\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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.

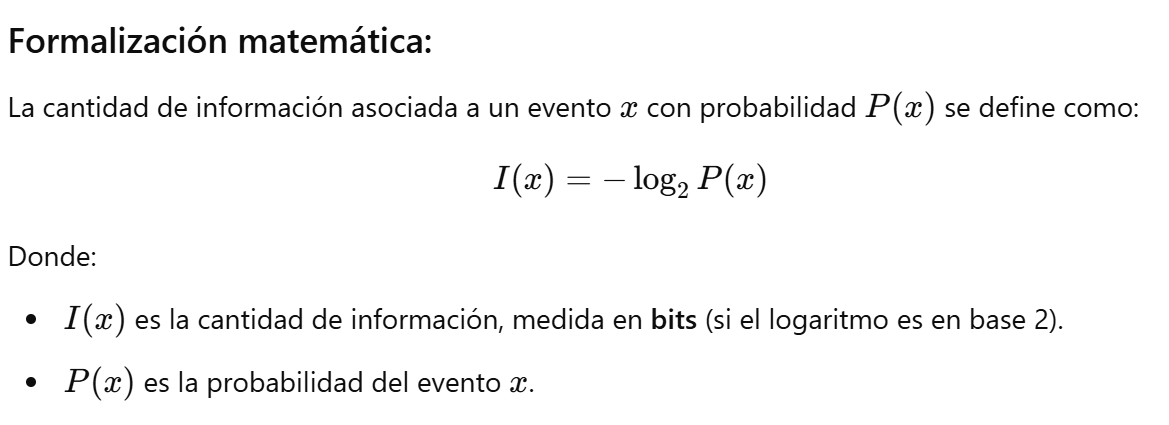
Cuando generamos una comprobamos si es “diferencial” con respecto a nuestro target. Recordad el EDA del Titanic..

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.*

**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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',

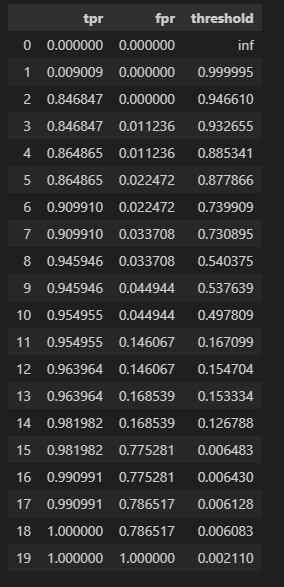
'month\_dau', 'week\_day', 'hour']]

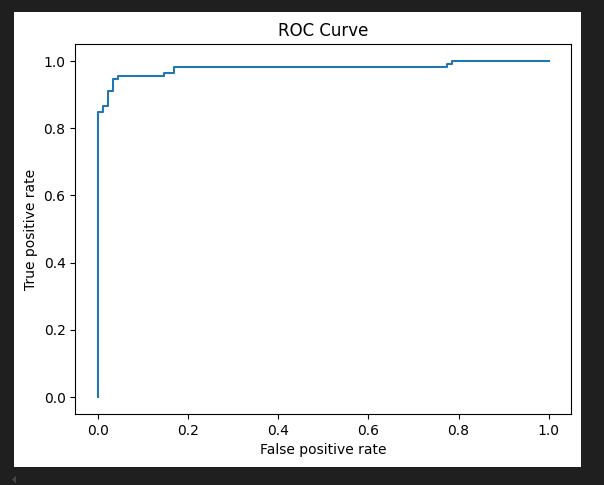
Conviene utilizar un StandardScaler. Escalas muy distintas entre las variables features.

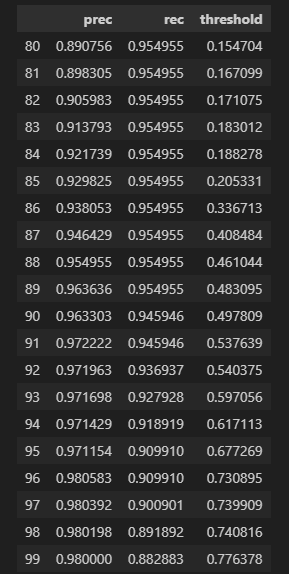
Crossvalidation: Train accuracy: 0.97 +/- 0.02

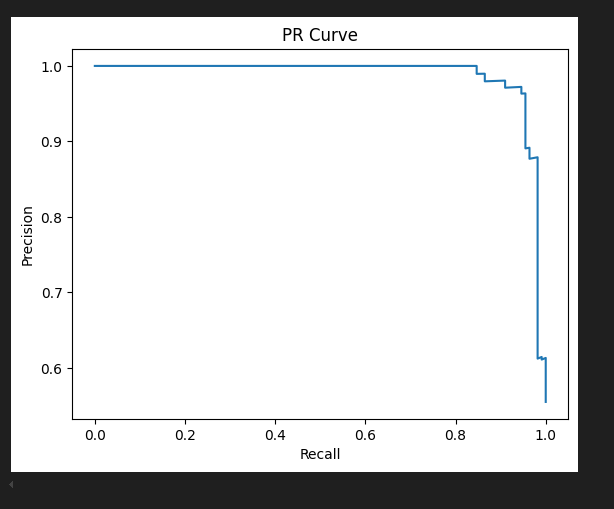
Test accuracy: 0.95

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









Confirmamos las hipótesis del feature selection:

