Bank Marketing Data Analysis

Understanding behaviors and trends of people who subscribe to term deposits

W200 Introduction to Data Science Programming
Summer 2022

Cecilia Li, Nishika Abeytunge, Venudhar Ravishankar
School of Information
University of California, Berkeley

Overview

A time deposit or term deposit is a deposit in a financial institution with a specific maturity date. In other words, subscribers lock away an amount of money for an agreed length of time during which they may not access the money until the term is up. The infamous Wells Fargo cross-selling scandal in 2019 involved opening millions of savings and checking accounts on behalf of their clients without consent. While its course of actions is abysmal, many banks engage in marketing their financial products, especially encouraging clients to deposit large sums of money into term deposits. With data science capabilities growing stronger, marketing analytics teams at financial institutions are now able to better predict the likelihood of a client subscribing to a term deposit. Various attributes ranging from age, demographics, income, etc. of different borrowers may be indicators of whether they subscribe.

Research Scope

The purpose of this project is to analyze a banking marketing campaign dataset in order to understand more about the term deposit subscription trends amongst the customers. We want to analyze relationships between factors and how they contribute to whether people subscribe to term deposits.

The report will focus on exploratory data analysis. Statistical analysis and machine learning predictive analytics are out of scope. Our process is as follows:

1. Import data: Import‘ bank additional’ dataset from data source and perform initial high-level analysis
2. Gauge the data: Look at the shape of the file, attributes, missing value, columns and their values respective to the outcome.
3. Clean the data: Remove irrelevant columns, deal with missing and incorrect values, turn categorical columns into dummy variables.
4. Analyze the data: Using Numpy packages to analyze relationships between variables and plot figures of trends and garner insights and findings.

Dataset Description

The data is related to direct marketing campaigns of a Portuguese banking institution from 2008 to 2010. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be (or not) subscribed. There are two datasets: 1) bank-additional-full.csv with over 40K examples 2) bank-additional.csv with 10% of the examples (around 4K), randomly selected from 1). For the purposes of this project, we will be using the file ‘bank-additional’.

There are 20 input variables ranging from age, marital status, income, education, whether they have had a personal loan/housing loan,whether or not they defaulted, etc. There is one output variable - whether the client has subscribed to a term deposit - in the binary form ‘y’ or ‘n’.

In Table 1, we show all names of features, their type, short description and values. Our EDA focuses on characterizing each of the features and their dependencies to result of the campaign.

Table 1: Features description of the Bank Marketing Data set.

Attributes	Kind	Attribute Illustration / Description	Values of attributes
age	numeric	age of client	values between 18 and 95
job	categorical	type of job	’management’, ’technician’, ’entrepreneur’, ’blue-collar’, ’unknown’, ’retired’, ’admin.’, ’services’,

			’self-employed’, ’unemployed’, ’housemaid’, ’student’
marital	categorical	marital status, note: ’divorced’ means divorced or widowed	’divorced’, ’married’, ’single’
education	categorical	degree of education	’primary’, ’secondary’, ’tertiary’, ’unknown’
default	binary	has credit in default?	’no’, ’yes’
balance	numeric	account balance	values between -8019 and 102127
housing	binary	has housing loan?	’no’, ’yes’
loan	binary	has personal loan?	’no’, ’yes’
contact	categorical	contact communication type	’cellular’, ’telephone, ’unknown’
day	numeric	day in month	Values between 1 and 31

month	categorical	last contact month of year	’Jan’, ’Feb’, ’Mar’, ’Apr’, ’May’, ’Jun’, ’Jul’, ’Aug’, ’Sep’, ’Oct’, ’Nov’, ’Dec’
duration	numeric	last contact duration, in seconds	values between 0 and 4918
campaign	numeric	number of contacts performed during this campaign and for this client (included last contact)	values between 1 and 63
p-days	numeric	number of days that passed by after the client was last contacted from a previous campaign, note: 999 means client was not previ- ously contacted	values between -1 and 871
previous	numeric	number of contacts performed before this campaign and for this client	values between 0 and 275
p-outcome	categorical	outcome of the previous market-	’failure’, ’other’,

		ing campaign	'success', 'un-known'
y	binary	has the client subscribed a term deposit?	'no', 'yes'

Data Cleaning

This data set consisted of 41,188 rows and 21 columns. Data set columns were named correctly however in some instances we decided it is more readable if we replace '.' with '_' since then it will not be misinterpreted when coding. So we renamed the columns.

There were no null values, however some of the categorical columns had values as " unknown", where we replaced these with mode values.

	age	job	marital	education	default	housing	loan	contact	month	day_of_
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	
1	57	services	married	high.scho	unknown	no	no	telephone	may	
2	37	services	married	high.school	no	yes	no	telephone	may	
3	40	admin.	married	basic.6y	no	no	no	telephone	may	
4	56	services	married	high.school	no	no	yes	telephone	may	

We analyzed the data types to see if they are on par with the descriptions. and reviewed the statistical analysis of the data set to identify abnormalities such as outliers. Overall all this data set was clean.

Categorical Value EDA

In this section, categorical value analysis focuses on:

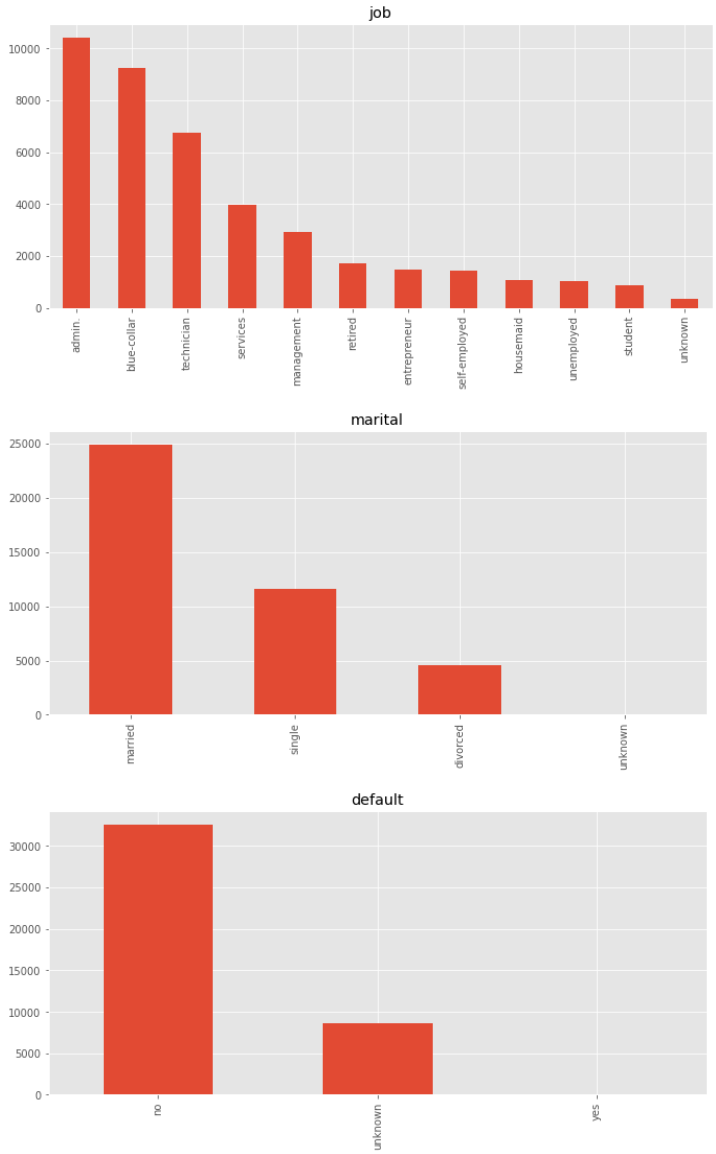
- Illustration of categorical value frequency distribution
- Illustration of distribution of categorical values in respective to output variable (campaign result)

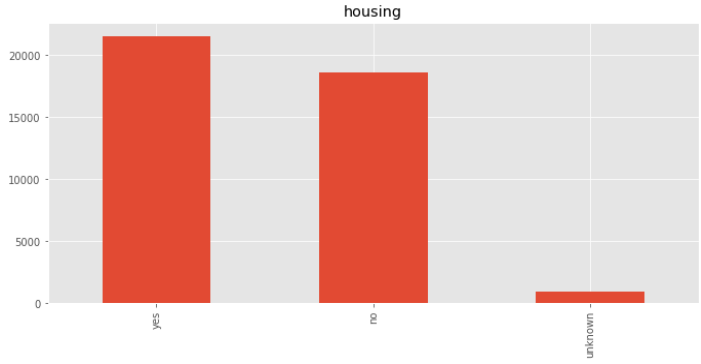
In order to better understand the distribution of input values, we first looked at the imbalance of the output variable (campaign results).

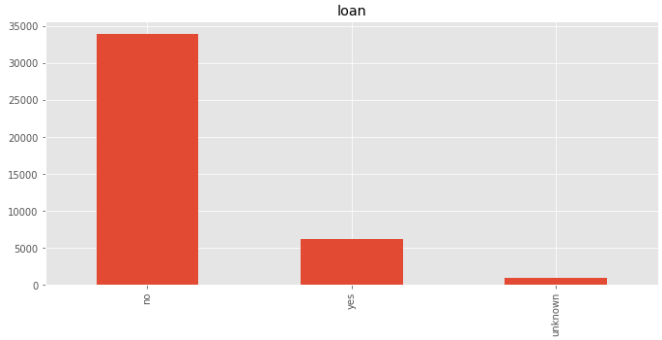
```
no     88.73
yes    11.27
Name: y, dtype: float64
```

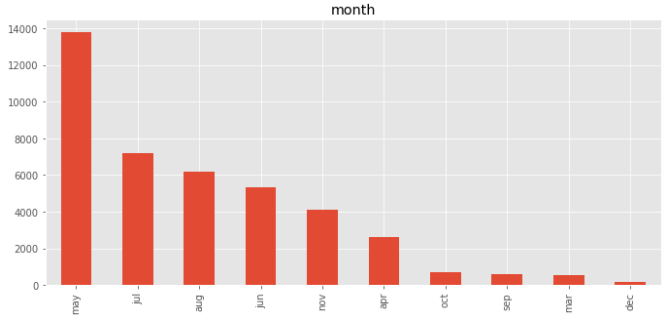
This variable distribution shows that the majority of users actually did not enroll in term deposits. As the next step, we should look at both the input value distributions as well as how they influence the campaign result.

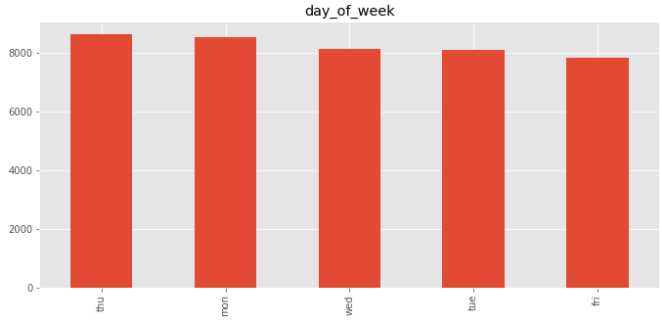
1) Illustration of categorical value frequency distribution







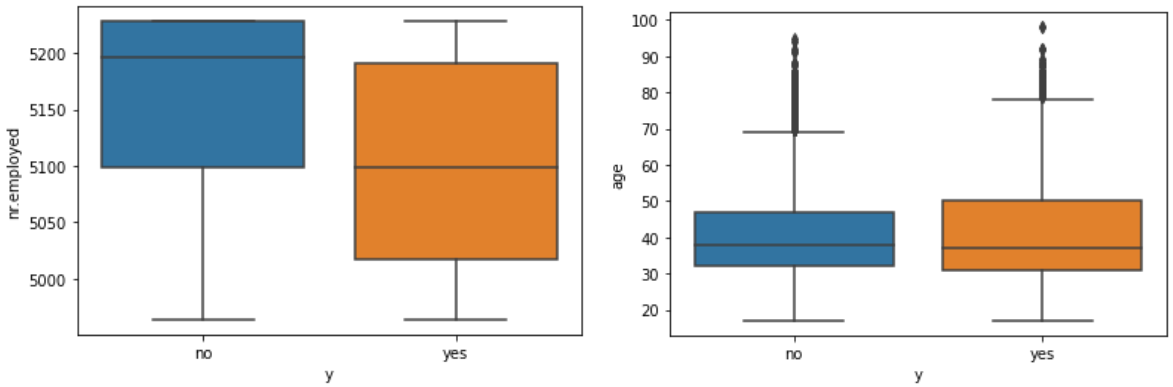




From analyzing and plotting data visually, we see that the majority of the customers are married. The top professions are blue-collar jobs, technicians, and administration - signaling a slightly

below average income. This speculation on income level, coupled with the stronger presence in housing loan application compared to personal loan application could explain why there is a strong imbalance in the result column as 'no', with people not subscribing to term deposits as they would need the limited cash towards other initiatives. Other interesting things to note are that a majority of customers were contacted in the month of May.

2) Illustration of distribution of categorical values in respective to output variable (campaign result)



We plotted the categorical variables like age and number of employees (indicating the size of the company) in box-plots according to the result of the campaign. 'Yes' means that the client subscribed to a term deposit, and 'no' indicates otherwise. The distribution for clients who subscribed to a term deposit is more diffused than not. We also analyzed the distribution of categorical variables in respect to whether the client subscribed to the term deposit. Interestingly, the outcome of subscription to term deposit is inverse with education level and work experience - those who have more years of education tend not to subscribe. Perhaps this is due to having additional knowledge of investments, market performances, and financial literacy that enables the people to make more strategic decisions with their money. People who subscribed to a term deposit also worked for smaller sized companies than people who did not subscribe, and this could be explained by the larger pension plans such as 401K larger sized companies offer to their employees. Employees with company offered-benefits such as 401K may invest their limited disposable income into those accounts compared to term deposits which may not offer a competitive rate and term deal.

Numerical Value EDA

This section will focus on univariate numerical analysis, that is, analyzing every variable that has continuous numeric values. We plotted these variables on histograms so that we can understand the distributions of the variables and make conclusions about the nature of the marketing data for each field, and also know what types of users we market to, and how each type of user behaved.

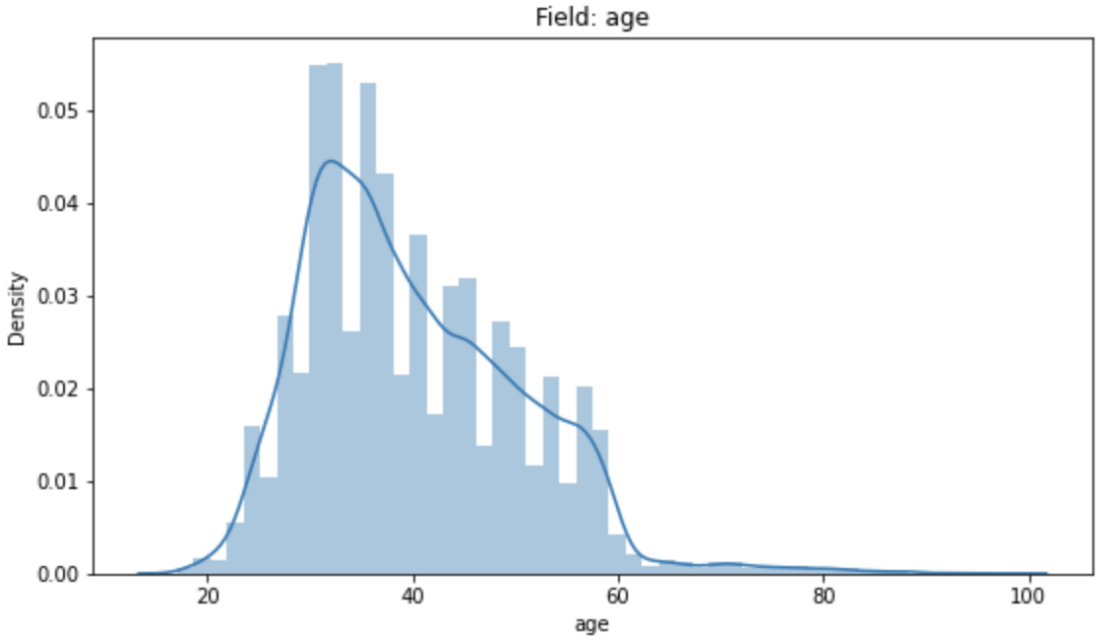
Selection and Plotting

We used numpy's number type to determine the continuous values, and then in a for loop, made a distribution plot for every variable that was numeric. Using `df.select_dtypes(include=np.number).columns` gives the list of columns with numeric values, and then in a for loop we run `sns.distplot(df[col])` to graph the distplot (histogram) of the values. See the below section for our analyses of these variables.

The distplot includes a best fit line for every bar graph, allowing us to see a rough estimation of the distribution. That's how we can more easily determine the skew of a graph.

Age

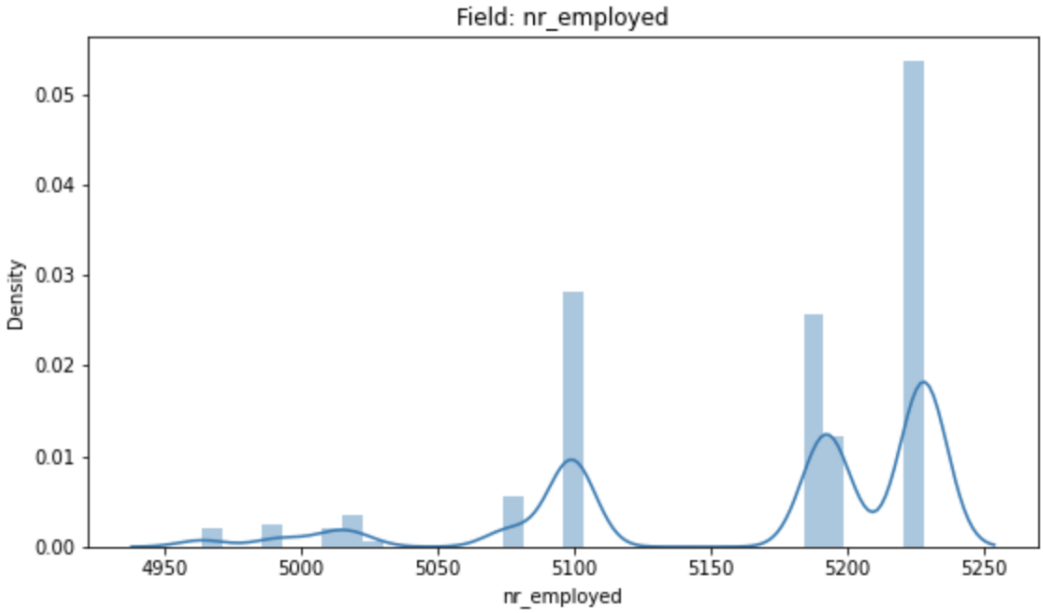
The clients who receive marketing campaigns largely lean younger. The median age is about 30, with some people also being over that, but the graph looks like this:



It’s not surprising that most of the customers of the bank who received our calls were on the younger side.

Number of employees

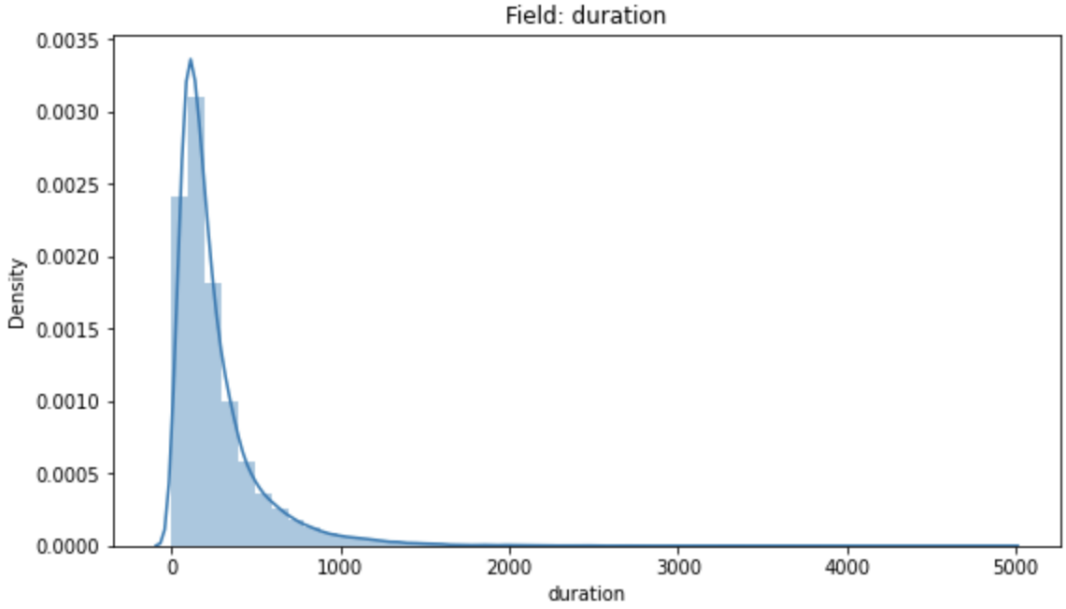
The number of people the bank employed is a less predictable distribution, since it is reported quarterly, and we don’t know of any upward or downward hiring trends. Also, it is not necessarily known that there is any impact on the number of employees and marketing success. Here we see the distribution of employee count per quarter:



This distribution is much more discrete, with an overall roughly upward trend, but not that noticeable. Even the guiding lines don’t point to a clear overall growth in size of the company.

Duration

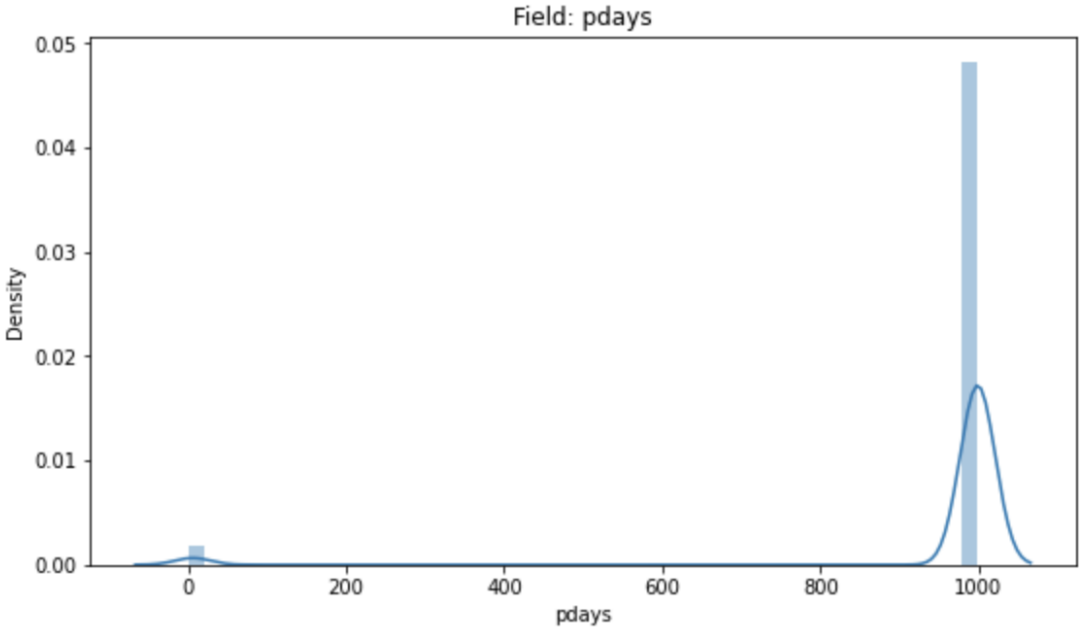
The calls also tend to lean shorter, leading to an expectedly right skewed distribution. It makes sense that most marketing calls wouldn’t take more than about 1-2 minutes (maybe 200 seconds at the most for the median), since callers either disconnect or are redirected elsewhere. Here is the distribution of durations:



This distribution leans to even smaller values than the others like age. The best fit line consistently grazes the top of every bar in the graph, meaning that the skewed curve fits the bars really well.

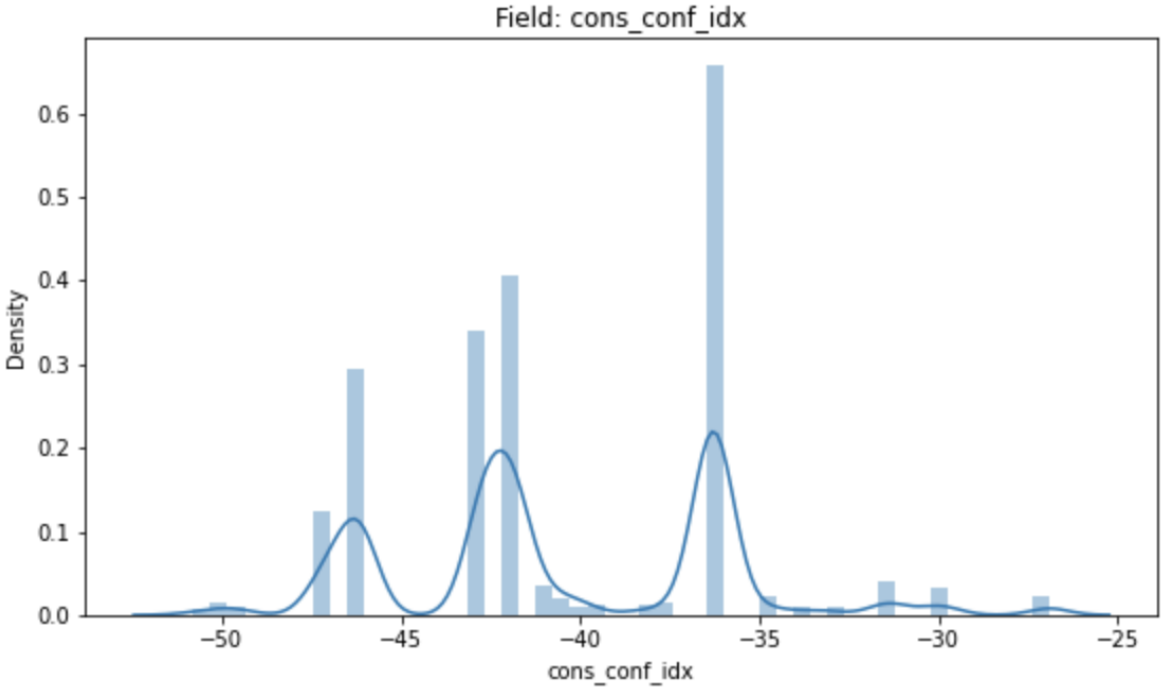
P-days

This field, as stated above, represents the number of days since the client was last contacted. A major plurality seems to be about 999, meaning that many of our clients were not previously contacted or it was the first call. For the remainder of clients (that we can see from the graph below), most of them only had about 1 day elapse between calls. This makes sense because we wouldn't really need more than a single call to get the message out, whether the client wants to stay in touch and consider anything or not. We can see the expectedly discrete distribution below, and are not surprised with how loosely the regression line fits the bar graph.



Other fields

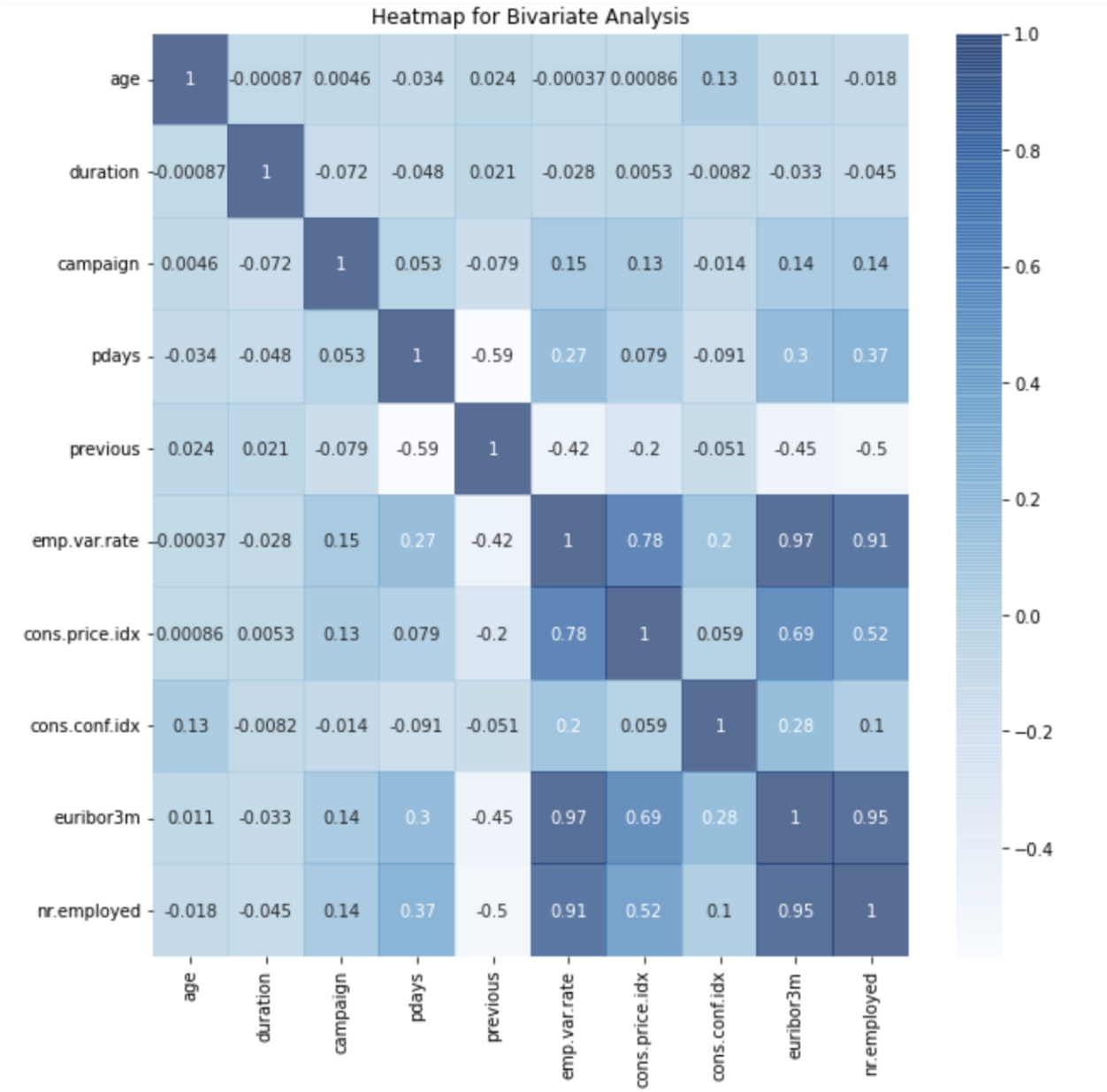
Most of the other fields have more continuous distributions like age, unlike p-days which are more discrete. One exception is the consumer confidence index, which is only available monthly and is more volatile than the other fields:



Bivariate Analysis

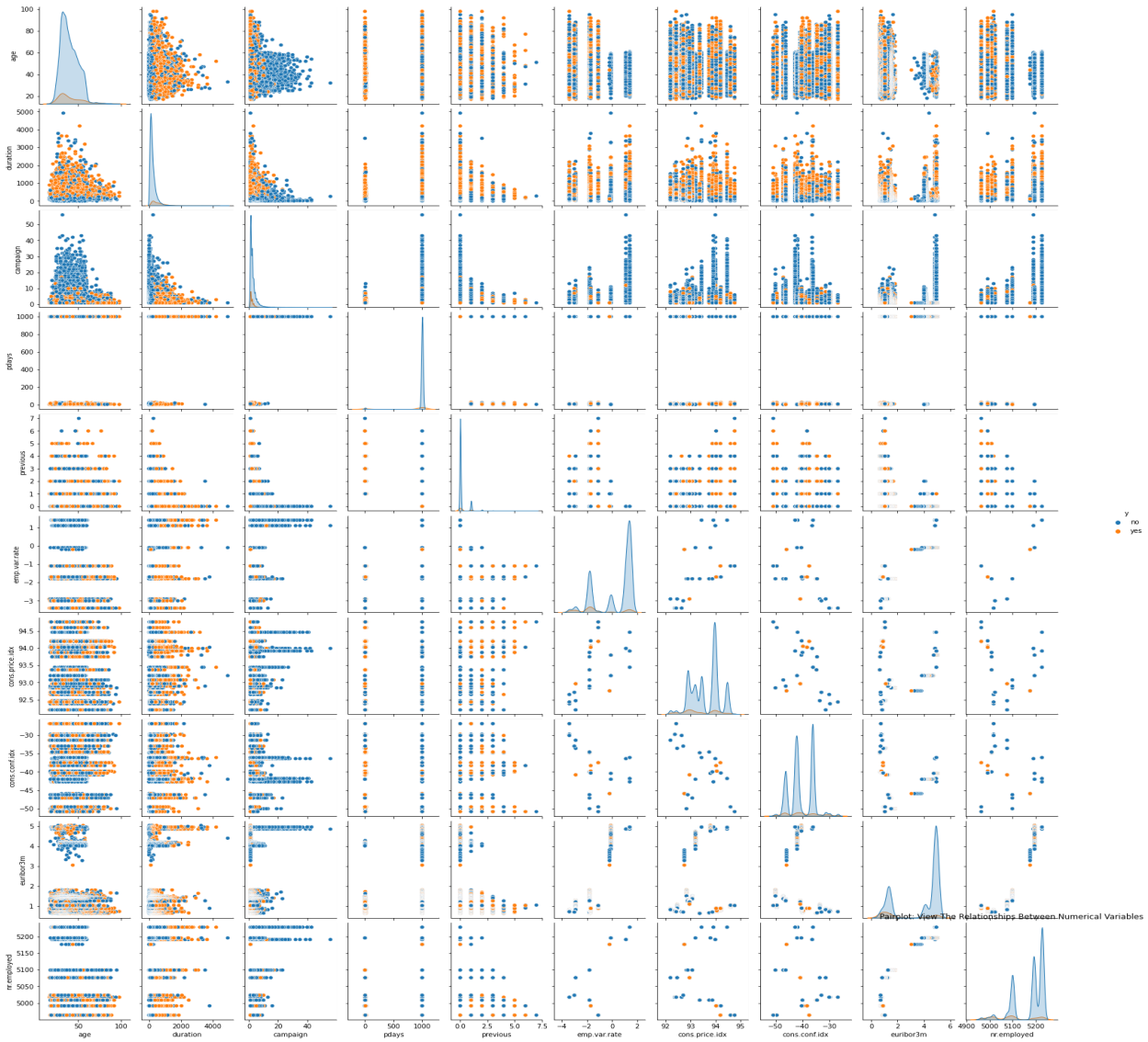
1. Bivariate Analysis of continuous variables

In this section we did a heat map to see if there is any correlation between numerical variables. Here darker the color, higher the positive correlation between variables. Lighter the color, the more negative the correlation is.



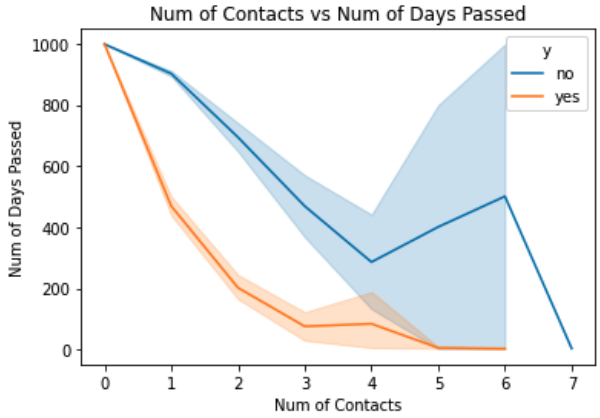
From this heatmap we can see there are some significant correlations between some variables. We can consider any correlation >=0.5 or <=0.5 are significant. However we will not include 'euribor3m' and 'cons.price.idx' in this analysis because those are mostly governed by the socio economic factors and we are not clear how the bank assigned the values to different customers. However these can be used in any future analysis with a machine learning approach. We will explore more on 'pdays' vs 'previous' , 'nr.employed' vs 'previous', ‘emp.var.rate' vs 'nr.employed'.

Before moving forward we also wanted to get a better understanding on how the relationships associated with each variable in a quick overview and did the below par plot with Seaborn. Hue is Orange for the successful and Blue is for unsuccessful customers from the previous campaign.



From this pairplot also we see that we can interpret a relationship only with the variables in the bottom right corner.
Below graphs show our relationship between 'nr.employed' (number of employees) vs 'previous'(number of contacts customers made before this campaign) . Data set might have included the number of employees in the customer's company to analyze the size of the organization and hence the stability of the employment. However this data set has very little standard deviation in the number of employees field. It is visible that the more stable the company that the customer works for , the fewer contacts they had with the bank. when the company size is smaller, they make more contacts , meaning this can indicate them attempting to terminating their deposits.

Shown below is the relationship between 'pdays' (number of days passed after previous campaign) vs 'previous' (number of contacts customers made before this campaign) . It is visible that there were few contacts made with that time passed from the previous campaign. However there is a higher confidence interval between 3-5 contacts and the number of days around 200. Which makes us think that customers may be withdrawing their deposits after 6 months. This might be a question about customer retention so as the term the deposits are available. or promote more of one year term deposits.

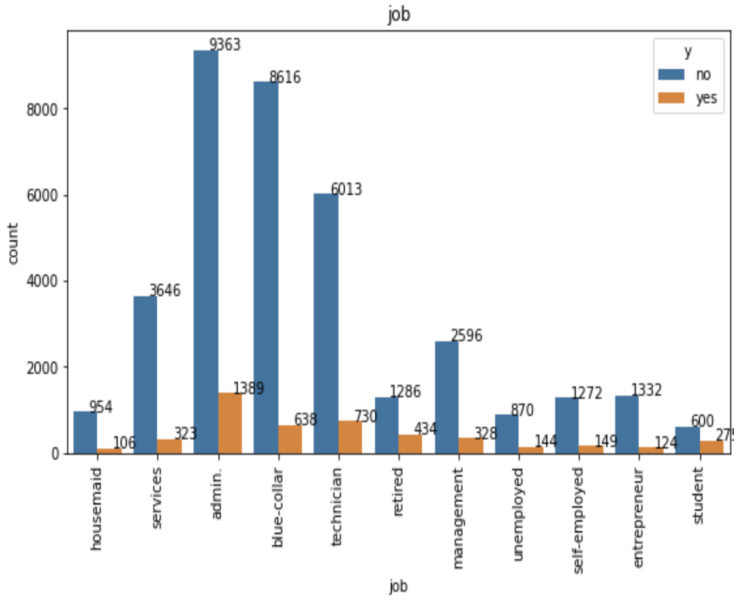


Below graph shows ‘emp.var.rate'(employment variation rate) vs 'nr.employed'(number of employees) and we can interpret that the bigger the company size, the less employee job changes. For successful campaigns we can suggest target customers with better employment stability.

2. Categorical Variables

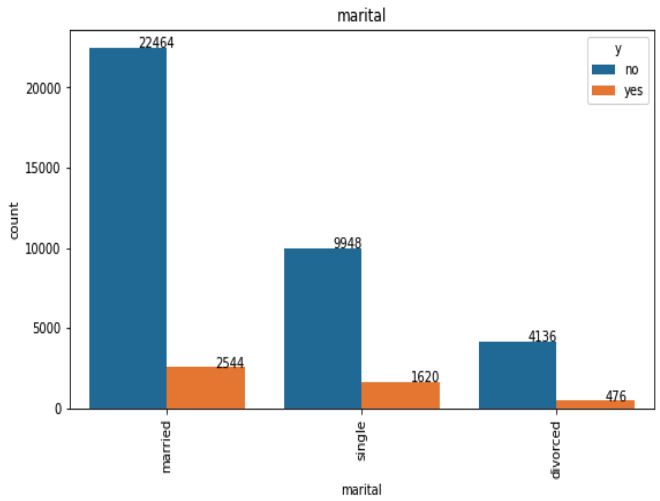
We analyzed the relationship of the outcome of the previous campaigns to categorical data in the dataset.

Job: Below plot shows higher numbers of acceptance are from admin, blue-collar and technician jobs. When we look at the ratio of acceptance[=Accepted customers/total Customers] students and the retired employees show the highest acceptance rate. Since this is a campaign for term deposits we can understand that people who do higher paid jobs may have selected other investment opportunities and who may hold short term infusions, are looking to deposit their money in these kinds of short term investments.



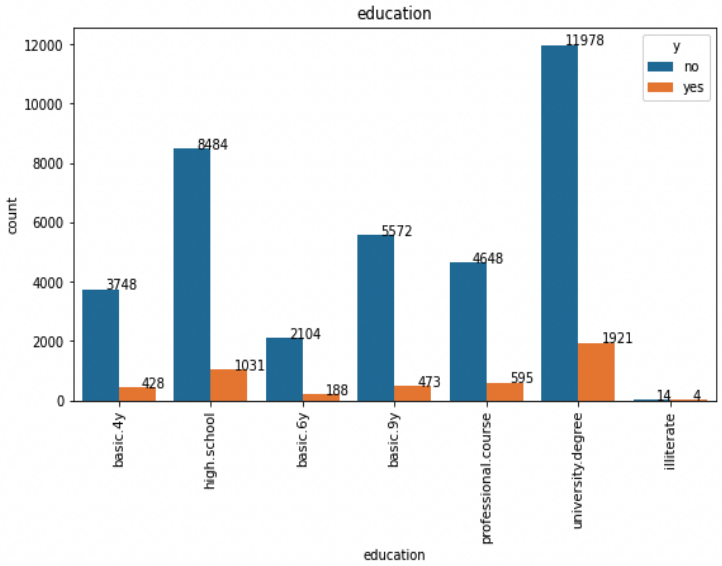
y	no	yes	accep_ratio
job			
student	600	275	0.31
retired	1286	434	0.25
unemployed	870	144	0.14
admin.	9363	1389	0.13
management	2596	328	0.11
technician	6013	730	0.11
housemaid	954	106	0.10
self-employed	1272	149	0.10
entrepreneur	1332	124	0.09
services	3646	323	0.08
blue-collar	8616	638	0.07

Marital Status: We can notice that the data set has a higher number of married customers and so the highest number of success in the previous campaign is also coming from married customers. However, the highest rate of acceptance is coming from single customers.



y	no	yes	accep_ratio
marital			
single	9948	1620	0.14
divorced	4136	476	0.10
married	22464	2544	0.10

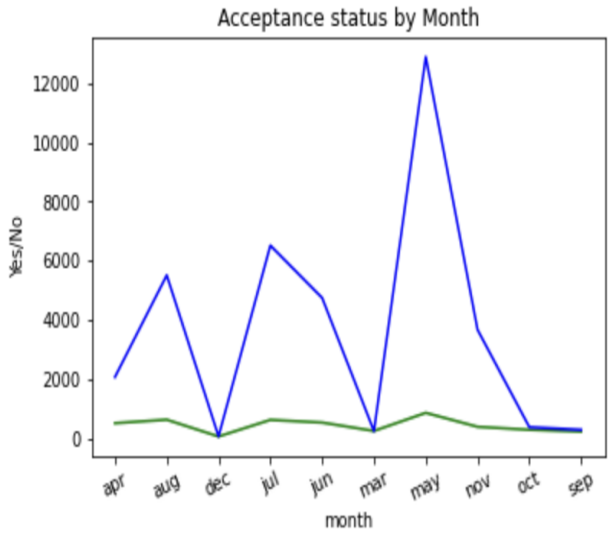
Education: This data set is skewed towards customers with a university degree. So we can observe the highest number of customers who accepted offers are having a university degree. This category of education comes as the category having the 2nd highest rate of acceptance.Who comes first in acceptance rate is customers who are illiterate. There have been 18 illiterate customers in the data set and four of them have accepted the offer to term deposits. This can be an isolated incident. We are not clear if they marked illiterate as if the customer is not illiterate in English language. In that case there can be situations where customers who run their own businesses which do not require literacy in English have money with them to invest in term deposits and are not knowledgeable enough for other types of investments .



y	no	yes	accep_ratio
education			
illiterate	14	4	0.22
university.degree	11978	1921	0.14
high.school	8484	1031	0.11
rofessional.course	4648	595	0.11
basic.4y	3748	428	0.10
basic.6y	2104	188	0.08
basic.9y	5572	473	0.08

Housing(mortgage), Loan: We can expect that there can be a direct negative relationship of accepting a term deposit if the customer is having a mortgage or a loan. However based on this data set we do not see a significant accet of customers having a mortgage or a loan to accept a term deposit.

Contact Month, Day: in below graph green line shows the accepted customers and blue line shows customers who did not accept the term deposits. Looks like in may back had a major campaign but majority of customers did not accept the term deposits. Acceptance rate is low as 0.06. Based on the acceptance rate, successful months are March, September, October and December. Except the month of October, all others are at the end of quarters and customers may have more money in hand due to end quarter earnings like bonuses, etc.



y	no	yes	accep_ratio
month			
mar	270	276	0.51
dec	93	89	0.49
sep	314	256	0.45
oct	403	315	0.44
apr	2093	539	0.20
aug	5523	655	0.11
jun	4759	559	0.11
nov	3685	416	0.10
jul	6525	649	0.09
may	12883	886	0.06

Conclusion & Next Steps

Based on our analysis we can advise the bank in selecting or promoting to customers as considering the below suggestion so that they can have a better success rate.

- Employment : More stability in employment considering less job changes and more medium size companies that are smaller in size. Approach students and retired customers more than in previous campaign
- Contact with the customer: On average spending time on contacting 3-4 times sufficient to change the decision positively
- Marital status : Include more single customers in the marketing campaign
- Education : Approach customers with University degrees and also recommend to find out more on customers who are illiterate and accepted the term deposits in last campaign
- Campaign months : Plan to initiate campaign in quarterly, end of the quarter may have more access rate.

These conclusions are based on the dataset that was provided and assume this is for specific zip codes. There can be differences that affect these decisions based on socio economic factors related to residency in different zip codes. So applying this to a bigger population will require further analysis of a larger data set.

Furthermore, our analysis suggests that people with higher income and higher work qualifications tend not to be subscribed to term deposits. As banks, it is important to capture a group of ultra wealthy individuals or young-professionals with an elongated career span. Financially, it is suggested that banks look into making the term deposit product more competitive, on par with larger company pension plans like 401K to attract additional investments from individuals. Strategically, offering financial advising services to individuals may be a bonus for customers to recognize the optimistic return on investment.

There are many possibilities to go further advanced with this research project. As we further progress into the course of the program, we could begin to leverage advanced statistical concepts and machine learning techniques to predict whether or not a user may subscribe to a term deposit based on their demographics, age, income, and other personal data.

Appendix

Link to Notebook

https://colab.research.google.com/drive/1Y0YyOEM6GeCLwLSYCXYXw_nBWT8D5QxI#scrollTo=TyyJsMbJKcVq

Link to Presentation

https://docs.google.com/presentation/d/14nIiSdUQ3HcO21S_hVYCHHfGk4ePYiLwZ0PshvuBo5I/edit?usp=drive_web&ouid=111331645813924038502

Data Structure

bank client data:
1 - age (numeric)
2 - job : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown")
3 - marital : marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)
4 - education (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown")
5 - default: has credit in default? (categorical: "no","yes","unknown")
6 - housing: has housing loan? (categorical: "no","yes","unknown")
7 - loan: has personal loan? (categorical: "no","yes","unknown")
related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: "cellular","telephone")
9 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
10 - day_of_week: last contact day of the week (categorical: "mon","tue","wed","thu","fri")
11 - 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. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")
social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: "yes","no")

References

- https://www.kaggle.com/code/sarthakniwate13/portuguese-bank-marketing-campaign-prediction/notebook
- https://github.com/onehungrybird/Bank-Marketing-DataSet-Analysis/blob/master/Bank_Marketing_DataSet_Analysis.ipynb
- https://notebooks.githubusercontent.com/view/ipynb?browser=unknown_browser&color_mode=auto&c[…]ory_id=337863188&repository_type=Repository&version=0
- https://github.com/jianwenwu/Bank-Marketing-Prediction/blob/master/Banking%20Marketing.ipynb
- https://www.kaggle.com/code/sarthakniwate13/portuguese-bank-marketing-campaign-prediction/notebook
- https://www.researchgate.net/publication/339988208_Data_Analysis_of_a_Portuguese_Marketing_Campaign_using_Bank_Marketing_data_Set