SAMPLE - SUPERSTORE SALES

Entrega Proyecto Final

Equipo de trabajo:
- Lombardini Gabriela
- Sosa Alonso Florencia Victoria

Índice:

1- Contexto y audiencia:

La base de datos seleccionada consiste en información sobre ventas y envíos en EEUU de la empresa superstore.

Además de los precios, también incluye cantidades, descuentos, regiones de compra, datos de compradores (como el tipo de comprador según su tarea), categoría y sub categoría del producto, medio de envió, fecha de compra y envió, etc.

La dataset es de uso público y puede encontrarse en Tableau
https://community.tableau.com/s/question/0D54T00000CWeX8SAL/sample-superstore-sales-excelxls

Lo que buscamos con este análisis es poder tener una mejor comprensión de las preferencias del cliente para poder focalizar en aquellas categorías que mas rentabilidad le proporciona a la compañía. Para esto proponemos a continuación una serie de gráficos que les permitirán tener una mejor comprensión del análisis realizado.

Contexto

En el ámbito de las ventas, se observa una amplia variación en las compras de ciertos productos. Aunque nuestra compañía cuenta con una amplia gama de productos disponibles para la venta, hemos identificado que las cantidades vendidas difieren significativamente entre las categorías de "Muebles" y "Tecnología" en comparación con la categoría de "Material de Oficina".

Audiencia

Este análisis está dirigido a nuestros valiosos clientes con el objetivo de proporcionar información que les permita identificar patrones de ventas y maximizar sus ganancias.

Limitaciones

En cuanto a las limitaciones del análisis, hasta el momento no se han identificado restricciones significativas que puedan afectar la precisión de los resultados. Sin embargo, es importante tener en cuenta que los datos utilizados se basan en la información disponible y pueden estar sujetos a

ciertas limitaciones inherentes al proceso de recopilación de datos. Estamos comprometidos a brindar la información más precisa y útil posible, dada la disponibilidad de los datos recopilados.

2- Hipótesis/preguntas de interés:

El objetivo de este análisis es poder responder a las siguientes preguntas:

¿Qué región vende más?

¿Cuál es la categoria mas comprada?

¿Que tipo de comprador es el mas habitual?

¿Cuáles son las 5 sub categorias que más facturaron?

¿Cualés son los mayores descuentos por categoria y sub categoria?

3-Metadata

El dataset contiene datos sobre las ventas y la facturación de la compañía, extraídos de su base de datos. La compañía proporcionó esta información a este equipo de científicos de datos para que la analizara y generara nueva información relevante.

- row_ID : Número de identificación de la fila.
- order_ID : Código de identificación de la orden.
- order_date : Fecha en que se realizó la orden.
- ship_date : Fecha en que se envió la orden.
- ship_mode : Modo en que se envió.
- customer_ID : Código de identificación del cliente.
- customer_name : Nombre del cliente
- segment : Clasificación de cliente segun tipo de consumidor.
- country/region : País donde reside el cliente.
- city : Ciudad donde reside el cliente.
- state : Estado donde reside el cliente.
- postal_code : Código postal del domicilio del cliente.
- region : Ubicación cardinal del socio en la región.
- product_ID : Código de indentificación del producto comprado.
- category : Categoría del tipo de producto comprado.
- sub_category : Sub-categoria del producto comprado.
- product_name : Nombre del producto comprado.
- sales : Precio de la venta.
- quantity : Cantidad del producto vendido.
- discount : Descuentos ofrecidos.

- profit : Ganancia obtenida.

El dataset contiene `21` variables con `9993` registros, entre los que se encuentra información por categoría, subcategoría y región con sus respectivas cantidades, montos de ventas, descuentos aplicados y ganancias obtenidas.

Tabla resumen de la información:

#	Column	Non-Null Count	Dtype
0	row_ID	9994 non-null	int64
1	order_ID	9994 non-null	object
2	order_date	9994 non-null	datetime64[ns]
3	ship_date	9994 non-null	datetime64[ns]
4	ship_mode	9994 non-null	object
5	customer_ID	9994 non-null	object
6	customer_name	9994 non-null	object
7	segment	9994 non-null	object
8	country/region	9994 non-null	object
9	city	9994 non-null	object
10	state	9994 non-null	object
11	postalcode	9983 non-null	float64
12	region	9994 non-null	object
13	product_ID	9994 non-null	object
14	category	9994 non-null	object
15	sub_category	9994 non-null	object
16	product_name	9994 non-null	object
17	sales	9994 non-null	float64
18	quantity	9994 non-null	int64
19	discount	9994 non-null	float64
20	profit	9994 non-null	float64

Estadísticas descriptivas iniciales:

	row_ID	postalcode	sales	quantity	discount	profit
count	9994.000000	9983.000000	9994.000000	9994.000000	9994.000000	9994.000000
mean	4997.500000	55245.233297	229.858001	3.789574	0.156203	28.656896
std	2885.163629	32038.715955	623.245101	2.225110	0.206452	234.260108
min	1.000000	1040.000000	0.444000	1.000000	0.000000	-6599.978000
25%	2499.250000	23223.000000	17.280000	2.000000	0.000000	1.728750
50%	4997.500000	57103.000000	54.490000	3.000000	0.200000	8.666500
75%	7495.750000	90008.000000	209.940000	5.000000	0.200000	29.364000
max	9994.000000	99301.000000	22638.480000	14.000000	0.800000	8399.976000

4- Análisis Exploratorio:

Se aplicaron distintas técnicas para mejorar la calidad y la utilidad de los datos en el análisis, identificamos y corregimos errores, inconsistencias, duplicados, valores faltantes o atípicos en los conjuntos de datos. Al limpiar los datos, se eliminan las anomalías que podrían afectar negativamente a los resultados del análisis o a la precisión de los modelos de aprendizaje automático.

Luego de aplicar dichas técnicas de limpieza de datos obtenemos información más clara y confiable:

Estadísticas descriptivas finales

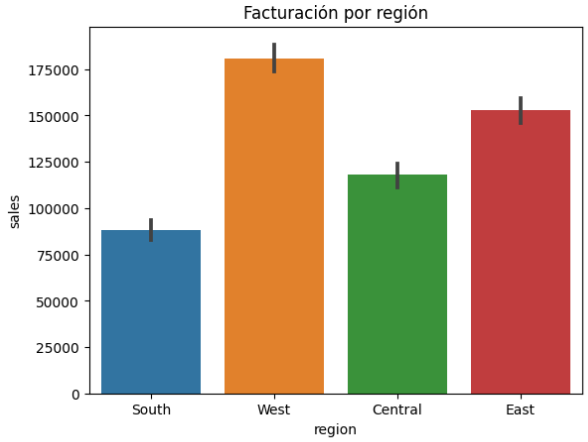
	sales	quantity	discount	profit	sales_proc	quantity_proc	discount_proc	profit_proc	year_order
count	7863.000000	7863.000000	7863.000000	7863.000000	7863.000000	7863.000000	7863.000000	7863.000000	7863.000000
mean	68.668357	3.574971	0.160198	9.402337	63.518085	2.522447	0.101494	5.074110	2019.733308
std	75.440016	2.125231	0.217531	37.690150	63.090638	0.683768	0.099621	32.105509	1.123835
min	0.444000	1.000000	0.000000	-766.012000	0.444000	1.000000	0.000000	-766.012000	2018.000000
25%	14.212000	2.000000	0.000000	1.670400	14.212000	2.000000	0.000000	1.670400	2019.000000
50%	36.288000	3.000000	0.200000	6.965400	36.288000	3.000000	0.200000	6.965400	2020.000000
75%	97.000000	5.000000	0.200000	18.873100	97.000000	3.000000	0.200000	18.873100	2021.000000
max	310.443000	14.000000	0.800000	149.760000	195.396000	3.000000	0.200000	28.320900	2021.000000

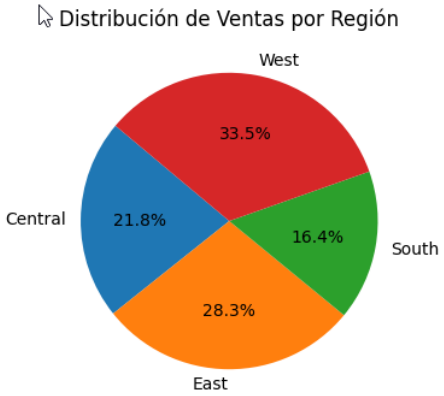
Hacemos un agrupado con categoría y subcategoría para poder analizar los datos:

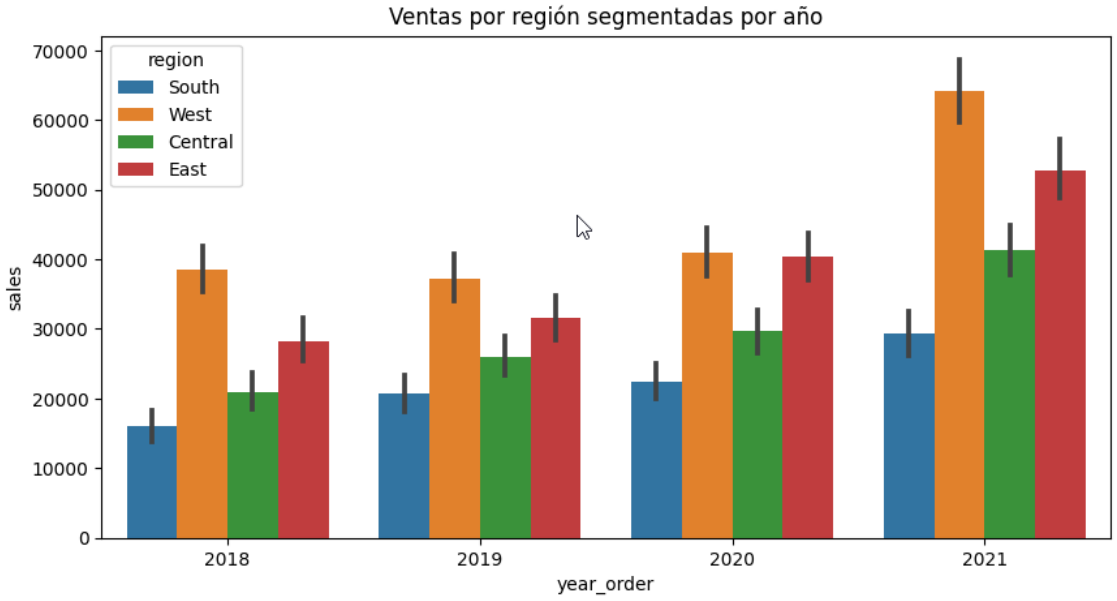
category	sub_category	sales	quantity	discount	profit	sales_proc	quantity_proc	discount_proc	profit_proc
Furniture	Bookcases	187.413656	2.721739	0.242522	-31.314173	159.029042	2.269565	0.135217	-35.142557
	Chairs	172.812494	2.643123	0.175465	5.206019	151.825543	2.234201	0.149814	1.369382
	Furnishings	63.973238	3.583524	0.142563	8.942734	60.662254	2.572082	0.081693	5.516059
	Tables	185.048495	2.183486	0.300459	-33.138597	157.302009	2.000000	0.176147	-35.185911
Office Supplies	Appliances	86.526272	3.422043	0.193011	-0.492628	80.983449	2.500000	0.086559	-6.519669
	Art	32.333956	3.775484	0.075613	7.950558	31.407752	2.588387	0.075613	6.709850
	Binders	35.248275	3.874372	0.383345	1.516647	34.051444	2.598708	0.157502	-1.298349
	Envelopes	55.846130	3.516260	0.080488	23.812429	53.173081	2.520325	0.080488	15.108820
	Fasteners	13.995516	4.213953	0.081860	4.395593	13.995516	2.674419	0.081860	4.395593
	Labels	27.341373	3.849112	0.070414	12.016714	26.684852	2.556213	0.070414	9.598185
	Paper	52.177363	3.792020	0.075212	22.580222	50.170594	2.564516	0.075212	14.809100
	Storage	96.376955	3.324459	0.083195	5.931313	88.190300	2.447587	0.083195	4.039689
	Supplies	39.607976	3.339286	0.077381	3.972305	37.851917	2.470238	0.077381	3.900262
Technology	Accessories	101.639755	3.515924	0.082803	23.041602	94.305917	2.528662	0.082803	14.334025
	Copiers	299.990000	1.000000	0.000000	89.997000	195.396000	1.000000	0.000000	28.320900
	Machines	137.230103	2.827586	0.400000	-32.950921	118.782483	2.344828	0.158621	-40.505266
	Phones	125.598207	3.135870	0.156522	15.463151	111.978808	2.403986	0.130435	9.654175

¿Qué región vende más?

En la suma total de ventas por región podemos ver que la zona oeste es la que tiene un mayor capital vendido.



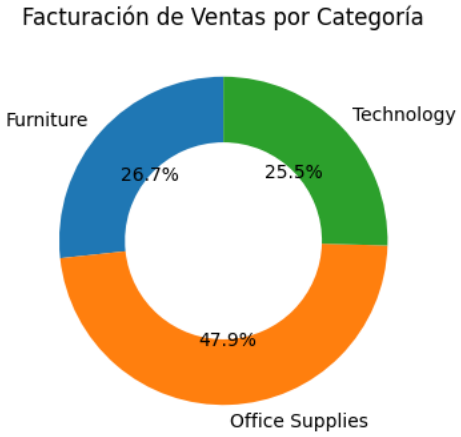


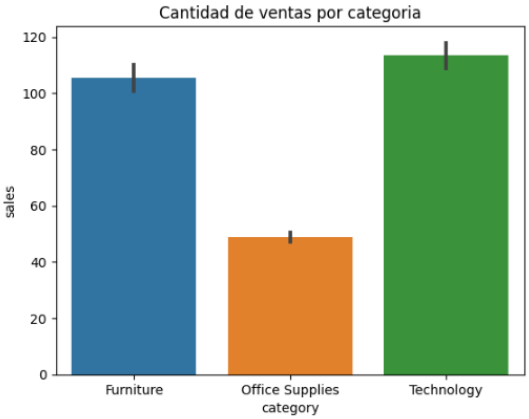


¿Cuál es la categoría mas comprada?

El gráfico entre categorías y ventas indica que los artículos tecnológicos son los mas vendidos, tanto en suma total como en cantidad.

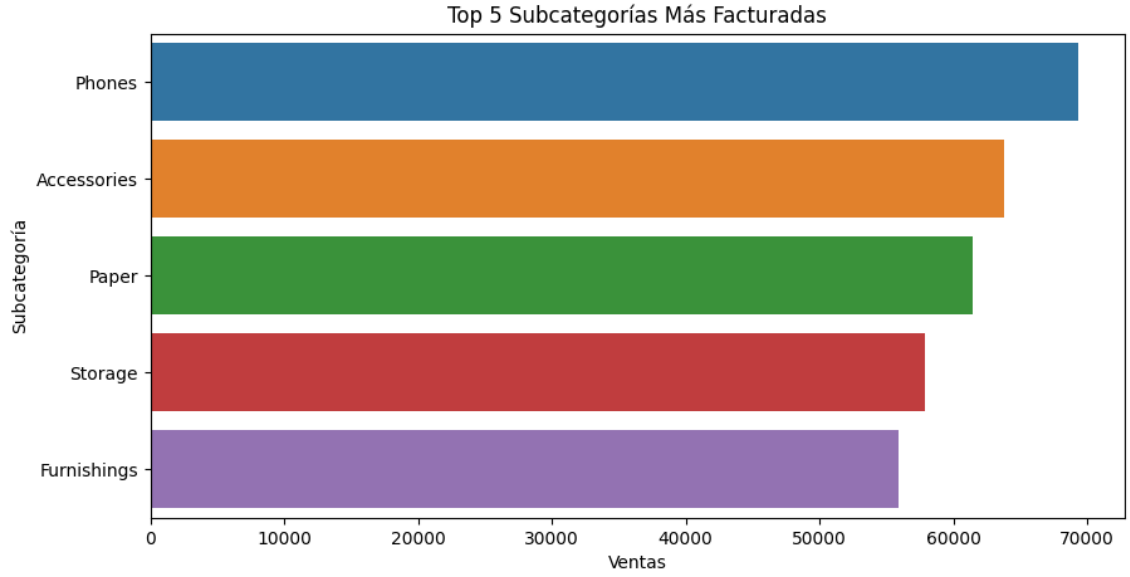
Dato extra: si analizamos los suplementos de oficina, vemos que son segundos en cuanto a suma total vendida y últimos (por mucha diferencia) si tenemos en cuenta la cantidad vendida. Lo que indicaría que vendiendo menos productos se generan mayores valores de venta.





¿Cuáles son las 5 sub_categorias que mas facturaron?

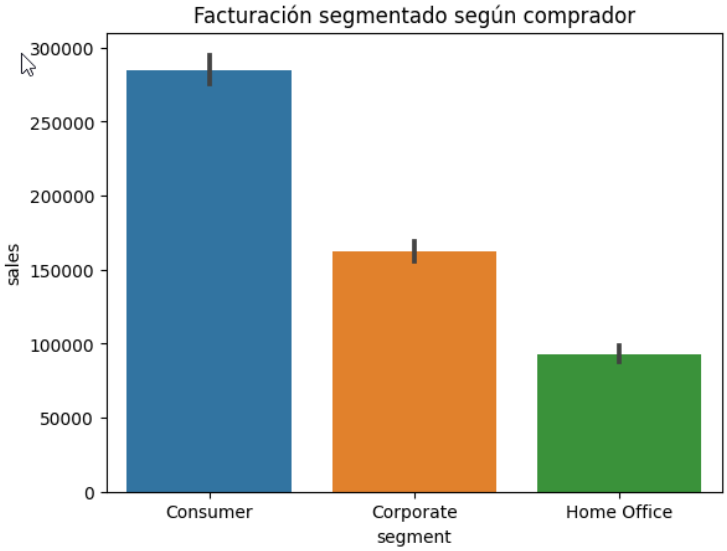
Si analizamos el gráfico entre las ventas y las sub_categorias, podemos visualizar que el artículo sillas es el que más facturó, seguido por teléfonos, almacenamiento, carpetas y accesorios (respectivamente).



¿Que tipo de comprador es el mas habitual?

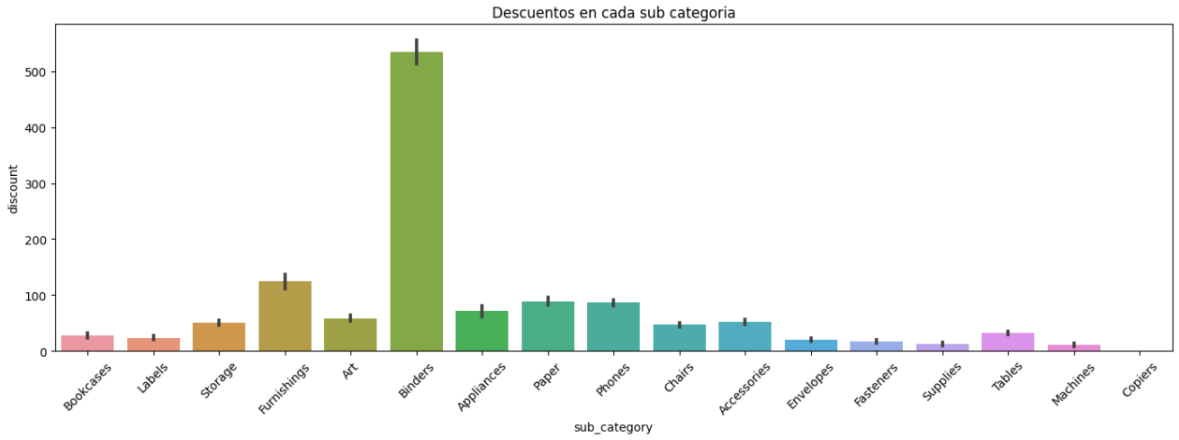
El recuento que nos muestra este gráfico indica que el consumidor final es el principal cliente de nuestra base de datos, luego el cliente corporativo y por último el home office.

El mismo orden se respeta en suma facturada por tipo de cliente.

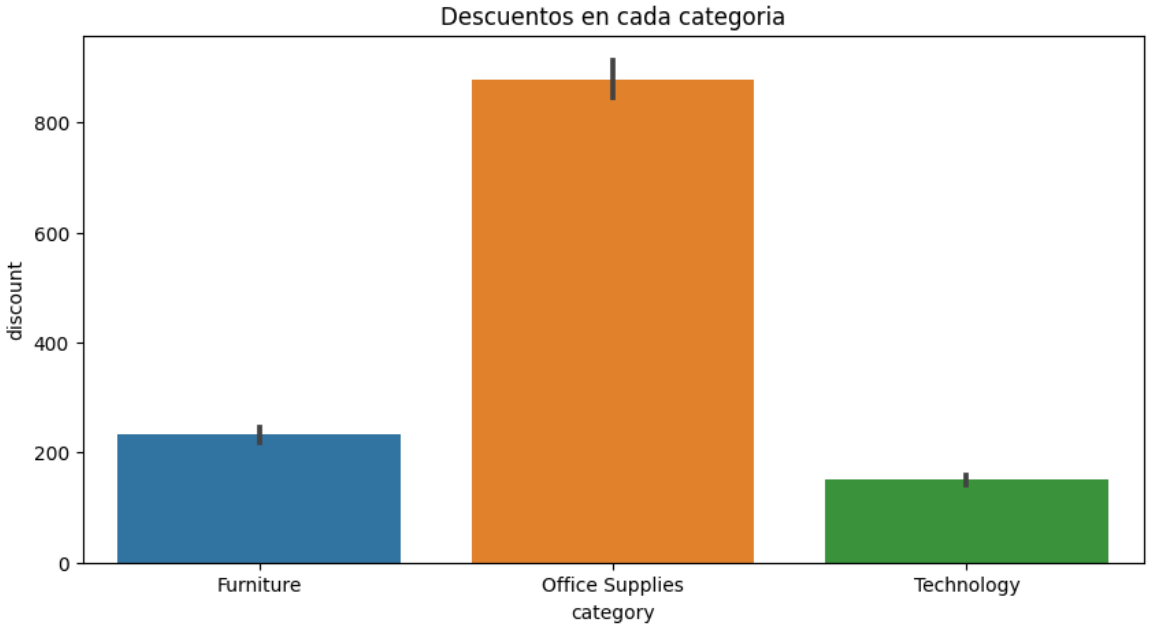


¿Cualés son los mayores descuentos por categoria y sub-categoria?

El gráfico anterior entre sub_categoria comprada y descuento ofrecido nos indica que: los artículos que reciben mayor descuento son carpetas, sillas, papel y teléfonos (en ese orden).



En cuanto al gráfico entre categorías y descuentos, los suplementos de oficina son los que reciben el mayor descuento.



5- Algoritmos elegidos

A fin de poder aplicar modelos de ML, se generó codificación a las variables categóricas, es decir se les asignó un valor numérico, mediante el algoritmo get_dummies.

Los modelos entrenados son los siguientes

a. Regresión Lineal
b. Árbol de Decisión
c. Random Forest.

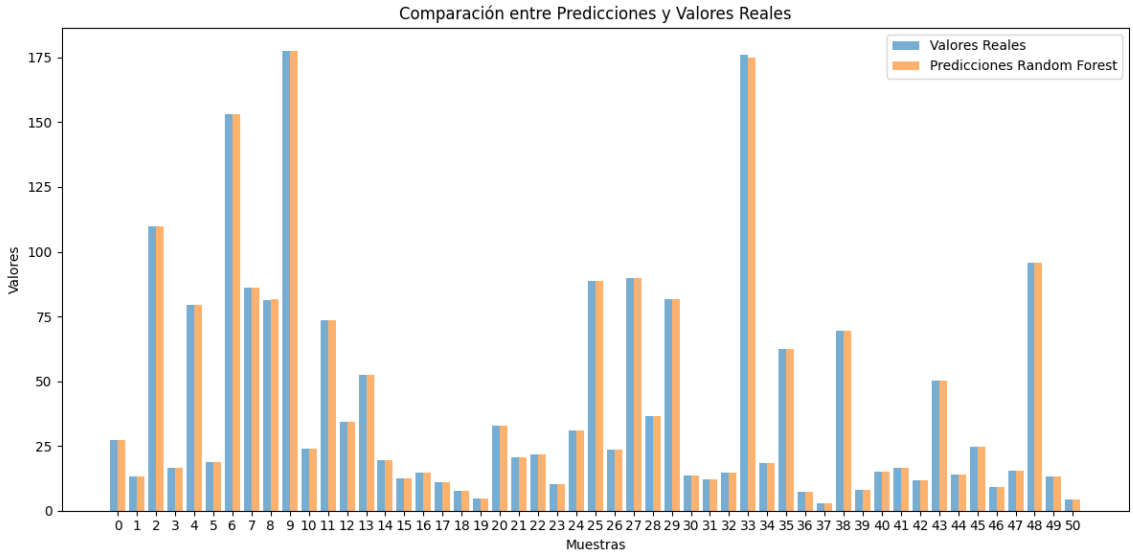
Analizando los resultados obtenidos en el proyecto de data science hasta el momento, podemos realizar las siguientes conclusiones:

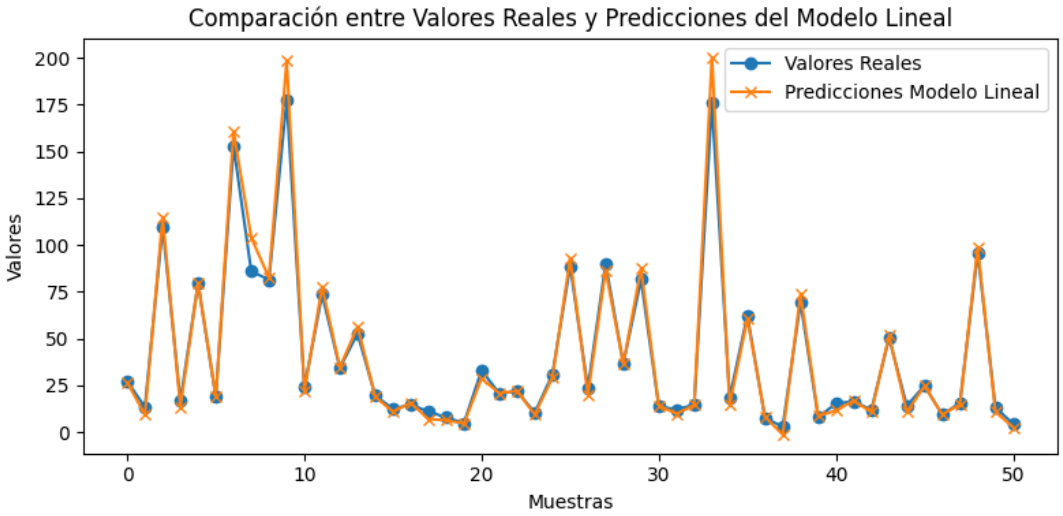
Durante el entrenamiento del modelo se han utilizado tres modelos de regresión para predecir el ingreso por ventas. Los resultados del Mean Squared Error (MSE) para cada modelo son los siguientes:

a. Linear Regression: 78.15
b. Decision Tree Regressor: 68.67
c. Random Forest Regressor: 41.76

El Random Forest Regressor obtiene el menor MSE, lo que indica que es el modelo que mejor se ajusta a los datos hasta el momento.

Este mismo indicador se repite en los siguientes gráficos.





Podemos observar que en la comparación de resultados reales y muestras, los valores se respetan en un gran porcentaje. Lo que indica un buen funcionamiento del modelo.

6-Validaciones del modelo:

- Validación simple

```
MSE:  0.01056234000114808
MAE:  0.021774378237525528
R2:  0.20994274292127402
```

- K-Fold Cross Validation:

```
MSE promedio usando K-Fold Cross Validation: 100.41778178331197
```

- Stratified- K fold

	sales	quantity	discount	profit	sales_proc	Consumer	Corporate	Home Office	Furniture	Office Supplies	...	Rhode Island	South Carolina	South Dakota	Tennessee	Texas	Utah	Vermont	Virginia	Washington	West Virginia
2	14.620	2	0.0	6.8714	14.620	0	1	0	0	1	...	0	0	0	0	0	0	0	0	0	0
4	22.368	2	0.2	2.5164	22.368	1	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
6	7.280	4	0.0	1.9656	7.280	1	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
29	124.200	3	0.2	15.5250	124.200	1	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
31	86.304	6	0.2	9.7092	86.304	1	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9958	7.300	2	0.0	2.1900	7.300	0	0	1	0	1	...	0	0	0	0	0	0	0	0	0	0
9976	249.584	2	0.2	31.1980	195.396	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	0
9980	85.980	1	0.0	22.3548	85.980	1	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
9989	25.248	3	0.2	4.1028	25.248	1	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
9992	29.600	4	0.0	13.3200	29.600	1	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0

1033 rows × 79 columns

- LOOCV (Leave One Out Cross-Validation)

```
Iteracion:  1  Accuracy: 0.989351403678606
Iteracion:  2  Accuracy: 0.9883833494675702
Iteracion:  3  Accuracy: 0.9883833494675702
Iteracion:  4  Accuracy: 0.9893410852713178
Iteracion:  5  Accuracy: 0.9893410852713178
Accuracy promedio: 0.9889600546312763
```

- LOOCV para regresión

```
[Parallel(n_jobs=1)]: Done  49 tasks      | elapsed:    2.1s
[Parallel(n_jobs=1)]: Done 199 tasks      | elapsed:    9.6s
[Parallel(n_jobs=1)]: Done 449 tasks      | elapsed:   20.8s
[Parallel(n_jobs=1)]: Done 799 tasks      | elapsed:   35.8s
[Parallel(n_jobs=1)]: Done 1249 tasks      | elapsed:   55.3s
[Parallel(n_jobs=1)]: Done 1799 tasks      | elapsed:  1.3min
[Parallel(n_jobs=1)]: Done 2449 tasks      | elapsed:  1.8min
[Parallel(n_jobs=1)]: Done 3199 tasks      | elapsed:  2.3min
[Parallel(n_jobs=1)]: Done 4049 tasks      | elapsed:  2.9min
[Parallel(n_jobs=1)]: Done 4999 tasks      | elapsed:  3.6min
MAE: 0.017 (0.092)
```

El resultado muestra que el modelo tiene un Error Absoluto Medio promedio de 0.017, lo que indica un buen rendimiento en la predicción de los valores objetivo. Sin embargo, la desviación estándar de 0.092 sugiere cierta variabilidad en los errores de predicción entre las iteraciones de la validación cruzada

Ánalisis y comparación de las validaciones

Realizamos distintos tipos de validaciones para corroborar las predicciones obtenidas:

1. Validación Simple:

Se ha aplicado la validación simple utilizando el modelo Random Forest Regressor. Los resultados obtenidos son: MSE: 0.01411 MAE: 0.02379 R2: 0.12808

Estos valores indican un bajo error cuadrático medio y un coeficiente de determinación (R2) que es bastante bajo. El modelo puede no estar generalizando de manera adecuada, y su capacidad para predecir el ingreso por ventas puede ser limitada.

2. K-Fold Cross Validation:

Se ha aplicado K-Fold Cross Validation utilizando el modelo Linear Regression. El MSE promedio obtenido es de 100.42. Este resultado indica que el modelo tiene un error medio elevado al generalizar en diferentes conjuntos de validación, lo que sugiere que podría estar sufriendo de sobreajuste.

3. Leave One Out Cross-Validation (LOOCV):

Se ha aplicado LOOCV utilizando el modelo Linear Regression. El MSE promedio obtenido es de 106.50. Este resultado muestra un error medio relativamente alto, lo que sugiere que el modelo puede no estar generalizando bien a datos no vistos.

4. Stratified K-Fold Cross Validation:

Se ha intentado aplicar Stratified K-Fold Cross Validation utilizando el modelo RandomForestClassifier. Sin embargo, parece que los datos no se han estratificado correctamente debido a que se está utilizando un modelo de regresión (RandomForestRegressor) en lugar de un modelo de clasificación (RandomForestClassifier). Por lo tanto, los resultados obtenidos aquí no son válidos para interpretar.

7-Conclusión

En resumen, a través de un análisis detallado de los datos y la implementación de modelos de Machine Learning, se logró comprender mejor las tendencias y patrones en los datos de ventas, y se construyó un modelo capaz de predecir ingresos por ventas con un rendimiento satisfactorio. Estos resultados podrían ser utilizados para la toma de decisiones en estrategias de ventas y marketing.

Dado que la categoría de suplementos de oficina es la más vendida, recomendamos crear ofertas o promociones que incentiven la compra cruzada de productos de otras categorías. Por ejemplo, se podría ofrecer un descuento en productos de papelería o limpieza con la compra de un producto de suplementos de oficina. De esta manera, se podría aumentar el valor promedio de los pedidos y fidelizar a los clientes. Además, la segmentación de clientes nos ayuda a identificar y redirigir las estrategias de marketing hacia los segmentos más rentables y tomar medidas específicas para mejorar el compromiso con los segmentos menos activos. Así, se podría optimizar el uso de los recursos y maximizar el retorno de la inversión.

Futuras líneas:

Análisis de Sensibilidad: Realiza un análisis de sensibilidad para identificar las variables más influyentes en las predicciones de ventas. Esto podría ayudarte a centrar tus esfuerzos en áreas clave para mejorar las estrategias de ventas.

Análisis Temporal: Incorpora información temporal en el modelo para prever tendencias estacionales o patrones a lo largo del tiempo, lo que podría proporcionar ideas para estrategias estacionales y promocionales más efectivas.

Segmentación de Clientes Avanzada: Utiliza técnicas de segmentación de clientes más avanzadas, como clustering o modelos de segmentación mixtos, para identificar grupos de clientes con comportamientos de compra similares y personalizar aún más tus estrategias de ventas.

Análisis de Causas Raíz: Si hay discrepancias entre las predicciones del modelo y las ventas reales, realiza un análisis de causas raíz para entender mejor por qué ocurren estas diferencias y tomar medidas correctivas basadas en los hallazgos.

Estas líneas de investigación te permitirán profundizar en diferentes aspectos de tu proyecto, mejorar tus modelos y tomar decisiones más informadas para impulsar las ventas y el rendimiento en general.