Présentateur

Amath SALL

Elève ingénieur des travaux statistiques

El Moustapha DIOP

Elève ingénieur des travaux statistiques

Diakhou NDAW

Elève ingénieur des travaux statistiques

Predict future sales

February 2th

PLAN

1. INTRODUCTION
2. Exploration des données
3. Traitement des données manquantes et extrémes
4. Future engineering
5. Evaluation
6. Evaluation

Introduction

Rappel de la problématique

Démarche

Problématique

On est des data-scientists chez Oshop ,
 un réseau de magasins de vente d’articles.

Dans le réseau on y vend, quotidiennement dans chaque magasin des boutiques des articles.

Il est demandé aux data-scientists du groupe de mettre en place un **modèle de prédiction des ventes totales** pour chaque produit et chaque magasin au cours du mois prochain.

01

Analyse exploratoire de données

Informations générales

Bases de données

sales_train.csv

	date	date_block_num	shop_id	item_id	item_price	item_cnt_day
0	02.01.2013	0	59	22154	999.00	1.0
1	03.01.2013	0	25	2552	899.00	1.0
2	05.01.2013	0	25	2552	899.00	-1.0

items.csv

	item_name	item_id	item_category_id
0	! ВО ВЛАСТИ НАВАЖДЕНИЯ (ПЛАСТ.) D	0	40
1	!ABBYY FineReader 12 Professional Edition Full...	1	76
2	***В ЛУЧАХ СЛАВЫ (UNV) D	2	40

item_categories.csv

	item_category_name	item_category_id
0	PC - Гарнитуры/Наушники	0
1	Аксессуары - PS2	1
2	Аксессуары - PS3	2

shops.csv

	shop_name	shop_id
0	!Якутск Орджоникидзе, 56 фран	0
1	!Якутск ТЦ "Центральный" фран	1
2	Адыгея ТЦ "Мега"	2

Base finale

	date_block_num	shop_id	item_id	item_cnt_month	item_price	item_category_id
0	0	0	32	6.0	221.0	40
1	0	0	33	3.0	347.0	37
2	0	0	35	1.0	247.0	40

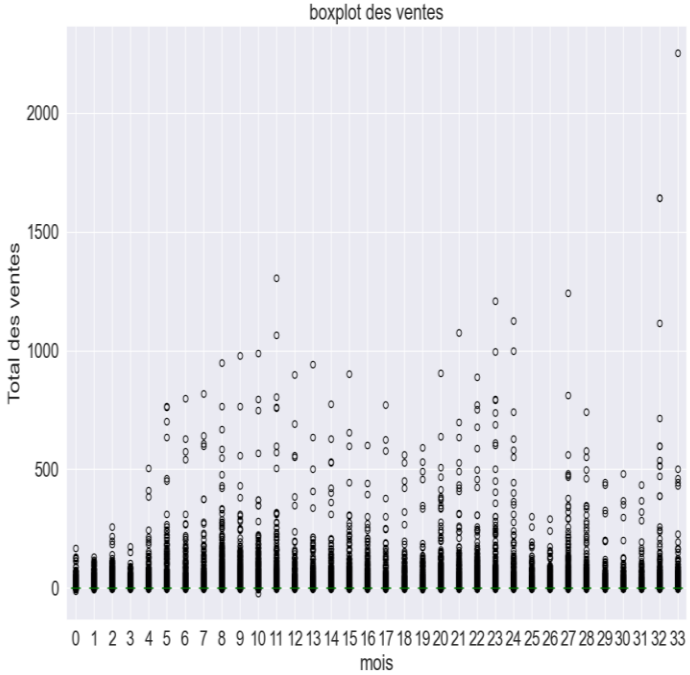
1 609 124 individus

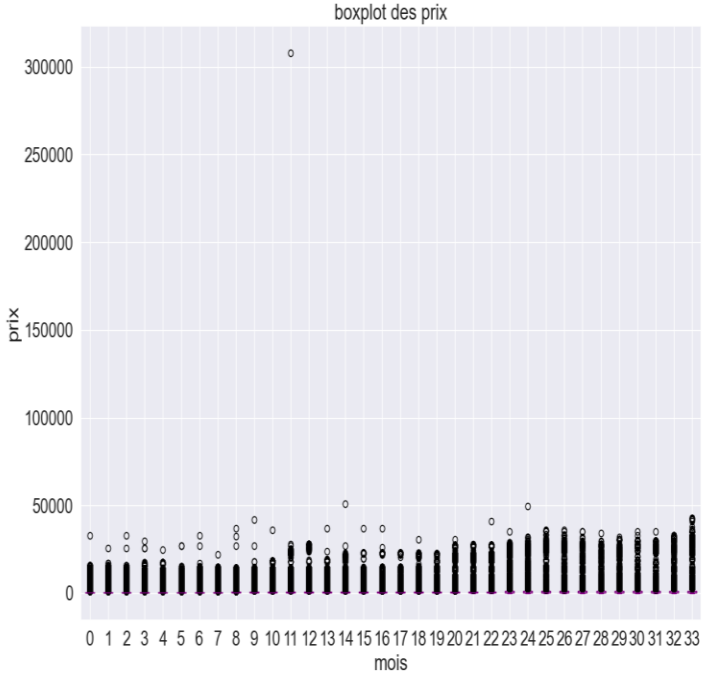
22 170 Articles

84 Catégories d'articles

60 Boutiques

Distribution des ventes et des prix par mois





Ventes

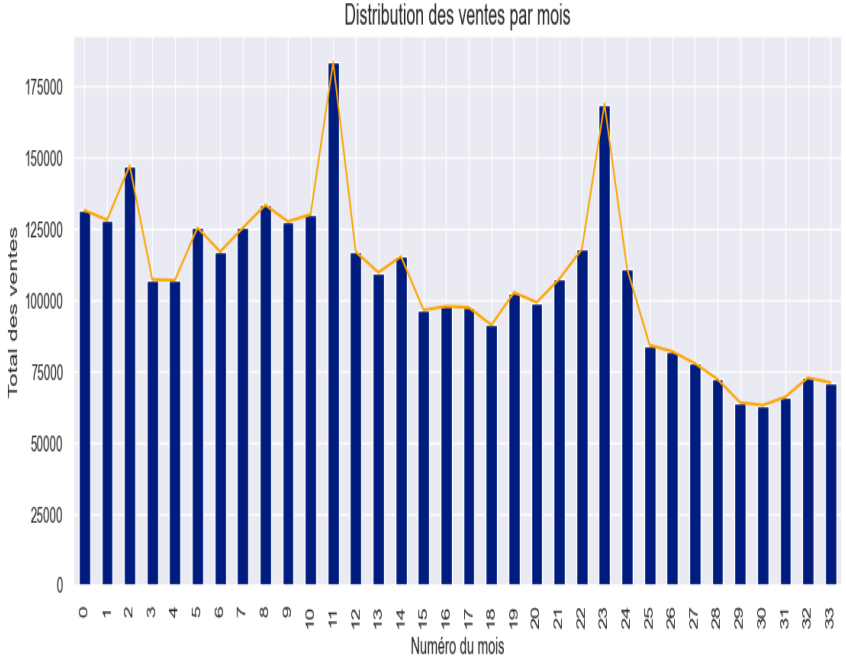
Moyenne	2
Maximum	2253

Prix

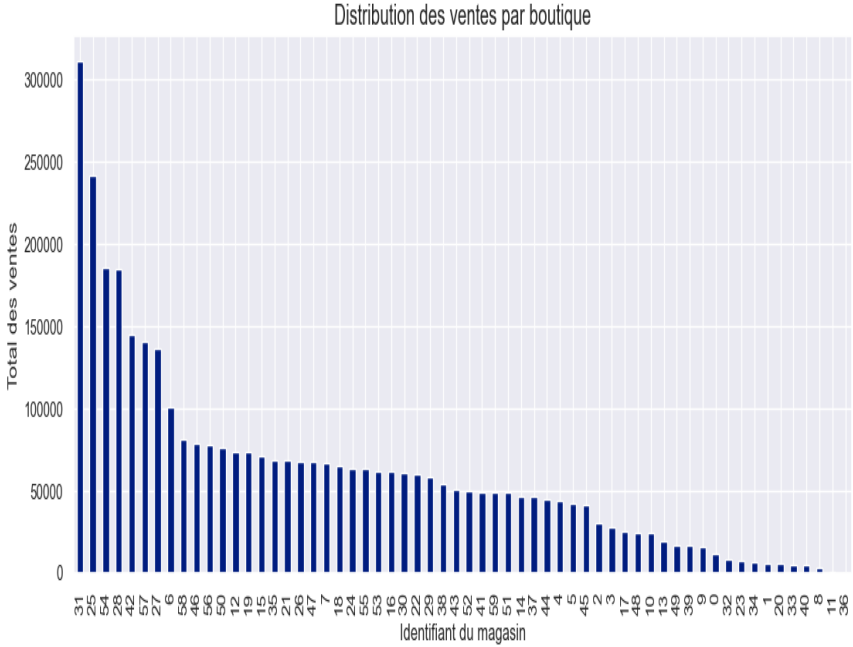
Moyenne	790
Maximum	307980

Répartition des ventes

a Par mois

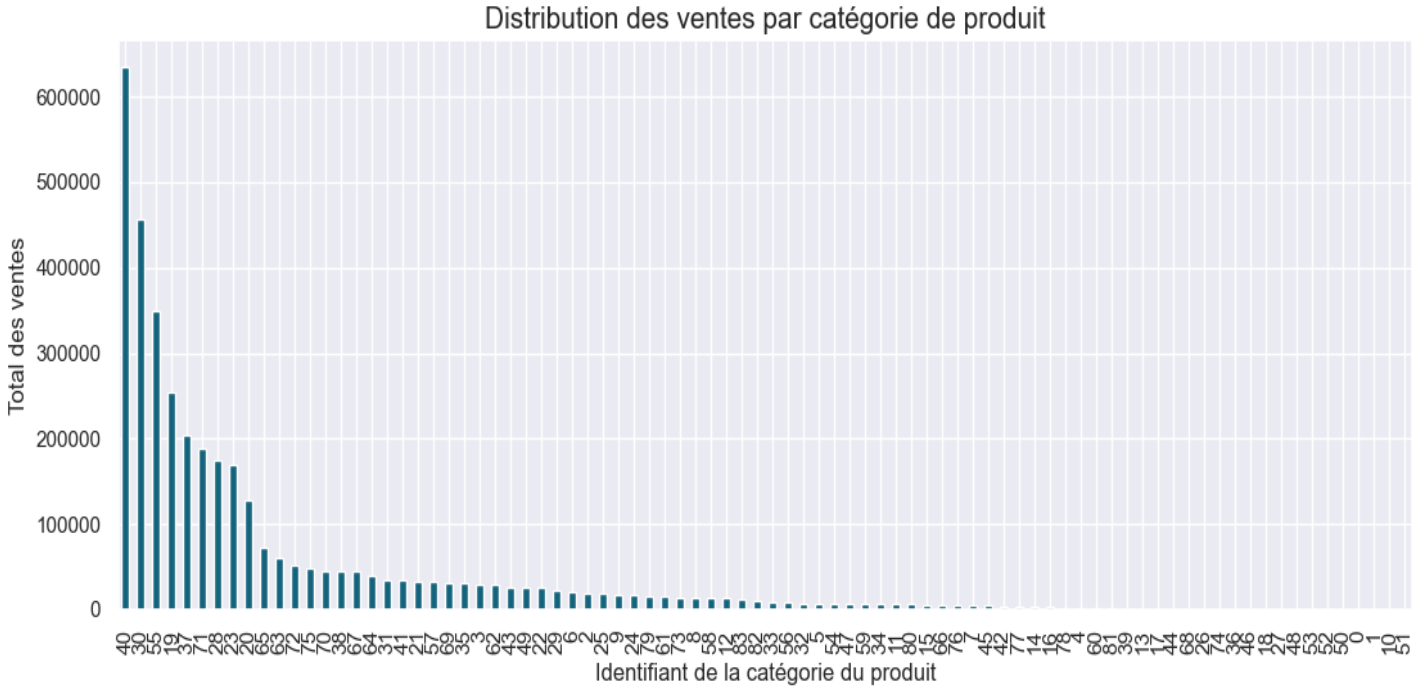


b Par boutique



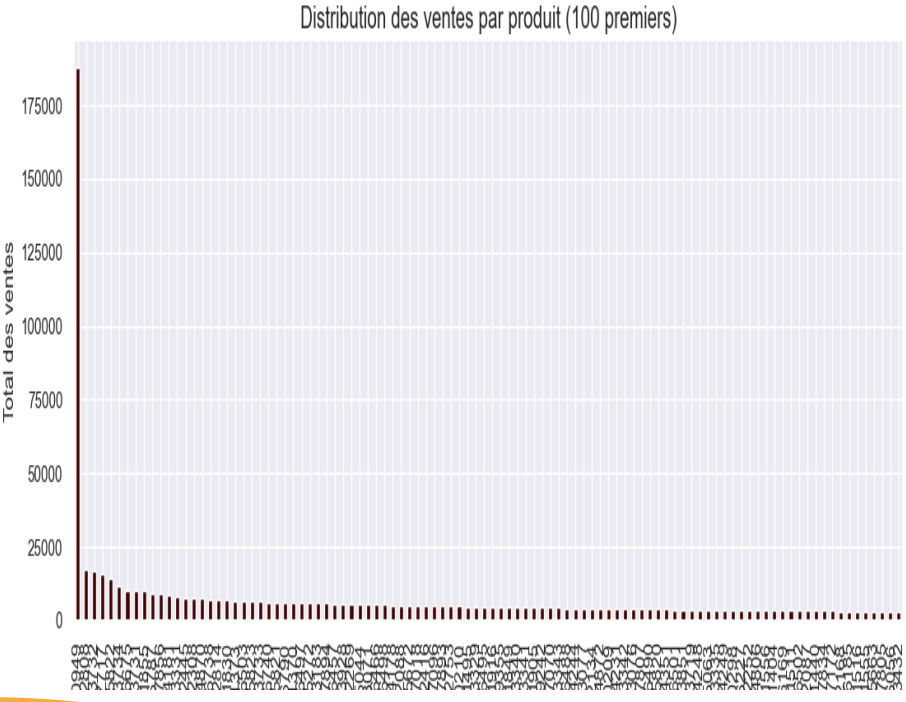
Répartition des ventes

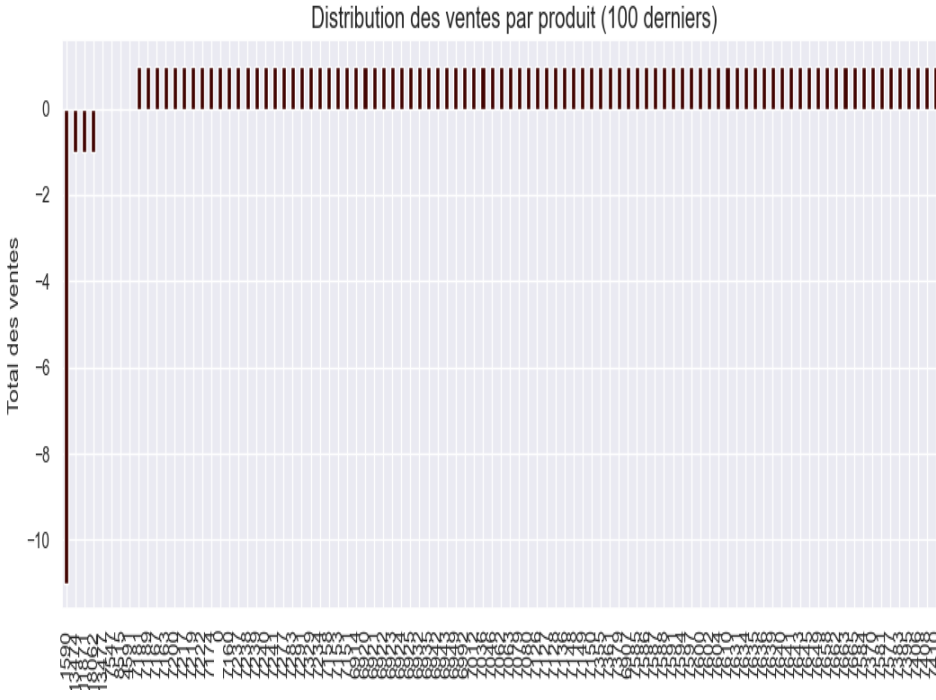
c Par catégorie



Répartition des ventes

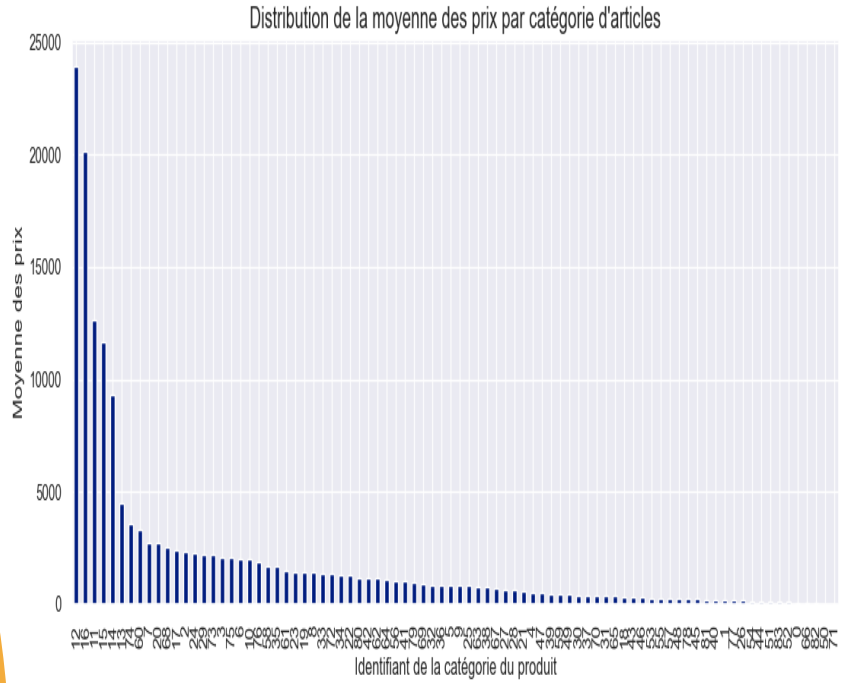
Par produit



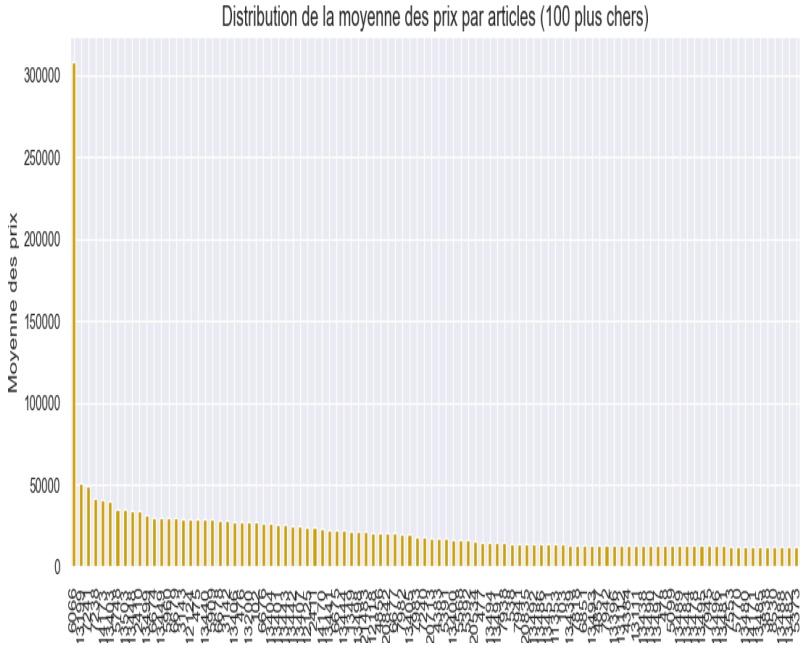


Répartition des prix

a Par catégorie



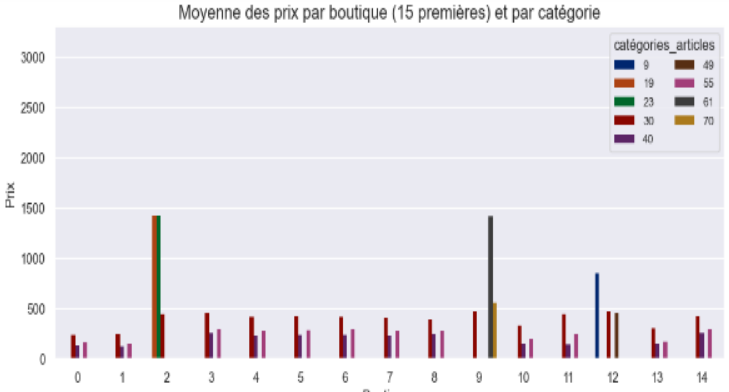
b Par produit

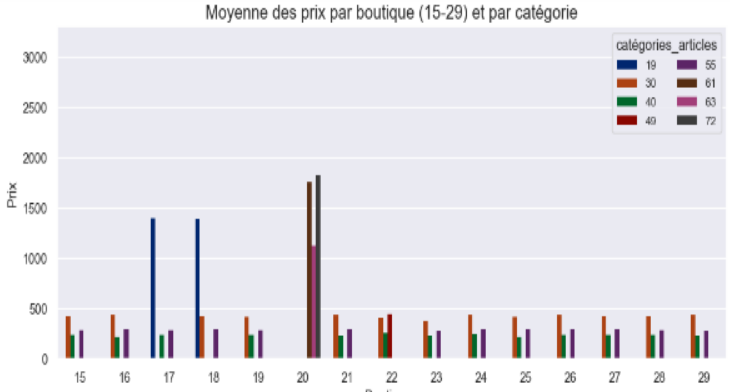


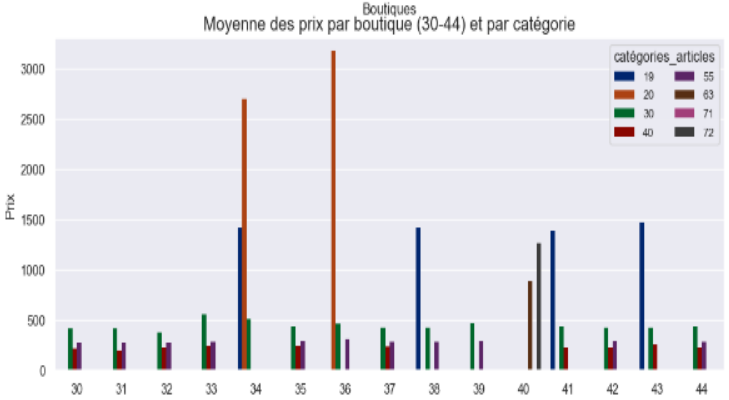
Répartition des prix

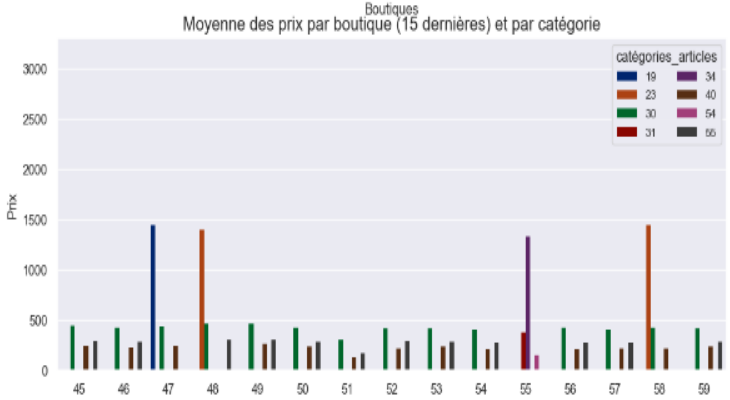
C

Par boutique et par catégorie (les 3 premières les plus vendus)









03

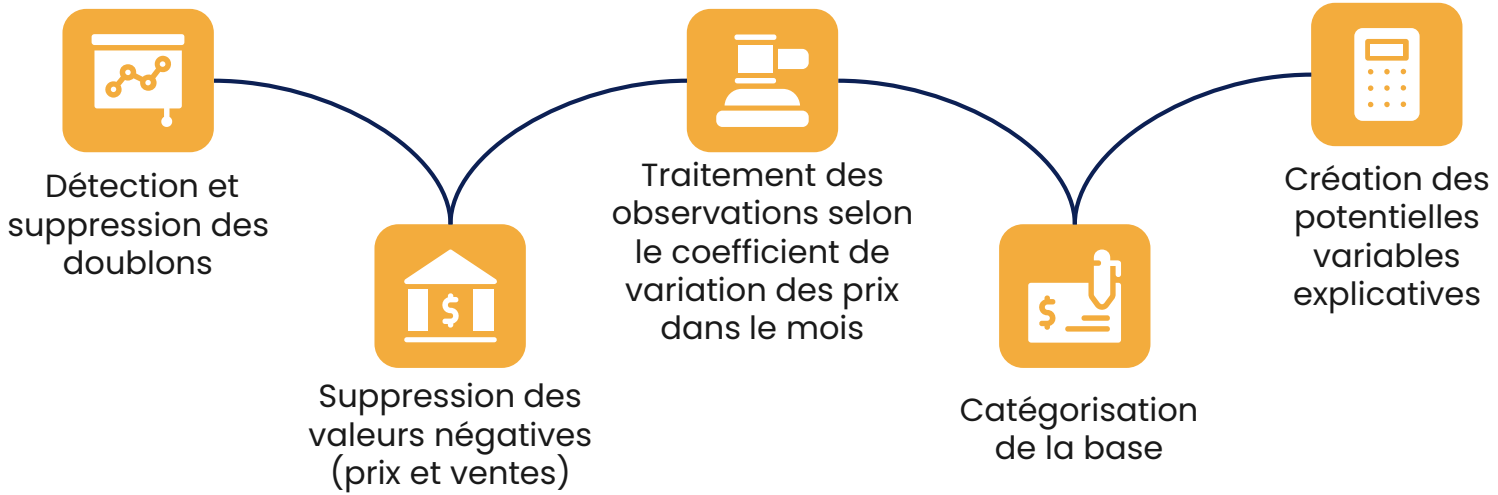
Data cleaning

Data cleaning et Feature Engeneering

L'analyse descriptive a montré que la base contenait beaucoup d'incohérences, et surtout qu'il n' y avait définitivement pas assez de variables pour faire tourner un modèle

Objectifs

Rendre la base plus digeste pour un modèle ML



Imputation des données manquantes

Pour les ventes retardées , les données manquantes sont remplacées par 0

L'imputation des prix est faite en 4 étapes succesives

1

Regroupement selon le mois, l'article

2

Regroupement selon le mois, la catégorie de l'article

3

Regroupement selon le mois, la boutique

4

Regroupement selon le mois

Modélisation

Comment choisir le modèle le plus adapté ?

RÉGRESSION

BAGGING

Random Forest

BOOSTING

- ✓ XGBOOST
- ✓ LIGHTGBM

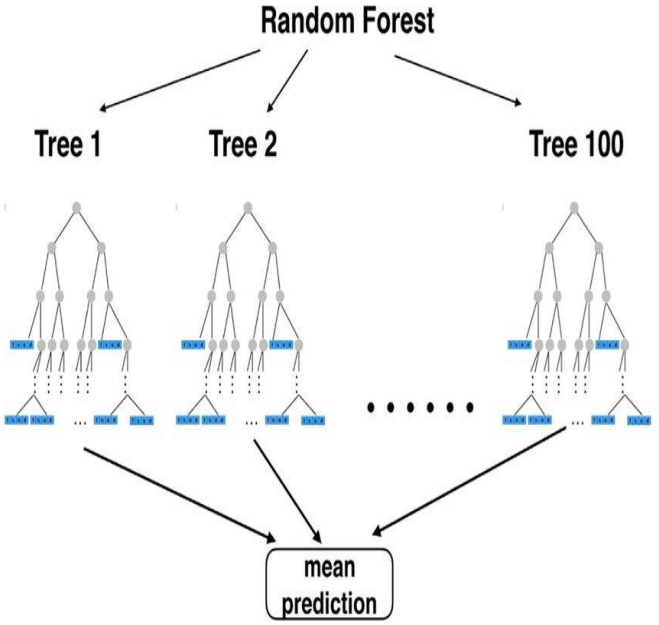
01

Random Forest

Random Forest

- Proposé par Leo Breiman en 2001

- algorithme qui **se base sur l'assemblage d' un ensemble d'arbres de décision** utilisés pour prédire une quantité ou une probabilité



02

XGBOOST

03

LightGBM

Un algorithme de boosting

LightGBM

LightGBM (*Light Gradient Boosting Machine*), et publié en 2016.

Basé sur des algorithmes d’arbre de décision, il est utilisé pour le classement, la classification et la régression

les arbres de décision sont construits en **fractionnant les observations** en fonction des valeurs des futures. L'algorithme CART recherche la meilleure répartition.

Comment choisir la répartition optimale ?

Cette méthode est lente lorsque la taille est grande

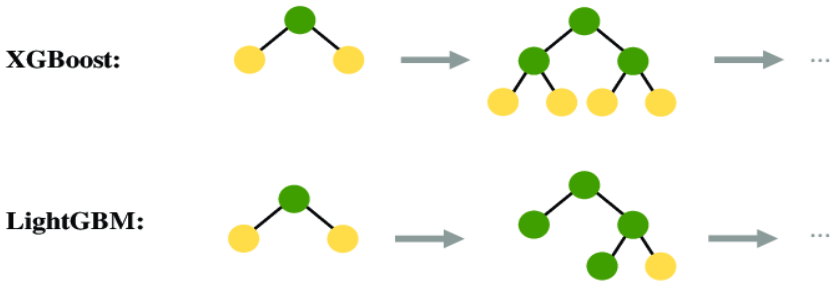
Avec LightGBM, on vise à résoudre ce problème de temps et de puissance de calcul

Donc on introduit une nouvelle méthode de recherche de la bonne structure d'arbre pour chaque apprenant ajouté.

LightGBM

LIGHTGBM VS XGBOOST

LightGBM fait pousser l'arbre par **feuille** tandis que l'autre algorithme se développe par niveau.

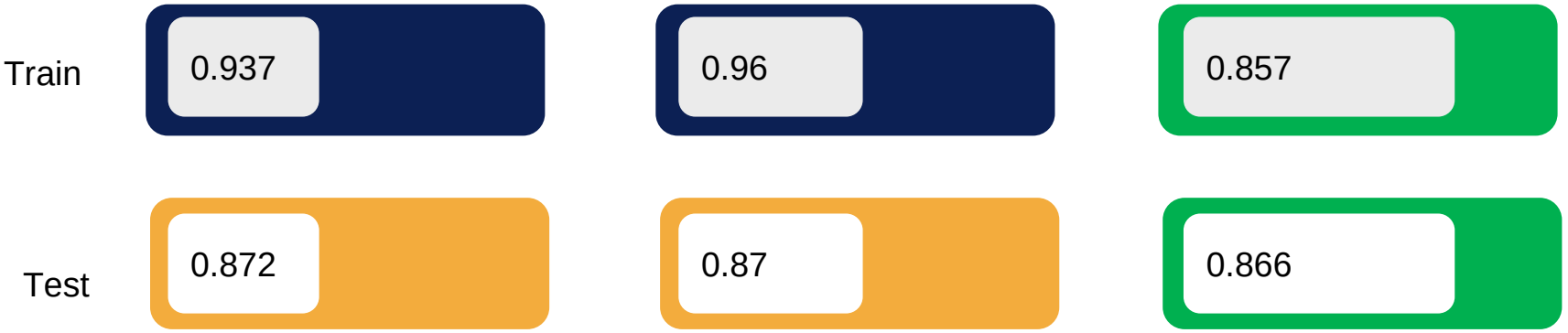


Il n y a pas de vainqueur au niveau de la performance les deux modèles sont très performants, par contre XGBoost est mieux adapté au jeux de données de petite taille, on risque **d'avoir un sur-apprentissage** avec de grands volumes de données tandis que LightGBM est son opposé.

LightGBM est **beaucoup plus rapide** que XGBoost. Grâce à sa méthode de réduction de dimension (EFB), il est capable de gagner en matière de puissance de calcule tout en conservant la même précision de XGBoost.

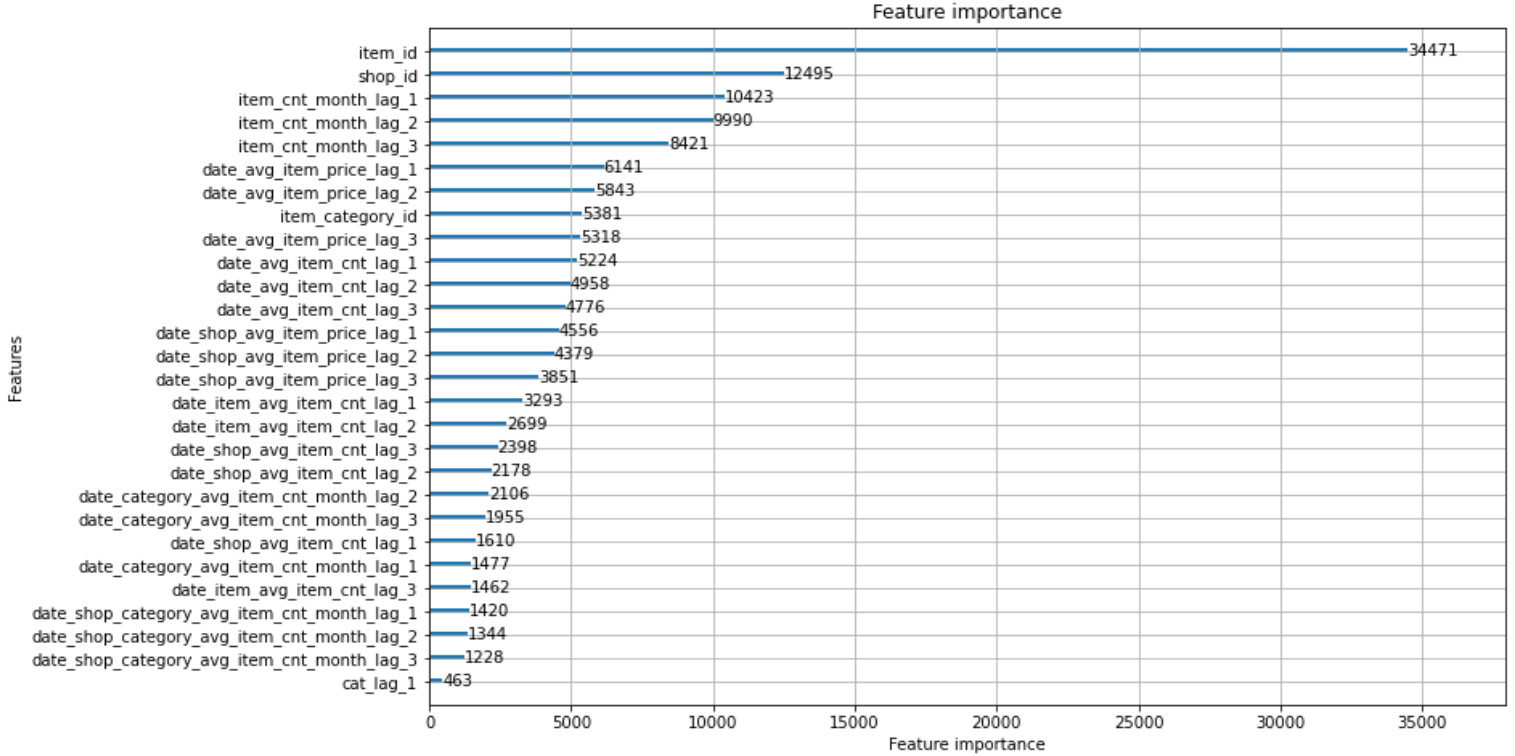
Evaluation

Score(r2) des modèles



On choisit le modèle LIGHTGBM, en effet ,les autres font du sur-apprentissage

Future importance



LIGHTGBM

CONCLUSION

MERCI POUR VOTRE ATTENTION !