المدرسة الوطنية للعلوم التطبيقية
ⵜⵉⵎⵙⵙⵓⵜ ⵜⴰⵏⴰⵎⵓⵔⵜ ⵜⵉⵎⴰⵡⴰⵙⵉⵏ ⵜⵉⵏⵉⵙⵉⵏⵉ
Ecole Nationale des Sciences Appliquées de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⴰⵎⴷ ⴱⵏ ⵄⴱⴷⵍⵍⴰⵀ
Université Sidi Mohamed Ben Abdellah

École Nationale des Sciences Appliquées de Fès

Filière
Ingénierie logicielle et intelligence artificielle (ILIA)

Rapport de Projet

Prédiction des Ventes et les tendances d'un store

Réalisé par :

ZOUAK Douae
EL HAOUARI Chahd

Encadrant académique :

Pr. IDRISSI KHAMLICHI Youness

Année Universitaire 2024-2025

Table des matières

Résumé du projet

Dans un contexte où la compétitivité des entreprises repose de plus en plus sur la capacité à anticiper les fluctuations du marché, ce projet s'inscrit comme une réponse innovante à un défi analytique majeur : prédire avec précision l'évolution des ventes en tenant compte de comportements d'achat complexes, de tendances variables et d'effets externes souvent difficiles à modéliser. L'objectif était clair : transformer des données historiques brutes en une source de valeur stratégique, capable de guider la prise de décision et d'optimiser les actions commerciales.

Pour relever ce défi, nous avons développé un système intelligent de prévision des ventes capable de capturer la richesse et la diversité des dynamiques temporelles associées aux séries de données réelles. Là où les méthodes traditionnelles, telles que ARIMA ou SARIMA, se heurtent à leurs limites face à la présence de multiples saisons, de tendances irrégulières ou d'effets événementiels soudains, notre solution se distingue par sa flexibilité, sa robustesse et son adaptabilité.

Au cœur de notre démarche repose une approche hybride, combinant des modèles avancés tels que Prophet et TBATS avec l'intégration habile de variables exogènes décrivant le contexte d'achat : jours spéciaux, promotions, variations saisonnières, ou encore changements de comportement des consommateurs. Cette personnalisation du modèle permet d'aller bien au-delà d'une simple extrapolation : elle offre une compréhension profonde des mécanismes sous-jacents aux ventes.

L'innovation majeure du projet réside ainsi dans la capacité de notre système à détecter et interpréter les rythmes cachés, à s'adapter aux irrégularités de la réalité, et à fournir des prédictions fiables même dans des environnements instables. Grâce à cette architecture intelligente, nous obtenons non seulement une amélioration significative de la précision, mais également une vision analytique plus riche et plus exploitable.

Ce travail démontre qu'une combinaison judicieuse entre modèles statistiques modernes, ingénierie des variables et analyse contextuelle constitue un levier puissant pour améliorer la planification commerciale. Le système développé s'impose ainsi comme un outil stratégique, transformant de simples données en insights concrets et actionnables, au service de la performance et de la prise de décision éclairée.

1 Contexte du Projet

1.1 Introduction

Dans le secteur hautement concurrentiel du retail, la capacité à anticiper les fluctuations de la demande constitue un avantage stratégique déterminant. Les entreprises doivent gérer un équilibre délicat entre disponibilité des produits, limitation des coûts de stockage et satisfaction client. Une mauvaise estimation des ventes peut engendrer des ruptures de stock, des surstocks coûteux ou une allocation inefficace des ressources humaines et matérielles. Dans ce contexte, la prévision des ventes se positionne comme un pilier fondamental de la performance opérationnelle et financière.

Face à la complexité croissante des comportements d'achat — influencés par la saisonnalité, les promotions, les tendances du marché ou encore les événements socio-économiques — les méthodes traditionnelles ne suffisent plus. Les modèles simples, souvent statiques, échouent à capturer la diversité et la dynamique des variations de la demande. Le besoin d'outils plus flexibles, plus intelligents et capables d'intégrer une multitude de facteurs devient alors essentiel.

C'est dans ce cadre que s'inscrit notre projet : concevoir un système avancé de prédiction des ventes capable de fournir une estimation fiable et détaillée des ventes futures d'un magasin spécifique sur un horizon de trois mois. Cette solution vise à offrir aux gestionnaires une vision claire de l'évolution probable de la demande, afin d'améliorer la planification des stocks, d'optimiser les achats, de dimensionner les équipes et de soutenir les décisions stratégiques.

Au-delà de la simple projection statistique, notre système ambitionne de capturer les mécanismes profonds et les relations cachées qui structurent les données historiques. Il s'agit non seulement de prédire, mais aussi de comprendre. En intégrant des modèles modernes de séries temporelles ainsi que des variables contextuelles pertinentes, ce projet répond à un besoin réel du secteur retail : disposer d'outils de prévision robustes, adaptatifs et capables de suivre le rythme d'un environnement en constante évolution.

Ce contexte met ainsi en lumière l'importance cruciale de solutions prédictives avancées pour permettre au magasin étudié — et, par extension, à tout acteur du retail — de prendre des décisions éclairées, d'améliorer son efficacité opérationnelle et de renforcer sa compétitivité à long terme.

Ce rapport est structuré autour de trois axes complémentaires permettant de présenter de manière cohérente l'ensemble du travail réalisé. Le premier axe est consacré aux fondements théoriques, où sont exposés les concepts clés, les principes méthodologiques et les modèles de prévision qui constituent la base scientifique du projet. Le deuxième axe détaille la méthodologie adoptée ainsi que les modèles retenus, en décrivant les étapes de préparation des données, les choix techniques effectués et les justifications qui ont guidé la construction du système prédictif. Enfin, le troisième axe porte sur la réalisation pratique et les résultats obtenus, illustrant la mise en œuvre concrète de la solution, les performances mesurées et les insights dégagés. Cette organisation vise à offrir une lecture claire, progressive et rigoureuse du développement du projet.

2 Les Séries Temporelles

Les séries temporelles représentent un type particulier de données où les observations sont enregistrées de manière ordonnée dans le temps, généralement à intervalles réguliers. Contrairement aux données classiques, elles intègrent une dimension temporelle qui leur confère des caractéristiques propres, telles que la tendance (évolution globale à long terme), la saisonnalité (variations cycliques régulières), les cycles économiques plus irréguliers, ainsi que la bruit ou variabilité aléatoire. L'analyse des séries temporelles vise à comprendre ces composantes afin de modéliser les mécanismes qui gouvernent l'évolution d'un phénomène au fil du temps et de pouvoir en prédire les valeurs futures. Cette discipline s'appuie sur des outils statistiques et mathématiques sophistiqués permettant de détecter des patterns, d'isoler les effets significatifs et de gérer la dépendance entre les observations successives, ce qui distingue fondamentalement les séries temporelles des données indépendantes et identiquement distribuées. Grâce à cette approche, il devient possible de produire des prévisions fiables et exploitables dans de nombreux domaines tels que la finance, le retail, la météorologie ou encore l'ingénierie.

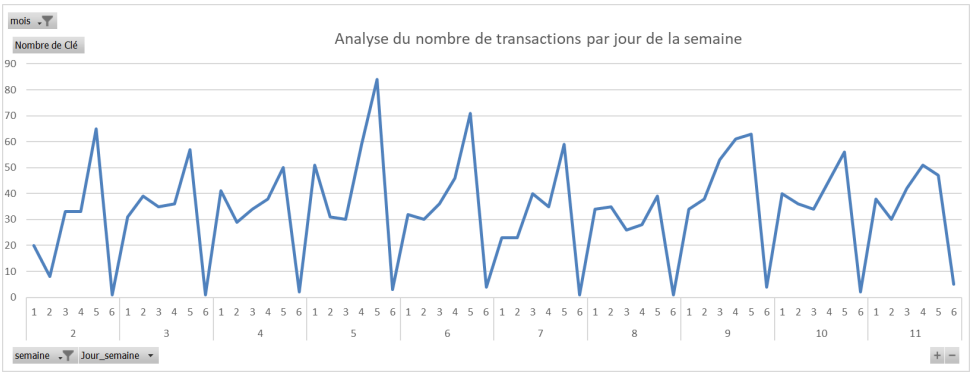


FIGURE 1 – Exemple de série temporelle

2.1 Décomposition des Séries Temporelles

Une série temporelle peut être décomposée en trois composantes principales :

2.1.1 Tendance (Trend)

La tendance d'une série temporelle correspond au mouvement général ou à l'évolution à long terme du phénomène observé, indépendamment des fluctuations saisonnières ou du bruit aléatoire. Elle permet de saisir la direction globale des observations sur une période étendue et constitue un indicateur fondamental pour la planification et la prise de décision stratégique. La tendance peut se manifester de différentes manières :

- **Croissante** : les valeurs de la série augmentent progressivement sur le long terme, ce qui peut refléter par exemple une augmentation continue des ventes due à la croissance du marché ou à la fidélisation de la clientèle.
- **Décroissante** : les valeurs diminuent sur la période étudiée, ce qui peut signaler une perte de clientèle ou une diminution de la demande pour certains produits.
- **Stable** : la série ne présente pas de variation significative à long terme, indiquant une certaine régularité ou maturité du marché.

D'un point de vue mathématique, la tendance peut être modélisée par :

$$T_t = f(t) + \epsilon_t$$

où T_t représente la valeur observée à l'instant t, $f(t)$ est une fonction du temps capturant le mouvement principal de la série, et ϵ_t représente les fluctuations aléatoires ou résiduelles.

La modélisation rigoureuse de la tendance est essentielle pour isoler les variations saisonnières et le bruit, et ainsi produire des prévisions fiables et interprétables.

2.1.2 Saisonnalité (Seasonality)

La saisonnalité d'une série temporelle correspond aux variations périodiques qui se répètent à intervalles réguliers au sein des données. Ces fluctuations reflètent des comportements récurrents liés à des facteurs temporels prévisibles, tels que les saisons de l'année, les mois, les semaines ou même des cycles journaliers. Par exemple, dans le commerce de détail, les ventes de certains produits peuvent augmenter systématiquement pendant les fêtes de fin d'année ou les soldes, tandis que d'autres connaissent des pics hebdomadaires selon les habitudes des consommateurs.

Mathématiquement, la saisonnalité peut être représentée par :

$$S_t = S_{t+kP}$$

où S_t est la composante saisonnière à l'instant t, **P** la période de répétition (par exemple 12 pour des données mensuelles présentant un cycle annuel), et k un entier qui représente le nombre de cycles écoulés. Cette formulation exprime que la variation saisonnière observée à un moment donné se reproduira de manière similaire après chaque période **P**.

La modélisation de la saisonnalité permet de distinguer les fluctuations prévisibles de la tendance globale ou du bruit aléatoire, offrant ainsi une meilleure compréhension de la dynamique de la série. Elle est essentielle pour produire des prévisions précises, car elle capture les motifs répétitifs inhérents aux données et permet d'anticiper les pics et creux attendus dans le futur.

2.1.3 Bruit (Noise)

Le bruit d'une série temporelle correspond aux variations aléatoires et imprévisibles qui ne suivent aucun schéma systématique. Contrairement à la tendance ou à la saisonnalité, le bruit ne présente pas de régularité et reflète l'influence de facteurs exceptionnels, ponctuels ou non observés qui affectent temporairement le phénomène étudié. Dans le contexte des ventes, cela peut inclure des événements inattendus comme des conditions météorologiques inhabituelles, des incidents logistiques, ou des promotions non planifiées.

Mathématiquement, le bruit est souvent modélisé comme une variable aléatoire ϵ_t suivant une distribution normale centrée réduite :

$$\epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

où 0 représente l'espérance nulle et σ^2 la variance qui quantifie l'intensité des fluctuations aléatoires. Cette modélisation permet de considérer le bruit comme une composante

résiduelle, indépendante de la tendance et de la saisonnalité, qui encapsule l'incertitude inhérente à la série.

La prise en compte du bruit est cruciale pour l'évaluation de la fiabilité des prévisions. En isolant les fluctuations aléatoires, il devient possible de se concentrer sur les motifs structurels significatifs et d'améliorer la précision des modèles prédictifs, tout en fournissant des intervalles de confiance pour les prédictions futures.

2.2 Types des séries temporelles

Les séries temporelles peuvent être classées en deux grandes catégories selon leur comportement statistique dans le temps : les séries stables (stationnaires) et les séries non stables (non stationnaires). Cette distinction est fondamentale car elle conditionne le choix des modèles de prévision et les transformations nécessaires avant toute modélisation.

2.2.1 Séries temporelles stationnaires (stables)

Une série est dite stationnaire lorsque ses propriétés statistiques — moyenne, variance, autocorrélation — restent constantes dans le temps. Autrement dit, les fluctuations observées sont centrées autour d'un niveau stable et ne présentent ni tendance marquée ni saisonnalité structurelle.

Caractéristiques principales :
— Moyenne constante au fil du temps.
— Variance stable (pas d'augmentation ou de diminution progressive de la dispersion).
— Autocorrélations décroissant rapidement vers zéro.
— Absence de tendance ou de saisonnalité persistante.

Ces séries sont généralement plus faciles à modéliser et adaptées aux modèles linéaires comme AR, MA ou ARMA. Elles représentent les dynamiques où les processus sous-jacents sont relativement réguliers et prévisibles.

2.2.2 Séries temporelles non stationnaires (non stables)

Une série est non stationnaire lorsque ses propriétés statistiques varient dans le temps. C'est le cas le plus fréquent dans les données réelles, notamment en économie, finance ou retail.

Caractéristiques principales :
— Présence d'une tendance (hausse ou baisse prolongée).
— Variances changeantes, souvent croissantes dans le temps.
— Cycles saisonniers persistants.
— Autocorrélations élevées sur de nombreux décalages (lags).

Types courants de non-stationnarité :
— Tendance linéaire ou non linéaire.
— Saisonnalité régulière (hebdomadaire, mensuelle, annuelle...).
— Changement structurel (rupture, changement de comportement).
— Variance non constante (hétéroscédasticité).

Ces séries nécessitent souvent des transformations pour devenir stationnaires :
— Différenciation simple ou saisonnière (modèles ARIMA, SARIMA).
— Transformation Box-Cox.
— Extraction de la tendance ou des composantes saisonnières.

Importance de cette distinction : La stationnarité est un pré-requis essentiel pour de nombreux modèles statistiques traditionnels. Les modèles modernes (Prophet, TBATS, LSTM, etc.) peuvent traiter directement la non-stationnarité, mais comprendre la nature de la série permet d'améliorer la qualité de la prévision et d'ajuster les choix méthodologiques.

2.3 Modèles des séries temporelles

Une série temporelle peut être décomposée en ses composantes principales : tendance T_t, saisonnalité S_t et bruit ϵ_t. Selon la nature des interactions entre ces composantes, deux modèles principaux sont utilisés :

— **Modèle additif** :

$$Y_t = T_t + S_t + \epsilon_t$$

Dans ce modèle, les effets de la tendance, de la saisonnalité et du bruit s'ajoutent linéairement. Il est adapté lorsque l'amplitude des variations saisonnières reste à peu près constante au fil du temps.

— **Modèle multiplicatif** :

$$Y_t = T_t \times S_t \times \epsilon_t$$

Ici, les composantes se multiplient, ce qui permet de représenter des séries où les variations saisonnières augmentent ou diminuent proportionnellement au niveau de la tendance. Ce modèle est particulièrement pertinent pour des séries dont les fluctuations deviennent plus importantes à mesure que la valeur globale augmente.

Le choix entre ces deux modèles dépend de la structure des données et est crucial pour obtenir des prévisions précises et fiables.

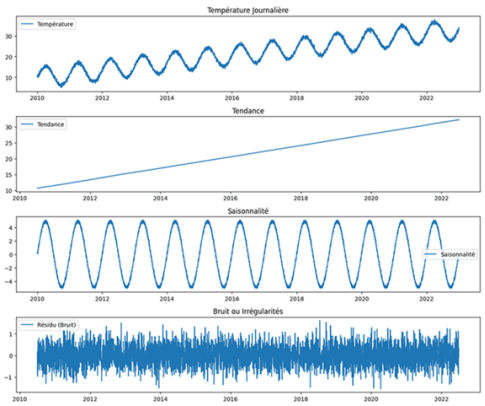


FIGURE 2 – Exemple de décomposition d'une série temporelle

3 Modèles de Prédiction

3.1 Modèles Classiques

Les modèles classiques de prévision des séries temporelles constituent la base de l'analyse prédictive avant l'essor des approches hybrides et des algorithmes modernes. Ils reposent sur des méthodes statistiques rigoureuses permettant de capturer les tendances, la saisonnalité et les cycles des données historiques pour estimer les valeurs futures.

3.1.1 Modèle AR (AutoRegressif)

Le modèle AR(p) (Auto-Régressif d'ordre p) est une méthode classique de prévision des séries temporelles qui repose sur l'hypothèse que la valeur future d'une variable peut être expliquée comme une combinaison linéaire de ses valeurs passées. Mathématiquement, il est défini par :

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

où :

- p : représente l'ordre du modèle, c'est-à-dire le nombre de valeurs passées (ou retards) utilisées pour la prévision.
- φ_i : sont les coefficients à estimer, reflétant l'influence de chaque valeur passée sur la valeur future.
- ε_t : est un bruit blanc, représentant les fluctuations aléatoires et imprévisibles de la série.
- c : est une constante qui peut être ajoutée pour ajuster le niveau moyen de la série.

Le modèle AR est particulièrement efficace pour capturer les dépendances temporelles linéaires dans une série, permettant ainsi de produire des prévisions basées sur l'historique récent. Cependant, il est limité lorsque la série présente des effets saisonniers complexes ou des comportements non linéaires.

3.1.2 Modèle MA (Moving Average)

Le modèle MA(q) (Moyenne Mobile d'ordre q) est une approche classique de prévision des séries temporelles qui se concentre sur l'influence des erreurs passées (ou résidus) sur la valeur actuelle de la série. Il est défini par la relation suivante :

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

où :

- q : est l'ordre du modèle, c'est-à-dire le nombre d'erreurs passées prises en compte.
- θ_i : sont les coefficients à estimer, qui pondèrent l'influence de chaque erreur passée sur la valeur actuelle.
- ε_t : représente le bruit blanc à l'instant t, supposé indépendant et identiquement distribué.
- μ : est la moyenne de la série.

Le modèle MA est particulièrement utile pour corriger les effets d'autocorrélation dans les résidus et pour lisser les fluctuations aléatoires. Contrairement au modèle AR, qui s'appuie sur les valeurs passées de la série, le MA met l'accent sur les chocs aléatoires récents pour améliorer les prévisions.

3.1.3 Modèle ARMA (AutoRegressive Moving Average)

Le modèle ARMA(p, q) combine les principes des modèles AR (Auto-Régressif) et MA (Moyenne Mobile) pour capturer à la fois les dépendances linéaires des valeurs passées et l'influence des erreurs passées sur la série temporelle. Il est défini par :

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

où :
— p : est l'ordre AR, soit le nombre de valeurs passées prises en compte.
— q : est l'ordre MA, soit le nombre d'erreurs passées utilisées.
— θ_i et ε_t sont les coefficients à estimer pour les parties AR et MA respectivement.
— ε_t : représente le bruit blanc à l'instant t.
— c : est une constante qui ajuste le niveau moyen de la série.

Le modèle ARMA est efficace pour modéliser des séries stationnaires présentant à la fois des autocorrélations des valeurs passées et des chocs aléatoires persistants. Il constitue une extension naturelle des modèles AR et MA, offrant une plus grande flexibilité pour représenter la dynamique réelle des séries temporelles, tout en restant limité aux séries stationnaires.

3.1.4 Mesures de dépendance série-temporelle

Analyse de l'Autocorrélation (ACF) et choix de l'ordre p

L'Analyse de l'Autocorrélation, représentée par la fonction ACF (Autocorrelation Function), est un outil statistique essentiel pour comprendre la structure interne d'une série temporelle. Le graphique de l'ACF mesure la corrélation entre la série et ses valeurs passées à différents retards, appelés lags.

Que représente le graphe de l'ACF ? Le graphe de l'ACF affiche, pour chaque lag k, le degré de corrélation entre X_t et X_{t+k}.

$$\rho(k) = \frac{\mathrm{Cov}(X_t, X_{t+k})}{\sqrt{\mathrm{Var}(X_t)\mathrm{Var}(X_{t+k})}}$$

Un lag correspond à un décalage temporel.
— **Lag 1** : corrélation entre aujourd'hui et hier.
— **Lag 2** : corrélation entre aujourd'hui et avant-hier.

Dans un graphe ACF, chaque barre représente la force de la corrélation entre la série et sa version décalée d'un certain nombre de périodes (lag). Plus la barre est haute, plus la relation entre le présent et le passé est forte. Pour savoir si cette relation est réellement

importante ou simplement due au hasard, le graphique contient deux lignes horizontales appelées *bornes de significativité*.

Si une barre reste entre ces deux lignes, cela signifie que la corrélation observée peut être expliquée par des variations aléatoires — elle n'est donc pas statistiquement significative. En revanche, si une barre dépasse ces lignes, cela indique que la corrélation est trop forte pour être due au hasard : elle devient alors statistiquement significative. Cela veut dire que ce lag a un véritable impact sur la série et qu'il peut être utile ou nécessaire pour la modélisation (par exemple pour choisir l'ordre d'un modèle AR ou MA).

PACF (Partial AutoCorrelation Function)

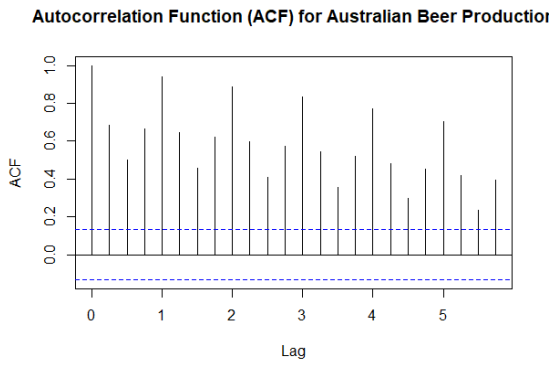
La fonction d'autocorrélation partielle (PACF) est un outil statistique essentiel dans l'analyse des séries temporelles. Elle permet de mesurer la corrélation entre une valeur de la série et une valeur décalée dans le temps (lag k), tout en éliminant l'effet des corrélations intermédiaires. Autrement dit, alors que l'ACF (Autocorrelation Function) évalue la corrélation totale entre y_t et y_{t-k}, la PACF isole uniquement la corrélation directe entre ces deux points, en neutralisant l'influence des lags 1, 2, ...,k1.

Mathématiquement, la PACF au lag k correspond au coefficient de corrélation résiduelle obtenu après avoir régressé séparément :

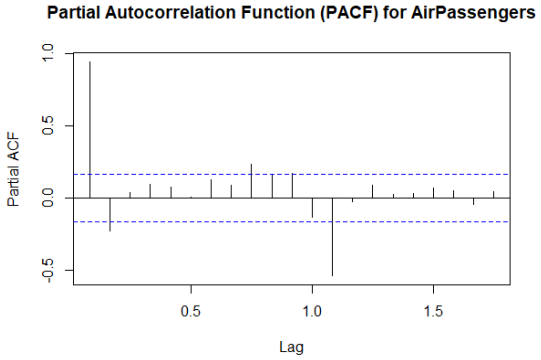
— y_t sur y_{t-1},..., y_{t-k-1}.
— y_{t-k} sur y_{t-1},..., y_{t-k-1}.

Puis, en calculant la corrélation entre les résidus des deux régressions. Cette approche permet de capturer uniquement la part de la relation qui n'est pas expliquée par les dépendances intermédiaires.

Sur le plan interprétatif, la PACF joue un rôle crucial dans la sélection des paramètres des modèles ARIMA, notamment pour déterminer l'ordre p du composant autorégressif (AR). Généralement, un décrochage net (cut-off) après le lag p dans le graphique PACF indique qu'un modèle AR(p) est approprié. Contrairement à l'ACF, qui peut montrer une décroissance progressive, la PACF permet d'identifier plus clairement l'étendue de la dépendance directe dans la série.



(a) Fonction d'autocorrélation (ACF)



(b) Fonction d'autocorrélation partielle (PACF)

FIGURE 3 – Graphiques ACF et PACF pour déterminer p et q

3.2 Modèles Saisonniers : SARIMA

Le modèle SARIMA (Seasonal AutoRegressive Integrated Moving Average) constitue une extension du modèle ARIMA destinée à capturer à la fois la dynamique non saisonnière et la structure saisonnière d'une série temporelle. Noté SARIMA(p,d,q)(P,D,Q), il combine des composantes autorégressives (AR), de différenciation (I), et de moyennes mobiles (MA), aussi bien au niveau ordinaire qu'au niveau saisonnier.

La formulation générale du modèle est donnée par :

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - \sum_{i=1}^{P} \Phi_i L^{is})(1 - L)^d (1 - L^s)^D X_t = (1 + \sum_{i=1}^{q} \theta_i L^i)(1 + \sum_{i=1}^{Q} \Theta_i L^{is})\varepsilon_t$$

où :
— L est l'opérateur de retard (lag), défini par $LX_t = X_{t-1}$.
— p, d, qsont respectivement les ordres autorégressif, de différenciation, et de moyenne mobile.
— P, D, Q sont les ordres des composants saisonniers.
— s désigne la période saisonnière (par exemple 12 pour des données mensuelles, 7 pour des données journalières hebdomadaires).
— ϵ_t représente un bruit blanc.

Le modèle introduit deux types de transformations :

— **Les composantes non saisonnières**

  — **AR(p)** : capture les dépendances directes à court terme.
  — **I(d)** : supprime les tendances en appliquant une différenciation ordinaire $(1 - L)^d$.
  — **MA(q)** : modélise les chocs instantanés et leur propagation.

— **Les composantes saisonnières**

  — **SAR(P)** : incorpore des dépendances espacées de s périodes via le terme $1 - \sum_{i=1}^{P} \Phi_i L^{is}$.
  — **SI(D))** : élimine une structure saisonnière répétitive par la différenciation $(1 - L^s)^D$.
  — **SMA(Q)** : prend en compte les chocs saisonniers via le terme $1 + \sum_{i=1}^{Q} \Theta_i L^{is}$.

En combinant ces mécanismes, SARIMA permet de modéliser des séries présentant à la fois tendance, autocorrelations de court terme, et répétitivité saisonnière. Ce modèle est particulièrement performant pour les séries où les motifs saisonniers sont réguliers et où l'effet saisonnier influence fortement les valeurs futures.

3.3 Modèles Modernes

Les modèles modernes de prévision des séries temporelles se distinguent des approches statistiques classiques par leur capacité à capturer des dynamiques complexes, non linéaires et souvent influencées par de multiples facteurs externes. Contrairement aux modèles traditionnels qui reposent principalement sur la stationnarité et les dépendances

linéaires, les modèles modernes — tels que Prophet, TBATS, ou encore les modèles basés sur le machine learning et le deep learning — traitent les séries temporelles comme des systèmes structurellement riches. Ils décomposent les données en plusieurs composantes (tendance, saisonnalités multiples, effets d'événements) et utilisent des méthodes avancées comme les transformations Box-Cox, les Fourier series, les réseaux neuronaux ou l'analyse des changements de régime pour modéliser les variations complexes. Leur architecture flexible permet d'intégrer à la fois des variables exogènes, des irrégularités, des points de rupture et des comportements non stationnaires, offrant ainsi une capacité prédictive plus robuste et mieux adaptée aux environnements réels où les données évoluent de manière dynamique et imprévisible.

3.3.1 Prophet (Facebook)

Prophet est un modèle de prévision développé par Facebook (Meta) conçu spécialement pour analyser les séries temporelles présentant à la fois tendance non linéaire, forte saisonnalité, et irrégularités structurelles. Il repose sur une décomposition strictement additive, ce qui permet d'interpréter facilement les contributions de chaque composante du modèle. Prophet est particulièrement efficace pour les données à longue durée, avec des motifs saisonniers complexes et des perturbations comme les jours fériés.

Le modèle s'écrit sous la forme :

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

où :

- $g(t)$ représente la tendance (trend), modélisée à l'aide de fonctions linéaires ou logistiques permettant de capturer des croissances non linéaires et d'intégrer des points de changement (« changepoints ») où la dynamique évolue brusquement.
- $s(t)$ correspond aux saisonnalités multiples, telles que les variations journalières, hebdomadaires ou annuelles. Prophet utilise pour cela des séries de Fourier qui permettent de modéliser des motifs saisonniers complexes et lisses.
- $h(t)$ modélise l'effet des jours fériés et événements spéciaux, grâce à une composante explicite qui peut ajouter ou soustraire des effets locaux sur des dates fixes ou variables.
- ϵ_t est un terme d'erreur capturant les variations aléatoires et les anomalies ponctuelles.

L'un des atouts majeurs de Prophet est sa robustesse aux valeurs aberrantes et sa capacité à intégrer automatiquement ou manuellement des changements structurels dans la série. De plus, la séparation claire des composantes facilite l'interprétation et améliore la flexibilité pour les prévisions réelles, notamment dans les applications de commerce, finance, séries web, et planification opérationnelle.

3.3.2 TBATS

Le modèle **TBATS** (acronyme pour Trigonometric, Box-Cox transformation, ARMA errors, Trend and Seasonal components) est une méthode avancée de prévision conçue pour analyser des séries temporelles présentant plusieurs saisonnalités complexes, souvent irrégulières ou non entières, comme celles observées dans le e-commerce, les transports, l'énergie ou les données web. Il constitue l'un des modèles les plus performants lorsque la

série présente plusieurs cycles simultanés, de longues périodes saisonnières ou des variations saisonnières changeantes dans le temps.

Mathématiquement, une version simplifiée du modèle BATS/TBATS peut être exprimée comme :

$$y_t^{(\omega)} = \ell_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t$$

où :

- ω est le paramètre de transformation Box-Cox, permettant de stabiliser la variance et de rendre la série plus stationnaire
- ℓ_t représente le niveau de la série.
- b_t correspond à la tendance, éventuellement amortie par le paramètre ϕ.
- $s_t^{(i)}$ désigne la composante saisonnière associée à la période m_i, TBATS étant capable d'intégrer plusieurs périodicités simultanées.
- d_t suit un modèle ARMA, permettant de modéliser les erreurs résiduelles de façon plus flexible.

L'innovation majeure de TBATS réside dans sa modélisation des saisonnalités à l'aide de fonctions trigonométriques, ce qui le rend particulièrement adapté aux séries présentant des périodicités longues ou non régulières — une limite majeure des modèles classiques comme SARIMA. De plus, grâce à la transformation Box-Cox et à l'intégration d'un modèle ARMA pour les erreurs, TBATS offre une excellente robustesse face aux données bruyantes et aux structures saisonnières non stationnaires.

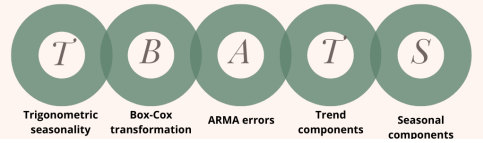


FIGURE 4 – Trigonometric, Box-Cox transformation, ARMA errors, Trend and Seasonal components

4 Réalisation du Projet

4.1 Description et préparation du jeu de données

Le projet s'appuie sur un jeu de données transactionnel comprenant des informations détaillées sur les commandes et les clients. Les principales colonnes incluent : **Customer ID** identifiant unique du client, **Customer Status** représentant le statut du client (Gold, Platinum, Silver), **Date Order was placed et Delivery Date** indiquant respectivement la date de commande et la date de livraison, **Order ID et Product ID** identifiants uniques de la commande et du produit, **Quantity Ordered** quantités commandées, **Total Retail Price for This Order** valeur totale de la commande et **Cost Price Per Unit** prix coûtant unitaire du produit.

Avant l'analyse, un travail de nettoyage et de prétraitement des données a été réalisé afin de s'assurer de la qualité et de la cohérence des informations. Les valeurs manquantes ont été traitées, les doublons supprimés, et certaines colonnes ont été harmonisées pour simplifier l'interprétation. Par exemple, les statuts des clients ont été standardisés pour uniformiser les catégories, remplaçant les mentions "GOLD", "PLATINUM" et "SILVER" par "Gold", "Platinum" et "Silver". Cette étape a permis de garantir une cohérence dans les analyses statistiques et graphiques ultérieures.

4.2 Analyse exploratoire des données

L'analyse exploratoire a consisté à étudier les tendances générales, la saisonnalité et la volatilité des ventes dans le temps. Des visualisations graphiques ont été produites pour observer l'évolution des ventes et des quantités commandées sur la période étudiée. L'objectif était d'identifier les motifs récurrents, les pics saisonniers et les anomalies. Cette exploration a révélé une forte saisonnalité récurrente ainsi qu'une tendance globale des ventes qui augmentait ou diminuait selon les périodes, justifiant le choix d'un modèle multiplicatif pour la prévision, capable de prendre en compte la variation proportionnelle des composantes saisonnières par rapport au niveau global de la série.

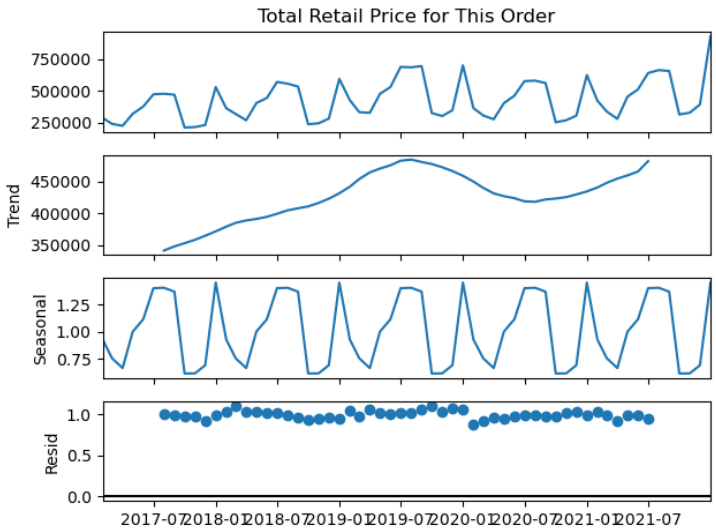


FIGURE 5 – Visualisation des tendances, saisonnalité et du bruit

Des visualisations supplémentaires ont permis d'étudier le comportement des clients selon leur statut. La répartition des commandes totales par catégorie — Silver, Gold, Platinum — a été représentée afin de comprendre le poids de chaque segment dans le chiffre d'affaires. Cette étape a mis en évidence la contribution significative des clients Gold et Platinum dans les ventes totales, fournissant ainsi un aperçu stratégique sur les segments prioritaires à cibler.

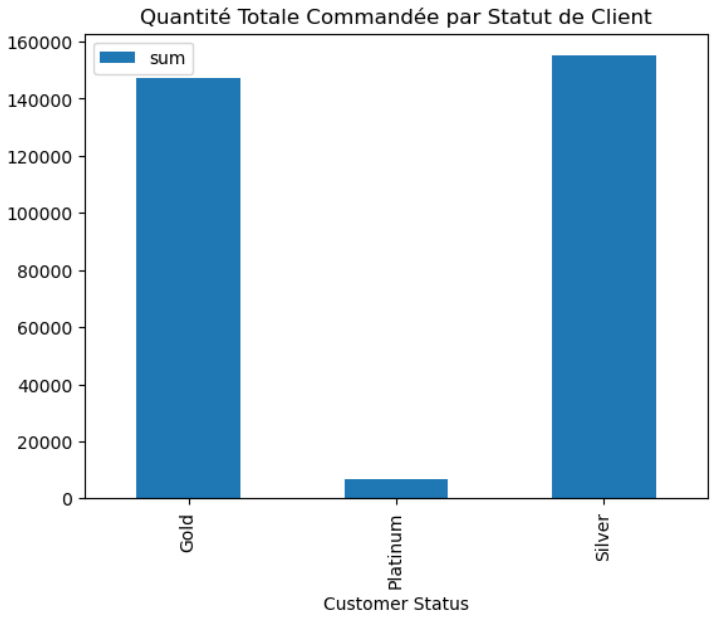


FIGURE 6 – Visualisation des comportement des clients

4.3 Analyse des produits et fusion des données

Pour enrichir l'analyse, le jeu de données des commandes a été combiné avec le jeu de données produit, comprenant les colonnes : Product ID, Product Line, Product Category, Product Group, Product Name, ainsi que les informations sur les fournisseurs (Supplier Country, Supplier Name, Supplier ID). La fusion de ces deux jeux de données a permis de créer un ensemble complet liant les commandes aux caractéristiques des produits.

Cette approche a facilité l'analyse des produits les plus vendus, des catégories et groupes de produits dominants, ainsi que des relations entre les ventes et les caractéristiques du produit. Par exemple, il a été possible d'identifier quelles lignes de produits généraient le plus de ventes, quelles catégories étaient les plus populaires. Cette étape a également permis de détecter des tendances dans les préférences des clients, en croisant les informations sur le statut des clients avec les produits achetés, offrant ainsi une vision stratégique complète pour l'optimisation des stocks et la planification commerciale.

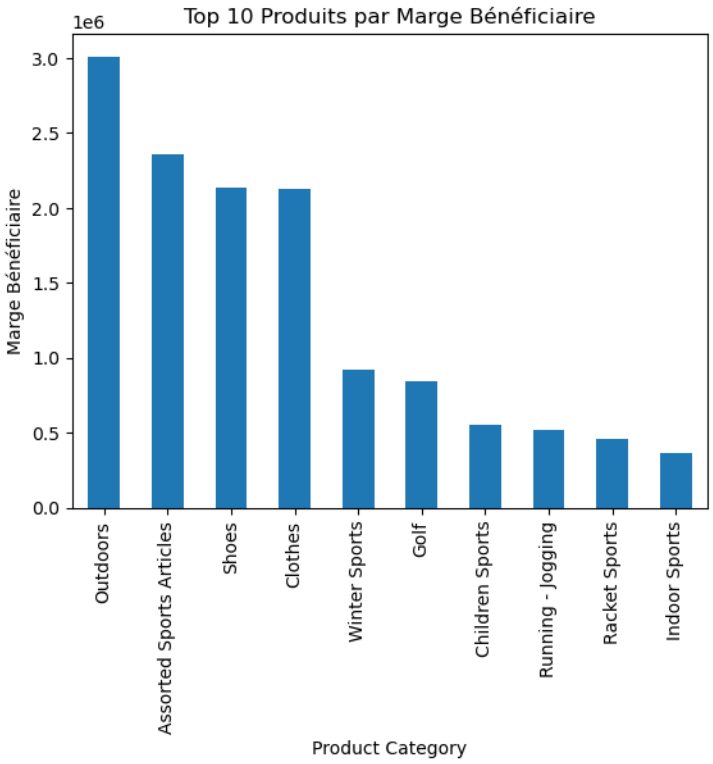


FIGURE 7 – Top 10 Catégories par Marge Bénéficiaire

4.4 Synthèse de l'étape de préparation et d'exploration

L'ensemble de ces étapes — nettoyage des données, harmonisation des statuts clients, visualisation des tendances et de la saisonnalité, fusion des jeux de données et analyse des produits — a permis d'obtenir un jeu de données fiable, riche et exploitable. Ces préparations ont posé les bases pour la construction de modèles de prévision robustes, en garantissant que les informations exploitées reflètent fidèlement la réalité des ventes et des comportements clients. La combinaison des analyses temporelles et des analyses produit/segment client offre une compréhension complète du marché, essentielle pour développer des stratégies de prévision précises et opérationnelles.

4.5 Feature Engineering : enrichissement de la série temporelle

Après le nettoyage et la préparation du jeu de données, une étape essentielle a consisté à enrichir la base par un feature engineering ciblé, permettant d'améliorer la capacité prédictive des modèles. Plusieurs nouvelles variables ont été créées, notamment des features temporelles et des lags, chacune jouant un rôle spécifique dans la compréhension des dynamiques de vente.

4.5.1 Features de date

Les données temporelles brutes (par exemple Date Order was placed) ont été transformées en plusieurs colonnes dérivées permettant d'extraire des informations utiles :

Les nouvelles features temporelles ajoutées — Day, Month, Year, Day of Week et Week of Year — permettent d'enrichir la série en capturant différents patterns temporels : le Day

révèle les variations quotidiennes liées au cycle du mois (comme les hausses en début ou fin de mois), le Month modélise les effets saisonniers tels que les périodes de fêtes ou promotions, tandis que le Year met en évidence les tendances longues et l'évolution structurelle des ventes. Le Day of Week est essentiel pour saisir les comportements hebdomadaires (pics le week-end, creux en milieu de semaine), et enfin le Week of Year aide à capturer des cycles hebdomadaires plus larges ou des campagnes commerciales récurrentes.

Ces nouvelles colonnes permettent aux modèles modernes (Prophet, TBATS, Machine Learning) de comprendre plus finement les patterns temporels présents dans les ventes.

4.5.2 Features de lag

Les lags représentent les valeurs passées de la variable cible (ex : ventes du jour J-1, J-7...), et sont indispensables pour capturer la dépendance temporelle :

Les variables Lag 1, Lag 7 et Lag 30 enrichissent la série en intégrant sa mémoire interne : Lag 1 (vente de la veille) permet de capturer la dynamique très courte des ventes, particulièrement utile dans les séries fortement autocorrélées ; Lag 7 (vente de la semaine précédente) met en évidence les comportements hebdomadaires récurrents ; et Lag 30 (vente du mois précédent) aide à intégrer les cycles mensuels ainsi que les variations intra-mensuelles.

Ces features enrichissent le modèle en lui permettant d'intégrer la mémoire du système : les ventes récentes influencent souvent celles du futur immédiat, surtout dans les contextes retail.

4.6 Modélisation Time Series : préparation et visualisation

Avant l'entraînement des modèles de prévision, la série temporelle a été agrégée par jour afin de définir clairement la variable cible :

4.6.1 Définition des ventes journalières

Les données brutes contenant les commandes ont été regroupées pour calculer le total des ventes par jour, ce qui constitue la structure standard en prévision temporelle. Cette agrégation permet de lisser les fluctuations intra-journalières et d'obtenir une vision précise de l'évolution quotidienne de l'activité du magasin.

4.6.2 Visualisation de la série temporelle

Une représentation graphique de la série a ensuite été réalisée afin d'observer :
— la tendance globale (augmentation ou diminution des ventes).
— la présence de saisonnalité (hebdomadaire, mensuelle, annuelle).
— les variations brusques ou anomalies.

Cette étape visuelle est déterminante car elle permet de confirmer les caractéristiques observées lors de l'analyse exploratoire :
— Une tendance évolutive.
— Une saisonnalité marquée.

— Des fluctuations proportionnelles au niveau global, indiquant qu'un modèle multiplicatif est plus adapté.

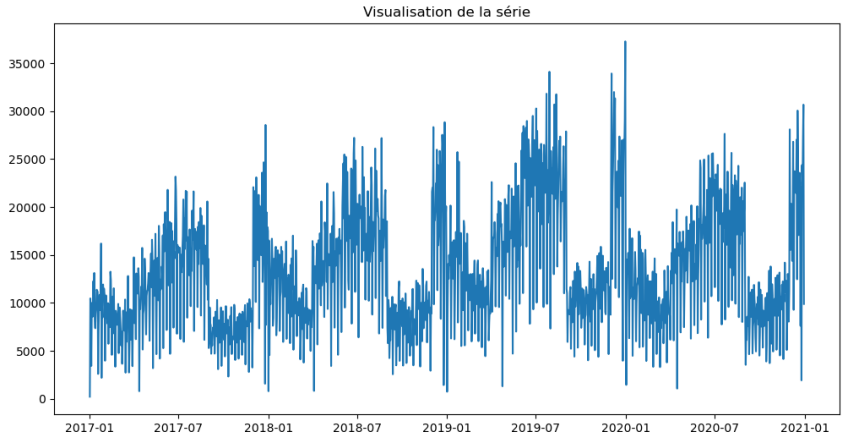


FIGURE 8 – Visualisation de la série

4.7 Test des modèles

4.7.1 Modèle ARMA

Après avoir préparé la série temporelle avec les features temporelles et les lags, la première étape de modélisation a consisté à tester les modèles **ARMA** afin de capturer les dépendances linéaires à court terme. Pour déterminer les paramètres p et q, nous avons utilisé les graphiques ACF et PACF. Ces visualisations permettent de mesurer respectivement la corrélation globale et partielle entre les valeurs présentes et leurs retards.

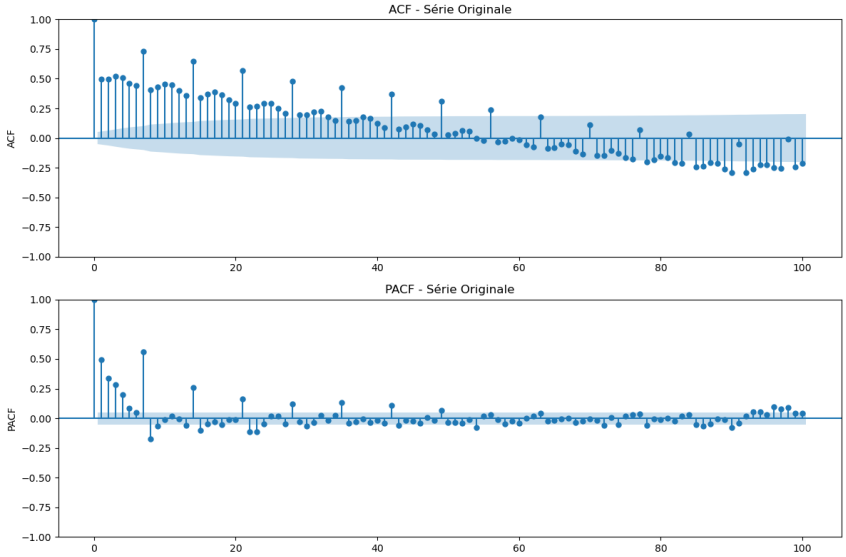


FIGURE 9 – Visualisation des courbes ACF et PACF

Lors de l'exploration, le graphique PACF a montré des pics significatifs aux premiers lags, suggérant une influence directe des valeurs passées sur la série. Cependant, les graphiques ne présentaient pas une décroissance nette et rapide vers zéro, ce qui indique que

la série n'était pas purement stationnaire et que des effets saisonniers persistants pouvaient masquer les signaux de dépendance. Malgré la sélection initiale des paramètres ARMA, les prédictions obtenues restaient imprécises et incapables de reproduire fidèlement les motifs réels des ventes.

4.7.2 Modèle ARIMA et choix du différentiation

Pour traiter la non-stationnarité, nous avons ensuite testé le modèle ARIMA, en expérimentant plusieurs ordres de différentiation **d**. L'objectif était de rendre la série plus stable et d'éliminer la tendance globale afin que les composantes **AR** et **MA** puissent modéliser correctement les fluctuations résiduelles.

FIGURE 10 – Choix du différentiation

Après plusieurs essais, nous avons constaté que d = 1 offrait les meilleurs résultats en termes de stabilité et de cohérence des résidus. Néanmoins, l'analyse des graphiques ACF et PACF des résidus montrait toujours des corrélations significatives sur plusieurs lags. La décroissance des autocorrélations ne tendait pas vers zéro à partir d'un certain rang, ce qui traduit la présence de motifs saisonniers forts et récurrents que le modèle ARIMA, basé sur des dépendances linéaires et une seule différentiation, ne peut pas capturer efficacement.

4.7.3 Modèle SARIMA

Face à la forte saisonnalité identifiée, le modèle SARIMA a été testé. SARIMA est conçu pour gérer des séries avec une saisonnalité unique en combinant des composantes ARIMA ordinaires avec des composantes AR, I, MA saisonnières. Nous avons donc choisi des paramètres saisonniers correspondant à la période identifiée dans la série.

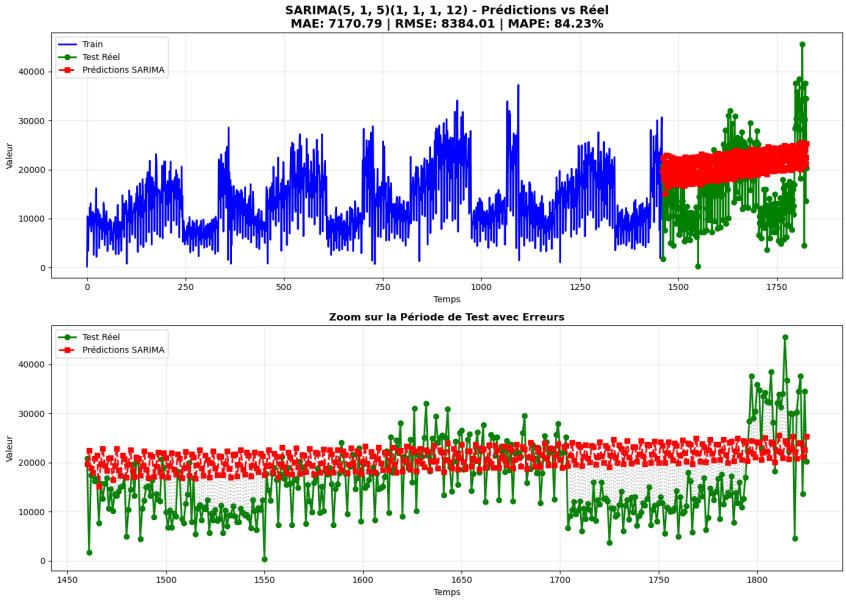


FIGURE 11 – Modèle SARIMA

Cependant, malgré ce réglage, les prédictions SARIMA se sont révélées très insatisfaisantes. Les courbes prévisionnelles ne suivaient pas le comportement réel des ventes, et les erreurs étaient très élevées. L'analyse approfondie a montré que la série présentait plusieurs saisonnalités simultanées — par exemple des cycles hebdomadaires et annuels coexistant — ce qui dépasse les capacités de SARIMA, limité à une seule saisonnalité. Les résidus du modèle présentaient encore des motifs structurés, confirmant que SARIMA n'était pas adapté à ce type de série multi-saisonnière.

4.7.4 Modèles modernes de prévision : Prophet et TBATS

Après les résultats insatisfaisants des modèles classiques, nous avons orienté la modélisation vers des modèles modernes capables de gérer plusieurs saisonnalités et des effets non linéaires. L'objectif était d'améliorer la précision des prévisions tout en capturant les motifs complexes observés dans les ventes quotidiennes.

Modèle Prophet

Le modèle Prophet repose sur une approche additive, décomposant la série en tendance, saisonnalités multiples et effets spécifiques comme les jours fériés. Pour exploiter pleinement ses capacités, nous avons enrichi la série avec des événements particuliers connus pour influencer les ventes, tels que le Black Friday et la rentrée scolaire. Ces variables contextuelles permettent au modèle de prévoir plus précisément les pics ou baisses significatives autour de ces dates.

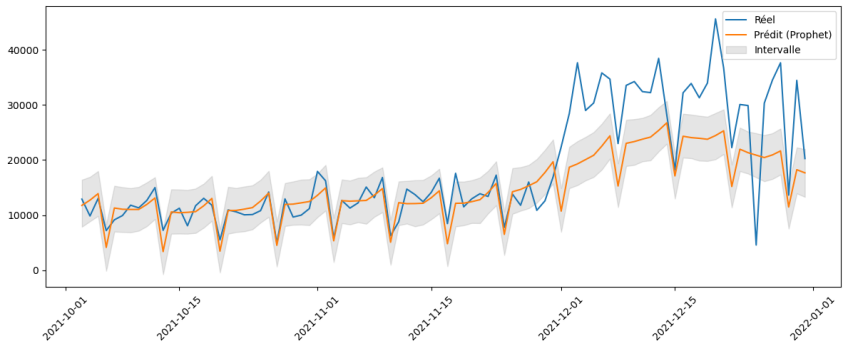


FIGURE 12 – Modèle Prophet

Les visualisations de la série montrent clairement que Prophet capture :

— la tendance globale des ventes.
— les pics hebdomadaires et mensuels grâce aux composantes saisonnières.
— les effets ponctuels des jours spéciaux.

Grâce à cette approche, les prévisions produites sont très proches des valeurs réelles, avec des erreurs significativement réduites par rapport aux modèles ARIMA et SARIMA. Prophet offre également l'avantage d'une interprétation intuitive, chaque composante de la série étant explicitement représentée.

Modèle TBATS

Le modèle TBATS (Trigonometric, Box-Cox, ARMA errors, Trend and Seasonal components) est particulièrement performant pour les séries présentant plusieurs saisons et des cycles irréguliers. Il modélise la série à l'aide de saisonnalités trigonométriques multiples, d'une tendance amortie et d'un modèle ARMA sur les résidus pour capturer les variations aléatoires.

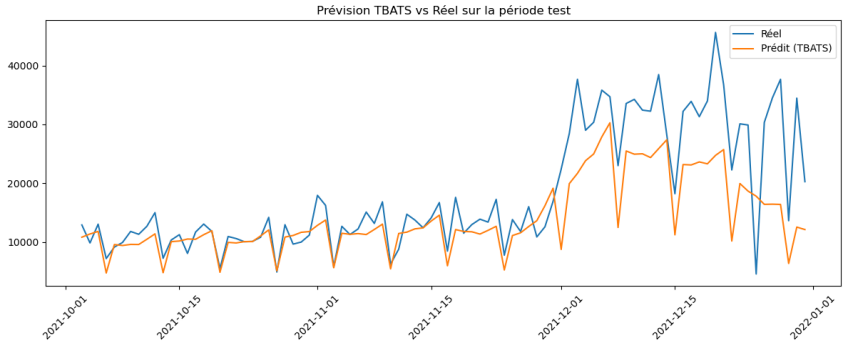


FIGURE 13 – Modèle TBATS

Nous avons appliqué TBATS sur la même série journalière, sans inclure explicitement les événements comme Black Friday, car le modèle intègre déjà des motifs saisonniers complexes de manière automatique. Les résultats ont montré que TBATS pouvait reproduire correctement la plupart des cycles saisonniers simultanés. Cependant, la précision sur les pics liés à des événements ponctuels était légèrement inférieure à Prophet, notamment pour les dates ayant un impact très marqué sur les ventes.

4.8 Résultats Finaux et Comparaison

Modèle	RMSE	MAE	MAPE (%)	Temps d'entraînement
TBATS	7,127.64	4,830.64	24.28	182s
Prophet	**6,704.94**	**4,537.51**	**23.98**	144s

TABLE 1 – Comparaison finale des performances

4.9 Application Streamlit

Après avoir sélectionné le modèle Prophet comme solution optimale, la dernière étape du projet a consisté à intégrer ce modèle dans une interface interactive afin de permettre aux utilisateurs de visualiser et d'exploiter facilement les prévisions de ventes futures. Pour cela, nous avons développé une application web légère avec Streamlit, une bibliothèque Python simple et efficace dédiée à la création d'interfaces de data science.

Le modèle entraîné a été importé directement depuis son fichier sauvegardé (Pickle), ce qui permet de l'utiliser sans réentraînement et de garantir une cohérence totale entre les résultats obtenus lors de l'expérimentation et ceux proposés dans l'application. L'interface offre aux utilisateurs la possibilité de sélectionner une plage de dates futures, puis d'obtenir instantanément une prédiction du volume de ventes attendu pour ces périodes.

Grâce à Streamlit, nous avons également enrichi l'application avec plusieurs visualisations dynamiques, notamment l'évolution prévue des ventes, les intervalles de confiance, ainsi que la décomposition de la tendance et des saisonnalités détectées par le modèle. Ces éléments permettent non seulement d'interpréter les prévisions, mais aussi de comprendre les facteurs structurels sous-jacents aux variations anticipées.

L'application constitue ainsi un outil opérationnel, facilement utilisable par l'équipe commerciale ou les décideurs. Elle facilite la planification des stocks, l'anticipation des pics de demande et l'optimisation des stratégies de vente. Cette étape finalise le projet en transformant le modèle de prévision en une solution pratique, accessible et directement exploitable au sein de l'entreprise.

FIGURE 14 – Interface d'acceuil

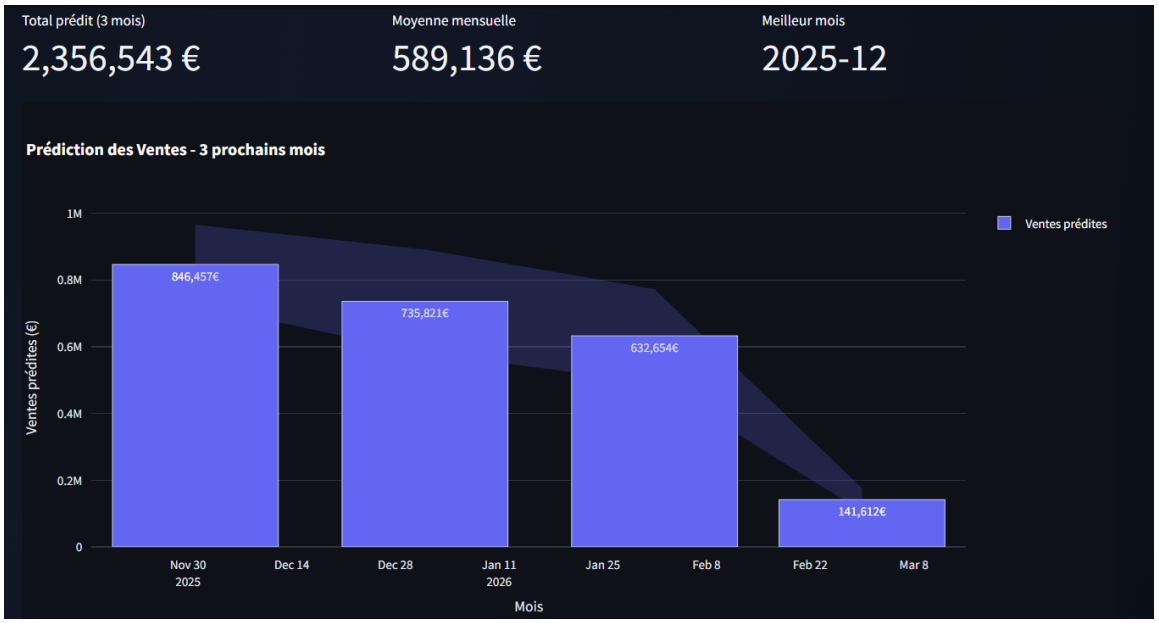


FIGURE 15 – Prédictions des 3 prochains mois

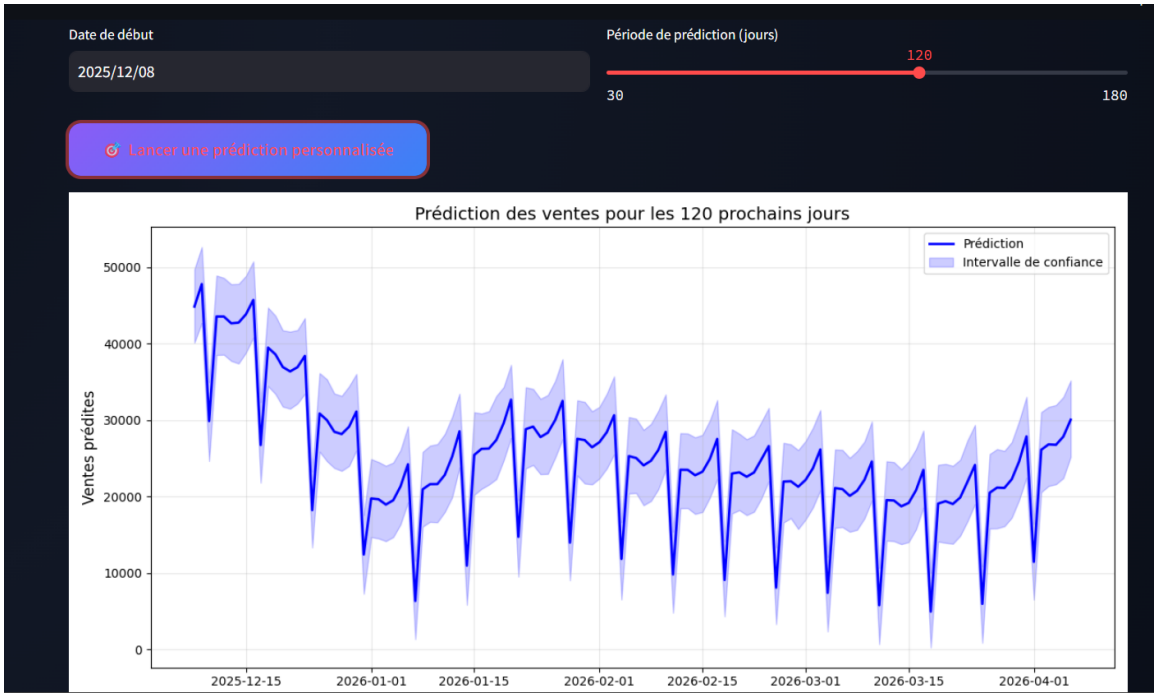


FIGURE 16 – Paramètres avancés

5 Conclusion

Ce projet a permis de développer une solution complète de prévision des ventes, intégrant à la fois une analyse approfondie des données, la sélection et le test de modèles statistiques et modernes, ainsi qu'une interface interactive pour l'exploitation des résultats. L'étude a débuté par un nettoyage minutieux des données et une exploration détaillée des tendances, de la saisonnalité et des comportements des clients et des produits. Cette étape a permis d'identifier les motifs récurrents et les anomalies, et de préparer des features pertinentes, telles que les variables temporelles, les lags et les événements spéciaux comme le Black Friday ou la rentrée scolaire.

Les modèles classiques (AR, MA, ARMA, ARIMA, SARIMA) ont été testés, mais ont montré leurs limites face à la complexité de la série temporelle, notamment la forte saisonnalité et la présence de multi-saisonnalités. Ces observations ont conduit à l'utilisation de modèles modernes comme TBATS et Prophet, capables de capturer simultanément plusieurs cycles saisonniers et d'intégrer des variables contextuelles. Les comparaisons de performance ont révélé que Prophet offrait les prévisions les plus précises et robustes, en reproduisant fidèlement la tendance globale, les cycles saisonniers et les effets des événements ponctuels.

Enfin, l'intégration du modèle Prophet dans une interface Streamlit a permis de rendre les prévisions directement exploitables par les utilisateurs, offrant un outil pratique pour la planification des ventes, la gestion des stocks et la prise de décisions stratégiques. Ce projet démontre que la combinaison d'un travail rigoureux sur les données, d'une sélection méthodique des modèles et d'une interface interactive peut transformer des données brutes en insights opérationnels fiables, directement applicables dans un contexte commercial.

En conclusion, cette démarche illustre l'importance d'adapter les méthodes de prévision aux caractéristiques spécifiques des séries temporelles et aux besoins des utilisateurs, en privilégiant la flexibilité et l'interprétabilité pour obtenir des résultats à la fois précis et exploitables.

A Annexes

A.1 Code Source

Le code source complet est disponible sur GitHub : `https://github.com/douae-zouak/Trend-Prediction`