UpGrad: Assignment

Rohit Kr. Bose

27 February 2019

1 Data preprocessing

- All categorical text data is converted to numeric data.
- Unknown data is replaced with random data sampled from its column.

2 Features

- To see which variables are important in determining the target variable (y)
- Random Forest Classifier is used to compute feature importances

age	0.0617
job	0.0311
marital	0.0153
education	0.027
default	8.297e-06
housing	0.0113
loan	0.0079
contact	0.0118
month	0.0209
day_of_week	0.0266

duration	0.413
campaign	0.0295
pdays	0.0175
previous	0.0137
poutcome	0.0106
emp.var.rate	0.0536
cons.price.idx	0.0235
cons.conf.idx	0.0324
euribor3m	0.1002
nr.employed	0.0903

- It can be observed that **duration** is an important feature. However, this feature cannot be controlled before a call; hence we do not act upon this.

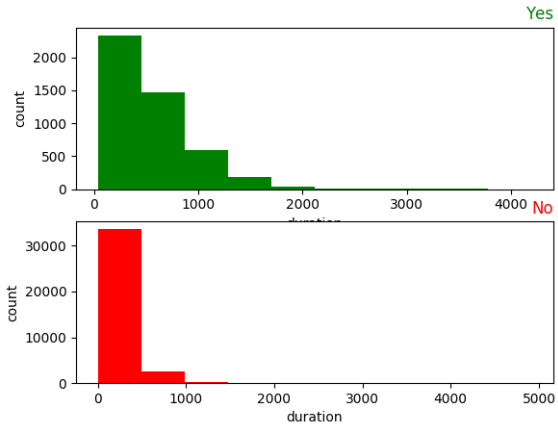

Figure 1: Duration

3 Task

Investigation of the dataset provided two important features that help in our task:

3.1 Campaign

Campaign : Number of contacts performed during this campaign and for this client

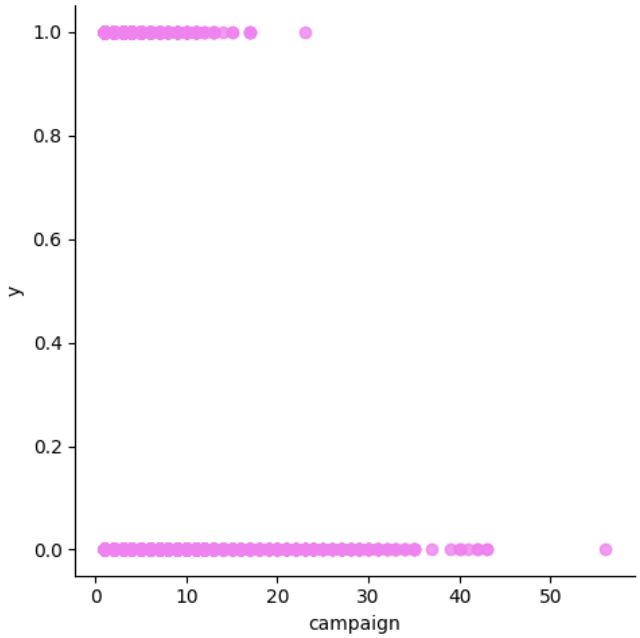

Figure 2: Campaign

Evidently, there is a drop from 1 to 0 i.e. from customers saying yes to no after a certain threshold i.e. after a certain number of calls. This makes this feature interesting.
We ran a loop that found the optimum number of calls (k) for which the customer remains interested. It is not of our best interest to call after it. Although it does increase prospects overall, it increases the cost. We need to optimize jointly for both parameters. As given in the problem, reduction in market cost is given by $X\%$ and $Y\%$ of prospects are to be acquired. Let us call the conversion rate C. **Conversion rate** is calculated as the number of potential customers divided by the number that actually subscribe. The average value of C is known to be 2% in the industry, hence we use that.

k	X	Y	C
1	61.05	49.57	0.06
2	38.79	75.67	0.05
3	26.52	88.04	0.04
4	19.30	93.41	0.03
5	14.59	95.99	0.02
6	11.38	97.61	0.02
7	9.11	98.43	0.01
8	7.43	98.79	0.01
9	6.13	99.16	0.01
10	5.09	99.42	0.01
11	4.27	99.68	0.01
12	3.62	99.74	0.00

k	X	Y	C
13	3.08	99.83	0.01
14	2.63	99.85	0.00
15	2.25	99.89	0.00
16	1.91	99.89	0.00
17	1.62	99.98	0.01
18	1.39	99.98	0.00
19	1.19	99.98	0.00
20	1.01	99.98	0.00
21	0.87	99.98	0.00
22	0.74	99.98	0.00
23	0.63	100.00	0.01
24	0.54	100.00	0.00

From this table, **6** appears to be the best stopping threshold. After this, the conversion rate drops below 2%. Hence, a seventh call is not advisable, as after any call after 6 has a very low conversion rate. Note that this strategy will lead to **11.38%** reduction in cost and still we can retain **97.61%** of prospects.

3.2 Job

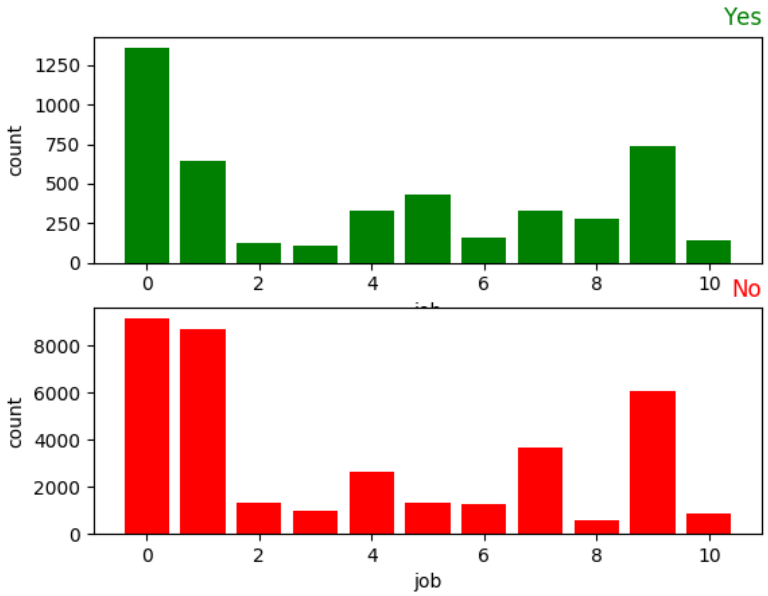

Figure 3: Job

Here, it can be observed that categories 2 (entrepreneur) and 3 (housemaid) have very less representation. If we stop calling these people altogether, we have the following results.

Category removed	X	Y
Entrepreneur	3.53	97.33
Housemaid	2.67	97.72
E + H	6.20	95.04

Evidently, it is not possible to jointly optimize here. If we prefer cost reduction over customer base, we can stop calling people from both categories; else we should simply stop calling housemaids.