Analysis of Customer Response to Fixed Term Savings Account Campaign Based on Personal Characteristics

Contents

FIGURE

TABLE

Introduction

This project aims to analyse the data from an international bank which held campaign to promote the fixed term saving account. The primary objective of this project is to identify the potential customers who would respond positively to the campaign. We were given two datasets; one with the campaign contact information and the response of each customer, other dataset which comprises the personal characteristics of the customers. Upon examining these datasets, we must build a predictive model to understand the factors influencing customer responses and to improve the effectiveness of future campaigns.

Campaign dataset has 4 variables related to the customer id, contact method, duration of each call and the response of each customer. While the personal dataset has 10 variables related to the customer id, age, customer home region, job, education, marital status, default which indicates whether the customer's credit is in default or not, balance that each customer holds, and the last two tells whether the customer has any housing loan or personal loan. This dataset consists of 33909 observations.

Pre-processing

Missing values

Missing values: campaign

Variable name	Description	Value	Variable type	No of missing values	No of outliers (Low)	No of outliers (High)
Cust_id			Numeric	-	-	-
Contact		1-Mobile, 2- telephone 3- Unknown	Nominal	0	0	1568
Duration			Numeric	0	0	0
Response		0-No 1-Yes	Nominal	0	0	0

Table 1 Missing values and outliers of campaign data

Missing values: personal

Variable name	Description	Value	Variable type	No of missing values	No of outliers (Low)	No of outliers (High)
Cust_id			Numeric	-	-	-
Age			Numeric	0	32	720
Region		0-North East 1-South West 2-East of England 3-London 4-south east 5-north west 6-west midland 7-Yorkshire and the Humber 8-east midlands	Nominal	0	1	1159
Job		1-admin 2-others 3-entrepreneur 4-domestic worker 5-mangement 6-retired 7-self -employed 8- services 9-student 10-technician 11-unemployed 12-unknown	Nominal	0	0	0
Marital		1-others 2-married 3-single	Nominal	0		
Education		1-primary 2-secondary	Nominal	0		

		3-tertiary 4-unknown				
Default		1-no 2-yes	Nominal	0		
Balance			Numeric	0		
Housing		1-no 2-yes	Nominal	0		
Loan		1-no 2-yes	Nominal	0		

Table 2 Missing values and outliers for personal data

Table 1 and table 2 depicts the missing values and outliers of both the datasets. From table 1 it is evident that there are no missing values with high outlier of 1568 for contact variable. Table 2 of personal characteristics shows there are no missing values as well, but age has 32 low outliers and 720 high outliers. Region has 1 low outlier and 1159 high outliers. Since there are no missing values there is no need to handle them.

Explanatory Data Analysis

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
custID	33909	2	45211	22667.10	13040.886
contact	33909	1	3	1.64	.896
duration	33909	0	4918	257.61	256.435
response	33909	0	1	.12	.321
Valid N (listwise)	33909				

Table 3 Descriptive statistics of campaign

The above table depicts the descriptive statistics of campaign dataset where we can note the mean of customer id(**22667.10**) which has the minimum value of 2 and maximum value of 45211. Since contact variable has got three methods the mean is **1.64** which tells more of mobile phone and telephone was used as contact method. The mean duration of each call being **257.61** seconds and the response mean is **0.12**.

The below table shows the descriptive statistics of personal dataset where mean of age and region being **40.97** and **4** respectively. job(**5.34**), marital(**2.17**), education(**2.22**), default(**1.02**) might play a vital role in the predictive analysis. The average balance a customer holds was found to be **1569.57**, and almost everyone has got either a housing loan or personal loan.

	N	Minimum	Maximum	Mean	Std. Deviation
custID	33909	2	45211	22667.10	13040.886
age	33909	18	95	40.97	10.628
region	33909	0	8	4.00	1.418
job	33909	1	12	5.34	3.269
marital	33909	1	3	2.17	.607
education	33909	1	4	2.22	.748
default	33909	1	2	1.02	.131
balance	33909	-7962	114438	1569.57	3420.725
housing	33909	1	2	1.56	.497
loan	33909	1	2	1.16	.367
Valid N (listwise)	33909				

Descriptive Statistics

Table 4 Descriptive statistics of personal

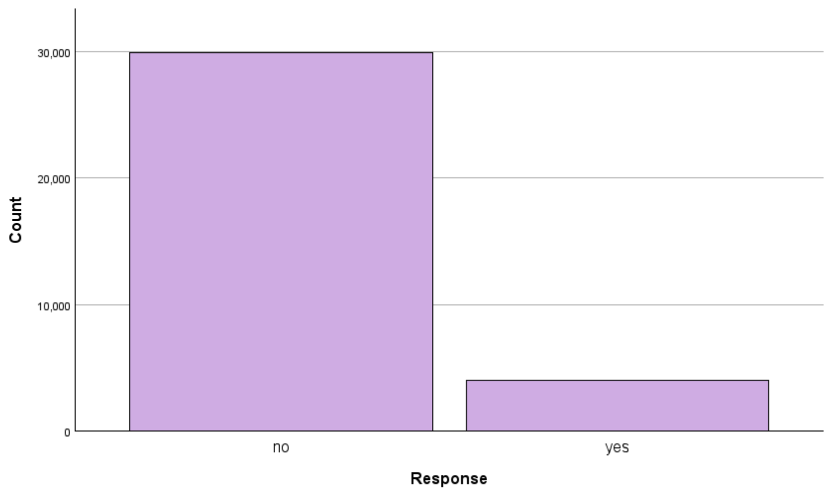


Figure 1 Count of responses

The above bar chart provides the evidence that most of the customers have the response of "No" with nearly 30000 customers. Before building the model, we were able to figure out the frequency of our dependent variable.

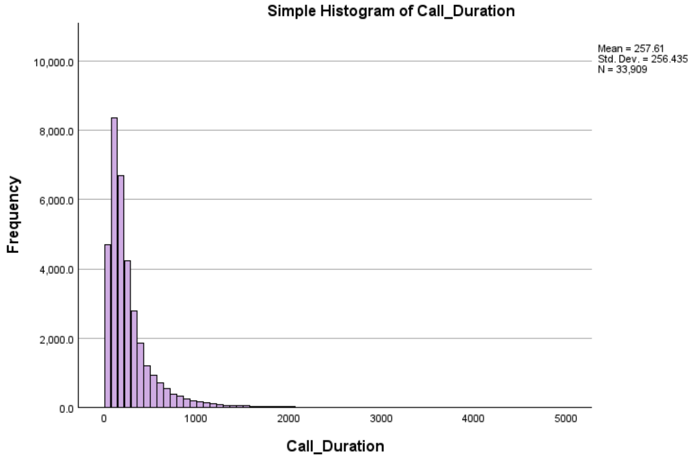


Figure 2 Histogram of call duration

The histogram in the figure 2 of the duration variable, the data appears to be right skewed with minimum amount of duration being 0 and maximum amount of duration per call contributing to 4918 seconds.

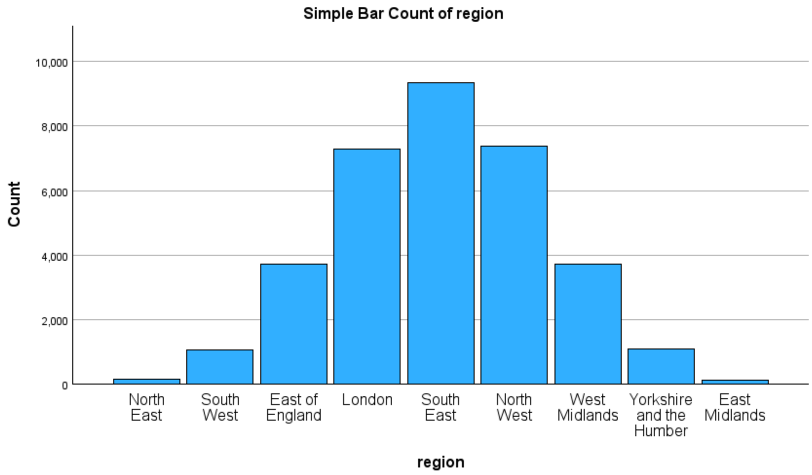


Figure 3 Home region of customers

The above figure depicts the home region of the customers with maximum number of customers living in Southeast region and the least from East Midlands.

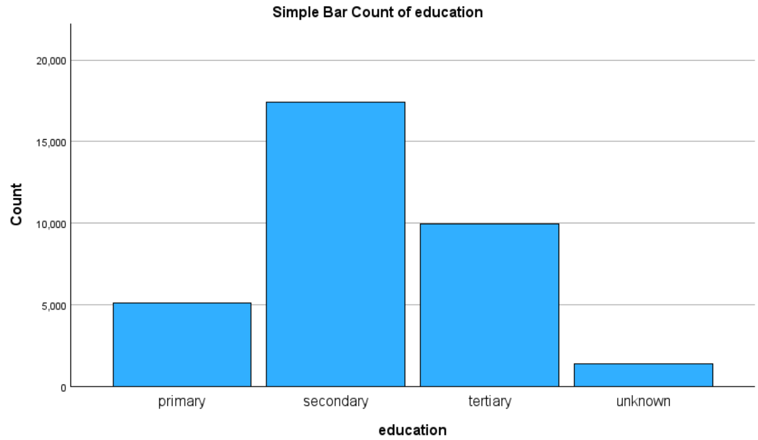


Figure 4 Educational qualification of customers

The above bar chart shows education level of each customer. Most of the customers were the ones with secondary education, the next with tertiary education, some of them were of primary education and rest of the customers were not unknown.

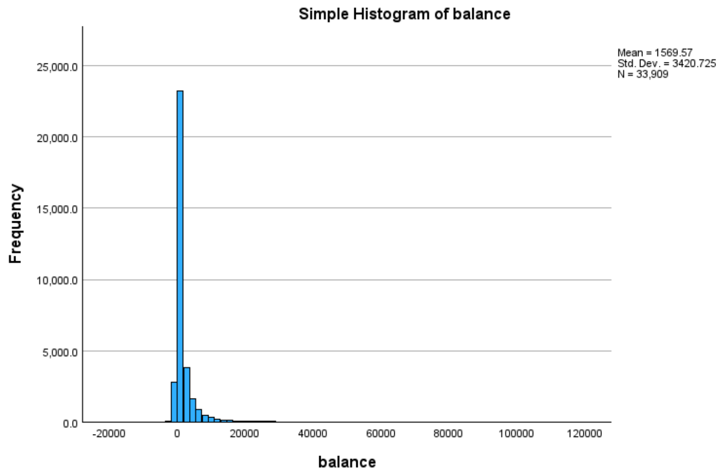


Figure 5 Histogram of balance of customer

The above histogram of balance variable, it appears that the data is highly skewed to the right with most balances concentrated around lower values and a long tail extending towards higher values.

Binning Variables

All the continuous variables were binned to build the logistic regression model. The two main method for binning is binning with equal width intervals and the other one is equal percentile intervals. Both the methods have its own pros and cons. Binning with equal width intervals is easy to implement and interpret, but then the data must be uniformly distributed, and it is less effective if the data is skewed. Whereas binning with equal percentile intervals ensures that each bin has equal number of data points, and it is suitable for skewed data as it provides balanced view of data distribution across all the bins. Since from above graphs it is evident that our data is skewed, the better option was chosen to be binning with equal percentile.

Balance (Binned)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<= 83	8488	25.0	25.0	25.0
	84 - 520	8480	25.0	25.0	50.0
	521 - 1655	8465	25.0	25.0	75.0
	1656+	8476	25.0	25.0	100.0
	Total	33909	100.0	100.0	

Table 5 Binned balance

Equal_Freq

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<= 103	8544	25.2	25.2	25.2
	104 - 180	8460	24.9	24.9	50.1
	181 - 318	8437	24.9	24.9	75.0
	319+	8468	25.0	25.0	100.0
	Total	33909	100.0	100.0	

Table 6 Binned call duration

Age (Binned)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<= 33	9776	28.8	28.8	28.8
	34 - 39	7740	22.8	22.8	51.7
	40 - 48	7935	23.4	23.4	75.1
	49+	8458	24.9	24.9	100.0
	Total	33909	100.0	100.0	

Table 7 Binned age

Table 5, table 6, and table 7 are the frequencies of the balance, call duration and age variables after they were binned. Each variable was binned with equal percentiles based on scanned cases, giving the number of cut points as 3 and equal width of 25%. Based on the information above, how each variable is categorized are seen.

Merging and splitting of dataset.

The two datasets were finally merged into a single dataset comprising of 33909 observations with 17 variables including all the binned variables. Once merged, the dataset was split into training and testing set. The split was 70% training set and 30% test set, and the “8888” was used as the randomisation seed.

	custID	contact	duration	age	region	job	marital	education	default	balance	housing	loan	Equal_balance	Equal_Frequence	Binned_age	response	Flag
1	2	unknown	151	44	London	technici...	single	secondary	no	34	yes	no	<= 83	104 - 180	40 - 48	no	Training
2	4	unknown	92	47	London	others	married	unknown	no	1751	yes	no	1656+	<= 103	40 - 48	no	Training
3	5	unknown	198	33	South East	unknown	single	unknown	no	1	no	no	<= 83	181 - 318	<= 33	no	Testing
4	6	unknown	139	35	London	manag...	married	tertiary	no	269	yes	no	84 - 520	104 - 180	34 - 39	no	Testing
5	7	unknown	217	28	Yorkshire a...	manag...	single	tertiary	no	520	yes	yes	84 - 520	181 - 318	<= 33	no	Training
6	9	unknown	50	58	South East	retired	married	primary	no	141	yes	no	84 - 520	<= 103	49+	no	Training
7	10	unknown	55	43	West Midla...	technici...	single	secondary	no	690	yes	no	521 - 1655	<= 103	40 - 48	no	Training
8	12	unknown	137	29	West Midla...	admin	single	secondary	no	453	yes	no	84 - 520	104 - 180	<= 33	no	Training
9	13	unknown	517	53	South East	technici...	married	secondary	no	7	yes	no	<= 83	319+	49+	no	Training
10	14	unknown	71	58	South West	technici...	married	unknown	no	83	yes	no	<= 83	<= 103	49+	no	Training
11	15	unknown	174	57	North West	services	married	secondary	no	188	yes	no	84 - 520	104 - 180	49+	no	Training
12	17	unknown	98	45	North West	admin	single	unknown	no	15	yes	no	<= 83	<= 103	40 - 48	no	Training
13	18	unknown	38	57	South East	others	married	primary	no	60	yes	no	<= 83	<= 103	49+	no	Training
14	19	unknown	219	60	North West	retired	married	primary	no	70	yes	no	<= 83	181 - 318	49+	no	Testing
15	20	unknown	54	33	North West	services	married	secondary	no	0	yes	no	<= 83	<= 103	<= 33	no	Training
16	21	unknown	262	28	West Midla...	others	married	secondary	no	841	yes	yes	521 - 1655	181 - 318	<= 33	no	Training
17	22	unknown	164	56	South East	manag...	married	tertiary	no	906	yes	no	521 - 1655	104 - 180	49+	no	Training
18	23	unknown	160	32	South East	others	single	primary	no	27	yes	yes	<= 83	104 - 180	<= 33	no	Training
19	25	unknown	181	40	South East	retired	married	primary	no	0	yes	yes	<= 83	181 - 318	40 - 48	no	Training
20	26	unknown	172	44	London	admin	married	secondary	no	-433	yes	no	<= 83	104 - 180	40 - 48	no	Training

Table 8 Splitting of dataset

Response Model

Initial Model

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Contact_Information			336.716	2	<.001	
	Contact_Information(1)	1.307	.072	334.155	1	<.001	3.697
	Contact_Information(2)	1.290	.110	138.543	1	<.001	3.634
	Region_Cust			7.902	8	.443	
	Region_Cust(1)	.024	.590	.002	1	.968	1.024
	Region_Cust(2)	.460	.440	1.095	1	.295	1.584
	Region_Cust(3)	.601	.426	1.995	1	.158	1.824
	Region_Cust(4)	.507	.423	1.436	1	.231	1.661
	Region_Cust(5)	.509	.423	1.452	1	.228	1.664
	Region_Cust(6)	.475	.423	1.261	1	.262	1.608
	Region_Cust(7)	.595	.426	1.951	1	.163	1.813
	Region_Cust(8)	.379	.441	.738	1	.390	1.461
	Job			118.111	11	<.001	
	Job(1)	.321	.293	1.197	1	.274	1.379
	Job(2)	-.010	.292	.001	1	.972	.990
	Job(3)	-.222	.318	.489	1	.484	.801
	Job(4)	-.200	.324	.378	1	.539	.819
	Job(5)	.079	.291	.074	1	.785	1.082
	Job(6)	.719	.297	5.869	1	.015	2.052
	Job(7)	-.010	.310	.001	1	.974	.990
	Job(8)	.022	.298	.005	1	.941	1.022
	Job(9)	.919	.308	8.915	1	.003	2.507
	Job(10)	.011	.291	.001	1	.970	1.011
	Job(11)	.179	.309	.334	1	.563	1.196
	Marital_Status			35.560	2	<.001	
	Marital_Status(1)	-.109	.085	1.646	1	.200	.896
	Marital_Status(2)	-.329	.058	31.949	1	<.001	.720
	Education			23.078	3	<.001	
	Education(1)	-.146	.132	1.215	1	.270	.864
	Education(2)	-.013	.117	.012	1	.912	.987
	Education(3)	.257	.124	4.319	1	.038	1.293
	Default(1)	-.028	.209	.018	1	.894	.973
	Housing_loan(1)	.657	.049	177.661	1	<.001	1.930
	Personal_loan(1)	.475	.074	41.632	1	<.001	1.608
	Balance (Binned)			98.360	3	<.001	
	Balance (Binned)(1)	-.687	.070	97.080	1	<.001	.503
	Balance (Binned)(2)	-.266	.062	18.521	1	<.001	.767
	Balance (Binned)(3)	-.191	.060	10.168	1	.001	.826
	Equal_Freq			1827.218	3	<.001	
	Equal_Freq(1)	-3.767	.132	812.493	1	<.001	.023
	Equal_Freq(2)	-2.146	.067	1018.039	1	<.001	.117
	Equal_Freq(3)	-1.311	.053	602.814	1	<.001	.270
	Age (Binned)			2.729	3	.435	
	Age (Binned)(1)	-.008	.076	.012	1	.912	.992
	Age (Binned)(2)	-.060	.075	.636	1	.425	.942
	Age (Binned)(3)	-.101	.073	1.921	1	.166	.904
	Constant	-2.819	.564	25.033	1	<.001	.060

a. Variable(s) entered on step 1: Contact_Information, Region_Cust, Job, Marital_Status, Education, Default, Housing_loan, Personal_loan, Balance (Binned), Equal_Freq, Age (Binned).

Table 9 Initial logistic regression model

Logistic regression model was used as predictive model, and the above diagram shows the insignificant independent variables corresponding to the dependent variable. Contact methods are significant which increase the odds of positive response. Specific job categories such as retired and student increases the possibility of opening a fixed term saving account whereas other job categories are insignificant. Higher education, particularly tertiary is significant compared to other categories of education. Having a housing loan or personal loan would boost the response to be positive. Binned balance categories and binned duration categories(Equal_Freq) are significant, indicating varying response probabilities across different balance and duration ranges. In contrast, region, default status, and binned age categories do not significantly affect the response outcome. Overall, the model highlights the importance of contact method, job type, marital status, education, loan status, duration, and balance in predicting customer responses.

Final Model

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Contact_Information			347.079	2	<.001	
	Contact_Information(1)	1.321	.071	344.718	1	<.001	3.748
	Contact_Information(2)	1.301	.109	142.979	1	<.001	3.672
	Housing_loan(1)	.666	.048	193.790	1	<.001	1.947
	Personal_loan(1)	.481	.073	43.308	1	<.001	1.617
	Balance (Binned)			103.905	3	<.001	
	Balance (Binned)(1)	-.693	.069	102.362	1	<.001	.500
	Balance (Binned)(2)	-.263	.061	18.385	1	<.001	.769
	Balance (Binned)(3)	-.187	.060	9.844	1	.002	.830
	Equal_Freq			1829.012	3	<.001	
	Equal_Freq(1)	-3.760	.132	810.745	1	<.001	.023
	Equal_Freq(2)	-2.141	.067	1017.851	1	<.001	.118
	Equal_Freq(3)	-1.308	.053	603.157	1	<.001	.270
	Updated_Job			107.052	2	<.001	
	Updated_Job(1)	.658	.087	57.157	1	<.001	1.931
	Updated_Job(2)	.909	.120	57.024	1	<.001	2.481
	Updated_Educ(1)	.291	.049	35.925	1	<.001	1.338
	Marital_Update(1)	-.338	.047	51.853	1	<.001	.713
	Constant	-2.383	.107	494.172	1	<.001	.092

a. Variable(s) entered on step 1: Contact_Information, Housing_loan, Personal_loan, Balance (Binned), Equal_Freq, Updated_Job, Updated_Educ, Marital_Update.

Table 10 Final logistic model

Since the initial model had insignificant variable, further refinement was carried out by removing those from the model and again running the regression. The above results shows that all the variables included in the model are significant, hence this is our final model. All the variables such as contact information, housing loan, personal loan, balance, duration of the call, particular jobs like retired and student, tertiary education and married persons have a strong likelihood of giving a positive response.

Impact of variables

Every significant variable included in the final model adds to the probability of a successful campaign outcome. For example, compared to unknown contact methods, contacting customers via phone or mobile significantly increases the likelihood of receiving a positive response. In the same way, job categories for retired and students are linked to increased response probabilities. The likelihood of a positive response is increased by greater education levels, the presence of home or personal loans, and single status.

Performance Metrics

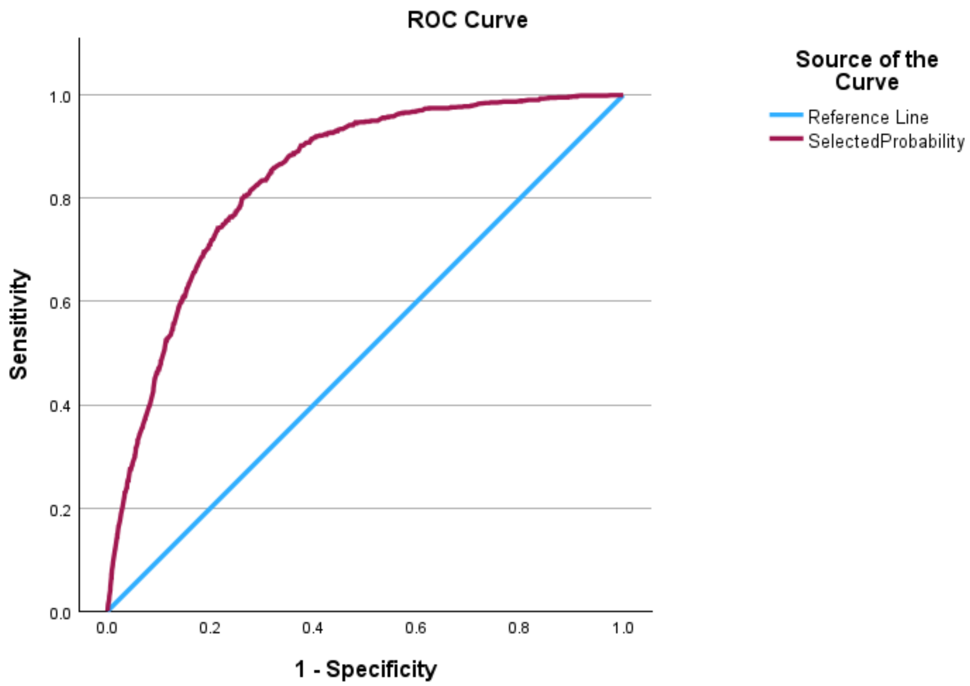


Figure 6 ROC Curve

Area Under the ROC Curve

Test Result Variable(s): Se

Area
.839

The test result variable (s): SelectedProbability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

Figure 7 AUC

The final logistic model was evaluated using various performance metrics such as ROC curve, AUC, and Precision-Recall curve by importing the logistic regression scorecard from the training model to the testing dataset which was used to assess its reliability. From figure 6 and figure 7, the ROC curve has demonstrated a significant deviation from the reference line with an AUC of **0.839**, suggesting that the model can effectively distinguish between positive and negative campaign responses.

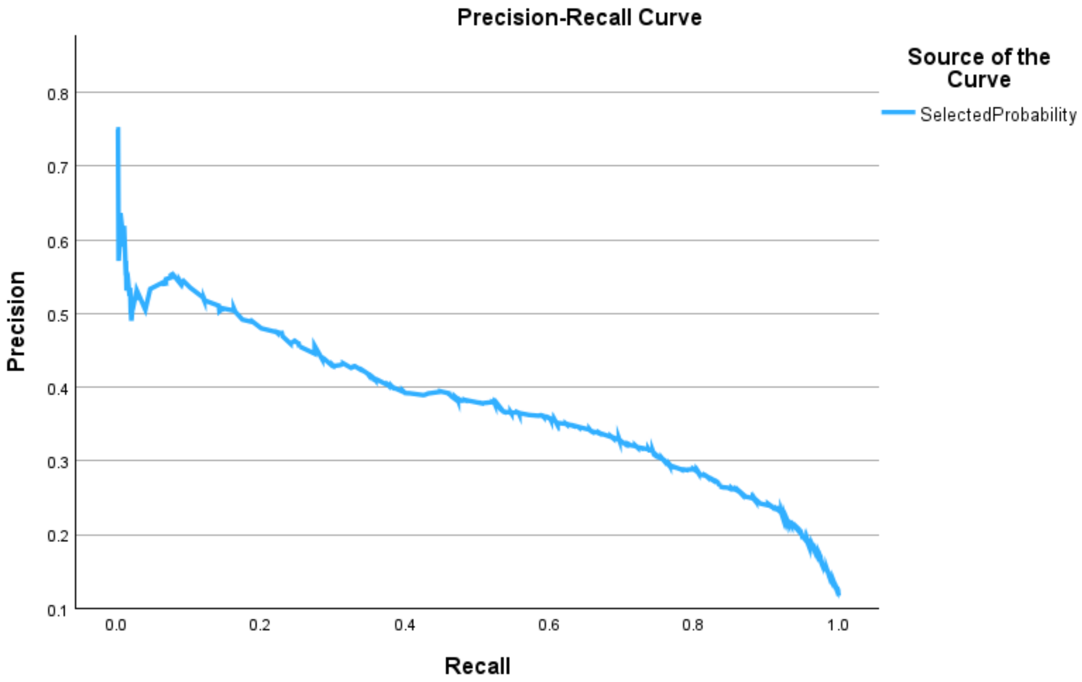


Figure 8 Precision-Recall curve

The Precision-Recall curve showed in the figure 8 high precision at lower recall levels, implying that the model is adept at identifying the most likely positive responses. As recall increases, precision decreases, which is typical in predictive models. These results underscore the model's reliability in predicting customer responses and support its application in optimizing the bank's marketing strategies.

From the above insights, the model predicts accurately, and it is reliable, hence the bank can use this model to target the customers effectively.

Marketing Campaign

Target Customers

There were several key factors influencing the likelihood of positive response from the regression results. From the results, the probable customers who we might want to target would be,

1. Customers who have provide their mobile and telephone contact information since customers who were contacted through mobile, and telephone showed significant positive response.
2. Customers who are either student or retired have shown the significant influence meaning that they are more receptive in opening a fixed term saving account.
3. The next set of customers would be married persons since they might be interested in saving plans for future stability.
4. Customers who have done their tertiary education will show positive response, as they would have studied about financial importance and know the need to open a long-term saving account.
5. Targeting customers who are financially stable since they have higher balance and might also own a housing loan or personal loan which brings them into opening a fixed term saving account.

Communication Channels

Even though we have identified the set of customers who would like to open a fixed term savings account in the bank, it is really important how they must communicate in order to pull them into opening the account.

1. **Mobile and Telephone Calls**: Since these methods showed a high response rate, a dedicated call centre team should be established to reach out to customers via these channels.
2. **Email Marketing**: For segments that may prefer less direct contact, well-crafted email campaigns can also be effective. Ensure the messages are personalized and targeted based on the segmentation criteria mentioned above.

Strategies

Personalized Customer Engagement

Craft highly personalized messages that resonate with each customer segment based on their unique characteristics such as job type, education level, and financial status. Use customer data to highlight the specific benefits of the fixed-term savings account that align with their needs and preferences.

Multi-channel Outreach

Mobile and Telephone Outreach: Deploy a dedicated call centre team to reach out to customers via mobile and telephone, which have been identified as the most effective contact methods. Schedule calls at convenient times to increase the likelihood of positive engagement.

Email Campaigns: Complement phone calls with personalized email campaigns that provide detailed information and easy sign-up options. Ensure that the email content is engaging, with clear calls to action and links to further resources.

Social Media Engagement: Leverage social media platforms to create awareness and engage with potential customers. Use targeted ads and posts to reach specific demographics identified in the analysis.

Data-Driven Adjustments

Real-Time Monitoring: Implement a robust monitoring system to track campaign performance in real-time. Use analytics to understand customer interactions and adjust strategies accordingly.

Feedback Loops: Collect customer feedback through surveys and direct interactions to continuously improve the campaign. Use this feedback to refine messages, offers, and engagement tactics.

Conclusion

Thus, from the above results of the logistic regression model, the factors which are significant and returns a positive response to the target variable was found out. Several methods to evaluate performance of the model was carried out and finally with all the findings a small marketing campaign was held to identify the potential customers who are more probable in opening a fixed term saving account in the bank.

Appendix

Univariate Statistics

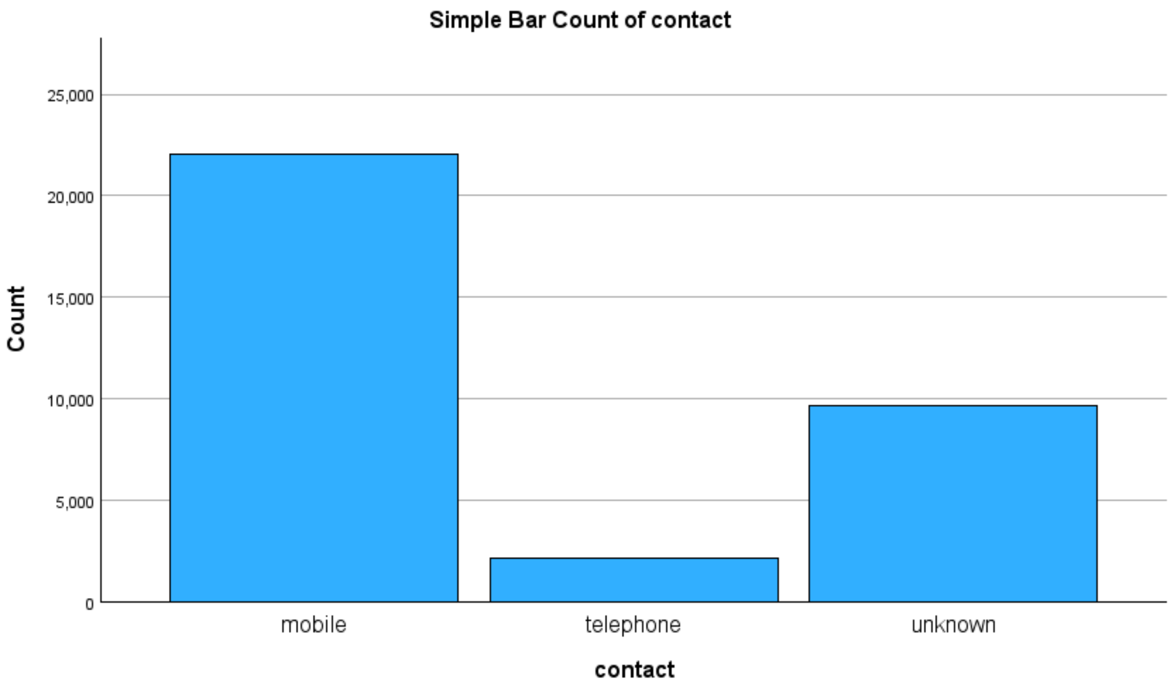
	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
duration	33909	257.61	256.435	0	.0	0	1568
contact	33909			0	.0		
response	33909			0	.0		

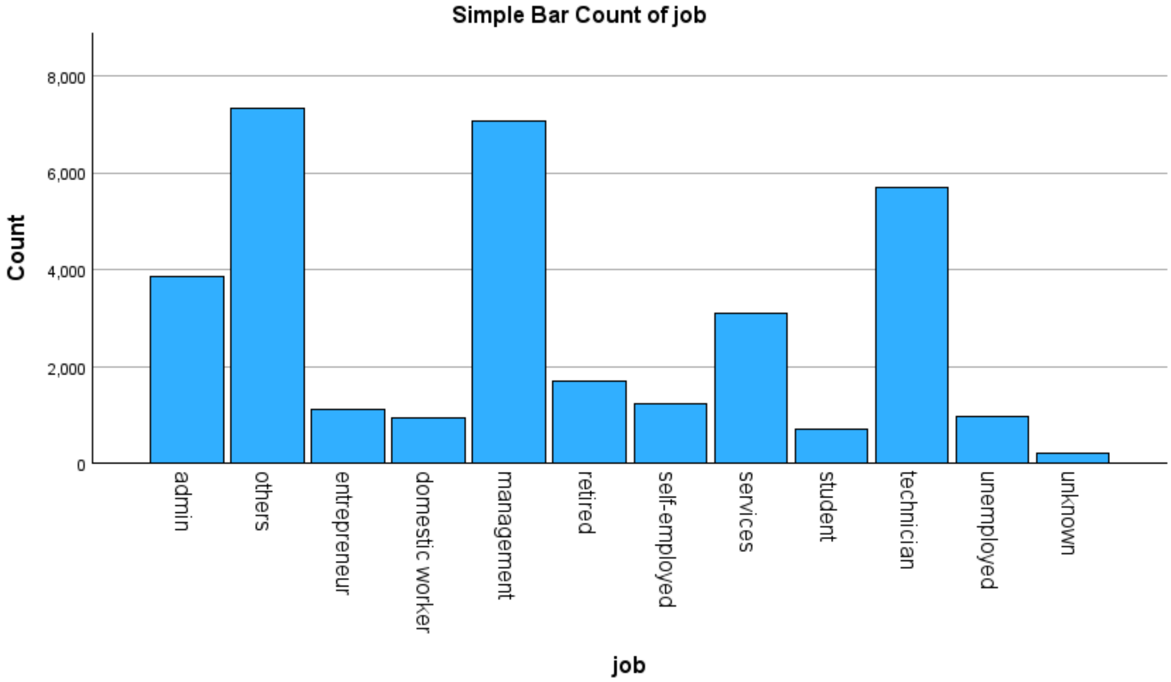
a. Number of cases outside the range (Mean - 2*SD, Mean + 2*SD).

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
age	33909	40.97	10.628	0	.0	32	720
balance	33909	1569.57	3420.725	0	.0	1	1159
region	33909			0	.0		
job	33909			0	.0		
marital	33909			0	.0		
education	33909			0	.0		
default	33909			0	.0		
housing	33909			0	.0		
loan	33909			0	.0		

a. Number of cases outside the range (Mean - 2*SD, Mean + 2*SD).





Simple Bar Count of job
Count
8,000
6,000
4,000
2,000
0
admin
others
entrepreneur
domestic worker
management
retired
self-employed
services
student
technician
unemployed
unknown
job