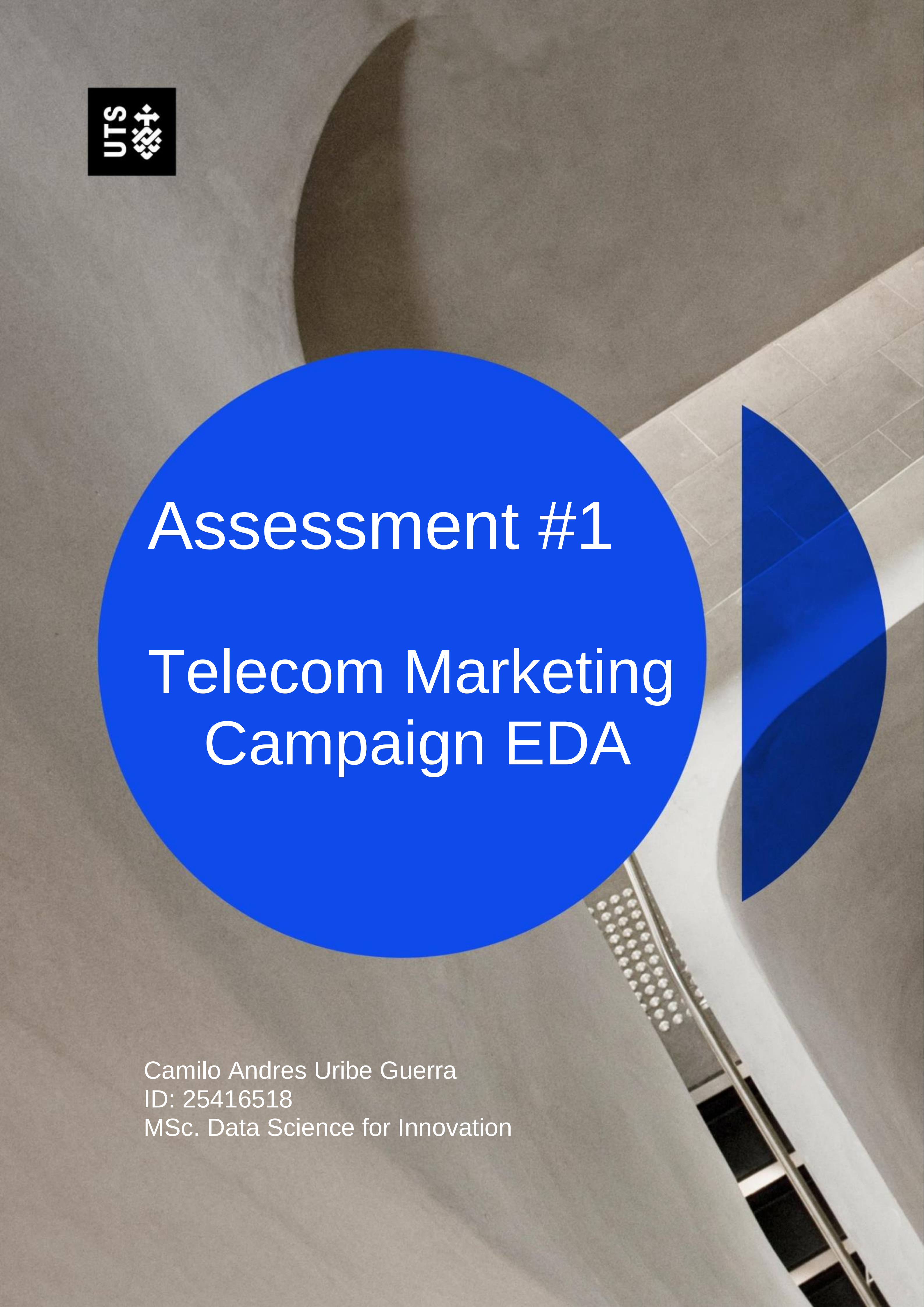


Assessment #1

Telecom Marketing Campaign EDA

Camilo Andres Uribe Guerra
ID: 25416518
MSc. Data Science for Innovation

Table of Contents

Table of Figures

Problem Formulation

Problem Formulation

Problem statement

This initial project assessment aims to conduct an Exploratory Data Analysis (EDA) for a marketing campaign by a mobile company that has recently launched a new subscription plan. The goal is to gain a comprehensive understanding of customer segmentation and identify the groups with the highest responsiveness to the campaign.

Rationale

In today's competitive market, telecommunications companies must continuously seek strategies to retain existing customers and attract new ones. In this context, mobile businesses face increasing competition to secure more subscribers through marketing campaigns that offer renewal discounts or sign-up incentives. However, the success of these plans depends on the ability to design marketing campaigns with a clear customer-focused objective.

This project examines a company's marketing campaign aimed at cross selling a new subscription plan to its customers and identifies the most responsive customer segments.

Primary Objective

- Identify key factors influencing customer subscription behavior in the mobile company's marketing campaign through exploratory data analysis (EDA).

Secondary Objectives

- dentify the key factors influencing customer behavior by analyzing categorical and numerical features through EDA and correlation/association methodologies.
- egment customers using demographic data to identify target groups that are most likely to subscribe to the mobile company.
- Analyze customer-interaction variables that influence subscription rates to enhance or improve campaign outcomes through better practices

Data Preprocessing

Data Preprocessing

Dataset Overview

The dataset utilized in this research is provided by the results of a marketing campaign conducted by the mobile telecom company. This dataset includes key customer attributes and interaction data which is the base to generate the customer segmentation.

The dataset was provided as a CSV file Telecom_Data.csv, which contains 41180 entries and 29 features. Each of the rows represents a customer, and the columns represent different attributes, i.e., age, job, marital status, among others. It also contains information relevant to the interaction that the call center agent had with the customer, i.e., the duration of the call, and whether or not the customer was contacted as part of a previous campaign. Besides the 20 features we have the target variable *y* which indicates whether the customer subscribed to the new plan following the marketing campaign. In addition to the data itself, a data dictionary (A. Data Dictionary) was provided, which contains detailed descriptions of each variable in the dataset.

Data preparation

After loading the dataset using Python and data analytics libraries such as pandas, numpy, and csv, the next step involved ensuring that the dataset was clean and ready for the exploratory data analysis (EDA) process. The dataset was first checked for null values, and no missing data was found.

On the other hand, with a closer look, 12 duplicate entries were found. These, however may depict different customers, as there is no unique identifier that sets it apart from one and another client. Thus, it may well be possible that these entries are valid even though they fall on the same features. Numeric data such as age, duration, and campaign were assigned the correct types for accurate analysis, and categorical variables were also properly assigned to ensure correct handling during processing. Invalid entries in numeric data were converted to NaN to prevent errors.

Furthermore, the categorical variables' unique values were reviewed to identify inconsistency or error. Similarly, numeric variables went through the same process in order to maintain data integrity prior to further analysis. erall, the data preprocessing ensured that the dataset was clean, properly structured, and ready for accurate exploratory data analysis. Also, the target variable was transformed using label encoding, with 1 indicating a customer subscribed and 0 indicating they did not subscribe,enabling better analysis and facilitating correlation calculations with numerical features.

Lastly, for some of the analysis, the unknown values were removed because they do not provide useful insights for decision-making when selecting variables. As a result, the functions used for plotting prior to processing are influenced by this removal.Top

Exploratory Data Analysis (EDA)

Exploratory Data Analysis – EDA

Now that the data has been initialized and pre-processed the next phase involves proceeding with the exploratory data analysis. EDA helps uncover the insights and generate the recommendations needed to solve the primary objective of this research. It will consist of two parts: one EDA will be about the categorical variables, and the other will be about the numerical ones. In addition to that, each set will be divided into customer attribute variables and interaction variables of the customer from the marketing campaign.

Target Variable

As mentioned before, the target variable (*y*) is a binary column, representing if the customer subscribed to the new plan. Figure 1, shows that the majority of customers, 88.74%, did not subscribe (no), while only 11.26% of customers subscribed to the new plan (yes). This imbalance shows that the dataset is skewed towards customers not subscribing, which may warrant consideration in further analyses beyond the scope of this EDA.

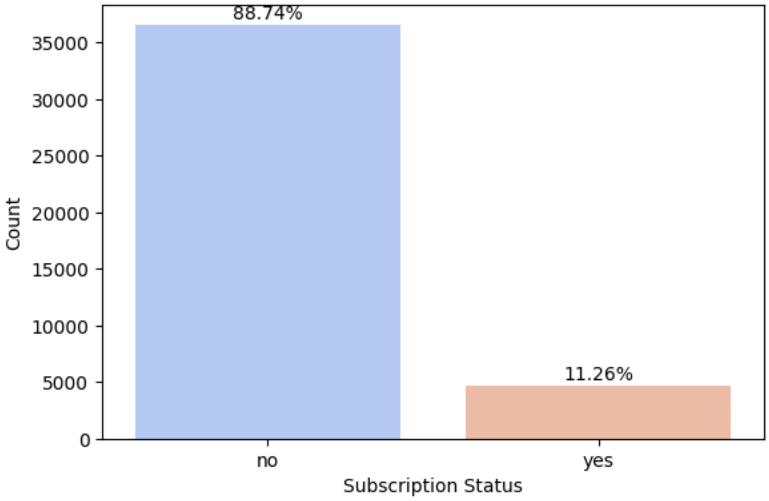


Figure 1. Distribution of subscription status after marketing campaign.

Numerical Features Analysis

The dataset contains important numerical variables that can be analyzed to understand the relationship with the target variable. Among all thefeatures, the only variable which would describe a customer directly is the age variable. The numerical variables may then be divided into company-related and interaction-related. Company-related variables i.e. nr.employed, which represents the number of employees, and emp.var.rate, which means employment variation rate. The interaction-related variables include: duration (in seconds) - call duration; campaign - number of contacts performed during this campaign and for this client; pdays - days passed after the last contact of this client in a previous campaign; previous - number of contacts performed before this campaign with the client; cons.price.idx - consumer price index; cons.conf.idx - consumer confidence index; euribor3m - Euribor 3 month rate.

To understand the relationships between the numeric variables, a correlation heatmap is presented in Appendix B. It shows that most of the relationships are not very strong. However, strong positive correlations are evident among the company-related variables: euribor3m, nr.employed, and emp.var.rate. This makes sense, as changes in the number of employees are naturally tied to variations in the employment rate. Other relationships, such as the negative relationship between pdays and previous, also provide insight into customer interaction patterns; customers who have been inactive for a longer period (higher pdays) are likely to have had fewer interactions in previous campaigns.

More importantly, Figure 2 presents the point-biserial correlation analysis, showing how each of the numerical features is related to the target variable (Kenney & Keeping, 1962). The most important finding is the strong positive relationship of the target variable with a value of 0.41 for duration-that is, longer call durations are associated with higher subscription probabilities. On the other hand, nr.employed (-0.35), pdays (-0.32), and euribor3m (-0.31) show a negative correlation. That is, a larger number of employees, larger periods of inactivity, and higher interest rates are associated with lower subscription rates.

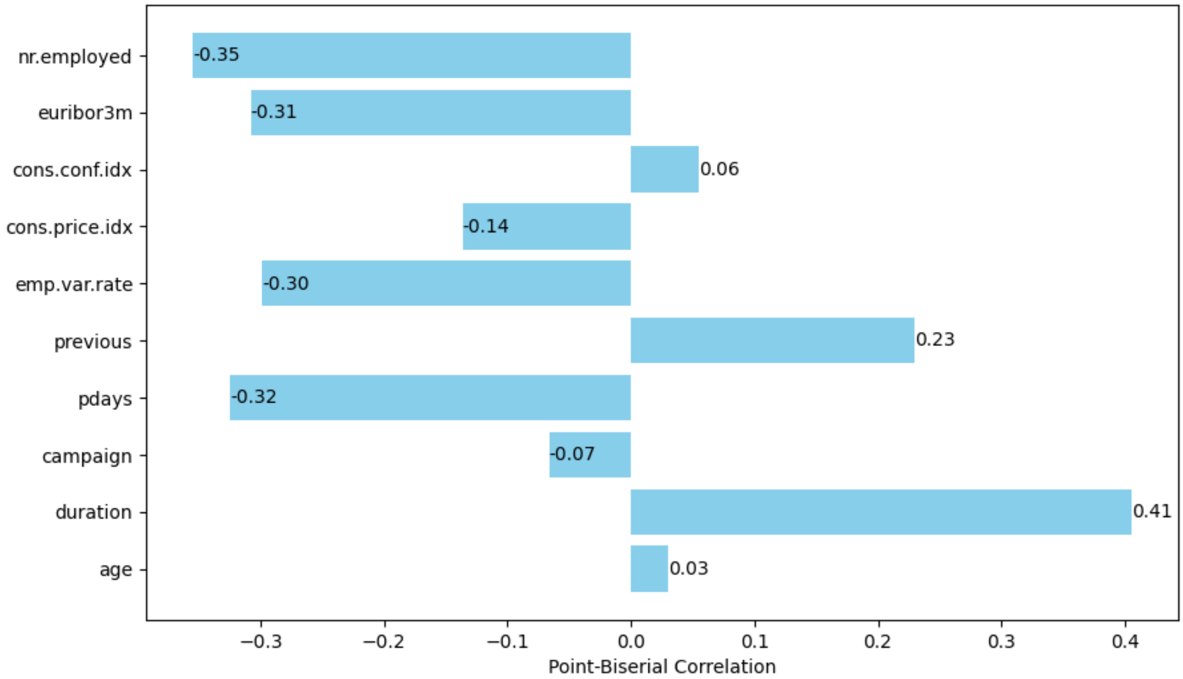


Figure 2. Point-Biserial correlation between subscriptions and numerical features.

Employability Effects

The employment-related variables, including *nr.employed* (number of employees) and *emp.var.rate* (employment variation rate), provide insights into the company's employees dynamics and their influence on customer subscription behavior. Both variables show a negative correlation with the target variable (*y*), indicating that periods with higher numbers of employees and greater changes in employment are associated with lower subscription rates.

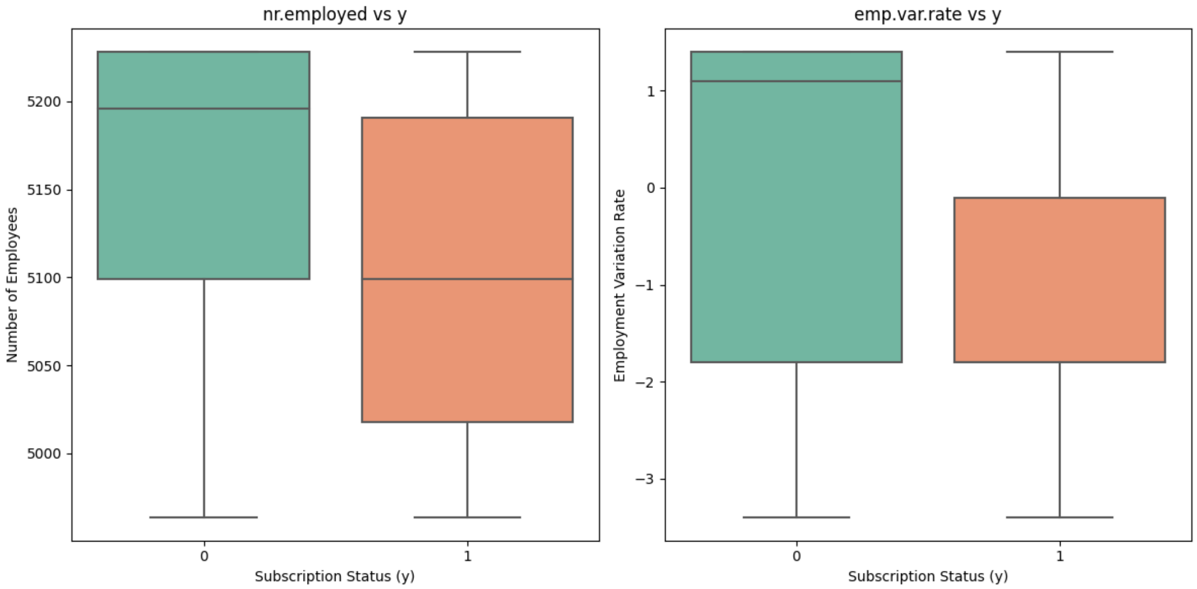


Figure 3. Boxplots of subscription vs employment variables.

Figure 3 presents the box plots of employment variables, The left plot shows that the non-subscribers have higher numbers of employees, while subscribers are associated with lower numbers of employees, confirming the earlier negative correlation.

As shown by the right plot, subscribers had lower rates of variation in employment, while non-subscribers had higher rates. This could just be showing that workforce instability had an adverse effect on subscription rates perhaps due to changes in the quality of the contact-center service or even communication at the time. In conclusion, unstable employee numbers cannot provide effective subscriptions, as stable, experienced employees likely perform better than new hires. More customer-facing employees do not guarantee better results, so fewer but more effective interactions are preferred.

Economic Environment Relation

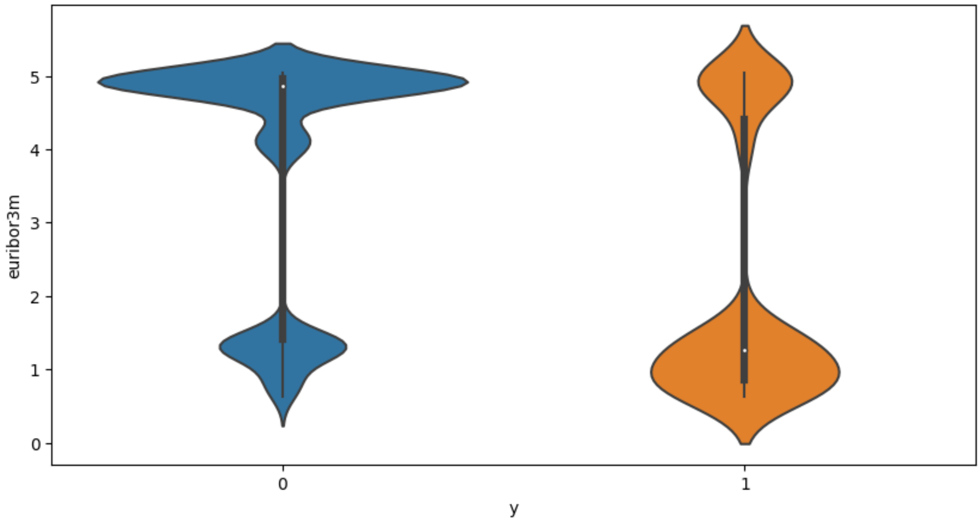


Figure 4. Violin plot of the Euribor3m vs target variable y.

This euribor3m variable stands for the 3-month Euro Interbank Offered Rate and reflects broader economic conditions. A negative correlation of -0.31 testifies that higher interest rates correspond to a worse subscription rate. Subscribers tended to face lower Euribor rates, as Figure 4 violin plot shows, indicating favorable economic conditions for the customers' decisions. Thus, company should focus targeting potential customers during periods of lower interest rates so consumers are more likely to subscribe.

Customer-interaction features

Regarding interaction variables, the most relevant feature is the duration of the call, which has the strongest correlation with the subscription outcome. Duration is highly left-skewed with a mean of 258 seconds and a maximum value of 4918 seconds; because of this, the logging of this variable would be beneficial for modelling. From the boxen plot of Figure 5, subscribers (1) had longer call duration, and for them, the median duration was much higher than non-subscribers (0). For instance, the 75th percentile of subscribers exceeds 319 seconds, and the average duration for non-subscribers is very clear that the longer the conversation and therefore the more emotionally invested, the better it generates subscriptions. Hence, the company should train their employees to focus in having longer and more engaging interactions because this is clearly linked to the subscription results at the end.

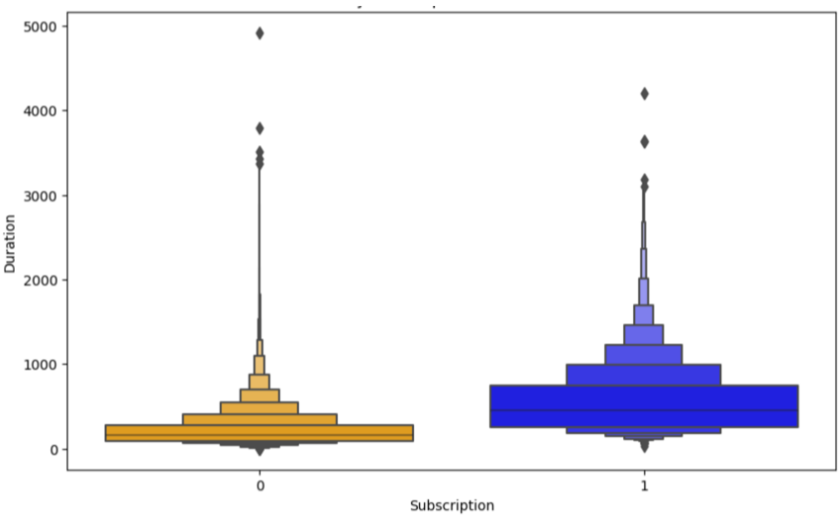


Figure 5. Boxen plot of relationship of call duration with subscription

The variables **pdays** (days since last contact) and **previous** (number of contacts during past campaigns) provide insights into the influence of prior customer interactions on subscription behavior.

It can be observed from the box plot shown on the left-hand side of Figure 6 that, generally customers who had subscribed had lower pdays values as compared to who do not subscribe. This will mean that more recent interaction provides a positive influence on subscription rates. Analysis excludes the 999 values, which represent customers who were not contacted during the previous campaign. More information is conveyed by the right-hand side plot that discretizes pdays: customers called in the past are much more likely to subscribe - over 63% subscription rate, versus about 9% rate for customers never called before.

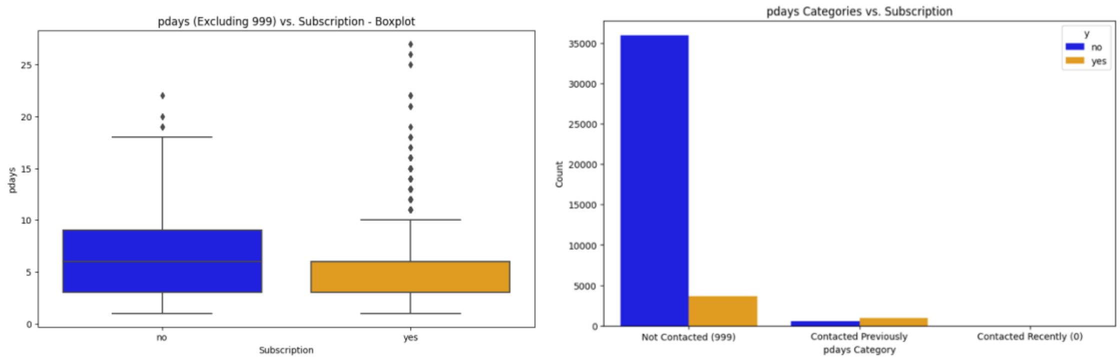


Figure 6. Relationship plots for pdays vs target variable (y)

As shown in Appendix C. Number of Previous Contacts vs y., customers with more prior contacts tend to have higher subscription rates, reinforcing the value of consistent communication. The increase in subscription likelihood as the number of contacts grows suggests that follow-up efforts significantly impact customer engagement and conversion. This shows that there is a limit value of previous contacts so it is better not to exceed more than 5 contacts and should be good to contact the customers more than 3 times before starting a new marketing campaign.

The remaining variables, such as age and the (cons.conf.idx), show very weak correlations with the target, meaning changes in these factors won't significantly affect subscription outcomes. While the consumer price index (cons.price.idx) has a small negative relationship (-0.14), and the number of contacts during the campaign shows a negative but low correlation (-0.07), their overall impact on the likelihood of subscription remains lower compared to the variables mentioned before.

Categorical Feature Analysis

The dataset contains some categorical features which are helpful for understanding both customer characteristics and the interactions. The customer related variables can divide in two groups: demographic data like job, marital status, education and financial status, such as default, housing, loan. On the other hand, interaction features such as contact type, month, day of week, outcome of the previous campaign.

The relationship between categorical variables and the target variable can be assessed using Cramér's V, which measures the strength of association between nominal features and the target (Siegel & Castellan, 1988). This analysis identifies which variables most influence subscription outcomes.

As shown in Figure 7. Crammers V association with y, poutcom*e* (previous campaign result) and *month* (last contact month) exhibit the strongest associations with subscription status, highlighting their importance in understanding customer behavior. Other variables, such as *job* and *contact*, show moderate associations, while *loan* and *housing* have weaker associations, suggesting a lesser impact on subscription likelihood.

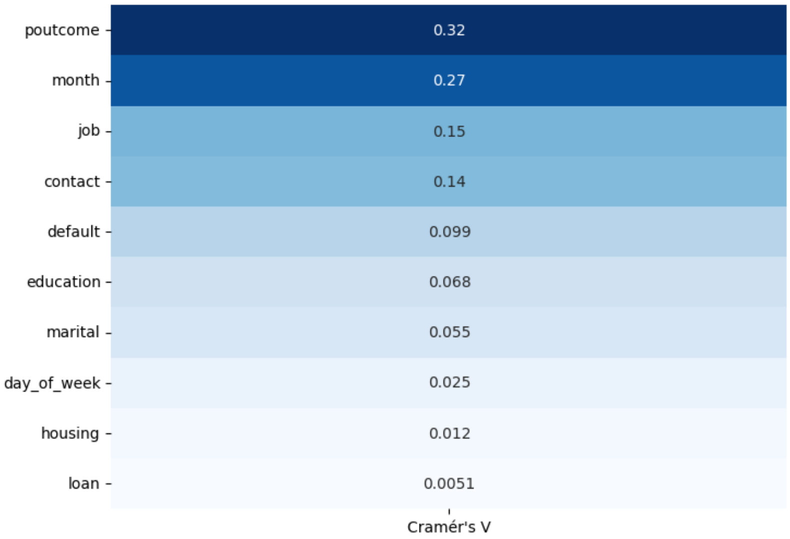


Figure 7. Crammers V association with y

Demographic Influence

Among the demographic features, job shows a notable level of association with subscription status. As illustrated in Figure 8. Bar proportion plot of job vs y target variable., students and retired individuals demonstrate the highest subscription rates, at 31% and 25%, respectively. These two groups are significantly more likely to respond to the campaign compared to other professions. For instance, unemployed and self-employed customers show lower subscription rates, at 14% and 11%, while blue-collar and services workers exhibit even lower responsiveness, around 7%-8%. This suggests that the campaign resonates more with specific professionals, potentially due to lifestyle or financial factors.

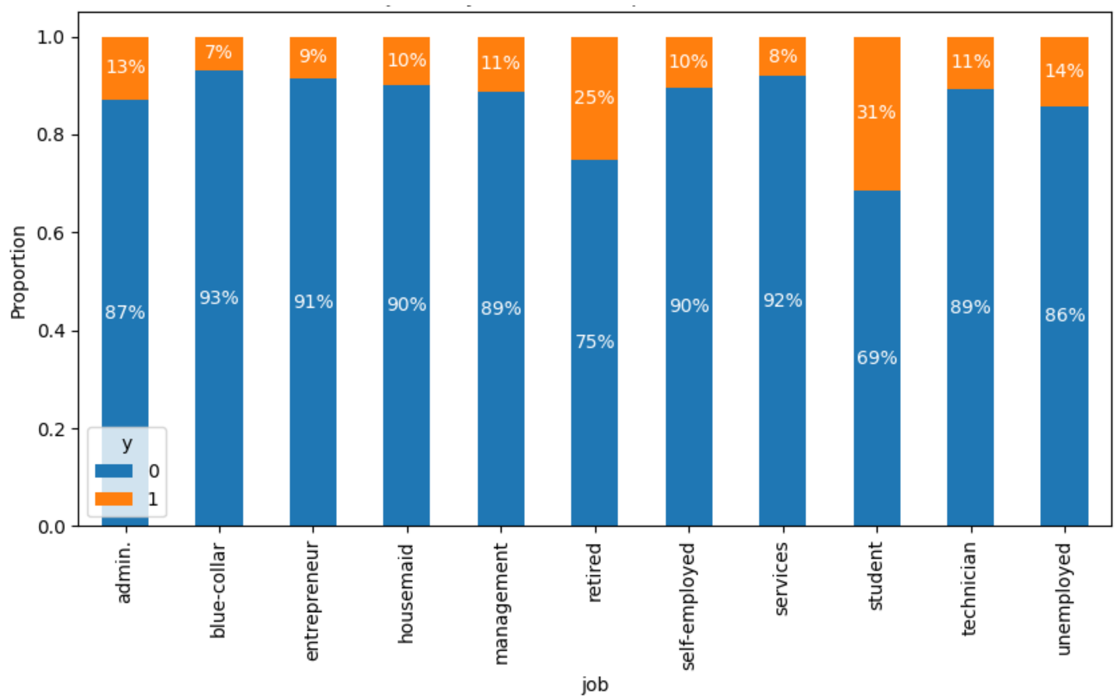


Figure 8. Bar proportion plot of job vs y target variable.

From the remaining demographical variables, marital status and education are also less linked to subscription status. Among marital status, single customers are slightly more likely to subscribe than married or divorced. As for education, illiterate customers or those with university degrees present a higher subscription rate, having a maximum for illiterates of 22%. However, the spread across the education level is relatively small. (Appendix D. Remaining demographical variable influence.)

Customer-interaction features

The categorical customer-interaction features such as the outcome of the previous campaign, month, and type of contact, all give higher insights into how prior engagements and timing affected subscription rates. Among these, poutcome is one of the most indicative variables. From Figure 9. Pie charts representing the relationship between poutcomne with y.customers with a successful outcome in the last campaign have a much higher probability of subscribing again, 65%, in fact. In contrast, customers for whom the previous outcome was a failure or who were not contacted at all have subscription rates of 14% and 9%, respectively. This indicates the prime driver of subsequent subscriptions is successful prior engagement. In conclusion, focusing on these previously engaged groups can yield better results compared to targeting new customers.

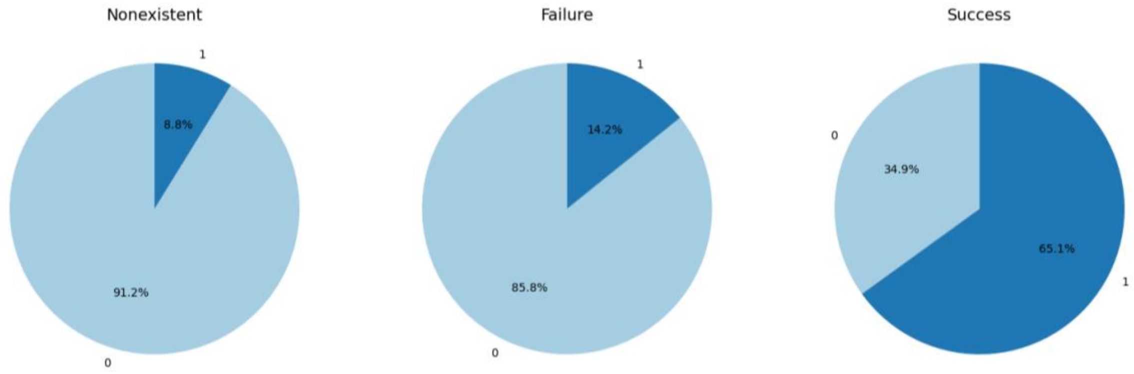


Figure 9. Pie charts representing the relationship between poutcomne with y.

By examining the month and day of the week columns, which represent when the customer was contacted, it is evident that the month column holds significance, while the day of the week does not influence the customer's final decision. March, September, October, and December are the most engaging months, accounting for over 50% of the total subscribers in the dataset. In contrast, the day of the week shows no variation in subscription rates (Appendix E. Month and day customer contact relationship with y.).

Cellular contact has proven much more effective than telephone contact, with 14.73% of customers contacted via cellular subscribing, compared to only 5.23% from telephone contact. This difference is further supported by a Pearson correlation of 0.14, showing a clear preference for cellular contact in successful subscriptions.

Financial Influence

he financial variables, including default, housing, and loan, show minimal influence on subscription behavior. In the case of the default variable, there are no customers with a "yes" value, leaving only "no" and "unknown," which do not provide any valuable insights. Similarly, loan and housing exhibit no significant correlation, making these features less relevant for predicting customer subscriptions.

Conclusions

Conclusions

The main objective of this analysis was to identify the key factors influencing customer subscriptions to mobile plans. The EDA generated the following insights:

1. Focus on longer calls on mobile phones: Subscription rates are higher with longer call durations, especially on mobile phones. Train employees to generate emore engaging and meaningful interactions, extended conversations.
2. Target previously successful and recently contacted customers: Customers who responded positively to previous campaigns are highly likely to subscribe again. Re-target these individuals, and follow up with recently contacted customers. Try to contact more 3-5 times to improve conversion rates at the end.
3. Maintain stability under employees: A consistent and experienced customer-facing employees perform significantly better at securing subscriptions. Rather than increasing the number of employees, focus on getting better quality interactions with well-trained, stable personnel.
4. Prioritize customers that are under the next groups: Focus marketing efforts on students and retirees, as they show the highest subscription rates. Education levels, such as a university degree or lack of formal education, also correspond with higher responses. Additionally, single customers are more likely to subscribe than married ones.
5. Increase campaigns during specific months and economic stability: Launch major campaigns during March, September, October, and December, the months with the highest engagement. Timing is crucial, and low interest rates along with a stable consumer price index are favorable for increased marketing efforts.
6. Financial factors such as loan status, housing, and default are less indicative of subscription behavior. Unless further data suggests otherwise, these variables may be deprioritized in customer targeting.

References

Siegel, S., & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.). McGraw-Hill.

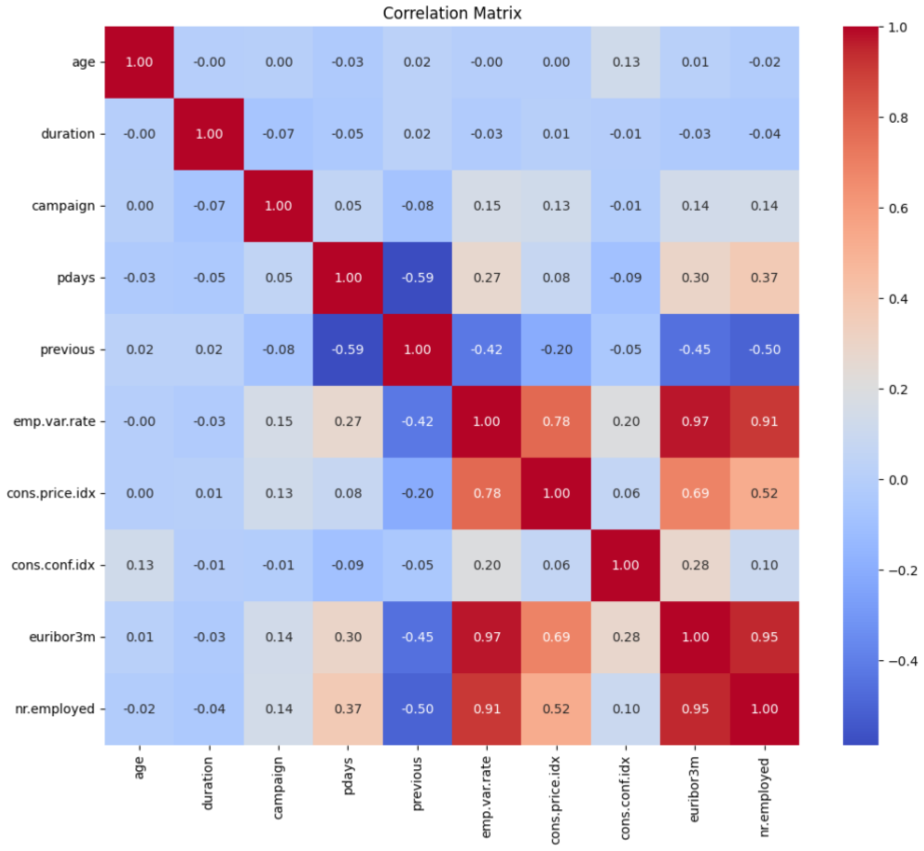
Kenney, J. F., & Keeping, E. S. (1962). Mathematics of statistics (Part 1, 3rd ed.). Van Nostrand.

Appendix

A. Data Dictionary

Variable Name	Description
age	Age
job	Type of job
marital	Marital status
education	Level of education
default	Has credit in default
balance	Average yearly balance
housing	Has a housing loan
loan	Has a personal loan
contact	Contact communication type
day	Day of contact
month	Month of contact
duration	Last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known.
campaign	Number of contacts performed during this campaign and for this client
pdays	Number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
previous	Number of contacts performed before this campaign and for this client
poutcome	Outcome of the previous marketing campaign
emp.var.rate	employment variation rate - quarterly indicator (numeric)
cons.price.idx	consumer price index - monthly indicator (numeric)
cons.conf.idx	consumer confidence index - monthly indicator (numeric)
euribor3m	euribor 3 month rate - daily indicator (numeric)
nr.employed	number employed - quarterly indicator (numeric)
y	Did the client subscribe to a Telecom plan?

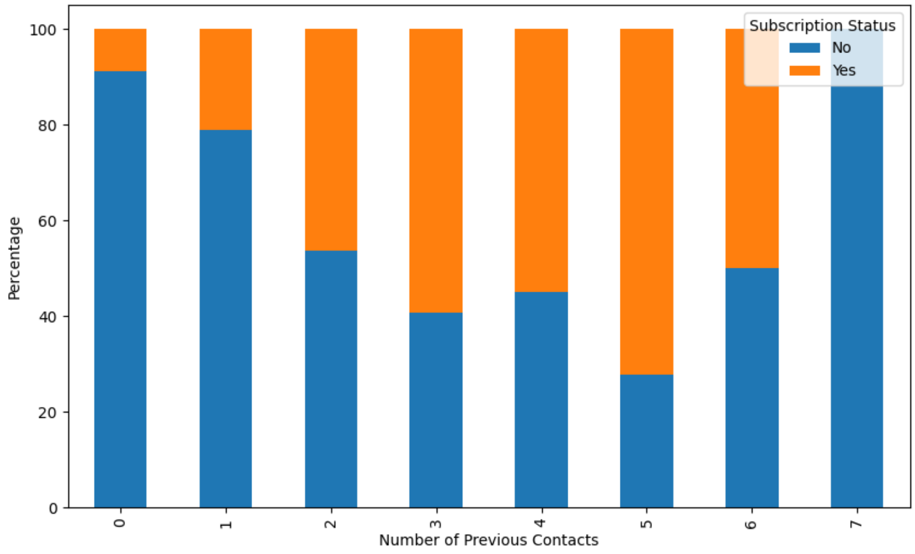
B. Numeric correlation heatmap



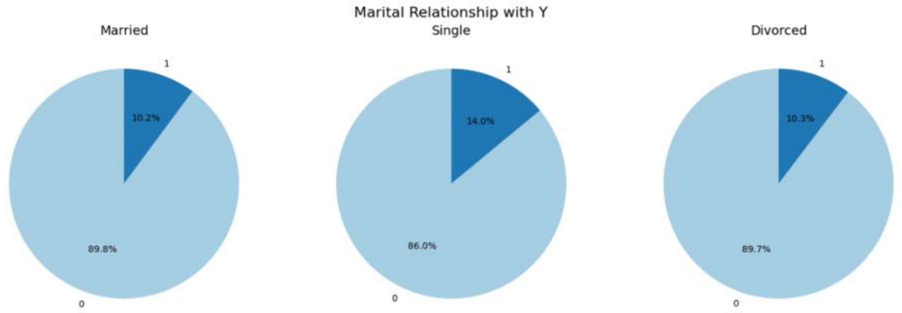
Correlation Matrix

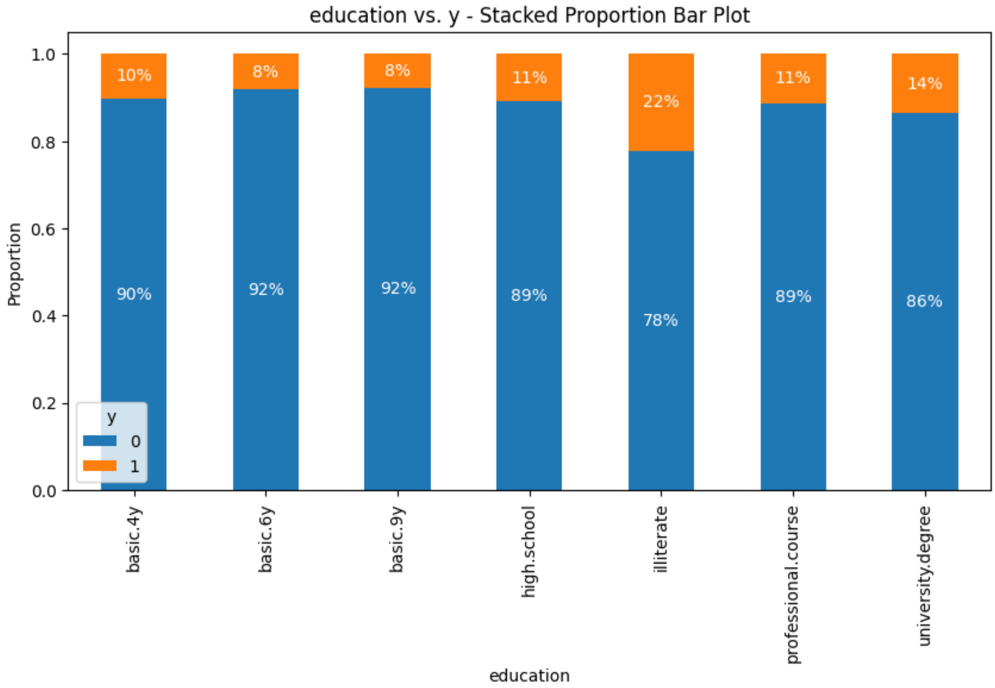
	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1.00	-0.00	0.00	-0.03	0.02	-0.00	0.00	0.13	0.01	-0.02
duration	-0.00	1.00	-0.07	-0.05	0.02	-0.03	0.01	-0.01	-0.03	-0.04
campaign	0.00	-0.07	1.00	0.05	-0.08	0.15	0.13	-0.01	0.14	0.14
pdays	-0.03	-0.05	0.05	1.00	-0.59	0.27	0.08	-0.09	0.30	0.37
previous	0.02	0.02	-0.08	-0.59	1.00	-0.42	-0.20	-0.05	-0.45	-0.50
emp.var.rate	-0.00	-0.03	0.15	0.27	-0.42	1.00	0.78	0.20	0.97	0.91
cons.price.idx	0.00	0.01	0.13	0.08	-0.20	0.78	1.00	0.06	0.69	0.52
cons.conf.idx	0.13	-0.01	-0.01	-0.09	-0.05	0.20	0.06	1.00	0.28	0.10
euribor3m	0.01	-0.03	0.14	0.30	-0.45	0.97	0.69	0.28	1.00	0.95
nr.employed	-0.02	-0.04	0.14	0.37	-0.50	0.91	0.52	0.10	0.95	1.00

C. Number of Previous Contacts vs y.

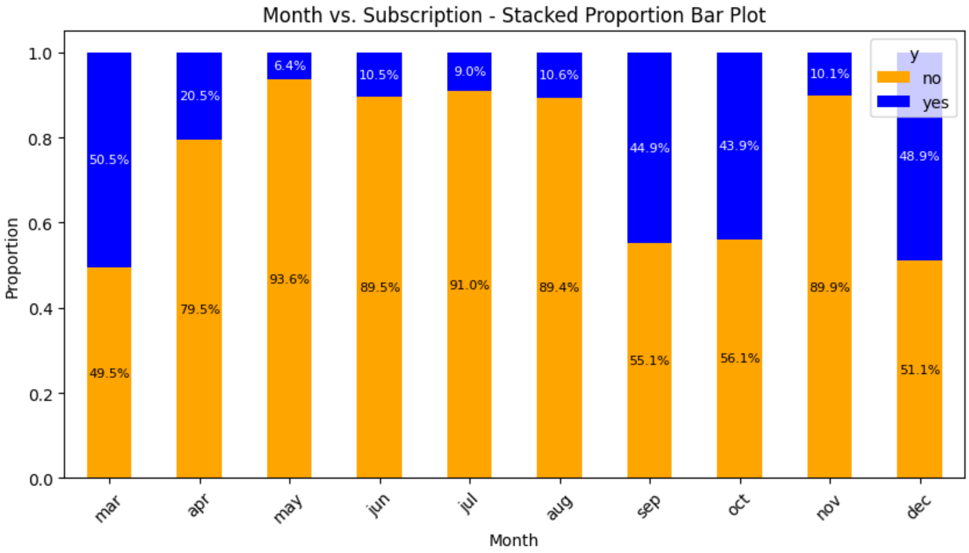


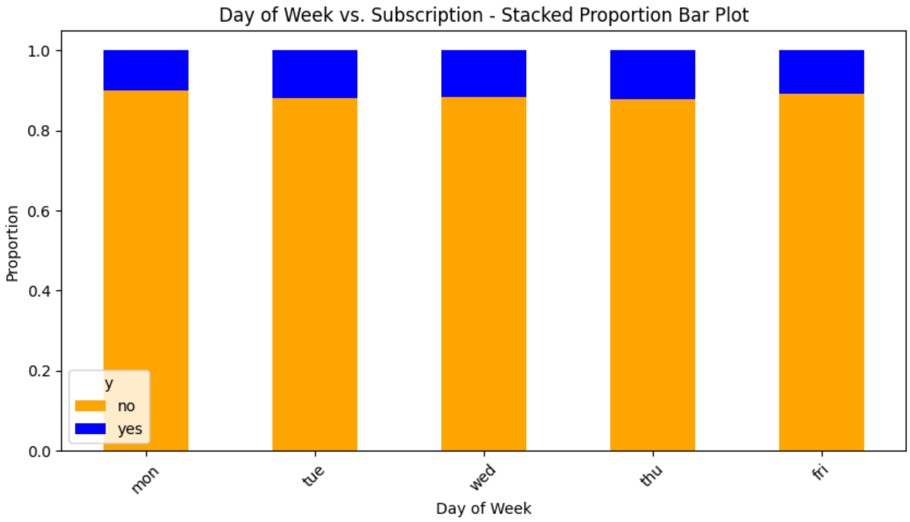
D. Remaining demographical variable influence.





E. Month and day customer contact relationship with y.


Month vs. Subscription - Stacked Proportion Bar Plot
Proportion
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
Month
y
no
yes


Day of Week vs. Subscription - Stacked Proportion Bar Plot
Proportion
mon
tue
wed
thu
fri
Day of Week
y
no
yes