Marketing Campaigns Classification

Guanhua Zhu

Brown University

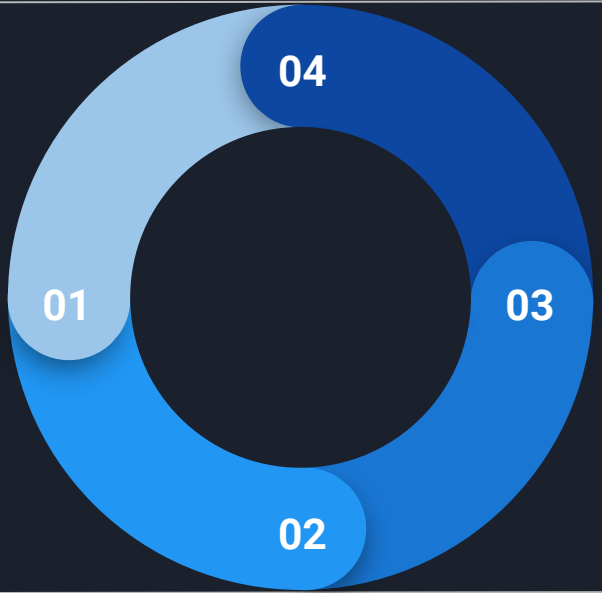
October 21, 2019

https://github.com/gzhu7/bank-marketing-campains-classification

Overview

Intro

Data&Preprocessing



Q&A

EDA

Overview

-To predict: whether the clients of a Portuguese bank will subscribe a term deposit at the bank

-The problem is a classification problem. The goal is to use classification methods to predict if a client will subscribe a term deposit (target variable, binary).

In business world, a portion of clients usually contribute much more than others to the total sales of deposits or financial products. One major task for banks is to find such target clients. The patterns shown on such group can be used to find other potential loyal clients with similar background or sharing same interests. Therefore, the prediction can be applied on various clients groups and it helps to seek for people who have higher probability to subscribe a term deposit.

-The dataset is initially posted on the paper *A Data Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems (2014)* by Sergio Moro, Paulo Cortez and P.Rita.

Data&Preprocessing

The data is related with direct marketing campaigns of a banking institution. THe marketing campaigns were based on phone calls. The dataset includes 41188 rows of data and 21 columns. There are 10 numeric feature variables and 10 categorical feature variables in the original dataset.

It includes the bank client data, variables related with the campaigns of the Portuguese banking institution, and the social-economic context context attributes.

01 Client data: age, job, marital status, education, credit in default, housing loan, personal loan

02 Variables related with marketing campaign: contact communication type, last contact month, last contact day of the week, last contact duration, number of cantacts performed, number of days passed by after the client was last contacted, number of contacts performed before campaign, outcome of previous campaign

03 Social-economic context attributes: employment variation rate(quarterly), consumer price index(monthly), consumer confidence index(monthly), the Euro Interbank Offered Rate(daily), and total number of employees(quarterly)

Data&Preprocessing

Missing data

no na value.

For the variable *pdays--number of days that passed by after the client was last contacted from a previous campaign*, if the client was not previously contacted, pday = 999. So a binary variable indicating whether a client was previously contacted (*pdayscon)* was used instead.

OneHotEncoder

Apply OneHotEncoder on: job, marital, education(since there is unknown group), default, housing, loan, contact, poutcome, month, day_of_week, pdayscon

Data&Preprocessing

Scaler

Apply StandardScaler on: duration, campaign, previous, age, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed

(They are all numeric and follow a tailed distribution.)

-Label

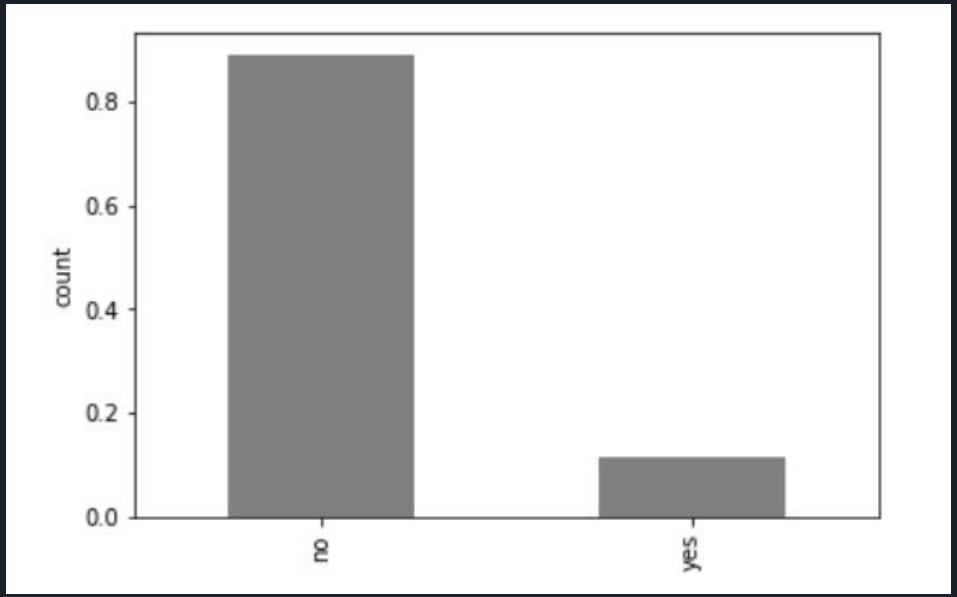
Apply LabelEncoder on the target variable: y (dependent variable)(41188)

-Preprocessed feature datasets:X(41,188*64)

EDA

Dependent Variable: y, has the client subscribed the term deposit?

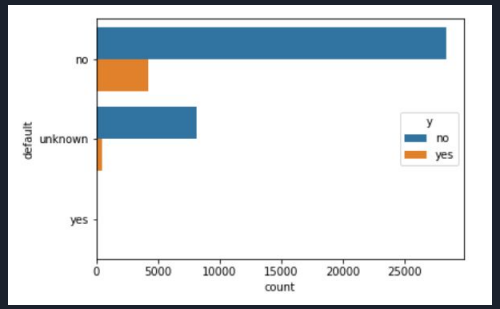
-Imbalanced

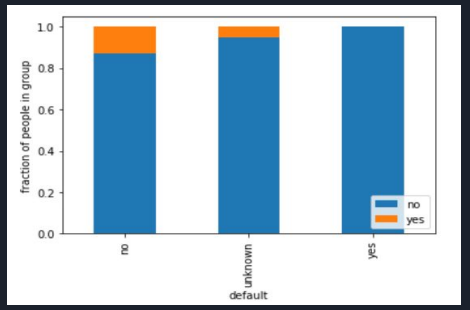


EDA

Has credit in default?(default,cat)

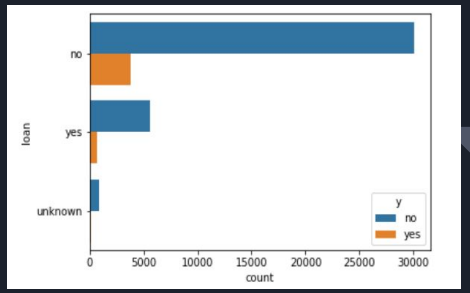
- Not likely to be chosen to the campaigns if value





Has personal loan?(loan,cat)

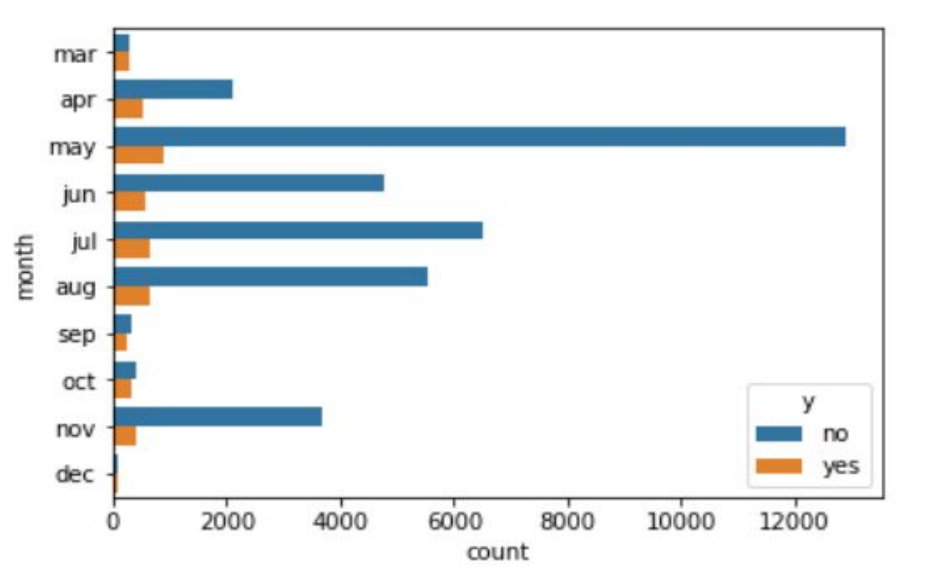
- Same case



EDA

Last contact month of year?(month,cat)

-Clearly seasonality, more in spring, less in winter



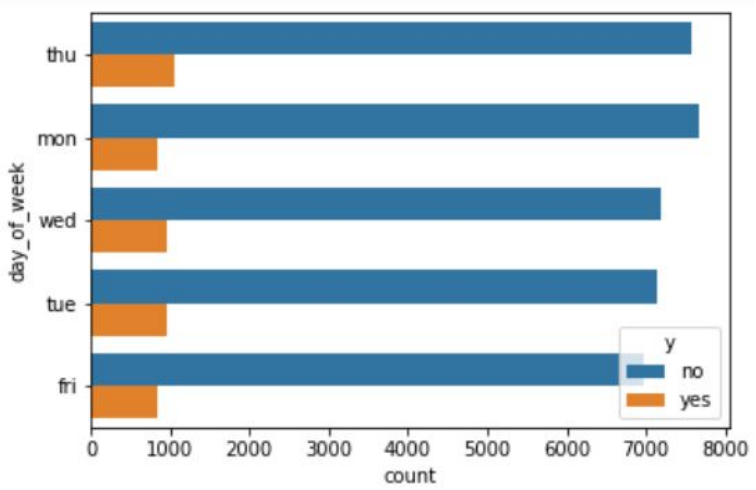
EDA

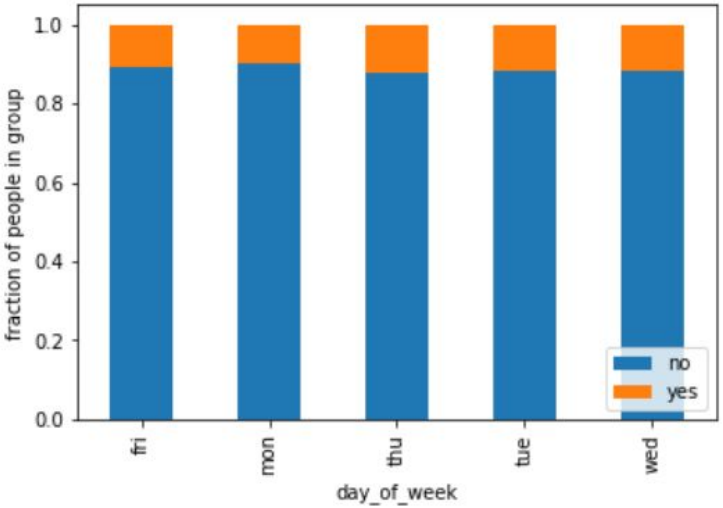
Last contact day of the week?

(day_of_week,cat)

-Nearly randomly distributed

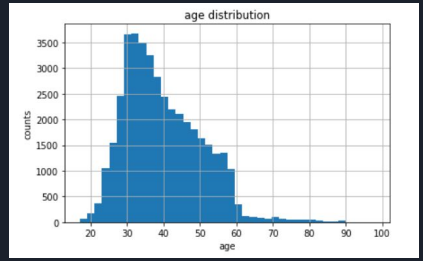
-No contact bias in each weekday of a week





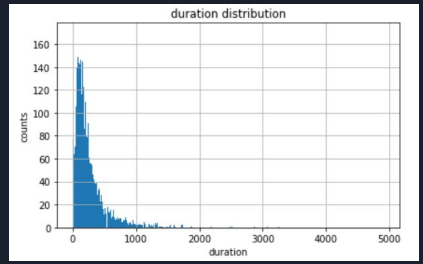
EDA

age(num)

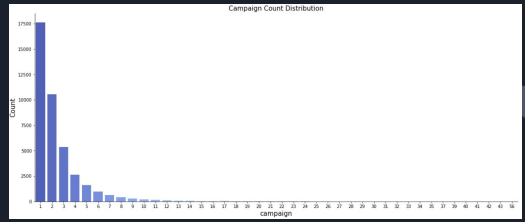


Duration:last contact durarion,

in seconds,highly affects the output

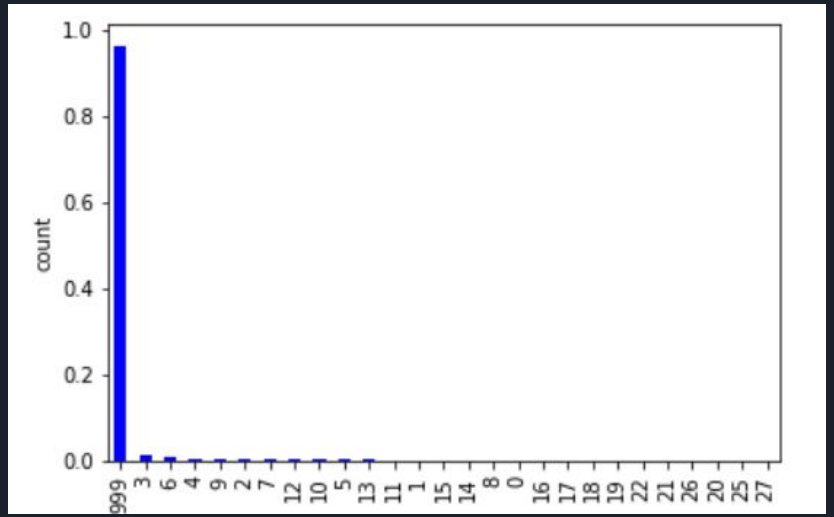


Campaign:number of contacts performed during this campaign (num)



EDA

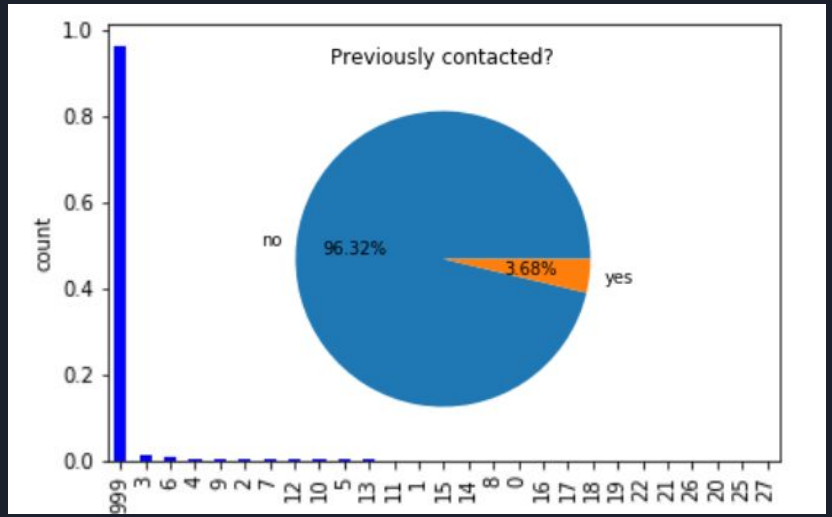
Pdays : number of days passed by after the client was last contacted(num,999 means client was not previously contacted)



EDA

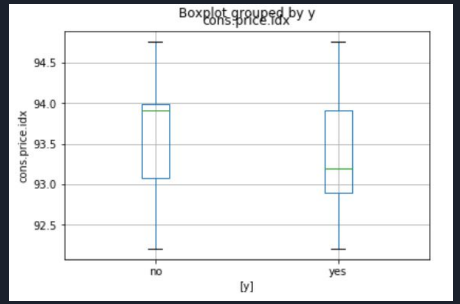
Pdays : number of days passed by after the client was last contacted(num,999 means client was not previously contacted)

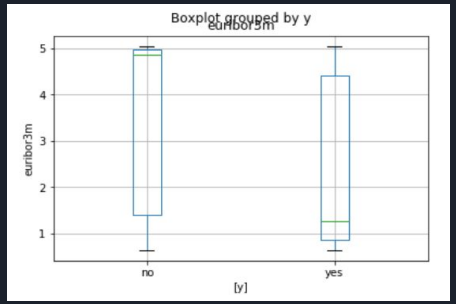
-use variable: pdayscon



EDA

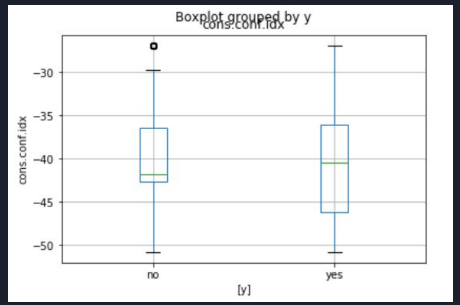
Consumer price index(num,quarterly)

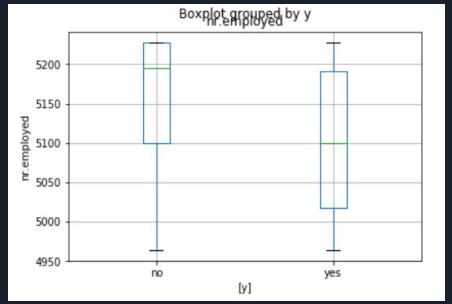




Euro interbank offered rate(num,daily)

Consumer confidence index(num,monthly)

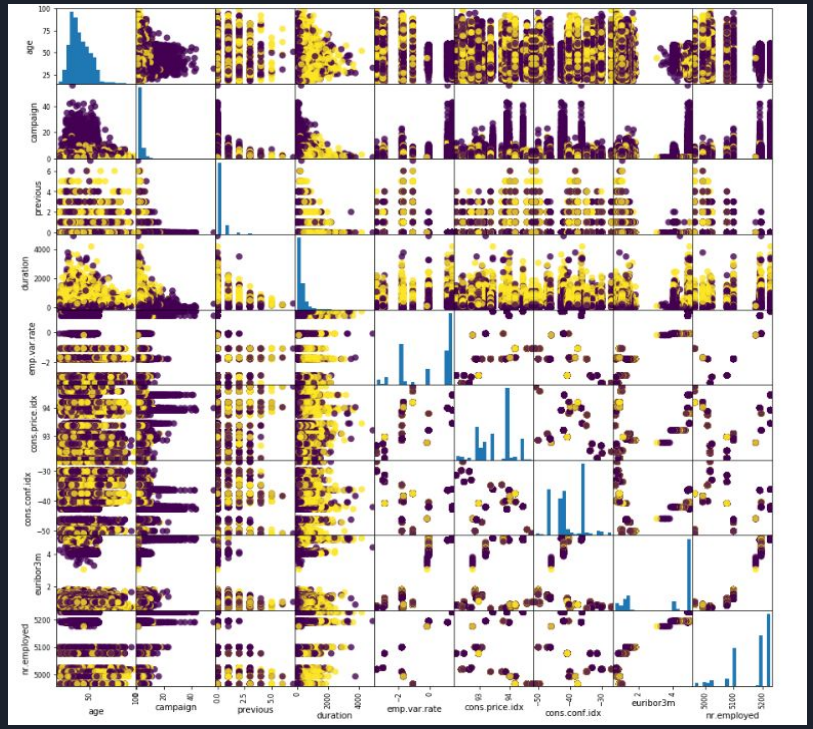




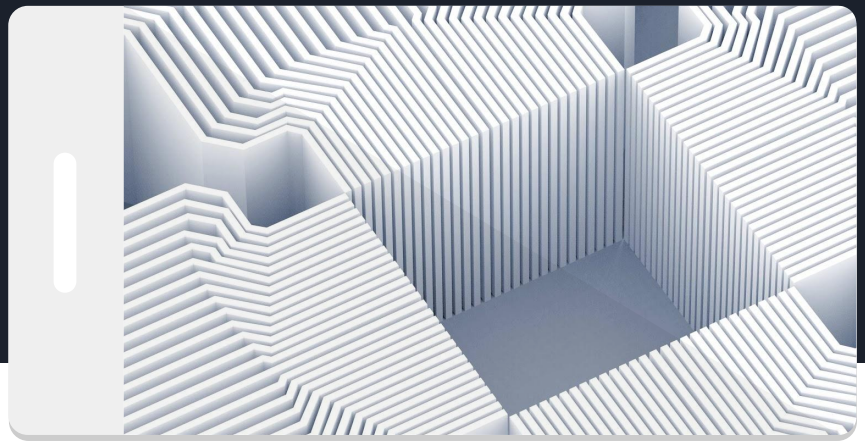
Total number of employees(num,quarterly)

EDA

Interactions between features



Questions?



Thank you!