Descuentos y artículos vendidos en MELI

Carmen Dení

Pasos para la formación del dataset

- Esta el módulo *data_experts/data_pipeline.py*
- Descargar data del endpoint https://api.mercadolibre.com/sites/MLM/search?q={query}
- Query varía sobre cadenas de artículos electrónicos

```
{
  "site_id": "MLM",
  "country_default_time_zone": "GMT-05:00",
  "query": "{query}",
  "paging": - {
    "total": 164,
    "primary_results": 164,
    "offset": 0,
    "limit": 50
  },
  "results": - [
    - {
      "id": "MLM1555606695",
      "title": "Querybaya G (extracto De Guayaba) C/60 Tabs Naturex",
      "condition": "new",
      "thumbnail_id": "661967-MLM52048392060_102022",
      "catalog_product_id": null,
      "listing_type_id": "gold_special",
      "sanitized_title": "querybaya-g-extracto-de-guayaba-c60-tabs-naturex",
      "permalink": "https://articulo.mercadolibre.com.mx/MLM-1555606695-querybaya-g-extracto-de-guayaba-c60-tabs-naturex-_JM",
      "buying_mode": "buy_it_now",
      "site_id": "MLM",
      "category_id": "MLM438195",
      "domain_id": "MLM-SUPPLEMENTS",
      "thumbnail": "http://http2.mlstatic.com/D_661967-MLM52048392060_102022-I.jpg",
      "currency_id": "MXN",
      "order_backend": 1,
      "price": 125,
      "original_price": null,
      "sale_price": - {
        "price_id": "",
        "amount": 125,
        "conditions": - {
          "eligible": true,
          "context_restrictions": [
          ],
          "start_time": null,
          "end_time": null
        },
        "currency_id": "MXN",
        "exchange_rate": null,
        "payment_method_prices": [
```

sold_quantity no esta!! 😱 😱 😱

Pero available quantity si!! 😏

```
      },
      "available_quantity": 1,
      "official_store_id": 5170,
      "official_store_name": "Farmacias BS",
      "use_thumbnail_id": true,
      "accepts_mercadopago": true,
      "variation_filters": - [
        "FLAVOR"
```

- Buscamos initial_quantity en el endpoint https://api.mercadolibre.com/items/{item_id}

- Obtenemos sold_quantity = initial_quantity – available_quantity

- Obtenemos el descuento como (original_price-prince)/original_Price

- Agregamos features como:

Categóricas

	count	unique	top	freq
buying_mode	1344	1	buy_it_now	1344
query	1344	40	teclado%20gaming	49
condition	1344	2	new	1342
brand	1344	414	Genérica	52
domain_id	1344	42	MLM-HEADPHONES	117
listing_type_id	1344	3	gold_special	774
is_official_store	1344	2	True	733
free_shipping	1344	2	True	1108
store_pick_up	1344	1	False	1344

Numéricas

|	min	mean	max
price	45.54	4482.535335	103999.00
initial_quantity	1.00	2202.347470	168400.00
available_quantity	1.00	113.807292	5000.00
sold_quantity	0.00	2088.540179	163400.00
installment_quantity	1.00	17.294643	24.00
installment_rate	0.00	25.061339	45.03
days_until_discount_end	0.00	2.651042	32.00
msi	1.00	4.494048	15.00
discount	0.00	19.673690	80.00

{
  "id": "MLM1555606695",
  "site_id": "MLM",
  "title": "Querybaya G (extracto De Guayaba) C/60 Tabs Naturex",
  "seller_id": 457895842,
  "category_id": "MLM438195",
  "official_store_id": 5170,
  "price": 125,
  "base_price": 125,
  "original_price": null,
  "currency_id": "MXN",
  "initial_quantity": 425,
  "sale_terms": [...],
  "buying_mode": "buy_it_now",
  "listing_type_id": "gold_special",
  "condition": "new",
  "permalink": "https://articulo.mercadolibre.com.mx/MLM-1555606695-querybaya-g-extracto-de-guayaba-c60-tabs-naturex-_JM",
  "thumbnail_id": "661967-MLM52048392060_102022",
  "thumbnail": "http://http2.mlstatic.com/D_661967-MLM52048392060_102022-I.jpg",
  "pictures": [...],
  "video_id": null,
  "descriptions": [
  ],
  "accepts_mercadopago": true,
  "non_mercado_pago_payment_methods": [
  ],
  "shipping": {...},
  "international_delivery_mode": "none",
  "seller_address": {...},
  "seller_contact": null,
  "location": {
  },
  "coverage_areas": [
  ],
  "attributes": [...],
  "listing_source": "",
  "variations": [...],
  "status": "active",
  "sub_status": [
  ],
  "tags": [...],
  "warranty": "Sin garantía",
  "catalog_product_id": null,
  "domain_id": "MLM-SUPPLEMENTS",
  "parent_item_id": null,
  "deal_ids": [...],
  "automatic_relist": false,

Aquí tenemos initial_quantity!! 🙌

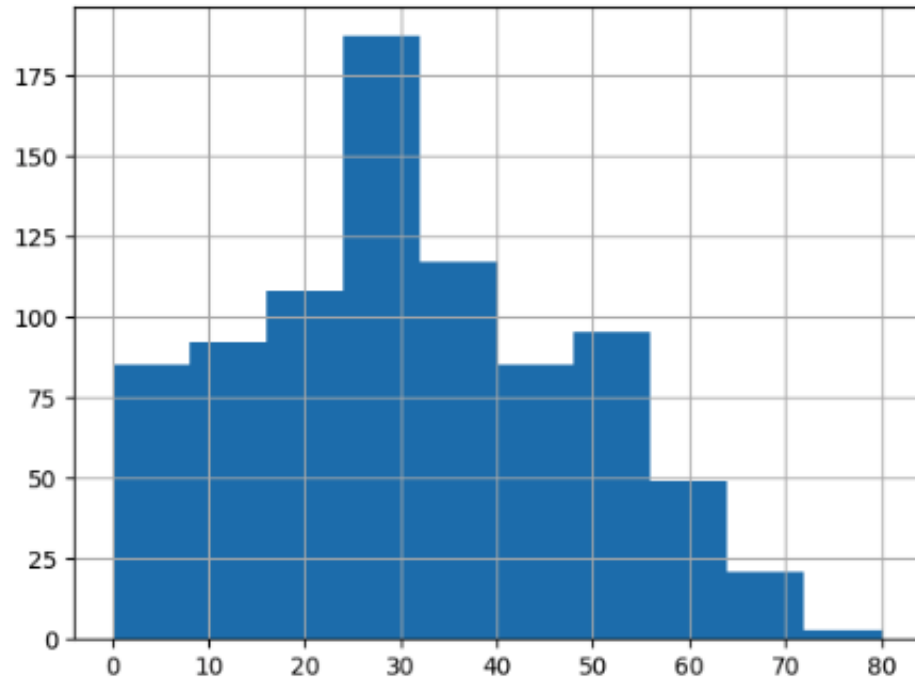# Relevant numerical_features
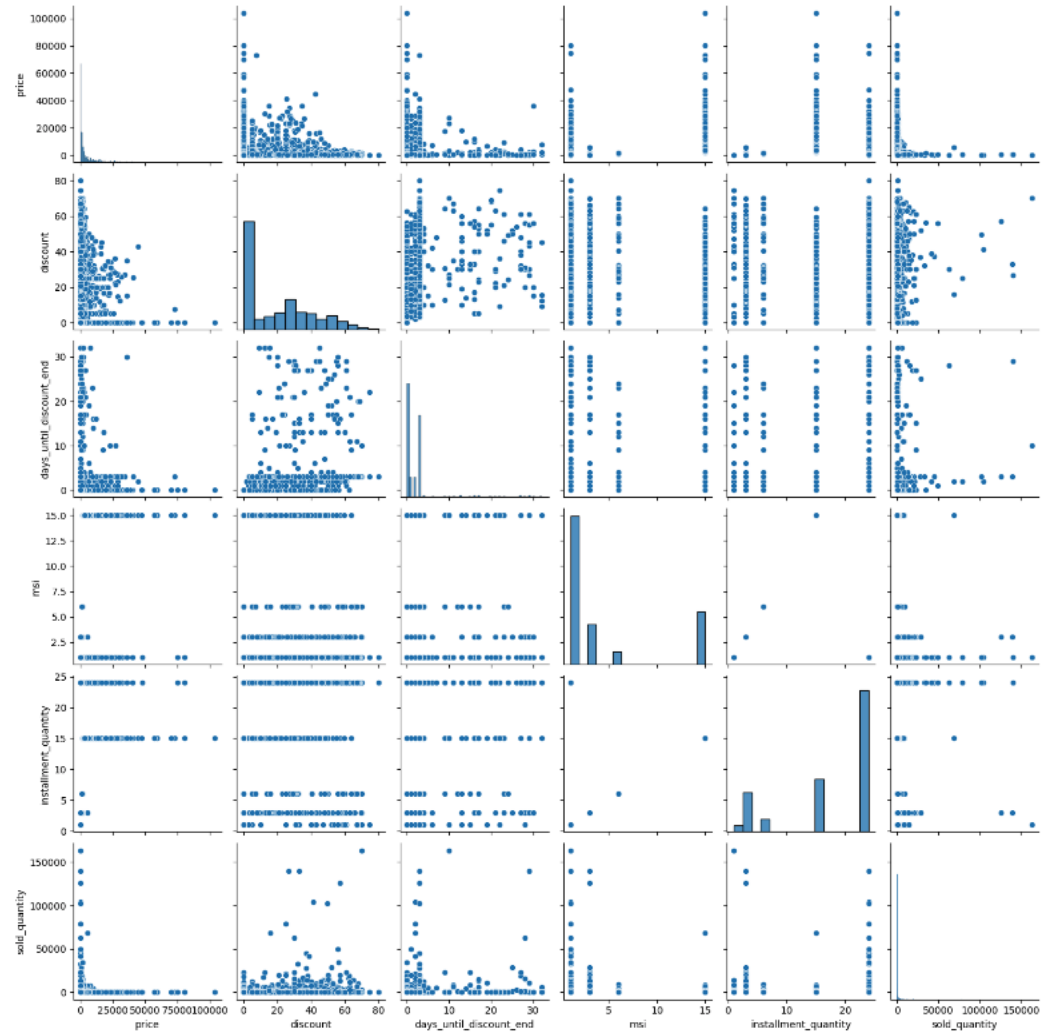
Relevant categorical features
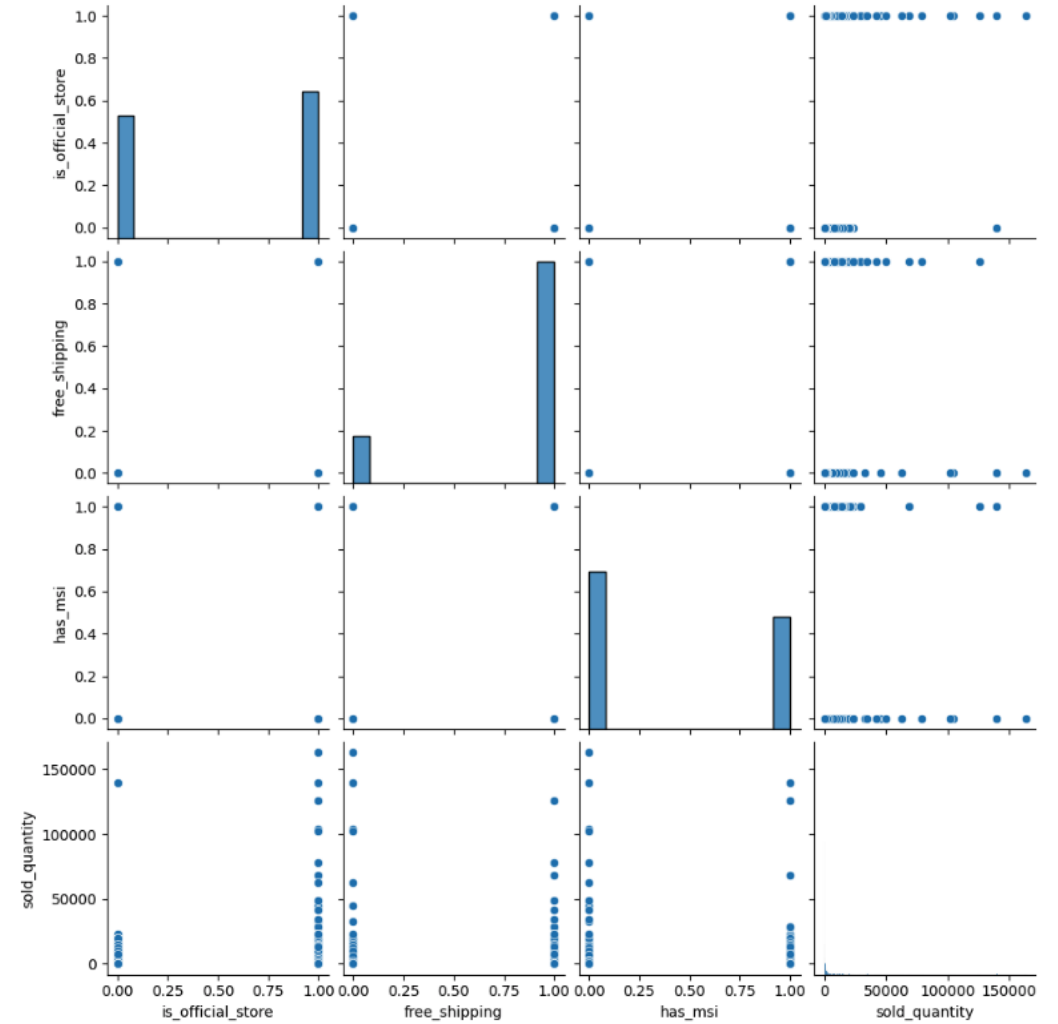
Algunos hayázgos acerca del descuento

Distribución total (quitando el descuento 0)

Categorías con más artículos con descuento

MLM-WRISTWATCHES	1.000000
MLM-TELEPHONES	1.000000
MLM-HOVERBOARDS	1.000000
MLM-HOME_ALARMS_AND_SENSORS	1.000000
MLM-TABLET_KEYBOARDS	1.000000
MLM-MOUSE_PADS	1.000000
MLM-DESKTOP_COMPUTER_KITS	0.952381
MLM-VEHICLE_WIRELESS_FM_TRANSMITTERS	0.875000
MLM-OFFICE_CHAIRS	0.866667
MLM-ROBOT_VACUUMS	0.850000
MLM-SMARTWATCHES	0.849057
MLM-COMPUTER_MICE	0.800000
MLM-SPEAKERS	0.787234
MLM-MOBILE_DEVICE_CHARGERS	0.757576
MLM-TABLETS	0.750000
MLM-COMPUTER_MONITORS	0.741176
MLM-HEADPHONES	0.735043
MLM-NOTEBOOKS	0.733333
MLM-ELECTRIC_SCOOTERS	0.714286
MLM-AUDIO_AND_VIDEO_CABLES_AND_ADAPTERS	0.692308
MLM-CELLPHONES	0.681818
MLM-LIGHT_BULBS	0.658537
MLM-VR_HEADSETS	0.631579

- Los descuentos más frecuentes son entre 20% y 40%
- El 63% de los artículos de electrónica tienen descuentos

Descuento promedio por Categoría

- Las categorías con descuentos mas grandes proporcionalmente son:
  - Mouse-pads
  - Teclados de Tablet
  - Transmisores FM
  - Relojes de muñeca
- Los artículos con mayor descuento son menos "modernos"

Categorías con descuento promedio mayor a 20%

domain_id	discount
MLM-MOUSE_PADS	49.002374
MLM-TABLET_KEYBOARDS	46.227430
MLM-VEHICLE_WIRELESS_FM_TRANSMITTERS	40.125281
MLM-WRISTWATCHES	39.500000
MLM-HOVERBOARDS	35.635611
MLM-DESKTOP_COMPUTER_KITS	34.571292
MLM-TABLETS	34.411089
MLM-SMARTWATCHES	32.813160
MLM-OFFICE_CHAIRS	31.936712
MLM-COMPUTER_MICE	31.140032
MLM-MOBILE_DEVICE_CHARGERS	27.652481
MLM-AUDIO_AND_VIDEO_CABLES_AND_ADAPTERS	24.926539
MLM-HEADPHONES	24.683916
MLM-ROBOT_VACUUMS	24.454760
MLM-SURVEILLANCE_CAMERAS	24.296727
MLM-ELECTRIC_SCOOTERS	24.143747
MLM-SPEAKERS	22.365876
MLM-LIGHT_BULBS	22.050601
MLM-COMPUTER_MONITORS	21.446463

¿Los artículos a msi son más caros? (intereses escondidos)

- Si pasa para el 40% de las categorías

has_msi	price		msi_more_expensive?	diff
domain_id	False	True		
MLM-DIGITAL_CAMERAS	23146.430769	28901.788049	True	5755.35728
MLM-DESKTOP_COMPUTER_KITS	10578.633333	15040.833333	True	4462.2
MLM-PROJECTORS	977.715556	5392.808824	True	4415.093268
MLM-GAME_CONSOLES	5859.803548	9984.724375	True	4124.920827
MLM-WRISTWATCHES	229.54	2999.25	True	2769.71
MLM-COMPUTER_MONITORS	3745.430545	6334.083333	True	2588.652788
MLM-CELLPHONES	2460.0305	4864.287083	True	2404.256583
MLM-SPEAKERS	1168.015455	3092.736429	True	1924.720974
MLM-TELEVISIONS	6829.195	8502.503143	True	1673.308143
MLM-TABLET_KEYBOARDS	279.99	699.65	True	419.66
MLM-ROBOT_VACUUMS	3959.859167	4316.581786	True	356.722619
MLM-PC_KEYBOARDS	914.849615	1264.39	True	349.540385
MLM-MOBILE_DEVICE_CHARGERS	309.413585	539.896923	True	230.483338
MLM-LIGHT_BULBS	218.875135	301.62	True	82.744865
MLM-WIRELESS_ANTENNAS_AND_ADAPTERS	146.0	192.468333	True	46.468333
MLM-HEADPHONES	475.777753	515.8675	True	40.089747
MLM-SURVEILLANCE_CAMERAS	450.504444	481.782353	True	31.277908

## ¿Los artículos a msi son más caros? (intereses escondidos)

- Si pasa para el 40% de las categorías

	price		msi_more_expensive?	diff
has_msi	False	True		
domain_id				
MLM-DIGITAL_CAMERAS	23146.430769	28901.788049	True	5755.35728
MLM-DESKTOP_COMPUTER_KITS	10578.633333	15040.833333	True	4462.2
MLM-PROJECTORS	977.715556	5392.808824	True	4415.093268
MLM-GAME_CONSOLES	5859.803548	9984.724375	True	4124.920827
MLM-WRISTWATCHES	229.54	2999.25	True	2769.71
MLM-COMPUTER_MONITORS	3745.430545	6334.083333	True	2588.652788
MLM-CELLPHONES	2460.0305	4864.287083	True	2404.256583
MLM-SPEAKERS	1168.015455	3092.736429	True	1924.720974
MLM-TELEVISIONS	6829.195	8502.503143	True	1673.308143
MLM-TABLET_KEYBOARDS	279.99	699.65	True	419.66
MLM-ROBOT_VACUUMS	3959.859167	4316.581786	True	356.722619
MLM-PC_KEYBOARDS	914.849615	1264.39	True	349.540385
MLM-MOBILE_DEVICE_CHARGERS	309.413585	539.896923	True	230.483338
MLM-LIGHT_BULBS	218.875135	301.62	True	82.744865
MLM-WIRELESS_ANTENNAS_AND_ADAPTERS	146.0	192.468333	True	46.468333
MLM-HEADPHONES	475.777753	515.8675	True	40.089747
MLM-SURVEILLANCE_CAMERAS	450.504444	481.782353	True	31.277908

Entrenamiento

- Target sold_quantity
- 3 experimentos:
  - Todos los features (E0)
  - Quitando initial_quantity (E1)
  - Quitando initial_quantity y available_quantity (E2)
- Modelos a probar:
  - RandomForestRegressor
  - XGBMRegressor
  - LGBMRegressor
  - Neural Networks (Keras-TensorFlow)
- Búsqueda de hiperparametros:
  - GridSearchCV
  - RandomizedSearchCV
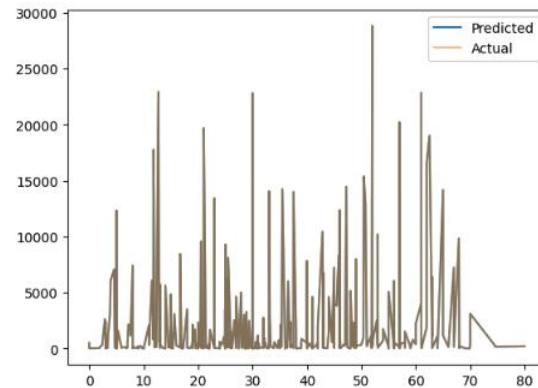 - Bayesian Optimization
- Score: R2 (más alto, mejor)
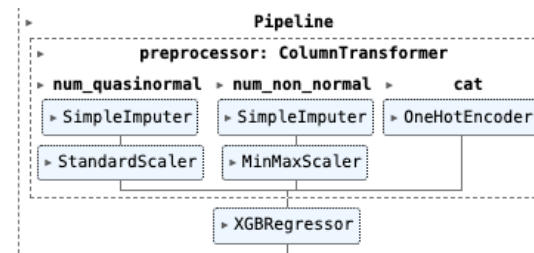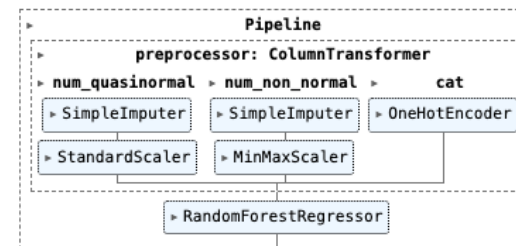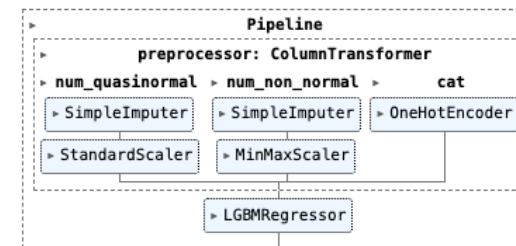
$$R2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})}, \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Mejor modelo para cada caso (E0)

- XGBoost
- Train r2_score= 1
- Test r2_score = 0.98
- Grid Search

Proyección para "discount": train set

Proyección para "discount": test set
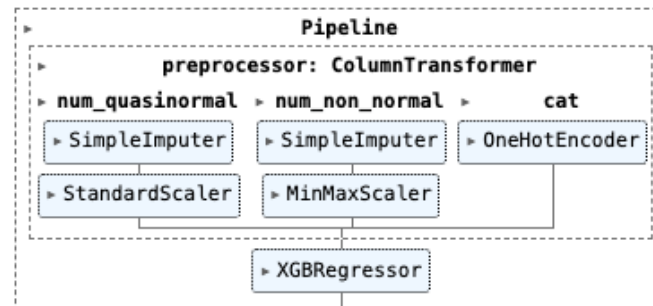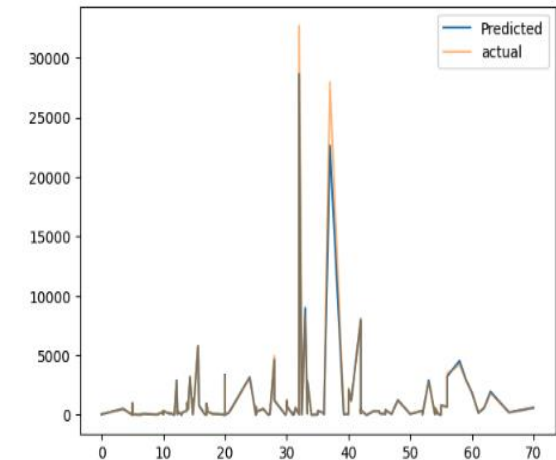
{'learning_rate': 0.055245405728306586,
 'max_depth': 5,
 'min_child_weight': 1,
 'n_estimators': 360}

Random Forest Feature importance

Top 10 Feature Importances
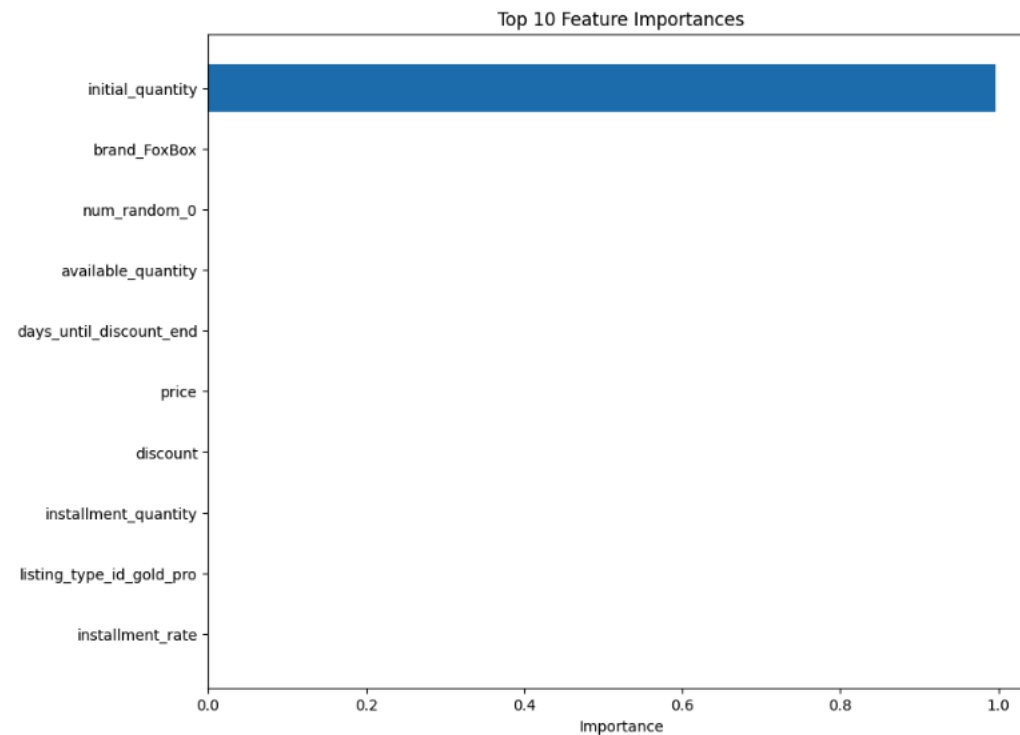
Feature	Importance
initial_quantity	1.00
brand_FoxBox	0.00
num_random_0	0.00
available_quantity	0.00
days_until_discount_end	0.00
price	0.00
discount	0.00
installment_quantity	0.00
listing_type_id_gold_pro	0.00
installment_rate	0.00
cat_random_0_C	0.00
listing_type_id_gold_special	0.00
is_official_store_True	0.00
query_reloj%20inteligente	0.00
domain_id_MLM-SMARTWATCHES	0.00
cat_random_0_A	0.00
domain_id_MLM-MOBILE_DEVICE_CHARGERS	0.00
cat_random_0_B	0.00
free_shipping_True	0.00
msi	0.00
is_official_store_False	0.00
query_bocina%20bluetooth	0.00

SHAP Feature importance

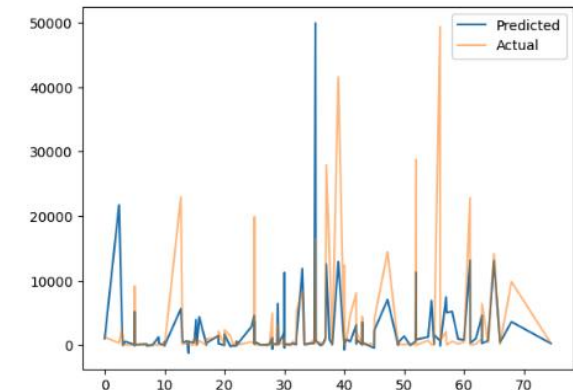Podemos ver que XGBoost aprovechó mejor
La info en available_quantity
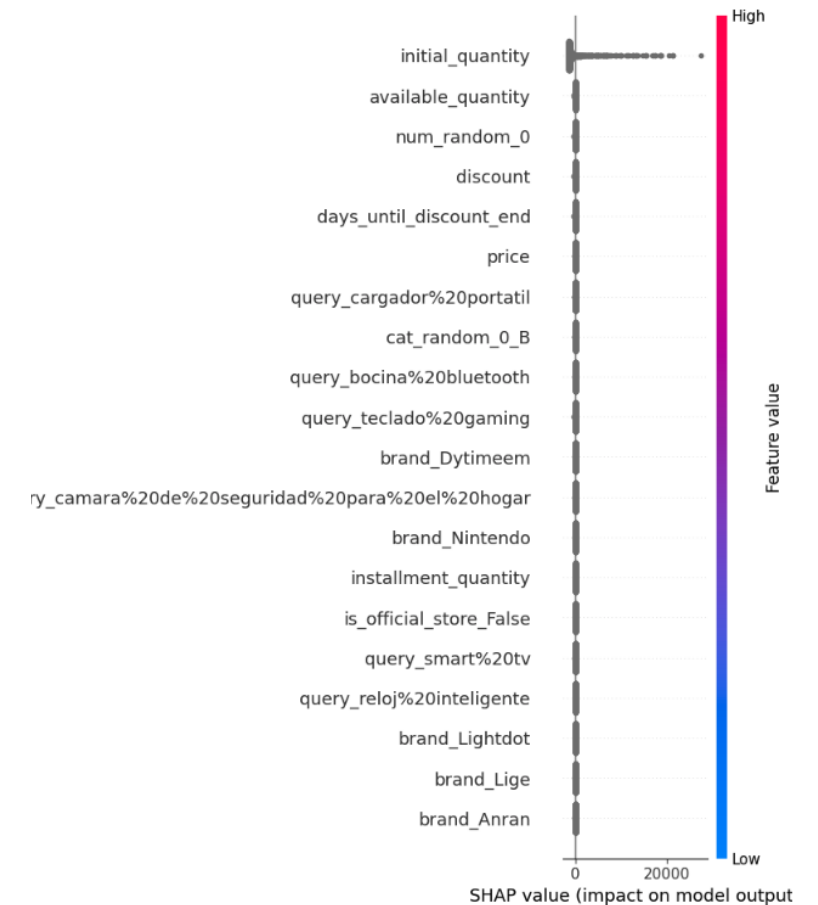
Mejor modelo para cada caso (E1)

- XGBoost
- Train r2_score= 0.82
- Test r2_score = 0.26
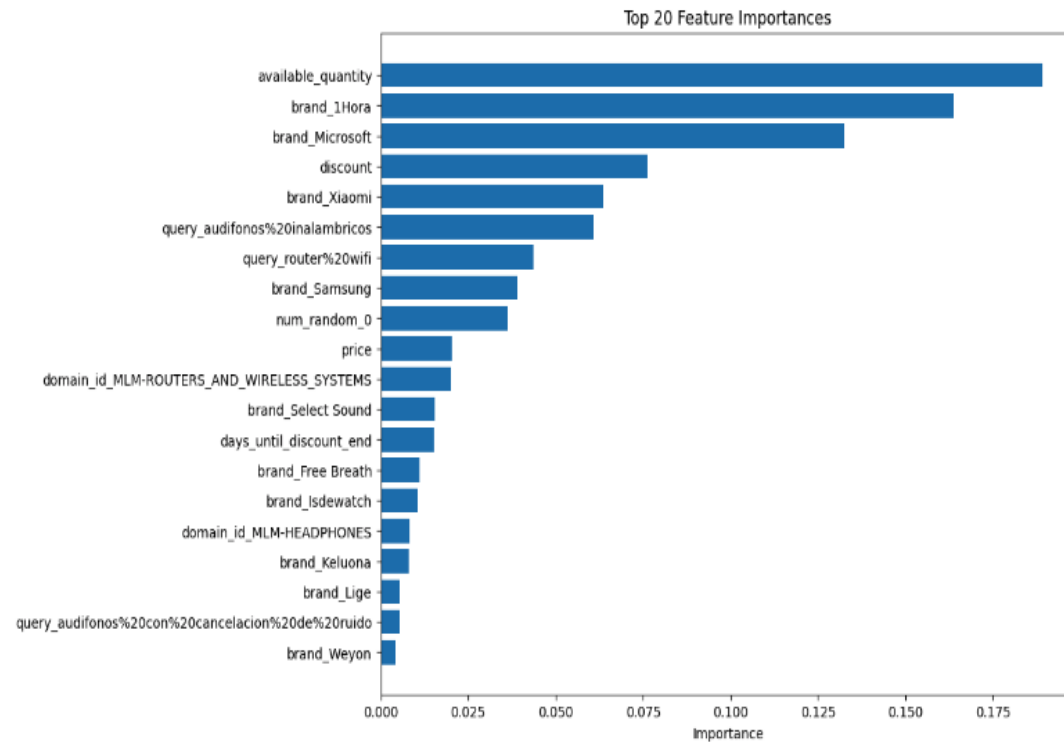- Bayesian Optimization

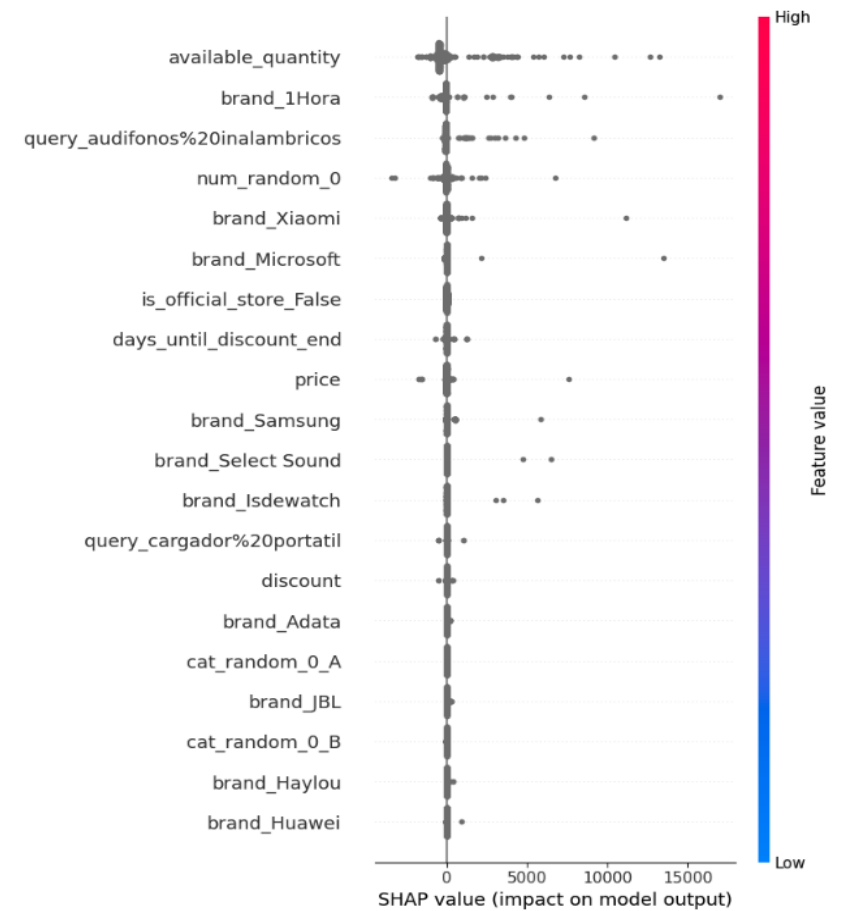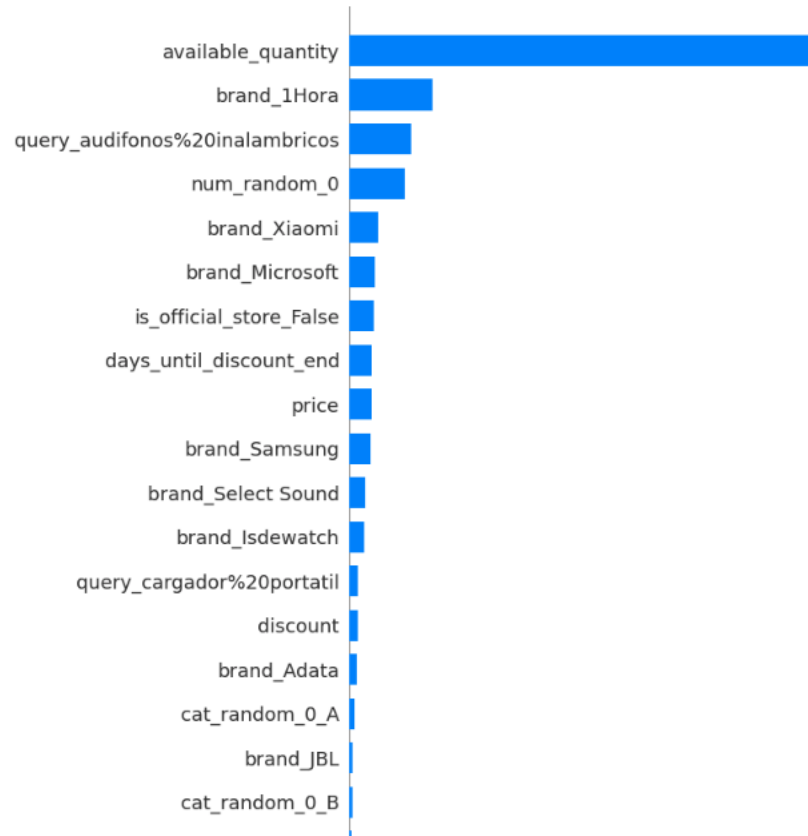Proyección para "discount": train set
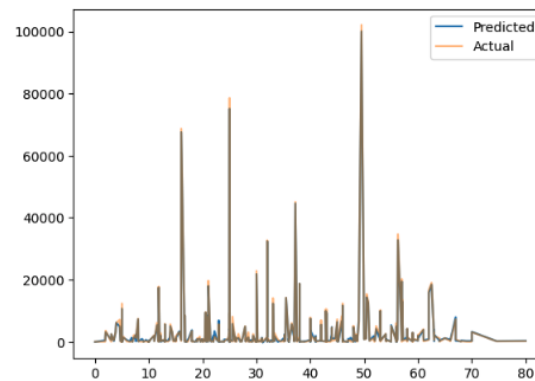
Proyección para "discount": test set
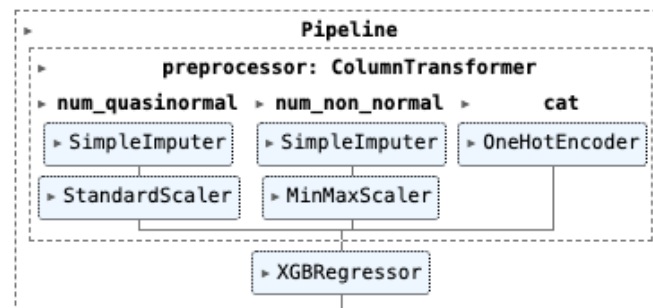
{'learning_rate': 0.01711241830402892,
 'max_depth': 9,
 'min_child_weight': 5,
 'n_estimators': 258}

Random Forest Feature importance

Top 20 Feature Importances

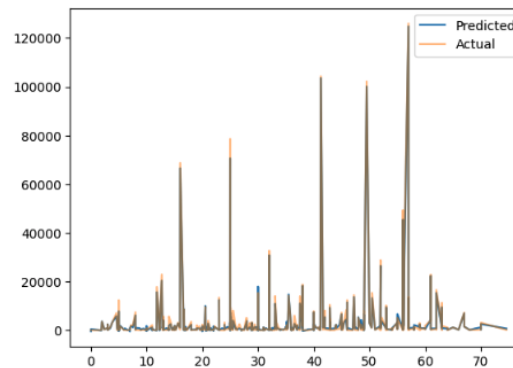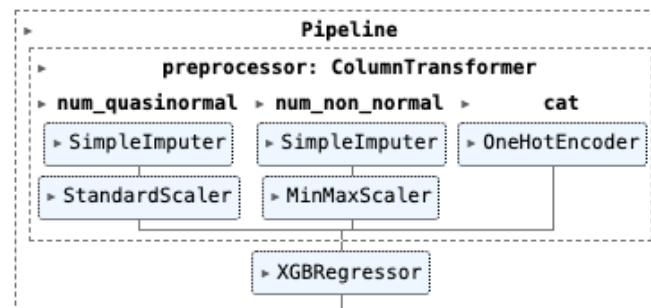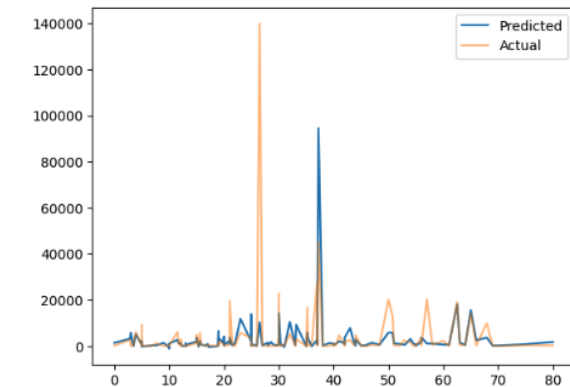| Feature	Importance
available_quantity	0.19
brand_1Hora	0.16
brand_Microsoft	0.13
discount	0.08
brand_Xiaomi	0.06
query_audifonos%20inalambricos	0.06
query_router%20wifi	0.04
brand_Samsung	0.04
num_random_0	0.04
price	0.02
domain_id_MLM-ROUTERS_AND_WIRELESS_SYSTEMS	0.02
brand_Select Sound	0.02
days_until_discount_end	0.02
brand_Free Breath	0.01
brand_Isdewatch	0.01
domain_id_MLM-HEADPHONES	0.01
brand_Keluona	0.01
brand_Lige	0.01
query_audifonos%20con%20cancelacion%20de%20ruido	0.01
brand_Weyon	0.00
brand_Huawei	0.00

SHAP Feature importance

Mejor modelo para cada caso (E2)

- XGBoost
- Train r2_score= 0.68
- Test r2_score = 0.12
- Bayesian Optimization
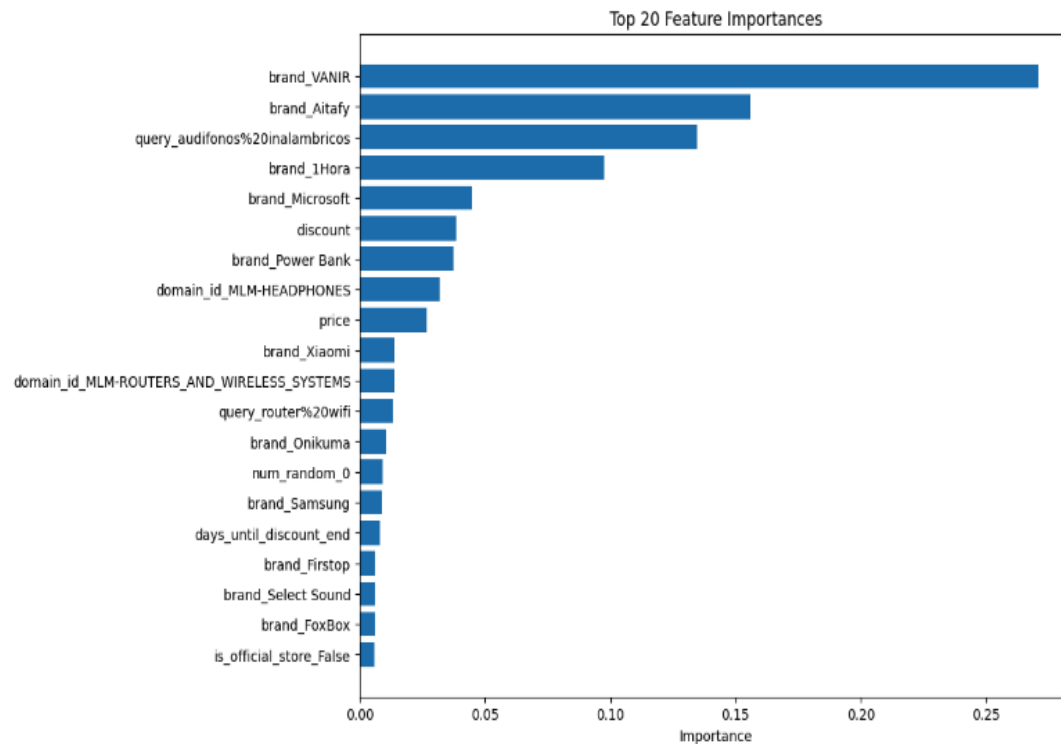
Proyección para "discount": train set

Proyección para "discount": test set


```
{'learning_rate': 0.01,
 'max_depth': 10,
 'min_child_weight': 5,
 'n_estimators': 283}
```

Random Forest Feature importance

Top 20 Feature Importances

| Feature	Importance
brand_VANIR	0.27
brand_Aitafy	0.16
query_audifonos%20inalambricos	0.13
brand_1Hora	0.10
brand_Microsoft	0.04
discount	0.04
brand_Power Bank	0.04
domain_id_MLM-HEADPHONES	0.03
price	0.03
brand_Xiaomi	0.01
domain_id_MLM-ROUTERS_AND_WIRELESS_SYSTEMS	0.01
query_router%20wifi	0.01
brand_Onikuma	0.01
num_random_0	0.01
brand_Samsung	0.01
days_until_discount_end	0.01
brand_Firstop	0.01
brand_Select Sound	0.01
brand_FoxBox	0.01
is_official_store_False	0.01
brand_Nanwei	0.01
is_official_store_True	0.00

SHAP Feature importance
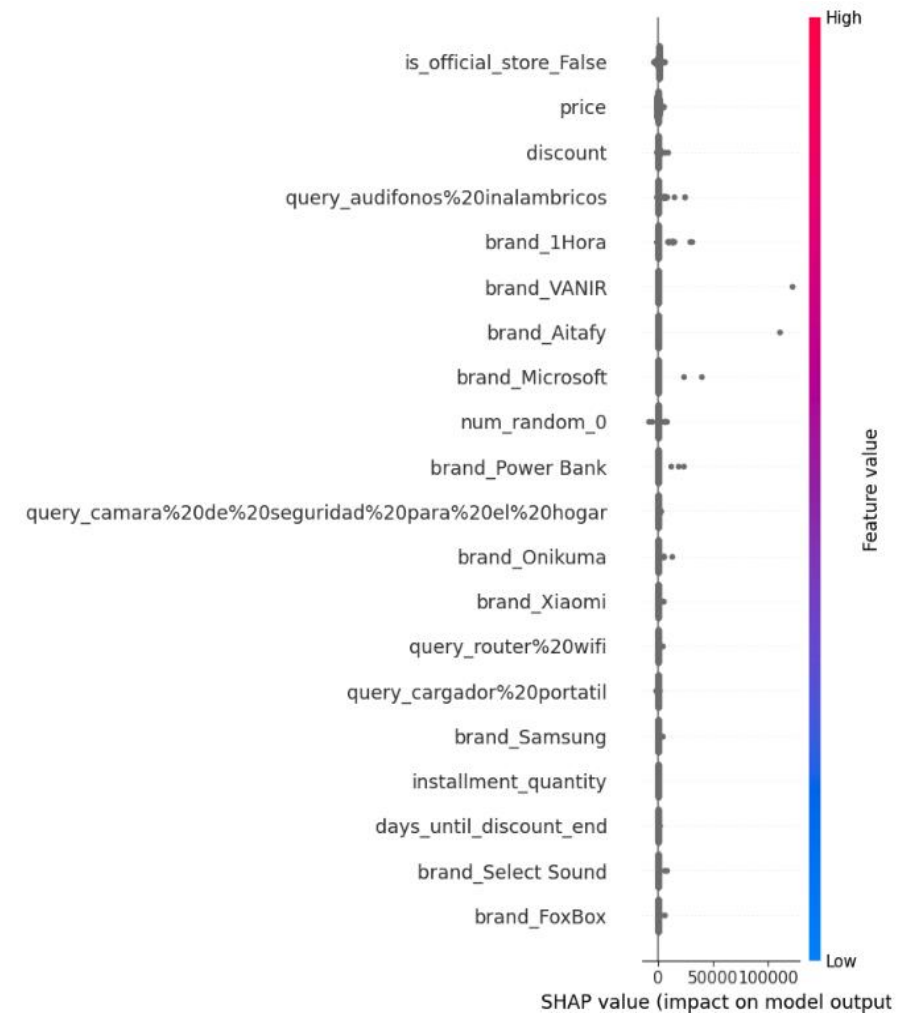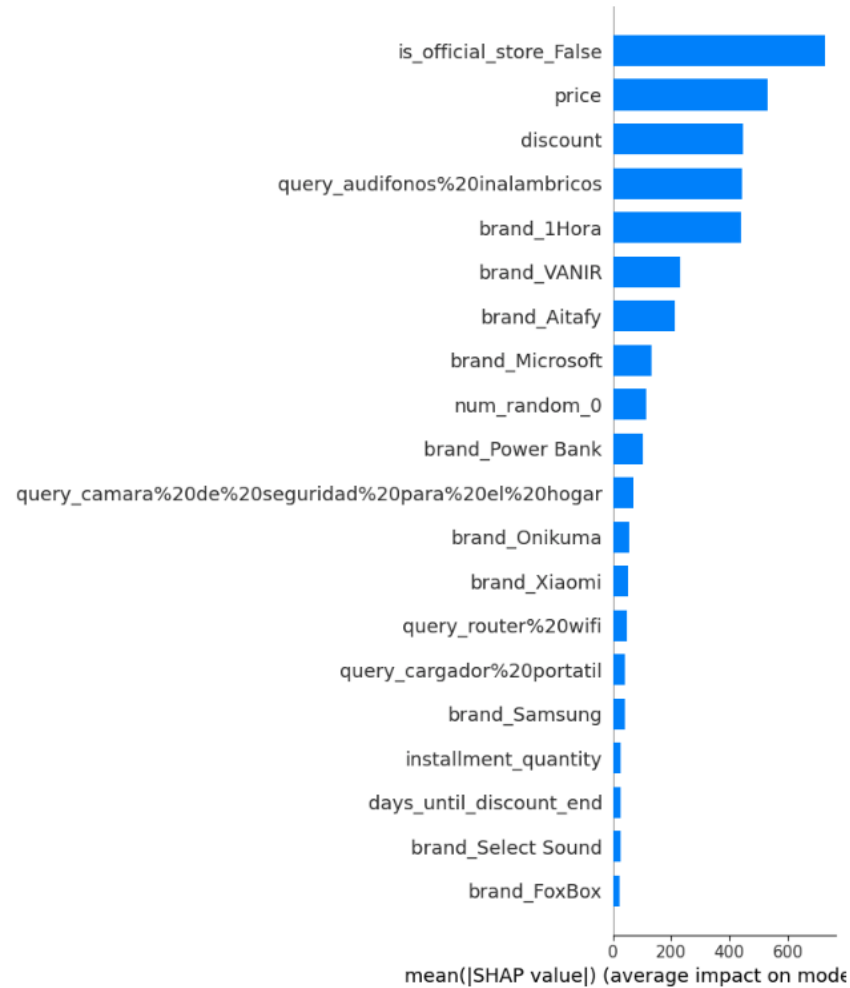
¿Realmente es mejor un modelo que otro?

- Compararemos el conjunto de valores de sus predicciones con una prueba estadística

Prueba t entre E1 y E2: Estadístico t = -1.4906, p-value = 0.1373
No podemos rechazar la hipótesis nula: No hay evidencia suficiente para decir que hay una diferencia significativa entre las predicciones de E1 y E2.

Comparación de Densidades - E1 vs E2
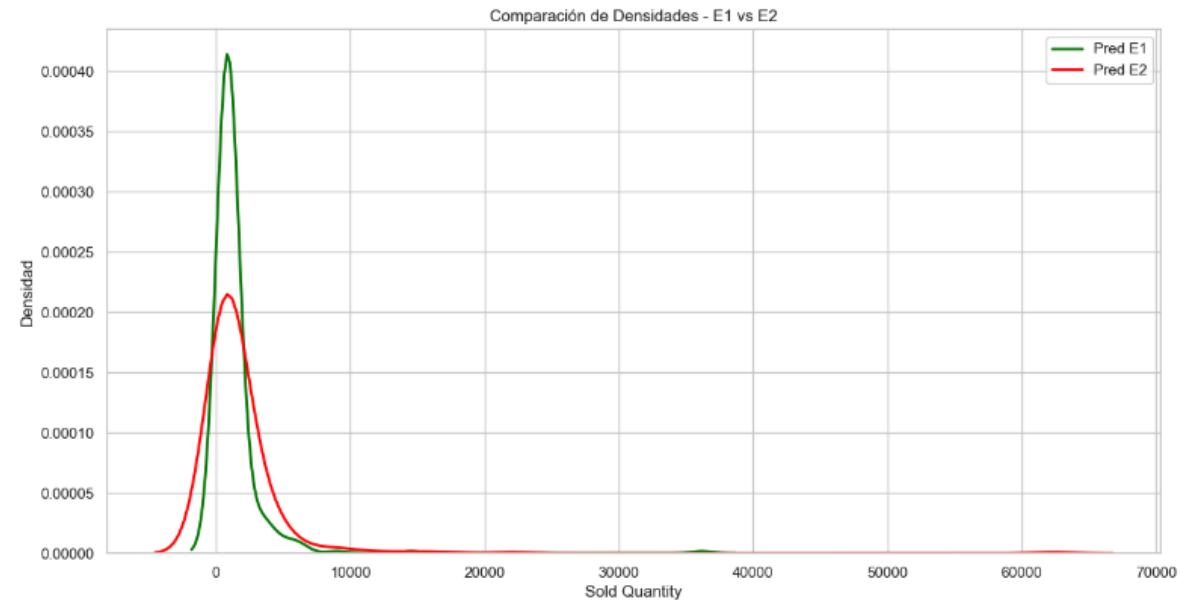
Siguientes pasos

- Determinar que features tendríamos en efecto en producción
- Hacer una búsqueda de hiper-parametros más exhaustiva para los algoritmos que tuvieron mejor performance