

TFG: Analítica de Clientes y Predicción de Demanda mediante Modelos de Machine Learning en el Sector Retail


Abel Mora Vázquez

Grado Ciencia de Datos Aplicada

Índice

 **Introducción**

 **Metodología**

 **Obtención de datos**

 **Análisis Exploratorio**

 **Procesamiento Lenguaje Natural**

 **Analítica de clientes**

 **Predicción de demanda**

 **Valoración y trabajos futuros**

Introducción

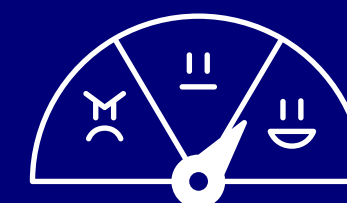
Sector Retail

El sector *retail* o **ventas al por menor** ofrece productos de manera directa al consumidor final. Los productos pasan del almacén a locales comerciales o digitales *e-commerce* donde el **clientx** los adquiere.



Ciencia de datos

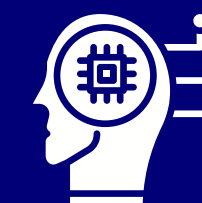
Obtiene de los **datos** recopilados *insights* relevantes para la empresa que permiten la optimización de stocks, incremento de beneficios, ajuste del precio a la demanda, la fidelización y satisfacción del consumidor y la captación de nuevxs clientxs.



Clientxs



Datos históricos



Predicción

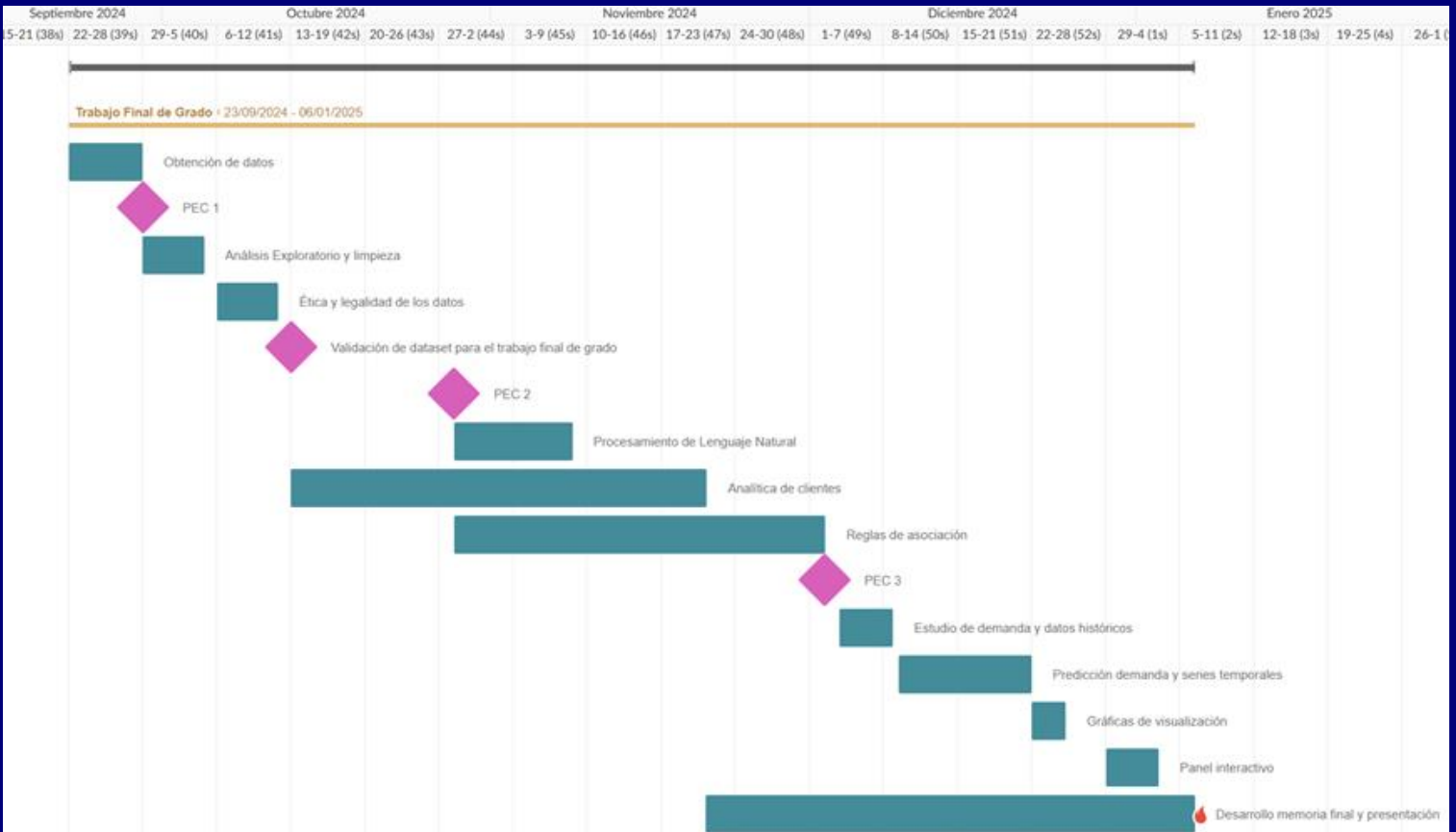
Introducción

Planificación

La planificación inicial ha sido distribuida en tareas a realizar incorporando posteriormente los hitos y actualizaciones.

Riesgos

La tabla de riesgos refleja los tipos de riesgos que pueden aparecer y junto a las medidas a tomar, forman la tabla de gestión de riesgos.



INTERNOS	EXTERNOS	TÉCNICOS
Datos insuficientes	Modificación sustancial de la situación laboral	Equipo técnico
Cumplimiento normativo	Modificación sustancial de salud propia familiar	Nuevas tecnologías

RECOGIDA Y MANIPULACIÓN

- Solicitud de datos a empresas
- Conjuntos de datos del sector retail
- Agrupación de datasets
- Preparación de datos

ANALÍTICA Y ALGORITMICA

- Análisis Exploratorio
- Procesamiento Lenguaje Natural
- Modelo de Clustering
- Reglas de asociación
- Series temporales

Metodología

ÉTICA Y LEGALIDAD

- Tratamiento de datos sensibles
- Cumplimiento normativo
- Análisis de impacto
- Lenguaje inclusivo

ARQUITECTURA Y VISUALIZACIÓN

- Memoria
- Presentación
- Presentación en vídeo
- Jupyter Notebook

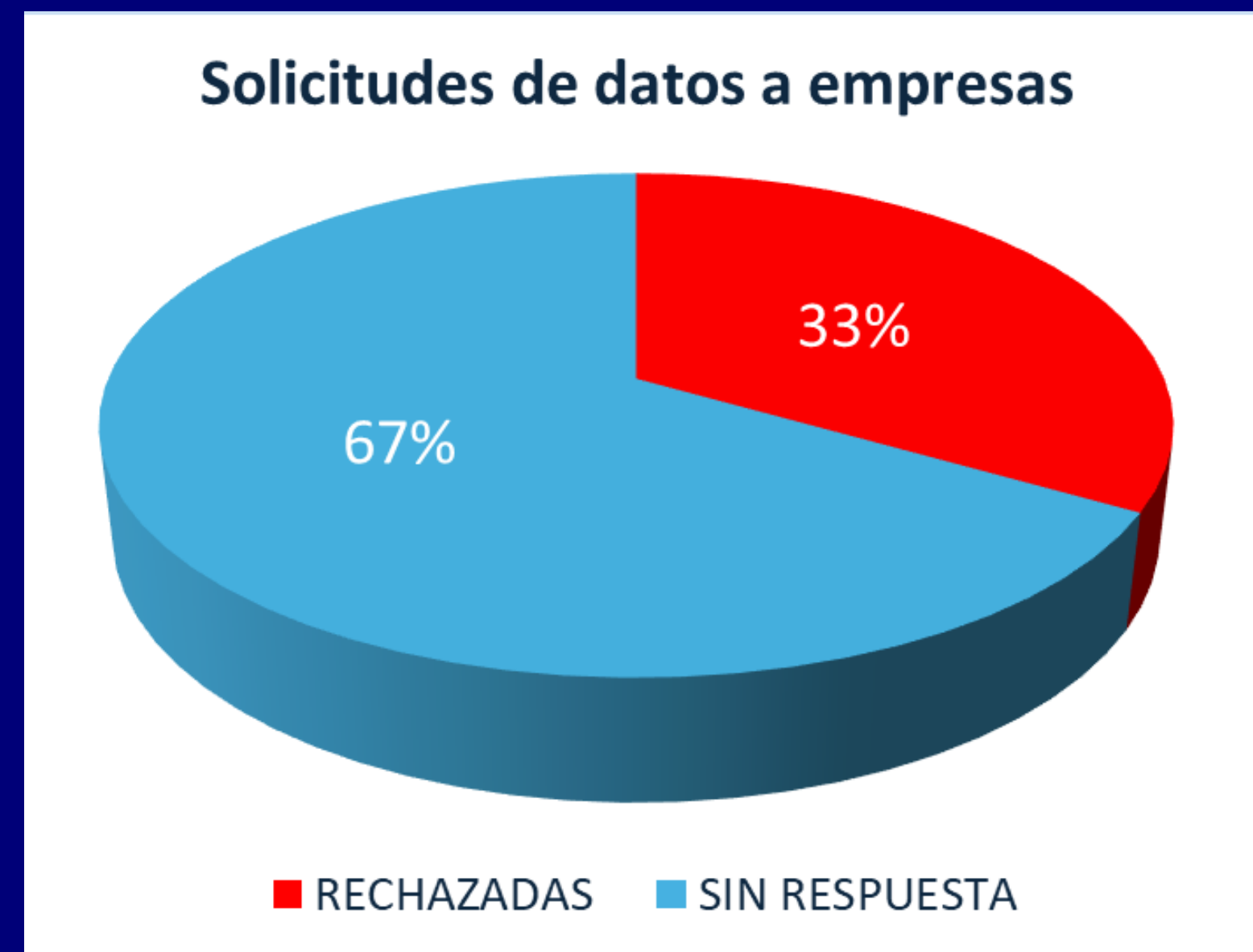
Obtención de datos

DATOS = VALOR

Las empresas ponen en valor los datos internos y son reacias a compartir datos debido a la información que se puede extraer de ellos y que son objetivo de este trabajo.

Las empresas han de cumplir con políticas de privacidad de datos aceptadas por los clientes.

Solicitudes realizadas a diferentes empresas en países y continentes distintos.



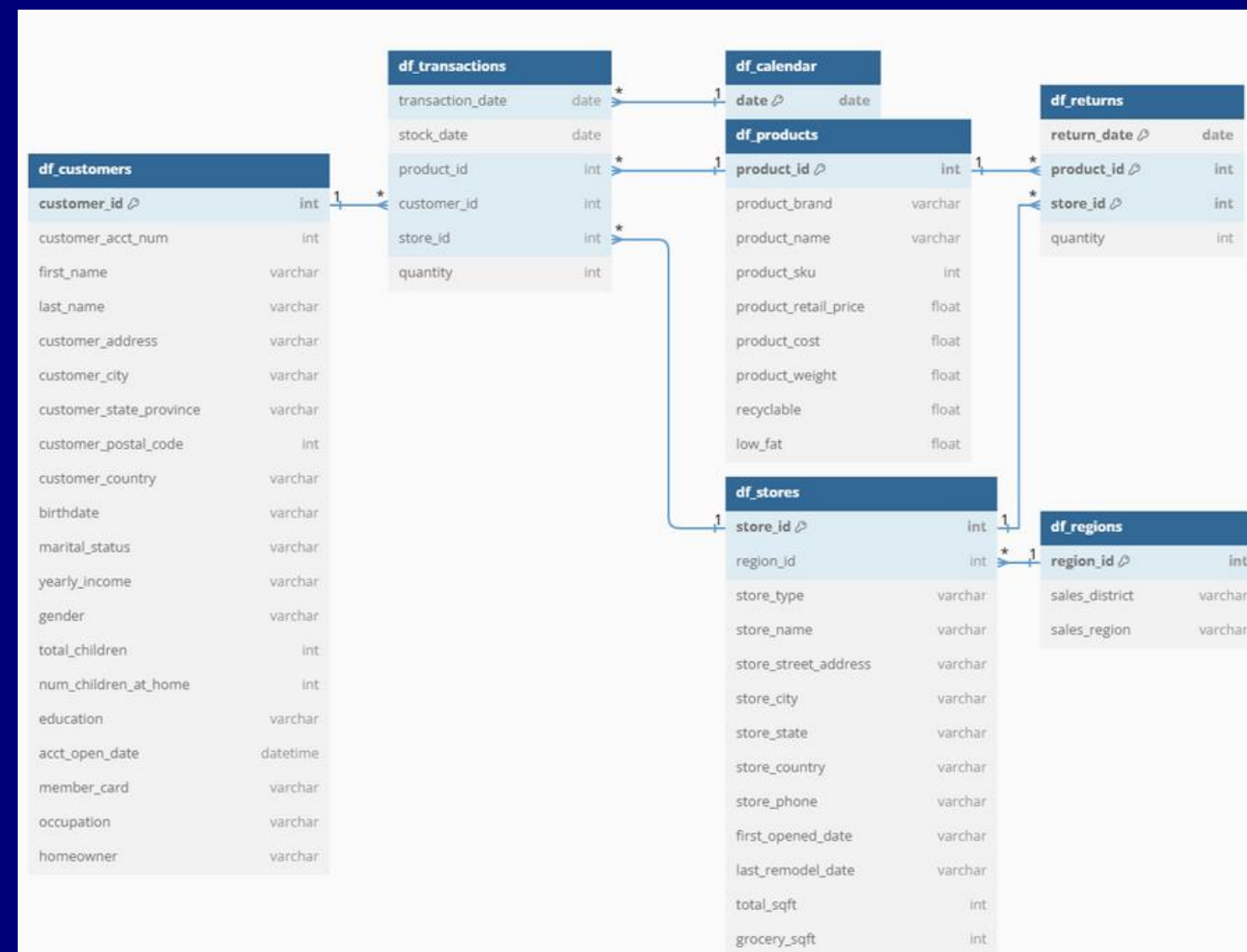
Obtención de datos

Datasets Maven Market

Conjunto de datasets de empresa del sector retail que opera distintos tipos de supermercados

Distribución en 8 datasets con variables comunes que permiten realizar la conexión entre ellos.

El contexto es la venta registrada durante los años 1997-1998 de clientes fidelizados mediante tarjeta de membresía.



Ética y legalidad

Sostenibilidad



Comportamiento ético Responsabilidad Social



Diversidad y derechos humanos



Análisis Exploratorio EDA

Características de los datasets: Validación de datos de cada dataset, número de filas, columnas, formato de los datos, existencia de valores nulos o duplicados.

Medidas de ética y legalidad: Tratar datos sensibles, adecuación de datos para minimizar impacto.

Investigación: Descubrir patrones o anomalías, revelar estructuras subyacentes, planteamiento de preguntas y líneas de desarrollo.



Procesamiento de Lenguaje Natural

Crear columna categoría de productos mediante el nombre del producto

Preprocesamiento de datos realizando limpieza y normalización de los datos

Análisis lingüístico con spaCy para obtener tokens y su etiqueta POS.

Clasificación semántica con modelo Zero-Shot de Hugging face mediante **BART** Bidirectional and Auto-Regresive Transformer

Clasificación manual mediante palabras clave de los productos no clasificados correctamente.

Categoría de productos				
Producto	Categoría	Producto	Categoría	Producto
Producto 1	Categoría 1	Producto 2	Categoría 2	Producto 3
Producto 4	Categoría 3	Producto 5	Categoría 4	Producto 6
Producto 7	Categoría 5	Producto 8	Categoría 6	Producto 9
Producto 10	Categoría 7	Producto 11	Categoría 8	Producto 12

Analítica de Clientxs

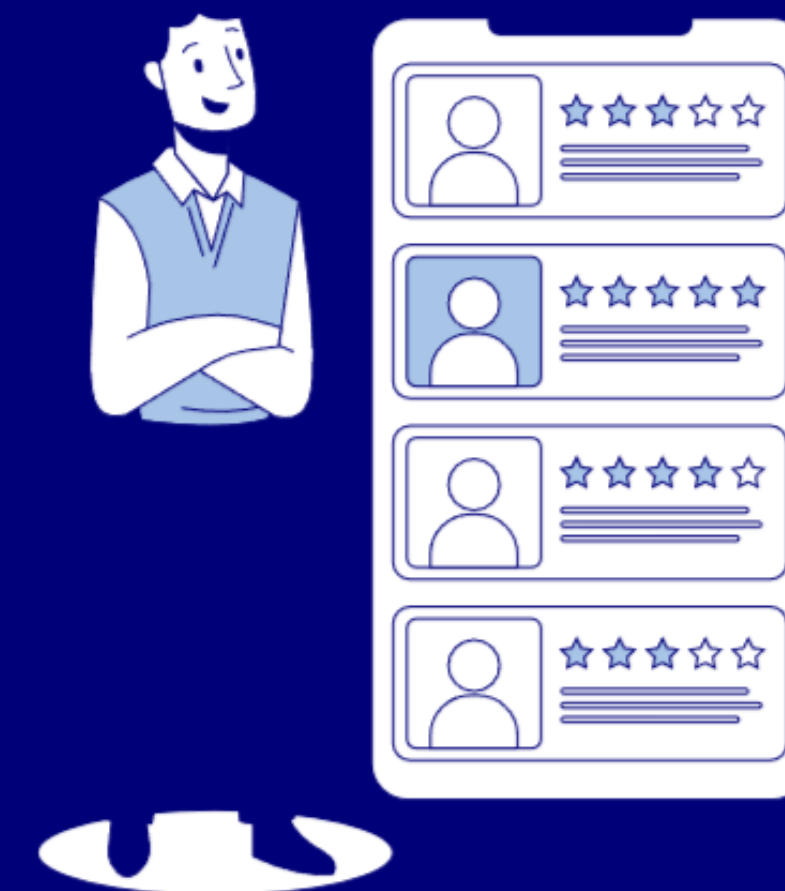
Entender el comportamiento de lxs clientxs facilitando la toma de decisiones a diferentes departamentos de la empresa.

Análisis sociodemográfico: Características de lxs clientxs, creación de nuevas variables y gráficas de distribución.

Perspectiva de género: Diferencias desigualdades entre géneros.

Segmentación de clientxs: Clusterizar consumidorxs con características similares mediante algoritmo K-means.

Reglas de asociación: Obtener combinación de productos frecuentes en las transacciones mediante algoritmo Apriori.



Analítica de Clientes

Silent Generation: Eliminar datos de clientx en caso de inoperatividad durante un periodo de 2 años.

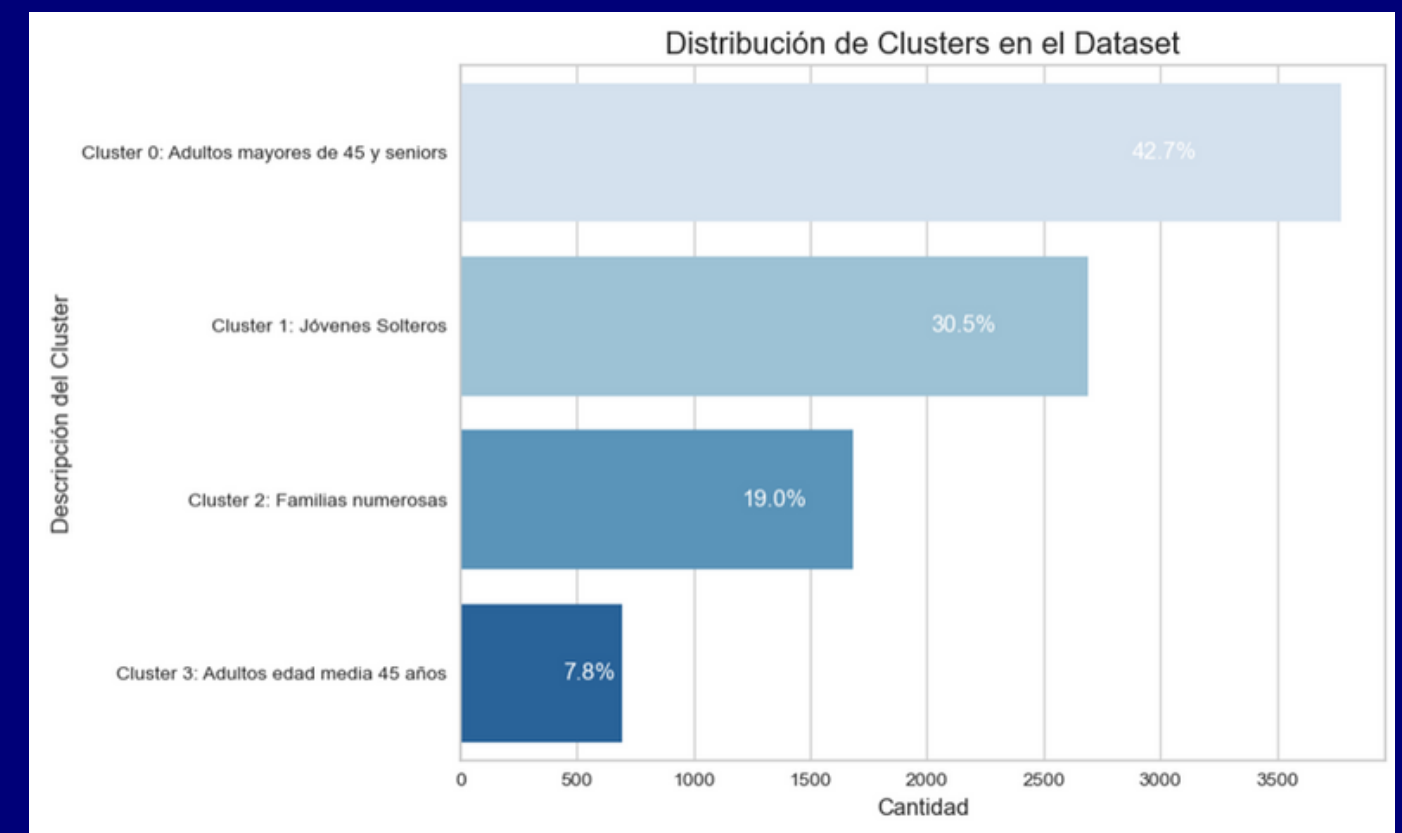
Baby Boomers: Campaña de recordatorio de identificación mediante tarjeta de membresía.
Incentivar el uso mediante descuentos.

Generación X: Campaña de marketing mediante redes sociales de beneficios de la membresía al comprar en los establecimientos.

Los datos sintéticos reflejan paridad en el análisis de la perspectiva de género.

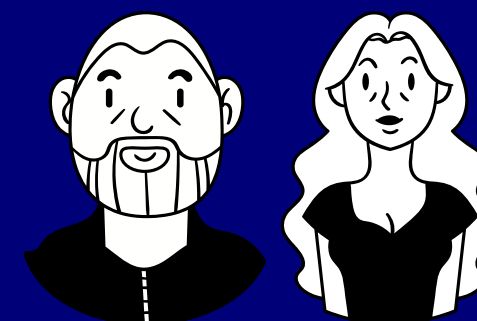
En base al análisis de los métodos de evaluación Elbow y Silhouette se escogen 4 clusters

Clientes que no han realizado compras por generación		
GENERACIÓN	NÚMERO DE CLIENTES	% RESPECTO A GENERACIÓN
Silent Generation	745	14.3%
Baby Boomers	387	13.5%
Generación X	307	13.8%



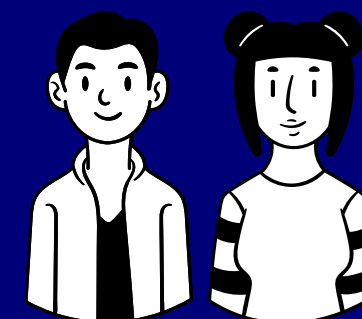
Adultxs mayores de 45 y seniors

Grupo de clientxs en edad previa a la jubilación y jubiladxs divorciadxs o viudxs y compras pequeñas.



Jóvenes Solterxs

Jóvenes solterxs realizan compras pequeñas.



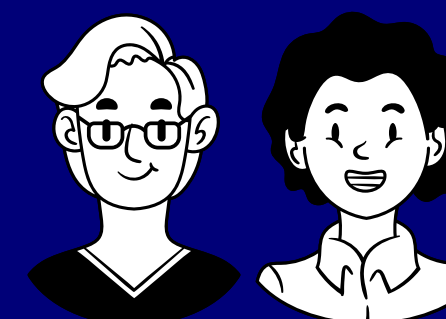
Familias numerosas

Familias con más de 3 hijxs realizan compras medias.



Adultxs

Adultxs alrededor de 45 años realizan compras grandes.

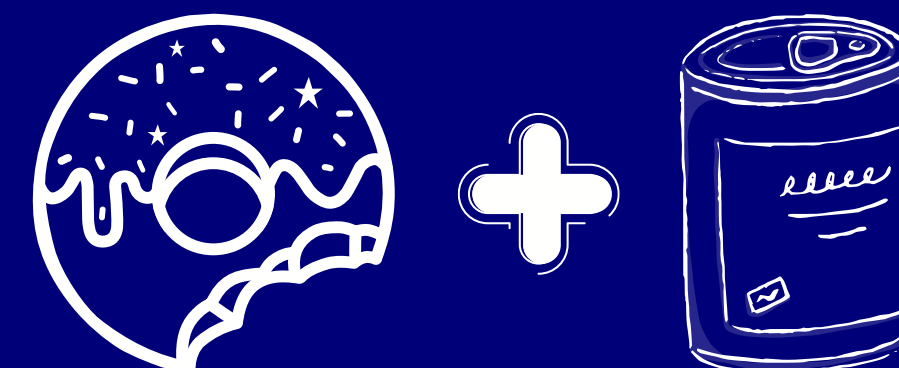


Las **reglas de asociación** reflejaron la falta de transacciones suficientes para obtener buenos resultados.

Regla 1 Bollería y Conservas

Aplicación posible: Necesidad venta de productos enlatados por cambio de formato.

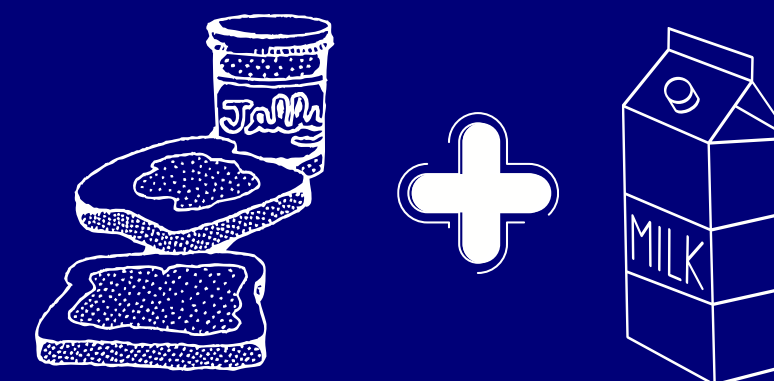
Realizar ofertas de productos de bollería incrementa la probabilidad de comprar productos enlatados.



Regla 2 Untables y Lácteos

Aplicación posible: Aumentar el volumen de ventas de productos lácteos debido a fecha de caducidad corta.

Ubicación en la misma zona que los productos untables.



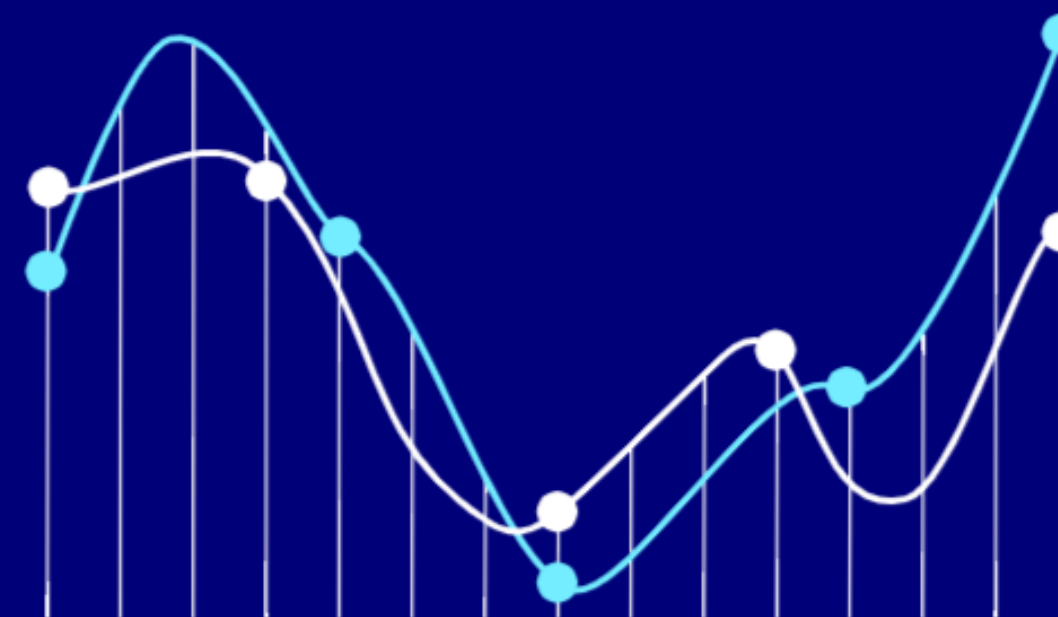
Series Temporales

Series temporales son un conjunto de datos medido en intervalo de tiempo regular y ordenado.

Las observaciones pasadas tienen influencia en las observaciones posteriores y futuras.

Permiten analizar tendencias, estacionalidades y ciclos en datos a lo largo del tiempo.

Disponer del mayor número de registros posibles para poder realizar una buena predicción.

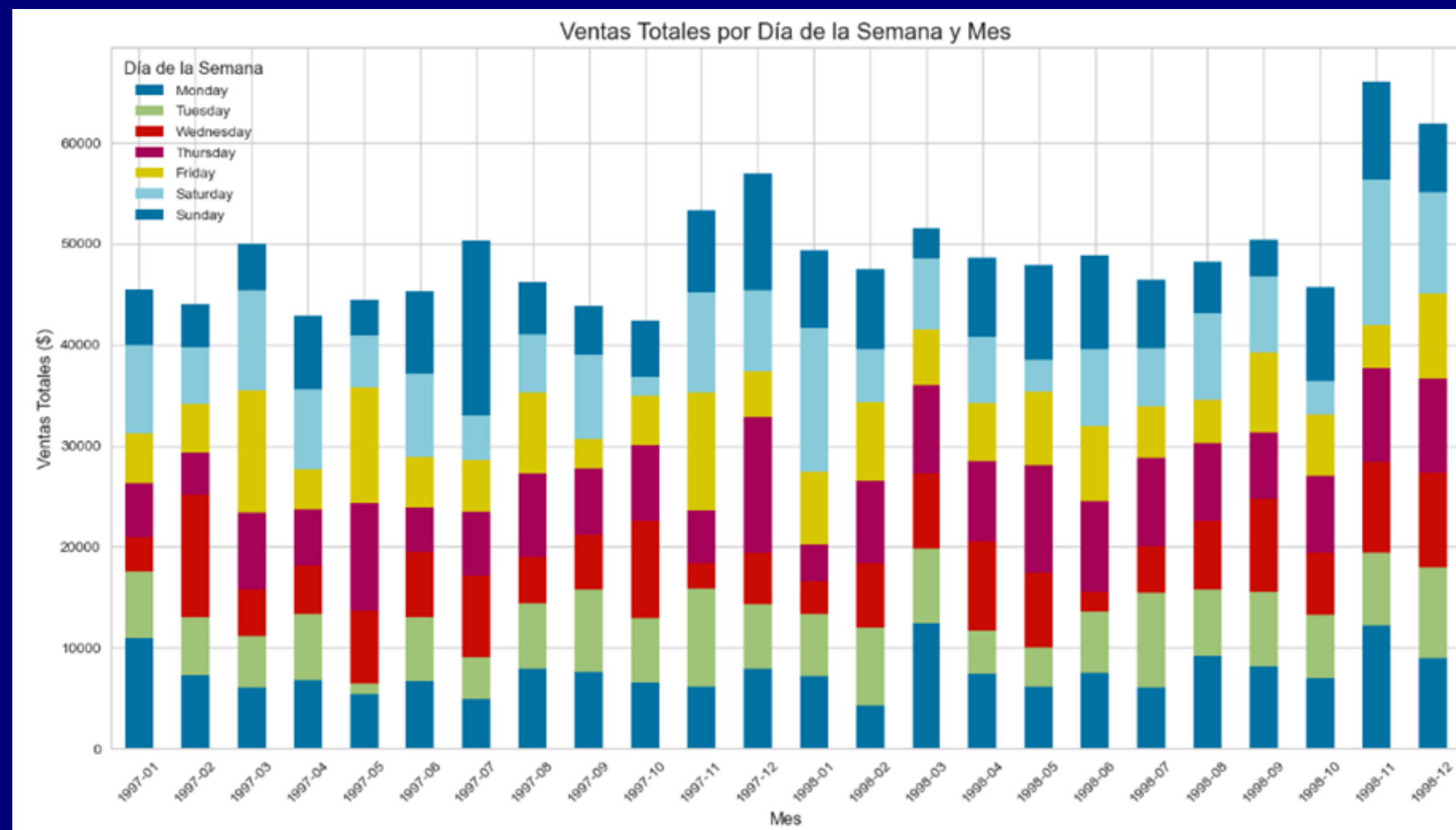


Series Temporales

Selección temporal de
supermercados USA

- Datos insuficientes
en **cantidad**
- Datos insuficientes
en **amplitud**

Imputación de datos
KNN neighbours



Predicción de demanda

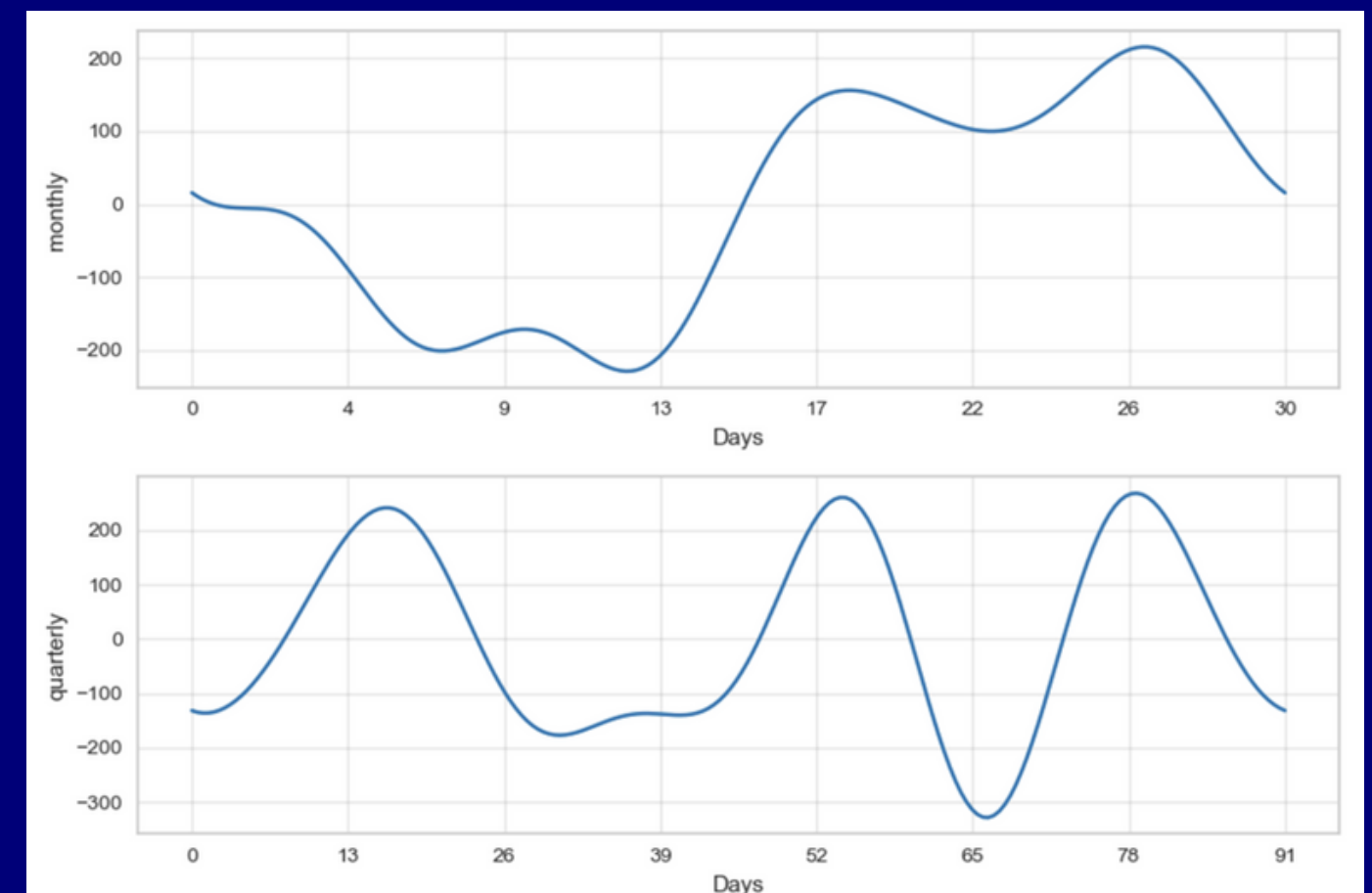
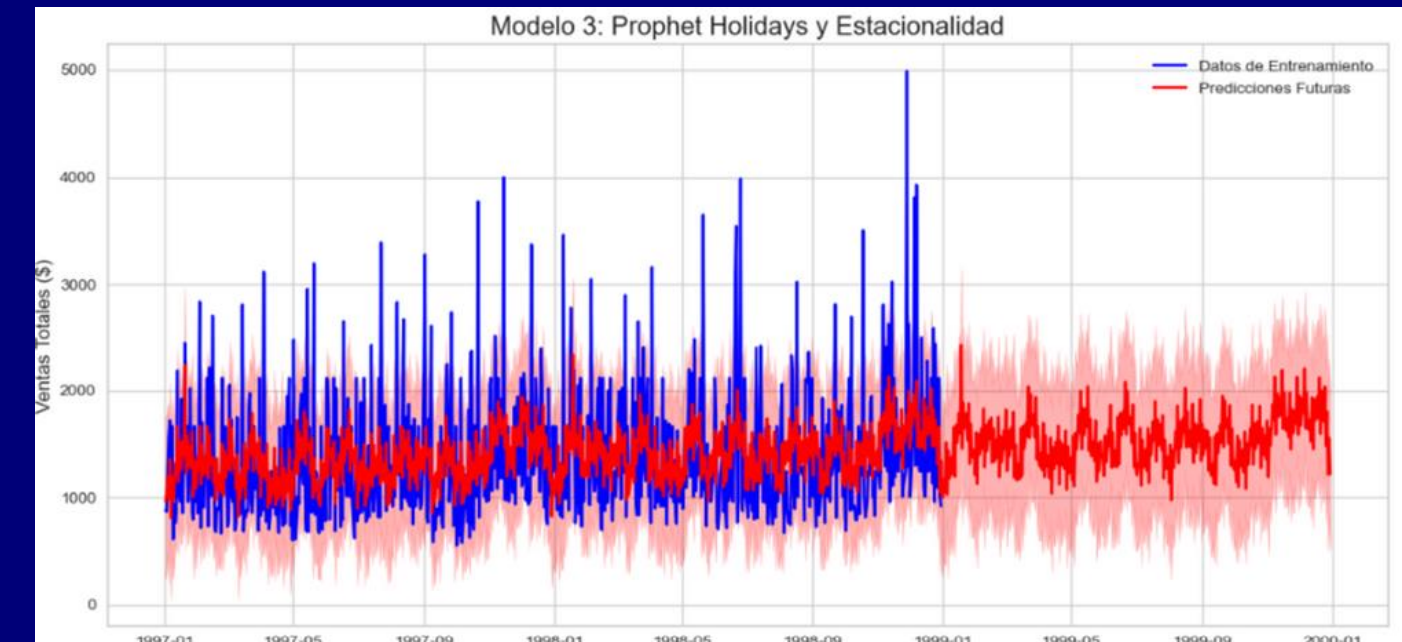
Modelo Prophet capta la estacionalidad y es robusto ante la falta de datos

Parametrización mediante diferentes estacionalidades, regresores y validaciones cruzadas

Métricas utilizadas para decidir el modelo con mejor desempeño RMSE y MAPE

Predicción de ventas con pendiente positiva

Porcentaje de errores elevado debido a las características de los datos



Valoración de resultados

Desarrollo de analítica de clientes

Minimización de impacto en sostenibilidad, comportamiento ético y diversidad

Análisis Exploratorio de datos en profundidad debido a falta de datos

Medidas de acción comerciales para incrementar el uso de tarjeta fidelización

Categorías de productos obtenidas mediante Procesamiento de lenguaje natural

Clusters de clientxs e identificación de clusters objetivo para aumento de venta

Reglas de asociación débiles detectando posibles casos de uso

Valoración de resultados

Desarrollo de predicción de demanda

Debido a la limitación de los datos disponibles los resultados no son los esperados

Modelos de Prophet para predicción de la facturación

Análisis de la estacionalidad y ciclicidad mermado debido a datos sintéticos que no reflejan consumos reales y la imputación de datos faltantes

Incapacidad de crear modelos de predicción de demanda por producto para optimizar stocks y el precio del producto que permita obtener mayor beneficio

Importancia de recopilar datos para las empresas

Trabajos futuros

- Clusterización de establecimientos cruzando datos sociodemográficos como la tasa de criminalidad e indicadores macroeconómicos
- Analítica de establecimientos para desarrollo de plan de remodelación optimizando la regla coste - beneficio
- Análisis de necesidades de los modelos para definir y mejorar las estrategias en la recogida de datos

GRACIAS POR SU ATENCIÓN

Autor: Abel Mora Vázquez

Tutor: Antonio Gutiérrez Blanco



UOC.universitat



@UOCuniversidad



UOCuniversitat