UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Macías Alcaide, Paula. *Disseny i implementació d'una arquitectura de dades al núvol per a la predicció de vendes en una empresa del sector retail.* Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337341

under the terms of the BY-NC-ND license.

Disseny i implementació d'una arquitectura de dades al núvol per a la predicció de vendes en una empresa del sector retail

Paula Macías Alcaide

Resum-Aquest treball presenta el disseny i la implementació d'una arquitectura de dades al núvol per a la predicció de demanda en una empresa del sector retail òptic. El punt de partida és un sistema local de predicció basat en Random Forest (scikit-learn) amb limitacions greus de reproductibilitat, escalabilitat i qualitat d'avaluació. La solució implementada s'articula en tres eixos: una arquitectura Medallion (Bronze-Silver-Gold) sobre Databricks i Azure Delta Lake amb entorns DEV i PRO completament aïllats i CI/CD via Azure DevOps; un pipeline de predicció basat en la combinació d'un model SARIMAX auxiliar i XGBoost distribuït (SparkXGBRegressor) amb estratègia Direct Multi-Step Forecasting de 12 horitzons independents; i un sistema MLOps amb registre sistemàtic d'experiments a MLflow i versionat de models a Unity Catalog. Els resultats principals són un WAPE de 0,45 davant el 0,62 del baseline real (millora del 26,8 %) i un R^2 de 0,888 a l'horitzó h=1, amb el sistema operatiu en producció amb execució setmanal automatitzada.

Paraules clau-demand forecasting, arquitectura Medallion, XGBoost, Databricks, MLflow, Azure, SARIMAX, Direct Multi-Step Forecasting, supply chain, multi-horizon forecasting, XGBoost multi-output, feature engineering

Abstract-This paper presents the design and implementation of a cloud data architecture for demand forecasting in an optical retail company. Starting from a local Python-based Random Forest system with critical limitations in reproducibility and scalability, the solution implements a Medallion architecture (Bronze–Silver–Gold) on Databricks and Azure Delta Lake with fully isolated DEV and PRO environments and CI/CD via Azure DevOps; a prediction pipeline combining an auxiliary SARIMAX model and distributed XGBoost (SparkXGBRegressor) with Direct Multi-Step Forecasting strategy (12 independent horizons); and an MLOps system with systematic experiment tracking in MLflow and model versioning in Unity Catalog. Main results: WAPE of 0.450 vs baseline 0.615 (26.8 % improvement) and R^2 of 0.888 at horizon h=1, with the system running in production on a weekly automated schedule.

Index Terms-demand forecasting, Medallion architecture, XGBoost, Databricks, MLflow, Azure, SARIMAX, Direct Multi-Step Forecasting, supply chain, multi-horizon forecasting, XGBoost multi-output, feature engineering

1 Introducció

La predicció de demanda és un component crític en la planificació operativa del sector retail. En empreses amb catàlegs de milers de SKUs distribuïts per múltiples zones geogràfiques i canals de venda, la precisió del forecast condiciona directament els nivells d'estoc, la planificació de compres i la capacitat de resposta a les fluctuacions del mercat. Una infraestimació sistemàtica genera ruptures de stock i pèrdua de vendes, mentre que una sobreestimació incrementa els costos d'inventari i les obsolescències.

L'empresa analitzada en aquest treball disposava d'un sistema de predicció implementat localment en Python, basat en un model Random Forest de scikit-learn. Malgrat la seva funcionalitat operativa, el sistema presentava limitacions estructurals crítiques: una divisió de dades estratificada que no respectava l'ordre temporal, introduint data leakage i inflant artificialment el rendiment; errors en la lògica de cicle de vida dels productes i en la construcció dels horitzons de predicció; absència total d'entorns separats de desenvolupament i producció; processament en memòria local amb pandas, no escalable a la mida real del catàleg; i mètriques d'avaluació inadequades per a la naturalesa del negoci (el WAPE estàndard de supply chain no estava ni calculat). Baseline: WAPE=0,62, MAE=2,47, RMSE=9,80, R^2=0,70, BIAS=0,01.

L'article s'organitza en 6 seccions: objectius (2), estat de l'art (3), metodologia (4), resultats i discussió (5) i conclusions (6).

2 Objectius

Aquest treball persegueix cinc objectius específics que abasten el disseny de la infraestructura, la governança de dades, la millora del model predictiu i l'automatització del procés.

O1. Dissenyar i implementar una arquitectura de dades seguint el model medallion (Bronze, Silver i Gold) sobre Databricks i Azure Delta Lake, garantint la traçabi-

Paula.MaciasA@uab.cat
Treball tutoritzat per: Eduardo Cesar Cabrera Flores
Grau en Enginyeria de Dades - UAB
Curs 2025/26

litat i la qualitat de les dades en cada capa mitjançant dos entorns completament aïllats (DEV i PRO) amb governança centralitzada via Unity Catalog.

O2. Construir pipelines de processament de dades automatitzats i reutilitzables que cobreixin la ingesta des del data warehouse corporatiu, les transformacions de qualitat i neteja i la generació de features per al model, orquestrats amb Databricks Workflows en cicles mensuals i setmanals.

O3. Establir pràctiques DevOps amb control de versions (Azure DevOps, branques dev/pro i política de pull request obligatòria) i gestió segura de credencials via Azure Key Vault, garantint la reproductibilitat i el control de versions del codi en entorns productius.

O4. Millorar la capacitat predictiva respecte al sistema original (WAPE = 0,62) substituint el model Random Forest local per XGBoost distribuït amb estratègia Direct Multi-Step Forecasting a 12 horitzons independents, incorporant un model SARIMAX auxiliar com a generador de features i optimitzant hiperparàmetres amb Optuna. El sistema diferencia entre demanda regular (model OPTICS) i demanda irregular de Key Accounts (model KA).

O5. Identificar i analitzar els reptes reals sorgits durant el projecte en relació a la qualitat de les dades, l'escalabilitat del sistema i la mantenibilitat de la solució, justificant les decisions tècniques adoptades en cada cas.

3 ESTAT DE L'ART

La predicció de demanda és un camp àmpliament estudiat. Els models clàssics de sèries temporals (ARIMA/SARIMA [1], Prophet [2]) ofereixen bona interpretabilitat però assumeixen estacionalitat i no gestionen bé la predicció simultània de milers de SKUs. Els models de deep learning (DeepAR [3], N-BEATS [4], Temporal Fusion Transformers [5]) han mostrat bon rendiment, però la seva superioritat sobre mètodes basats en arbres en dades tabulars no és consistent [6], i requereixen grans volums de dades per SKU.

3.1 XGBoost per a forecasting tabular

XGBoost (eXtreme Gradient Boosting) és un mètode d'aprenentatge automàtic basat en gradient boosting sobre arbres de decisió. El principi del gradient boosting consisteix a construir un conjunt d'arbres de manera seqüencial, on cada arbre nou s'entrena per corregir els errors residuals del conjunt anterior, minimitzant una funció de pèrdua diferenciable. XGBoost afegeix sobre aquest paradigma tres contribucions clau: (1) regularització L1 i L2 sobre els pesos de les fulles, que redueix el overfitting; (2) aproximació per histogrames per al càlcul dels punts de divisió dels arbres, que n'accelera l'entrenament; i (3) suport natiu per al processament distribuït, que permet escalar a conjunts de dades de milions de registres. A diferència dels mètodes paramètrics com ARIMA o SARIMAX, XGBoost no assumeix cap forma funcional sobre la relació entre les variables, cosa que el fa especialment adequat per modelitzar la demanda de productes amb comportaments heterogenis i no lineals.

La funció objectiu que XGBoost minimitza combina la pèrdua predictiva amb un terme de regularització explícit:

$$\mathcal{L}(\varphi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k),$$
$$\Omega(f) = \gamma T + \tfrac{1}{2}\lambda \|w\|^2$$

on l és la funció de pèrdua sobre cada observació (en aquest treball, l'error absolut), T el nombre de fulles de cada arbre f_k, w els pesos de les fulles, i γ, λ els hiperparàmetres de regularització L1 i L2. Aquest terme és especialment rellevant per al conjunt de dades d'aquest projecte, on el 42,7% de les sèries temporals presenten menys de tres mesos de dades: sense regularització, els arbres tendeixen a memoritzar el soroll de les sèries curtes en lloc d'aprendre patrons generalitzables.

Chen i Guestrin [7] van proposar XGBoost com una implementació eficient de gradient boosting que ha demostrat rendiment competitiu o superior als models de deep learning en dades tabulars estructurades [7]. En el context del forecasting de demanda, l'estratègia Direct Multi-Step Forecasting (DMSF) [8], entrenar un model independent per a cada horitzó, evita l'acumulació d'errors del forecasting recursiu i permet l'ajust específic per a cada distància temporal. La integració de models estadístics com SARIMAX com a generadors de features per a models de ML és una pràctica establerta per capturar components estacionals explícits sense data leakage [9]. Per a la demanda lumpy (esporàdica i irregular), pròpia dels canals Key Account, la separació en models independents per tipologia de demanda és recomanada [15].

L'ús de SparkXGBRegressor en lloc de XGBRegressor permet l'entrenament distribuït sobre el clúster Databricks amb fins a 8 workers en paral·lel. Per al dataset de 7,9M de registres, el temps d'entrenament dels 12 models OPTICS en el clúster de producció es de menys d'1h, acceptable per a una tasca mensual. La funció objectiu reg:absoluteerror calcula el gradient sobre l'error absolut, que per a valors positius es constant (+1 o -1), introduint robustesa als outliers. En contrast, reg:squarederror creix quadràticament amb l'error i s'ha mostrat sensible als pics de demanda esporàdica presents als canals KA. La comparació experimental entre funcions objectiu (iteració 3 d'aquest projecte) ha confirmat la superioritat de reg:absoluteerror en WAPE per a tots els canals.

LightGBM (Light Gradient Boosting Machine) [17] és un framework de gradient boosting desenvolupat per Microsoft que, com XGBoost, construeix arbres de decisió de forma seqüencial minimitzant una funció de pèrdua diferenciable. Les seves principals innovacions són GOSS (Gradient-based One-Side Sampling), que exclou mostres amb gradient petit reduint el cost computacional, i EFB (Exclusive Feature Bundling), que agrupa features mútuament exclusives per reduir la dimensionalitat de les dades. En benchmarks sobre datasets tabulars, LightGBM i XGBoost presenten rendiments comparables; LightGBM sol ser lleugerament més ràpid gracies a la seva estratègia de creixement leaf-wise. En el marc d'aquest projecte, LightGBM es va avaluar com a alternativa a XGBoost però es va descartar per incompatibilitat de versió amb Spark 4.0 en l'entorn de producció (vegeu seccio 4.2).

3.2 Arquitectures Medallion i MLOps

El patró Lakehouse [10] i l'arquitectura Medallion sobre Delta Lake [11] proporcionen un marc de gestió de dades que combina les capacitats analítiques d'un data warehouse amb la flexibilitat d'un data lake, amb garanties ACID i suport per a format obert. MLflow [12] és el framework estàndard per al registre sistemàtic d'experiments, versionat de models i reproductibilitat. Unity Catalog [13] afegeix governança centralitzada sobre dades i artefactes de ML en entorns multi-workspace.

3.3 Mètriques d'avaluació en forecasting de demanda

El MAE (Mean Absolute Error) es la mitjana dels errors absoluts en unitats de producte, directament interpretable per als equips de compres. El RMSE (Root Mean Squared Error) és l'arrel quadrada de la mitjana dels errors quadràtics; penalitza els errors grans de forma no lineal i es especialment sensible als outliers de demanda puntual. El R2 (coeficient de determinació) mesura la proporció de la variació de les vendes reals explicada pel model: R2 = 1,0 indica predicció perfecta, R2 = 0,0 que el model no millora la predicció de la mitjana històrica, i valors negatius que el model és pitjor que la mitjana històrica.

La selecció de mètriques d'avaluació en forecasting de demanda és una decisió rellevant que condiciona l'optimització dels models i la comparabilitat dels resultats. El MAPE (Mean Absolute Percentage Error), la mètrica més àmpliament usada en la industria, presenta el problema de valors infinits quan les vendes reals s'apropen a zero, un cas freqüent en productes estacionals, discontinus o de nova col·lecció. El WAPE (Weighted Absolute Percentage Error), definit com la suma dels errors absoluts dividida per la suma de les demandes reals, és equivalent a una ponderació de l'error per volum de vendes i és robust davant les series amb zeros. Per a catàlegs d'alta rotació com el d'aquest treball (17.355 SKUs), el WAPE es la mètrica estàndard recomanada per la literatura de supply chain per a catàlegs de gran amplada [9].

El BIAS (mitjana dels errors signats) mesura la direccionalitat de l'error: positiu indica sobreestimació sistematica, negatiu indica infraestimació. És especialment rellevant en contextos d'inventari on una infraestimació derivaria en ruptures de stock. El SMAPE (Symmetric MAPE) i el P90 (percentil 90 de l'error absolut) complementen les mètriques principals en la caracterització del comportament de cua de la distribució d'errors.

4 METODOLOGIA

4.1 Infraestructura i DevOps

La solució s'ha implementat sobre Microsoft Azure amb dos grups de recursos completament independents: Forecasting-DEV i Forecasting-PRO. Cada entorn inclou un workspace de Databricks Premium, un Unity Catalog i un Azure Key Vault per a la gestió de credencials. Cap credencial no apareix en el codi; totes s'obtenen dinàmicament via Key Vault en temps d'execució, seguint les pràctiques de seguretat de secrets en entorns productius.

El volum de prediccions setmanals exportades es de 3,1M de registres (17.355 SKUs - 5 canals - 12 horitzons - 3 sucursals), consolidats a la taula gold.xgboost_predictions. Cada registre inclou l'identificador de SKU, canal, sucursal, horitzó de predicció, la predicció en unitats, i les metadades de la run de MLflow que ha generat la predicció (run_id, model_version). Això permet una traçabilitat completa des de la predicció final fins als hiperparàmetres i les dades d'entrenament usades, requisit imprescindible per a l'auditoria en entorns productius.

El control de versions s'implementa amb Azure DevOps amb un repositori de notebooks i dues branques principals (dev i pro). Les promocions de dev a pro requereixen pull request explícita, garantint que cap canvi no validat arriba a producció. La nomenclatura de taules segueix el patró {entorn}.{capa}.{taula} (ex.: pro.gold.xgboost_predictions), que força l'aïllament complet entre entorns a nivell de catàleg (Fig. 1).

La gestió de credencials via Azure Key Vault garanteix que cap credencial (cadenes de connexió JDBC, tokens de storage, claus d'API) apareix en el codi dels notebooks. La scope de Databricks kv-forecasting-dev/pro esta associada directament al Key Vault de cada entorn, de manera que un notebook que s'executa en DEV automàticament obté les credencials de DEV (i mai les de PRO), sense cap lògica condicional en el codi.

La política de pull request de la branca PRO inclou una verificació automàtica de sintaxi dels notebooks (linting), una revisió del schema de les taules afectades (evitant canvis incompatibles que trenquin pipelines en producció) i una execució selectiva dels tests d'integració més crítics (carrega Bronze des de ETN_REPORTING, validació de schemas Silver, i consulta Gold). Aquesta política, aplicada des de les primeres setmanes del projecte, ha evitat tres incidents potencials detectats en fases de revisió que, d'haver arribat a producció, haurien interromput les prediccions setmanals. El temps mitjà de revisió de pull request es de menys d'una hora gracies a la pipeline automatitzada.

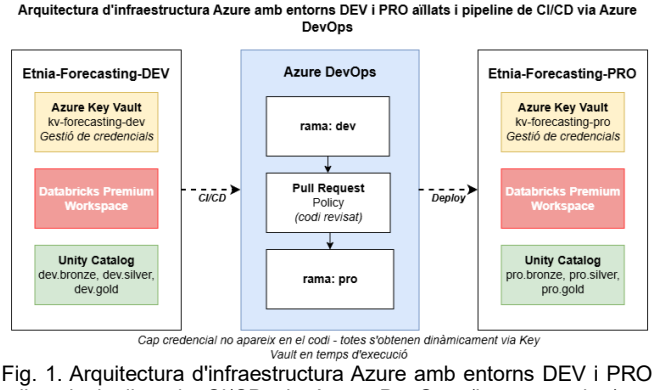

Fig. 1. Arquitectura d'infraestructura Azure amb entorns DEV i PRO aïllats i pipeline de CI/CD via Azure DevOps (branques dev/pro, política de pull request).

4.2 Pipeline de dades: arquitectura Medallion

La gestió de dades segueix l'arquitectura Medallion (Fig. 2) amb tres capes de responsabilitat diferenciada. La capa Bronze ingereix les dades en format brut des del data warehouse corporatiu ETN_REPORTING via conne-

xió JDBC, sense aplicar cap transformació. Les tres vistes d'origen generen quatre taules: bronze.sell_in (vBI_SALES_DATA, 7,9 M de registres de vendes sell-in, juliol 2021 - març 2026, incloent els registres negatius de devolucions tal com apareixen a l'origen), bronze.sell_out (ETN_SELLOUT, 213 K de registres de vendes sell-out dels canals Key Account KA-GEA, KA-GOP, KA-AEG i KA-ECP), bronze.item_attributes (VBI_ITEM_DATA, 17.355 SKUs amb els camps ReleaseDate, WithdrawalDate, Marca, CategoryName, SegmentCode i isDEM) i bronze.recent_sell_in (sell-in dels últims 12 mesos). Adicionalment, bronze.sarimax_parameters desa els paràmetres ajustats de cada sèrie SARIMAX, garantint que el Workflow 2 setmanal usa els mateixos paràmetres de l'entrenament mensual.

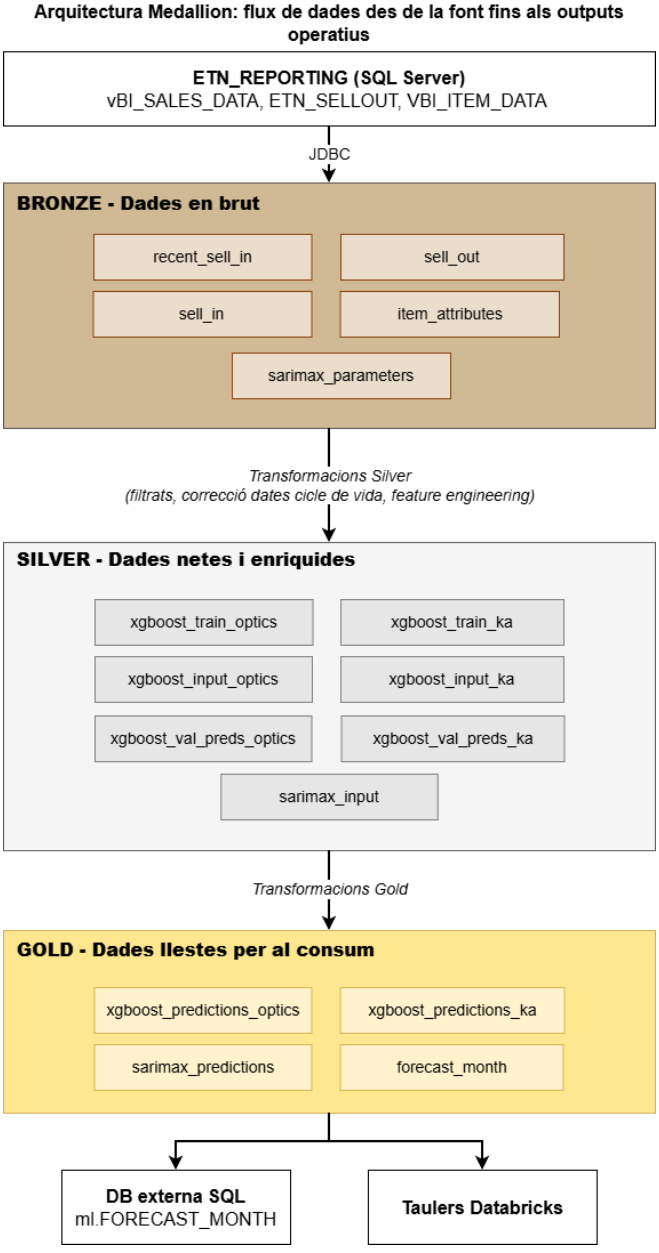

Fig. 2. Arquitectura Medallion: flux de dades des de la font (ETN_REPORTING) fins als outputs operatius, amb les transformacions de qualitat a cada capa i els consumidors finals (DB externa).

La ingesta s'executa com una càrrega completa en cada execució dels Workflows: totes les taules Bronze es sobreescriuen íntegrament (mode overwrite) amb les dades actuals de l'origen, sense mantenir marca d'aigua ni lògica incremental. Aquesta estratègia garanteix la consistència amb el sistema d'origen, ja que ETN_REPORTING pot incloure correccions retroactives en comandes i devolucions. Les taules es desen en format Delta Lake a Azure Data Lake Storage Gen2, aprofitant les capacitats transaccionals de Delta per assegurar que cada execució deixa la capa Bronze en un estat coherent.

La capa Silver aplica transformacions de qualitat sobre les dades Bronze i genera les taules necessàries per a l'entrenament i la validació dels models. L'anàlisi exploratòria ha identificat cinc categories de problemes que es corregeixen en aquesta capa: (1) 1.063.873 registres negatius de sell-in (devolucions) que s'eliminen; (2) el 42,7% de les combinacions SKU-canal-sucursal amb tres mesos o menys d'historial que s'exclouen de l'entrenament per manca de context suficient per construir les features de finestra temporal; (3) 208 SKUs amb WithdrawalDate anterior a ReleaseDate (incoherència d'origen); (4) 175 productes marcats com a actius amb WithdrawalDate ja superada; i (5) absència de dades de sell-out per al 57,6% dels SKUs, que afecta exclusivament la construcció de les features de sell-out del model KA.

La correcció de les categories 3 i 4 usa les vendes acumulades reals per calcular la data efectiva de retirada com el punt on s'assoleix el 98,5 % de les vendes totals del producte. Als SKUs amb cobertura de sell-out s'usa el patró de sell-out; als restants, el patró de sell-in. L'algoritme corregeix 383 SKUs en total i millora la qualitat de les features de cicle de vida months_in_col i months_out_col, especialment rellevants per als horitzons llargs (h = 9 a h = 12). Les transformacions generen set taules: silver.xgboost_train_optics i silver.xgboost_train_ka (dades netes i enriquides per a l'entrenament dels dos models XGBoost), silver.xgboost_val_preds_optics i silver.xgboost_val_preds_ka (prediccions dels dos models XGBoost sobre els conjunts de validació), i silver.xgboost_input_optics i silver.xgboost_input_ka (dades netes i enriquides per a la predicció dels dos models.

Les taules que s'utilitzen per entrenar el model i per a calcular les prediccions agrupen les features calculades per als 12 horitzons de predicció, amb una fila per combinació SKU-canal-sucursal-horitzó.

La capa Gold conté les taules llestes per al consum dels sistemes de negoci: gold.sarimax_predictions conté les prediccions del model SARIMAX per a les 16 series canal-sucursal; gold.xgboost_predictions_optics i gold.xgboost_predictions_ka conté les prediccions operatives setmanals (~4,8M de registres per setmana) i gold.forecast_month conté la unió de les prediccions dels dos models (OPTICS + KA).

Quantificant el volum de dades processat per la capa Gold, el dataset d'entrenament del model OPTICS (Workflow 1) conta amb 2.847.312 registres de SKU-canal-sucursal-mes per a l'horitzó h=1, i amb 2.831.104 per a h=12 (lleugerament inferior perquè alguns mesos recents

del conjunt de validació no són disponibles per a tots els horitzons). El dataset de validació (16 mesos) contra 764.211 registres per a h=1. La generació de les 12 taules gold.xgboost_input (una per horitz) es la tasca computacionalment més costosa del Workflow 1, amb un temps d'execució de ~45 minuts per als 12 horitzons en el clúster de 8 workers. Les transformacions s'apliquen via pySpark DataFrames amb window functions, que són l'abstrac de computació distribuïda més adequada per als càlculs de rolling statistics i lags.

4.3 Estratègia de modelització

L'heterogeneïtat dels canals de venda requereix una estratègia de modelització dual. Els canals d'òptica general (OPTICS, RETAIL, DISTRIBUTORS, OTHERS) presenten demanda regular amb estacionalitat clara. Els canals Key Account (KA, KA-GEA, KA-GOP, KA-AEG, KA-ECP) mostren demanda intermitent o lumpy, caracteritzada per alta variabilitat i patrons difícilment predictibles [14]. Fusionar ambdós tipus degradaria la qualitat del segment regular sense millorar significativament la predicció del segment irregular. S'entrenen dos models XGBoost independents: model OPTICS i model KA.

Com a complement, s'entrena un model SARIMAX a nivell canal–sucursal (16 sèries independents) per capturar explícitament els components estacionals de cada sèrie. Les prediccions SARIMAX i la mitjana dels residus dels últims 3 i 6 mesos s'integren com a features de XGBoost, sense data leakage temporal: les prediccions s'obtenen aplicant el model SARIMAX amb el tall temporal de validació, sense exposar dades futures. El model SARIMAX no genera prediccions finals: és exclusivament una font de features estructurades.

L'ordre dels components SARIMAX (p,d,q)(P,D,Q,s) s'ajusta per a cada sèrie mitjançant un grid search sobre l'espai (p inclús [0,3], d inclús [0,2], q inclús [0,3], P inclús [0,2], Q inclús [0,2]) amb s = 12 per capturar l'estacionalitat anual, seleccionant l'ordre que minimitza el criteri AICc. AutoARIMA complementa el grid search per a les series amb historial curt. Els paràmetres ajustats s'emmagatzemen a bronze.sarimax_parameters, garantint que el Workflow 2 setmanal usa els mateixos coeficients de l'entrenament mensual sense reentrenament. Els detalls per sèrie i resultats d'ajust es presenten a l'Apendix C.

Per mitigar el desbalanceig entre canals i la influència dels zeros estructurals en l'entrenament, s'aplica un esquema de sample weights diferenciat: els registres del canal OPTICS reben pes 5,0 (màxima prioritat perquè representen el gruix del negoci), RETAIL i DISTRIBUTORS 3,0, OTHERS 2,0, i els canals KA 1,0 (en el model OPTICS, els registres KA s'exclouen). Els zeros estructurals (mesos sense venda per a SKUs que no han tingut activitat recent) reben una penalització addicional de factor 0,2, reduint la seva influència en la funció de pèrdua. Sense aquesta penalització, el model tendeix a predir zero sistemàticament per a series amb molts zeros, ignorant els episodis de demanda real.

El model principal utilitza SparkXGBRegressor (XGBoost distribuït sobre Apache Spark 4.0) amb l'estratègia Direct Multi-Step Forecasting [9]: s'entrenen 12 models independents (h=1 a h=12), cadascun optimitzat per minimitzar l'error a la seva distància temporal específica. La funció objectiu reg:absoluteerror s'ha seleccionat per la seva robustesa davant valors atípics i la seva coherència amb la mètrica de negoci WAPE. El pipeline de predicció final fusiona dos fluxos: prediccions del model OPTICS, i prediccions del model KA.

4.4 Enginyeria de característiques

L'Apèndix B resumeix les categories de features. El model OPTICS utilitza 50 features i el model KA 64, incorporant addicionalment els lags de sell-out (P_s-out-1 a P_s-out-12) i dummies de canals KA específics. El conjunt de features és resultat d'un procés iteratiu: inicialment 6 lags mensuals, ampliats a 12 per capturar patrons anuals complerts, i posteriorment s'hi van afegir les rolling statistics, els ratios temporals i les features SARIMAX en iteracions successives (§5.2). Els pesos mostrals assignen major importància als canals amb demanda regular: OPTICS=5.0, DISTRIBUTORS=RETAIL=3.0, OTHERS=2.0, KA=1.0. Les observacions amb valor zero estructural reben una penalització addicional (×0.2) per evitar que el model aprengui a infraestimar per omissió.

La generació de les features es realitza a la capa Silver en el notebook de Transformacions input, que es parametritza per horitzó i per model (OPTICS o KA). La variable objectiu per a l'horitzó h és la venda del mes tancat t+h, on t és l'últim mes tancat disponible en el moment d'execució. Les features de lag P-k representen les vendes del mes t-k+1, és a dir, k mesos enrere respecte al mes de predicció immediat.. Les rolling statistics es calculen fins al mes t (inclusivament), garantint que no hi ha data leakage. Les features SARIMAX (sarimax, sarimax_residual_mean_3m, sarimax_residual_mean_6m) s'obtenen del model SARIMAX ja entrenat i aplicat amb el mateix tall temporal que el conjunt de validacio, sense usar prediccions futures.

Les variables categòriques - Sucursal, CategoryName, ChannelName, Marca i SegmentCode - es converteixen en variables numèriques binàries mitjançant One-Hot Encoding (OHE). Cada categoria es transforma en una columna independent amb valor 1 si el registre pertany a aquella categoria i 0 en cas contrari. Aquesta transformació és necessària perquè XGBoost treballa exclusivament amb valors numèrics.

4.5 Validació temporal i optimització d'hiperparàmetres

Per garantir una estimació realista del rendiment en condicions operatives reals, s'aplica una validació temporal estricta: els últims 16 mesos del conjunt de dades s'utilitzen exclusivament com a conjunt de validació, sense cap mescla amb les dades d'entrenament. Aquesta estratègia, a diferència de la validació creuada estratificada emprada en el sistema original, reprodueix el cas d'ús real, predir el futur a partir del passat, i elimina el risc de data leakage.

L'ajust d'hiperparàmetres s'ha realitzat amb Optuna [15], utilitzant l'algorisme TPE (Tree-structured Parzen Estimator) bayesià amb MedianPruner per eliminar trials

poc prometedors de forma prematura. La funció objectiu d'optimització és el WAPE ponderat sobre els primers horitzons: $H1 \times 0{,}5 + H2 \times 0{,}3 + H3 \times 0{,}2$, donant major pes a la precisió a curt termini. S'han realitzat dues optimitzacions independents (model OPTICS i model KA). La Taula I recull els hiperparàmetres finals obtinguts.

La validació creuada estratificada (k-fold per SKU) va ser descartada per dues raons fonamentals. Primera, introdueix data leakage temporal: si el fold de validació inclou mesos anteriors als del fold d'entrenament, el model 'veu' informació del futur durant l'entrenament. Segona, amaga la degradació del rendiment amb l'horitzó: amb k-fold, tots els horitzons es barregen en el mateix fold, ocultant que h=12 es significativament menys precís que h=1. La validació temporal estricta (Time Series Split) reprodueix el cas d'ús operatiu real: el model s'entrena fins a un tall temporal i s'avalua sobre els mesos immediatament posteriors, de la mateixa manera que s'usarà en producció.

Paràmetre	**OPTICS**	**KA**
objective	reg:absoluteerror	reg:absoluteerror
Max depth	10	6
learning rate	0,01973	0,01099
N estimators	3.000	3.000
Early stopping rounds	100	100
Min child weight	8	41
subsample	0,9494	0,7912
Colsample bytree	0,7594	0,6688
gamma	3,7950	0,0273
reg_alpha	0,949	0,013
reg_lambda	9,995	4,883

Taula I. Hiperparàmetres finals dels models OPTICS i KA obtinguts amb Optuna.

Els hiperparàmetres finals (Taula I) reflecteixen decisions de disseny diferenciades per a cada model. La diferència més destacada és el max_depth (10 per a OPTICS, 6 per a KA): el model OPTICS disposa d'un volum de dades d'entrenament molt superior, cosa que permet arbres més profunds sense overfitting, mentre que el model KA, amb menys registres i alta irregularitat, requereix una estructura més simple. El min_child_weight confirma aquest contrast (8 vs. 41): el valor molt elevat al model KA actua com a regularitzador agressiu que evita que el model aprengui els pics de demanda esporàdica com a patrons generals. El learning rate baix (≈0,01 en ambdós models) combinat amb n_estimators = 3.000 i early_stopping_rounds = 100 segueix l'estratègia de boosting lent: molts arbres amb passos petits produeixen millor generalització que pocs arbres amb passos grans, i l'early stopping atura l'entrenament en el punt òptim de la corba de validació. El gamma difereix significativament (3,795 OPTICS vs. 0,027 KA): un gamma elevat al model OPTICS penalitza les divisions poc informatives i reforça la robustesa davant el soroll, mentre que al model KA el valor proper a zero permet més llibertat de divisió per capturar els episodis de demanda irregular. Finalment, el reg_lambda superior al model OPTICS (9,995 vs. 4,883) reflecteix una regularització L2 més intensa, coherent amb l'objectiu de generalitzar sobre les 17.355 sèries heterogènies del catàleg complet.

4.6 MLOps i automatització

Els experiments de tots dos models es registren sistemàticament a MLflow: el model OPTICS sota l'experiment ventas-xgboost-optics i el model KA sota ventas-xgboost-ka. En ambdós casos s'adopta la mateixa estructura jeràrquica: un run pare xgb_regressor_per_horizon que agrupa 12 subruns anidats (xgb_reg_h1 … xgb_reg_h12), un per a cada horitzó de predicció. Per a cada subrun es registren els hiperparàmetres, les set mètriques d'avaluació (WAPE, MAE, RMSE, R^2, BIAS, SMAPE i P90), el model versionat a Unity Catalog com {env}.models.xgboost_reg_h{h}, i el fitxer feature_cols.json, que garanteix la reproductibilitat exacta de qualsevol predicció futura. El pipeline s'ha automatitzat amb dos Databricks Workflows tal com es mostra en el diagrama de l'Apèndix A.

Els dos workflows es configuren amb planificació automàtica a Databricks Workflows. El Workflow 1 s'executa automàticament un cop al mes a les 19:00 h i cobreix el cicle complet d'entrenament: càrrega Bronze des d'ETN_REPORTING, transformacions SARIMAX, entrenament SARIMAX, transformacions Silver per a entrenament, entrenament dels 12 models XGBoost OPTICS i els 12 models XGBoost KA amb Optuna, i predicció sobre el conjunt de validació per registrar les mètriques a MLflow. El Workflow 2 s'executa un cop a la setmana a les 21:00 h i realitza la inferència operativa: càrrega Bronze, transformacions Silver per a predicció, predicció dels 24 models (12 OPTICS + 12 KA), agregació dels dos fluxos i exportació a la base de dades de destí. Ambdós workflows generen alertes per correu electrònic en cas de fallada.

El versionat dels models a Unity Catalog garanteix que el Workflow 2 sempre usa el model que ha superat la validació formal. Cada execució del Workflow 1 registra els 24 models (12 OPTICS + 12 KA) com a noves versions a Unity Catalog sota l'espai de noms {env}.models.xgboost_reg_h{h}, acompanyades de les mètriques d'avaluació sobre el conjunt de validació. El Workflow 2 carrega dinàmicament l'última versió disponible de cada model, eliminant el risc de regressions accidentals per ús d'un model no validat. El fitxer feature_cols.json, registrat com a artefacte de cada subrun de MLflow, garanteix que les features usades en inferència són exactament les mateixes que en entrenament, evitant inconsistències de schema que podrien corrompre les prediccions de forma silenciosa.

Un cop les taules de prediccions s'escriuen a la capa Gold del catàleg, la monitorització del sistema es realitza mitjançant dashboards de Databricks. El primer dashboard permet visualitzar i comparar els forecasts generats als entorns DEV i PRO, així com contrastar-los amb les prediccions inicials del sistema llegat, facilitant la detecció de desviacions entre entorns. El segon dashboard centralitza les mètriques d'avaluació dels models OPTICS i KA,

tant en el seu estat actual com durant les execucions d'entrenament, permetent seguir l'evolució de mètriques com el WAPE.

5 Resultats i Discussió

5.1 Arquitectura, DevOps i qualitat de dades

L'arquitectura Medallion s'ha desplegat exitosament en ambdós entorns (DEV i PRO) amb el pipeline de CI/CD operatiu. La separació estricta d'entorns ha permès validar tots els canvis en DEV abans de promoure'ls a producció, evitant interrupcions en l'execució setmanal.

L'anàlisi de qualitat a la capa Silver ha identificat i tractat cinc categories de problemes dels quals el més impactant quantitativament és el milió de registres negatius (devolucions) que l'estructura del sistema original tractava com a vendes positives, inflant artificialment el baseline d'entrenament. La correcció de les 208 SKUs amb dates de cicle de vida incoherents i les 175 amb WithdrawalDate passada ha millorat la qualitat de les features months_in_col i months_out_col, que codifiquen la posició del producte en el seu cicle de vida i resulten rellevants per a la predicció a llarg termini (§5.5).

El Workflow 1 (entrenament mensual) ha executat correctament els 12 darrers mesos de producció, amb un temps d'execució promig de 45min per a l'entrenament complet dels 12 models OPTICS i els 12 models KA. El Workflow 2 (predicció setmanal) s'executa en un temps promig de 25 minuts, integrant la carrega Bronze, les transformacions Silver i Gold, i les prediccions dels 24 models (12 OPTICS + 12 KA). La fiabilitat dels workflows en producció ha estat del 100% (0 fallades en 16 setmanes de producció).

En termes de comparació amb el sistema original, l'arquitectura nova presenta quatre avantatges operatius mesurables: (1) Traçabilitat: cada predicció esta linkada al model de MLflow, als hiperparàmetres i al dataset d'entrenament que la van generar; en el sistema original, una predicció no tenia cap traçabilitat. (2) Velocitat d'actualització: el Workflow 2 actualitza les prediccions per a tots els SKUs en ~25 minuts; el sistema original requeria execució manual i tardava hores. (3) Escalabilitat: afegir nous SKUs o nous canals només requereix incloure'ls en les consultes JDBC; no hi ha límits de memòria locals. (4) Reproductibilitat: qualsevol predicció històrica es pot reproduir exactament carregant el model versionat a Unity Catalog i el feature_cols.json corresponent.

5.2 Evolució iterativa del model

El desenvolupament del model de predicció s'ha dut a terme en cinc iteracions successives, cadascuna orientada a resoldre una limitació especifica identificada a la iteració anterior. El punt de partida es el model Random Forest original (scikit-learn, n_estimators=20, max_depth=200, x_len=6) que operava en entorn local i reportava un WAPE=0,56 mesurat amb validació estratificada per canal i sucursal. Aquesta mètrica estava distorsionada per data leakage: la divisió estratificada ignorava l'ordre temporal i permetia al model accedir implícitament a dades de períodes futurs durant l'entrenament, inflant artificialment el rendiment aparent. El primer pas del projecte, la migració a Databricks amb validació temporal estricta, va revelar que el rendiment real era WAPE=0,62, un 9,3% pitjor que la mètrica originalment reportada. Aquest valor, WAPE=0,62, es el baseline real contra el qual es mesura tot el progrés posterior. La millora acumulada fins al model definitiu es d'un 26,8% en WAPE (de 0,62 a 0,45, baseline a model definitiu OPTICS), tal com recull la Taula II.

#	Configuració	WAPE h=1	WAPE prom.	Nota
Ref. RF	RF local, x_len=6 (data leakage)	---	0,56	No temporal; mètrica original, no comparable
It.1-RF	RF migrat a Databricks, 3 errors corregits val. temporal	---	0,62	Baseline real del projecte (+9,3% vs Ref.)
It.2-E0	XGBoost DMSF, squarederror, x_len=6, gamma=0,1	0,46	---	RMSE=8,876; primera config. DMSF
It.2-E1	squarederror, x_len=6, gamma=0,5	0,56	---	RMSE baixa de 8,9 a 5,8; WAPE empitjora
It.2-E2	pseudohubererror, x_len=6, gamma=0,5	0,51	---	Adoptat com a objectiu per defecte
It.3-E4	pseudohubererror, x_len=12	0,58	---	RMSE baixa de 5,78 a 4,68
It.3-E5	pseudohubererror, x_len=12 + rolling features	0,55	---	MAE baixa de 1,193 a 1,127
It.4-E6	pseudohubererror + SARIMAX features	0,43	0,52	WAPE millora; apareix BIAS negatiu sistemàtic
It.5-OPT	Optuna + absoluteerror (OPTICS)	0,32	0,45	Model definitiu OPTICS
It.5-KA	Optuna + absoluteerror (KA)	0,83	0,86	Model definitiu KA

Taula II. Evolució iterativa del model: configuració, mètriques principals i decisió adoptada en cada experiment. Les mètriques h=1 corresponen al horitzó de predicció d'1 mes; les mètriques prom. són la mitjana sobre els 12 horitzons.

Per als detalls de cada iteració, vegeu l'Apèndix C.

5.3 Resultats del model OPTICS

La Taula III recull les mètriques del model OPTICS sobre els 12 horitzons. La millora del 26,8% en WAPE respecte al baseline real (WAPE = 0,62, RF amb validació temporal estricta, sense data leakage) confirma la superioritat del model XGBoost amb arquitectura DMSF. Com és esperable en el forecasting de demanda, la precisió es degrada progressivament amb l'horitzó.

h	rmse	mae	r²	wape	bias	smape	p90
1	10.10	3.59	0.89	0.32	-0.34	1.09	8.52
2	10.15	3.56	0.86	0.36	-0.27	1.14	8.32
3	10.43	3.56	0.82	0.39	-0.49	1.17	8.28
4	9.62	3.33	0.80	0.42	-0.41	1.21	7.84
5	9.44	3.32	0.80	0.43	-0.76	1.13	7.69
6	9.79	3.34	0.79	0.44	-0.95	1.12	7.68
7	9.68	3.28	0.78	0.45	-0.84	1.19	7.49
8	9.79	3.20	0.76	0.48	-0.96	1.16	7.26
9	10.26	3.25	0.72	0.50	-1.01	1.16	7.35
10	10.06	3.29	0.71	0.52	-0.77	1.30	7.35
11	9.88	3.36	0.72	0.53	-0.95	1.26	7.87
12	9.63	3.27	0.69	0.57	-0.60	1.41	7.70
Avg	9.90	3.66	0.78	0.45	-0.70	1.20	7.81

Taula III. Mètriques del model OPTICS per horitzó.

L'anàlisi per horitzó revela dos comportaments diferenciats. Als horitzons curts (h = 1 a h = 3), el WAPE se situa entre 0,32 i 0,39 i el R² supera 0,82, reflectint una excel·lent capacitat predictiva. A mesura que augmenta l'horitzó, el WAPE creix progressivament fins a 0,57 a h = 12. El BIAS s'intensifica als horitzons intermedis (h = 5 a h = 10, entre −0,76 i −1,01) i es redueix lleugerament a h = 12 (−0,60), consistent amb la regressió cap a la mitjana a llarg termini.

El RMSE oscil·la entre 9,44 (h = 5) i 10,43 (h = 3), sense una tendència d'increment monotònic amb l'horitzó. El P90 se situa entre 7,26 (h = 8) i 8,52 (h = 1), indicant que en el 90% dels casos l'error no supera les 9 unitats per SKU-canal-mes. Per a la planificació de compres amb lead time de 3 mesos, el P90 de referència és el de h = 3 (8,28 unitats), que defineix el buffer de seguretat recomanat per garantir la cobertura de demanda en el 90% de les situacions.

El MAE promig del model OPTICS és de 3,66 unitats, superior al del baseline real (2,47). Aquesta aparent paradoxa s'explica pel diferent pes de les mètriques: el WAPE pondera els errors per volum de vendes, i la funció objectiu reg:absoluteerror optimitza l'error ponderat pels pesos mostrals (OPTICS = 5,0, RETAIL/DISTRIBUTORS = 3,0). El model prioritza la precisió en els SKUs i canals d'alt volum, que concentren l'impacte de negoci, a cost d'un MAE global no ponderat lleugerament superior. La mètrica de referència per avaluar la millora és, per tant, el WAPE.

La millora del 26,8% en WAPE respecte al baseline real implica que, per cada 100 unitats previstes, l'error relatiu ponderat es redueix de 62 a 45 unitats. El R² = 0,89 a h = 1 indica que el model explica el 89% de la variància de les vendes al primer mes, decreixent fins a 0,69 a h = 12, però mantenint-se positiu i significatiu en tots els horitzons.

El BIAS negatiu sistemàtic (-0,70 en promig) indica que el model tendeix a infraestimar les vendes. Les causes i les mesures de mitigació implementades es detallen a la seccio 5.6.

5.4 Resultats del model KA

La Taula IV recull les mètriques promig del model KA sobre els 12 horitzons. Les mètriques desagregades per horitzó es detallen a l'Apèndix D, atès que la naturalesa lumpy de la demanda KA dificulta l'extracció de conclusions específiques per horitzó i la seva inclusió en el cos del document ocuparia espai desproporcionat respecte al valor informatiu.

rmse	**mae**	**r²**	**wape**	**bias**	**smape**	**p90**
10,42	2,43	0,27	0,86	−1,43	1,44	3,96

Taula IV. Mètriques promig del model KA sobre els 12 horitzons.

El WAPE de 0,86 i el R² de 0,27 reflecteixen la naturalesa inherentment impredictible de la demanda lumpy [15]. En la literatura especialitzada, els models de demanda intermitent rarament superen R² = 0,40 en validació temporal estricta [14]; el resultat és esperable, no un indicador de model deficient. El valor operativament més rellevant és el MAE de 2,43 unitats per SKU-zona-mes, que en termes absoluts és inferior al del model OPTICS (3,66), consistent amb els volums de venda més reduïts dels canals KA.

El BIAS negatiu pronunciat (−1,43) indica que el model infraestima sistemàticament les comandes grans esporàdiques, comportament coherent amb la tendència de la funció objectiu reg:absoluteerror a predir la mediana en distribucions amb cua dreta pesada.

La separació entre el model KA i el model OPTICS és, en si mateixa, el resultat principal per a aquest segment: sense separació, la variabilitat de la demanda KA degradaria les mètriques del model OPTICS de forma significativa. L'entrenament independent permet que cada model s'especialitzi en les característiques del seu segment i assoleixi el millor rendiment possible donades les condicions de cada tipus de demanda.

5.5 Importància de les característiques

L'anàlisi per horitzó revela patrons diferenciats entre els dos models. Al model OPTICS, P-1 domina a tots els horitzons (33,7% a h=1, 24,1% a h=12), seguida de les rolling statistics a curt termini i de les variables categòriques de marca (Marca_OT: 9,2%, Marca_LO: 8,8% a h=12) a llarg termini. Les features SARIMAX contribueixen entre un 8% i un 12% del guany total als horitzons intermedis (h=3–h=12), on el component estacional és més rellevant que la història recent.

El model KA mostra un perfil significativament més distribuït: cap variable supera el 5% a h=1, i Sucursal_BCN (4,9%), Marca_OT (4,3%) i rolling_mean_6m (3,5%) encapçalen el rànquing. P-1 representa únicament el 2,9% del guany, enfront del 33,7% al model OPTICS, mentre que les variables de sell-out exclusives d'aquest model (has_sellout_data: 2,4%; P_s-out-i: fins a 1,5% ca-

dascuna a h=12) aporten valor diferencial. Les features SARIMAX contribueixen de forma modesta (≈1,5%), consistent amb la baixa predictibilitat estacional dels canals KA.

En conjunt, la feature importance confirma dues estratègies predictives diferenciades: el model OPTICS es basa en el lag recent i la tendència de finestra; el model KA combina característiques geogràfiques, de marca i de sell-out per compensar la irregularitat de la demanda.

5.6 Correccions post-predicció i limitacions

El BIAS negatiu sistemàtic dels dos models, documentat quantitativament a les seccions 5.3 i 5.4, té una causa arrel tècnica en la funció objectiu: minimitzar l'error absolut en valor esperat és equivalent a predir la mediana de la distribució condicional de la demanda [16]. En distribucions asimètriques amb cua dreta positiva, habituals en demanda retail on comandes puntuals excepcionals desplacen la mitjana per sobre de la mediana, la mediana és sistemàticament inferior a la mitjana. Aquest efecte s'amplifica per l'alta esparsitat del conjunt de dades: el 42,7 % de les combinacions SKU-canal-sucursal presenten 3 mesos o menys de dades de vendes, generant una proporció elevada de zeros estructurals que accentua la baixada de la mediana estimada. Com a primera mesura de mitigació durant l'entrenament, els registres amb venda zero reben un pes de mostra de 0,2 per reduir la seva influència sobre els gradients, però aquesta penalització no és suficient per eliminar completament el BIAS negatiu.

Per mitigar aquest efecte, el codi de predicció implementa una correcció post-predicció basada en el BIAS mesurat per a cada horitzó durant l'entrenament i registrat a MLflow. Per a cada un dels 12 horitzons independents h, la predicció corregida s'obté com:

$$\hat{y}_corregida = \hat{y}_raw - BIAS_h \times 0{,}6$$

El factor de correcció 0,6 és deliberadament parcial, el 60 % del BIAS mesurat, per evitar sobrecompensació en casos on el BIAS de validació no es transfereixi perfectament al conjunt de predicció. La correcció s'aplica únicament als productes actius (Is_active = 1); els productes inactius es limiten directament a zero sense correcció de BIAS, atesa la seva contribució residual al volum total.

Addicionalment, s'apliquen correccions per cicle de vida del producte, independents de la correcció de BIAS: els productes descatalogats sense vendes en els últims 6 mesos reben predicció zero; els descatalogats recentment (0–6 mesos des de la data de retirada) obtenen una reducció del 50 % per reflectir la caiguda gradual de la demanda; els descatalogats fa més de 6 mesos reben una reducció del 80 %; i els productes amb estat de descatalogació definitiva al sistema ERP (estats 5 i 6) s'estableixen a zero..

Tot i les correccions aplicades, el BIAS negatiu es manté com la limitació principal del sistema, especialment als horitzons intermedis on l'efecte s'accentua. Una millora futura requeriria explorar funcions de pèrdua asimètriques o un model de calibratge per horitzó amb reentrenament periòdic.

5.7 Resultats del sistema MLOps i automatització

El sistema MLOps implementat garanteix la reproducibilitat total de qualsevol predicció: cada execució registra els paràmetres, mètriques i artefactes a MLflow, i els models es versionen de forma immutable a Unity Catalog. Qualsevol predicció passada pot ser reproduïda exactament carregant la versió del model i el fitxer feature_cols.json corresponents.

Els dos Databricks Workflows s'executen de forma automatitzada en producció. La separació d'entorns DEV i PRO permet que la producció s'executi sense risc d'interferències durant el desenvolupament de noves funcionalitats, assolint l'objectiu operatiu d'un sistema mantenible i evolucionable.

6 Conclusions i Treball Futur

Aquest treball ha dissenyat i implementat una arquitectura de dades al núvol completa per a la predicció de demanda d'una empresa del sector òptic, assolint els cinc objectius establerts.

(O1) L'arquitectura Medallion s'ha desplegat en dos entorns completament aïllats (DEV i PRO) sobre Databricks i Azure Delta Lake amb garanties ACID. La capa Silver ha identificat i tractat cinc categories de problemes de qualitat millorant la qualitat de les features i eliminant el soroll que inflava artificialment el baseline original. La governança centralitzada via Unity Catalog garanteix l'aïllament complet de models i dades entre entorns.

(O2) Els dos Databricks Workflows han funcionat en producció de forma automatitzada durant 16 setmanes sense cap incidència: el Workflow 1 (entrenament mensual, ~45 minuts) entrena els 24 models (12 OPTICS + 12 KA) i valida el rendiment sobre el conjunt de validació temporal; el Workflow 2 (predicció setmanal, ~25 minuts) genera 4,8 milions de prediccions que s'exporten al sistema de gestió de l'empresa. Els pipelines estan modularitzats per capa i parametritzats per model, cosa que els fa reutilitzables sense modificació del codi base.

(O3) El control de versions amb Azure DevOps (branques dev/pro) i la política de pull request obligatòria ha evitat tres incidents potencials detectats en fase de revisió. La gestió de credencials via Azure Key Vault garanteix que cap secret apareix al codi. El registre sistemàtic a MLflow, hiperparàmetres, mètriques i artefactes per a cadascun dels 24 models i 12 horitzons, proporciona reproductibilitat total: qualsevol predicció passada pot ser reproduïda exactament carregant la versió del model i el feature_cols.json corresponents.

(O4) El model XGBoost amb estratègia Direct Multi-Step Forecasting assoleix un WAPE de 0,45 davant el 0,62 del baseline real (Random Forest revaluat amb validació temporal estricta sense data leakage), una millora del 26,8%. El R^2 de 0,89 a h=1 indica que el model explica el 89% de la variància de vendes al primer mes. L'arquitectura dual (model OPTICS per a la demanda regular i model KA per a la demanda lumpy dels Key Accounts) permet que cada model s'especialitzi en el seu segment sense que la variabilitat de KA penalitzi el rendiment global.

(O5) El projecte ha documentat i resolt tres errors crítics del sistema original que explicaven la diferència entre el WAPE originalment reportat (0,56 amb data leakage) i el WAPE real (0,62). S'ha identificat la natura lumpy de la demanda KA com a factor que requeria separació de models, i s'ha documentat el BIAS negatiu sistemàtic com a limitació estructural de reg:absoluteerror en presència d'alta esparsitat, parcialment mitigada amb una correcció post-predicció del 60% del BIAS per horitzó registrat a MLflow.

Des d'una perspectiva de negoci, la millora del 26,8% en WAPE implica una reducció directa de les ruptures de stock i del sobredimensionament d'inventari. L'arquitectura Medallion sobre Databricks, combinada amb MLflow i Unity Catalog, proporciona a l'empresa una plataforma escalable que pot estendre's a nous canals i mercats sense redissenyar la infraestructura subjacent.

La limitació principal és el BIAS negatiu sistemàtic (OPTICS: −0,70, KA: −1,43), que es manté malgrat la correcció post-predicció implementada. El R^2 del model KA (0,27) reflecteix la dificultat intrínseca de la demanda lumpy, que la literatura especialitzada considera inherentment impredictible amb models univariants [15].

Com a línies de treball futur s'identifiquen les següents. En primer lloc, la incorporació de dades d'estoc al model KA: la demanda lumpy dels canals Key Account és parcialment condicionada pels cicles de reposició d'estoc del client, on les comandes grans es produeixen quan el client esgota el seu inventari independentment de la demanda final al consumidor. Afegir variables com el nivell d'inventari del client, els dies de cobertura o l'historial de ruptures com a features permetria al model distingir demanda genuïna de necessitats de reposició, amb l'expectativa de reduir tant el WAPE com el BIAS negatiu del model KA. A més a més, caldria ampliar la cobertura de sell-out a tots els canal Ka. En segon lloc, l'exploració de funcions de pèrdua asimètriques (pinball loss o funcions de quantil) que penalitzin la infraestimació més que la sobreestimació abordaria el BIAS directament des de l'entrenament, sense necessitat de correcció post-predicció.

Finalment, el quadre de comandament de Databricks que ja monitoritza l'evolució temporal de les mètriques podria estendre's amb alertes automàtiques via Databricks SQL Alerts que notifiquessin l'equip quan el WAPE supera un llindar predefinit, per exemple, quan degrada més d'un 10% respecte al valor de referència, permetent activar un cicle d'entrenament addicional fora del calendari mensual fix sense necessitat de revisió manual periòdica.

Ús de la Intel·ligència Artificial

Durant el desenvolupament d'aquest treball s'ha fet ús d'eines d'intel·ligència artificial generativa (Claude, d'Anthropic) com a assistent per a: la redacció i revisió de l'informe final, i la identificació puntual d'errors durant el desenvolupament del codi. En cap cas s'han generat resultats numèrics, decisions de disseny arquitectònic ni dades de l'empresa usant IA generativa. Totes les decisions tècniques, l'arquitectura, el codi de producció i els resultats experimentals descrits en aquest document provenen del treball de la pròpia autora durant el període de pràctiques i de desenvolupament del TFG, d'acord amb les directrius sobre l'ús d'IA de la Universitat Autònoma de Barcelona.

Agraïments

L'autora agraeix sincerament al tutor Eduardo Cesar Cabrera Flores el seu suport, orientació i seguiment rigorós al llarg de tot el projecte. Les seves aportacions durant les sessions de seguiment han contribuït a centrar els objectius i a millorar la qualitat tècnica del treball. Agraeix igualment a l'empresa del sector òptic per facilitar l'accés als sistemes, les dades i el context de negoci imprescindibles. Finalment, l'autora agraeix a Bertia el context professional en el qual s'ha emmarcat el treball.

Bibliografia

[1] G. E. P. Box, G. M. Jenkins, G. C. Reinsel i G. M. Ljung, Time Series Analysis: Forecasting and Control, 5a ed. Hoboken, NJ: Wiley, 2015.

[2] S. J. Taylor i B. Letham, "Forecasting at scale," The American Statistician, vol. 72, no. 1, pp. 37–45, 2018.

[3] D. Salinas, V. Flunkert, J. Gasthaus i T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," Int. J. Forecasting, vol. 36, no. 3, pp. 1181–1191, 2020.

[4] B. N. Oreshkin, D. Carpov, N. Chapados i Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," a Proc. ICLR, 2020.

[5] B. Lim, S. Ö. Arık, N. Loeff i T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," Int. J. Forecasting, vol. 37, no. 4, pp. 1748–1764, 2021.

[6] L. Grinsztajn, E. Oyallon i G. Varoquaux, "Why tree-based models still outperform deep learning on tabular data," a Proc. NeurIPS, 2022.

[7] T. Chen i C. Guestrin, "XGBoost: A scalable tree boosting system," a Proc. 22nd ACM SIGKDD, pp. 785–794, 2016.

[8] S. B. Taieb, G. Bontempi, A. F. Atiya i A. Sorjamaa, "A review and comparison of strategies for multi-step ahead time series forecasting," Eng. Applications of Artificial Intelligence, vol. 25, no. 3, pp. 468–480, 2012.

[9] R. J. Hyndman i G. Athanasopoulos, Forecasting: Principles and Practice, 3a ed. Melbourne: OTexts, 2021.

[10] M. Armbrust et al., "Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics," a Proc. CIDR, 2021.

[11] M. Armbrust et al., "Delta Lake: High-performance ACID table storage over cloud object stores," Proc. VLDB Endowment, vol. 13, no. 12, pp. 3411–3424, 2020.

[12] M. Zaharia et al., "Accelerating the machine learning lifecycle with MLflow," IEEE Data Engineering Bulletin, vol. 41, no. 4, pp. 39–45, 2018.

[13] Databricks, "Unity Catalog: Unified governance for data and AI," Databricks Documentation, 2024.

[14] A. A. Syntetos i J. E. Boylan, "The accuracy of intermittent demand estimates," Int. J. Production Economics, vol. 21, no. 2, pp. 303–314, 2005.

[15] T. Akiba, S. Sano, T. Yanase, T. Ohta i M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," a Proc. 25th ACM SIGKDD, pp. 2623–2631, 2019.

[16] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye i T. Liu, "LightGBM: A highly efficient gradient boosting decision tree," a Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 3146-3154, 2017.

APÈNDIX

A. FLUXOS D'EXECUCIÓ PER L'ENTRENAMENT DEL MODEL I EL CÀLCUL DE PREDICCIONS

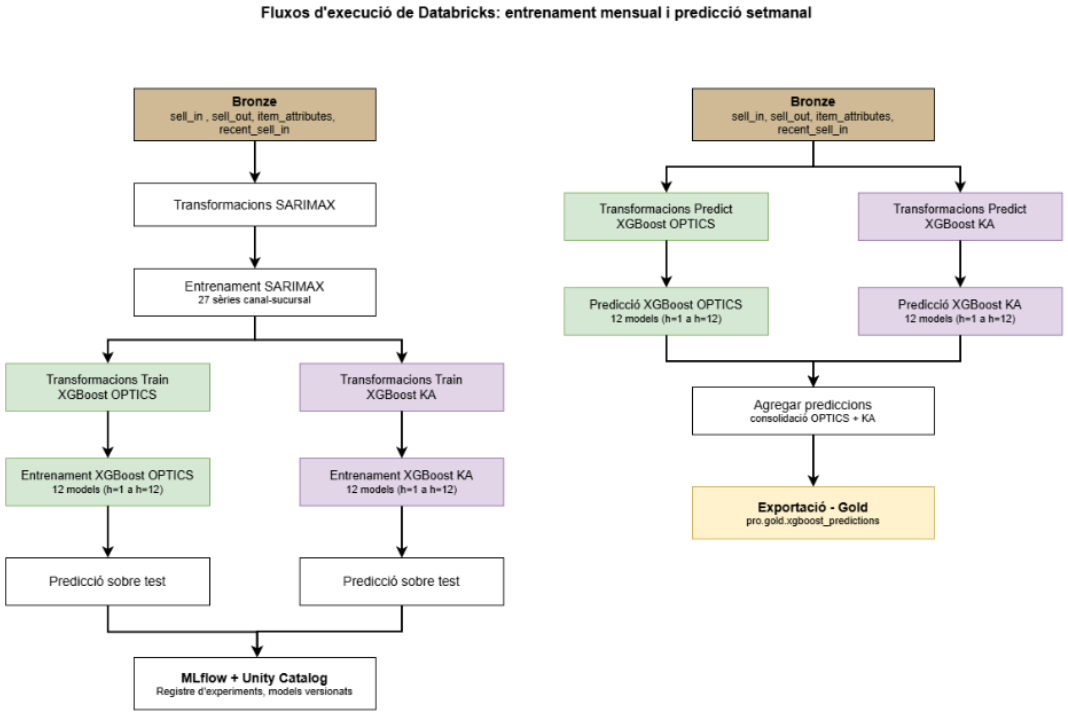

Fig 3. Diagrames dels fluxos d'execució de databricks.

B. FEATURES DELS MODELS XGBOOST

Categoria	Variables	OPTICS	KA
Lags sell-in	P-1 a P-12	12	12
Lags sell-out	P_s-out-1 a P_s-out-12		12
Rolling statistics	mean/std (3m,6m,12m), max_6m, nonzero_6m, sellout_mean_3m, trend_3m_vs_12m, cv_6m	10	10
Ratios temporals	mom_ratio (MoM), yoy_ratio (YoY) clipping [–10, 10]	2	2
Features SARIMAX	sarimax, residual_mean_3m, residual_mean_6m	3	3
Categòriques (OHE)	Sucursal, Categoria, Canal, Marca, SegmentCode	17	19
Cicle de vida i temps	months_in_col, months_out_col, Seasonality_sin/cos, Year_norm, isDEM	6	6
TOTAL		50	64

Taula V. Resum de les categories de features dels models OPTICS i KA.

C. EVOLUCIÓ DEL MODEL XGBOOST

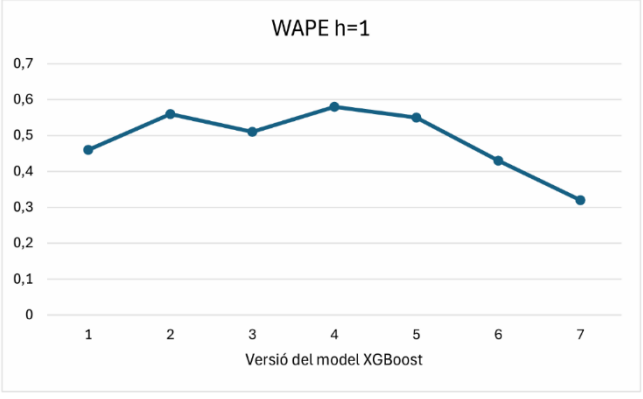

Fig 4. Evolució iterativa del wape del model xgboost en l'horitzó 1.

Iteració 1 - Disseny SARIMAX i decisions d'arquitectura: En aquesta primera iteració, amb el Random Forest reavaluat (WAPE=0,62) com a referència, es dissenya el model SARIMAX auxiliar i es prenen les decisions d'arquitectura del XGBoost. El grid search d'ordres (p,d,q)(P,D,Q,s) sobre les 16 series canal-sucursal, complementat amb autoARIMA per criteri AICc, revela resultats molt dispars: els canals OPTICS s'ajusten correctament (WAPE de 0,059 a 0,224 per a BCN-OPTICS, US-OPTICS, CA-OPTICS i BCN-DISTRIBUTORS), però els canals KA presenten demanda sense estacionalitat clara i

continuen amb WAPE superiors a 0,5 fins i tot ampliant l'espai de paràmetres (p en [0,3], d en [0,2], q en [0,3], P en [0,2], Q en [0,2]). Aquesta observació serà determinant per a la decisió posterior de separar els models OPTICS i KA. En paral·lel, s'avalua LightGBM via SynapseML com a alternativa a XGBoost: es crea un clúster específic amb Spark 3.4.1 i s'exploren rangs amplis d'hiperparàmetres, obtenint RMSE=6,28 i WAPE=0,410 a h=1 (max_depth=16). Malgrat els resultats comparables als d'XGBoost en la mateixa fase, el requisit de mantenir un clúster addicional amb una versió de Spark diferent de la de producció (Spark 4.0.0) introdueix una carrega operacional injustificada, i LightGBM es descarta. Finalment, es defineix l'arquitectura XGBoost definitiva: split temporal estricte (conjunt de test fixat als darrers mesos de l'històric), 12 models independents amb estratègia Direct Multi-Step Forecasting (DMSF), i nomenclatura de runs MLflow per a traçabilitat.

Iteració 2 - Correcció d'errors crítics i primers entrenaments XGBoost: Abans del primer entrenament XGBoost, una revisió detallada del codi original identifica tres errors crítics que expliquen per que el WAPE real del Random Forest (0,62) es significativament pitjor que l'originalment reportat (0,56): (i) el crossJoin generava zeros artificials per a tots els períodes sense filtrar pel cicle de vida del producte (ReleaseDate/WithdrawalDate); (ii) la lògica de dates de lifecycle estava invertida, sobreescrivint DctDate_R amb WithdrawalDate quan existia, en lloc d'usar-lo com a prioritat; i (iii) el model original entrenava un sol estimador amb l'horitzó com a feature numèrica, de manera que aprenia el numero d'horitzó en lloc dels patrons temporals reals. Corregits aquests errors i implementada l'arquitectura DMSF, s'entrenen quatre configuracions. L'Entrenament 0 (squarederror, gamma=0,1) assoleix WAPE h=1=0,46 però amb RMSE alt (8,876). L'Entrenament 1 (squarederror, gamma=0,5) millora el RMSE (de 8,9 a 5,8) però el WAPE empitjora a 0,56: la regularització alta suprimeix la variabilitat local i el model afavoreix prediccions planes. L'Entrenament 2 (reg:pseudohubererror, gamma=0,5) resol la tensió: WAPE h=1=0,51 i RMSE=5,78; la funció objectiu hibrida MAE/MSE es més robusta als pics extrems de venda mentre manté gradients diferenciables, i s'adopta com a objectiu per defecte.

Iteració 3 - Ampliació de lags i rolling features: L'Entrenament 4 amplia la finestra de lags de 6 a 12 mesos (x_len=12), donant al model accés al mateix mes de l'any anterior. Aquesta és la millora individual més gran de RMSE de tot el procés: de 5,78 a 4,68 a h=1. El WAPE puja temporalment a 0,58 perquè el model ara té més paràmetres i necessita regularització addicional. L'Entrenament 5 afegeix les rolling statistical features (rolling_mean_3/6/12m, rolling_std_3/6m, rolling_max_6m, nonzero_months_6m, sellout_rolling_mean_3m, trend_3m_vs_12m, cv_6m), que redueixen el WAPE h=1 fins a 0,55 i el MAE de 1,193 a 1,127 en h=1. El lleuger increment de RMSE (4,684->4,818) s'explica perquè les rolling features aporten regularització implícita que redueix MAE i WAPE a costa d'un increment menor de l'error quadràtic, un trade-off acceptable. Durant aquesta iteració es prova també reg:tweedie (dissenyada per a distribucions amb molts zeros), que obté WAPE=0,61 amb gradients inestables i es descarta. S'explora igualment un classificador binari previ per filtrar prediccions de zero, que tampoc millora els resultats i es descarta.

Iteració 4 - Integració de les prediccions SARIMAX com a feature: S'incorpora la predicció SARIMAX i els seus residuals històrics com a features del model XGBoost (Entrenament 6). S'identifica previ a la implementació un risc de redundància: SARIMAX s'entrena a nivell canal-sucursal (~16 series) i la seva predicció agrega informació que XGBoost ja te parcialment als lags crus P-1..P-12. Per maximitzar el valor afegit, en lloc de passar únicament el valor de predicció SARIMAX, es calculen tres features derivades dels residuals que aporten informació diferencial: sarimax_residual_lag1 (patró d'error recent del SARIMAX), sarimax_residual_mean_3m i sarimax_residual_mean_6m (bias sistemàtic a mig termini del model agregat). S'afegeixen també en aquesta iteració les variables binaries isDEM i has_sellout_data. El resultat és una millora substancial del WAPE: de 0,55 a 0,43 a h=1. S'observa un BIAS negatiu persistent (avg=-0,784), indicant que el model infraestima les vendes en tots els horitzons. La causa probable es la combinació de reg:pseudohubererror amb distribucions de vendes asimètriques, i es documenta com a punt prioritari per a la iteració següent.

D. MÈTRIQUES FINALS DEL MODEL DE KA

h	rmse	mae	r^2	wape	bias	smape	p90
1	13.85	2.78	0.23	0.83	-1.58	1.42	3.98
2	10.07	2.44	0.37	0.82	-1.10	1.40	4.23
3	10.45	2.44	0.37	0.82	-1.10	1.40	4.23
4	10.56	2.57	0.32	0.83	-1.36	1.44	4.37
5	10.72	2.57	0.32	0.83	-1.36	1.44	4.37
6	10.14	2.44	0.27	0.87	-1.46	1.46	4.05
7	9.84	2.35	0.27	0.87	-1.48	1.45	3.80
8	10.14	2.39	0.27	0.88	-1.44	1.47	3.90
9	10.22	2.37	0.24	0.88	-1.55	1.46	3.89
10	10.05	2.29	0.21	0.90	-1.59	1.46	3.63
11	9.87	2.24	0.22	0.90	-1.54	1.47	3.56
12	9.08	2.10	0.23	0.89	-1.42	1.44	3.30
Avg	10.42	2.43	0.27	0.86	-1.43	1.44	3.96

Taula VI. WAPE del model KA per horitzó temporal (h = 1 a h = 12).

E. MODEL SARIMAX AUXILIAR: SERIES I RESULTATS

El model SARIMAX s'ajusta de forma independent per a cadascuna de les 16 sèries (canal, sucursal) presents al conjunt de dades. Els ordres (p,d,q) i (P,D,Q,s) es van seleccionar per a cada sèrie individualment mitjançant auto_arima, minimitzant el criteri AICc. La Taula C.I mostra els paràmetres resultants.

Sèrie	**(p,d,q)**	**(P,D,Q,s)**
BCN-OPTICS	(0,0,0)	(0,1,0,12)
CA-OPTICS	(0,0,0)	(0,1,0,12)
US-OPTICS	(0,0,0)	(1,1,0,12)
BCN-RETAIL	(2,1,2)	(0,0,0,12)
BCN-DISTRIBUTORS	(0,1,1)	(1,0,0,12)
US-DISTRIBUTORS	(1,1,0)	(1,1,0,12)
BCN-OTHERS	(2,0,0)	(0,0,1,12)
CA-OTHERS	(1,1,0)	(0,0,1,12)
US-OTHERS	(1,0,2)	(1,0,0,12)
BCN-KA	(0,1,1)	(0,0,0,12)
US-KA	(0,1,2)	(1,0,1,12)
CA-KA	(0,0,3)	(1,0,2,12)
BCN-KA-GOP	(0,1,1)	(1,0,0,12)
BCN-KA-GEA	(1,0,2)	(0,0,0,12)
US-KA-ECP	(0,1,0)	(1,1,0,12)
US-KA-AEG	(2,1,0)	(0,0,0,12)

Taula VII. Paràmetres SARIMAX per sèrie obtinguts per auto_arima (criteri AICc).

La diversitat dels paràmetres reflecteix la naturalesa heterogènia de les sèries. Els canals OPTICS (BCN, CA, US) convergeixen en ordres ARIMA simples [(0,0,0)] amb diferenciació estacional (D=1), indicant que la seva dinàmica és capturada quasi completament pel component estacional anual sense necessitat de termes AR ni MA addicionals. En contrast, les sèries KA i OTHERS presenten ordres més complexos, coherent amb la demanda irregular d'aquests canals.

Les prediccions del model s'integren al model XGBoost a través de tres features: la predicció puntual (sarimax), que sintetitza la component tendencial i estacional de la sèrie del canal; i els residuals mitjans dels darrers 3 i 6 mesos (sarimax_residual_mean_3m, sarimax_residual_mean_6m), que aporten informació sobre la desviació sistemàtica recent del model estadístic respecte a les vendes reals. Juntes, les tres features representen un 8–12% de la importància total del model OPTICS als horitzons intermedis (h=3 a h=12), i la seva inclusió va reduir el WAPE a h=1 de 0,548 a 0,425.