

PRÉVISION DES VENTES

2026

MACHINE LEARNING

Réalisé par :

Khadidiatou DIAKHATE

Aissatou Sega DIALLO

Jacques ILLY

Haba Fromo FRANCIS

Dior MBENGUE

Elèves ingénieurs statisticiens économistes

Chargée du cours:

Mme Mously DIAW, Lead ML Engineer

Plan

1 CONTEXTE

3 PRÉSENTATION DES DONNÉES

5 PRÉTRAITEMENT ET FEATURE ENGINEERING

7 PRÉSENTATION DU DASHBOARD

2 PROBLÉMATIQUE ET OBJECTIFS

4 ANALYSE EXPLORATOIRE

6 MODÉLISATION

8 LIMITES ET RECOMMANDATIONS

Introduction

Contexte

La grande distribution est un secteur caractérisé par de très forts volumes de vente, une grande diversité de produits et des marges souvent faibles.

Dans ce contexte, une mauvaise anticipation de la demande entraîne soit des ruptures de stock, soit du surstock, ce qui génère des pertes financières importantes, surtout pour les produits périssables.

La prévision de la demande devient donc un enjeu stratégique pour améliorer la performance des supermarchés.



Corporación Favorita

Corporación Favorita est l'une des plus grandes chaînes de supermarchés en Équateur.

Elle possède un réseau de magasins variés et commercialise des milliers de produits, allant des produits alimentaires de base aux produits frais et périssables.

La demande varie fortement selon le type de magasin, la localisation, la période de l'année et les campagnes de promotion.

Pourquoi le Machine Learning ?

Les méthodes classiques de prévision atteignent leurs limites face à :

- la taille massive des données,
- le nombre élevé de facteurs explicatifs,
- les relations non linéaires entre variables.

Le Machine Learning permet de modéliser ces relations complexes et de mieux s'adapter à la diversité des comportements de vente.

?) 2. PROBLÉMATIQUE

La problématique centrale du projet ainsi :

Comment prédire avec précision les ventes journalières par produit et par magasin, en tenant compte des effets temporels, promotionnels, structurels et contextuels afin d'optimiser la gestion des stocks et la prise de décision ?



OBJECTIFS

Objectifs opérationnels

- améliorer la planification des stocks ;
- réduire les ruptures et le surstock ;
- fournir des prévisions exploitables par les équipes métier.

Objectifs scientifiques

- comparer plusieurs approches de modélisation ;
- analyser l'impact des variables explicatives ;
- évaluer la robustesse des modèles dans le temps.

Organisation du projet

Objectif principal

Développer un modèle de Machine Learning capable de prédire les ventes journalières de manière fiable.

Structure du travail

- Analyser et préparer les données
- Concevoir des variables explicatives pertinentes
- Comparer plusieurs modèles de prédiction
- Sélectionner et interpréter le meilleur modèle
- Valoriser les résultats via un dashboard

Présentation des données

Données : vue d'ensemble

Les données proviennent de Corporación Favorita et sont organisées en plusieurs fichiers décrivant :

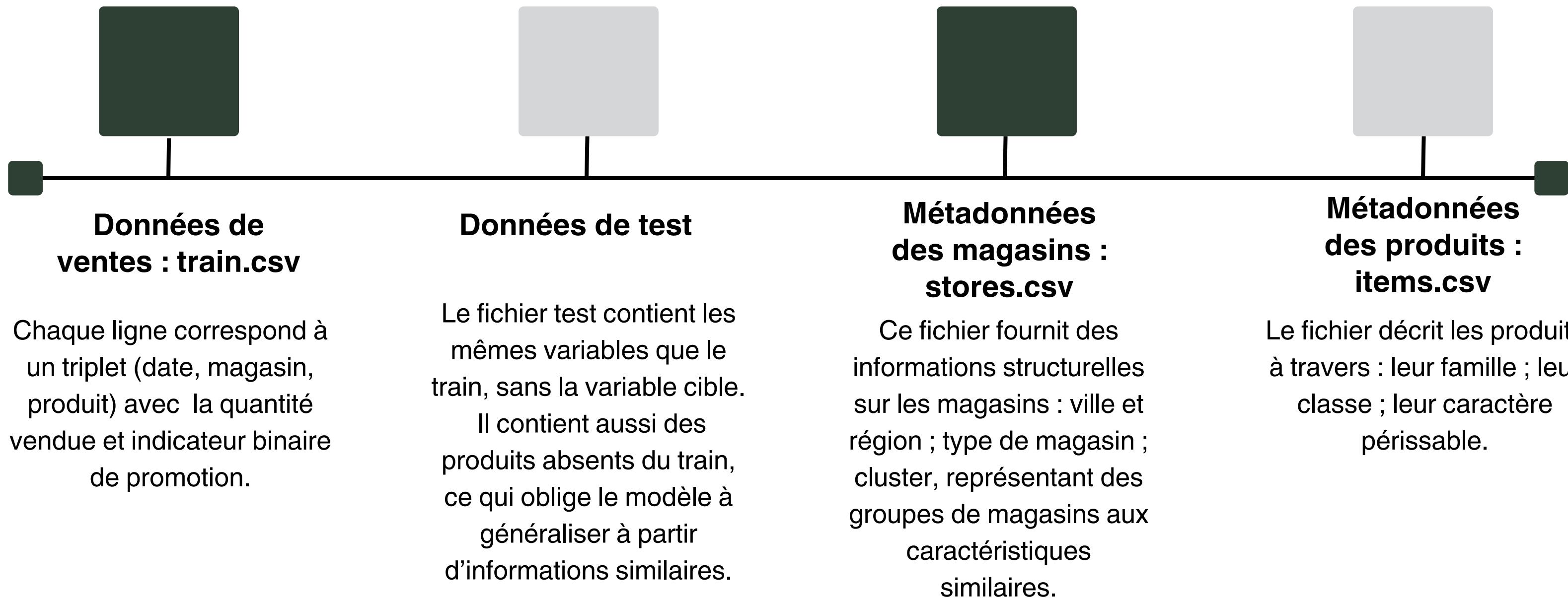
- les ventes,
- les magasins,
- les produits,
- l'environnement économique,
- les jours fériés.

Cette organisation impose un travail préalable de jointure et de cohérence des données, essentiel avant toute modélisation



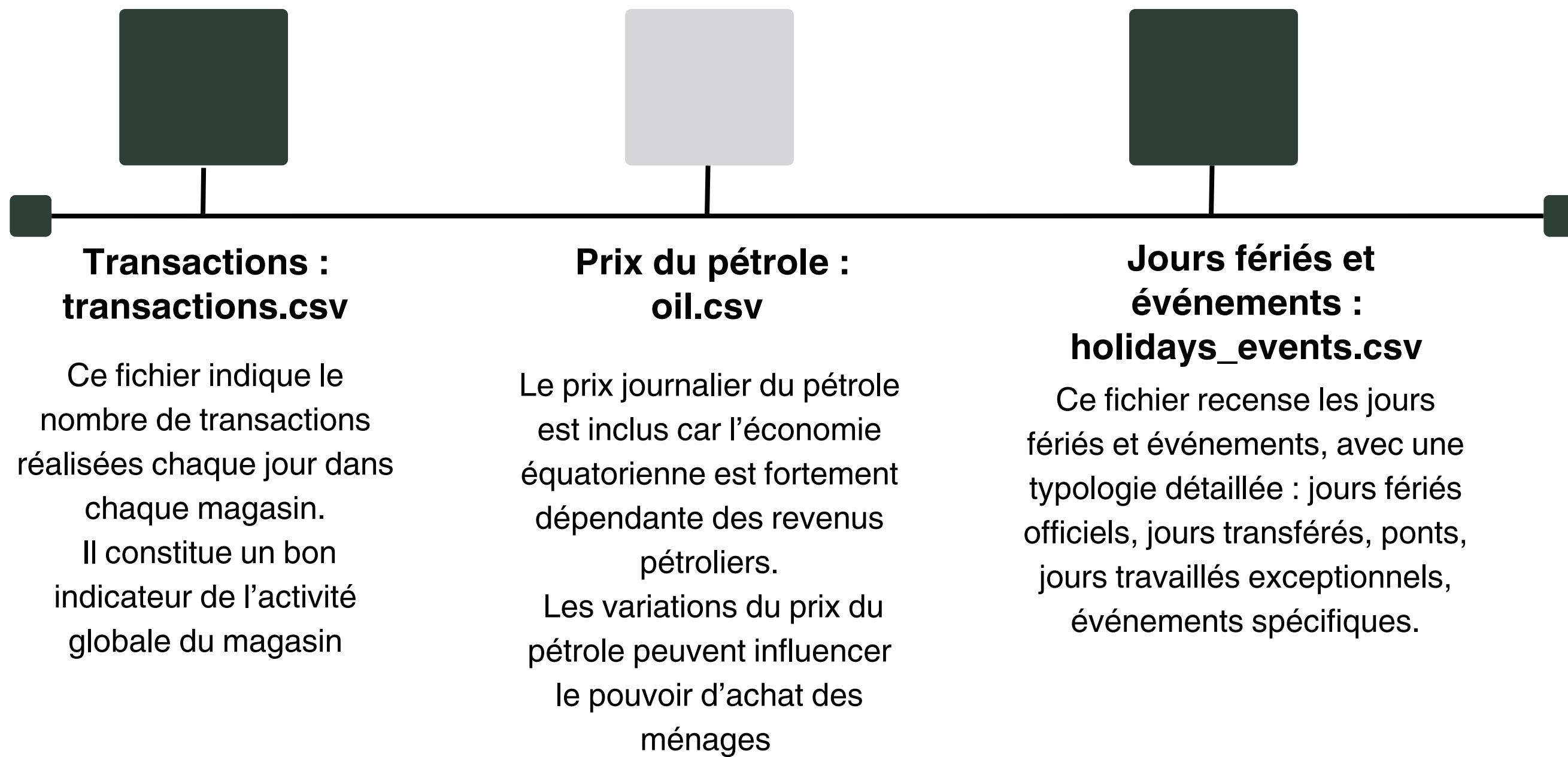


Présentation des données (1/2)





Présentation des données (2/2)



Granularité et variable cible

Les données sont observées à la granularité :
Jour × Magasin × Produite.

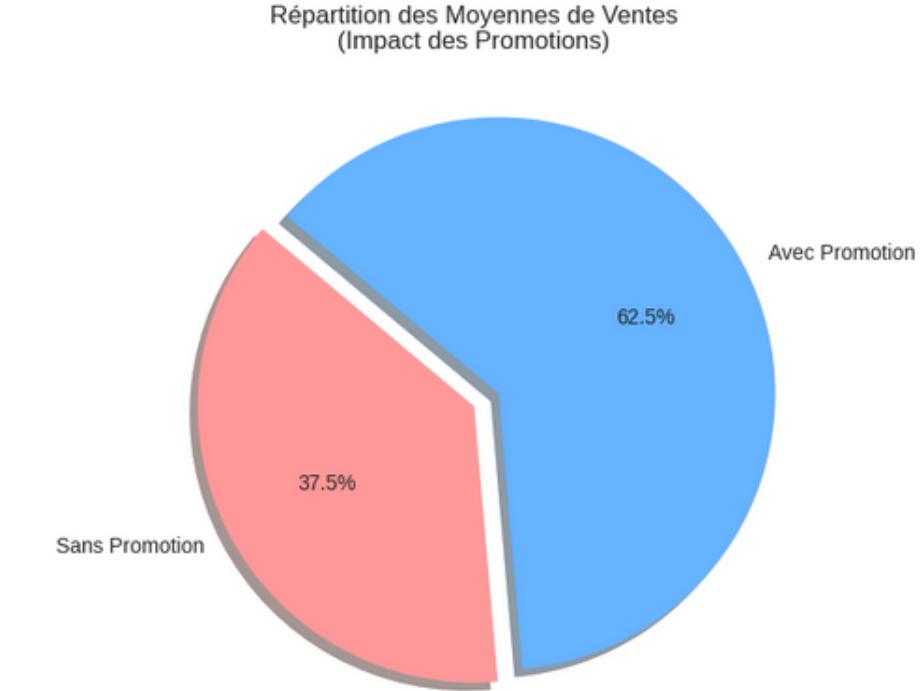
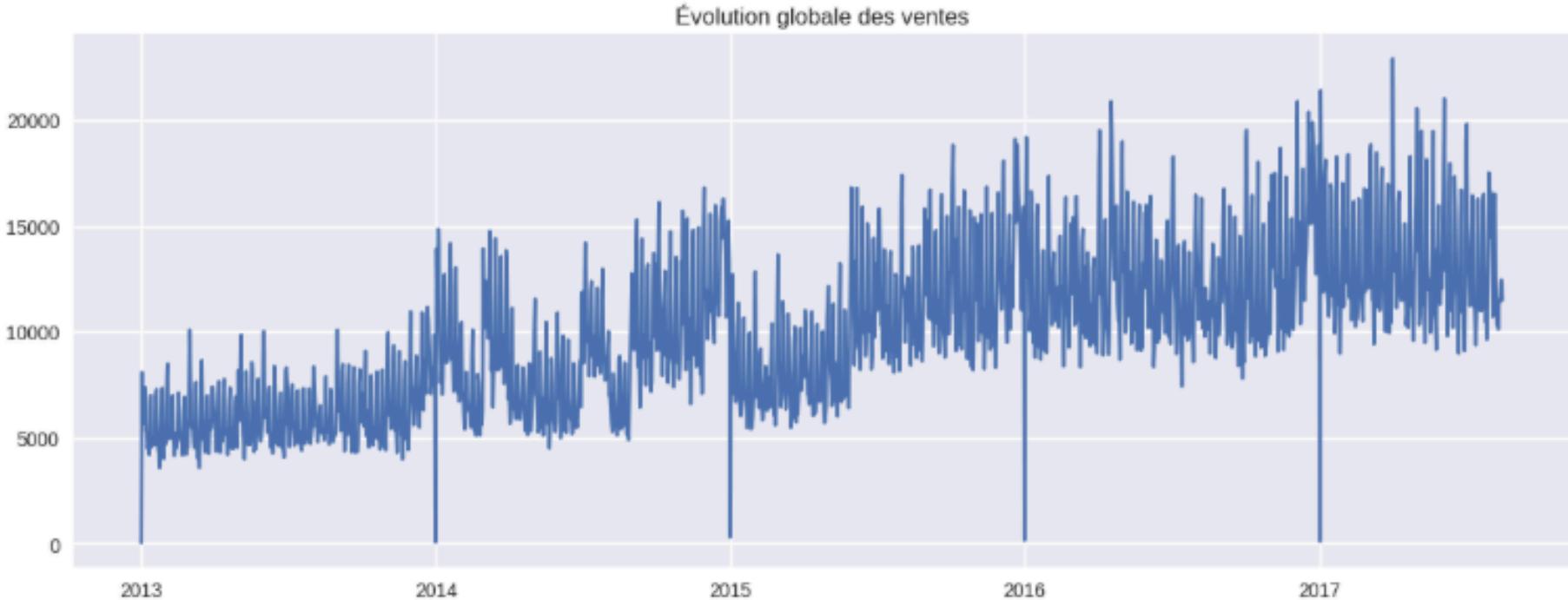
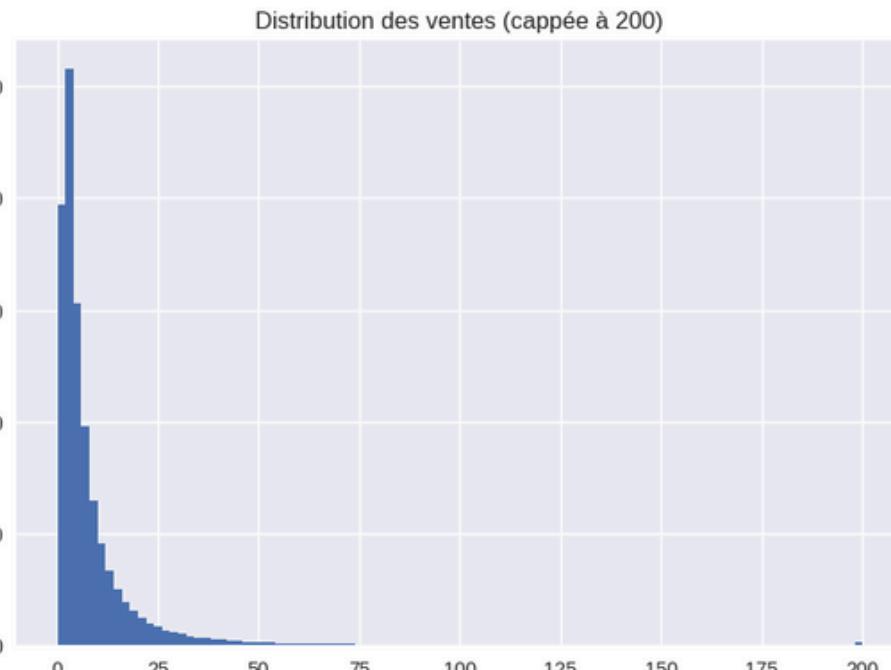
La variable cible est : unit_sales : quantité vendue par jour pour un produit donné dans un magasin donné.

Elle peut être entière ou réelle (produits vendus au poids), négative (retours) et les ventes égales à zéro ne sont pas observées dans les données.



Analyse exploratoire

Analyse exploratoire (1/2)

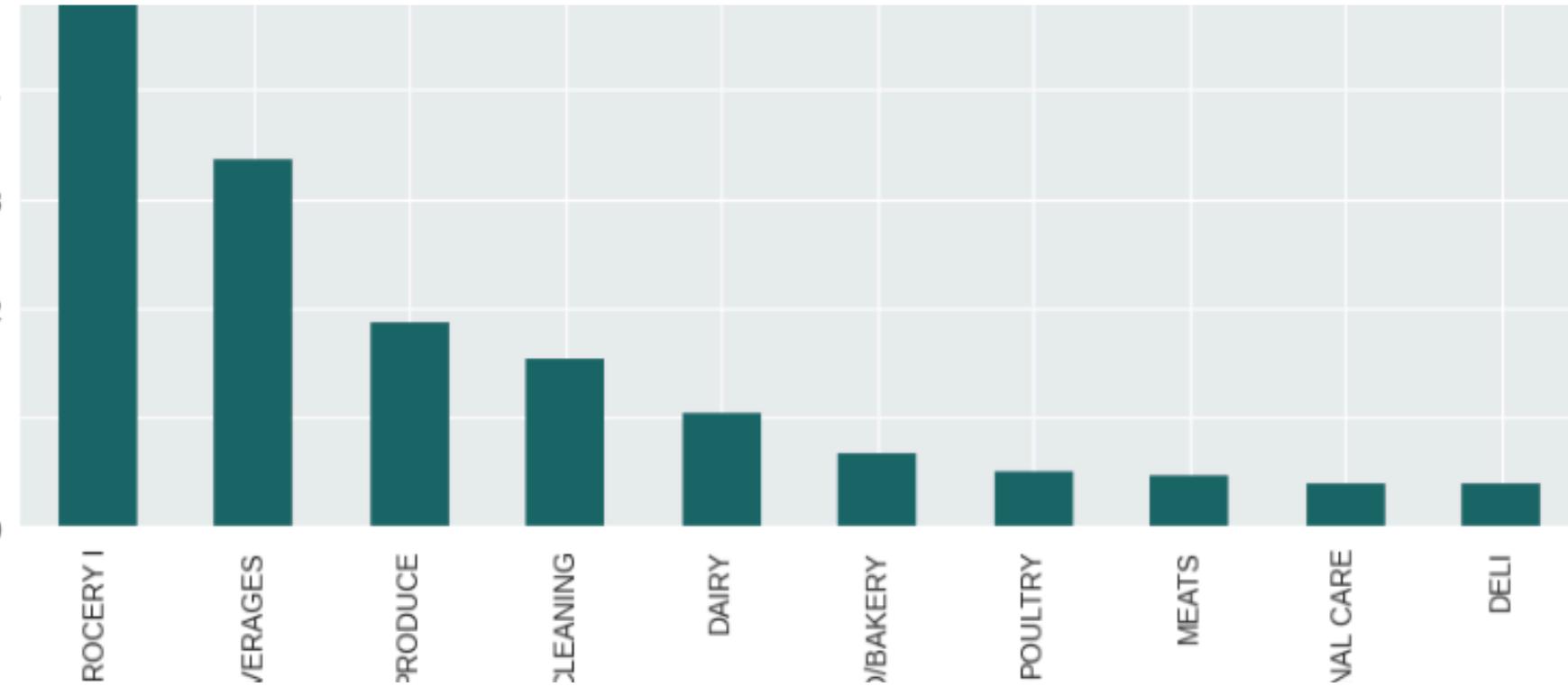


La variable cible est fortement asymétrique à droite, avec une majorité de faibles ventes et quelques valeurs extrêmes rares.

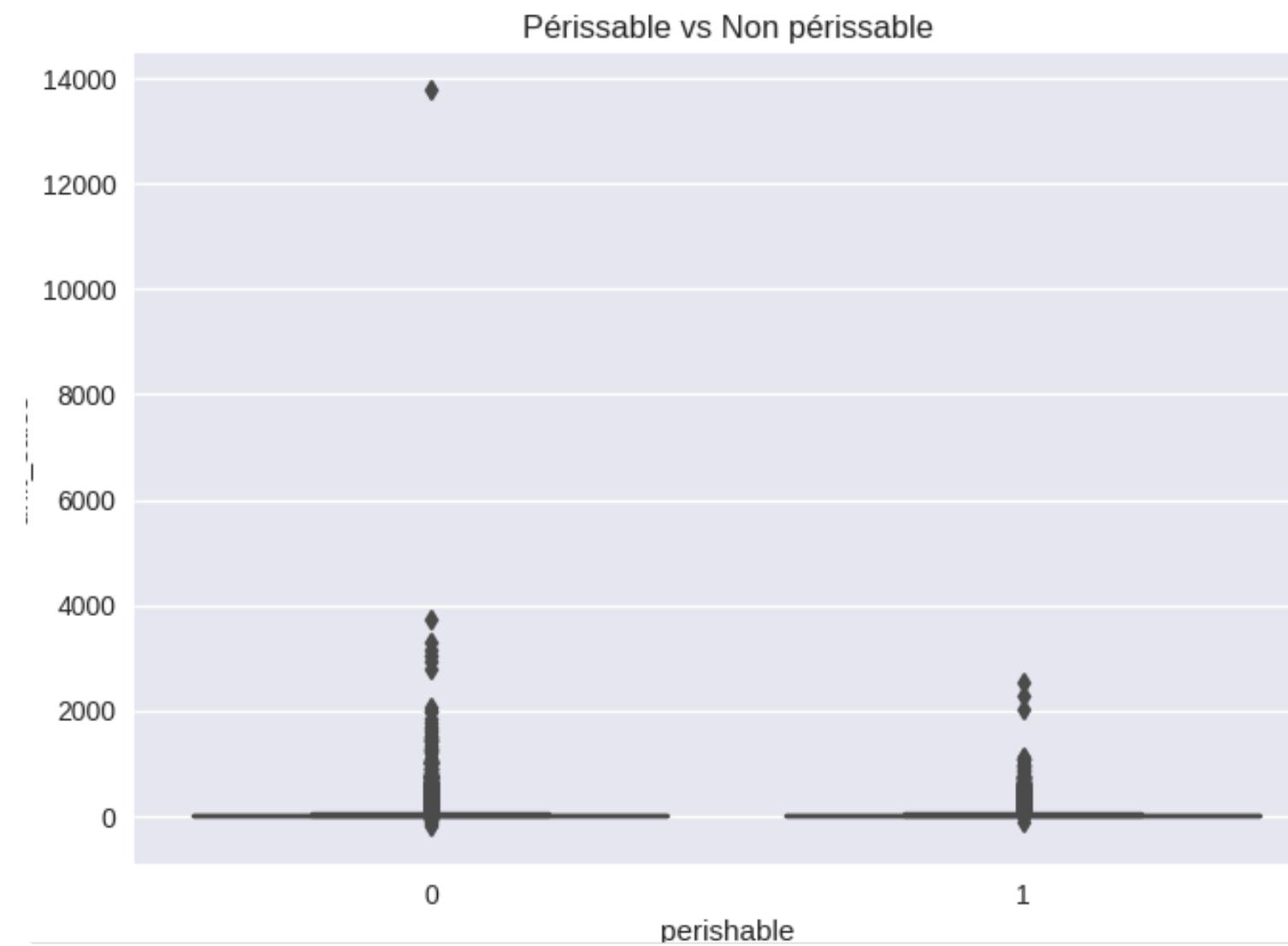
L'évolution des ventes montre une tendance globale à la hausse entre 2013 et 2017, avec une saisonnalité marquée et des baisses récurrentes en début d'année. Cette dynamique indique la nécessité d'un modèle tenant compte des effets temporels.

Les promotions ont un impact majeur sur les ventes, représentant plus de 60 % des volumes et augmentant fortement les quantités vendues par produit.

Analyse exploratoire (2/2)



Les catégories GROCERY I et BEVERAGES dominent largement les volumes de ventes et constituent le principal moteur de l'activité. Les autres familles, bien que significatives, contribuent de manière plus marginale aux transactions totales.



Les produits périssables présentent des niveaux de vente légèrement plus élevés que les produits non périssables, avec une dispersion plus importante et de nombreux outliers dans les deux catégories..

EDA

[Lien vers le Notebook EDA](#)

Prétraitement et feature engineering

Prétraitement et feature engineering (1/2)



PRÉTRAITEMENT DES DONNÉES

Le prétraitement a permis de nettoyer les données brutes issues de plusieurs sources. Les valeurs manquantes ont été traitées selon leur signification métier, les valeurs aberrantes ont été analysées et corrigées lorsque nécessaire, et les formats des variables, notamment temporelles, ont été harmonisés. Cette étape a permis d'obtenir des bases cohérentes et exploitables.



FEATURE ENGINEERING

Le feature engineering a consisté à enrichir les données en créant de nouvelles variables explicatives à partir de l'information existante. Ces variables permettent de mieux capturer la dynamique temporelle des ventes ainsi que les effets des promotions et du contexte.

- variables temporelles (jour, mois, semaine, week-end)
- indicateurs de promotion
- variables retardées (lags)
- moyennes mobiles (rolling statistics)
- encodage des variables catégorielles

Prétraitement et feature engineering (2/2)



FUSION DES SOURCES DE DONNÉES

Les différentes sources de données (ventes, produits, magasins, promotions, transactions, prix du pétrole et jours fériés) ont été fusionnées afin de construire une table finale unique. Cette table regroupe l'ensemble des variables explicatives nécessaires à la modélisation.

- une observation = jour × magasin × produit
- toutes les informations sont centralisées



RÉSULTAT DU PIPELINE DE PRÉPARATION

À l'issue de ce pipeline de prétraitement et de feature engineering, une table finale propre, cohérente et enrichie a été obtenue.

Cette table est directement utilisée pour l'entraînement et l'évaluation des modèles de Machine Learning.

Modélisation

Lien vers le Notebook

Comparaison des modèles

	Model	MSE Train	RMSE Train	MAE Train	MAPE Train (%)	R ² Train	MSE Test	RMSE Test	MAE Test	MAPE Test (%)	R ² Test
0	Ridge (alpha=1.0)	0.33	0.57	0.44	34.19	0.57	0.35	0.59	0.45	34.94	0.52
1	Lasso (alpha=0.01)	0.33	0.57	0.44	34.39	0.57	0.35	0.59	0.45	35.06	0.52
2	ElasticNet (alpha=0.01, l1_ratio=0.5)	0.33	0.57	0.44	34.30	0.57	0.35	0.59	0.45	34.94	0.52

	Model	MSE Train	RMSE Train	MAE Train	MAPE Train (%)	R ² Train	MSE Test	RMSE Test	MAE Test	MAPE Test (%)	R ² Test
0	LightGBM	0.28	0.53	0.41	32.01	0.63	0.31	0.56	0.43	32.61	0.58
1	CatBoost	0.29	0.53	0.41	32.39	0.62	0.31	0.56	0.43	32.93	0.58
2	Naive	NaN	NaN	NaN	NaN	NaN	1.01	1.00	0.86	80.97	-0.37
3	Seasonal Naive	NaN	NaN	NaN	NaN	NaN	4.83	2.20	1.43	113.48	-5.55

Dashboard

Lien vers le Dashboard

Limites et recommandations

Limites du modèle et axes d'amélioration pour une meilleure précision.



1 LIMITES DU PROJET

Le projet présente certaines limites liées à la complexité du problème et aux contraintes techniques. Les données utilisées sont volumineuses, ce qui implique des temps de calcul élevés et limite l'exploration exhaustive de modèles et d'hyperparamètres. De plus, le modèle est entraîné sur un horizon temporel donné et peut être sensible à des changements structurels futurs.

2 RECOMMANDATIONS

Afin de répondre aux contraintes liées au volume des données et à la complexité du modèle, plusieurs pistes d'amélioration peuvent être envisagées. L'optimisation du pipeline de traitement, l'exploration de modèles plus légers et la mise en place de solutions de calcul adaptées permettraient d'améliorer l'efficacité et la scalabilité du système.

**Merci
de votre attention**