SALES PREDICTION

Junior Rojas Fernandez - Data Analyst

Link: Exposición del video

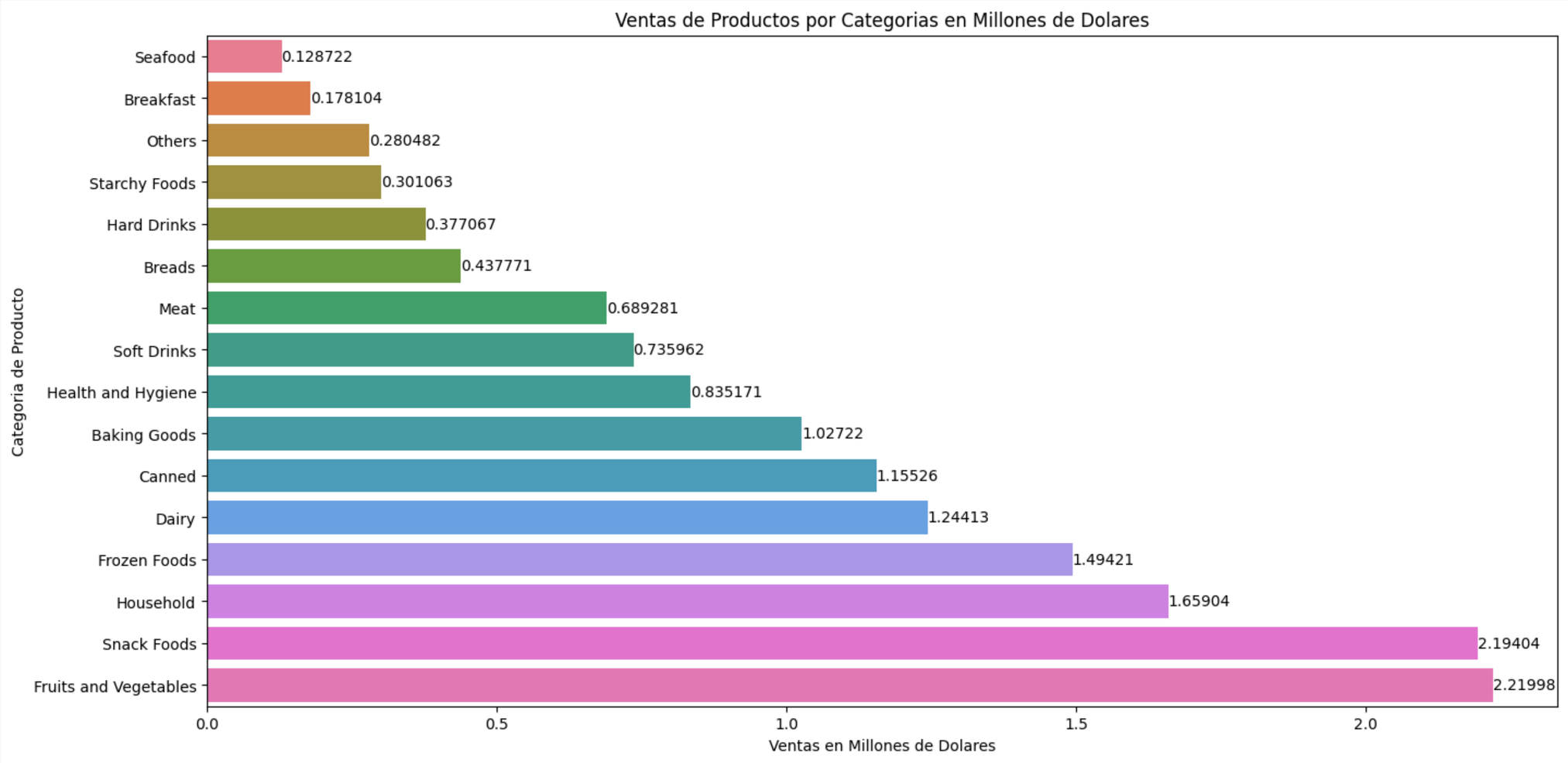
ABRIL/ 2024

Introducción

Nuestro conjunto de datos comprende diversas características relacionadas con artículos de venta y puntos de venta, como pesos de los artículos, visibilidad, tipos y detalles de los puntos de venta. El objetivo de este análisis es construir modelos de regresión que puedan predecir con precisión las ventas basadas en estas atribuciones.

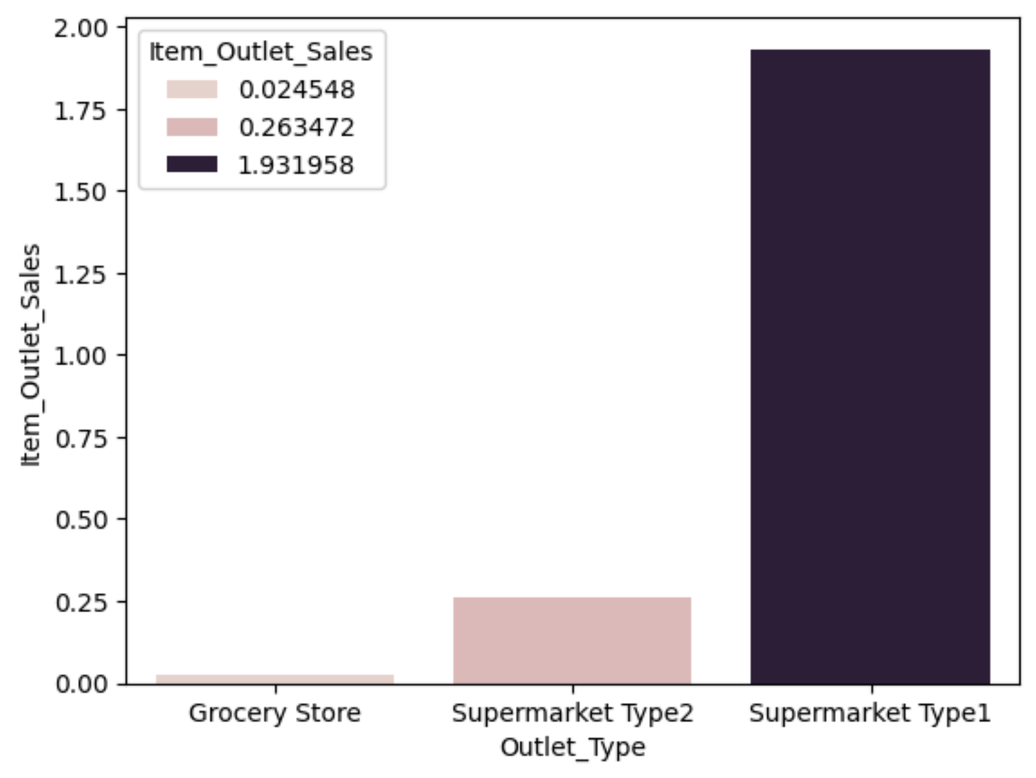
En esta presentación, compararemos el rendimiento de diferentes modelos de regresión, incluyendo Regresión Lineal, Vecinos más Cercanos (KNN), Bosques Aleatorios (Random Forest) y Árbol de Decisión (Decision Tree), basados en métricas como R^2 (coeficiente de determinación), MSE (error cuadrático medio) y RMSE (error cuadrático medio raíz). Nuestro objetivo es identificar el modelo más efectivo para predecir ventas en nuestro contexto minorista.

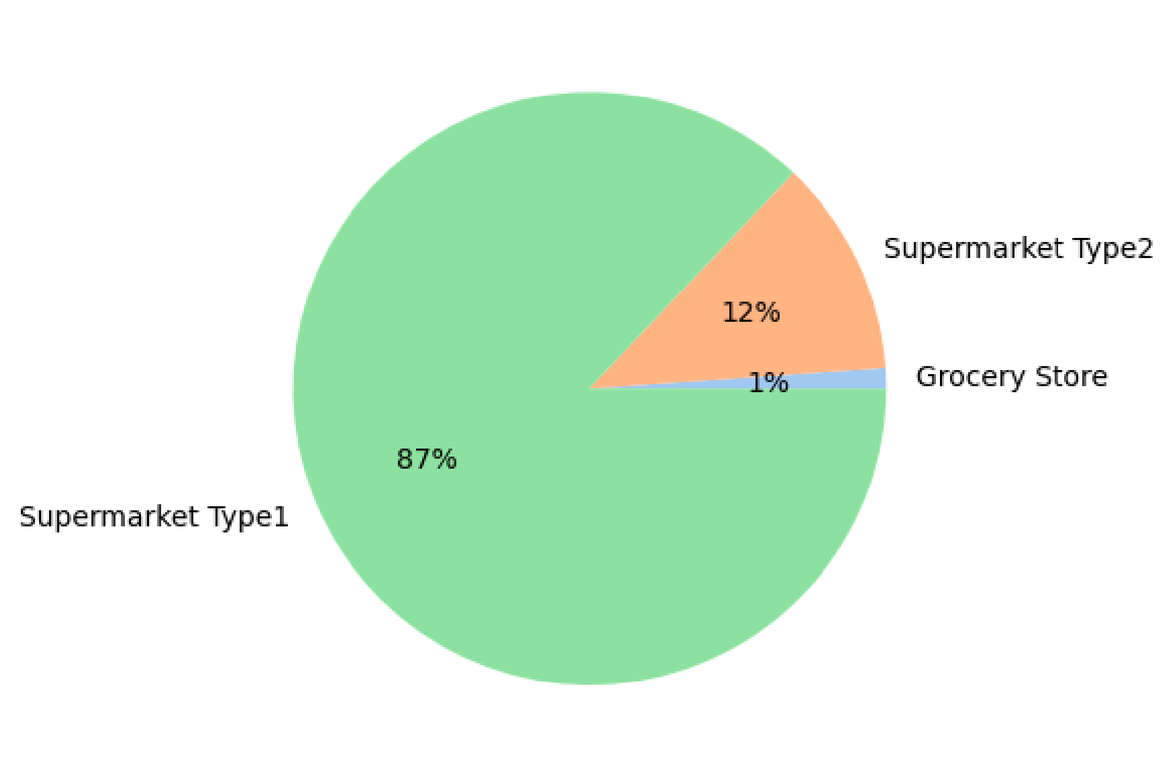
Tipos de Productos



Se observa que el top 3 de productos que mas ingresos trae son las fruta y vegetales, snack foods y Household. Se analizara cada uno de estos productos.

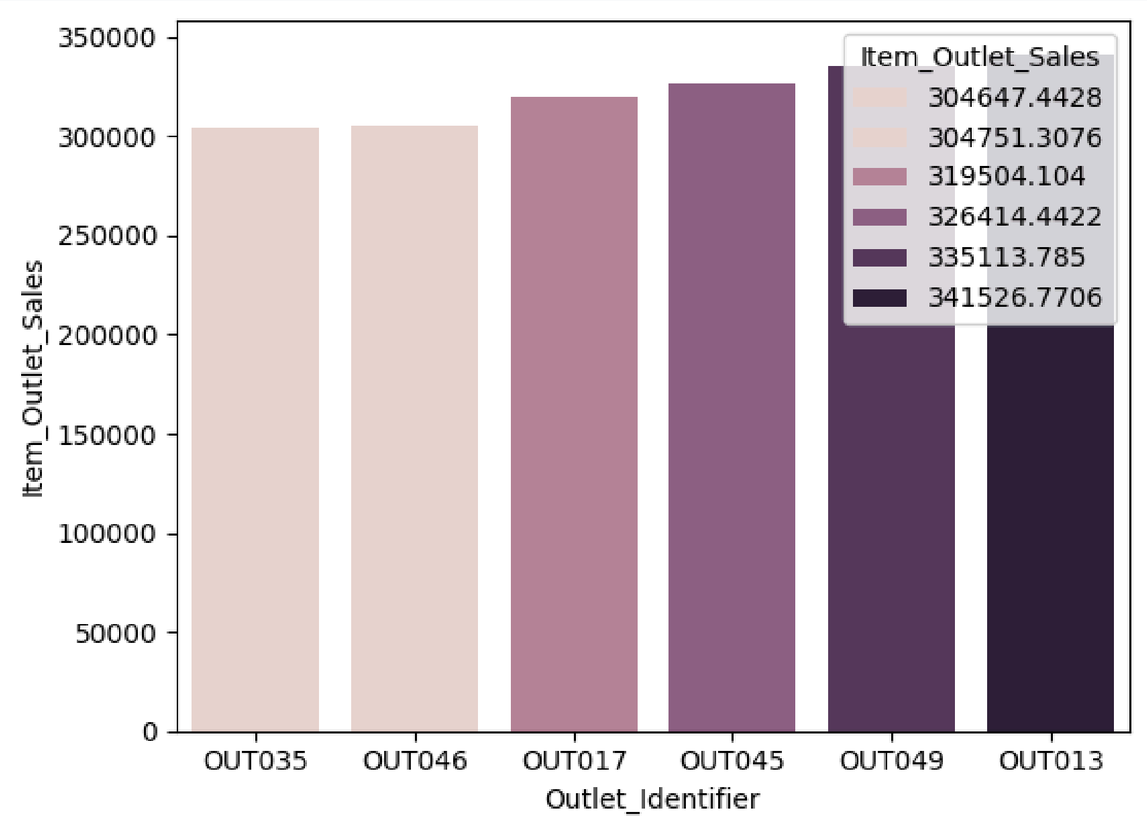
Cantidad de Mercado por Tipo de Local





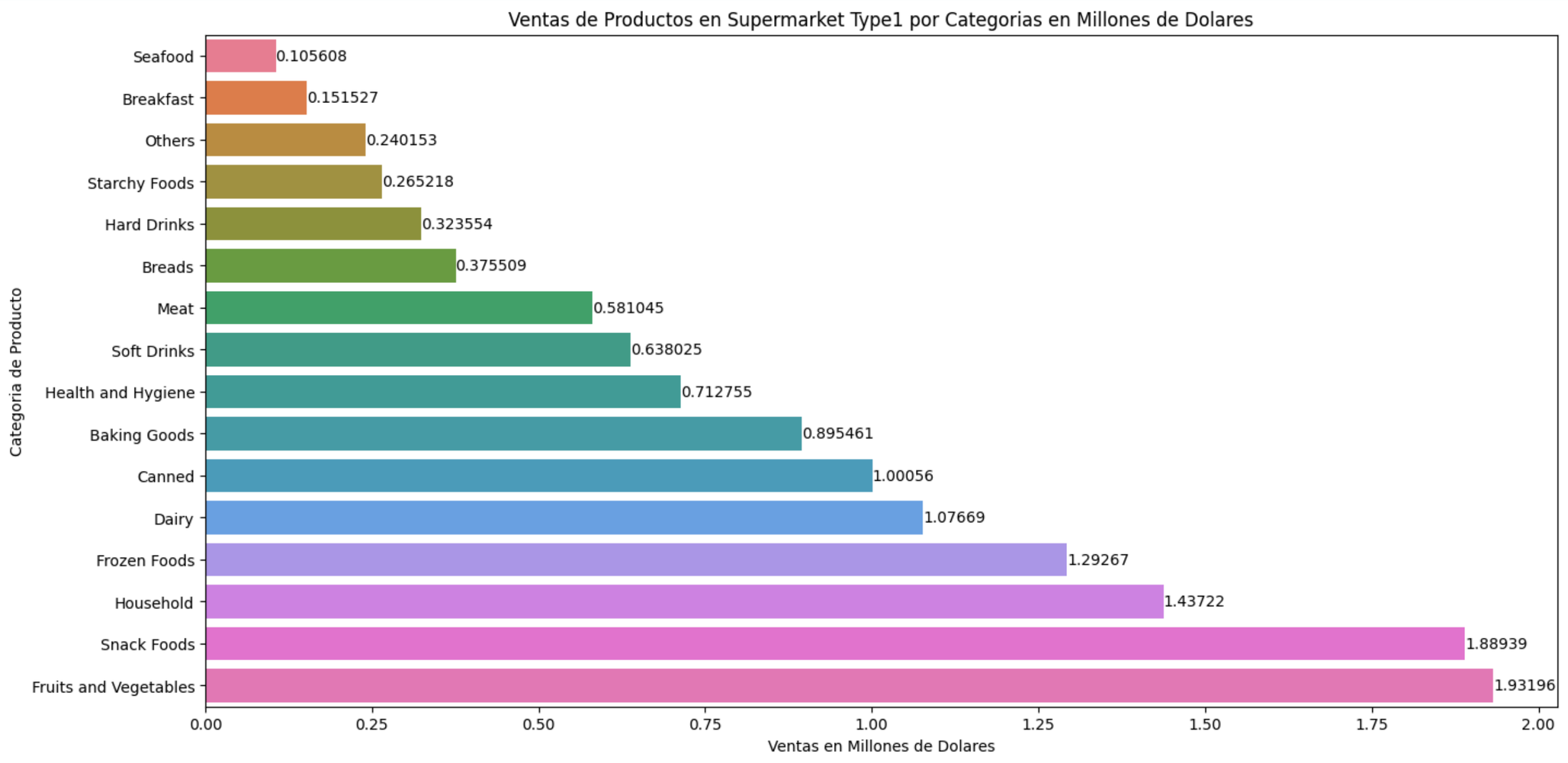
Podemos observar, que los supermercados tipo1 acaparan un 87% de las ventas respecto de los otros tipos de mercado.

Análisis de Supermercados Tipo 1



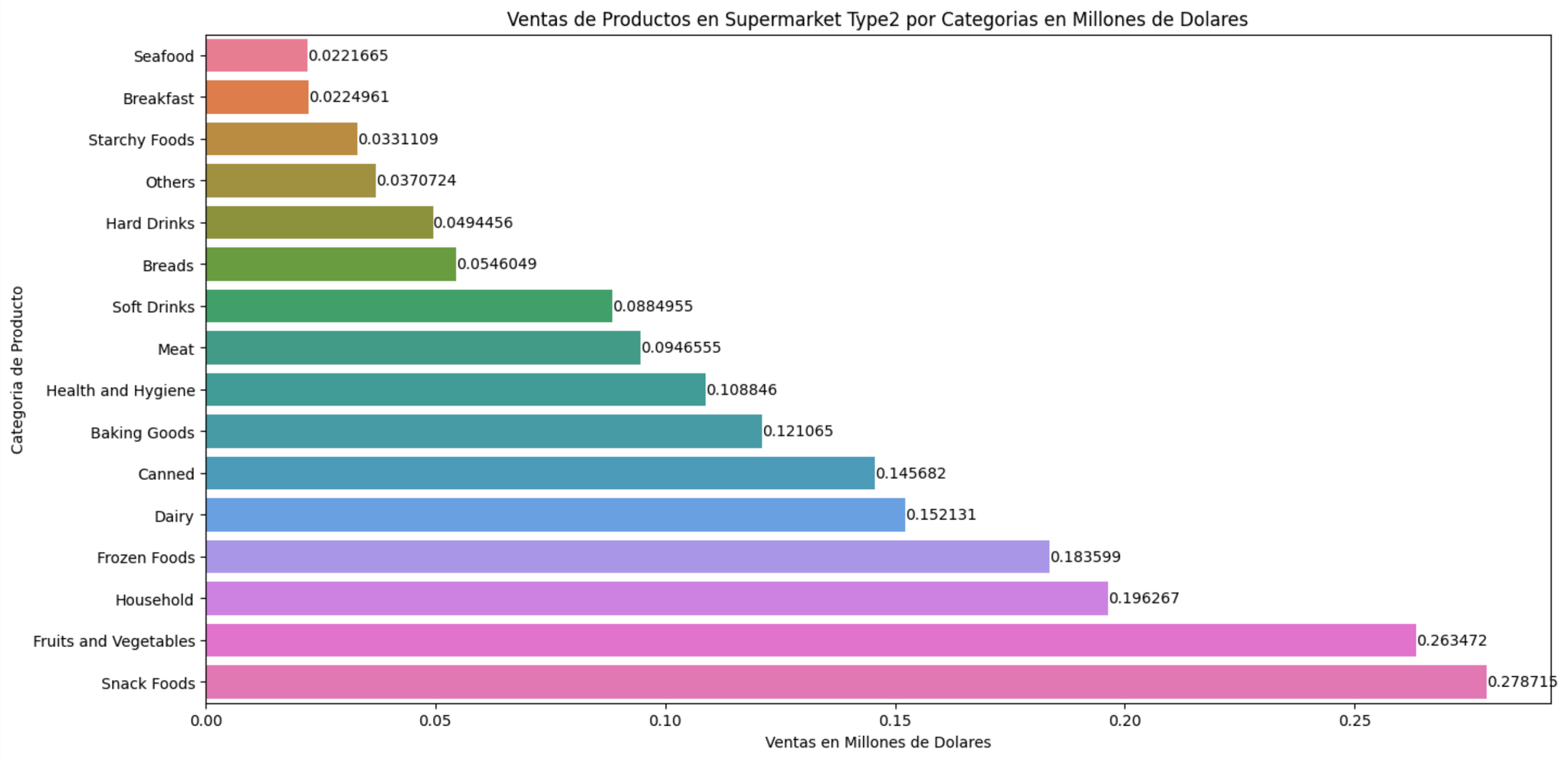
Podemos conluir que los supermerkados del tipo 1 presentan ventas similares para el tipo de producto verduras y frutas. Ademas, que este tipo de supermarket tipo 1 abarca el 87% del mercado de verduras y frutas.

Categorias en Supermarket Tipo 1



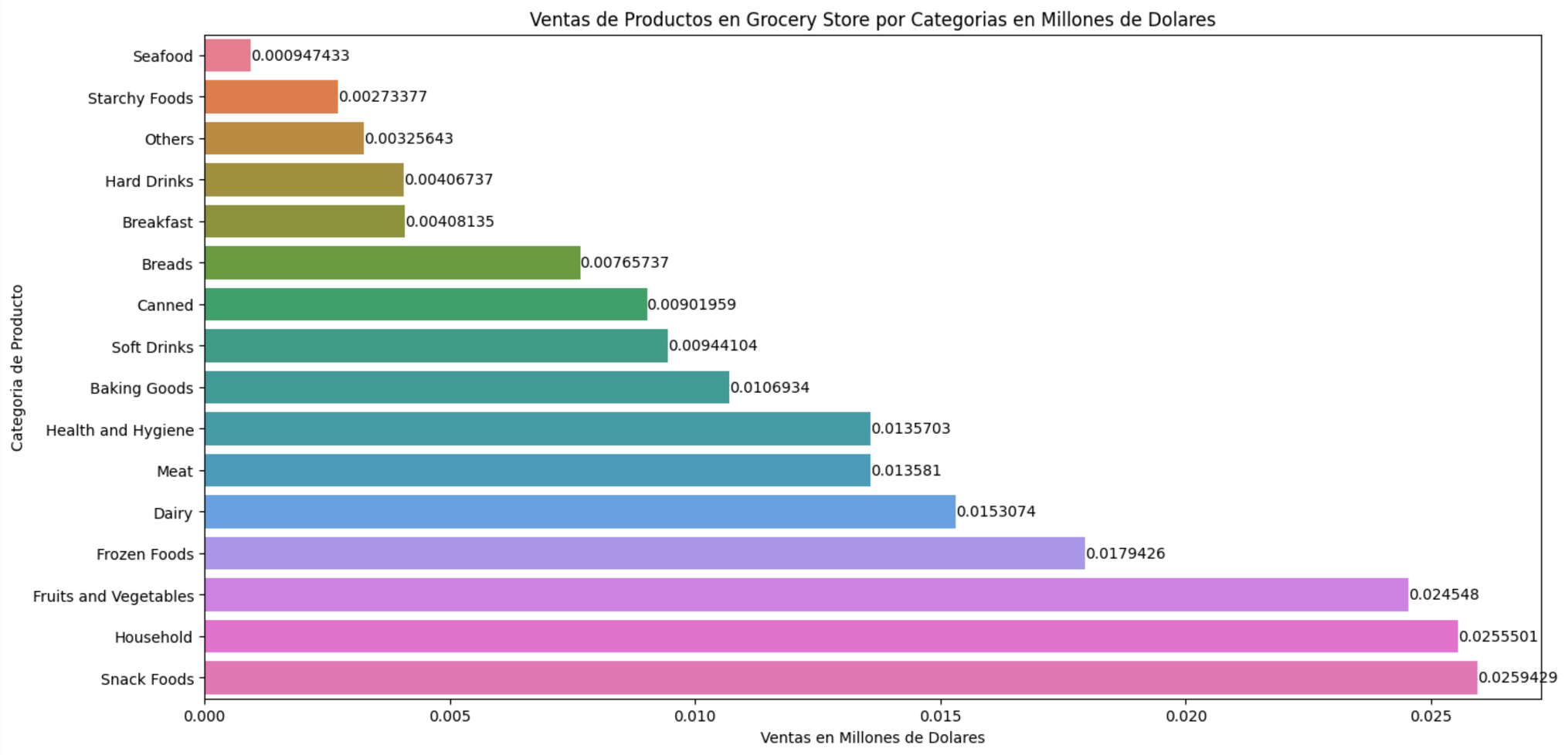
Se observa que el top 3 de productos que mas ingresos trae son las fruta y vegetales, snack foods y Household.

Categorias en Supermarket Tipo 2



Se observa que el top 3 de productos que mas ingresos trae son las fruta y vegetales, snack foods y Household.

Categorias en Grocery Store



Los 3 tipos de mercados presentan los mismos tipos de productos y el top 3 de productos es el mismo en los tres casos.

Predicción de ventas

Nuestro conjunto de datos abarca una variedad de características relacionadas con los artículos y los puntos de venta, que son fundamentales para comprender las tendencias y patrones que influencian las ventas. Como se ha observado que todos los tipos de supermarket comparten las mismas categorias de productos se ha intendado realizar una prediccion de las ventas con todos los datos.

Compararemos el rendimiento de diferentes modelos de regresión, incluyendo Regresión Lineal, Vecinos más Cercanos (KNN), Bosques Aleatorios (Random Forest) y Árbol de Decisión (Decision Tree), basados en métricas como R^2 (coeficiente de determinación), MSE (error cuadrático medio) y RMSE (error cuadrático medio raíz). Nuestro objetivo es identificar el modelo más efectivo para predecir ventas en nuestro contexto minorista.

Datos para el entrenamient

No.	Column	Non-Null Count	Dtype
0	Item_Weight	8519 non-null	float64
1	Item_Fat_Content	8519 non-null	object
2	Item_Visibility	8519 non-null	float64
3	Item_Type	8519 non-null	object
4	Item_MRP	8519 non-null	float64
5	Outlet_Establishment_Year	8519 non-null	int64
6	Outlet_Size	8519 non-null	object
7	Outlet_Location_Type	8519 non-null	object
8	Outlet_Type	8519 non-null	object
9	Item_Outlet_Sales	8519 non-null	float64

Datos que se tomaron en cuenta para el entrenamiento.

Resultados obtenidos

Modelo	R^2 Entrenamiento	R^2 Prueba	MSE	RMSE
KNN	0.768	0.760	813052.21	901.69
Random Forest (RF)	0.834	0.796	615540.74	784.56
Árbol de Decisión (DT)	0.817	0.755	746823.66	864.19
Regresión Lineal	0.776	0.793	618035.53	786.15

Resultado obtenidos con los diferentes modelos entrenados.

Conclusiones

- KNN: El modelo tiene un rendimiento decente, con un coeficiente de determinación R^2 de alrededor de 0.76 tanto en entrenamiento como en prueba. Sin embargo, el RMSE es relativamente alto, indicando que las predicciones pueden tener errores considerables en términos absolutos.

- RF: Los bosques aleatorios (Random Forest) muestran un mejor rendimiento que KNN, con un coeficiente de determinación R^2 de aproximadamente 0.80 en prueba y un RMSE menor que el de KNN, lo que sugiere predicciones más precisas en términos absolutos.

- DT: El árbol de decisión (Decision Tree) tiene un rendimiento similar al de KNN en términos de coeficiente de determinación R^2, pero con un RMSE ligeramente más bajo. Esto indica que puede tener un rendimiento algo mejor que KNN en términos de precisión de predicciones.

- LR: Buen rendimiento, con un coeficiente de determinación R^2 alto tanto en entrenamiento como en prueba, y un RMSE similar al de Random Forest.

Recomendaciones

- Realizar análisis de relevancia de características: Identificar columnas que no están contribuyendo significativamente al entrenamiento de nuestros modelos puede optimizar el proceso al eliminar variables irrelevantes o con bajo impacto en las predicciones, mejorando la eficiencia y precisión de nuestros modelos.
- Obtener más datos por local: Realizar una recolección adicional de datos específicos por cada local puede mejorar la precisión de nuestros modelos predictivos al considerar características y patrones únicos de cada ubicación.
- Optimizar la toma de datos: Modificar la metodología de captura de datos para asegurar una recopilación específica y detallada por periodos de tiempo, permitiendo así entrenar modelos predictivos más precisos y con mayor capacidad de generalización.

GRACIAS.

MARZO/ 2024