Maestría en Ciencia de Datos

Laboratorio de Implementación III

Trabajo Final

Docentes:

Gustavo Denicolay

Alumno

Lautaro Ruiz

10-08-2025

Contenido

Introducción

Este trabajo se desarrolló en el marco de la materia Laboratorio de Implementación III, de la Maestría en Ciencia de Datos. El objetivo principal fue construir y evaluar modelos predictivos para estimar la venta en toneladas (“tn”) de 780 productos durante febrero de 2020. El desafío incluyó la participación en una competencia de Kaggle, donde se midió el desempeño de las predicciones generadas por distintas técnicas.

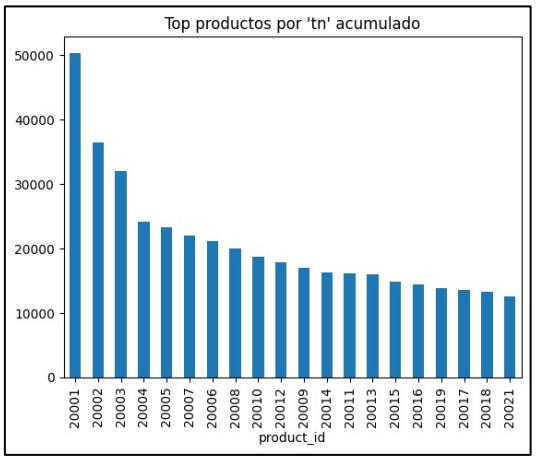
La problemática abordada es relevante para la industria, ya que permite anticipar la demanda y optimizar la gestión de inventarios, logística y producción. El dataset utilizado refleja un escenario real, con retos asociados a la calidad, granularidad y volumen de datos, y requirió integrar técnicas estadísticas y de machine learning.

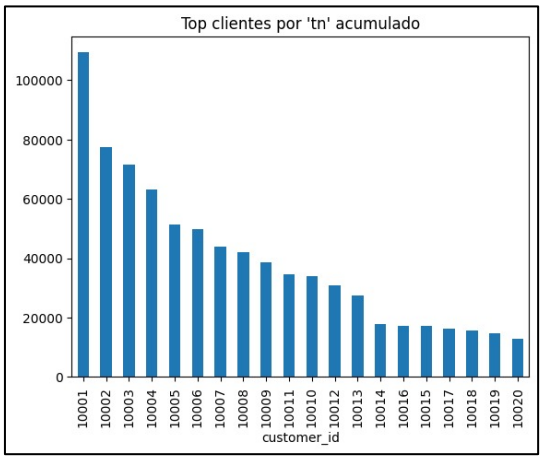
Descripción del Dataset Original

El dataset original cuenta con aproximadamente 3 millones de registros, con información a nivel producto, cliente y periodo (mes-año) desde enero de 2017 hasta diciembre de 2019. La variable de interés es la cantidad vendida en toneladas (“tn”).

También se considera un catálogo de productos que incluye variables categóricas como “cat1”, “cat2”, “cat3”, “brand” y la descripción del producto (“descripcion”), permitiendo una segmentación de productos que podría ayudar durante la etapa del modelado.

El análisis exploratorio mostró un marcado desbalance de la variable “tn” tanto para los productos como para los clientes: el 20% de las toneladas totales provienen de solo 10 productos (de un total de 1.233), y el 45% de las toneladas son aportadas por solo 10 clientes (de 590).



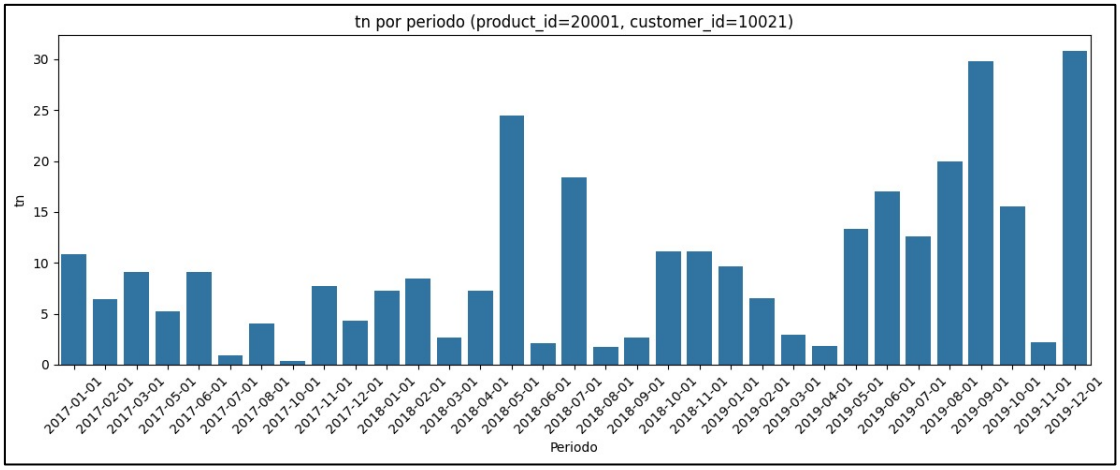


Por otro lado, en las series temporales de toneladas vendidas por cliente y producto, se observó que cuando un cliente no compraba un producto en un mes determinado, no existía un registro de dicha transacción. Esto genera dos situaciones:

1. Un cliente suele comprar un producto, pero en ciertos meses no lo hace (ausencia temporal).
2. Un cliente nunca compra un producto, aunque sí compra otros productos ese mes (ausencia estructural).

Estas observaciones son claves para el tratamiento de los datos y la construcción de los modelos.

Por ejemplo, esto puede apreciarse en el caso del cliente 10021, que no realizó compras del producto 20001 en diciembre de 2018, ni en otros meses de 2017.



Técnicas aplicadas (base de análisis a nivel producto)

Para estimar la demanda de febrero de 2020, se probaron técnicas que requerían definir el nivel de agregación de los datos.

Algunas de ellas necesitaban una base consolidada a nivel de producto, es decir, agrupando las toneladas vendidas de todos los clientes para cada período. Estas técnicas son:

a) Medias móviles

Esta técnica es robusta y fácil de implementar, pero no captura tendencia ni estacionalidades complejas.

Se calcularon las medias de los últimos n meses (siendo n = 3, 12, 24 y 32) para estimar la demanda de toneladas futura. El mejor resultado se obtuvo utilizando la media de los últimos 12 meses.

b) Auto-ARIMA

Se ajustaron modelos de series de tiempo para cada producto, buscando mejorar la predicción respecto a la media móvil. Sin embargo, los resultados no superaron a las medias.

c) Regresión Lineal para productos “mágicos”

Se identificaron productos con un comportamiento específico (llamados mágicos) y se ajustó un modelo de regresión lineal (sin intercepto) a partir de los datos de 2018 para predecir 201902.

Luego, los coeficientes obtenidos se utilizaron para predecir las toneladas totales de los productos en 202002.

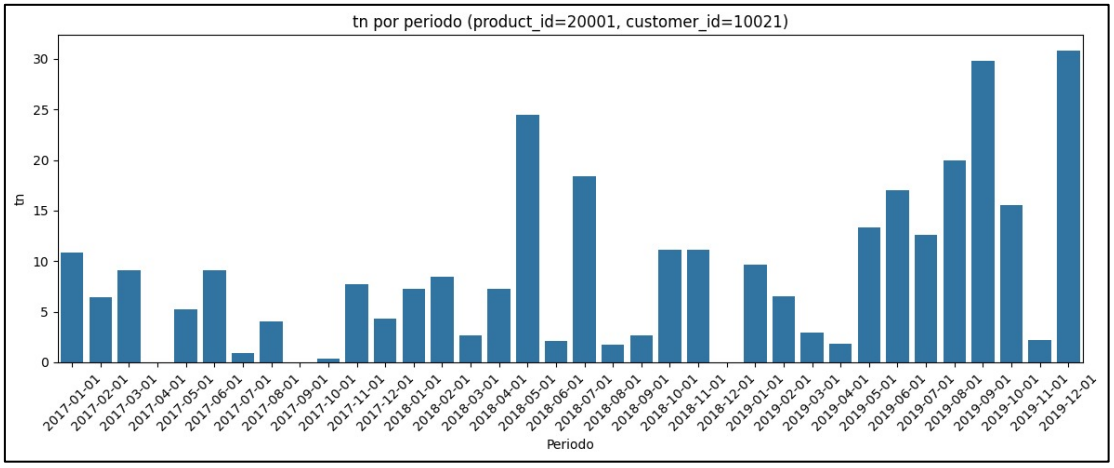
Finalmente, solo se tomaron estas predicciones para los productos mágicos, mientras que para los productos “no mágicos” se consideró como predicción la media móvil de los últimos 12 meses.
Este modelo es más importante a delante ya que forma parte de la solución final, por lo que será referenciado como “regresión lineal con los productos mágicos”.

Técnicas aplicadas (base de análisis a nivel producto-cliente)

Se realizó un modelado más complejo mediante la técnica LightGBM (LGBM). Para este, se requirió un procesamiento más extenso y sofisticado, reflejado en las notebooks de “armado de la base completa”, “feature engineering”, “post procesamiento” y “modelado”.

a) Armado de la base completa

Se completaron con ceros los registros ausentes para indicar que un cliente no compró un producto en un mes determinado. Además, se incluyeron los períodos a predecir para facilitar la posterior ingeniería de características (feature engineering) y la división del dataset.
De esta forma, el conjunto de datos original pasó de 3 millones de registros a aproximadamente 20 millones.
Este paso fue clave para que el modelo pudiera aprender tanto de compras como de ausencias.
Se puede observar entonces como se completa la serie para el cliente del ejemplo anterior (cliente 10021), que no realiza compras del producto 20001 en diciembre 2018, ni en otros meses del 2017.



b) Feature Engineering (FE)

Se generaron 72 nuevas variables, incluyendo lags, medias móviles, tendencias, variables temporales (“mes”, “trimestre”, “año”), variables externas (series de ipc y dólar), y variables de comportamiento histórico por cliente y producto. El proceso de FE fue iterativo y se validó el aporte de cada variable mediante un análisis de importancia.

c) Definición del target

La variable objetivo (“target”) se definió como la diferencia entre las toneladas vendidas en el período actual (t) y las vendidas dos períodos antes (t-2), es decir:

$target = tn_t - tn_{t-2}$.

d) Separación del conjunto de datos para el entrenamiento y validación del modelo

Para realizar el entrenamiento del modelo se tomaron en consideración los datos de los periodos 201708 hasta 201906, y se validaron con los periodos desde 201907 hasta 201912.

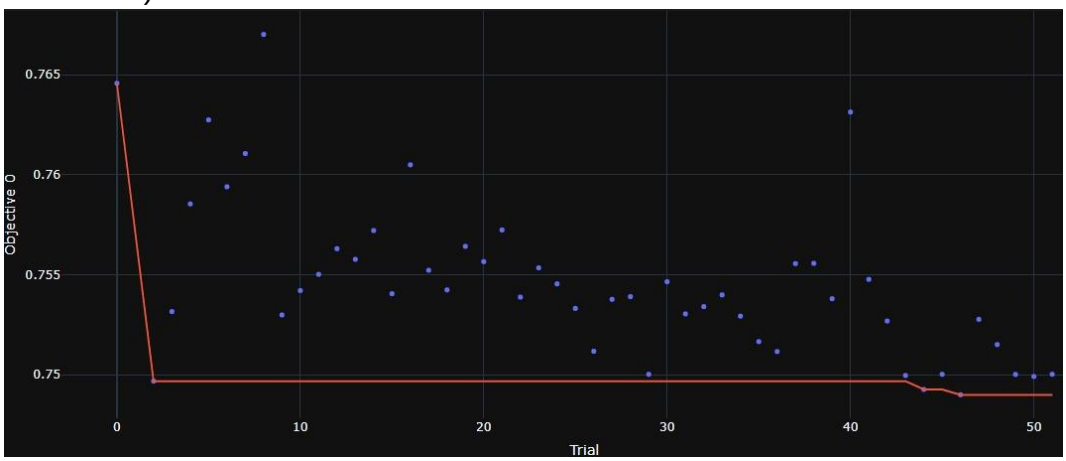
e) Optimización de hiperparámetros con optuna

Se optimizaron hiperparámetros claves como learning rate, número de hojas, profundidad máxima, entre otros. Se emplearon diferentes semillas para robustecer el ajuste y controlar la aleatoriedad del modelo ajustado con los mismos hiperparámetros.

A partir de pruebas realizadas, se consideró la siguiente configuración inicial de hiperparámetros:

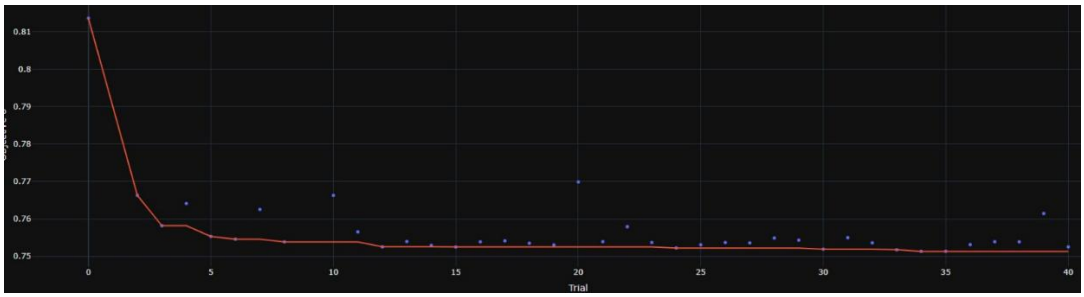
- 'learning_rate': 0.01687419508606365
- 'num_leaves': 123
- 'max_depth': 11
- 'colsample_bytree': 0.8290262199183336
- 'subsample': 0.721454421913562
- 'min_child_samples': 55
- 'reg_alpha': 0.6187094704138628
- 'reg_lambda': 0.795407032102261
- 'n_estimators': 551

El uso de optuna permitió automatizar la búsqueda y mejorar el desempeño del modelo en la base de validación, como puede verse en la salida de los ajustes (con optuna-dashboard):



Se puede observar que la “mejor” configuración de hiperparámetros surge en la segunda iteración del proceso.

Previo a este, se habían realizado otras optimizaciones como se puede ver a continuación que permitieron llegar al conjunto inicial definido más arriba:

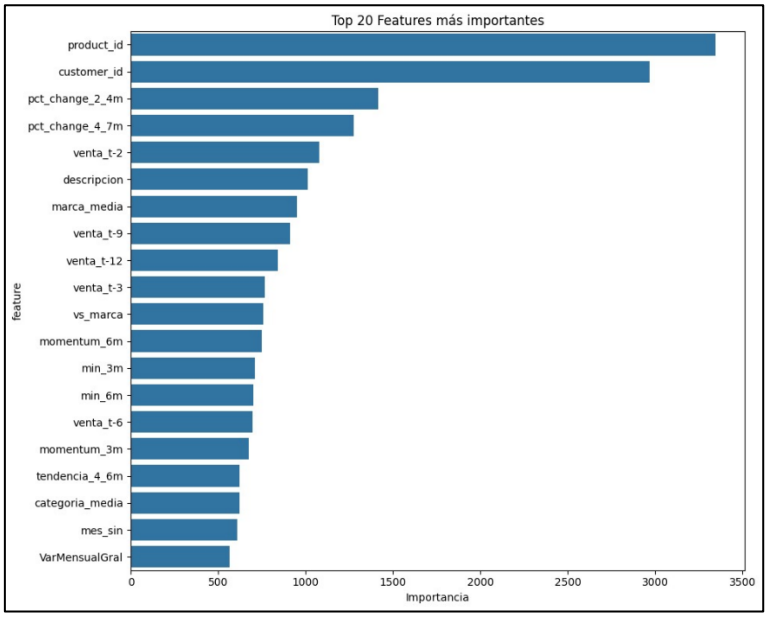


La configuración final de los hiperparámetros es:

- 'learning_rate': 0.033298202348152534
- 'num_leaves': 121
- 'max_depth': 12
- 'colsample_bytree': 0.7702280076072465
- 'subsample': 0.9451048961270575
- 'min_child_samples': 63
- 'reg_alpha': 0.8261485784018764
- 'reg_lambda': 0.2801548449048477
- 'n_estimators': 335

f) Importancia de variables

Los modelos LGBM ajustados permitieron realizar un análisis importancia de las variables consideradas para el entrenamiento, destacando la importancia de los indicadores de producto y cliente. A su vez, entre las cinco variables más importantes se encontraron los cambios porcentuales en las ventas entre los meses t-2 y t-4, y entre t-4 y t-7, así como el total de toneladas vendidas en los dos meses previos (t-2).



g) Ajuste del modelo final y alternativas aplicadas

A partir de la “mejor” configuración de hiperparámetros, se entrenaron múltiples modelos LGBM con diferentes semillas y se realizó un ensamble de los mismos para mejorar la estabilidad y el desempeño del modelo.

Además, se probaron distintas alternativas como aplicar LGBM a todos los productos, o solo al top 20% según volumen total de toneladas y combinando con las técnicas más simples probadas para el resto de los productos.

La alternativa con mejor desempeño consistió en utilizar el modelo de LGBM para top de productos que acumulan el 20% de las toneladas totales, y para el resto considerar las predicciones del modelo de “regresión lineal con los productos mágicos”.

Limitaciones

Algunas de las limitaciones que podrían identificarse en cuanto al desarrollo del trabajo:

- Selección de la métrica de optimización: La optimización con Optuna se realizó con una métrica distinta a la utilizada en la competencia de Kaggle. Si bien se hicieron pruebas preliminares cambiando la métrica, estas fueron poco exhaustivas y no mostraron mejoras. Una evaluación más profunda podría haber permitido un mejor alineamiento con el objetivo final.
- Definición de los conjuntos de validación: El conjunto de validación se definió de forma fija (201907–201912), sin evaluar múltiples ventanas temporales ni enfoques de validación tipo rolling window. Esto puede haber limitado la capacidad del modelo para generalizar a periodos futuros.
- Uso de ponderaciones en el entrenamiento: Se realizaron pruebas simples utilizando sample_weight (sin mejoras), pero no se exploraron en profundidad esquemas de ponderación que favorezcan productos o clientes con mayor relevancia estratégica.
- Cobertura de técnicas adicionales: No se incluyeron técnicas basadas en redes neuronales recurrentes (LSTM/GRU) ni enfoques híbridos de series temporales, que podrían haber capturado patrones no lineales complejos.
- Factores externos no integrados: Algunas variables macroeconómicas o de mercado adicionales (competencia, clima, feriados específicos) no fueron incluidas.

Conclusiones

El desarrollo de este trabajo permitió explorar y comparar diversas técnicas de predicción de la demanda, desde enfoques simples como medias móviles y regresiones lineales específicas, hasta modelos de gradient boosting como LightGBM optimizados con Optuna.

Si bien los modelos más complejos ofrecieron cierta mejora en métricas internas, en la competencia de Kaggle las diferencias respecto a enfoques más simples fueron menores a lo esperado. Esto sugiere que, en este problema particular, la complejidad del modelo no necesariamente se traduce en un beneficio proporcional en desempeño, probablemente debido a la naturaleza y calidad de los datos.

La estrategia con mejor desempeño resultó ser un enfoque híbrido, combinando las predicciones de un modelo LGBM, para el top 20% de productos por ventas en toneladas, con las del modelo de “regresión lineal con los productos mágicos”.