Escribe aquí
tu frase favorita.

E indica aquí su autor

Agradecimientos

Me gustaría agradecer...

También quiero destacar...

Por último...

Índice general

Índice de figuras

Índice de tablas

Índice de algoritmos

Resumen

Este Trabajo Fin de Máster aborda el problema de predicción de ventas semanales en el sector minorista a partir de la serie histórica de ventas y un conjunto de variables exógenas (calendario, festivos e indicadores macroeconómicos). El objetivo principal es diseñar un protocolo experimental reproducible y comparar, bajo condiciones homogéneas, varios enfoques de modelado capaces de incorporar covariables externas: modelos estadísticos (SARIMAX), un modelo aditivo a escala (Prophet), un enfoque probabilístico basado en RNN (DeepAR) y dos modelos neuronales globales (LSTM y Transformer). Para garantizar una comparación justa, se establecen particiones temporales coherentes y un proceso de ingeniería de características causal que evita fuga de información (lags y estadísticos móviles construidos exclusivamente con pasado). La evaluación se realiza mediante MAE, RMSE y sMAPE, reportando tanto resultados globales como por tienda.

Los resultados muestran que la incorporación de variables exógenas mejora el rendimiento en escenarios donde las covariables aportan señal relevante y están disponibles a futuro (especialmente las de calendario). En general, los modelos globales neuronales y DeepAR tienden a beneficiarse de la información compartida entre tiendas, mientras que SARIMAX y Prophet ofrecen interpretabilidad y un comportamiento robusto en tiendas con patrones estables. Además, se incluyen experimentos de robustez para estudiar el comportamiento del sistema ante series cortas y ante cambios estructurales relacionados con un choque económico. Finalmente, se discuten limitaciones clave (supuesto *oracle* de ciertas covariables, sensibilidad a cambios de distribución) y se proponen líneas futuras para llevar los modelos a un entorno operativo.

Palabras clave: series temporales; predicción de demanda; variables exógenas; modelos globales; aprendizaje profundo.

Abstract

This Master's Thesis tackles the problem of weekly retail sales forecasting using historical sales together with a set of exogenous variables (calendar, holidays, and macroeconomic indicators). The main goal is to design a reproducible experimental protocol and to compare, under homogeneous conditions, several forecasting approaches that can explicitly leverage external covariates: statistical models (SARIMAX), a scalable additive model (Prophet), a probabilistic RNN-based model (DeepAR), and two global neural architectures (LSTM and a Transformer). To ensure a fair comparison, the work enforces consistent temporal splits and a leakage-safe feature engineering process, where target-derived features (lags and rolling statistics) are computed causally using only past information. Model accuracy is evaluated using MAE, RMSE, and sMAPE, reporting both global results and per-store performance.

The results indicate that adding exogenous variables improves performance when covariates carry predictive signal and are available for the forecasting horizon (especially calendar-related features). Overall, DeepAR and global neural models benefit from shared information across stores, whereas SARIMAX and Prophet provide interpretability and robust behavior for stores with stable patterns. The thesis also includes robustness experiments to analyze performance under short histories and under distribution shifts related to an economic shock. Finally, key limitations are discussed (notably the *oracle* assumption for some covariates and sensitivity to regime changes), and future work is proposed to move towards a realistic operational setting.

Keywords: time series; demand forecasting; exogenous variables; global models; deep learning.

Introducción 1

La predicción de demanda es una pieza central en la gestión de cadenas de suministro y operaciones minoristas: afecta a decisiones de inventario, planificación de personal, reposición y diseño de promociones. En series temporales reales, las ventas presentan estacionalidad (anual y semanal), efectos calendario (festivos y eventos), tendencias, así como perturbaciones externas asociadas a cambios macroeconómicos o a choques de oferta y demanda. Este Trabajo Fin de Máster se centra en el pronóstico de ventas semanales por tienda, utilizando información histórica y un conjunto de variables exógenas.

1.1. Motivación

En entornos multi-tienda es habitual disponer de decenas o centenares de series con comportamientos heterogéneos. Modelar cada tienda de forma independiente puede ser subóptimo cuando hay series cortas o ruidosas; en cambio, los *modelos globales* explotan patrones compartidos y pueden mejorar la generalización. Al mismo tiempo, existe una necesidad práctica de incorporar covariables externas (por ejemplo, festivos o indicadores económicos), lo que favorece enfoques que acepten regresores.

1.2. Planteamiento del problema

Se aborda un problema de predicción multiserie: para cada tienda, se desea predecir las ventas en un horizonte fijo (semanas futuras) a partir del historial de ventas y covariables. El proyecto prioriza una comparación *homogénea* de modelos: se seleccionan métodos capaces de manejar variables exógenas y se define un protocolo de validación temporal común para todos.

1.3. Aportaciones

Las principales contribuciones del trabajo son:

- Un protocolo experimental reproducible para comparar modelos con regresores exógenos en ventas semanales.

- Una implementación unificada de experimentos y métricas (MAE, RMSE, sMAPE) con resultados globales y por tienda.
- Un análisis de robustez ante series cortas y ante cambios estructurales asociados a un choque económico.
- Integración de modelos representativos: SARIMAX, Prophet (Taylor y Letham, 2018), DeepAR (Flunkert et al., 2017) y arquitecturas neuronales globales (LSTM (Hochreiter y Schmidhuber, 1997) y Transformer (Vaswani et al., 2017)).

1.4. Estructura de la memoria

El documento se organiza de la siguiente forma. El capítulo de **Estado del arte** revisa los enfoques relevantes para series temporales con covariables. El **Marco teórico** introduce los conceptos necesarios (validación temporal, fuga de información, métricas y regularización). A continuación, se definen los **Objetivos**, la **Metodología** y el diseño experimental. Finalmente, se presentan **Resultados y Discusión**, **Conclusiones** y **Limitaciones y líneas futuras**.

Estado del arte

2

La predicción de series temporales en entornos reales (como el comercio minorista) es un problema clásico y a la vez vigente: la demanda presenta estacionalidad, tendencias, efectos calendario y cambios estructurales asociados a promociones, festivos o choques macroeconómicos. En este contexto, la literatura ofrece dos grandes familias de enfoques: (i) modelos estadísticos interpretables basados en estructuras lineales y (ii) modelos de aprendizaje automático, especialmente redes neuronales profundas, que aprenden representaciones no lineales a partir de grandes colecciones de series.

2.1. Modelos estadísticos para series temporales

Los modelos *AutoRegressive Integrated Moving Average* (ARIMA) y sus extensiones siguen siendo una referencia por su interpretabilidad y por la claridad con la que separan componentes autorregresivos, de media móvil e integración. Una extensión habitual para incorporar covariables externas es la regresión dinámica, que combina regresión (con variables exógenas) y errores modelados mediante ARIMA; este enfoque se conoce comúnmente como *Seasonal ARIMA with eXogenous regressors* (SARIMAX). Su fortaleza reside en la capacidad de capturar dependencias temporales y, a la vez, incluir información explicativa adicional (por ejemplo, festivos o indicadores económicos), siempre que las covariables estén disponibles sin fuga de información. Una exposición moderna y aplicada puede encontrarse en Hyndman y Athanasopoulos (2021), mientras que el tratamiento clásico se recoge en Box et al. (2015).

2.2. Modelos aditivos y Prophet

Prophet es un modelo aditivo diseñado para series con estacionalidad y efectos de calendario, optimizado para su uso a escala y con un *workflow* robusto ante datos faltantes y cambios de tendencia. Además, permite incorporar regresores externos (*regressors*) de forma directa, lo que lo hace atractivo en escenarios donde ciertas variables explicativas se conocen a futuro o pueden estimarse con suficiente antelación. Taylor y Letham (2018) describen su formulación y su enfoque de diseño orientado a la operación en entornos industriales.

2.3. Modelos neuronales: RNN y aprendizaje global

En aplicaciones con múltiples series (p. ej., ventas por tienda) es común adoptar enfoques *globales*: un ´unico modelo aprende patrones compartidos y diferencias sistemáticas entre series, lo que permite generalizar mejor cuando hay series cortas o ruidosas. Entre los métodos neuronales más influyentes se encuentran las *Recurrent Neural Network* (RNN) y, en particular, las *Long Short-Term Memory* (LSTM), capaces de modelar dependencias de largo alcance mediante compuertas que mitigan el desvanecimiento del gradiente (Hochreiter y Schmidhuber, 1997). Para mejorar la capacidad de generalización, el *Dropout* (regularización estocástica) (Dropout) es un regularizador estándar que aproxima un ensamble de subredes y reduce el sobreajuste (Srivastava et al., 2014).

2.4. Pronóstico probabilístico y DeepAR

En el ´ambito del pronóstico probabilístico, DeepAR propone una formulación autoregresiva basada en RNN que produce distribuciones predictivas condicionales, permitiendo cuantificar incertidumbre y soportar covariables (dinámicas y estáticas) (Flunkert et al., 2017). Este tipo de modelos resulta especialmente adecuado cuando se requiere una estimación de riesgo o cuando se desea comparar no solo el error medio sino también la calibración de la incertidumbre.

2.5. Transformers y atención

Los Transformers han demostrado un gran rendimiento en secuencias mediante mecanismos de atención que permiten modelar dependencias sin recurrencia, facilitando el paralelismo y el aprendizaje de relaciones a largo plazo. La arquitectura fundacional se introdujo en Vaswani et al. (2017). En series temporales, las variantes basadas en atención se han convertido en una línea activa de investigación, especialmente en contextos multivariantes y con gran cantidad de covariables.

2.6. Evaluación y comparación de modelos

La comparación objetiva de modelos exige protocolos de validación temporal y métricas acordes al problema. Medidas como MAE, RMSE y sMAPE son habituales; sin embargo, cada una enfatiza propiedades distintas (sensibilidad a outliers, penalización cuadrática, o comparabilidad relativa), por lo que conviene reportar varias y discutir su interpretación (Hyndman y Athanasopoulos, 2021; Hyndman y Koehler, 2006).

Marco teórico 3

Este capítulo resume los conceptos teóricos necesarios para comprender el diseño experimental y la selección de modelos del proyecto: formulación del problema de pronóstico, papel de las variables exógenas, prevención de fuga de información y métricas de evaluación.

3.1. Formulación del problema

Sea y_t la variable objetivo (ventas semanales) para una tienda determinada. El objetivo es construir un modelo que, dado un historial de $y_{1:t}$ y un conjunto de covariables $\mathbf{x}_{1:t+h}$, produzca predicciones $\hat{y}_{t+1:t+h}$ para un horizonte h.

En este proyecto se considera un escenario con covariables exógenas relacionadas con calendario (p. ej., semana del año, indicadores de festivos) y con factores económicos. Para mantener la comparabilidad, se priorizan modelos que acepten covariables externas de manera explícita (SARIMAX, Prophet con regresores, DeepAR con covariables y modelos neuronales multivariantes).

3.2. Variables exógenas y supuestos de disponibilidad

Un aspecto crítico al usar variables exógenas es distinguir entre:

- **Covariables conocidas a futuro**: calendario y festivos planificados.
- **Covariables no conocidas a futuro**: indicadores macroeconómicos publicados con retraso o inciertos.

En la parte experimental se adopta un supuesto *oracle* para ciertas variables (se asume que pueden utilizarse en el horizonte de predicción). Este supuesto permite estudiar el *potencial* de mejora al incorporar información externa, pero debe interpretarse como un límite superior de rendimiento. Las implicaciones y limitaciones de este supuesto se discuten en el capítulo de limitaciones.

3.3. Fuga de información y validación temporal

En series temporales, la fuga de información aparece cuando una transformación usa información del futuro para construir variables en el pasado. Para evitarlo, las características

derivadas del objetivo (lags, medias móviles) deben construirse de forma *causal*, es decir, utilizando y_τ solo para predecir instantes posteriores.

Asimismo, los esquemas de partición deben respetar el orden temporal. En lugar de validación aleatoria, se emplean particiones temporales y, cuando procede, evaluación *walk-forward* (entrenar en un prefijo y validar en un segmento posterior) (Hyndman y Athanasopoulos, 2021).

3.4. Métricas de error

Para evaluar el rendimiento se reportan tres métricas complementarias:

- **MAE** (MAE): $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. Es interpretable en unidades originales y robusta ante outliers comparada con RMSE.
- **RMSE** $(RMSE)$: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. Penaliza más los errores grandes.
- **sMAPE** $(sMAPE)$: una versión simétrica del MAPE que facilita comparaciones relativas cuando las escalas varían; conviene interpretarla con cautela cuando hay valores cercanos a cero (Hyndman y Koehler, 2006).

3.5. Regularización y sobreajuste en redes neuronales

Los modelos neuronales tienden a sobreajustar cuando la cantidad de datos por serie es limitada o cuando la señal tiene alta varianza. El *Dropout* (regularización estocástica) (Dropout) es un mecanismo de regularización ampliamente usado que reduce co-adaptaciones entre neuronas y mejora la generalización (Srivastava et al., 2014).

3.6. Atención para series temporales

En arquitecturas tipo Transformer, la atención permite ponderar de forma adaptativa qué instantes pasados resultan más relevantes para predecir el futuro (Vaswani et al., 2017). En contextos con múltiples covariables, estos mecanismos pueden capturar interacciones complejas entre señales.

Objetivos

4

4.1. Objetivo general

Diseñar y evaluar un sistema de predicción de ventas semanales por tienda que integre variables exógenas y permita comparar, bajo un protocolo común y reproducible, modelos estadísticos y de aprendizaje profundo.

4.2. Objetivos específicos

1. **Definir un protocolo experimental temporal** (particiones, horizonte y métricas) que evite fuga de información y permita comparar modelos de manera justa.

2. **Implementar y entrenar modelos con capacidad de covariables exógenas**, incluyendo SARIMAX, Prophet, DeepAR, LSTM y Transformer.

3. **Evaluar el rendimiento global y por tienda** utilizando MAE, RMSE y sMAPE, identificando patrones de error y casos de fallo.

4. **Analizar la robustez** del sistema ante series cortas y ante cambios estructurales relacionados con un choque económico.

5. **Documentar las conclusiones** y proponer mejoras para un despliegue más realista (disponibilidad de covariables, recalibración y detección de drift).

Datos y covariables

5

5.1. Descripción del conjunto de datos

El trabajo utiliza el conjunto de datos *Walmart Store Sales*, con observaciones semanales de ventas agregadas por tienda. Cada registro se identifica por el par `(Store, Date)` y la variable objetivo es `Weekly_Sales`. El dataset contiene 45 tiendas, 143 semanas (desde el 5 de febrero de 2010 hasta el 26 de octubre de 2012) y 6435 observaciones en total.

5.1.1. Variables disponibles

El fichero incluye las siguientes columnas:

- **Identificadores**: `Store` (tienda), `Date` (fecha semanal).
- **Objetivo**: `Weekly_Sales` (ventas semanales).
- **Exógenas observadas**:
  - `Holiday_Flag` (indicador binario de semana festiva)
  - `Temperature`
  - `Fuel_Price`
  - `CPI`
  - `Unemployment`

En este dataset no se observan valores perdidos en ninguna de las variables.

5.2. Taxonomía de covariables y disponibilidad a futuro

Un aspecto central de este TFM es que la utilidad de las covariables depende de su **disponibilidad en el horizonte de predicción**. Por ello, distinguimos:

- **Covariables conocidas a futuro** (*known-future*): variables de calendario derivadas de la fecha (semana del año, mes, año) y, en escenarios realistas, un calendario de festivos publicado con antelación.

- **Covariables observadas (no garantizadas) a futuro**: variables como temperatura, precio del combustible, CPI o desempleo pueden no estar disponibles con certeza en el futuro. En consecuencia, su uso directo en predicción implica una hipótesis adicional (por ejemplo, una predicción exógena previa o un escenario *oracle*).

Esta distinción guía el diseño experimental y la discusión de resultados para evitar conclusiones no transferibles a un entorno de producción.

5.3. Ingeniería de características

A partir de las variables originales se construyen características causales, con especial cuidado para evitar **fuga de información**. Se usan tres grupos:

- **Lags del objetivo**: `lag_1`, `lag_2`, `lag_4`, `lag_8` y `lag_52`.
- **Estadísticos móviles**: `roll_mean_4`, `roll_mean_8`, `roll_std_8` (y ventanas adicionales según configuración).
- **Calendario**: `weekofyear`, `month` y `year`.

Todas las características derivadas del objetivo se calculan únicamente con valores **anteriores** al instante de predicción.

5.4. Normalización y representación

Los modelos neuronales globales (LSTM y Transformer) operan sobre secuencias multivariantes y emplean normalización (*standardization*) para estabilizar el entrenamiento. Los modelos estadísticos (SARIMAX y Prophet) consumen regresores en su escala original, pudiendo beneficiarse de transformaciones adicionales; en este trabajo se prioriza la comparabilidad del pipeline.

5.5. Horizonte de evaluación

La evaluación principal considera un horizonte de test de 39 semanas, correspondiente al tramo 2012-02-03 a 2012-10-26. Este horizonte se utiliza para comparar modelos bajo el mismo conjunto de fechas y tiendas.

Metodología

6

Este capítulo describe el protocolo experimental y las decisiones metodológicas para comparar modelos con covariables exógenas en un contexto multiserie. El objetivo es asegurar una evaluación **justa** y **auditable**, evitando fugas de información y documentando supuestos (especialmente sobre disponibilidad de exógenas).

6.1. Formulación del problema

Sea $y_{s,t}$ la venta semanal (`Weekly_Sales`) de la tienda s en la semana t. Dados un historial de longitud L y un conjunto de covariables $\mathbf{x}_{s,t}$, se desea estimar $\hat{y}_{s,t+h}$ para $h \in \{1, \ldots, H\}$. En la evaluación principal se toma $H = 39$ semanas.

6.2. Partición temporal y horizonte de test

La evaluación se realiza respetando el orden temporal (no se mezclan observaciones futuras en entrenamiento). Para la comparación principal se define un tramo final de **39 semanas** (desde 2012-02-03 hasta 2012-10-26) como conjunto de test. De este modo, todos los modelos se comparan sobre las mismas fechas y tiendas.

6.3. Prevención de fuga de información (anti-leakage)

La fuga de información en series temporales puede ocurrir de forma sutil, especialmente al construir características o al tratar covariables exógenas. Para minimizar riesgos, se adoptan las siguientes reglas:

- **Causalidad en características**: lags y estadísticos móviles del objetivo se calculan usando solo semanas $< t$.
- **Aislamiento entre splits**: cualquier transformación ajustada con datos (p. ej., escalado) se ajusta solo con entrenamiento y luego se aplica a validación/test.
- **Exógenas y horizonte**: el uso de exógenas en $t + h$ requiere explicitar si dichas exógenas son conocidas a futuro (calendario) o si se asume un escenario *oracle*.

Estas prácticas son coherentes con recomendaciones estándar de validación en forecasting (Hyndman y Athanasopoulos, 2021).

6.4. Características y configuración común

La configuración común se basa en:

- **Exógenas**: `Holiday_Flag`, `Temperature`, `Fuel_Price`, `CPI`, `Unemployment`.
- **Lags**: 1, 2, 4, 8, 52.
- **Ventanas móviles**: 4, 8, 12.
- **Calendario**: `weekofyear`, `month`, `year`.
- **Lookback**: 52 semanas para modelos secuenciales globales.

6.5. Diseño experimental (E1–E5)

Se implementan varios experimentos que cubren escenarios complementarios: [COM-

Tabla 6.1: ***Experimentos definidos para evaluación.*** *En todos los casos se respetan particiones temporales y se reportan métricas globales y por tienda.*

ID	Esquema	Objetivo
E1	Walk-forward	Evaluar despliegue con recalibración frecuente: para cada semana de test se entrena con el pasado y se predice 1 paso.
E2	Holdout final (39)	Comparación directa en un horizonte fijo de 39 semanas (2012-02-03 a 2012-10-26).
E3	*Leave-one-store-out*	Medir capacidad de generalización inter-tienda en modelos globales (entrenar en todas menos una).
E4	Train-grupo / test-grupo	Robustez al transferir entre grupos: entrenar en tiendas de mayor volumen y evaluar en 10 tiendas de menor media.
E5	Shock exógeno	Sensibilidad ante cambios estructurales: perturbar desempleo en test y observar degradación relativa.

PLETAR: indicar qué experimentos se ejecutaron completamente en el entorno de entrega y el coste computacional aproximado.]

6.6. Métricas

Se reportan MAE, RMSE y sMAPE. Además de las métricas globales (agregadas), se calculan métricas por tienda para caracterizar heterogeneidad y detectar series difíciles.

6.7. Reporte y trazabilidad

Las ejecuciones generan artefactos en `outputs/` (predicciones, métricas y figuras) y registran metadatos (semilla, features y librerías). Esta trazabilidad facilita auditar resultados y regenerar las tablas/figuras integradas en la memoria.

Modelos

7

Este capítulo describe los modelos evaluados, con énfasis en cómo incorporan covariables exógenas y en qué medida permiten aprendizaje *global* (multiserie) frente a aprendizaje por serie.

7.1. Modelos estadísticos con exógenas

7.1.1. SARIMAX

SARIMAX extiende los modelos ARIMA estacionales incluyendo regresores exógenos. En esta memoria se utiliza como *baseline* interpretable: permite modelar dependencia temporal (componentes autorregresivas y de medias móviles) y, a la vez, capturar el efecto de covariables. Su uso es habitual en regresión dinámica y econometría (Box et al., 2015; Hyndman y Athanasopoulos, 2021).

7.1.2. Prophet con regresores

Prophet es un modelo aditivo que combina tendencia, estacionalidades y efectos de festivos. Además, permite añadir regresores externos. En este trabajo se emplea Prophet como alternativa robusta y de rápida calibración, especialmente útil cuando se desea interpretabilidad y un flujo de modelado ágil (Taylor y Letham, 2018).

7.2. Modelos probabilísticos y de aprendizaje profundo

7.2.1. DeepAR

DeepAR es un modelo autoregresivo basado en RNN que aprende una distribución predictiva condicional y se ha consolidado como referencia en forecasting probabilístico multiserie. Su arquitectura permite incorporar covariables dinámicas (por ejemplo, exógenas por semana) (Flunkert et al., 2017).

7.2.2. LSTM global

Las redes LSTM (Hochreiter y Schmidhuber, 1997) modelan dependencias de largo plazo y son una base habitual en series temporales. En este TFM se entrena una red global sobre

múltiples tiendas, de modo que el modelo comparte parámetros y puede transferir patrones entre series.

7.2.3. Transformer global

Los Transformers basados en atención (Vaswani et al., 2017) permiten modelar dependencias a largo plazo sin recurrencia explícita. En forecasting, su atractivo reside en la capacidad de combinar señales heterogéneas (lags, calendario y exógenas) y capturar interacciones no lineales.

7.3. Entrada de características y horizonte

Los modelos globales reciben, para cada tienda, una secuencia multivariada de longitud **lookback**. En la configuración principal se usa `lookback = 52` (aprox. un año), lo que permite capturar estacionalidad anual y patrones de medio plazo. El horizonte de predicción evaluado es de 39 semanas.

7.4. Regularización y consideraciones prácticas

En los modelos neuronales se emplea Dropout para reducir sobreajuste (Srivastava et al., 2014). Además, se incorporan mecanismos defensivos para evitar fallos cuando una tienda tiene historia efectiva menor que el *lookback* requerido (reducción adaptativa del lookback), de forma que el entrenamiento y la inferencia sean robustos en presencia de series más cortas.

7.5. Escenarios de covariables (realista vs. *oracle*)

Para interpretar correctamente el valor de las exógenas, se contemplan dos escenarios conceptuales:

- **Escenario *oracle***: se asume que todas las covariables exógenas son conocidas durante el horizonte:
  - `Holiday_Flag`
  - `Temperature`
  - `Fuel_Price`
  - `CPI`
  - `Unemployment`
- **Escenario realista**: solo se consideran covariables garantizadas (p. ej., calendario), o bien se introduce un modelo auxiliar para anticipar exógenas.

En los resultados se discute explícitamente qué conclusiones dependen de la hipótesis *oracle*.

Resultados y Discusión

8

8.1. Resumen de resultados

En esta sección se resumen los resultados obtenidos en los experimentos y se discuten los patrones observados. Para facilitar la comparación entre modelos, se reportan métricas globales (agregadas) y métricas por tienda. La discusión se centra en dos ejes: (i) el impacto de incorporar variables exógenas y (ii) la capacidad de generalización de modelos globales frente a enfoques por serie.

8.2. Comparación global entre familias (configuración base)

Como punto de partida, la Tabla 8.1 compara familias de modelos con una configuración base.

Tabla 8.1: ***Comparación global entre modelos (configuración base).*** *Métricas agregadas sobre todas las tiendas y semanas del horizonte de test. Menor es mejor.*

Modelo	MAE	RMSE	sMAPE
DeepAR (exógenas)	59 573.68	86 287.62	6.041
LSTM global (exógenas)	75 827.53	98 575.42	8.016
Prophet (regresores)	142 509.72	186 757.80	12.139
SARIMAX (exógenas)	124 157.68	169 022.47	10.889
Transformer global (exógenas)	113 420.92	147 263.71	9.863

La Figura 8.1 complementa la tabla, facilitando una lectura visual conjunta.

8.3. Ranking global por MAE

La Figura 8.2 muestra el ranking de los mejores modelos según MAE. Esta visualización es ́util para identificar configuraciones competitivas y comparar la estabilidad de rendimiento.

model	MAE	RMSE	sMAPE
deepar_exog	59573.6846	86287.6172	6.0413
lstm_exog	75827.5277	98575.4242	8.0161
prophet_regressors	142509.7176	186757.8048	12.1389
sarimax_exog	124157.6803	169022.47	10.8893
transformer_exog	113420.9156	147263.7065	9.8635

Figura 8.1: ***Comparación global de métricas (MAE/RMSE/sMAPE).*** *Menor es mejor.*

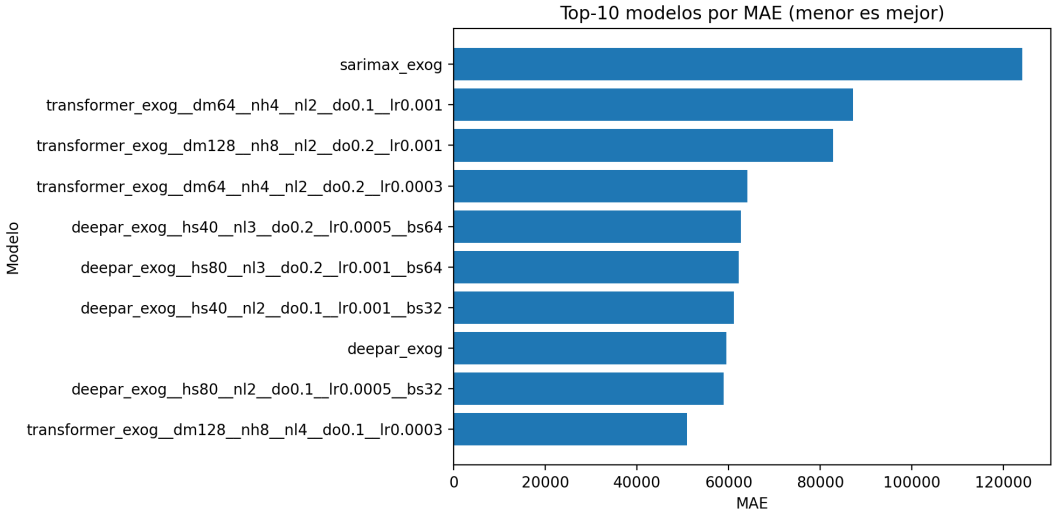

Figura 8.2: ***Top 10 modelos por MAE.*** *Menor es mejor.*

8.4. Tabla comparativa de métricas

La Tabla 8.2 recoge MAE, RMSE y sMAPE para los modelos mejor posicionados según el resumen experimental.

8.5. Comparación normalizada de las top-5 configuraciones

La Figura 8.3 compara las top-5 configuraciones tras normalizar cada métrica para facilitar una lectura conjunta. En general, un modelo puede dominar en MAE pero no necesariamente en RMSE (más sensible a errores grandes) o en sMAPE (métrica relativa).

Modelo	MAE	RMSE	sMAPE
transformer_exog__dm128__nh8__nl4__do0.1__lr0.0003	50,989.60	68,451.73	5.53
deepar_exog__hs80__nl2__do0.1__lr0.0005__bs32	58,954.49	82,957.67	6.16
deepar_exog	59,573.68	86,287.62	6.04
deepar_exog__hs40__nl2__do0.1__lr0.001__bs32	61,163.80	89,421.24	6.13
deepar_exog__hs80__nl3__do0.2__lr0.001__bs64	62,260.56	89,886.14	6.19
deepar_exog__hs40__nl3__do0.2__lr0.0005__bs64	62,739.17	89,809.35	6.42
transformer_exog__dm64__nh4__nl2__do0.2__lr0.0003	64,173.73	85,342.68	6.69
transformer_exog__dm128__nh8__nl2__do0.2__lr0.001	82,876.35	104,778.98	11.05
transformer_exog__dm64__nh4__nl2__do0.1__lr0.001	87,233.72	105,921.99	10.93
sarimax_exog	124,157.68	169,022.47	10.89

Tabla 8.2: ***Top-10 modelos globales por MAE (E2).***

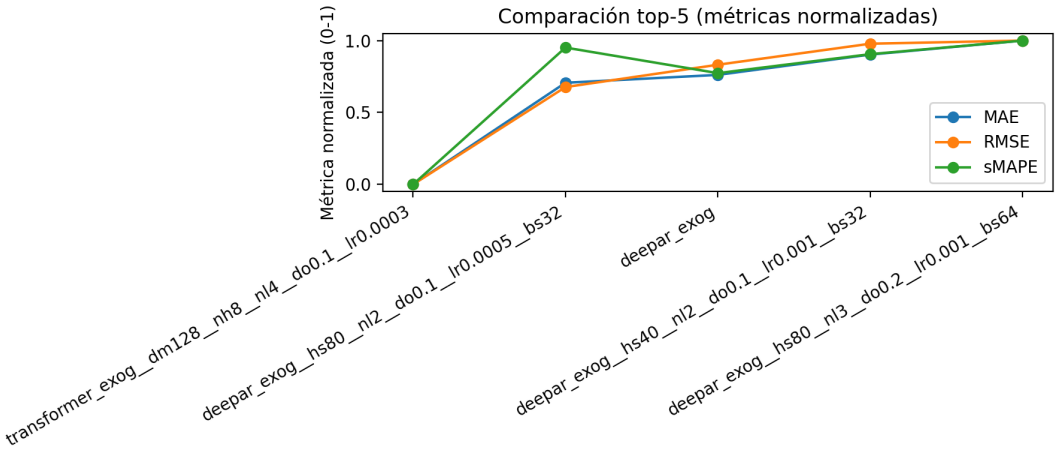

Figura 8.3: ***Top 5: comparación normalizada de MAE, RMSE y sMAPE.*** *Menor es mejor.*

8.6. Discusión

De forma cualitativa, se observan los siguientes patrones:

- **Beneficio de covariables**: las variables de calendario tienden a aportar señal consistente; su impacto es mayor en tiendas con estacionalidad marcada.
- **Modelos globales**: en presencia de múltiples tiendas, los enfoques globales (DeepAR, LSTM y Transformer) pueden beneficiarse de patrones compartidos, en especial cuando algunas series son cortas.
- **Interpretabilidad**: SARIMAX y Prophet aportan interpretabilidad y diagnóstico de componentes. En cambio, los modelos neuronales capturan no linealidades y efectos combinados, a costa de mayor complejidad.
- **Sensibilidad a cambios de régimen**: los errores tienden a aumentar ante cambios estructurales; este aspecto se analiza en experimentos de robustez y se retoma en limitaciones.

En conjunto, los resultados respaldan el uso de variables exógenas cuando su disponibilidad a futuro está garantizada o es estimable, y sugieren que el aprendizaje global es una estrategia adecuada en escenarios multiserie.

8.7. Ablaciones: valor de las exógenas y del calendario

Una cuestión clave es separar el aporte de:

- **Calendario** (conocido a futuro), y
- **Exógenas observadas** (no garantizadas a futuro sin hipótesis adicional).

En esta ablación se comparan tres conjuntos de variables: FS0 (solo *lags*), FS1 (*lags*+calendario) y FS2 (*lags*+calendario+exógenas). Los resultados globales (ventana E2, 45 tiendas, 1 755 puntos) se resumen en la Tabla 8.3.

Tabla 8.3: ***Ablación E0 (métricas globales).*** *Comparación de conjuntos de variables: FS0 (solo* lags*), FS1 (*lags*+calendario) y FS2 (*lags*+calendario+exógenas). Menor es mejor.*

model	feature_set	MAE	RMSE	sMAPE	MAE_store_macro
lstm_exog	FS0	498 804	615 528	60.96	498 804
lstm_exog	FS1	715 895	819 447	109.52	715 895
lstm_exog	FS2	484 668	596 657	54.88	484 668
transformer_exog	FS0	330 957	392 481	36.76	330 957
transformer_exog	FS1	490 951	577 818	57.92	490 951
transformer_exog	FS2	433 880	501 105	51.62	433 880

De forma agregada, el efecto es heterogéneo: (i) en `lstm_exog`, FS1 empeora, mientras que FS2 logra una mejora moderada frente a FS0 (MAE $\downarrow$ de $498{,}8\,\text{k}$ a $484{,}7\,\text{k}$); (ii) en `transformer_exog`, FS1 también degrada, y FS2 sólo recupera parcialmente pero queda por debajo de FS0 (MAE $330{,}9\,\text{k}$ en FS0 vs. $433{,}9\,\text{k}$ en FS2).
La Figura 8.4–8.6 complementa esta comparación con una lectura visual de las métricas.

8.8. Escenario realista vs. escenario *oracle*

La hipótesis *oracle* (exógenas conocidas en el horizonte) tiende a favorecer modelos con regresores ricos. En un despliegue realista, la comparación debe considerar o bien exógenas realmente conocidas (calendario, festivos), o bien el error compuesto de un pipeline que primero predice exógenas y después predice ventas.

En la ablación E0, el conjunto FS2 (que incluye exógenas observadas) asume implícitamente un escenario *oracle* para el horizonte de predicción. Por tanto, estos resultados deben interpretarse como una cota optimista del beneficio de las exógenas. Un análisis más realista requeriría (i) limitarse a covariables conocidas a futuro (FS1) o (ii) modelar explícitamente la incertidumbre de las exógenas mediante un submodelo o una estrategia de imputación/persistencia.

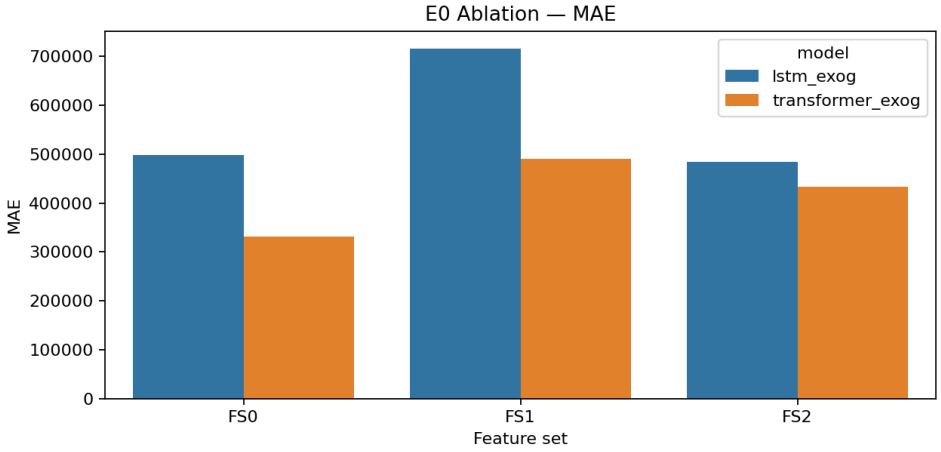

Figura 8.4: ***Ablación E0: MAE global por conjunto de variables.*** *Menor es mejor.*

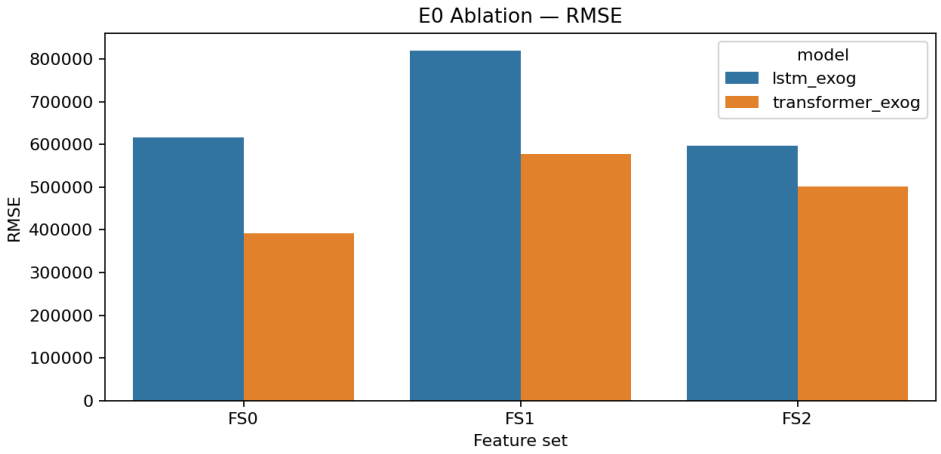

Figura 8.5: ***Ablación E0: RMSE global por conjunto de variables.*** *Menor es mejor.*

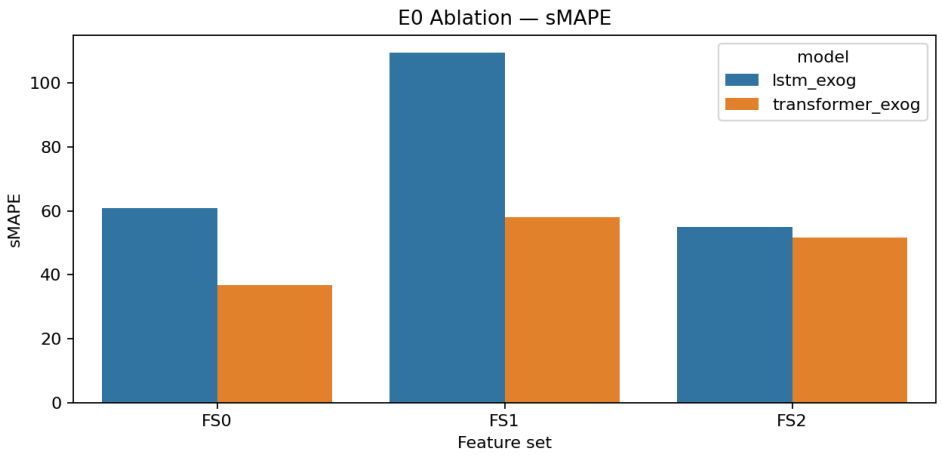

Figura 8.6: ***Ablación E0: sMAPE global por conjunto de variables.*** *Menor es mejor.*

8.9. Heterogeneidad por tienda

Además de la métrica global, el análisis por tienda permite identificar:

- tiendas donde los modelos globales transfieren bien patrones,
- tiendas con alta volatilidad o cambios de régimen,
- casos donde un modelo simple puede ser competitivo.

La Tabla 8.4 resume la dispersión del MAE por tienda (percentiles sobre las 45 tiendas del holdout E2). Esta vista permite observar cómo el beneficio de añadir covariables puede concentrarse en ciertas tiendas y no necesariamente reflejarse de forma uniforme.

Tabla 8.4: ***Ablación E0 (distribución del MAE por tienda).*** *Percentiles del MAE (p50/p75/p90) calculados sobre las 45 tiendas del conjunto de test. Menor es mejor.*

Modelo	FS	n tiendas	$p50$ MAE	$p75$ MAE	$p90$ MAE
lstm_exog	FS0	45	553 007	698 648	813 999
lstm_exog	FS1	45	757 800	1 014 507	1 120 013
lstm_exog	FS2	45	523 141	748 119	822 172
transformer_exog	FS0	45	396 475	432 771	509 918
transformer_exog	FS1	45	547 618	736 185	834 619
transformer_exog	FS2	45	518 800	632 329	690 790

Reproducibilidad y artefactos

9

La reproducibilidad es un objetivo transversal del trabajo. Se han diseñado scripts, estructuras de salida y metadatos para que los resultados puedan regenerarse y auditarse.

9.1. Estructura del proyecto

El repositorio se organiza en:

- **`src`**: implementación del pipeline de carga de datos, construcción de características, ejecución de experimentos y modelos.
- **`scripts`**: puntos de entrada para lanzar experimentos (p. ej., `run_all_experiments.py`).
- **`notebooks`**: notebooks exploratorios y de entrenamiento.
- **`outputs`**: predicciones, métricas, figuras y metadatos de ejecución.

9.2. Metadatos y configuración

Los experimentos registran metadatos (semilla, columnas exógenas, lags/ventanas, lookback e información de librerías). En particular, se utiliza **semilla 42** para controlar, en la medida de lo posible, la aleatoriedad.

9.3. Ejecución de experimentos

La ejecución está automatizada mediante scripts. Para regenerar resultados, el flujo esperado es:

1. Instalar dependencias (ver `requirements.txt`) y configurar el entorno.
2. Ejecutar el script de experimentos (p. ej., `scripts/run_all_experiments.py`) con los parámetros deseados.
3. Verificar que se generan predicciones, métricas y figuras bajo `outputs/`.

[COMPLETAR: comando exacto utilizado en la máquina de entrega, incluyendo argumentos y versión de CUDA/CPU si aplica.]

9.4. Artefactos reportados en la memoria

Esta memoria integra resultados mediante:

- Figuras en `Plantilla de memoria/figuras/`.
- Tablas LaTeX generadas a partir de CSV de métricas (p. ej., `summary_metrics.csv`).

[COMPLETAR: procedimiento reproducible (script) para regenerar la tabla `tabla_top10_metricas.tex` y la figura `top10_mae.png` desde los CSV.]

9.5. Compilación del documento

En Windows, la compilación se realiza con una secuencia manual robusta (sin `make`):

1. `pdflatex memoria.tex`
2. `makeglossaries memoria`
3. `bibtex memoria`
4. `pdflatex memoria.tex` (dos pasadas)

Este orden asegura que referencias cruzadas, bibliografía y glosarios se resuelvan correctamente.

Conclusiones

10

1. La incorporación de variables exógenas mejora el pronóstico cuando dichas covariables aportan señal predictiva y son coherentes con el horizonte de predicción.

2. La comparación bajo un protocolo temporal *común* (particiones consistentes y características causales) es esencial para evitar conclusiones sesgadas por fuga de información.

3. Los modelos globales (DeepAR y redes neuronales entrenadas con múltiples tiendas) tienden a beneficiarse del aprendizaje compartido, especialmente en tiendas con menos historial.

4. Modelos estadísticos como SARIMAX y enfoques aditivos como Prophet ofrecen una relación favorable entre interpretabilidad y rendimiento, resultando opciones robustas para entornos con restricciones operativas.

5. Los experimentos de robustez sugieren que los cambios estructurales pueden degradar significativamente el rendimiento, lo que motiva estrategias de monitorización y reentrenamiento.

Limitaciones y Perspectivas de Futuro

11

11.1. Limitaciones

- **Disponibilidad de covariables a futuro**: parte del análisis asume un escenario *oracle* para ciertas variables exógenas. En un despliegue real, algunas covariables (p. ej., indicadores macroeconómicos) pueden no estar disponibles con la misma antelación o pueden requerir su propia predicción.
- **Cambios de distribución**: los modelos pueden degradarse ante cambios estructurales (nuevos patrones de consumo, eventos extremos, cambios de políticas comerciales). Sin mecanismos de adaptación, el rendimiento observado en un periodo puede no extrapolar a periodos posteriores.
- **Heterogeneidad entre tiendas**: aunque los modelos globales ayudan, existen tiendas con dinámicas idiosincráticas (rupturas, aperturas/cierres, promociones no registradas) que limitan la capacidad de generalización.
- **Métricas agregadas**: un valor global puede ocultar degradaciones severas en subconjuntos de tiendas. Es necesario acompañar el análisis con estadísticos por tienda y segmentaciones.

11.2. Perspectivas de futuro

- **Pronóstico de covariables y escenario realista**: extender el sistema para predecir covariables no conocidas a futuro (o sustituirlas por variables proxy), evaluando la degradación frente al escenario *oracle*.
- **Modelos probabilísticos y calibración**: complementar las métricas puntuales con evaluación de distribuciones predictivas (calibración, intervalos) para soportar decisión bajo incertidumbre.
- **Detección de drift y reentrenamiento**: incorporar monitorización continua y políticas de reentrenamiento, con ventanas deslizantes y validación *walk-forward* periódica.
- **Segmentación por tiendas**: entrenar modelos por clúster o jerárquicos que exploten similitudes (zona geográfica, tamaño, perfil de ventas) para reducir heterogeneidad.

- **Explicabilidad**: añadir herramientas para analizar la contribución de covariables (p. ej., importancias o análisis de sensibilidad) y apoyar la toma de decisiones.

Lista de Acrónimos

ARIMA *AutoRegressive Integrated Moving Average*.

CPI *Consumer Price Index*.

Dropout *Dropout* (regularización estocástica).

LSTM *Long Short-Term Memory*.

MAE *Mean Absolute Error*.

RMSE *Root Mean Squared Error*.

RNN *Recurrent Neural Network*.

RNN *Recurrent Neural Network*.

SARIMAX *Seasonal ARIMA with eXogenous regressors*.

sMAPE *symmetric Mean Absolute Percentage Error*.

TFM Trabajo Fin de Máster.

Apéndice A: Detalles de configuración

A

A.1. Metadatos de ejecución

Se registran metadatos de configuración (semilla, features, librerías) en el fichero: `outputs/metadata_experiments.json`
[COMPLETAR: incluir aquí (o como tabla) el contenido relevante de metadatos para la entrega final.]

A.2. Lista de características

[COMPLETAR: tabla con features exactas usadas en el experimento principal.]

Apéndice B: Figuras adicionales

B

B.1. Distribución de errores

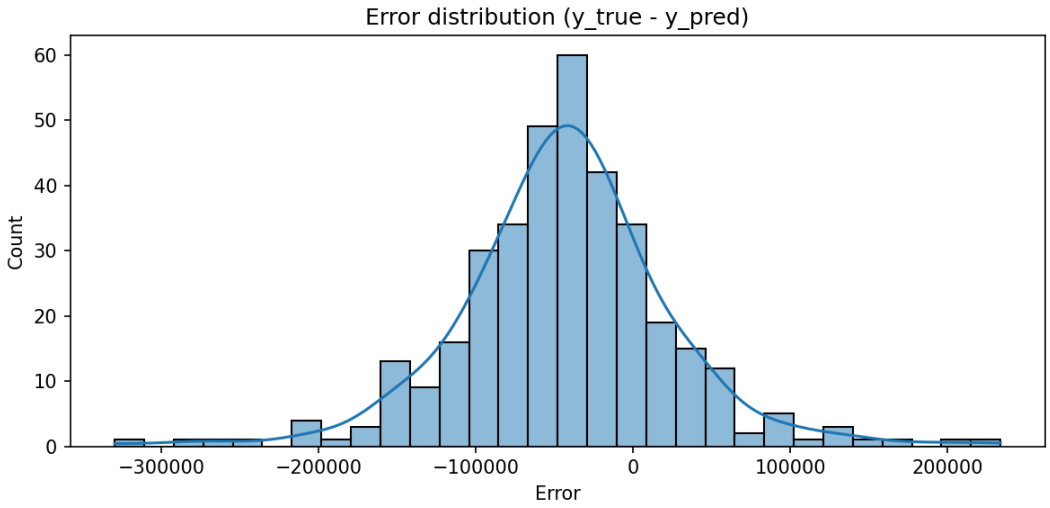

Figura B.1: ***Distribución de errores (Transformer).*** *Histograma/resumen de errores sobre el horizonte de evaluación.*

B.2. Ejemplos por tienda

[COMPLETAR: insertar 2–3 figuras de predicción vs. real por tienda para modelos representativos.]

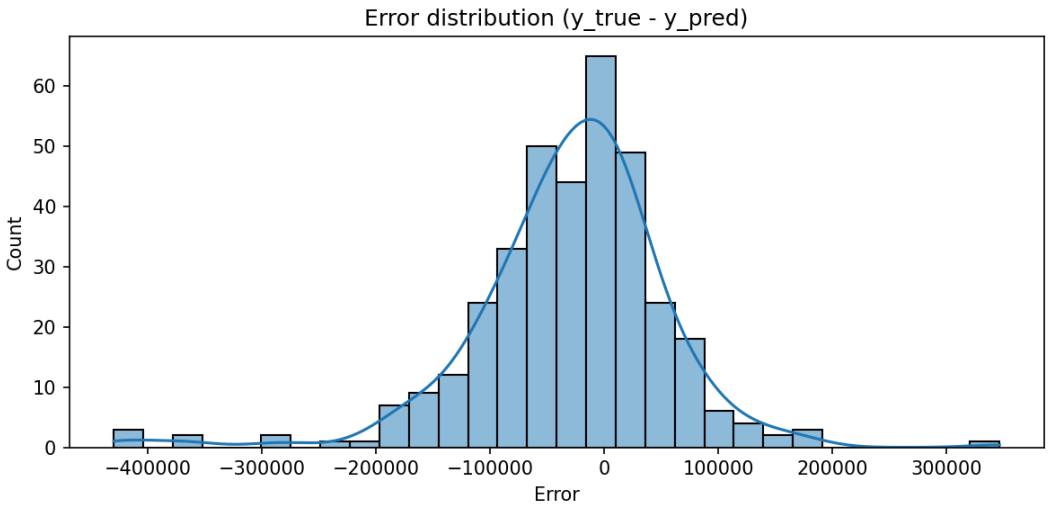

Figura B.2: ***Distribución de errores (DeepAR).*** *Histograma/resumen de errores sobre el horizonte de evaluación.*

Bibliografía

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., y Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. Wiley, 5 edition.

Flunkert, V., Salinas, D., y Gasthaus, J. (2017). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *arXiv preprint arXiv:1704.04110*.

Hochreiter, S. y Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hyndman, R. J. y Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 3 edition. Accedido el 31/01/2026.

Hyndman, R. J. y Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., y Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Taylor, S. J. y Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Łukasz., y Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.