TFM GRUPO 3 DSMarket

Para identificar grupos de productos similares, hacemos un clustering con los datos obtenidos.

Clustering Items

Grupo Indicadores	Indicador	Estadístico	items mayores ingresos	items Menos ppopulares	items Mas populares	items menor stock	Items mayor stock
			cluster				
Monetarios	Importe	Media	490060.217703	52275.911027	79173.810000	69017.473698	40545.694187
		Desviación	262606.777406	34061.750050	64135.916318	56019.199849	43063.602014
		Mínimo	268227.351000	16649.208000	10743.812500	1177.183000	558.879300
		Perc. 25	331979.753375	26586.228000	24916.443750	29717.820000	13952.177250
		Perc. 50	396496.862500	33937.164000	65327.550000	52054.683450	26906.352200
		Perc. 75	539644.158000	70484.184000	109293.756250	91029.285000	49886.899400
		Máximo	1691829.504000	128949.600000	208901.700000	280534.300000	332383.350000
Frecuencia	tiempo última venta	Media	1558.608333	565.133333	1197.466667	1604.168014	604.531277
		Desviación	378.087676	435.251958	446.493300	275.939326	312.900031
		Mínimo	112.000000	49.000000	357.000000	672.000000	0.000000
		Perc. 25	1335.250000	262.500000	913.500000	1400.000000	266.000000
		Perc. 50	1771.000000	469.000000	1281.000000	1757.000000	616.000000
		Perc. 75	1841.000000	843.500000	1526.000000	1834.000000	882.000000
		Máximo	1841.000000	1379.000000	1841.000000	1841.000000	1253.000000
Popularidad	Top ventas	Media	0.000000	0.000000	10.000000	0.000000	0.000000
		Desviación	0.000000	0.000000	0.000000	0.000000	0.000000
		Mínimo	0.000000	0.000000	10.000000	0.000000	0.000000
		Perc. 25	0.000000	0.000000	10.000000	0.000000	0.000000
		Perc. 50	0.000000	0.000000	10.000000	0.000000	0.000000
		Perc. 75	0.000000	0.000000	10.000000	0.000000	0.000000
		Máximo	0.000000	0.000000	10.000000	0.000000	0.000000
	Peores ventas	Media	0.000000	10.000000	0.000000	0.000000	0.000000
		Desviación	0.000000	0.000000	0.000000	0.000000	0.000000
		Mínimo	0.000000	10.000000	0.000000	0.000000	0.000000
		Perc. 25	0.000000	10.000000	0.000000	0.000000	0.000000
		Perc. 50	0.000000	10.000000	0.000000	0.000000	0.000000
		Perc. 75	0.000000	10.000000	0.000000	0.000000	0.000000
		Máximo	0.000000	10.000000	0.000000	0.000000	0.000000

Análisis Clustering Ítem

Hemos obtenido una segmentación de 5 clústeres.

- El primer clúster aglutina los ítems que más ingresos generan, los de mejor tiquet promedio y con el precio promedio más alto.
- El segundo clúster aglutina los ítems con peores ventas.
- El tercer clúster aglutina los ítems más vendidos.
- El cuarto es un clúster que aglutina los ítems de menor tiempo de stock. Y con un rango de precios mayor.
- El quinto clúster aglutina una gran cantidad de ítems que tiene un mayor tiempo en stock.

Resultados de la segmentación

Una vez observados los resultados, fijándonos en los departamentos a los que pertenecen los ítems, podemos concluir:

- En el clúster de “Ítems mayores ingresos”, la mitad pertenecen al departamento de “Supermarket”.
- Clúster “Ítems menos populares”, la gran mayoría pertenecen al departamento de “Supermarket_2”.
- Clúster “Ítems más populares”, la gran mayoría pertenecen al departamento de Home & Garden_1.
- Clústeres de “menor stock” y “mayor stock”, contienen ítems de todo tipo, son grupos más heterogéneos.

Clustering Tiendas

Grupo Indicadores	Indicador	Estadístico	Tiendas premium	De menor stock	Tiendas mas rentables	Tiendas 'marca blanca'
Monetarios	Importe	Media	20232861.365467	21372012.388300	38954694.439600	21288248.600200
		Desviación	2013287.663364	3027891.363913	nan	6255709.873942
		Mínimo	17918743.698600	18965575.217900	38954694.439600	14872474.908600
		Perc. 25	19558191.708250	19672117.930050	38954694.439600	18247095.034300
		Perc. 50	21197639.717900	20378660.642200	38954694.439600	21621715.160000
		Perc. 75	21389920.198900	22575230.973500	38954694.439600	24496135.446000
		Máximo	21582200.679900	24771801.304800	38954694.439600	27370555.732000
Evolucion precio	precios	Media	5.550726	5.444421	5.481252	5.494179
		Desviación	0.063498	0.030218	nan	0.009276
		Mínimo	5.513276	5.421843	5.481252	5.483791
		Perc. 25	5.514068	5.427257	5.481252	5.490454
		Perc. 50	5.514861	5.432671	5.481252	5.497117
		Perc. 75	5.569451	5.455710	5.481252	5.499374
		Máximo	5.624041	5.478749	5.481252	5.501631
Frecuencia	última venta	Media	0.000000	70.000000	0.000000	23.333333
		Desviación	0.000000	49.000000	nan	40.414519
		Mínimo	0.000000	21.000000	0.000000	0.000000
		Perc. 25	0.000000	45.500000	0.000000	0.000000
		Perc. 50	0.000000	70.000000	0.000000	0.000000
		Perc. 75	0.000000	94.500000	0.000000	35.000000
		Máximo	0.000000	119.000000	0.000000	70.000000
Mas popularidad	top ventas	Media	1016.333333	1016.333333	3049.000000	0.000000
		Desviación	1760.340971	1760.340971	nan	0.000000
		Mínimo	0.000000	0.000000	3049.000000	0.000000
		Perc. 25	0.000000	0.000000	3049.000000	0.000000
		Perc. 50	0.000000	0.000000	3049.000000	0.000000
		Perc. 75	1524.500000	1524.500000	3049.000000	0.000000
		Máximo	3049.000000	3049.000000	3049.000000	0.000000
Menos popularidad	low ventas	Media	0.000000	0.000000	0.000000	3049.000000
		Desviación	0.000000	0.000000	nan	0.000000
		Mínimo	0.000000	0.000000	0.000000	3049.000000
		Perc. 25	0.000000	0.000000	0.000000	3049.000000
		Perc. 50	0.000000	0.000000	0.000000	3049.000000
		Perc. 75	0.000000	0.000000	0.000000	3049.000000
		Máximo	0.000000	0.000000	0.000000	3049.000000

Análisis Clustering Tiendas

Hemos obtenido una segmentación de 4 clúster.

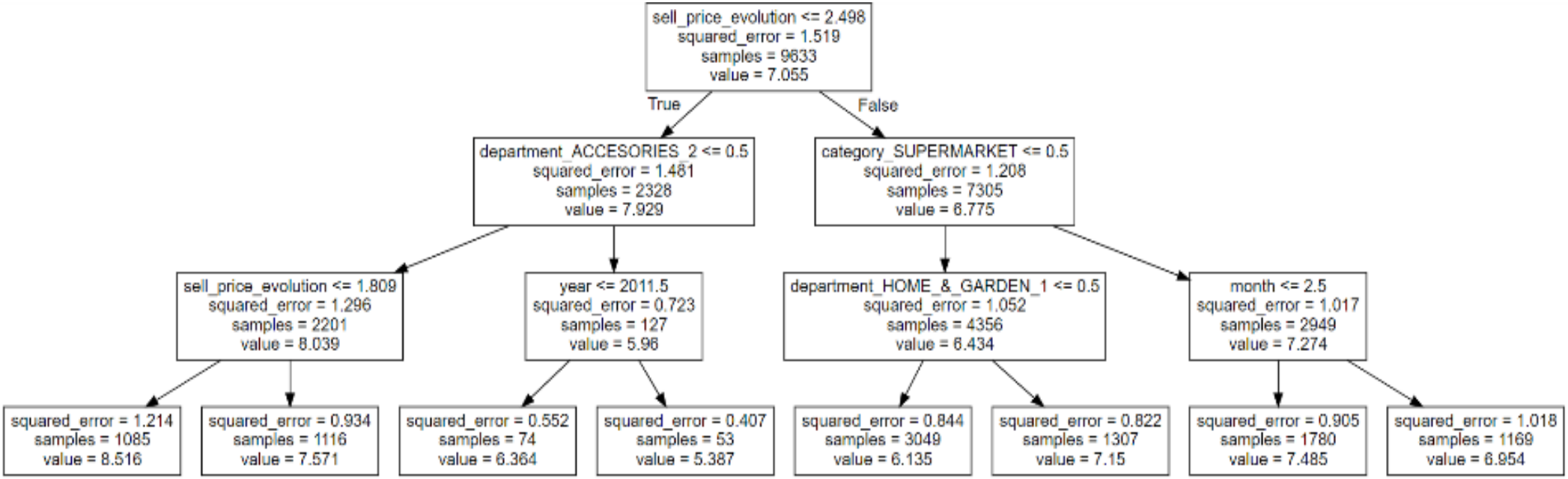
- El primer clúster es el que aglutina las tiendas con unos precios más altos y que más han variado.
- El segundo clúster aglutina las tiendas con un menor tiempo de stock de los productos.
- El tercer clúster aglutina las tiendas con mayores ingresos por ventas. Y con los productos top ventas.
- El cuarto clúster aglutina las tiendas con los productos con peores ventas.

Clúster / Tienda	Tiendas Premium	De menor stock	Tiendas más rentables	Tiendas marca blanca
Back_Bay				X
Brooklyn				X
Greenwich_Village				X
Harlem	X			
Midtown_Village	X			
Queen_Village		X		
Roxbury		X		
South_End		X		
Tribeca			X	
Yorktown	X			

Resultados de la segmentación

- El clúster de Tiendas Premium se corresponde con los barrios de moda, con mucha actividad comercial y de ocio.
- El clúster Tiendas de marca blanca se corresponde con zonas habitadas por población más joven, con tendencia a no comprar productos de marca.
- El clúster de Menor Stock, se corresponde con zonas residenciales.
- El clúster de las Tiendas más rentables, se compone por la tienda de Tribeca. Que pertenece a uno de los barrios más populares de Nueva York.

Modelo de predicción de ventas: Machine Learning - Regresión



- Árbol de decisión del algoritmo empleado para la predicción.
- Profundidad de 3 segmentaciones para evitar samples demasiado pequeñas.
- El año y el mes influyen en el precio de algunos productos.
- Según el departamento se observan patrones de ventas distintos.
- Los productos con menos sell_price_evolution tienden a ser más baratos.

Top Features

sell_price_evolution	0.488439
category_SUPERMARKET	0.217102
department_HOME_&_GARDEN_1	0.156331
department_ACCESORIES_2	0.082044
month	0.023079
department_HOME_&_GARDEN_2	0.013813
category_ACCESORIES	0.011585
year	0.006748
department_SUPERMARKET_3	0.000488
region_New York	0.000372
store_code_PHI_1	0.000000
store_code_PHI_2	0.000000
store_code_NYC_4	0.000000
store_code_PHI_3	0.000000
store_South_End	0.000000

dtype: float64

- Sell_price_evolution es la variable con más information gain para el algoritmo.

Predicción de ventas

item	Target	Prediction-RF
ACCESORIES_1_002	6.496775	6.131473
ACCESORIES_1_002	6.496775	6.131473
ACCESORIES_1_002	6.496775	6.131473
ACCESORIES_1_002	6.202536	6.131473
ACCESORIES_1_002	6.202536	6.131473
ACCESORIES_1_002	6.202536	6.131473
ACCESORIES_1_002	7.042286	6.131473
ACCESORIES_1_002	7.042286	6.131473
ACCESORIES_1_002	7.042286	6.131473
ACCESORIES_1_004	8.063063	6.131473

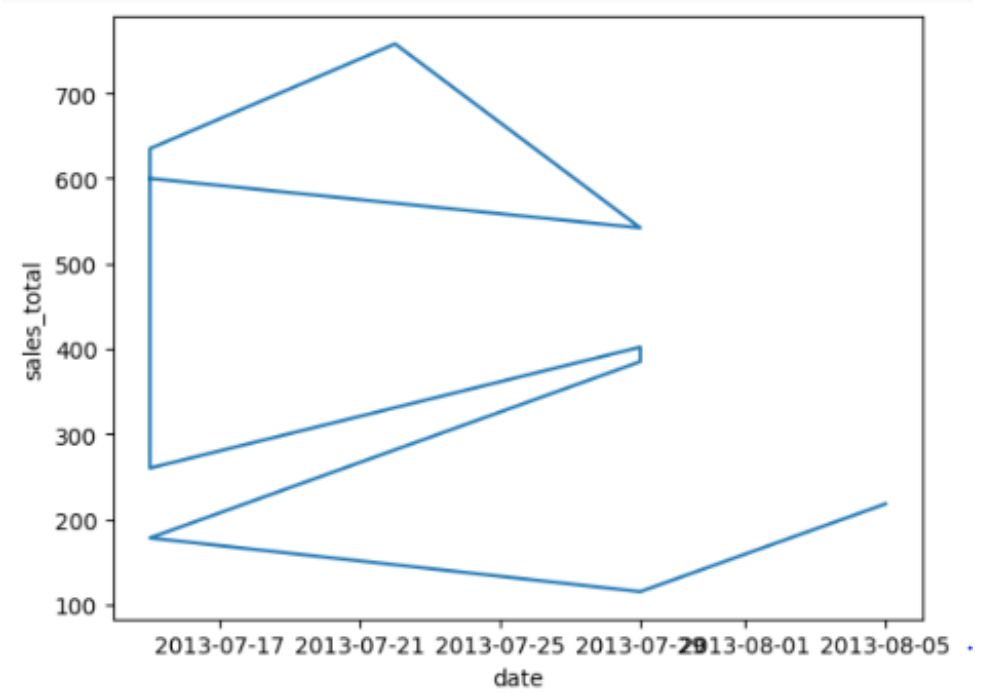
item	Target	Prediction-RF
SUPERMARKET_3_090	12.431226	8.560694
SUPERMARKET_3_090	12.431226	8.560694
SUPERMARKET_3_090	12.431226	8.560694
SUPERMARKET_3_090	11.691047	8.560694
SUPERMARKET_3_090	11.691047	8.560694
SUPERMARKET_3_090	11.691047	8.560694
SUPERMARKET_3_586	11.630308	8.560694
SUPERMARKET_3_586	11.630308	8.560694
SUPERMARKET_3_226	11.495047	8.560694
SUPERMARKET_3_226	11.495047	8.560694
SUPERMARKET_3_226	11.495047	8.560694
SUPERMARKET_3_226	11.495047	8.560694
SUPERMARKET_3_635	11.292566	8.560694
SUPERMARKET_3_635	11.292566	8.546624
SUPERMARKET_3_635	11.292566	8.560694
SUPERMARKET_3_226	11.277127	8.560694
SUPERMARKET_3_226	11.277127	8.560694
SUPERMARKET_3_226	11.277127	8.560694
SUPERMARKET_3_226	11.277127	8.560694
SUPERMARKET_3_377	11.209087	8.560694

- En la tabla izquierda los datos no están ordenados. En la tabla derecha sí están en orden ascendente.
- Buenas predicciones en valores bajos.
- Predicciones no tan ajustadas en valores altos.
- Posiblemente debido a la calidad y distribución de los datos analizados.

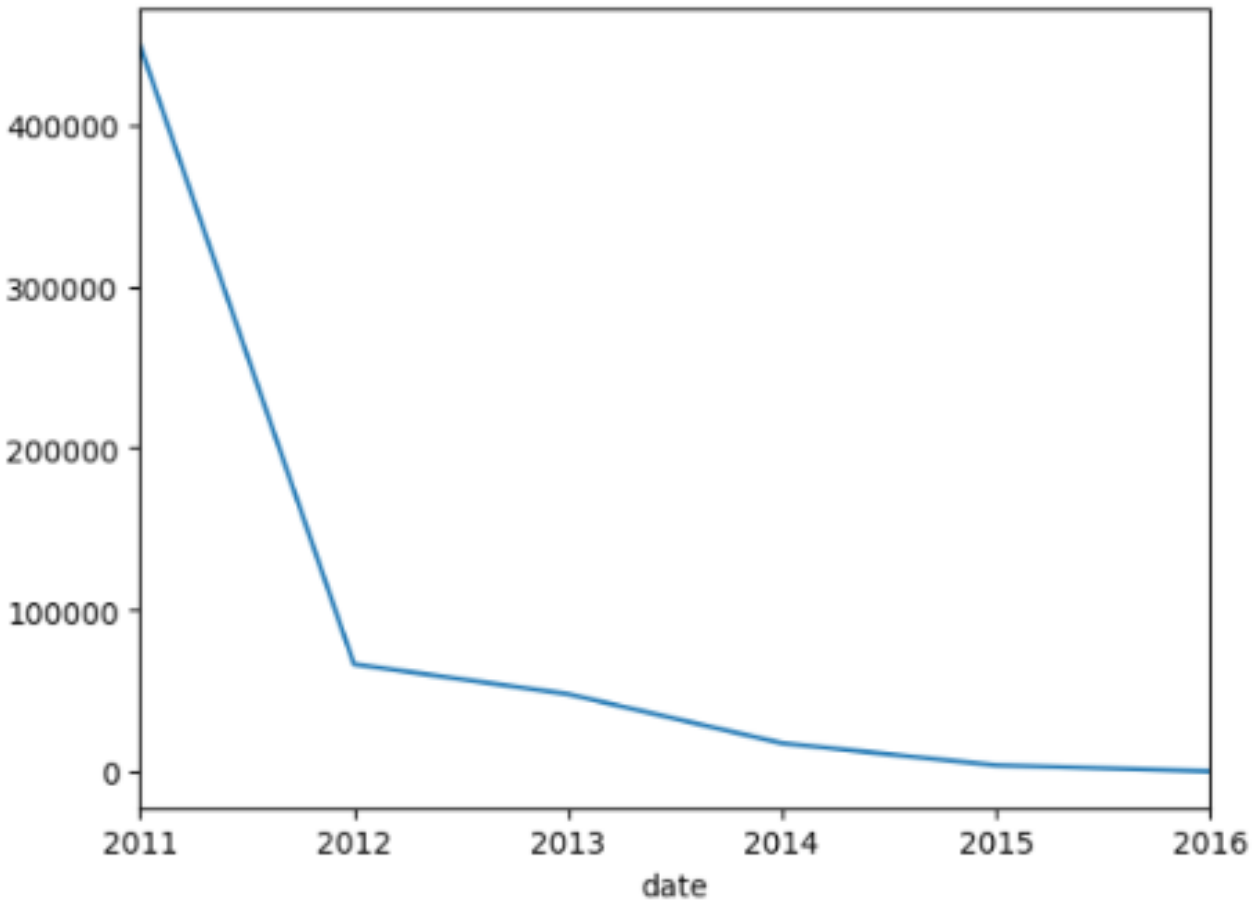
Conclusiones de la predicción de ventas enfoque ML Regresión.

- En train y test el modelo predecía bastante bien. Pero al ir a validación las predicciones empeoran.
- El modelo no dispone de suficientes datos para realizar la generalización.
- Rebalanceo de datos ampliando muestra, conseguimos valores un poco más altos. Pero las métricas no mejoran.

Modelo de predicción de ventas: Serie Temporal



- Gráfico ejemplo de ventas/ítem en un periodo de tiempo.
- Datos se concentran en un periodo muy pequeño de tiempo.
- No siguen una distribución uniforme de las ventas a lo largo del tiempo.



- Gráfico ejemplo de la media de todas las ventas en Nueva York.
- Se observa que las ventas bajan drásticamente.
- Es posible que sea debido a la falta de datos del muestreo.

Conclusiones de la predicción de ventas: enfoque Series Temporales

- Utilizamos algoritmo xgb.XGBregressor, para la predicción.
- El algoritmo da un error alto posiblemente debido a la poca dimensionalidad del dataframe.
- Se realiza un rebalanceo de los datos para aumentar las muestras tanto en train, test y validación. Ya que sin este rebalanceo, el algoritmo no disponía de suficientes datos para poder hacer la predicción.
- Al hacer rebalanceo, también baja el error. Tanto en validación como en train.

Conclusiones de la predicción de ventas: enfoque Series Temporales

- Esta es una muestra de las predicciones de ventas que hemos obtenido:

item	Index	Actual	Predicted
ACCESORIES_1_283	0	204	116.513893
ACCESORIES_2_132	1	184	148.374527
HOME_&_GARDEN_1_159	2	146	172.428635
HOME_&_GARDEN_1_242	3	175	141.315140
HOME_&_GARDEN_1_274	4	191	139.095245