Opitien Brume Pascal

Data Analyst Project

brumepascal@gmail.com

Analysis-of-Bank-Marketing-Data-Using-Python-Predictive-Analytics-

INTRODUCTION

1.1 Overview of the Project

In today's dynamic business environment, data plays a crucial role in shaping strategic decisions. This report aligns with this trend by delving into a dataset from a Portuguese banking institution focused on its direct marketing campaigns conducted via phone calls. My analysis aims to unravel the influence of various factors such as customer demographics, financial indicators, and specific campaign strategies on the likelihood of customers subscribing to a term deposit. The insights garnered from this study are intended to enhance the effectiveness of future marketing endeavors of the bank.

1.2 Objectives of the Analysis

The analysis aims to reveal insights into customer behavior and the effectiveness of the marketing strategies employed. The specific objectives are:

- To understand the demographic and characteristics of the individuals in the dataset. This includes an exploration of attributes such as age, job, marital status and education level.
- To investigate the impact of these characteristics on the likelihood of a customer subscribing to a term deposit. The analysis will specifically look at whether job background, education level, and the method of contact significantly influence the subscription rate.
- To examine customer responses across different demographic groups with focus on how response rates vary across different age groups based on contact communication type.
- To identify and report any additional patterns or insights that emerge from the data which could provide valuable implications for future marketing strategies.

1.3 Brief Description of the Dataset

The dataset sourced from the direct marketing campaigns of a Portuguese banking institution. It encapsulates a variety of variables that include demographic information (age, job, marital status, education), financial attributes (default, balance, housing, loan), campaign-related details (contact type, last contact day and month, campaign number, days since last contact, previous campaign outcomes), and the target variable indicating whether the customer subscribed to a term deposit

METHODOLOGY

Importing Libraries and Data Loading

2.1.1 Library Utilization:

The analysis began with importing essential Python libraries for handling and manipulating data, conducting statistical analysis, creating visualizations, and applying machine learning techniques. These libraries included:

- **pandas**: For data manipulation and analysis.
- **numpy**: For numerical operations.
- **matplotlib** and **seaborn**: For data visualization.
- **Sklearn**: For pre-processing data and implementing machine learning models.

2.1.2 Data Loading:

The dataset was loaded into a pandas Data Frame from a CSV file. A specific separator (**sep=';'**) was used to correctly parse the dataset as it used semicolons instead of commas. The initial few rows were displayed with .head() to get an overview of the data structure and types.

Data Inspection and Cleaning

2.2.1 Inspecting the Data:

- The dataset was inspected for missing values and data types of each column. It was observed that there were no missing values across all columns.
- The presence of duplicate entries was checked, with a total of 12 duplicate rows being identified. These were retained due to the lack of unique identifiers, ensuring that potentially valuable data was not lost.

2.2.2 Cleaning and Reformatting:

- The column 'y' was renamed to 'term_deposit' for clarity.
- The dataset underwent a reformatting process where values in categorical columns were renamed for better readability. For instance, education levels were renamed to more descriptive terms, and month and day names were spelled out fully.
- The presence of 'unknown' values in several key columns was addressed by treating them as separate categories. This approach was chosen to avoid significant data loss and to potentially uncover patterns associated with these 'unknown' values.

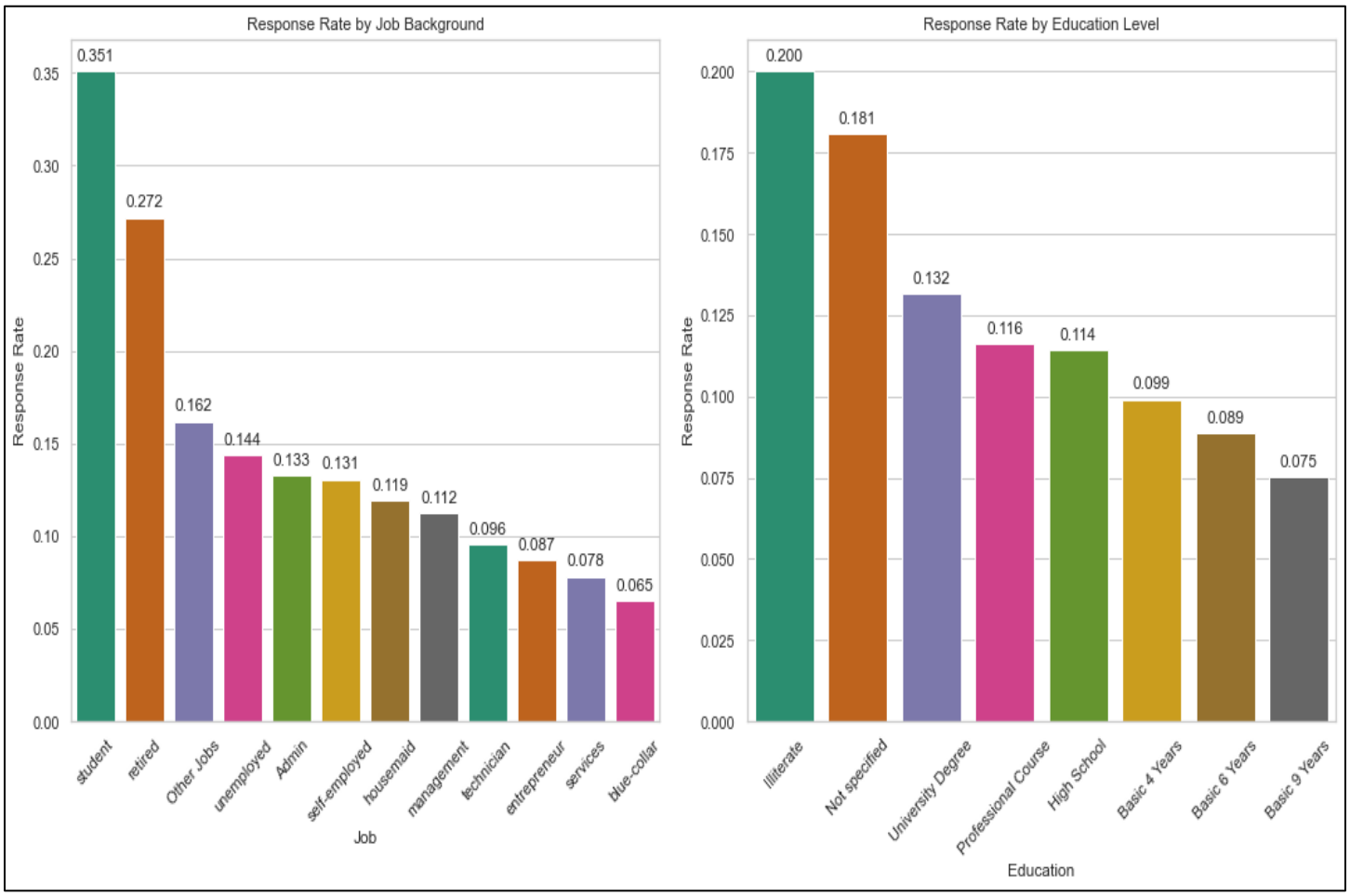
2.3 Data Sampling:

- To facilitate a manageable and efficient analysis, 20% of the data was randomly sampled. A random seed was set to ensure the reproducibility of the sample. This sampled subset was then used for the subsequent analyses in Task.

EXPLORATORY DATA ANALYSIS

This phase is crucial in understanding the complex interplay of variables that influence customer decisions regarding term deposit subscriptions.

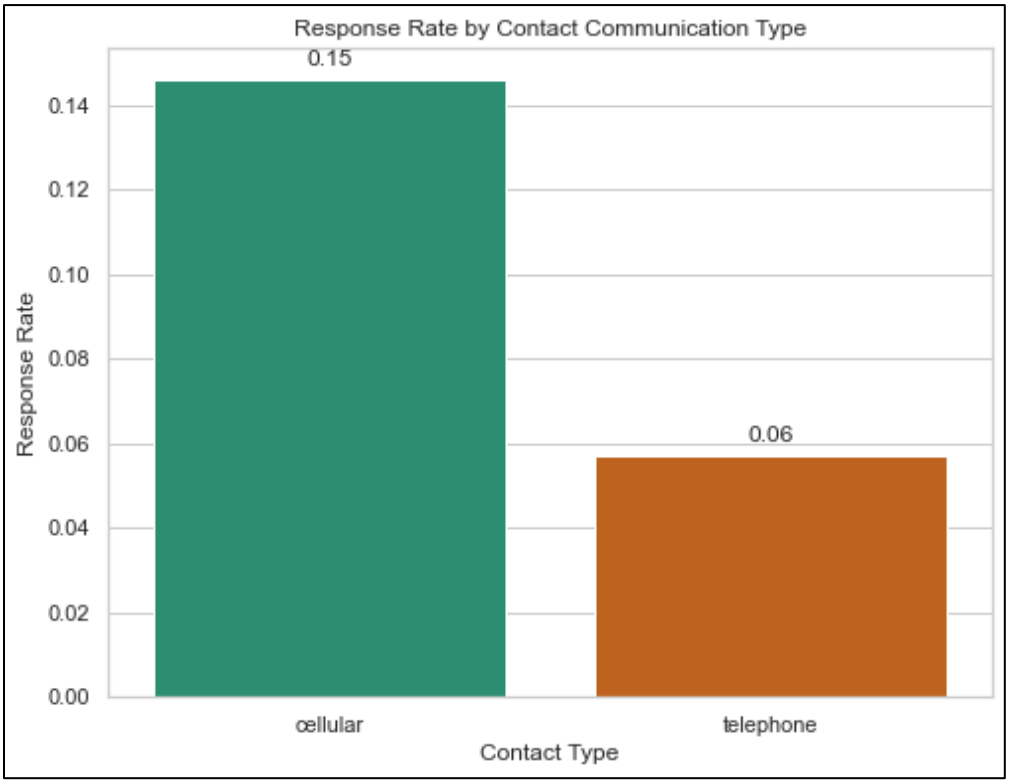
3.1 Response Rate by Job and Education: Response rates were calculated for different job categories and education levels.



- Higher response rates among students and retired individuals, indicating their receptiveness to marketing efforts.
- Lower response rates in job categories like entrepreneurs, services, and blue-collar workers.

- The highest response rate was observed in the 'Illiterate' education category, followed by those with university degrees and unspecified education levels.

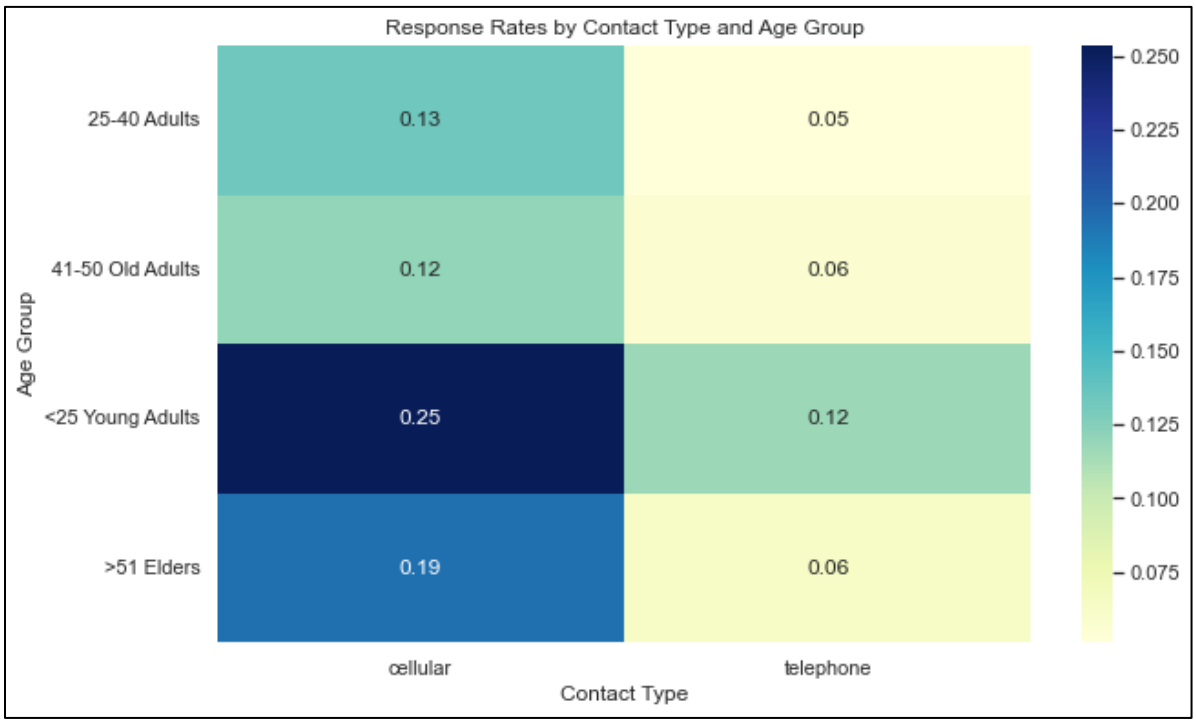
3.2 Response Rates by Contact Communication Type:

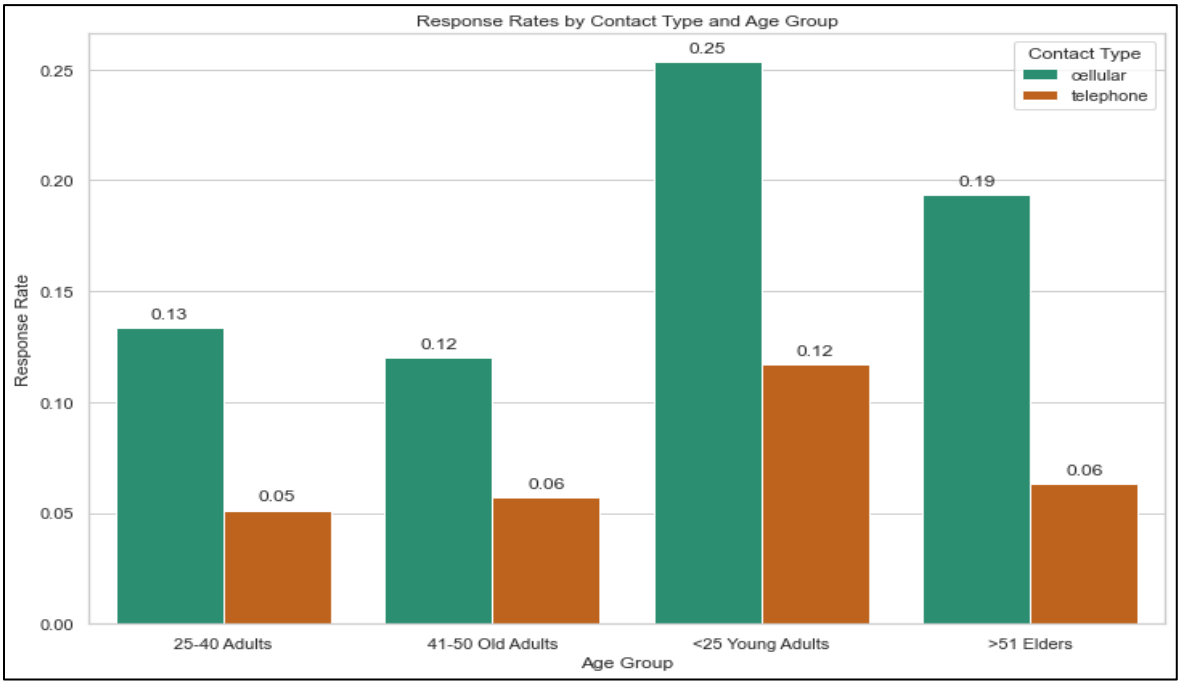


- Response rates were higher for cellular contacts (15%) compared to telephone contacts (6%).
- The data suggests a preference for cellular communication, possibly due to its convenience and personal nature.

3.3 Contact Communication Type-Responses Relationship Across Different Age Groups:

Age groups were defined (<25, 25-40, 41-50, >51) and response rates were analyzed by contact type.

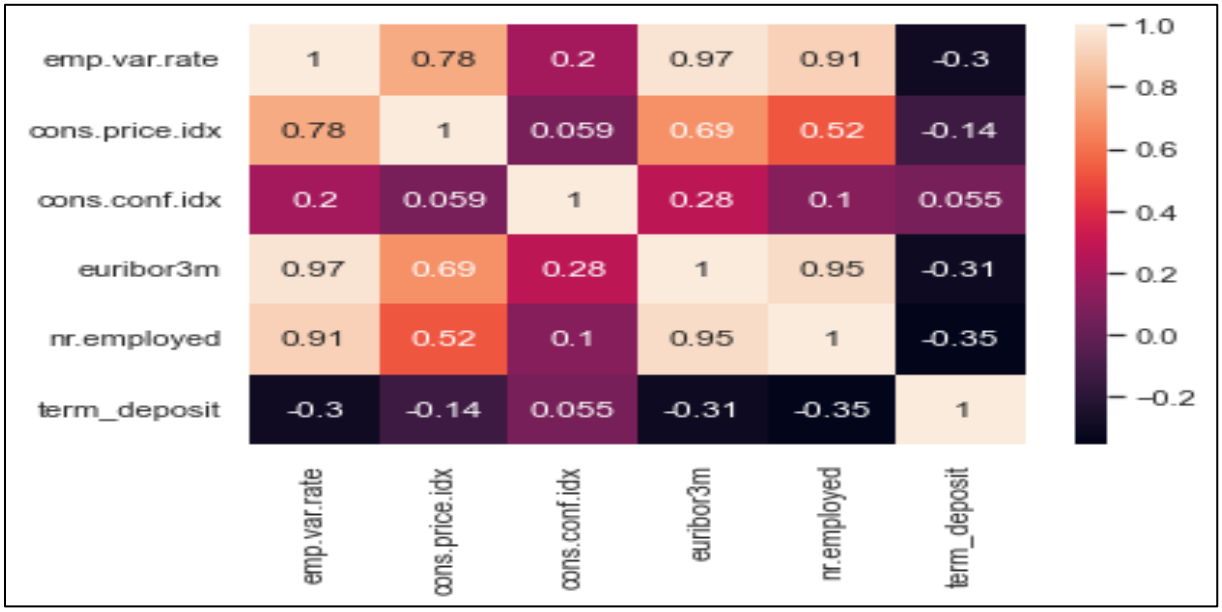




- The highest overall responsiveness in the <25 age group, especially for cellular contacts.
- The 25-40 age group exhibited the least responsiveness to telephone contacts.

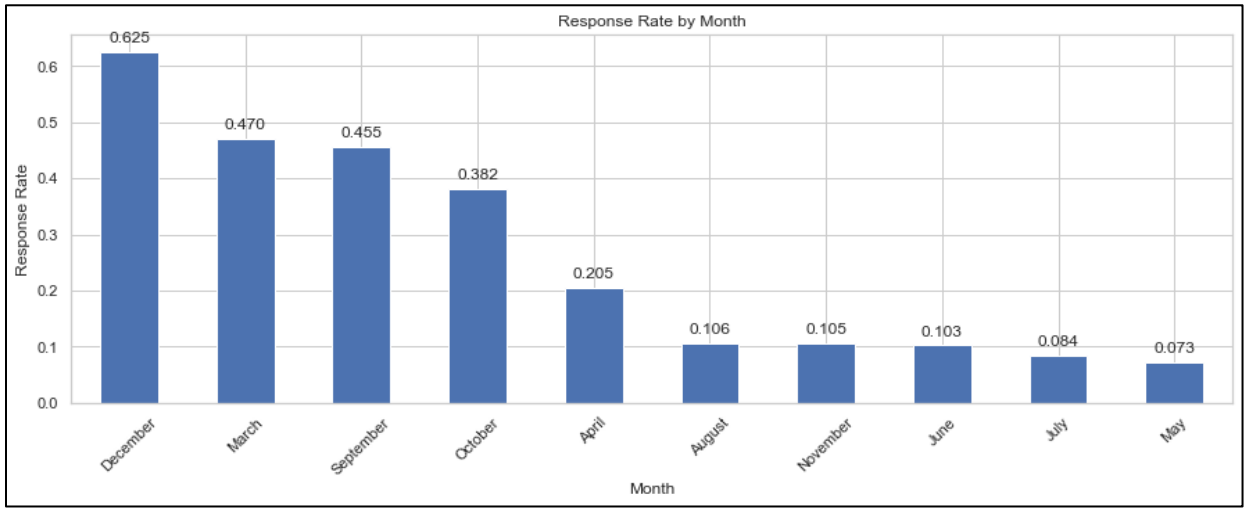
- Consistently lower response rates for telephones across all age groups, indicating a trend towards preference for cellular communication.

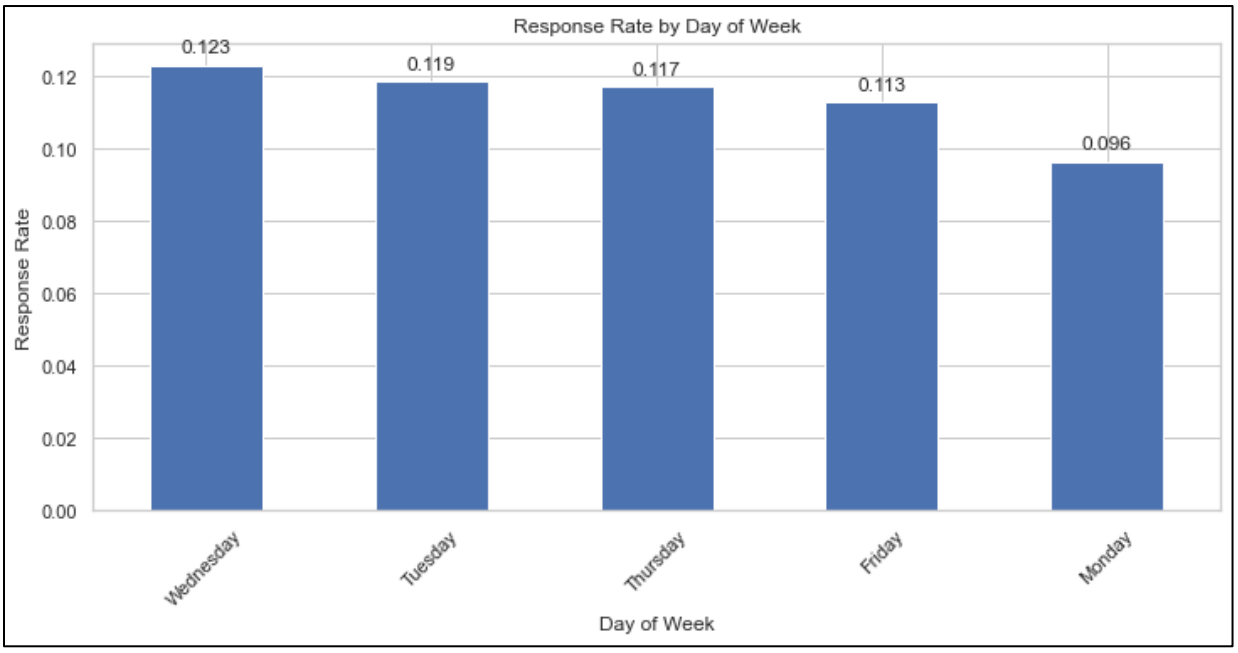
3.4 Patterns in Economic Indicators:



- Economic factors like employment variation rate, consumer price index, consumer confidence index, Euribor 3-month rate, and number of employed were correlated with term deposit subscription.
- A heatmap of the correlation matrix highlighted negative correlations between these economic indicators and term deposit subscriptions, suggesting their influence on customer decisions.

3.5 Patterns in Month and Day of Week by Response Rates:





- Response rates varied significantly across months, with December showing the highest response rate.
- The effectiveness of marketing strategies seemed to fluctuate with the time of the year, indicating the impact of seasonal behaviors on customer engagement.
- Analysis of response rates by day of the week revealed Wednesday as having the highest response rate, with Monday being the lowest.

MODEL BUILDING

4.1 Logistic regression supervised machine learning -

In this section I utilize logistic regression, a supervised machine learning technique to understand features that affect the likelihood of customers subscribing to a term deposit. This approach aims to identify key factors influencing customer decisions and enhance the effectiveness of the bank's marketing strategies.

- **Encoding and Normalization**: The dataset was preprocessed to transform categorical variables using **LabelEncoder** and normalize numerical features with **MinMaxScaler**. This step ensures that the model accurately interprets the data without bias towards variables with larger scales.
- **Feature Selection**: The initial model included a broad range of demographic and economic indicators. Significant features were later identified based on their statistical impact on the model's predictive capability.

4.2 Model Evaluation and Interpretation

Initial Model Findings

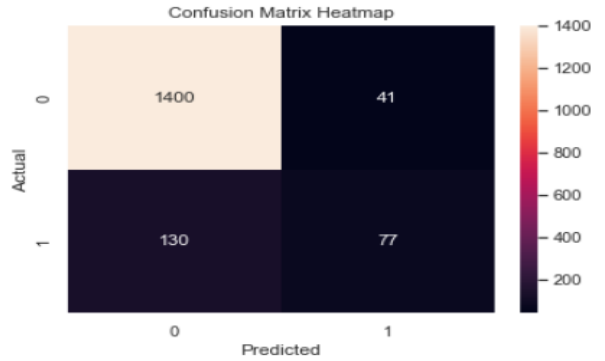
The initial model was evaluated based on accuracy, precision, recall, F1 score, ROC score, and a confusion matrix was generated. The key metrics are presented in Table 1 below.

Logit Regression Results

Dep. Variable:	term_deposit	No. Observations:	6590
Model:	Logit	Df Residuals:	6569
Method:	MLE	Df Model:	20
Date:	Sun, 24 Dec 2023	Pseudo R-squ.:	0.4106
Time:	22:46:25	Log-Likelihood:	-1349.8
converged:	True	LL-Null:	-2290.0
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3588	1.011	-0.355	0.723	-2.340	1.622
age	0.9467	0.360	2.631	0.009	0.242	1.652
job	-0.0073	0.013	-0.552	0.581	-0.033	0.019
marital	0.1618	0.092	1.759	0.079	-0.019	0.342
education	0.0520	0.022	2.405	0.016	0.010	0.094
default	-0.4661	0.165	-2.820	0.005	-0.790	-0.142
housing	0.0855	0.099	0.866	0.387	-0.108	0.279
loan	-0.1329	0.140	-0.952	0.341	-0.407	0.141
contact	-0.5094	0.161	-3.174	0.002	-0.824	-0.195
month	-0.1164	0.023	-5.012	0.000	-0.162	-0.071
day_of_week	0.0643	0.037	1.741	0.082	-0.008	0.137
duration	16.9912	0.657	25.875	0.000	15.704	18.278
campaign	0.7966	1.044	0.763	0.445	-1.249	2.842
pdays	-1.1172	0.391	-2.855	0.004	-1.884	-0.350
previous	-0.8337	0.793	-1.052	0.293	-2.387	0.720
poutcome	0.5285	0.186	2.840	0.005	0.164	0.893
emp.var.rate	-3.3584	0.855	-3.928	0.000	-5.034	-1.683
cons.price.idx	0.8990	0.810	1.110	0.267	-0.688	2.486
cons.conf.idx	-0.0062	0.421	-0.015	0.988	-0.832	0.820
euribor3m	2.7215	1.163	2.340	0.019	0.442	5.001
nr.employed	-4.5644	1.272	-3.589	0.000	-7.057	-2.072

```
Model: Logistic Regression
Accuracy: 89.62%
Precision: 65.25%
Recall: 37.20%
F1 score: 47.38%
ROC score: 67.18%
Confusion matrix:
 [[1400   41]
 [ 130   77]]
```



Significant Features Model

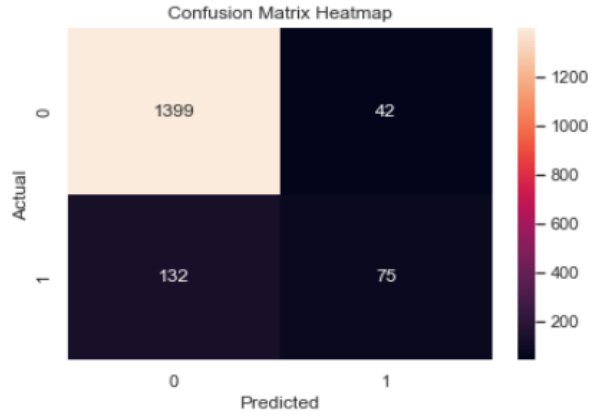
A second model incorporating only statistically significant features Based on p-values and coefficients was then evaluated. These includes 'duration', 'pdays', 'poutcome', 'emp.var.rate', 'euribor3m', 'nr.employed', 'age', 'education', 'default', 'contact', and 'month'. This model's performance metrics are compared against the initial model, as shown in Table 2 (attached below).

Logit Regression Results

Dep. Variable:	term_deposit	No. Observations:	6590
Model:	Logit	Df Residuals:	6578
Method:	MLE	Df Model:	11
Date:	Sun, 24 Dec 2023	Pseudo R-squ.:	0.4080
Time:	22:46:59	Log-Likelihood:	-1355.8
converged:	True	LL-Null:	-2290.0
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0356	0.420	0.085	0.932	-0.788	0.859
duration	17.0687	0.656	26.027	0.000	15.783	18.354
pdays	-0.7749	0.230	-3.362	0.001	-1.227	-0.323
poutcome	0.6727	0.137	4.899	0.000	0.404	0.942
emp.var.rate	-2.4159	0.522	-4.626	0.000	-3.439	-1.392
euribor3m	2.7604	0.691	3.994	0.000	1.406	4.115
nr.employed	-5.2970	0.540	-9.810	0.000	-6.355	-4.239
age	0.6268	0.326	1.921	0.055	-0.013	1.266
education	0.0506	0.021	2.353	0.019	0.008	0.093
default	-0.4759	0.165	-2.880	0.004	-0.800	-0.152
contact	-0.4442	0.144	-3.079	0.002	-0.727	-0.161
month	-0.1203	0.021	-5.721	0.000	-0.161	-0.079

```
Model: Logistic Regression with Significant Features
Accuracy: 89.44%
Precision: 64.10%
Recall: 36.23%
F1 score: 46.30%
ROC score: 66.66%
Confusion matrix:
 [[1399   42]
 [ 132   75]]
```



Cross-Validation

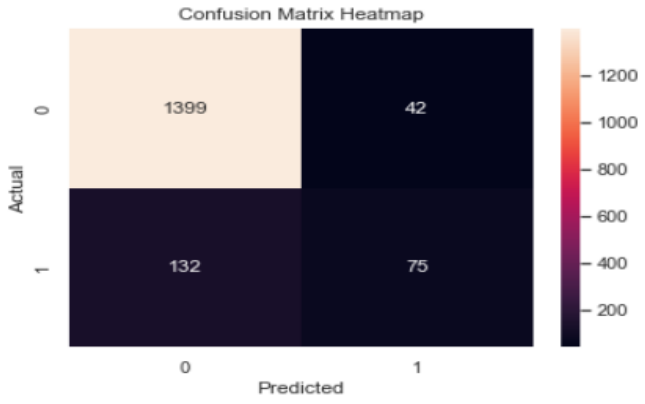
After constructing a new model with significant variables, it is important to cross validate to improve the efficiency of the model. this is important to improve reliability and prevent over-fitting. shown in Table 3 below.

Logit Regression Results

Dep. Variable:	term_deposit	No. Observations:	6590
Model:	Logit	Df Residuals:	6578
Method:	MLE	Df Model:	11
Date:	Sun, 24 Dec 2023	Pseudo R-squ.:	0.4080
Time:	22:52:20	Log-Likelihood:	-1355.8
converged:	True	LL-Null:	-2290.0
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0356	0.420	0.085	0.932	-0.788	0.859
duration	17.0687	0.656	26.027	0.000	15.783	18.354
pdays	-0.7749	0.230	-3.362	0.001	-1.227	-0.323
poutcome	0.6727	0.137	4.899	0.000	0.404	0.942
emp.var.rate	-2.4159	0.522	-4.626	0.000	-3.439	-1.392
euribor3m	2.7604	0.691	3.994	0.000	1.406	4.115
nr.employed	-5.2970	0.540	-9.810	0.000	-6.355	-4.239
age	0.6268	0.326	1.921	0.055	-0.013	1.266
education	0.0506	0.021	2.353	0.019	0.008	0.093
default	-0.4759	0.165	-2.880	0.004	-0.800	-0.152
contact	-0.4442	0.144	-3.079	0.002	-0.727	-0.161
month	-0.1203	0.021	-5.721	0.000	-0.161	-0.079

```
Accuracy: 89.44%
Precision: 64.10%
Recall: 36.23%
F1 score: 46.30%
ROC score: 66.66%
Confusion matrix:
 [[1399   42]
 [ 132   75]]
```



RESULT AND DISCUSSION

5.1. Do job backgrounds and education levels affect customers responding rate significantly?

- **Education:** The coefficient for education is 0.0506 with a p-value of 0.019. This suggests that education levels do have a statistically significant effect on the likelihood of a customer responding positively to a term deposit offer. The positive coefficient indicates that higher education levels are associated with a higher likelihood of subscription.
- **Job:** In the initial model, the coefficient -0.0073 with a p-value of 0.581 indicates that while job backgrounds were initially considered in the first model, they don't have a statistically significant effect on the likelihood of a customer responding to a term deposit offer. The high p-value suggests that the impact of job backgrounds is not strong enough to be considered significant effect.
- **Summary :** Education levels are a significant predictor of customer response rates to term deposit offers. Job backgrounds were not significant.

5.2. Is there a relationship between customer response rates and the contact communication type?

The coefficient for the contact variable is -0.4442 with a p-value of 0.002. This indicates a statistically significant relationship between the contact communication type and the customer response rate. The negative coefficient suggests that certain types of contact (telephone / cellular) are associated with a lower likelihood of a customer subscribing to a term deposit.

- **Summary** : There is a significant relationship between contact communication type and customer response rates.

5.3. Is the contact communication type -responses relationship identified in the previous question the same for all age groups?

```
Age Group: >51 Elders, Chi-square test P-value: 9.651556239834825e-11
Age Group: 25-40 Adults, Chi-square test P-value: 1.840350899186192e-17
Age Group: 41-50 Old Adults, Chi-square test P-value: 1.6912408433731314e-0
6
Age Group: <25 Young Adults, Chi-square test P-value: 0.006178376333066204
```

The chi-square test results indicate that there is a statistically significant relationship between contact type and term deposit response for each of the age groups:

- **Age Group >51 Elders:** The very low p-value (9.65e-11) suggests that there is a statistically significant association between contact type and term deposit responses in this age group.
- **Age Group 25-40 Adults**: a very low p-value (1.84e-17) indicates a strong statistical significance of the relationship between contact type and term deposit responses for adults aged 25 to 40.
- **Age Group 41-50 Old Adults**: The low p-value (1.69e-06) also indicates a significant association between contact type and response rates in this age group.
- **Age Group <25 Young Adults:** The p-value (0.0062) is less than 0.05, there is a statistically significant relationship in this age group although the level of significance is lower compared to the other groups.

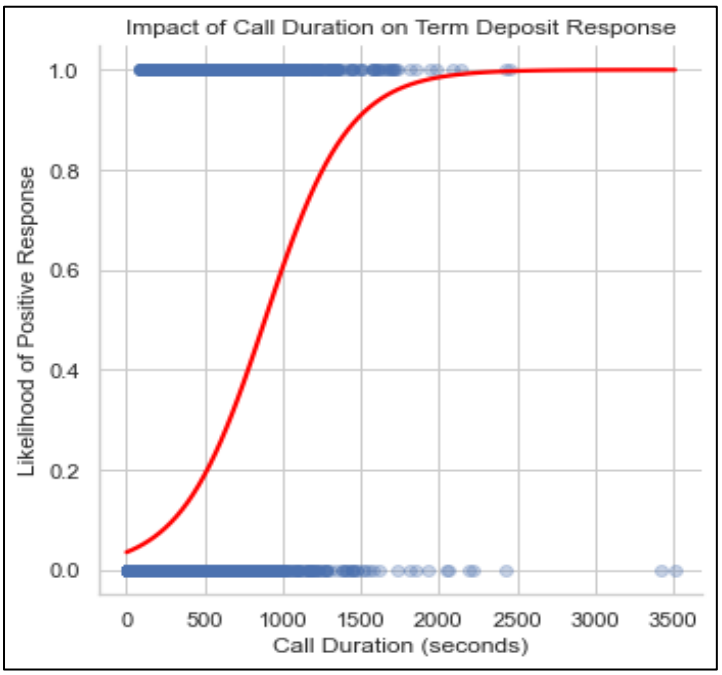
Overall, the results suggest that the type of contact (telephone / cellular) does indeed have a significant impact on the likelihood of a customer responding to a term deposit offer and this impact varies across different age groups.

5.4.1 Significant Impact of Call Duration on Customer Response:

- Coefficient: 17.0687
- P-value: <0.000
- Interpretation: Highly significant. Longer call durations increase the likelihood of a term deposit subscription.

The analysis strongly indicates that the duration of the call plays a pivotal role in predicting customer responses to term deposit offers. Specifically, the longer the duration of the call, the higher the likelihood of a positive response. This finding suggests that customers who engage in conversations longer with bank representatives might be more receptive to understanding the details of the term deposit offer and might have a higher chance of agreement.

This insight emphasizes the importance of quality engagement in customer interactions. It suggests that instead of focusing solely on the number of calls or contacts made, banks should also consider the depth and effectiveness of each interaction. Training customer service representatives to engage in more meaningful and informative conversations could increase the success rate of these marketing calls.

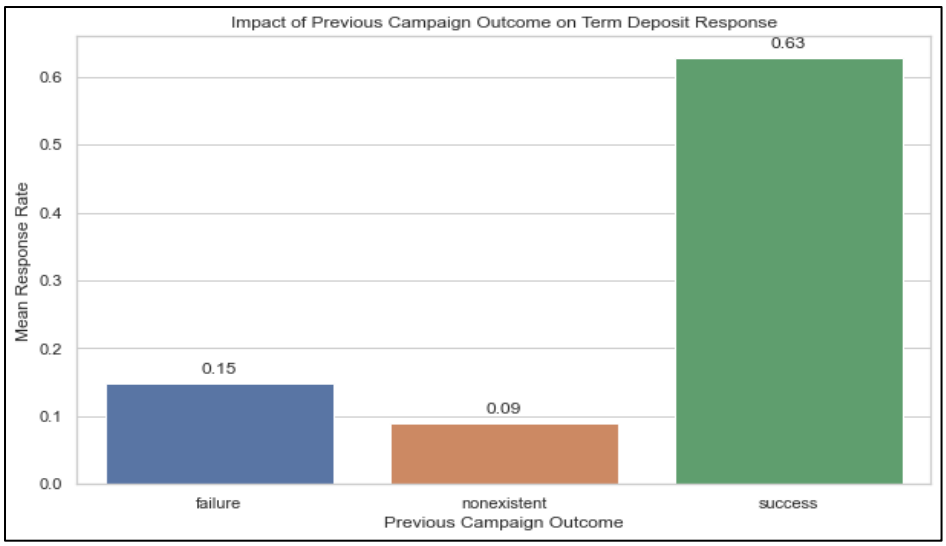


The graph also further validates the impact of call duration on the likelihood of a customer agreeing to a term deposit.

5.4.2 Significant Impact of pout come (Previous Campaign Outcome)

- Coefficient: 0.6727
- P-value: <0.000
- Interpretation: Highly significant positive impact. Successful outcomes in previous campaigns increase the likelihood of subscription.

Statistically "poutcome" in the logistic regression output is significant at $p < 0.05$, with a coefficient estimate of 0.6727. This positive coefficient indicates that a successful outcome in the previous campaign is a strong predictor of a positive response in the current campaign.



This bar plot shows three categories of previous campaign outcomes: failure, nonexistent, and success further shows the pattern with response to customer response. This substantial difference underscores the importance of successful customer engagement in previous campaigns for future campaign effectiveness.

CONCLUSION

In this report i conducted an in-depth analysis of a dataset from a Portuguese bank, focusing on the factors influencing customer decisions to subscribe to a term deposit. The study involved data manipulation, exploratory analysis, and predictive modeling using logistic regression. The findings provided valuable insights into customer behavior and response patterns, highlighting the role of demographics, communication methods, call duration, and previous campaign outcomes in influencing customer decisions.

RECOMMENDATIONS

Based on the analysis I can recommend the following

- **Demographic-Focused Marketing**: Develop marketing strategies that cater to specific customer groups while considering their educational background and preferred communication methods.
- **Emphasis on Mobile Communication**: Prioritize cellular communication for customer contacts, as it showed higher response rates compared to traditional telephone methods.
- **Investment in Quality Interactions**: Enhance customer service training to focus on longer and more engaging conversations, as longer call durations correlated with higher subscription rates.
- **Utilization of Past Campaign Data**: Analyze successful past campaigns to identify effective strategies that can be replicated in future effort