based on its exceptional performance among the evaluated models. According to Figure 5.1, it attained the highest AUC (0.8025), the lowest Average Squared Error (0.0770), and a comparatively low misclassification rate (10.08%), in addition to a robust Gini coefficient (0.6049), indicating strong generalizability and discriminative capability with unseen data.

Due to the class imbalance, we prioritised the F1-score, reconciling Precision and Recall (Appendix 5.2). The default cutoff of 0.50 yielded an F1-score of 0.3721, whereas the KS-optimised cutoff of 0.13 demonstrated markedly improved classification performance. At this threshold, the Forest model identified 560 true positives, 848 false positives, and 343 false negatives, yielding a Precision of 0.3977, a Recall of 0.6204, and an F1-score of 0.48. This enhancement validates the efficacy of modifying decision thresholds in imbalanced classification scenarios, as it optimises the identification of genuinely interested customers while reducing extraneous noise. This method conforms to optimal practices in imbalanced learning (Luque et al. 2019), wherein cutoff tuning is essential for improving actionable accuracy.

A tiered response strategy is advised in operational contexts, predicated on model predictions. According to Appendix 5.3, customers identified with high confidence (P_term_deposityes > 0.80) should be prioritised for direct communication through phone or personal banking representatives. Individuals within the moderate range (0.50–0.80) may be approached with tailored email campaigns, whereas the low-confidence cohort (below 0.50) might receive generic newsletters or be omitted from active promotion to enhance resource efficiency. This strategy enables the organisation to customise outreach intensity based on anticipated probability, enhancing overall campaign efficiency and return on marketing investment.

By adopting this precision-oriented strategy, financial institutions can optimise marketing investment returns while synchronising engagement initiatives with customer propensity. The Forest model, particularly its calibrated threshold, serves as a nexus between technical resilience and business outcomes.

6. Conclusion and Recommendations

This report utilised a comprehensive analytics methodology to reveal behavioural trends in term deposit subscriptions. Descriptive analysis identified occupation, timing, and market conditions as pivotal factors: administrative and technician positions exhibited elevated subscription rates (3.3% and 1.8%, respectively), with May recognised as the month with the highest conversions. Correlation analysis indicated that reduced interest rates (euribor3m, $r = -0.31$) and poor employment conditions (emp.var.rate, $r = -0.30$) positively affected the likelihood of subscriptions, implying that economic uncertainty stimulates demand for secure savings. Inferential testing corroborated these findings. A Chi-Square test established a statistically significant association between job type and subscription outcome (Chi-Square = 909.469, $p < .001$), whereas a t-test revealed that subscribers experienced fewer contact attempts on average (2.05 compared to 2.63, $p < .001$). These findings advocate a transition from extensive outreach to precise, strategically timed communication. The predictive model enhanced these insights. The Forest algorithm attained the highest AUC of 0.8025 and an F1-score of 0.48 at a cutoff of 0.13, surpassing other algorithms and facilitating tiered customer segmentation based on predicted likelihood.

In future campaigns, banks should emphasise strategic timing, tailor outreach by occupational group, and restrict excessive contact attempts. Nonetheless, constraints encompass data imbalance, lack of customer-specific financial metrics, and lack of insight into conversion timelines. Future research may investigate causal modelling, integrate demographic or psychographic data, and evaluate uplift modelling techniques to measure incremental effects. Improving model explainability and exploring ensemble methods may enhance prediction accuracy and foster trust in real-world applications.

References:

Struckell, EM, Ortegren, M, Burris, E & Rutherford, BN 2022, 'Financial literacy and self-employment – the moderating effect of gender and race', *Journal of Business Research*, vol. 139, pp. 639–653, viewed [date you accessed it], https://doi.org/10.1016/j.jbusres.2021.10.003.

Deloitte 2019, *Measuring and managing marketing effectiveness: Deloitte Georgia*, *Deloitte*, viewed 18 April 2025, https://www.deloitte.com/ge/en/services/consulting/perspectives/measuring-marketing-effectiveness-mroi.html.

Luque, A, Carrasco, A, Martín, A & de las Heras, A 2019, 'The impact of class imbalance in classification performance metrics based on the binary confusion matrix', *Pattern Recognition*, vol. 91, pp. 216–231, viewed [date you accessed it], https://doi.org/10.1016/j.patcog.2019.02.023.

Appendix:

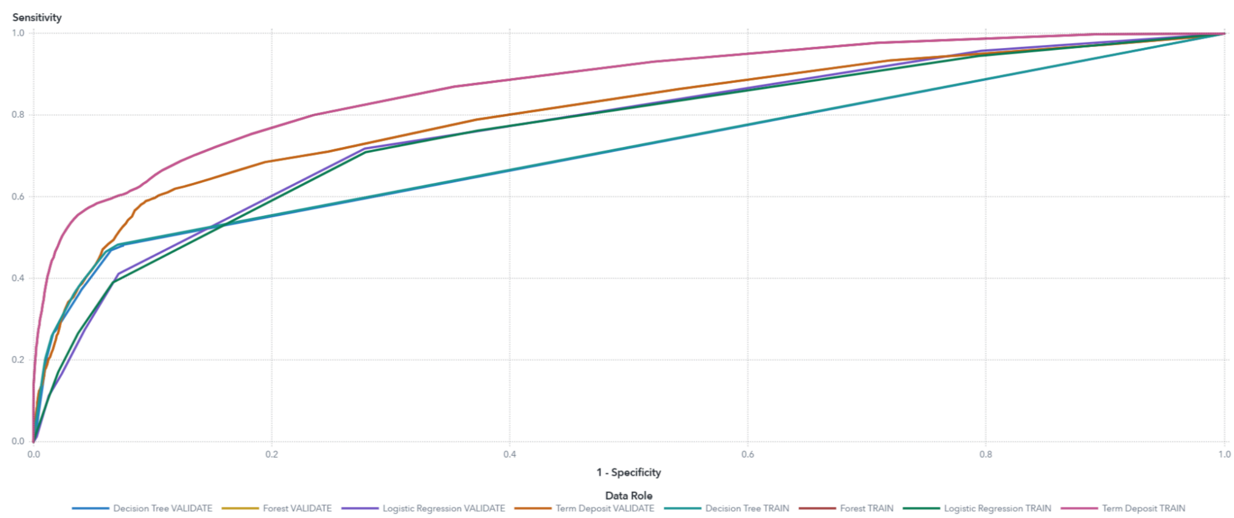
Appendix 5.1: Model Comparison Results betweem Forest and Other Algorithms

Model Comparison

Champion ↑	Name	Algorit...	KS (Youden)	Accuracy	Average Squared Error	Area Under ROC	Cumulative Lift	Cutoff	F1 Score	False Positive Rate	Gain	Gini Coefficient	ROC Separation	Lift
*	Term Deposit	Forest	0.5013	0.8992	0.0770	0.8025	4.5404	0.5000	0.3721	0.0206	3.5404	0.6049	0.2452	3.5659
	Logistic Regression	Logistic Regression	0.4403	0.8889	0.0857	0.7664	3.7920	0.5000	0.1860	0.0129	2.7920	0.5329	0.1001	3.4960
	Decision Tree	Decision Tree	0.4074	0.9028	0.0804	0.7109	4.3832	0.5000	0.3816	0.0167	3.3832	0.4218	0.2502	3.1230
	Forest	Forest	0.5013	0.8992	0.0770	0.8025	4.5404	0.5000	0.3721	0.0206	3.5404	0.6049	0.2452	3.5659

ROC Reports

View chart: ROC



Appendix 5.2: Model Evaluation Metrics (Forest Model)

Event Classification

View chart: Table

Cutoff	Cutoff Source	Target Name	Response	Event	Value	Training Freque...	Validation Freq...	Test Frequency	Training Percent...	Validation Perce...	Test Percentage
0.0800	KS	term_deposit	CORRECT	yes	True Positive	2,722	.	.	75.3182	.	.
0.0800	KS	term_deposit	INCORRECT	yes	False Negative	892	.	.	24.6818	.	.
0.0800	KS	term_deposit	CORRECT	no	True Negative	23,347	.	.	81.8131	.	.
0.0800	KS	term_deposit	INCORRECT	no	False Positive	5,190	.	.	18.1869	.	.
0.1300	KS	term_deposit	CORRECT	yes	True Positive	.	560	.	.	62.0155	.
0.1300	KS	term_deposit	INCORRECT	yes	False Negative	.	343	.	.	37.9845	.
0.1300	KS	term_deposit	CORRECT	no	True Negative	.	6,286	.	.	88.1133	.
0.1300	KS	term_deposit	INCORRECT	no	False Positive	.	848	.	.	11.8867	.
0.5000	Default	term_deposit	CORRECT	yes	True Positive	1,318	240	.	36.4693	26.5781	.
0.5000	Default	term_deposit	INCORRECT	yes	False Negative	2,296	663	.	63.5307	73.4219	.
0.5000	Default	term_deposit	CORRECT	no	True Negative	28,272	6,987	.	99.0714	97.9394	.
0.5000	Default	term_deposit	INCORRECT	no	False Positive	265	147	.	0.9286	2.0606	.

Based on the confusion matrix at cutoff = 0.13 (Validation Set):

- True Positives (TP): 560

- False Positives (FP): 848

- False Negatives (FN): 343

Precision (Positive Predictive Value):

= TP / (TP + FP)

= 560 / (560 + 848)

= 560 / 1408 ≈ 0.3977

Recall (Sensitivity):

= TP / (TP + FN)

= 560 / (560 + 343)

= 560 / 903 ≈ 0.6204

F1-Score (Harmonic Mean of Precision and Recall):

= 2 × (Precision × Recall) / (Precision + Recall)

= 2 × (0.3977 × 0.6204) / (0.3977 + 0.6204)

= 2 × 0.2468 / 1.0181 ≈ 0.4841

Appendix 5.3: Labelled Customer Data (upload to Turnitin)

Identifier	EM_EVENTPROBABILITY	EM_CLASSIFICATION	EM_PROBABILITY	P_term_deposityes	P_term_depositno	I_term_deposit
1	0.042055407	no	0.957944593	0.042055407	0.957944593	no
2	0.041459274	no	0.958540726	0.041459274	0.958540726	no
3	0.039968289	no	0.960031711	0.039968289	0.960031711	no
4	0.04342064	no	0.95657936	0.04342064	0.95657936	no
5	0.029187987	no	0.970812013	0.029187987	0.970812013	no
6	0.04463967	no	0.95536033	0.04463967	0.95536033	no
7	0.0398282	no	0.9601718	0.0398282	0.9601718	no
8	0.028962758	no	0.971037242	0.028962758	0.971037242	no
9	0.041505326	no	0.958494674	0.041505326	0.958494674	no
10	0.048088554	no	0.951911446	0.048088554	0.951911446	no
11	0.043062412	no	0.956937588	0.043062412	0.956937588	no
12	0.04691698	no	0.95308302	0.04691698	0.95308302	no
13	0.016839941	no	0.983160059	0.016839941	0.983160059	no
14	0.03871104	no	0.96128896	0.03871104	0.96128896	no
15	0.0329138	no	0.9670862	0.0329138	0.9670862	no
16	0.039793488	no	0.960206512	0.039793488	0.960206512	no
17	0.035161351	no	0.964838649	0.035161351	0.964838649	no
18	0.028756825	no	0.971243175	0.028756825	0.971243175	no
19	0.044993024	no	0.955006976	0.044993024	0.955006976	no
20	0.048108893	no	0.951891108	0.048108893	0.951891108	no
21	0.045604851	no	0.95439515	0.045604851	0.95439515	no
22	0.047900322	no	0.952099678	0.047900322	0.952099678	no
23	0.041822992	no	0.958177008	0.041822992	0.958177008	no
24	0.051453155	no	0.948546845	0.051453155	0.948546845	no
25	0.043709873	no	0.956290127	0.043709873	0.956290127	no
26	0.041873978	no	0.958126022	0.041873978	0.958126022	no
27	0.044315054	no	0.955684946	0.044315054	0.955684946	no
28	0.045632623	no	0.954367377	0.045632623	0.954367377	no
29	0.036461425	no	0.963538575	0.036461425	0.963538575	no
30	0.036978762	no	0.963021238	0.036978762	0.963021238	no
31	0.037646159	no	0.962353841	0.037646159	0.962353841	no
32	0.03963975	no	0.96036025	0.03963975	0.96036025	no