Análisis de publicaciones con descuento y desarrollo de un modelo predictivo de pronóstico de ventas.

Partiendo de las consignas dadas de:

- Analizar publicaciones con descuento en Mercado Libre (MELI) y armar un dataset con las mismas.

- Desarrollar un modelo utilizando el dataset armado para predecir la cantidad de ventas que tendrá una publicación.

Dataset:

Tras analizar la API de MELI y haber probado la extracción de publicaciones se vio que se devolvieron pocas con descuento se decidió utilizar un scrapper sobre la sección de descuentos de la página y posteriormente yendo directamente a la información de cada artículo en particular, siendo esta forma de extracción de datos en regla con el archivo robots.txt de MELI.

También se utilizó la API con códigos de artículos individuales para obtener datos con más facilidad y también se usó para los datos de los vendedores.

Uniendo las 3 fuentes de datos se logró obtener un dataset de 10000+ publicaciones con descuento.

Análisis de datos:

Al tener el dataset unido se arrancó revisando los tipo de datos, nulos y filtración por duplicados, normalizando donde era posible. También se trabajó armando features que ayudaran con la predicción del dato de interés (sold\_quantity).

Luego de ver la correlación de los datos generó una cierta sorpresa que la cantidad de imágenes de una publicación fuera una de las que más influenciara las ventas. También se vio que el ID del vendedor tenía efecto sobre las ventas, eso se debía al hecho que había vendedores repetidos y lo más probable es que un mismo vendedor maneje un nivel similar de ventas a través de varias de sus publicaciones. Uno de los datos que tenía un efecto relativamente bajo era el porcentaje de descuento.

Seleccionando los features que más efecto tuvieron sobre el número de ventas se avanzó al desarrollo del modelo.

Modelo predictivo:

Para la selección del modelo se partió de una lógica de una regresión lineal, pero que requería ciertas decisión no lineales lo cual concluyó en un árbol de decisiones con el algoritmo de Random Forest.

\* No se generó un baseline aunque se debería.

Realizando y graficando los resultados de MAE, MSE y RMSE variando la cantidad de estimadores de observar en los gráficos que con 50 era suficiente para llegar a un error mínimo y el accuracy más alto.

También se trabajó con una función para analizar la influencia que tuvieran los features sobre los resultados del modelo y se observaron algunos con importancia más alta que lo que mostraba el gráfico de correlación, pero eran features esperados (como cantidad de reseñas) por sus datos respecto del label.

Tras cambiar sold\_quantity y available\_quantity a features más categóricos (cada categoría manejaba un rango de cantidades) utilizando una tabla de la API, se pudo obtener un mejor resultado en las predicciones.

Conclusion:

Aunque el modelo no logró tener un accuracy alto (0.46 - 0.48) surge la duda si es posible manejar un grado alto de exactitud con los datos armados. Mismo el valor de ventas por su naturaleza es difícil de pronosticar teniendo en cuenta que debe haber factores externos de los mercados que influencian la compra de ciertos tipo de artículos o marcas.

Sin embargo, si los datos obtenidos se siguieran trabajando, como los datos del título y la descripción se podría ir mejorando la precisión del modelo y bajando el error. Otro dato que se podría traer es el texto de las reseñas con sus respectivas valoraciones, pero todas las mejoras dichas ya entran en el área de NLP que incrementa la complejidad para trabajar los datos.

En cuestión de otros algoritmos que podrían servir para generar estas predicciones con los mismos datos no se me ocurren reemplazos directos.