
MiniEDA

1000 filas
Perfectamente balanceado 500 y 500.

El enunciado no nos da más información así que utilizamos el accuracy como métrica.

Discusión? Qué nos interesa? Precision? Recall? Accuracy?

Más de cerca y discusión:
Timestamp: Time at which consumer clicked on Ad or closed window
Male: si tienen esa info probablemente será una red social. En tal caso deberían tener mucha más información. Teorías? (Correlación muy baja, por cierto)

DataPreparation, Feature Engineering and Feature Selection

Intentamos convertir features a numéricas y generar otras.

Cuando generamos una comprobamos si es “diferencial” con respecto a nuestro target. Recordad el EDA del Titanic..

LabelEncoder, get_dummies (OneHotEncoding): demasiado disperso.

Shannon: para ser importante tiene que ser diferencial. Lo que más información tiene.

*Según Claude Shannon, en su teoría de la información, **la información** se define como una medida de la reducción de la incertidumbre sobre un sistema o evento tras recibir un mensaje o dato. Shannon planteó que la cantidad de información transmitida está relacionada con la probabilidad de que ocurra el evento en cuestión: cuanto menos probable sea un evento, más información aporta cuando ocurre.*

Formalización matemática:

La cantidad de información asociada a un evento x con probabilidad $P(x)$ se define como:

$$I(x) = -\log_2 P(x)$$

Donde:

- $I(x)$ es la cantidad de información, medida en **bits** (si el logaritmo es en base 2).
- $P(x)$ es la probabilidad del evento x.

Describe:

- Distintas escalas -> Scaler

Descartamos:

- City_encoded por que tiene demasiados valores distintos
- Country por el mismo motivo

Aniadimos:

- Topic_solution porque parece discriminatoria (value_counts). Se podría hacer lo mismo con otras palabras pero ya veremos la forma de aproximar el problema de forma más sistemática. Baja correlación. En realidad la aniadimos porque nos ha costado trabajo generarla.
- month: diferencias de en torno al 5% (no muy altas)
- week_day: diferencias de en torno al 5% (no muy altas)
- month_dau: diferencias >10%
- hour: diferencias de entre 5 y 10 %
- (En el correlation matrix vemos que el coeficiente de Pearson de todas nuestras nuevas variables es muy cercano a 0).

Obviamos:

- La alta correlación entre Daily Internet Usage y Daily Time Spent on Site 0.52

```
X = df[['Daily Time Spent on Site', 'Age', 'Area Income',
       'Daily Internet Usage', 'Male', 'Topic_solution', 'month',
       'month_dau', 'week_day', 'hour']]
```

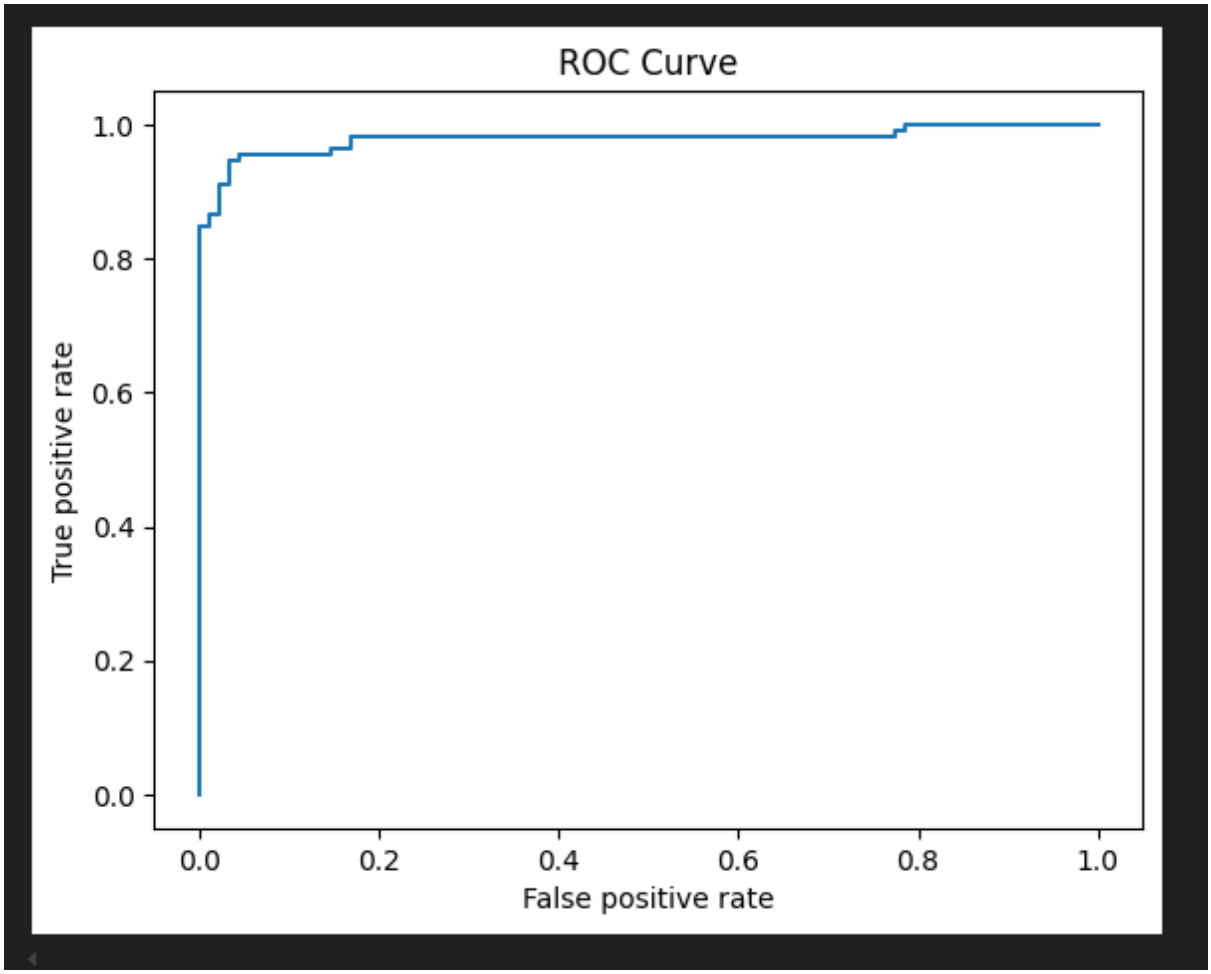
Conviene utilizar un StandardScaler. Escalas muy distintas entre las variables features.

Crossvalidation: Train accuracy: 0.97 +/- 0.02

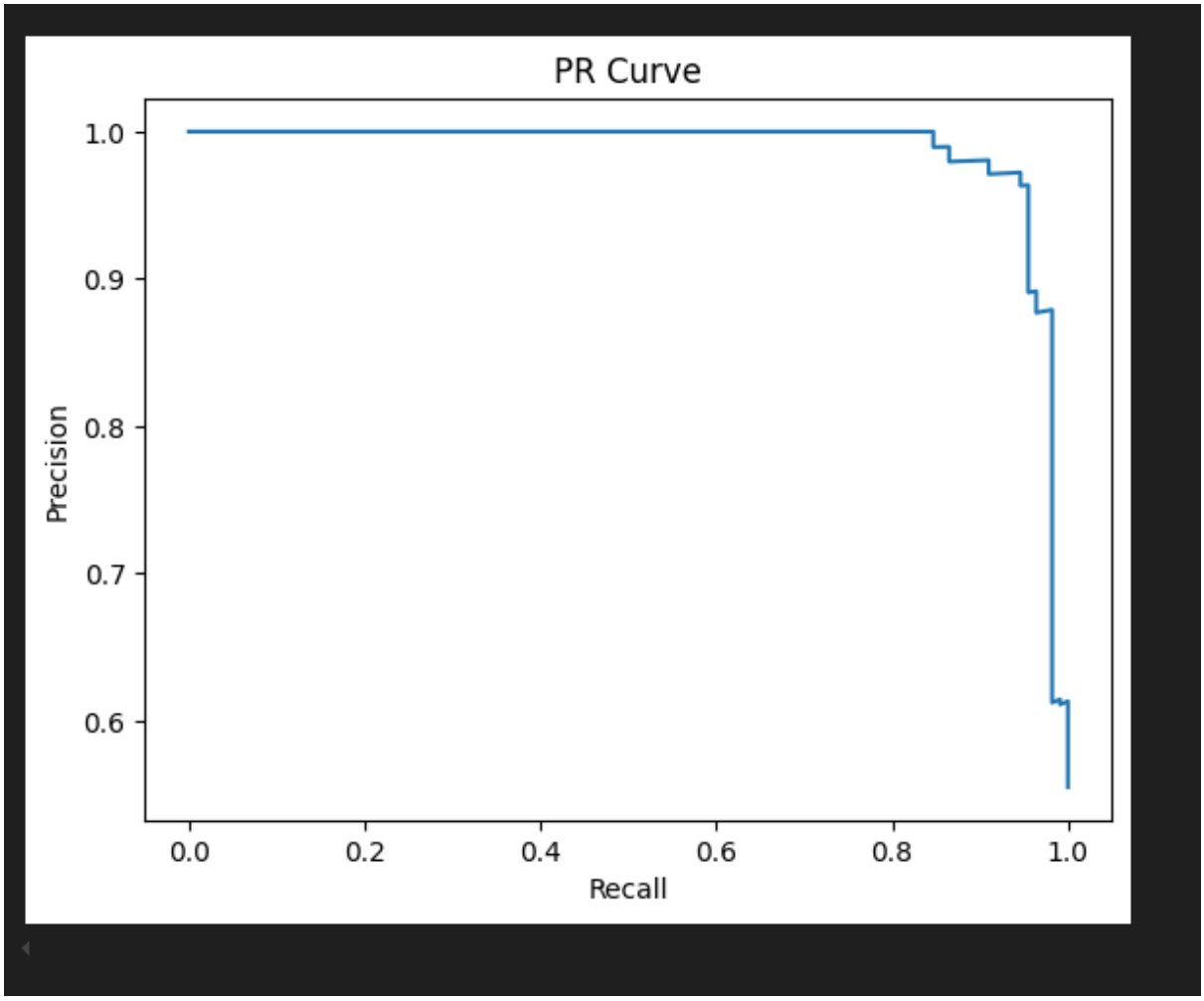
Test accuracy: 0.95

Damos por bueno el fit. Todavía no hemos visto formas de regularizar el LogisticRegression.

	tpr	fpr	threshold
0	0.000000	0.000000	inf
1	0.009009	0.000000	0.999995
2	0.846847	0.000000	0.946610
3	0.846847	0.011236	0.932655
4	0.864865	0.011236	0.885341
5	0.864865	0.022472	0.877866
6	0.909910	0.022472	0.739909
7	0.909910	0.033708	0.730895
8	0.945946	0.033708	0.540375
9	0.945946	0.044944	0.537639
10	0.954955	0.044944	0.497809
11	0.954955	0.146067	0.167099
12	0.963964	0.146067	0.154704
13	0.963964	0.168539	0.153334
14	0.981982	0.168539	0.126788
15	0.981982	0.775281	0.006483
16	0.990991	0.775281	0.006430
17	0.990991	0.786517	0.006128
18	1.000000	0.786517	0.006083
19	1.000000	1.000000	0.002110

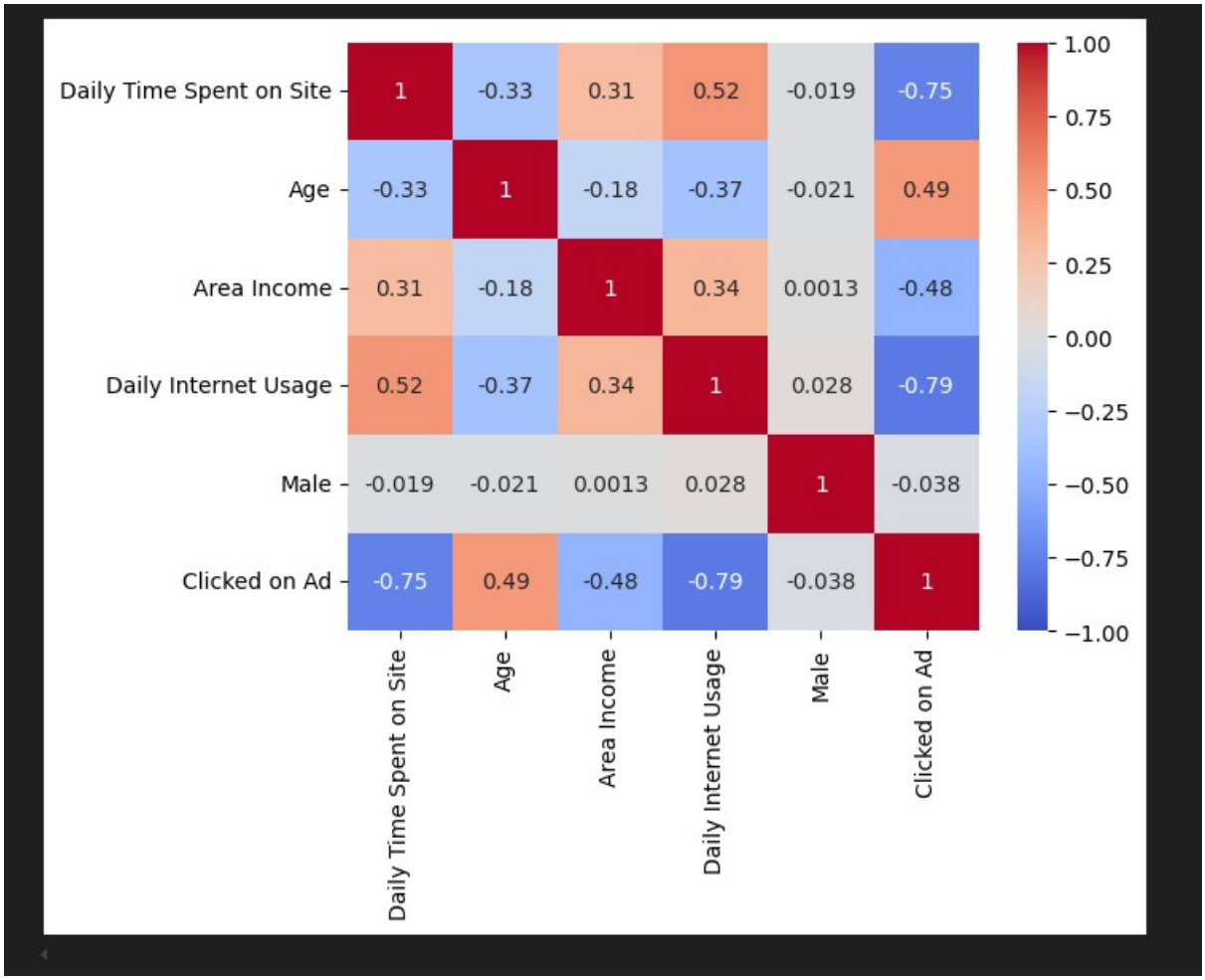

ROC Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

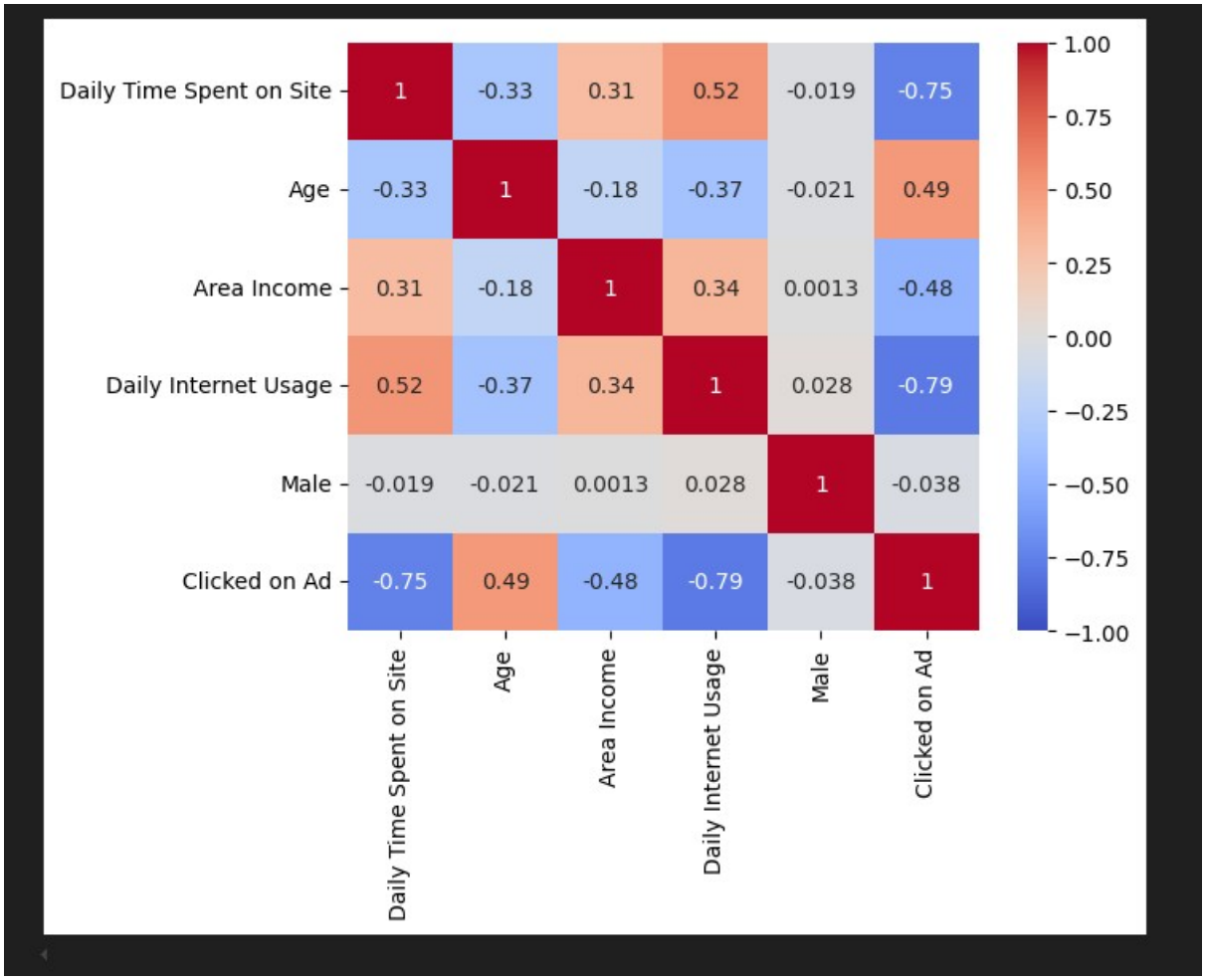
	prec	rec	threshold
80	0.890756	0.954955	0.154704
81	0.898305	0.954955	0.167099
82	0.905983	0.954955	0.171075
83	0.913793	0.954955	0.183012
84	0.921739	0.954955	0.188278
85	0.929825	0.954955	0.205331
86	0.938053	0.954955	0.336713
87	0.946429	0.954955	0.408484
88	0.954955	0.954955	0.461044
89	0.963636	0.954955	0.483095
90	0.963303	0.945946	0.497809
91	0.972222	0.945946	0.537639
92	0.971963	0.936937	0.540375
93	0.971698	0.927928	0.597056
94	0.971429	0.918919	0.617113
95	0.971154	0.909910	0.677269
96	0.980583	0.909910	0.730895
97	0.980392	0.900901	0.739909
98	0.980198	0.891892	0.740816
99	0.980000	0.882883	0.776378



Confirmamos las hipótesis del feature selection:

```
Feature Importances:
Daily Internet Usage: 2.8293
Daily Time Spent on Site: 2.5640
Area Income: 1.6707
Age: 1.2452
Male: 0.3295
month: 0.2574
week_day: 0.2314
month_dau: 0.1908
Topic_solution: 0.1542
hour: 0.1057
```


Daily Time Spent on Site
Age
Area Income
Daily Internet Usage
Male
Clicked on Ad
1 -0.33 0.31 0.52 -0.019 -0.75
-0.33 1 -0.18 -0.37 -0.021 0.49
0.31 -0.18 1 0.34 0.0013 -0.48
0.52 -0.37 0.34 1 0.028 -0.79
-0.019 -0.021 0.0013 0.028 1 -0.038
-0.75 0.49 -0.48 -0.79 -0.038 1
Daily Time Spent on Site
Age
Area Income
Daily Internet Usage
Male
Clicked on Ad
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Daily Time Spent on Site
Age
Area Income
Daily Internet Usage
Male
Clicked on Ad
1 -0.33 0.31 0.52 -0.019 -0.75
-0.33 1 -0.18 -0.37 -0.021 0.49
0.31 -0.18 1 0.34 0.0013 -0.48
0.52 -0.37 0.34 1 0.028 -0.79
-0.019 -0.021 0.0013 0.028 1 -0.038
-0.75 0.49 -0.48 -0.79 -0.038 1
Daily Time Spent on Site
Age
Area Income
Daily Internet Usage
Male
Clicked on Ad
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00