hack(io)

Proyecto 9:
Modelos de clustering y regresión para una empresa de e-commerce.

DAVID FRANCO

| DATA SCIENCE 2024

Objetivo y Herramientas

- Segmentar los datos mediante clustering y luego diseñar modelos de regresión específicos para cada segmento, utilizando las herramientas de sckitlearn.

hack(io)

Clusters

hack(io)

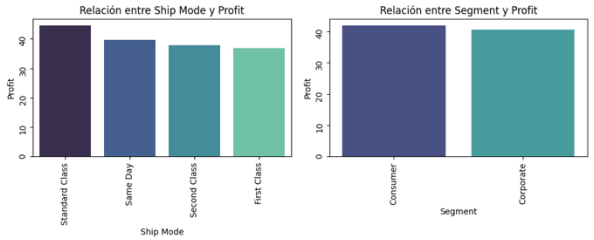
Segmentación generada por clientes

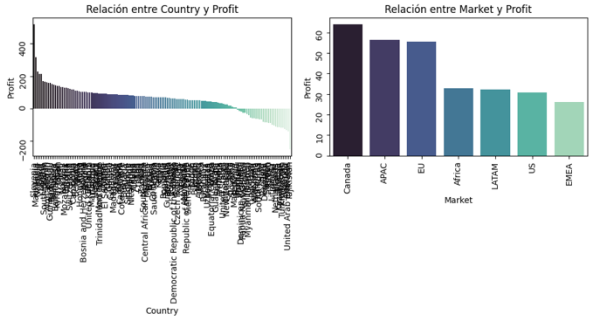
- **Cluster 0**
- Agrupa las ventas de los segmentos Consumer y Corporate, cuyas ventas se centran en tecnología y muebles.

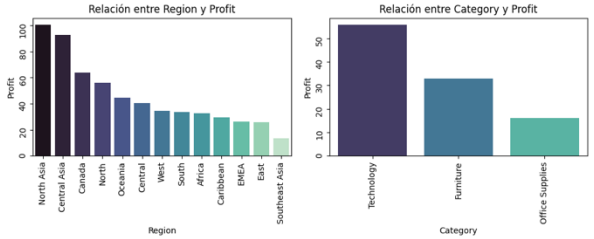
- **Cluster 1**
- Cubre exclusivamente al segmento de Home Office, aunque participa también de Consumer y Corporate. Su principal categoría de ventas son los suministros de oficina. Genera beneficios que superan en más de 2x los obtenidos por el cluster 1, sin olvidar que cuenta con cerca de 11 mil registros más.

Hallazgos del EDA

hack(io)







Variables categóricas

Hay un grupo de países que claramente generan pérdidas a la empresa, lo mismo que con la subcategoría de mesas.

En lo que respecta al ship mode, el tipo de envío no tiene relación directa con el beneficio, ocurre lo mismo con la prioridad, aunque en este caso, sorprende que, aquellos con prioridad crítica, generan un menor beneficio.

Asia (norte y central) es la región más rentable.

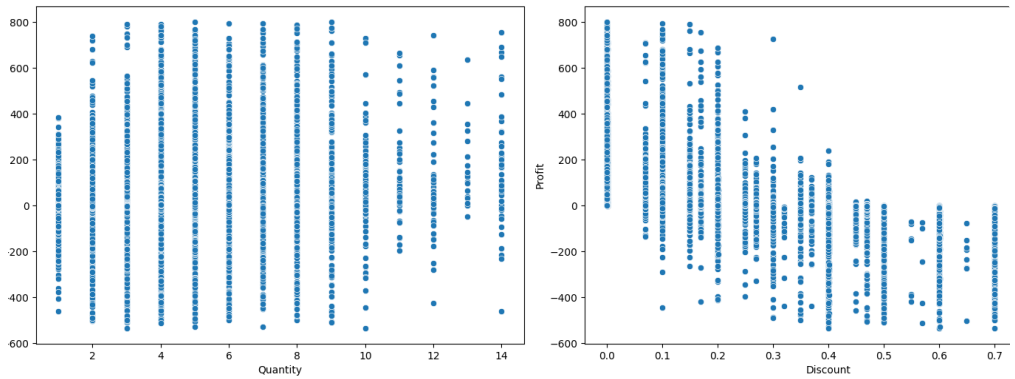
Aunque Canadá es el mercado con menores ventas, es el que mejores beneficios reporta.

La categoría más rentable es la de tecnología, seguida por los muebles.

Las subcategorías más rentables son las fotocopiadoras, los electrodomésticos y las estanterías para libros.

hack(io)

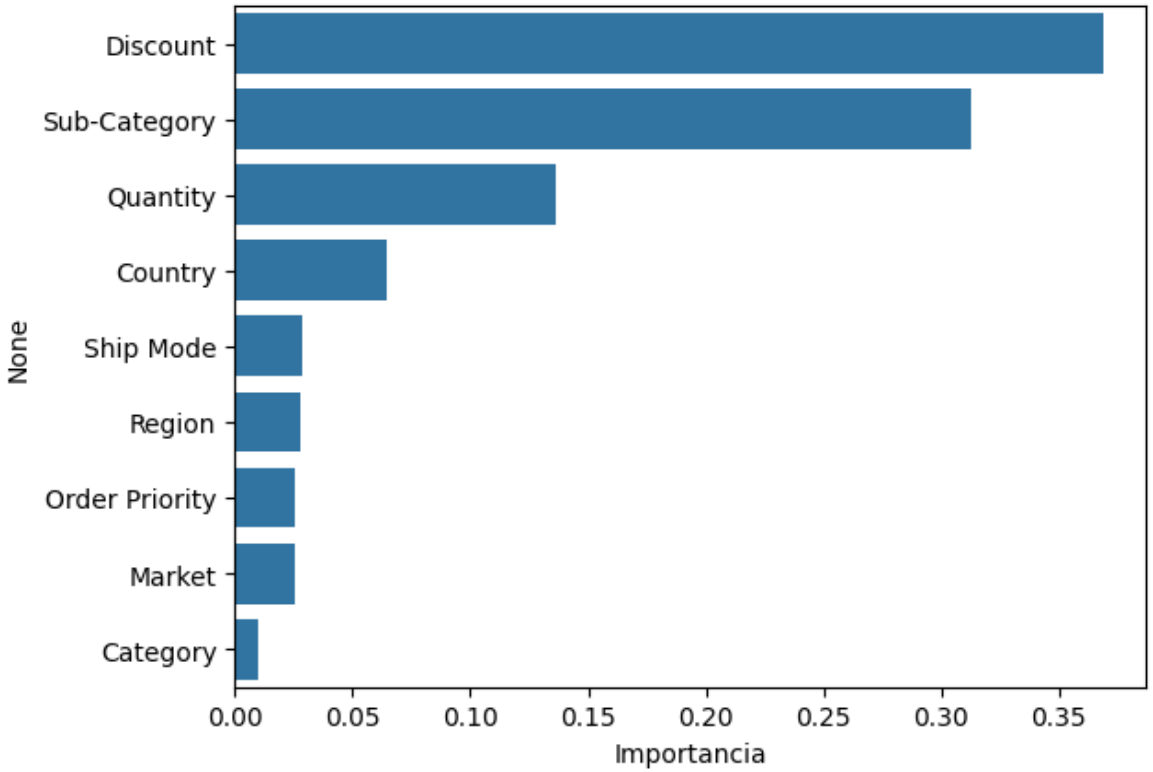
Variables numéricas



Cantidad: hasta 11 unidades no suele ser determinante del beneficio, aunque aquellos pedidos con una única unidad reportan beneficios mucho más pequeños. A partir de 11, en general, es menos probable tener pérdidas, aunque también, los beneficios suelen ser menores.

Descuento: hay una tendencia muy clara: a mayor descuento, mayor probabilidad de pérdidas. Con 0% de descuento, no hay pérdidas. Todos los pedidos con descuentos superiores al 40% generan pérdidas para la empresa.

Modelos elegidos:
XGBoost y Random Forest



Cluster 0	r2_score	RMSE
train	0.50	89.69
test	0.50	91.29

Cluster 1	r2_score	RMSE
train	0.49	43.29
test	0.50	43.52

hack(io)

Próximos pasos

CREAR UNA PLATAFORMA DE CONSULTA USANDO STREAMLIT, QUE INCLUYA UN MAPA DE CALOR CON LOS RESULTADOS DE BENEFICIOS POR REGIÓN.

DEDICAR MÁS TIEMPO A MEJORAR LOS RESULTADOS DE LOS MODELOS DE REGRESIÓN.

CREAR UN MODELO DE CLUSTERING QUE NO INCLUYA A LA VARIABLE "PROFIT".

CREAR UN MODELO DE CLUSTERING QUE NO ESTÉ BASADO EN CLIENTES, SINO EN PRODUCTOS.

hack(io)

hack(io)