CUSTOMER SPENDING PREDICTION

Optimizando el Gasto de los Clientes

Data Science Team

11 March, 2024

INTRODUCCIÓN

La necesidad de prever y optimizar el gasto de sus usuarios ha llevado a una empresa de comercio electrónico a buscar soluciones innovadoras. Como científicos de datos, hemos sido convocados para desarrollar un modelo de machine learning que pueda predecir con precisión cuánto gastará un usuario al visitar dicho sitio web

FASES DEL PROYECTO

Preprocesamiento de Datos

Se realizó el análisis del dataset. Se trato columnas con diccionarios internos. Se generaron nuevas variables y se realizó la correcta limpieza de datos y tratamiento de valores faltantes.

EDA y Construcción Modelos

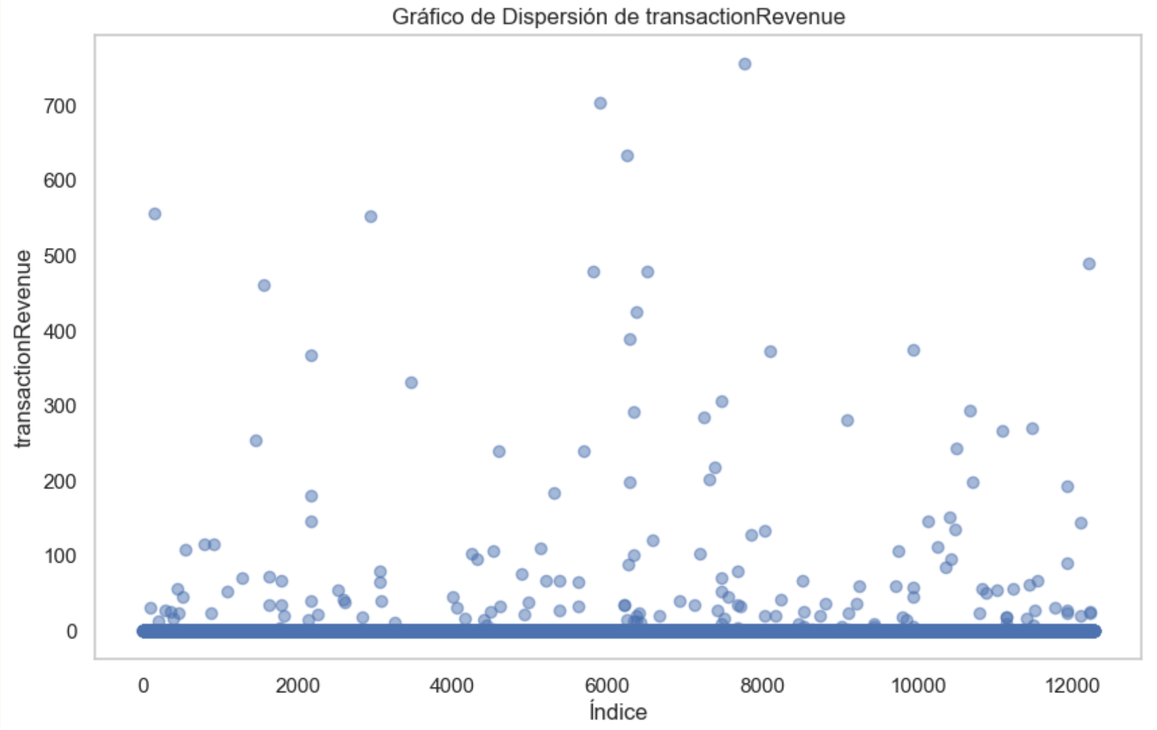
Se realizó la visualización de las variables. Se descartaron variables no relevantes para nuestro objetivo. Se estandarizaron nuestros datos y se generaron 4 modelos de regresión.

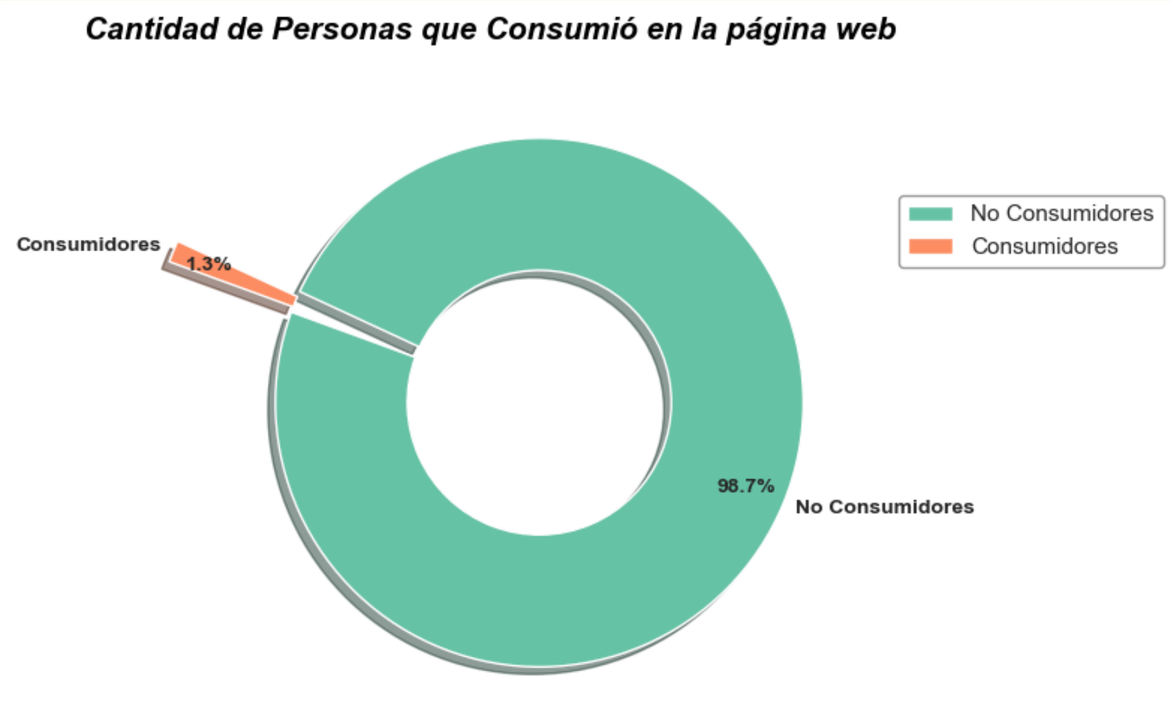
Evaluación y Selección del Modelo

Se evaluaron varias métricas de todos los modelos como lo son el RMSE, MAE Y R2.. Asimismo, se realizaron otras técnicas de evaluación como lo son la Visualización de Residuos y La Binarización de los valores reales y predichos.

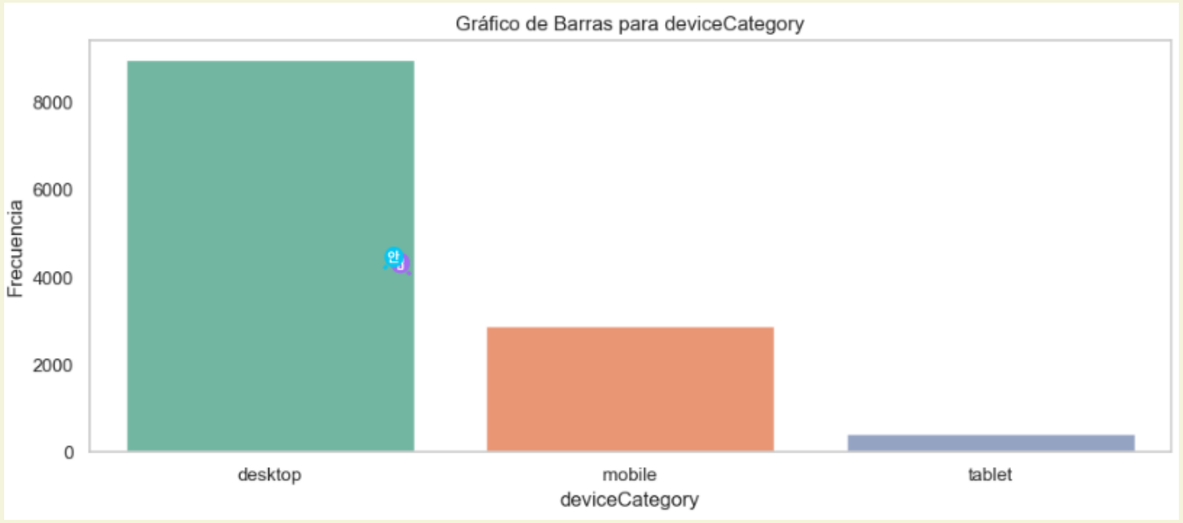
VARIABLE OBJETIVO

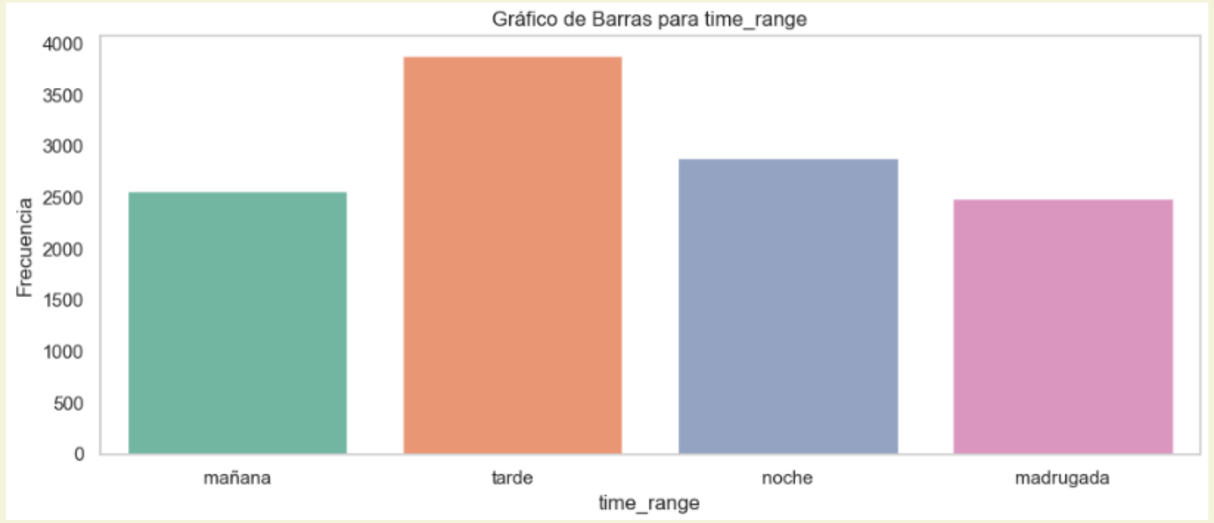
- Observamos que la cantidad de personas que consumieron al visitar la página web es mínima comparado con las personas que no consumieron nada. Esto nos indica que nuestra data está muy desbalanceada.
- En la gráfica de dispersión, notamos que casi todos nuestros datos se sitúan cerca al valor cero.
- La mayor cantidad de personas que visitan la página web no concretan la compra.
- Las personas que realizan un gasto elevado, es mínimo.

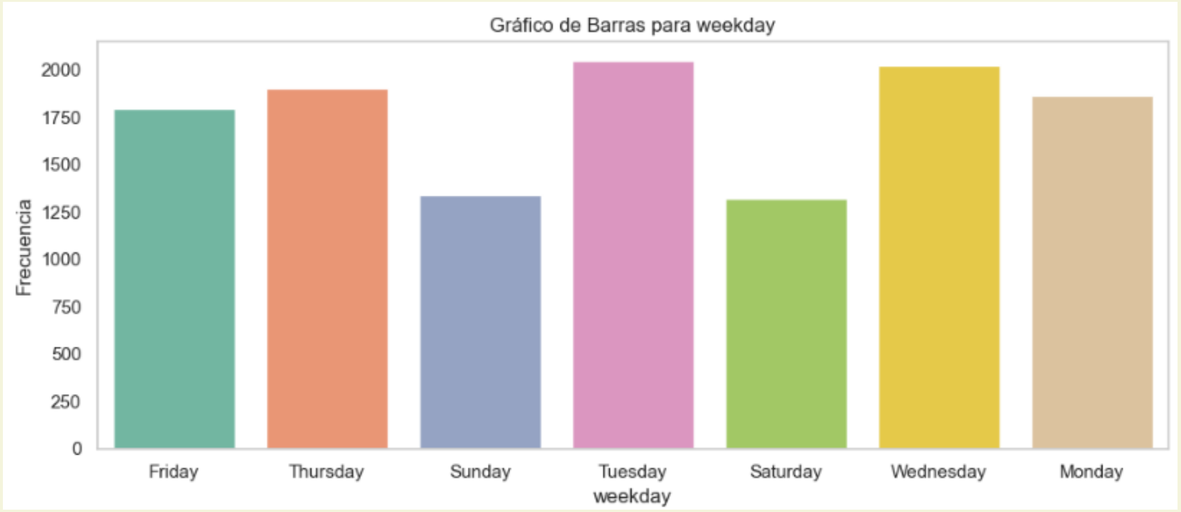


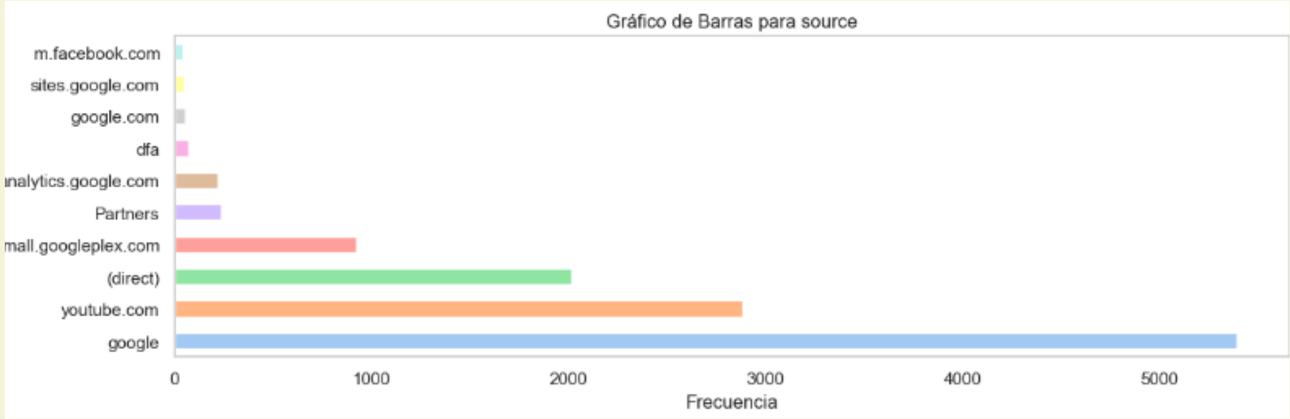


VISUALIZACIONES

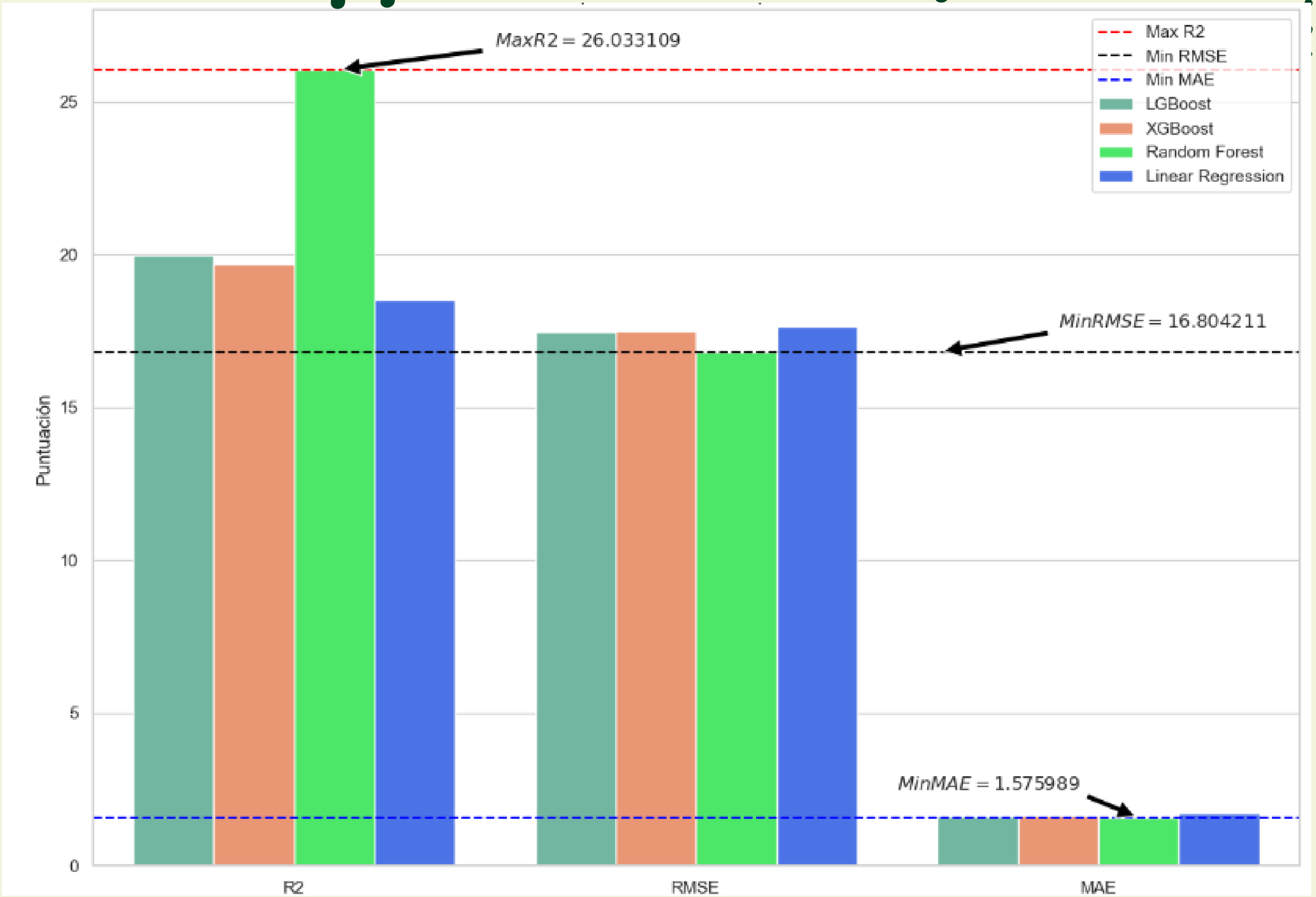






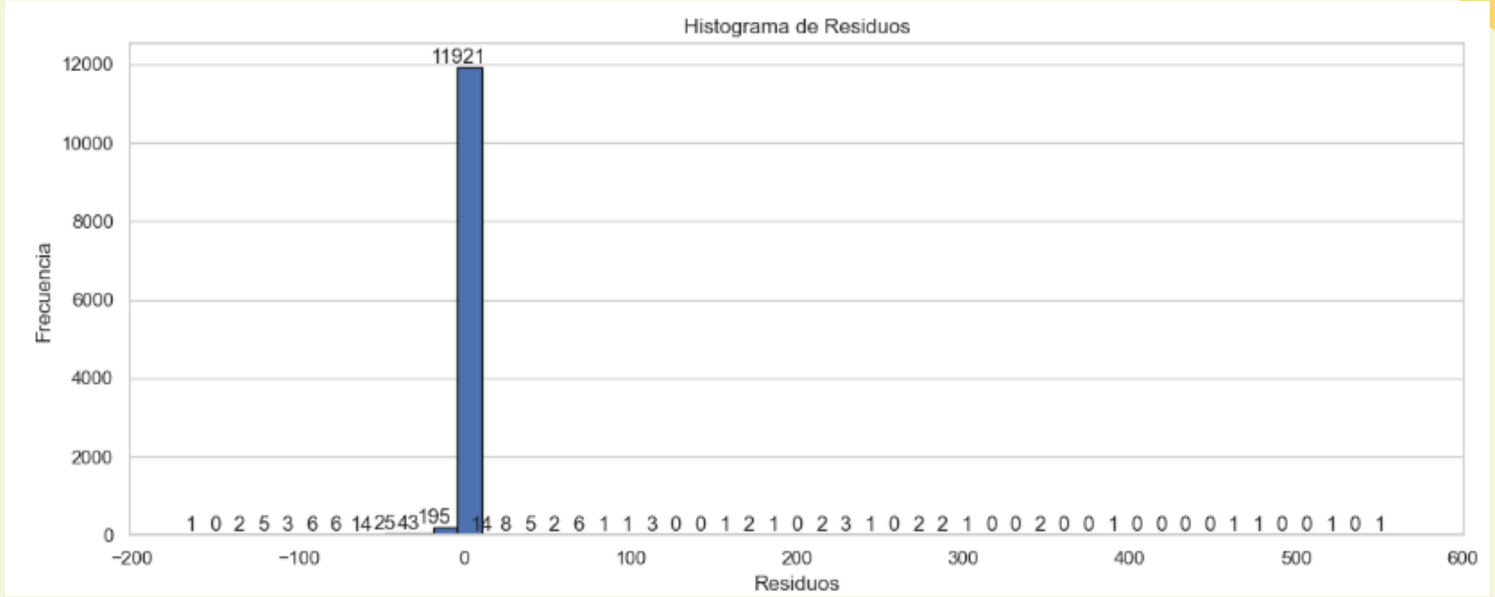


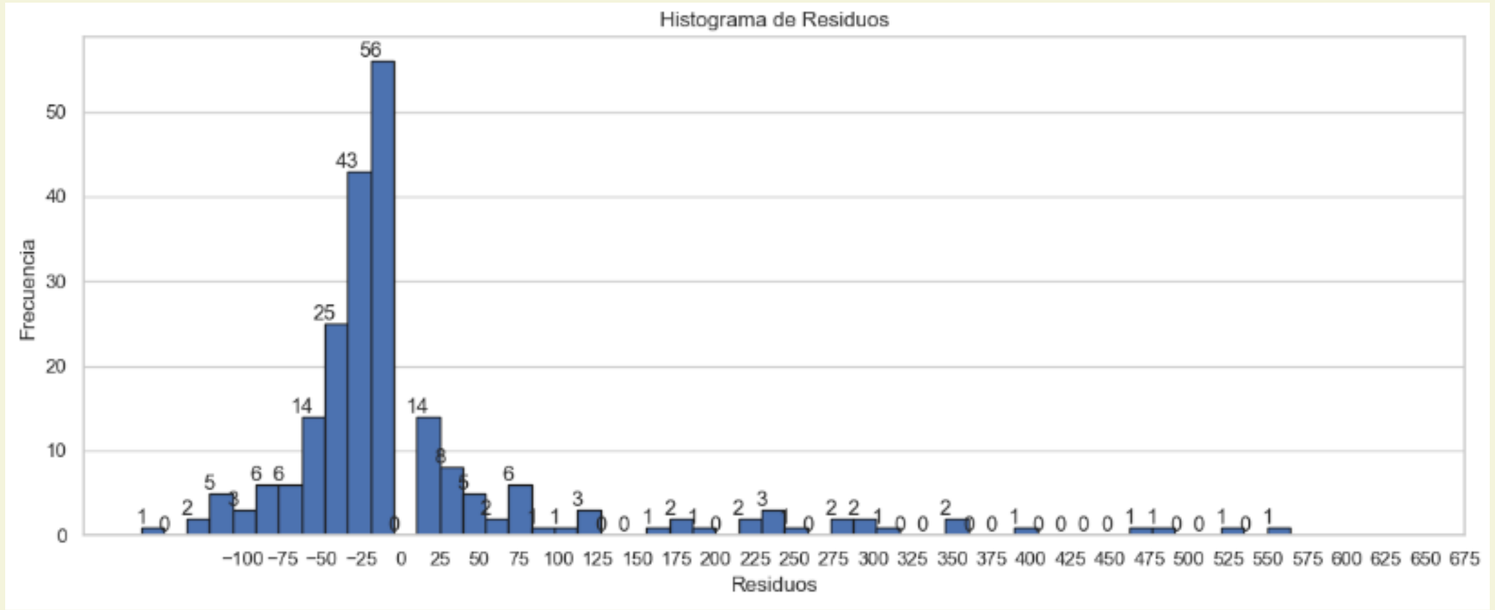
MODELO ELEGIDO



PUESTA EN PRODUCCIÓN

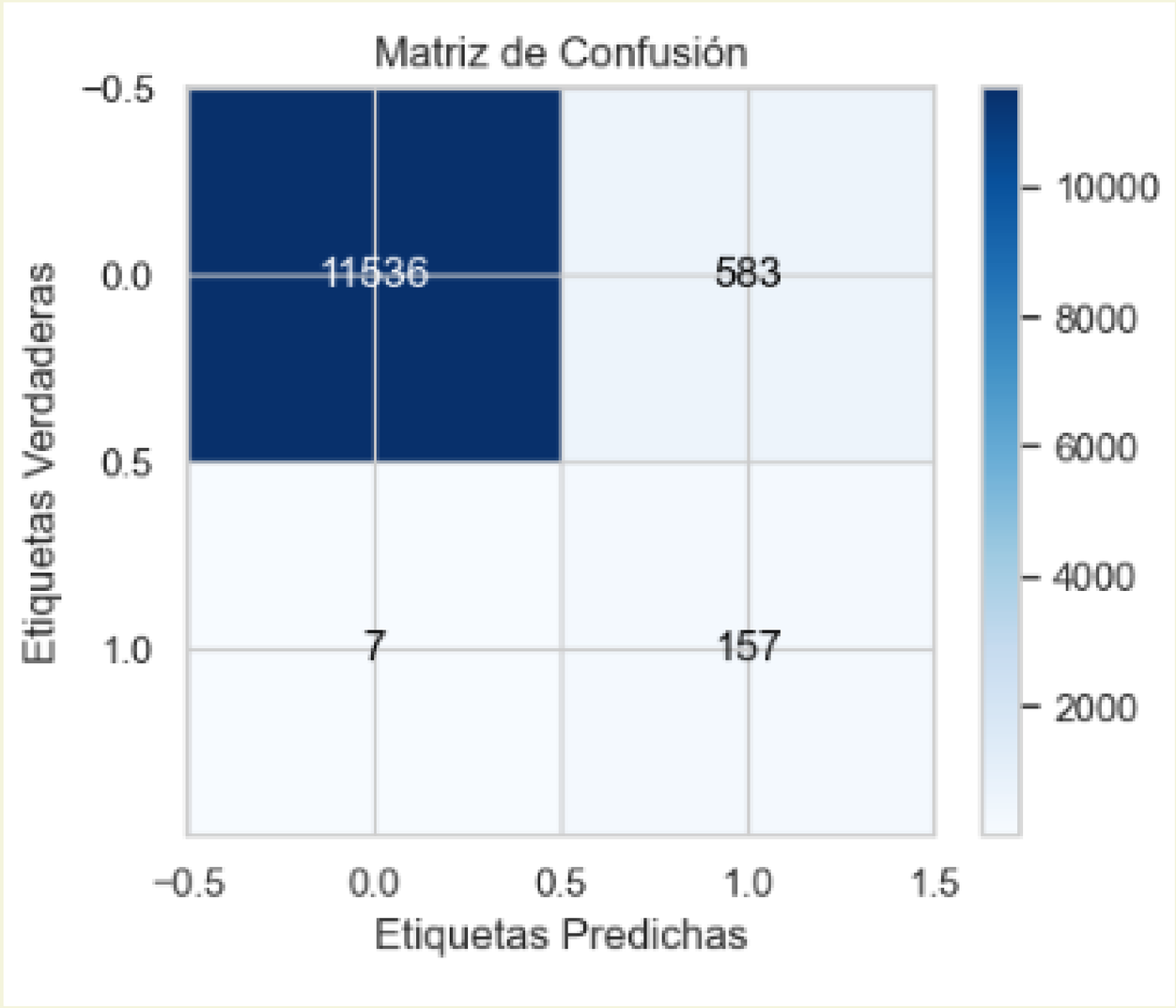
	transactionRevenue	predictions
11154	17.98	44.977159
2599	37.80	51.424945
448	55.96	86.949956
4410	14.48	1.411975
2168	39.68	77.835117
2838	18.99	47.411696
280	27.19	49.326383
7563	45.57	61.930408
10714	197.70	127.494114
4321	95.18	100.517935
10496	243.16	165.078359
7250	284.34	166.428028





PUESTA EN PRODUCCIÓN

```
Accuracy: 0.951966132052430 2
Precision: 0.2121621621621621 6
Recall: 0.957317073170731 7
ROC AUC: 0.954605396887370 9
```



PUESTA EN PRODUCCIÓN

10 a 40

	transactionRevenue	predictions	difference
4410	14.48	1.411975	13.068025
6253	15.19	1.916542	13.273458
4162	15.99	2.636140	13.353860
2599	37.80	51.424945	13.624945
11413	16.99	1.707270	15.282730
11936	23.97	40.097687	16.127687
4614	32.49	49.267862	16.777862
4486	24.73	46.694739	21.964739
280	27.19	49.326383	22.136383
1636	34.91	57.700760	22.790760
8028	19.75	42.583577	22.833577
6408	23.49	0.000000	23.490000
6663	19.95	43.781225	23.831225
11768	31.49	57.552992	26.062992
1785	33.59	6.599298	26.990702
11154	17.98	44.977159	26.997159
8738	20.78	48.378375	27.598375
2838	18.99	47.411696	28.421696
5629	32.94	4.412372	28.527628
3069	39.99	69.290032	29.300032
7679	34.99	65.281607	30.291607
7117	34.38	65.818693	31.438693
93	31.49	65.172405	33.682405
3246	10.63	44.467881	33.837881
4972	38.38	3.556969	34.823031
8821	36.78	0.000000	36.780000
9218	35.98	73.369134	37.389134
2168	39.68	77.835117	38.155117

40 a 100

	transactionRevenue	predictions	difference
9550	44.98	45.283292	0.303292
10843	55.99	53.323909	2.666091
5625	64.78	60.884402	3.895598
11234	55.99	60.134792	4.144792
2511	53.67	58.629736	4.959736
3995	44.79	49.941054	5.151054
4321	95.18	100.517935	5.337935
11035	53.97	48.144856	5.825144
7678	79.19	88.560831	9.370831
11535	89.99	80.264721	9.725279
11557	66.98	57.188601	9.771399
1635	71.85	61.981498	9.868502
9254	59.98	46.779602	13.200398
7476	69.99	84.260686	14.270686
4895	74.85	89.155893	14.305893
517	45.57	60.267345	14.697345
7563	45.57	61.930408	16.360408
5204	67.19	49.142497	18.047503
2595	41.72	63.086708	21.366708
6269	88.05	65.937881	22.112119
10434	95.80	73.239242	22.560758
1086	52.02	74.886773	22.866773
7473	52.49	79.291244	26.801244
11447	62.05	92.760952	30.710952
448	55.98	86.949956	30.969956
3058	64.94	103.464547	38.524547
10890	49.99	9.570614	40.419386

100 a 200

	transactionRevenue	predictions	difference
789	114.85	114.165593	0.484407
4525	105.48	104.687288	0.792712
10257	112.05	117.345276	5.295276
5135	109.57	119.080075	9.510075
7852	127.12	114.039381	13.080619
6573	119.99	106.664235	13.325765
4249	101.95	116.743119	14.793119
8032	132.58	115.749755	16.830245
6280	198.10	174.345644	23.754356
9752	105.49	77.781877	27.708123
924	115.88	144.297704	28.417704
10494	134.32	99.832855	34.487145
11537	192.87	154.122170	38.747830
6337	100.78	61.410026	39.369974
7189	102.35	146.339448	43.989448
10145	145.45	101.170779	44.279221
2165	145.99	96.961197	49.028803
2169	179.60	128.993584	50.606416
10714	197.70	127.494114	70.205886
12101	144.97	64.662835	80.307165
10416	151.80	66.556583	85.243417
557	107.94	205.035069	97.095069
5304	184.05	3.248536	180.801464

200 a 899

	transactionRevenue	predictions	difference
7316	200.89	191.077473	9.812527
7383	217.44	194.045312	23.394688
10496	243.16	165.078359	78.081641
1462	253.77	143.538403	110.231597
7250	284.34	166.428028	117.911972
11091	267.20	148.758840	118.441160
9091	279.97	160.943994	119.026006
4602	239.88	80.313958	159.566042
11473	269.55	87.609449	181.940551
6334	291.07	97.512312	193.557688
6278	389.12	174.317860	214.802140
2173	367.04	147.207776	219.832224
8102	372.65	143.150376	229.499624
3469	330.51	93.603098	236.906902
5687	239.60	0.000000	239.600000
10685	293.53	41.416128	252.113872
9951	374.85	97.300303	277.549697
6371	424.50	143.757094	280.742906
6516	478.40	184.014316	294.385684
7478	306.72	11.784837	294.935163
12220	489.20	184.787742	304.412258
1569	460.98	112.915200	348.064800
2945	552.50	192.548127	359.951873
154	556.61	154.676653	401.933347
5822	479.03	15.605872	463.424128

CONCLUSIONES

- Se construyeron 4 modelos de regresión: LGBoost, XGBoost, Random Forest y Linear Regression.
- Las métricas principales analizadas fueron R2, RMSE y MAE.
- Random Forest tuvo los valores más bajos en RMSE y MAE.
- Visualización de residuos mostró que Random Forest y XGBoost tenían mejores respuestas.
- Binarización de valores reales y predichos: Random Forest tuvo mejor precisión para la clase 0, pero errores para la clase 1.
- Se eligió Random Forest como el modelo preferido.
- En la implementación, hubo dificultades para reducir los falsos positivos en la clase 1 debido al desbalance de datos.