EDA (Exploratory Data Analysis)

Week 11 - LISUM 04

José Vicente Solorzano

29/11/21

Agenda

- Data understanding
- Clients
- Last contact
- Campaign
- Bivarial analysis of some features
- Recommended models

Data Understanding

Features are divided into 4 groups:

1. Data related to clients (age, job, marital, education, default, housing and loan).

2. Data related with the last contact of the current campaign (contact, month, day_of_week and duration).

3. Data related with campaigns (campaign, pdays, previous and poutcome).

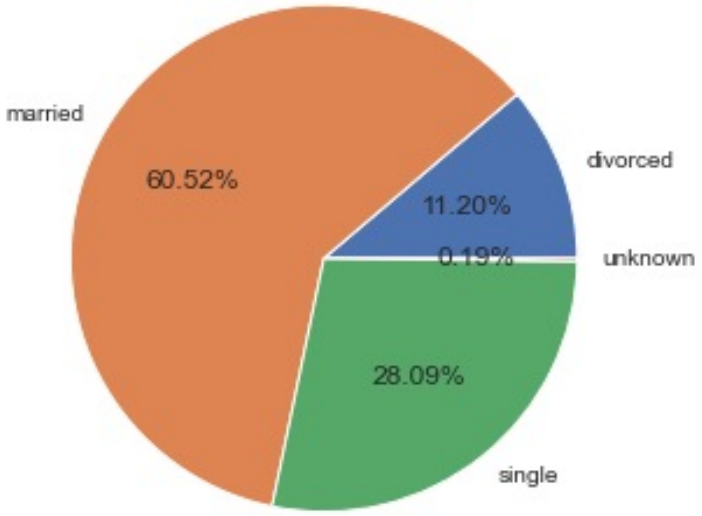
4. Data related with socio economic context attributes (emp.var.rate, cons.Price.idx, cons.conf.idx, euribor3m, nr.employed).

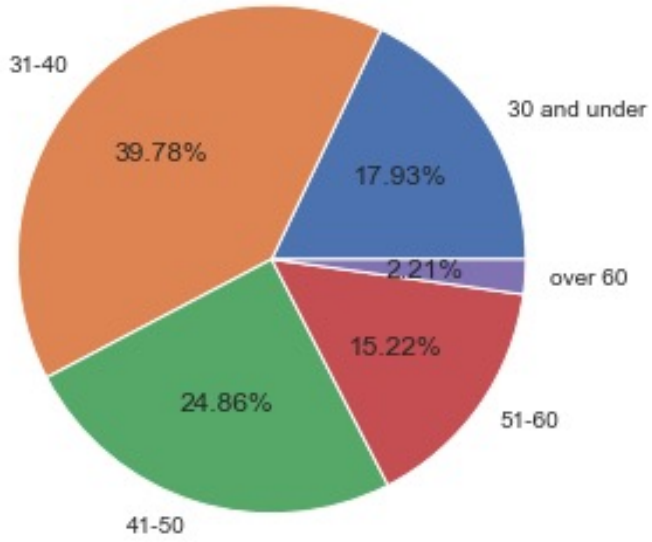
5. TARGET: “y”.

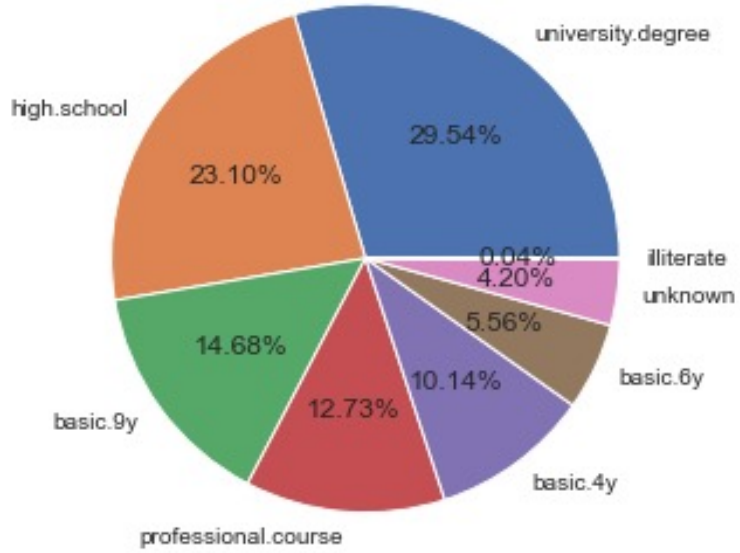
```
Number of rows:  41188  -  Number of columns:  21
```

```
['age', 'job', 'marital', 'education', 'default', 'housing', 'loan',
 'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays',
 'previous', 'poutcome', 'emp.var.rate', 'cons.price.idx',
 'cons.conf.idx', 'euribor3m', 'nr.employed', 'y'],
```

Client





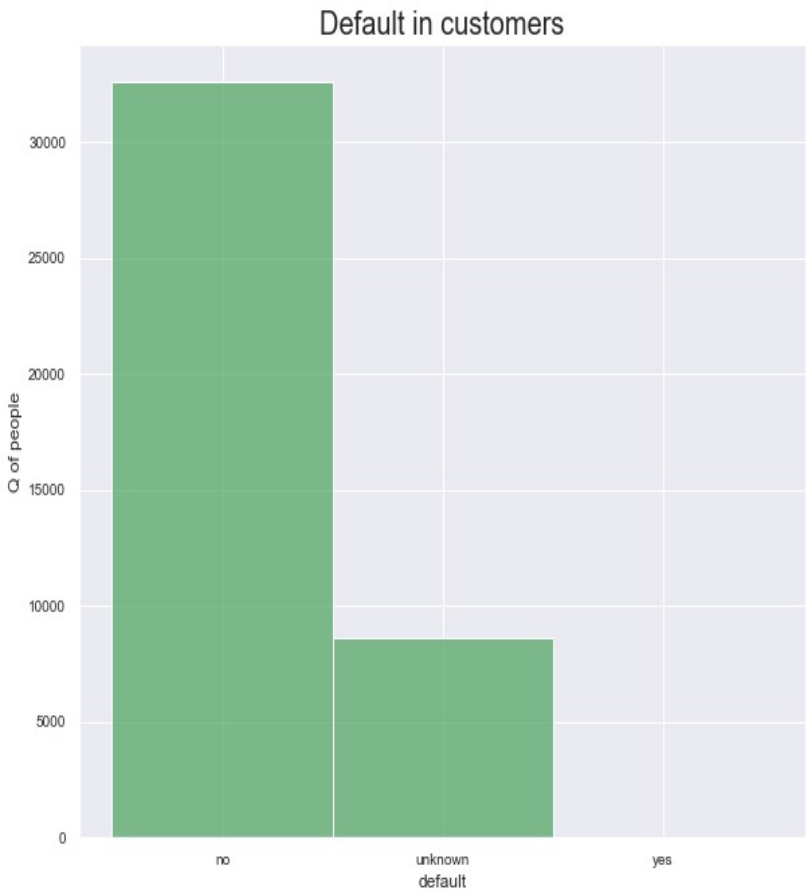


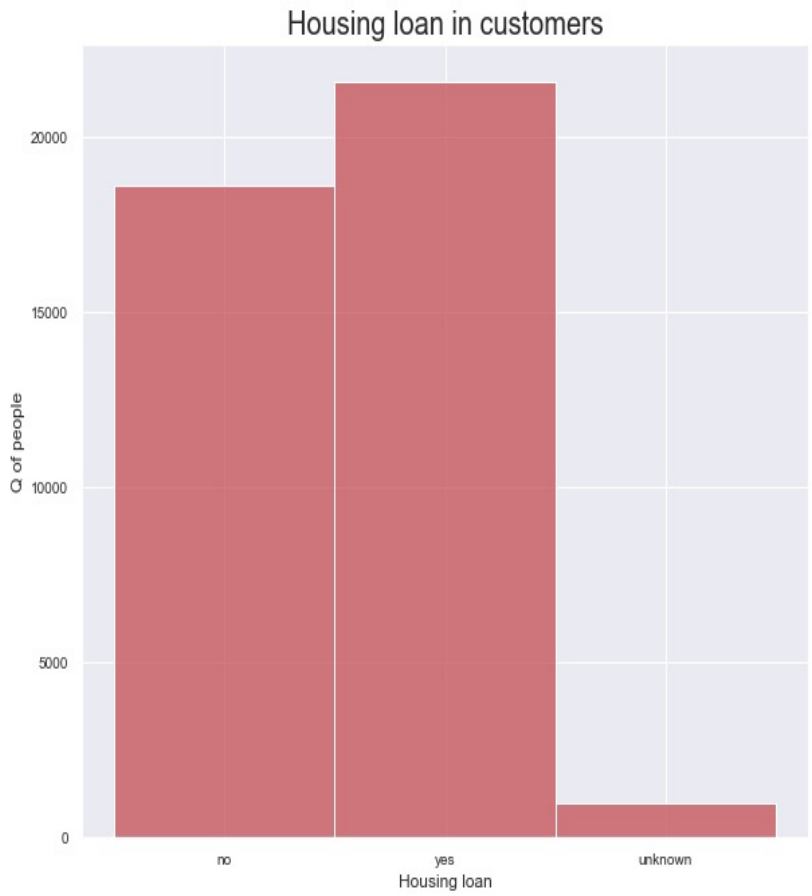
Customers job (%):

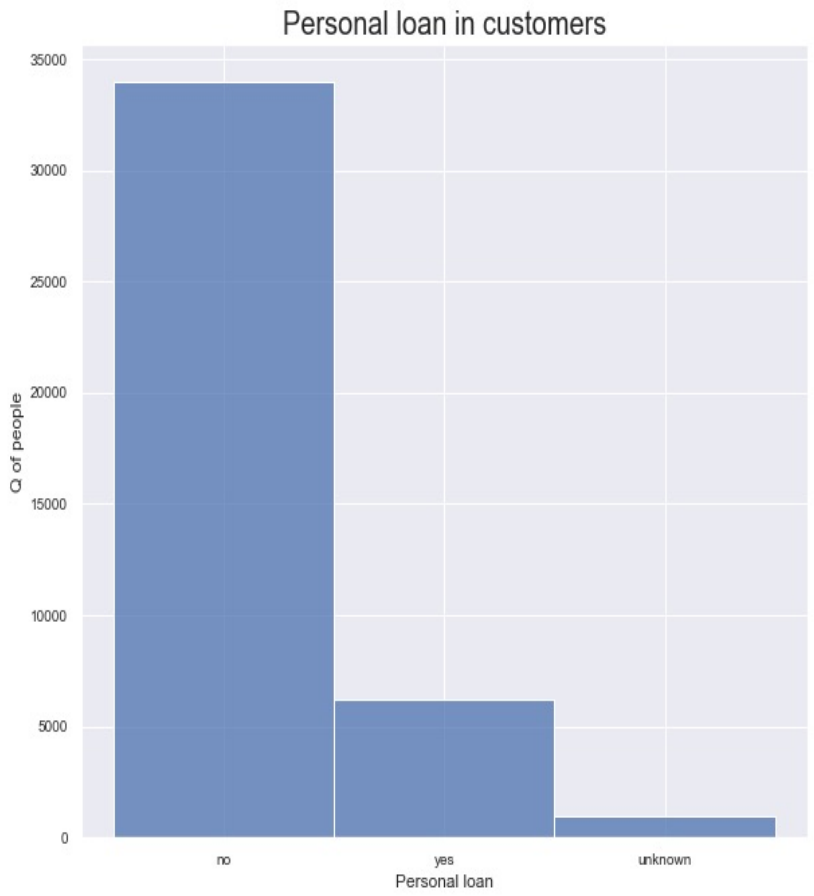
job	
admin.	25.303486
blue-collar	22.467709
technician	16.371273
services	9.636302
management	7.099155

Customer's demo:
60% are married.
64% are admin, blue-collar or technician.
65% are 31-50 yo.
65% are graduated from HS, University or professional course.

Customers and Default, Housing Loan and Personal Loan:

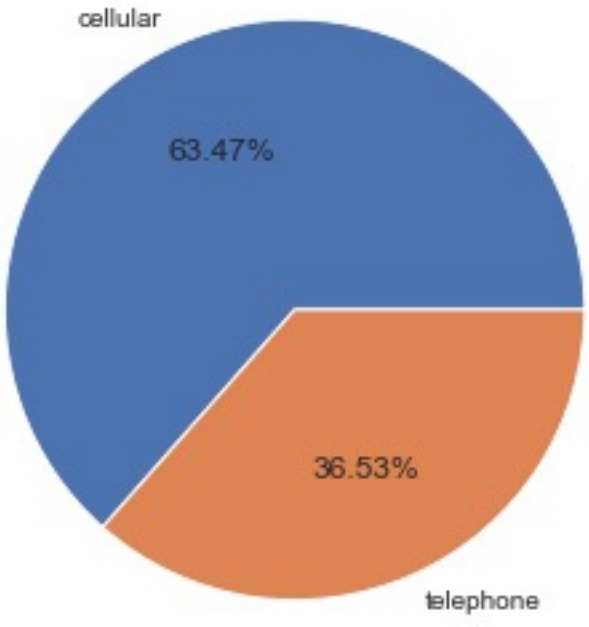






Last Contact

Data Glacier
Your Deep Learning Partner

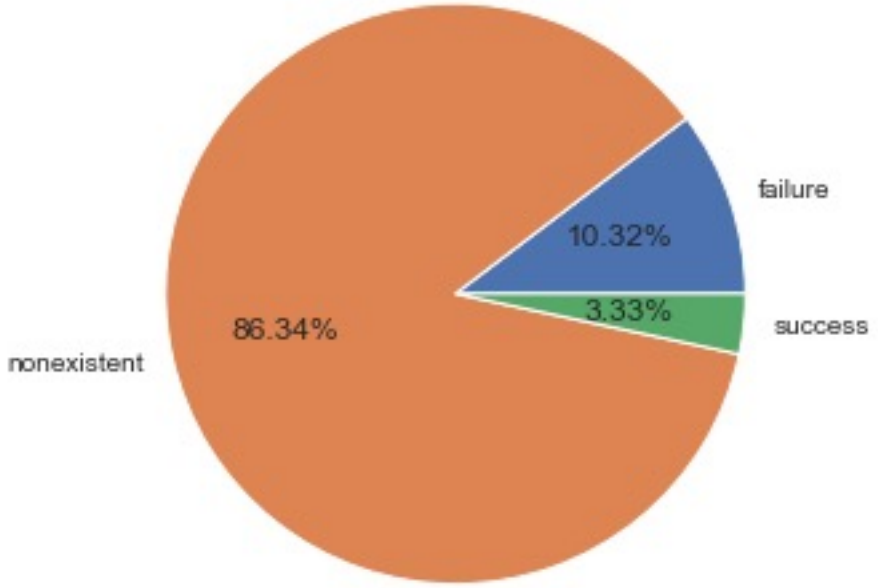


month	contact
may	33.429640
jul	17.417694
aug	14.999514
jun	12.911528
nov	9.956784

```
duration_range
0-200          55.258813
201-400        27.347771
401-600         8.983199
601-800         4.144411
801-1000        1.944741
over 1000       2.321064
```

63% of contacts to customers are done by celular.
78% of contacts to customers are done in 4 months (may, july, august and june).
82% of calls last less 400 seconds.

Campaign



campaign_range	
0-3	81.463047
4-6	12.695445
7-9	3.185394
10-12	1.279499
More than 12	1.376615

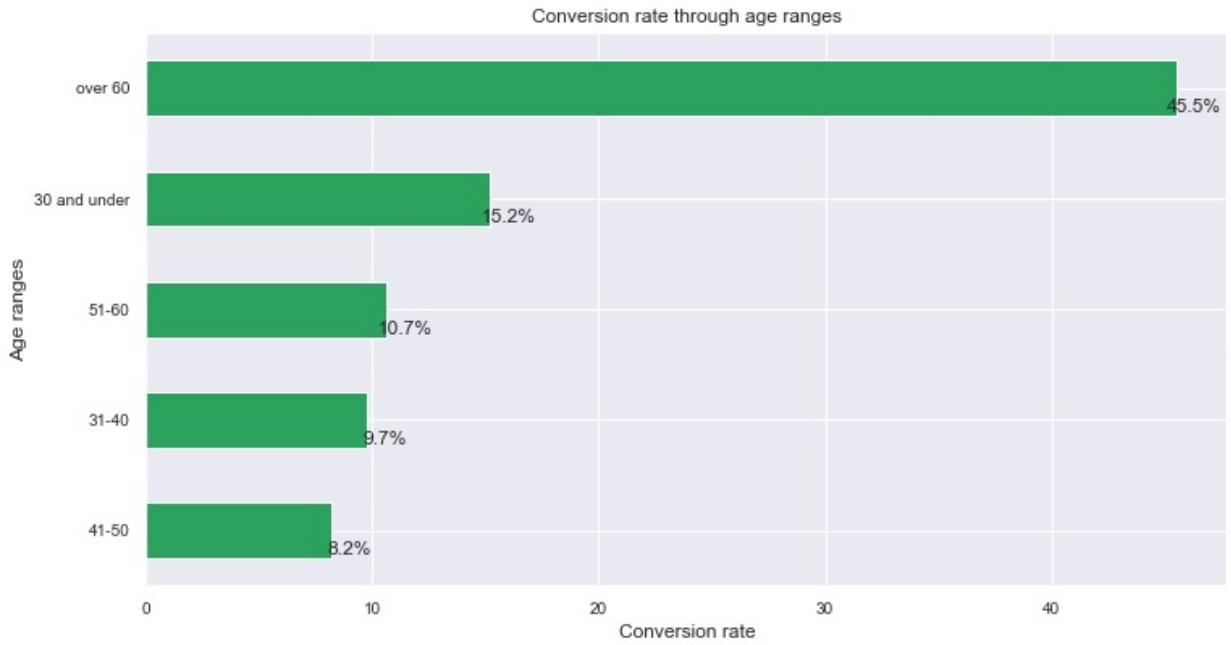
Name: campaign, dtype: float64

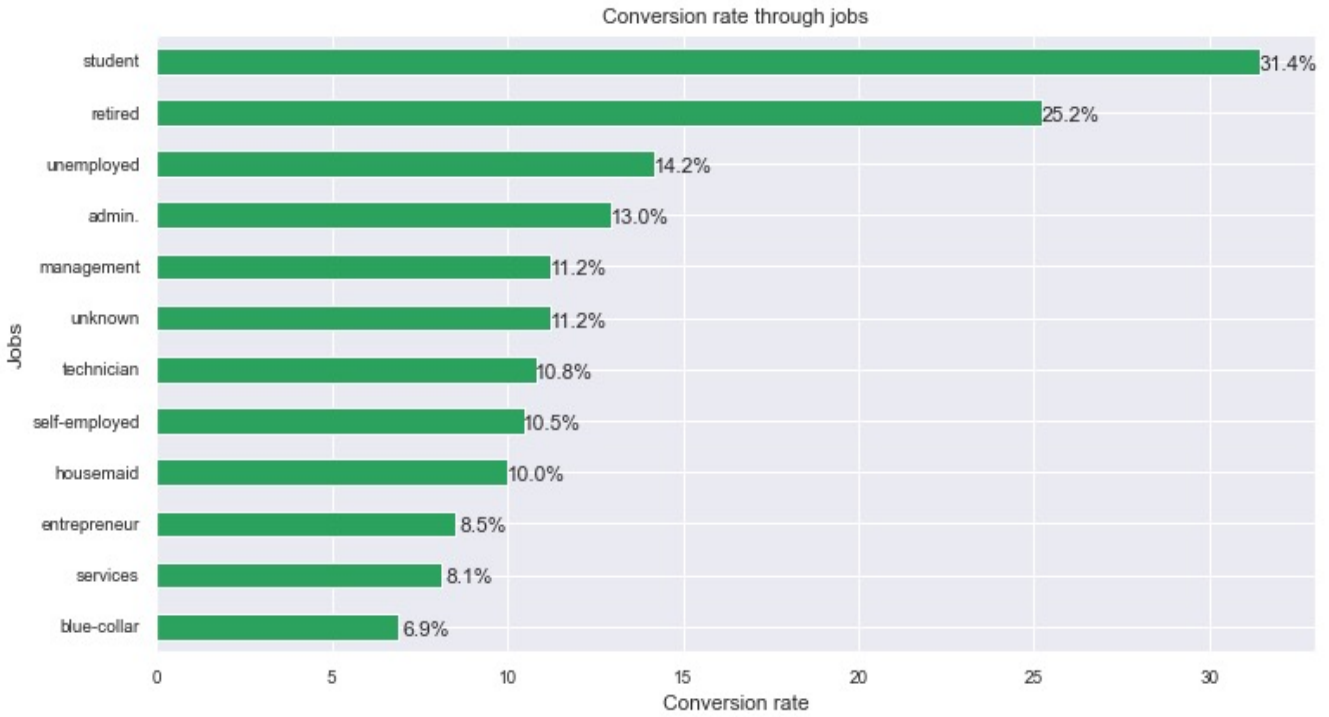
previous	
0	86.343110
1	11.073614
2	1.830630
3	0.524425
4	0.169952
5	0.043702
6	0.012139
7	0.002428

86% of customers dont have previous relation with the bank.
3,3% succes of the 14% of customers who have previous contacts with the bank.

Bivarial for Some features Age and job

Data Glacier
Your Deep Learning Partner





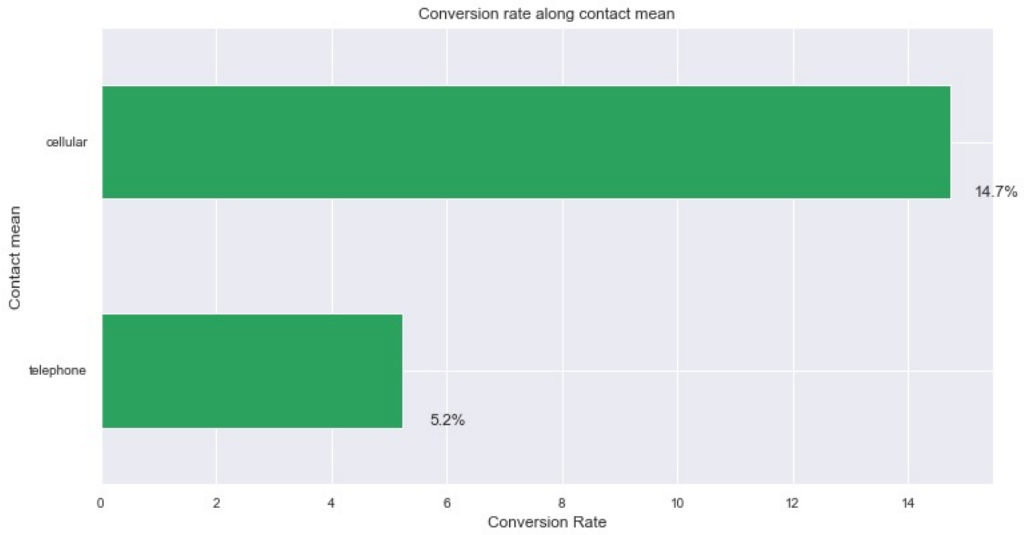
Jobs and target:

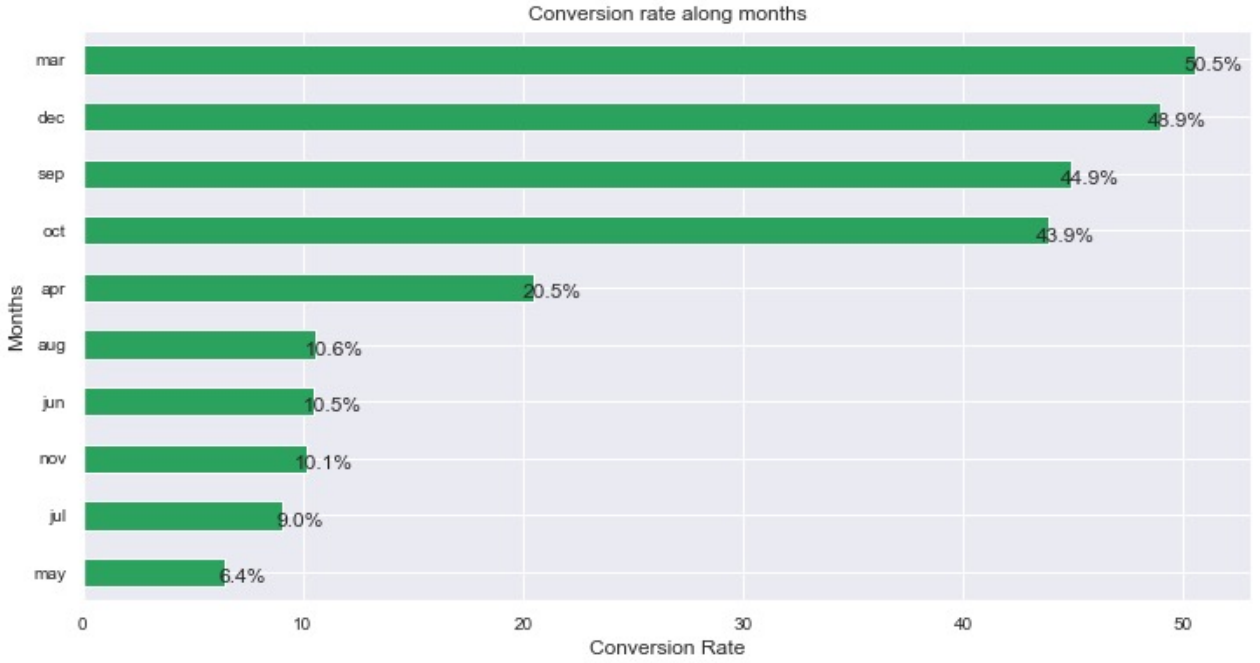
y	no	yes
job		
admin.	87.027442	12.972558
blue-collar	93.105684	6.894316
entrepreneur	91.483516	8.516484
housemaid	90.000000	10.000000
management	88.782490	11.217510
retired	74.767442	25.232558
self-employed	89.514426	10.485574
services	91.861930	8.138070
student	68.571429	31.428571
technician	89.173958	10.826042
unemployed	85.798817	14.201183
unknown	88.787879	11.212121

People under 30 and over 60 yo are customers with higher acceptance rate. 15% and 45% respectively.
Accordinly, student and retired are top Jobs in acceptance rate.

Bivarial for
Some features
Contact medium, month and call duration

Data Glacier
Your Deep Learning Partner



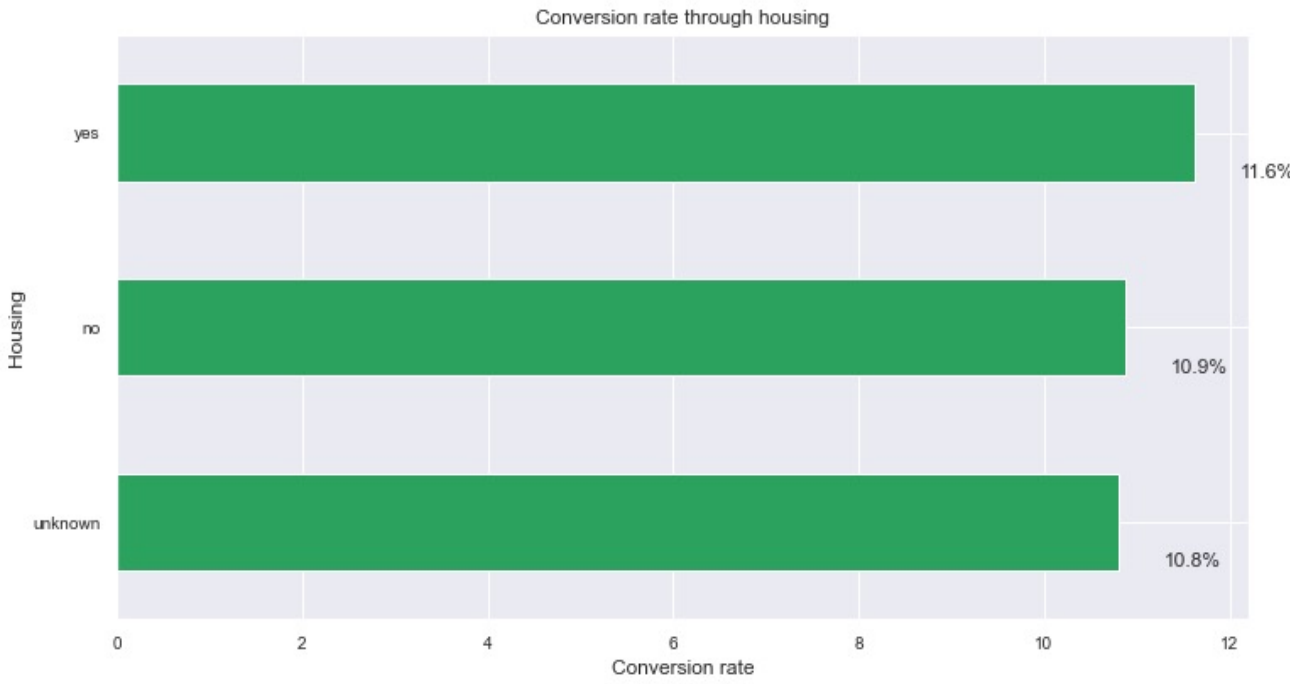


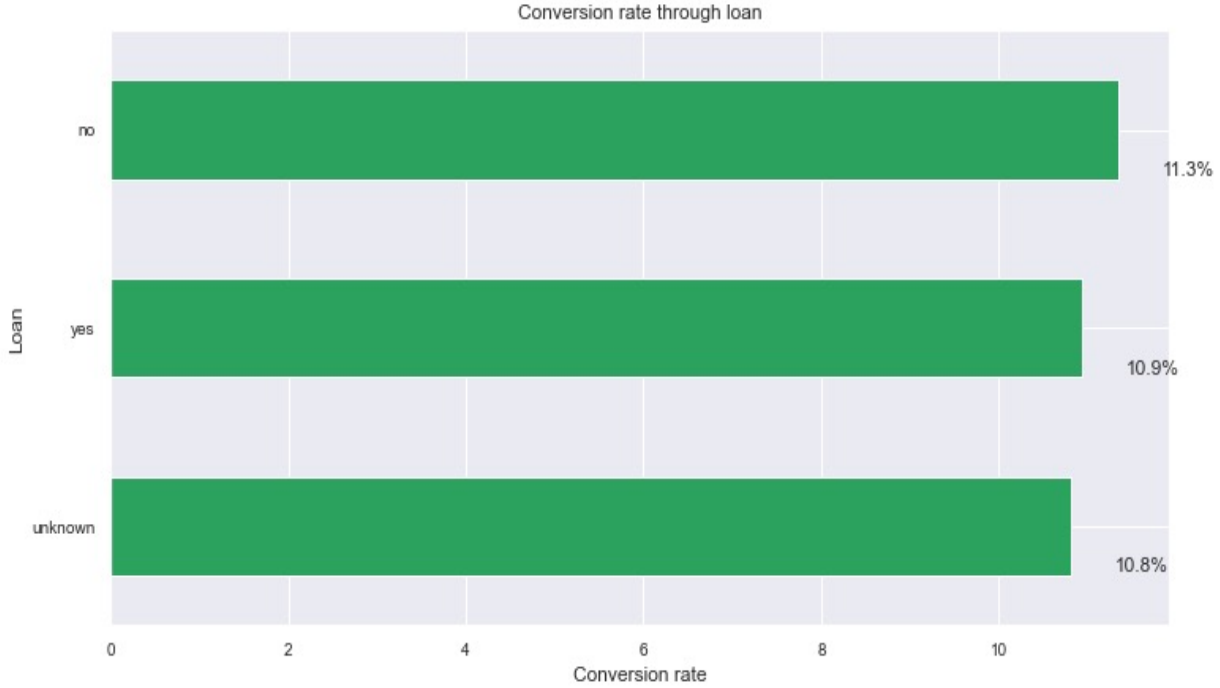
duration_range	y	no	yes	totals	% no	% yes
0-300		28466	1518	29984	94.937300	5.062700
301-600		6302	1438	7740	81.421189	18.578811
601-900		1259	924	2183	57.672927	42.327073
901-1200		329	439	768	42.838542	57.161458
1201-1500		115	192	307	37.459283	62.540717
1501-1800		36	69	105	34.285714	65.714286
1801-2100		22	34	56	39.285714	60.714286
2101-2400		6	8	14	42.857143	57.142857
More than 2400		13	18	31	41.935484	58.064516

15% of people contacted by celular accepted the term deposit, while just 5% of the ones contacted by telephone.
December, march, october and september have few customers, but with very high conversión rate (over 40%).
Longer call, higher conversión rate. Consider that a small quantity of people have calls for more tan 900 seconds.

Bivarial for Some features Default, Loan and Housing

Data Glacier
Your Deep Learning Partner





Default and target:

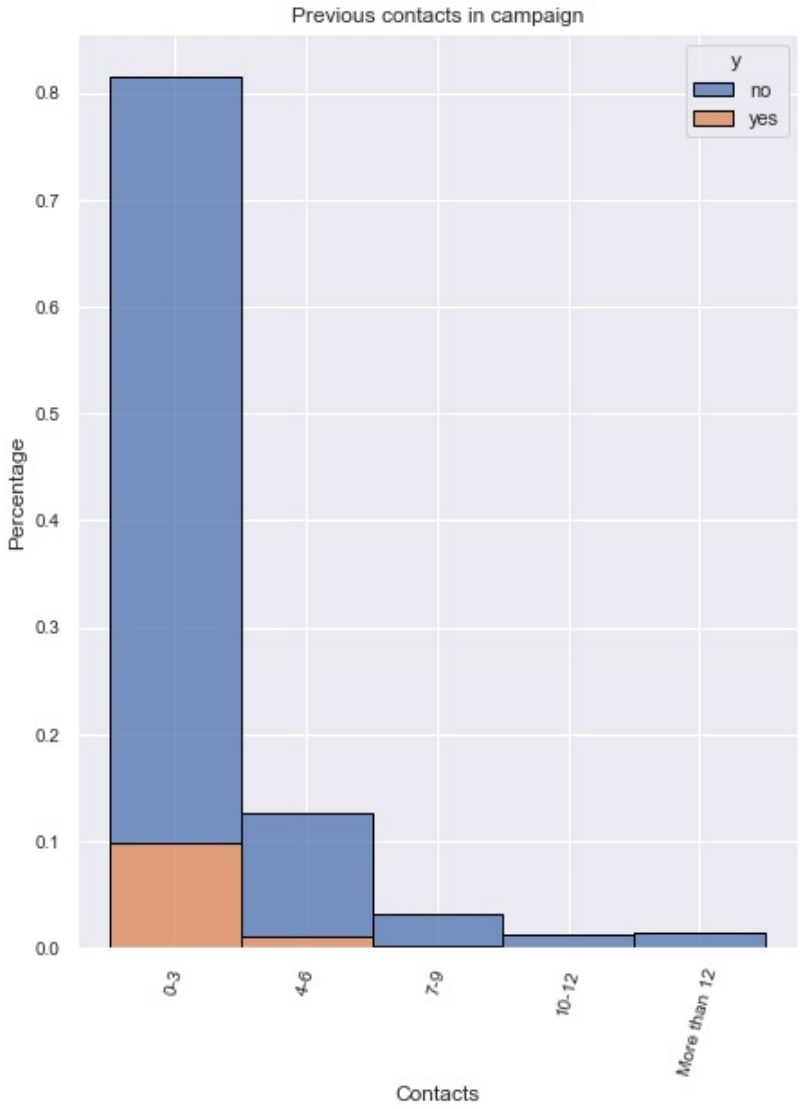
y	no	yes
default		
no	87.121026	12.878974
unknown	94.847040	5.152960
yes	100.000000	0.000000

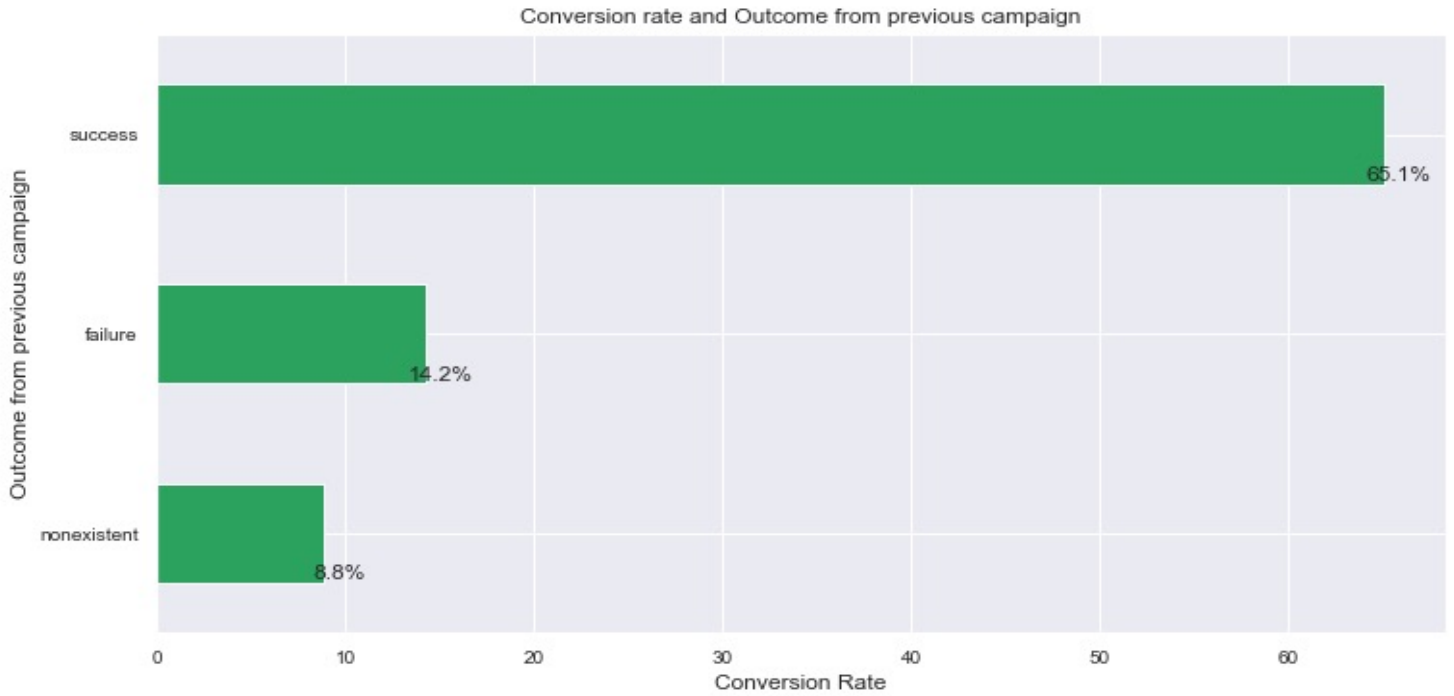
Conversion rate in Loan and Housing is similar for people with, without and unknown.

For Default is different. Conversion rate of “no default” people is 13%, and unknown is 5%.

Bivarial for Some features Campaign and target

Data Glacier
Your Deep Learning Partner



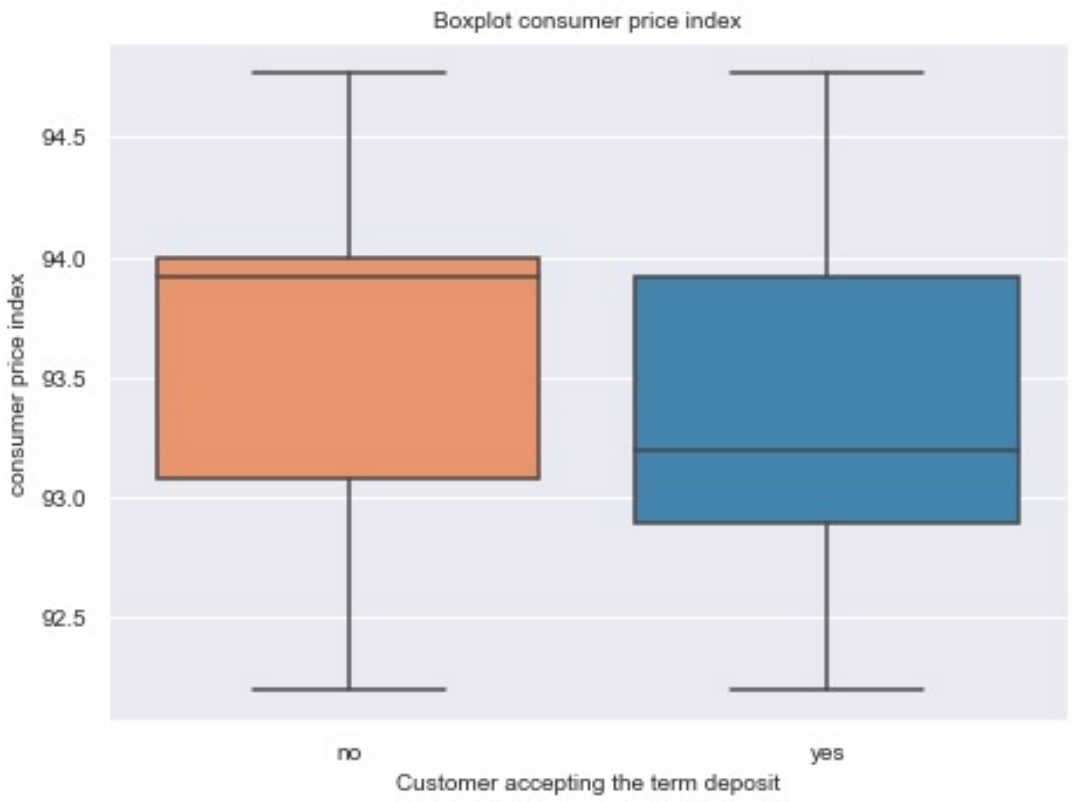


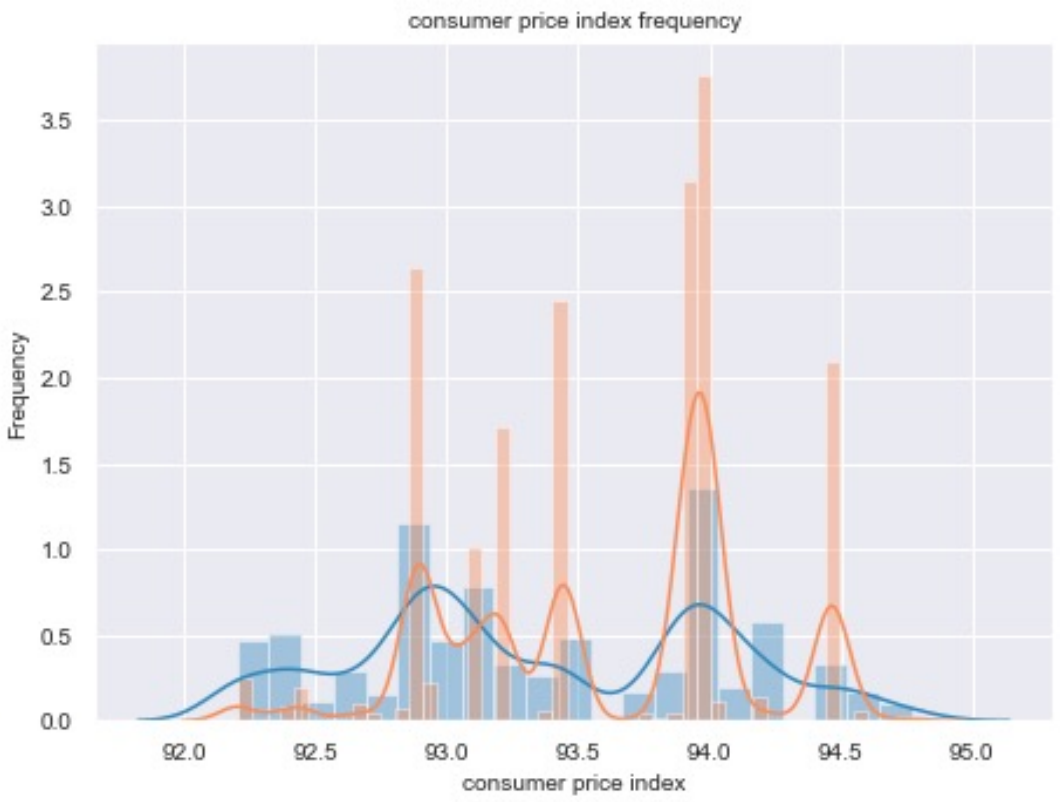
Less number of contacts during the campaign seems to be better for conversion… 12% for 0-3 contacts and 9% for 4-6.

Success in previous campaign is very important for conversión rate. 65% of customers with previous success, are accepting now the term deposit.

Bivarial for
Some features
Socioeconomics and target

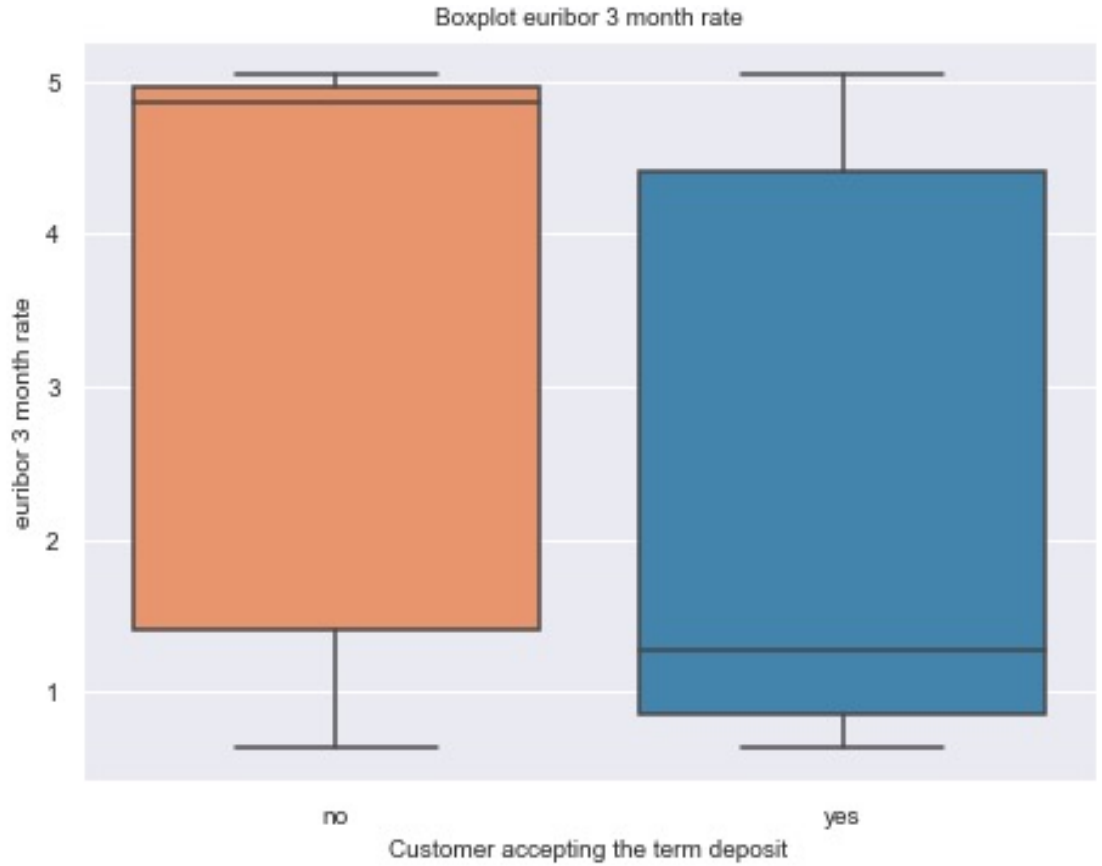
Data Glacier
Your Deep Learning Partner

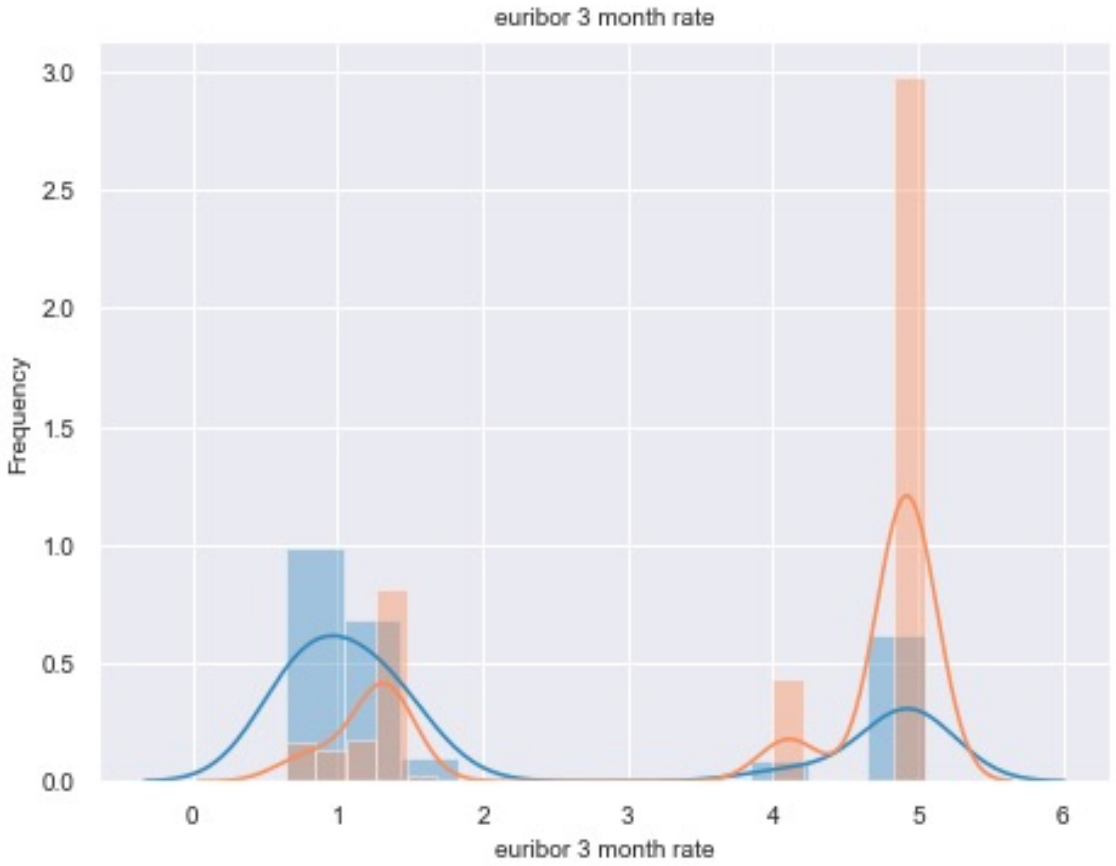




Bivarial for
Some features
Socioeconomics and target

Data Glacier
Your Deep Learning Partner





Recommended Model

Recommended Model

Data Glacier
Your Deep Learning Partner

We implemented pycaret library to know the model would perform better in our business problem.

The criterion was based mainly in AUC (area under the curve), the results were:

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.8960	0.8001	0.3245	0.5671	0.4126	0.3602	0.3772	0.6410
catboost	CatBoost Classifier	0.8976	0.7950	0.3104	0.5866	0.4057	0.3555	0.3772	6.8270
lr	Logistic Regression	0.8239	0.7908	0.6367	0.3468	0.4489	0.3548	0.3781	0.6580

We are going to try with the 2 best models:

1) Light Gradient Boosting Machine
2) CatBoost Classifier

Next week we are going to implement some Grid Search and hyperparameters optimization to get the ideal model.

Thank You