VIRTUAL INTERNSHIP

DATA SCIENCE

LISUM01

Data Understanding

Name: Samuel Alejandro Cueva Lozano
Email: scueval07@gmail.com
Country: Perú

27/07/2021

Sumario

Data understanding

Business Background

Client

- ABC bank: Portuguese banking institution

Problem Description

- ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to know whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Data Types and missing values

	Pandas types	Python types	Number of records	Number of missing values	% of missing values
age	int64	int	41188	0	0.0
job	object	str	41188	0	0.0
marital	object	str	41188	0	0.0
education	object	str	41188	0	0.0
default	object	str	41188	0	0.0
housing	object	str	41188	0	0.0
loan	object	str	41188	0	0.0
contact	object	str	41188	0	0.0
month	object	str	41188	0	0.0
day_of_week	object	str	41188	0	0.0
duration	int64	int	41188	0	0.0
campaign	int64	int	41188	0	0.0
pdays	int64	int	41188	0	0.0
previous	int64	int	41188	0	0.0
poutcome	object	str	41188	0	0.0
emp.var.rate	float64	float	41188	0	0.0
cons.price.idx	float64	float	41188	0	0.0
cons.conf.idx	float64	float	41188	0	0.0
euribor3m	float64	float	41188	0	0.0
nr.employed	float64	float	41188	0	0.0
y	object	str	41188	0	0.0

Proposed Approach

Duplicate values

To consider two o more records as duplicates these must have the same values in the most of the attributes (excluding unique identifier) or very similar values in the case that they are strings (e.g. ‘Samuel Cueva Lozano’ = ‘Samy Cueva Loz’ and the others attributes be the same).
A record in the data set will be considered duplicate only if there are records with the same value in each attribute.

Null values

The *Empty values, NaN, N/A values* and values like *“?”* will be treated as ***Null values***

Fields with ***too many null values*** would be removed or filled according to the following:

- Fields like *Age, Duration and Emp.var.rate* would be filled because they have missing values that depends on their hypothetical values or depend on other attributes.
- The rest of fields would be removed because they have missing values due to bad configuration, issues with data collection, or untraceable random reasons.

After cleaning at the field level, rows with null values would be removed.

Outlier detection

Outliers will be removed from numerical fields so that they don’t negatively affect the analysis.
A fixed threshold would be used for the *Age, Duration, Campaign, Pdays* and *Previous* attribute to avoid inconsistent data, then the IQR Score will be used to filter out the outliers in all attributes.

Data Transformation

The fields *Education, Month and Day_of_the_week* could be transformed to a more readable format for better understanding and of course the others categorical attributes will be transformed to feed the model later.

Problems in the data

Duplicate values

Using Pandas, the duplicate rows are as follows:

	age	job	marital	education	default	housing
1266	39	blue-collar	married	basic.6y	no	no
12261	36	retired	married	unknown	no	no
14234	27	technician	single	professional.course	no	no
16956	47	technician	divorced	high.school	no	yes
18465	32	technician	single	professional.course	no	yes
20216	55	services	married	high.school	unknown	no
20534	41	technician	married	professional.course	no	yes
25217	39	admin.	married	university.degree	no	no
28477	24	services	single	high.school	no	yes
32516	35	admin.	married	university.degree	no	yes
36951	45	admin.	married	university.degree	no	no
38281	71	retired	single	university.degree	no	no

Note: only some attributes are displayed due to the table size.

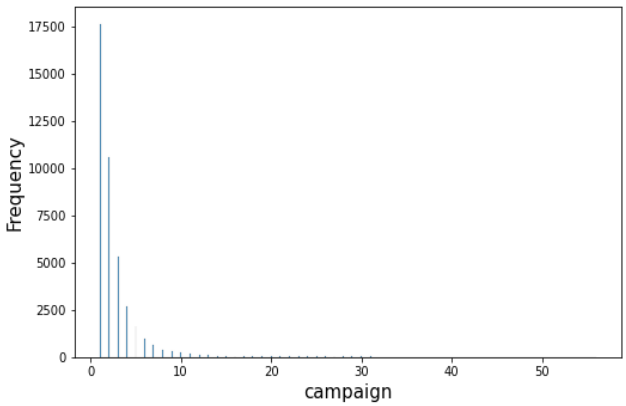
This rows will be removed

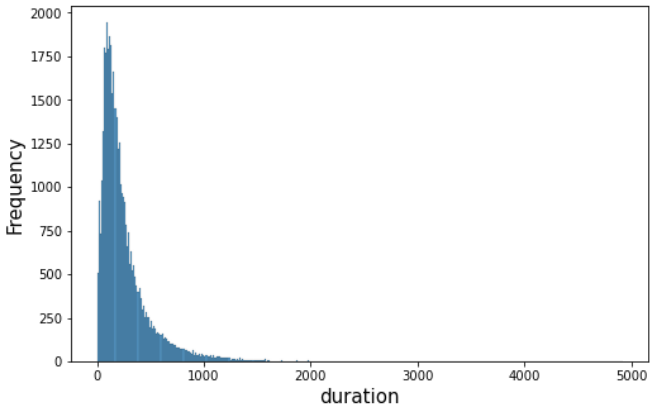
Outliers

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

Potential outliers in *Duration* and *Campaign*

Histograms for Campaign and Duration





A deeper analysis will be made in EDA