EDA

Data Science Project - “ABC Bank Marketing”

Sebastián Castro
Omar Jazouli
Freddy Tapia
Ignacio Solórzano

14-May-2021

- Quick variables review
- Outliers detection and treatment
- Missing values imputation
- Treatment of categorical and numerical variables
- Recommended models

Quick variables review

Clients Data

1. age: int
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')

Data related with the last contact

8. contact: contact communication type (categorical: 'cellular','telephone')
9. day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
10. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
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.

Quick variables review

Data related with the last campaign

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')

Data related with social and economic context¶

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)

Result of the current campaign

21. y - has the client subscribed a term deposit? (binary: 'yes','no')

Outliers detection and treatment

The methods selected to detect the outliers values were,

- **IQR**: is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts.

- **Z-score**: is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviation from the mean.

The methods selected to treat the outliers values were,

- **Elimination:**: with this approach we eliminate the observation which are outliers. We use the IQR and Z-score approach to detect those values

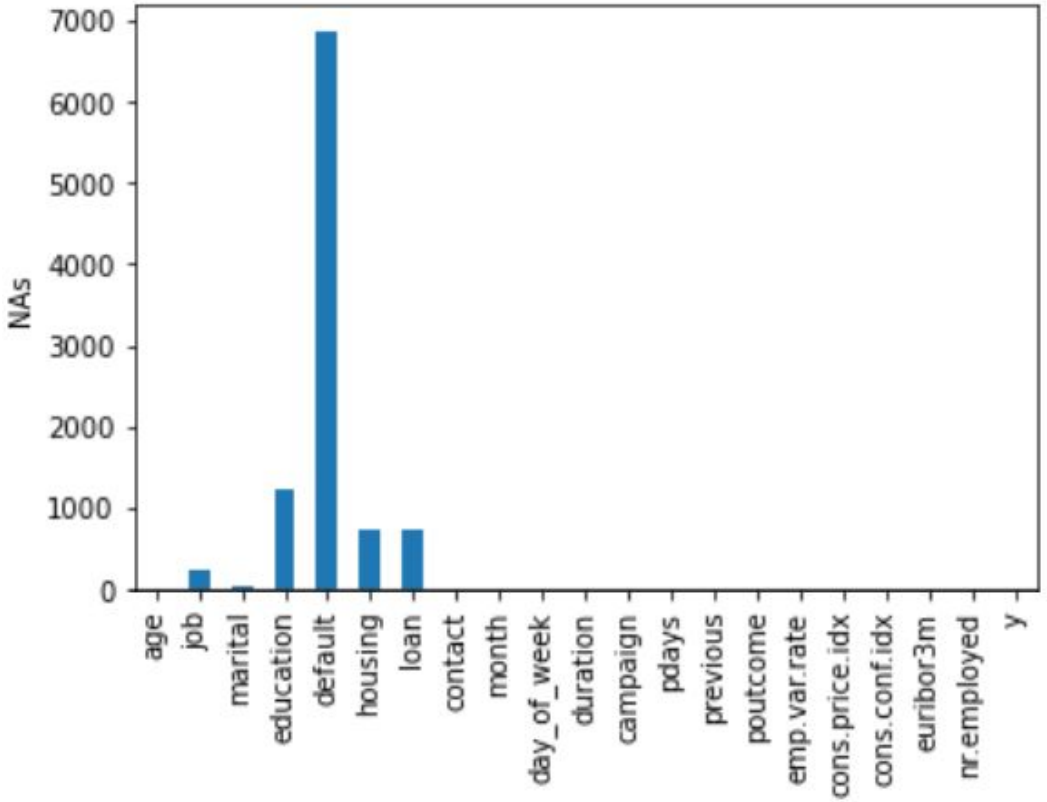
- **WoE**: is a process used as a benchmark to screen variables in the credit risk modeling projects such as probability of default. They help to explore data and screen variables. It is also used in marketing analytics project such as customer attrition model, campaign response model etc.

Missing values imputation

The method used to treat the NA values is,

- **Elimination:** we eliminate the observations with NA values. In the case of the variable “default” we eliminated it due to the high number of NA values.

- **Mean**: using the library Pycaret in Python, the default method to treat a NA value is replace it by the mean.



Treatment of categorical and numerical variables

Categorical variables

- **Bivariate analysis:** with this process we study the correlation that occur with each categorical variable and the target variable

Numeric Variables

- **Normalization:** the goal of normalization is to rescale the values of numeric columns in the dataset without distorting differences in the ranges of values or losing information. This process is provided by pycaret library.

Recommended models

We obtained two different data which were calculated with different methods,

Data 1 :

This data was obtained by eliminating the outliers detected by the IQR method. For this case the top three models obtained by the pycaret library are,

- Light Gradient Boosting Machine
- CatBoost Classifier
- Extreme Gradient Boosting

Data 2 :

This data was obtained by treating the outliers with the Woe method. For this case the top three models obtained by the pycaret library are,

- Light Gradient Boosting Machine
- Gradient Boosting Classifier
- Logistic Regression

Thank You