Exploratory Data Analysis

Bank Marketing Campaign

Jan 15, 2023

Group Name: Evolve Data
Name: Dmitry Sharukhin
Email: sharuhinda@gmail.com
Country: Russia
College/Company: Finval GC
Specialization: Data Science

Agenda

Problem Description

Overall Dataset Characteristics

Approach

EDA

Conclusions and Recommendations

Technical Details

Problem Description

Client:
ABC Bank

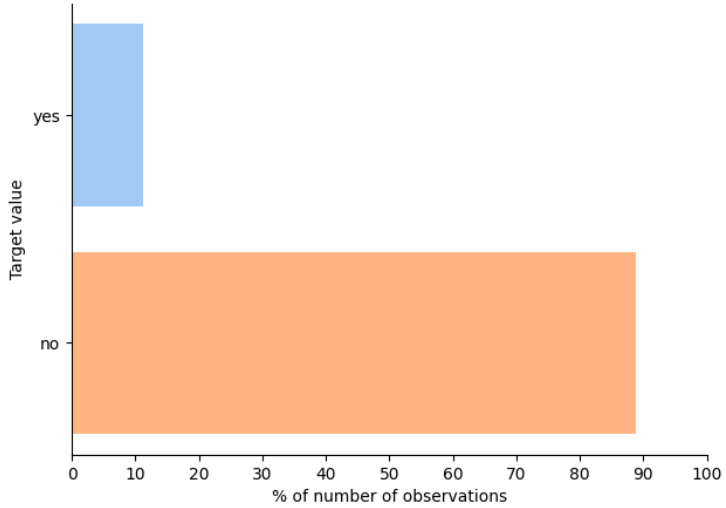
Description:
ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution)

Bank wants to use ML model to shortlist customer whose chances of buying the product are more so that their marketing channel (tele marketing, SMS/email marketing etc.) can focus only to those customers in order to save resources (which is directly involved in the cost (resource billing)) and their time

Objectives:

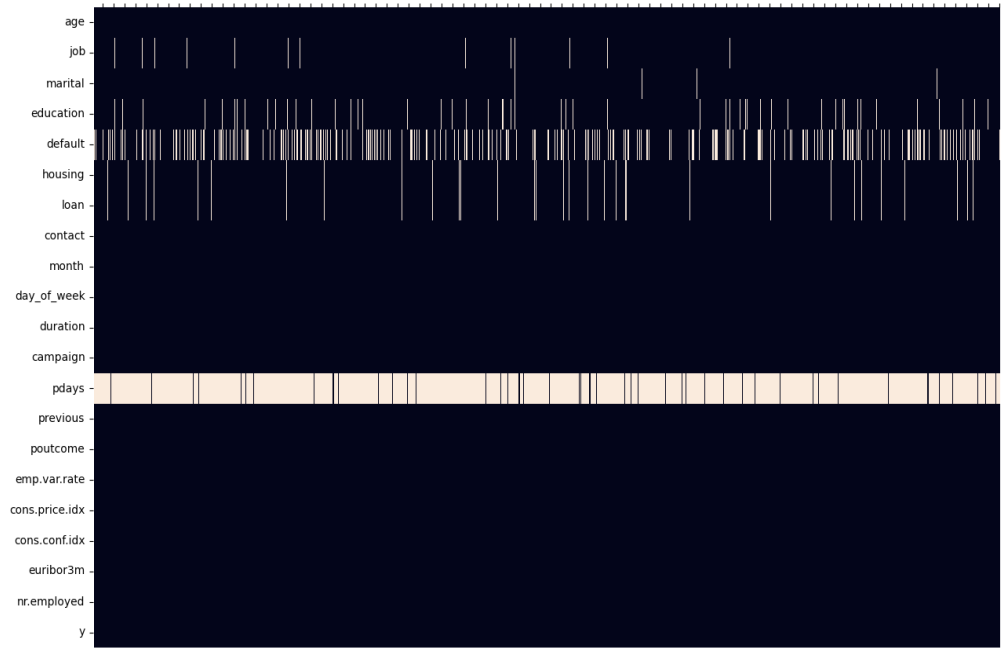
- Create ML model to shortlist customer whose chances of buying the product are more so that their marketing channel (tele marketing, SMS/email marketing etc.) can focus only to those customers

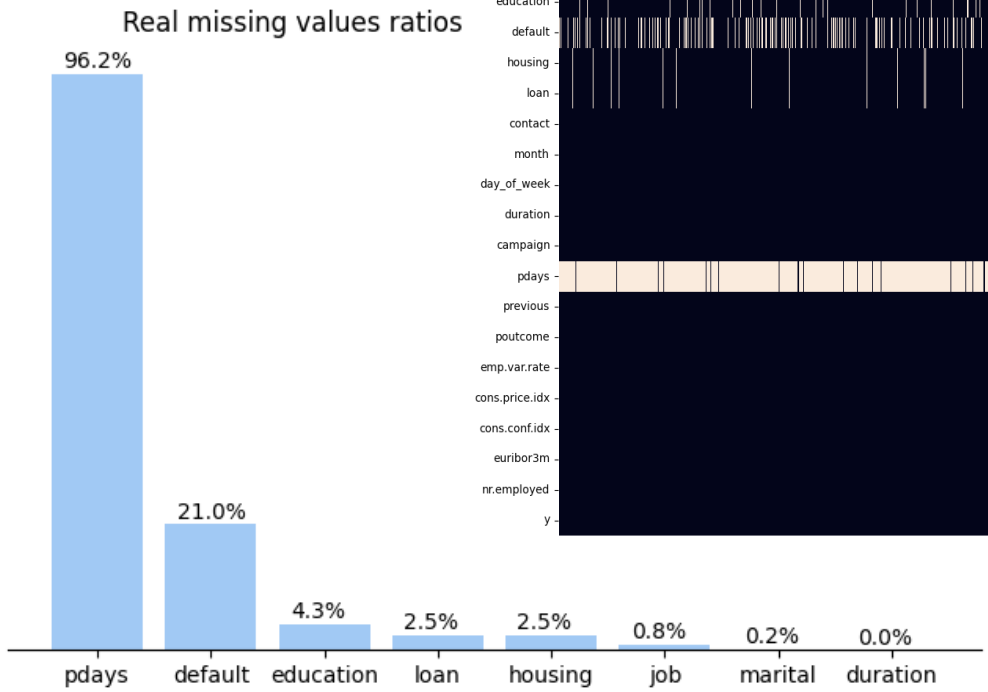
Overall Dataset Characteristics



The dataset is **HIGHLY imbalanced**
(positive labels << negative labels)

Some features contain **many missing-like values**
(these are 'unknown' and 999)





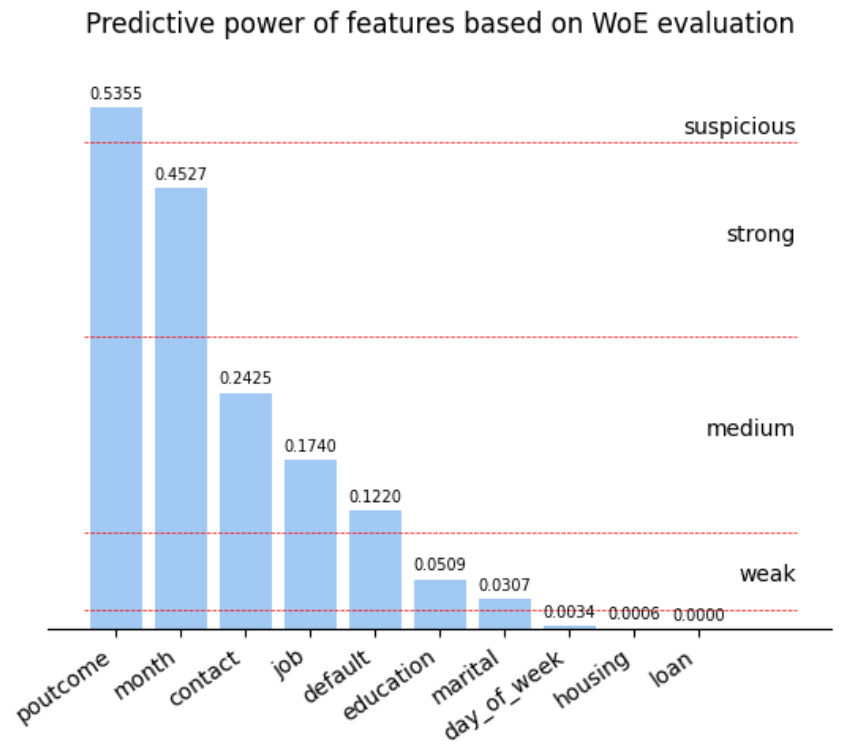
Data Glacier
Your Deep Learning Partner

Approach

For analysis purposes we divided all features on personal, marketing and macroeconomic categories

Because of low frequency of positive observations in the dataset we **must be very careful about dropping** rows.
Ideally not dropping at all

We will use the approach known as **Weight-of-Evidence** to keep all observations safe.
This approach is based on **relative odds of events** (odds to have ‘yes’ vs. odds to have ’no’) and can be used even for missing data considering it as a separate category

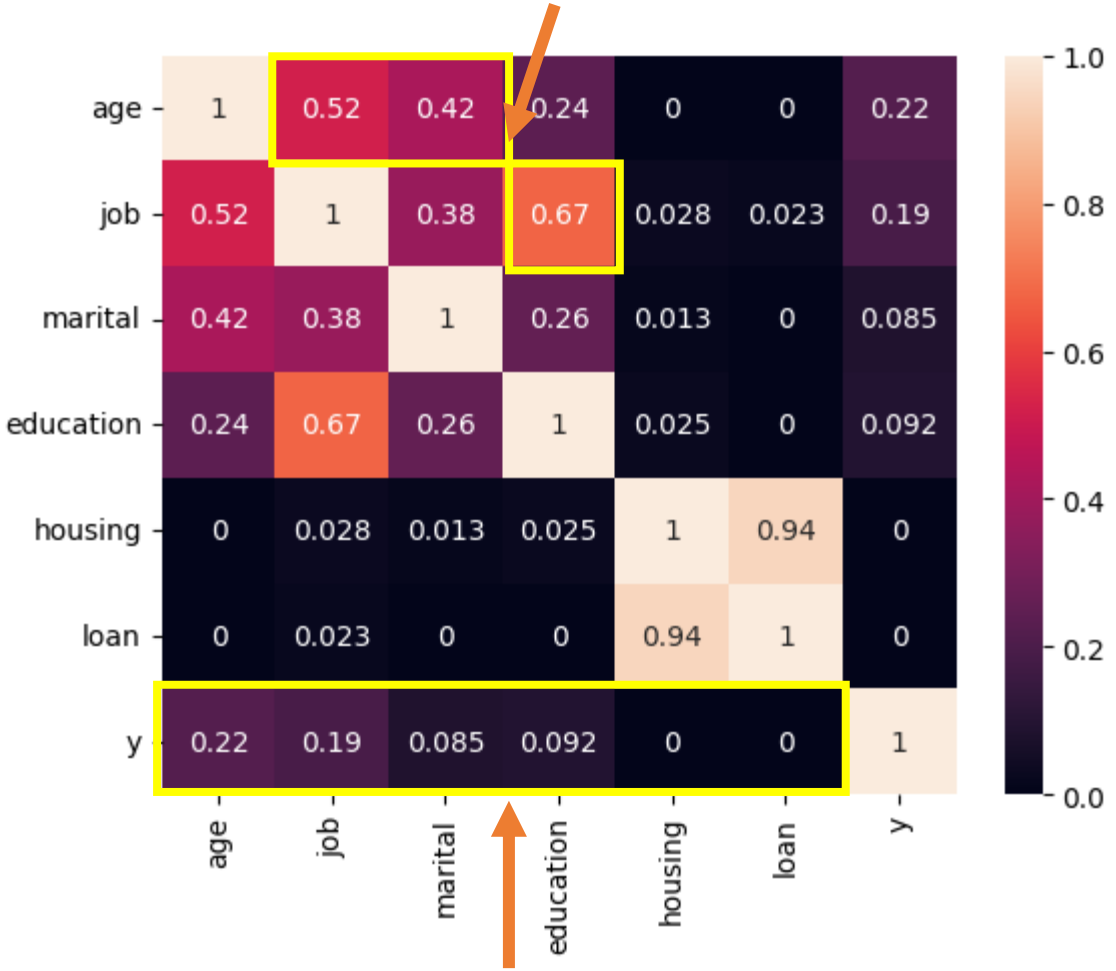


This approach also allows to evaluate the “usefulness” or predictive power of features

Data Glacier
Your Deep Learning Partner

EDA: Personal Features

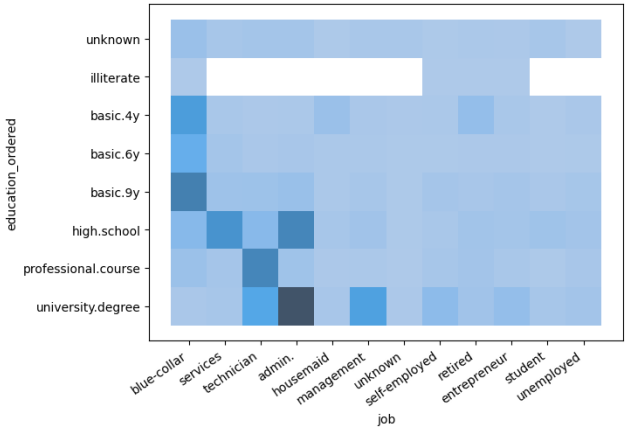
There are several features having relatively strong correlation (association strength) with each other.
It may help with imputing missing values



There are no features having high correlation (association strength) with target

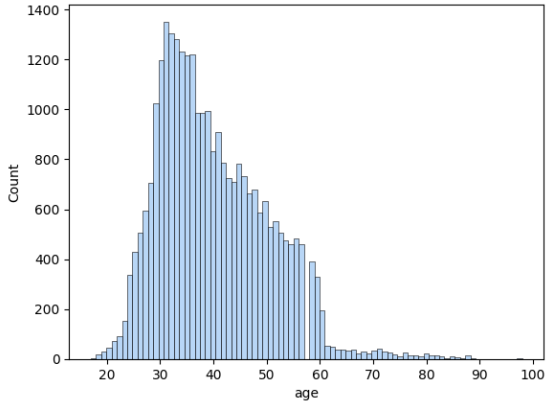
Data Glacier
Your Deep Learning Partner

EDA: Personal Features (continue)

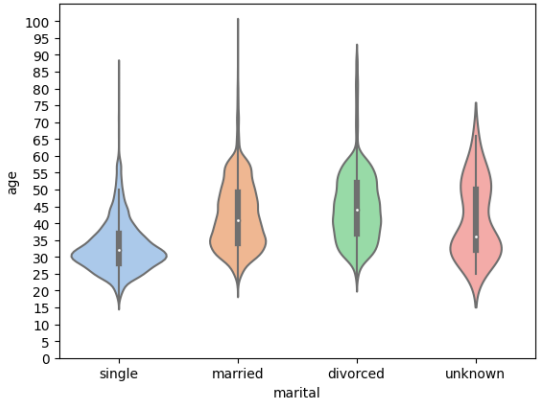


There are obvious modal values for jobs between education levels and vice versa we could use to impute missing values in these columns.

We are going to use ‘most frequent’ strategy to impute missing values for each job vs education correspondence



Frequency distribution of age is non-normal with significant drop after 60.
As savings are tend to be more attractive to elder people, we cannot drop that values. We will try to bin those values to gain additional predictive strength for model instead

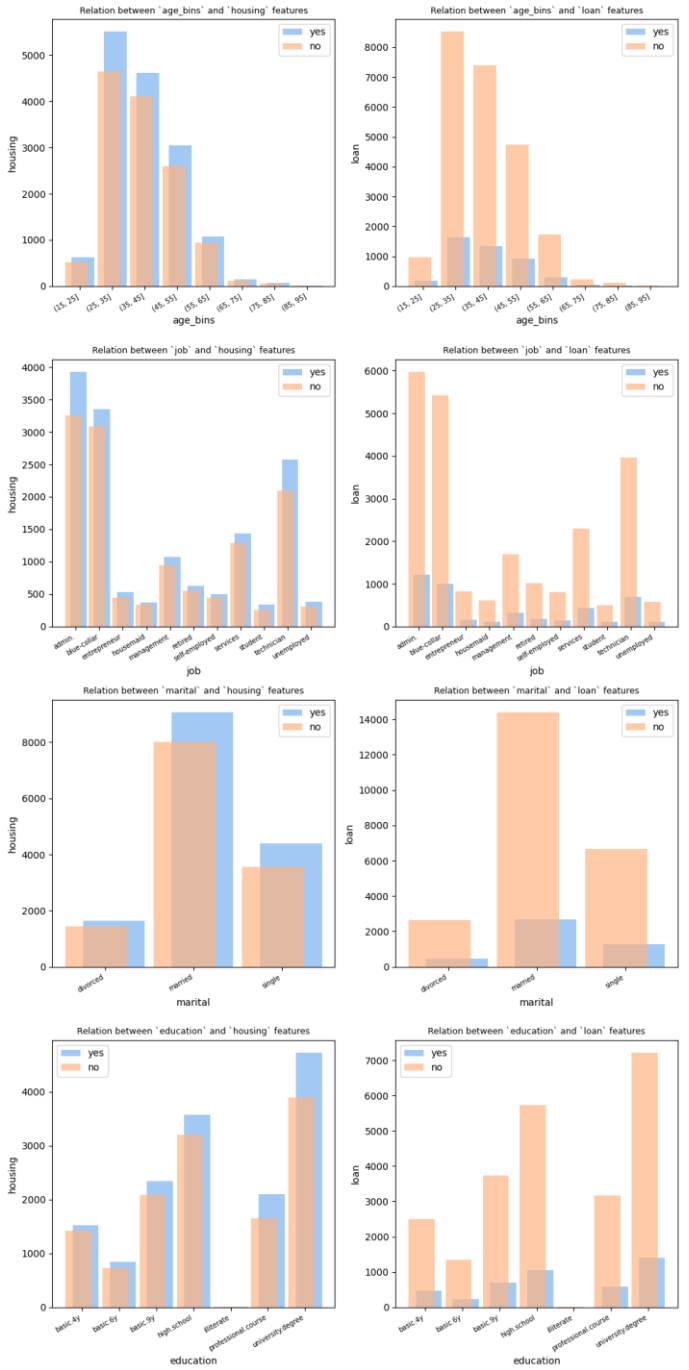


There are no clear-defined differences between groups in marital status.

We are going to use ‘most frequent’ strategy to replace ‘unknown’ values

Data Glacier
Your Deep Learning Partner

EDA: Personal Features (continue)



Most frequent values for personal and housing loans stay the same across other personal features

Using general ‘most frequent’ strategy to replace ‘unknown’ values is safe for these columns

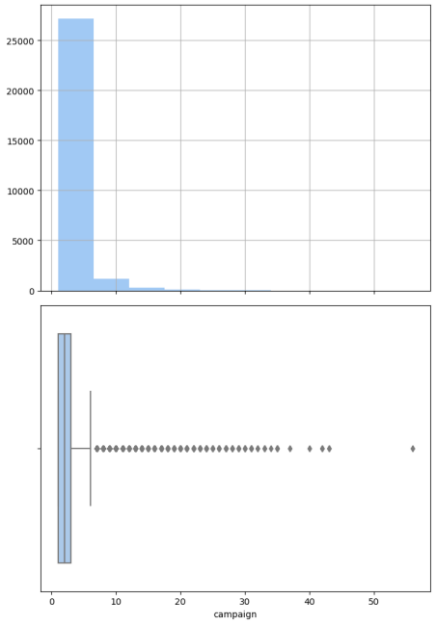
Data Glacier
Your Deep Learning Partner

EDA: Macroeconomic Features

There are several macroeconomic indicators with high correlation with each other. This can create undesirable consequences for some types of models

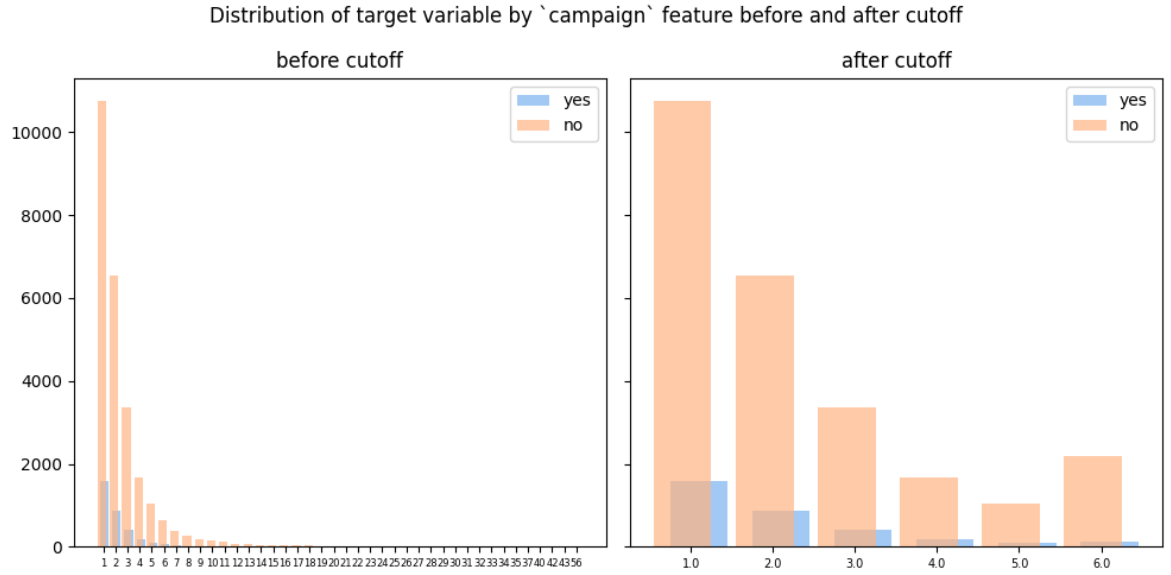
Data Glacier
Your Deep Learning Partner

EDA: Marketing Features



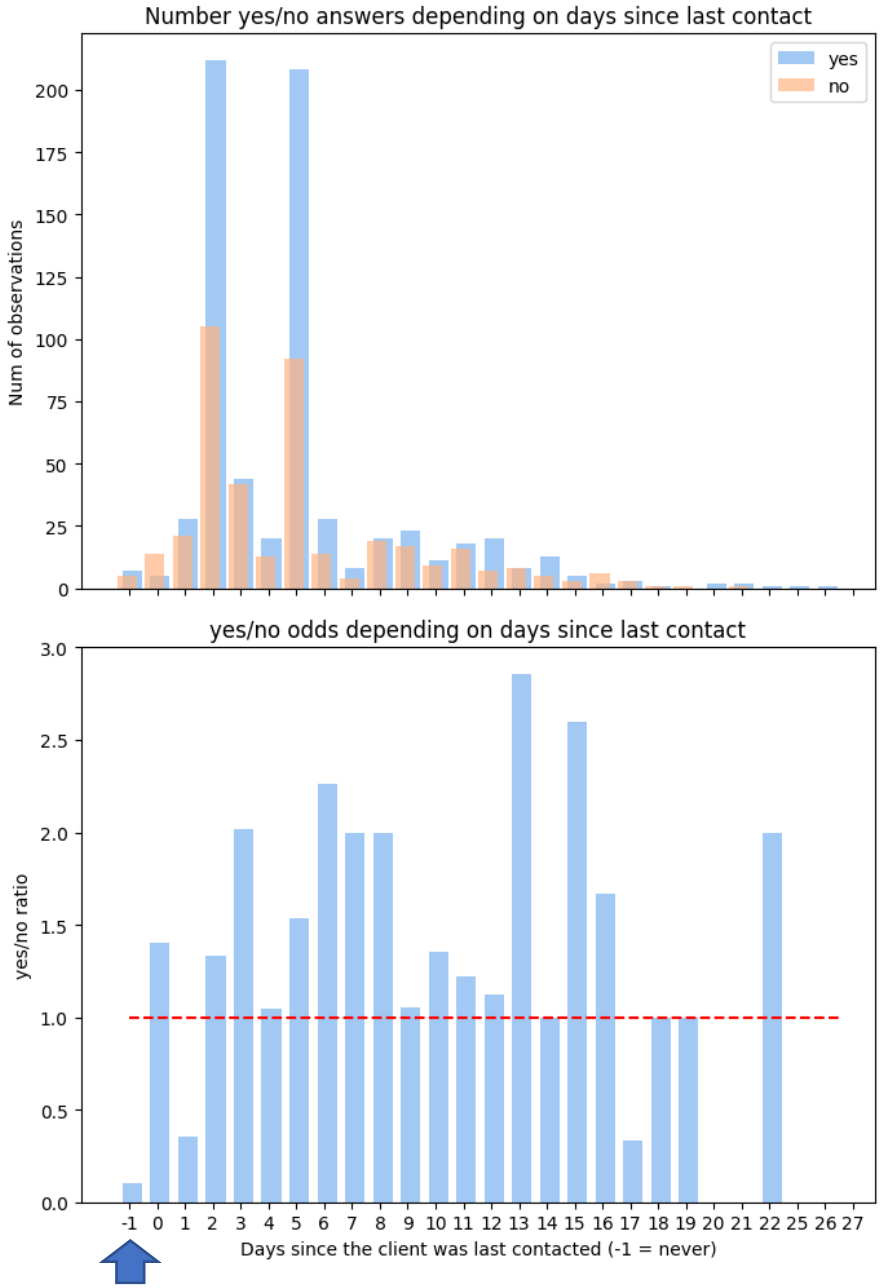
'**Campaign**' feature is hugely skewed (long right tail). We are going to cut-off outliers. We consider as outliers all values greater than Q3+1.5*IQR

It allows not to drop rows with potentially valuable info and creates a more reasonable set of values to apply Weight-of-Evidence evaluations



Data Glacier
Your Deep Learning Partner

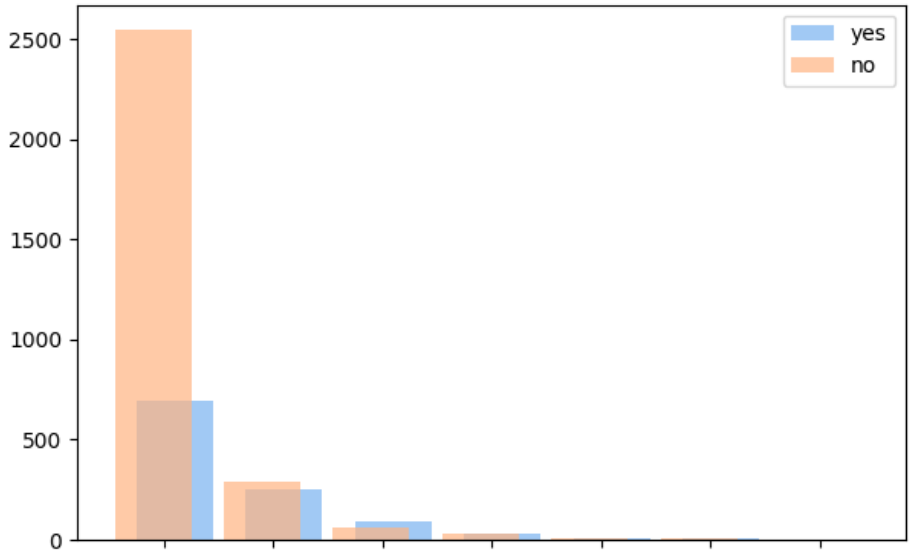
EDA: Marketing Features (continue)

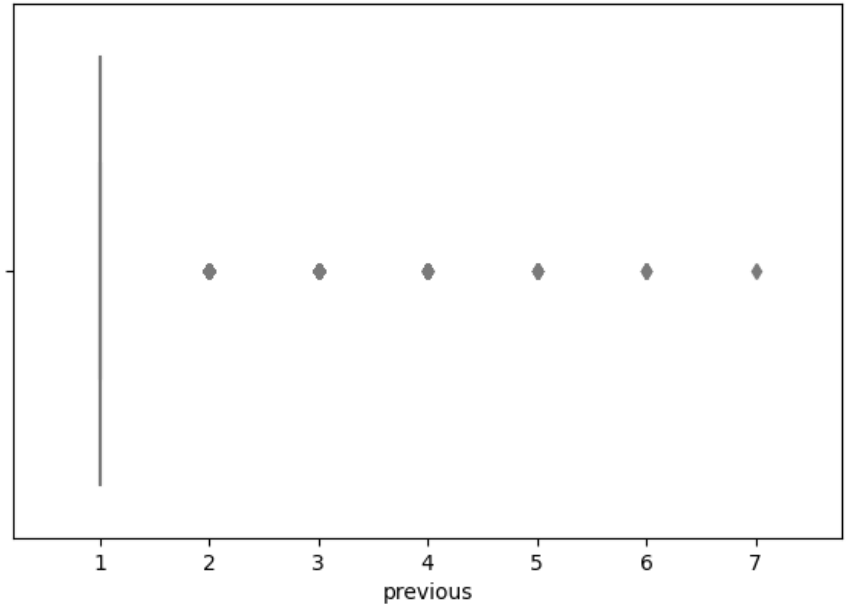


Overall odds to get positive result are much higher if the client was contacted before

Data Glacier
Your Deep Learning Partner

EDA: Marketing Features (continue)



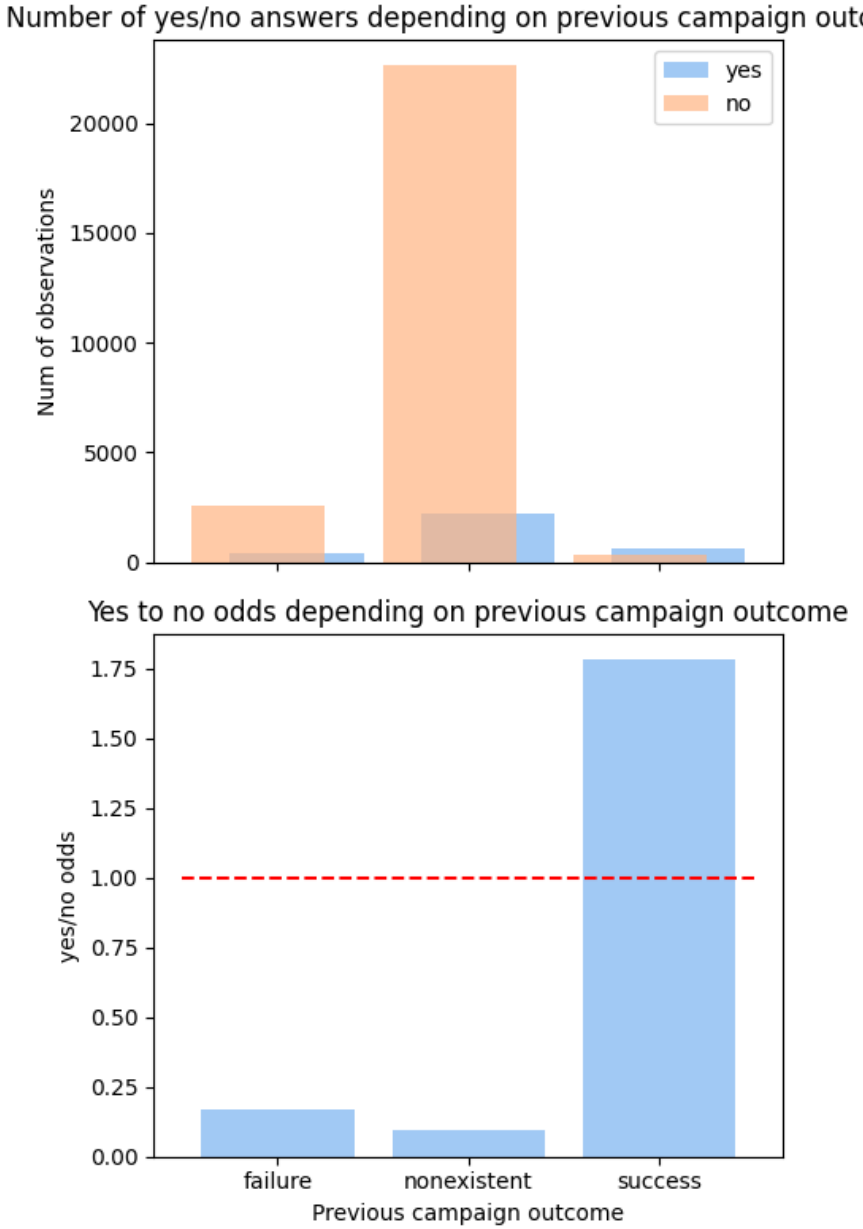


In order to use WOE approach here we have to provide each category contains >5% of sample size

	previous	y
	count	sum
previous		
0	24838	2181
1	3234	691
2	534	246
3	151	89
4	56	30
5	12	8
6	5	3
7	1	0

Considering 3 as a cutoff value will provide necessary category size

EDA: Marketing Features (continue)



Overall odds to get positive result are much higher if the previous campaign succeeded

Conclusions and Recommendations

Data Glacier
Your Deep Learning Partner

Conclusions:

- Based on the Weight of Evidence (WoE) approach the predictive power of features was evaluated.
- All features were left except ‘duration’
- Dataset is highly imbalanced so missing-like data was retained to encode it later

Final recommendations:

Perform features encoding using 3 main strategies (and their mixed options with dropping any of high correlated and/or low variance features) and compare their impact on model’s metrics:

- use WoE approach to encode all category and category-like features without imputation
- use WoE partially only on truly categorical features without imputation, convert numeric features to reasonable values and cutoff outlying values
- don’t use WoE on any feature, impute values and encode categorical features by ordinary way

Technical Details

Data Glacier
Your Deep Learning Partner

Problem type:
Binary classification
Requirement for model interpretability

Appropriate model types:
- Logistic Regression (simple linear model with direct evaluation of probability)
- Bayes Naïve Classifier (conditional probability-based type)
- Decision Tree (highly interpretable)
- Random Forest or other tree-based models like LightGBM and CatBoost (interpret using SHAP package)
- K Nearest Neighbors (explained by distance)

Notes:
- Highly imbalanced dataset will require classes weighting or using oversampling technique
- 3 highly correlated features found. They are 'euribor3m', 'nr.employed' and 'emp.var.rate' (Pearson's r > 0.9). This leads to potential multicollinearity problem for some ML model types

THANK YOU