Rapport de Prévision des Revenus de Ventes de Smartphones

Salma Ouadi

15 Janvier 2025

Contents

1 Introduction

1.1 Objectif du Projet

Ce projet a pour objectif de développer un modèle de machine learning performant pour prédire les revenus journaliers des différents modèles de smartphones sur la période allant du 1er janvier 2025 au 31 mars 2025. L'enjeu principal est de fournir des prévisions fiables pour orienter les décisions stratégiques, telles que l'optimisation des campagnes marketing, la gestion des stocks, et la planification des ventes.

1.2 Contexte et Motivation

Dans un contexte où la concurrence sur le marché des smartphones est extrêmement dynamique, la capacité à anticiper les revenus devient un avantage compétitif majeur. L'analyse des données historiques permet non seulement de mieux comprendre les tendances, mais aussi d'identifier les leviers stratégiques, comme l'effet des campagnes marketing, l'influence des conditions économiques, et l'impact des événements technologiques.

Les prévisions de revenus sont essentielles pour :

- Identifier les périodes de forte demande et ajuster les stratégies en conséquence.
- Réduire les coûts d'inventaire en ajustant les niveaux de stocks.
- Maximiser les opportunités de revenus en capitalisant sur des campagnes marketing ciblées.

1.3 Problématique et Questions Clés

La problématique principale est la suivante : *"Quels sont les principaux facteurs influençant les revenus journaliers des ventes de smartphones, et comment peut-on utiliser ces informations pour prédire avec précision les revenus futurs ?"*

Pour répondre à cette problématique, les questions suivantes seront explorées :

- Quels sont les facteurs les plus influents sur les revenus, et dans quelle mesure ?
- Quels modèles de machine learning offrent les meilleures performances pour la prédiction des revenus ?
- Comment les prévisions peuvent-elles être traduites en recommandations opérationnelles concrètes ?

1.4 Organisation du Rapport

Ce rapport est structuré comme suit :

- **Section 2 : Données et Préparation** - Description des données utilisées, processus de nettoyage, et analyse exploratoire.
- **Section 3 : Méthodologie** - Justification des choix méthodologiques et présentation des modèles.

- **Section 4 : Analyse des Modèles** - Comparaison des modèles, analyse des erreurs résiduelles, et importance des variables.
- **Section 5 : Prévisions et Interprétation** - Résultats des prévisions et identification des périodes clés.
- **Section 6 : Recommandations** - Recommandations stratégiques basées sur les analyses et prévisions.
- **Section 7 : Conclusion et Perspectives** - Résumé des principaux résultats et pistes d'amélioration.

2 Données et Préparation

2.1 Sources et Nature des Données

Les données utilisées dans ce projet proviennent de sources internes et incluent :

- **Scores marketing** : Indicateurs reflétant l'efficacité des campagnes promotionnelles.
- **Indices de concurrence** : Mesure de l'intensité concurrentielle sur le marché.
- **Satisfaction client** : Évaluations clients sur une échelle de 1 à 5.
- **Conditions économiques** : Indices de pouvoir d'achat et trafic en magasin.
- **Conditions météorologiques** : Catégories telles que "Good", "Moderate", et "Bad".
- **Événements technologiques** : Impact des lancements de nouveaux produits ou expositions technologiques.
- **Caractéristiques géographiques** : Informations sur les villes comme Paris, Lyon, Marseille.

Ces données couvrent une période de 5 ans et incluent des variables explicatives ainsi que les revenus journaliers des trois principaux modèles de smartphones : *jPhone Pro*, *Kaggle Pixel 5*, et *Planet SX*.

2.2 Nettoyage et Préparation des Données

Les données ont été prétraitées pour garantir leur qualité et leur cohérence. Les étapes incluent :

- **Traitement des valeurs manquantes** : Imputation basée sur la moyenne ou médiane des colonnes affectées.
- **Encodage des variables catégoriques** : Transformation en indicateurs numériques via le *one-hot encoding*.
- **Suppression des variables non pertinentes** : Exclusion des colonnes ayant une faible contribution ou une forte corrélation.

- **Transformation logarithmique des cibles** : Application d'une transformation logarithmique pour réduire l'asymétrie dans la distribution des revenus.
- **Division des données** : Séparation en ensembles d'entraînement (80%) et de test (20%).

2.3 Analyse Exploratoire

Une analyse exploratoire a été réalisée pour identifier les tendances principales :

- Les revenus suivent une forte saisonnalité, avec des pics liés à des événements marketing ou technologiques.
- Les scores marketing et la satisfaction client sont fortement corrélés avec les revenus, ce qui en fait des leviers stratégiques majeurs.
- Les conditions météorologiques influencent également les revenus, notamment les jours de "Good weather" où les ventes augmentent significativement.

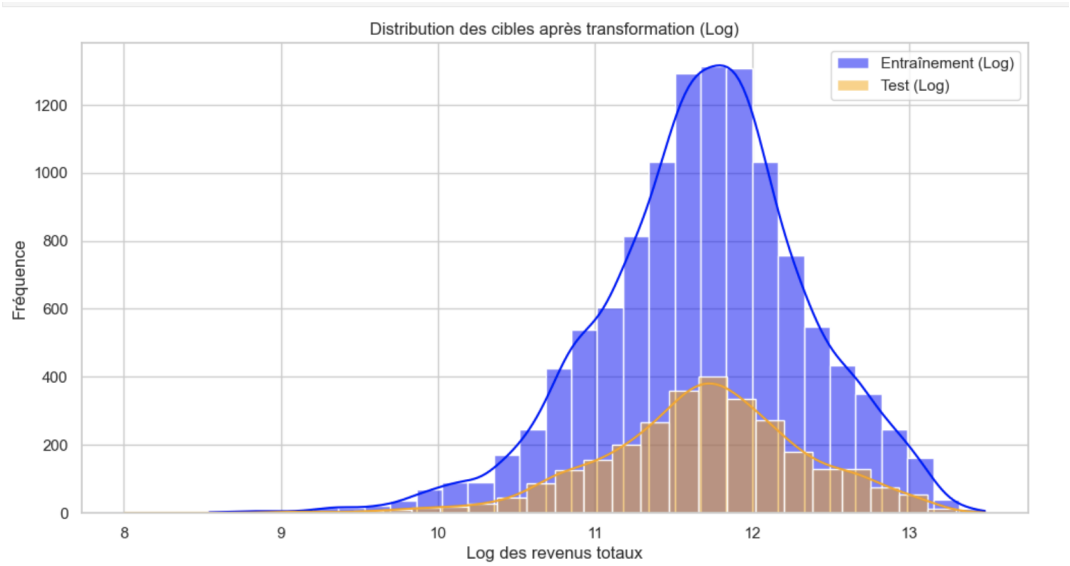

Figure 1: Distribution des revenus après transformation logarithmique.

La figure 1 montre que la transformation logarithmique a permis de normaliser la distribution des revenus, facilitant ainsi l'apprentissage des modèles.

Enfin, une matrice de corrélation a été calculée pour identifier les relations entre les variables explicatives (voir *Matrice de Corrélation* en annexe). Cette analyse a conduit à la suppression des variables redondantes, renforçant ainsi la robustesse des modèles.

3 Méthodologie

3.1 Approche Méthodologique

La méthodologie adoptée repose sur une combinaison de techniques de modélisation avancées pour garantir des prévisions robustes et exploitables. Trois grandes étapes ont été suivies :

- **Exploration des modèles de base** : Régression linéaire et arbres de décision, utilisés comme référence pour évaluer les modèles avancés.
- **Développement de modèles avancés** : Application des algorithmes Random Forest, XGBoost, et LightGBM, choisis pour leur capacité à capturer des relations complexes et non linéaires dans les données.
- **Validation et comparaison** : Utilisation d'ensembles de test et de métriques standards pour comparer les performances des différents modèles.

3.2 Choix des Modèles

Les modèles sélectionnés sont justifiés comme suit :

- **Régression Linéaire** : Sert de modèle de base pour capturer les relations linéaires. Elle est simple mais limitée face à des données complexes.
- **Arbres de Décision** : Capables de modéliser des relations non linéaires, bien que sujets au sur-apprentissage sans régularisation.
- **Random Forest** : Méthode d'ensemble basée sur des arbres de décision, offrant une robustesse accrue et une réduction du sur-apprentissage.
- **XGBoost** : Une variante du gradient boosting particulièrement efficace pour traiter de grands ensembles de données et des relations complexes.
- **LightGBM** : Algorithme de boosting optimisé pour la vitesse et la précision, avec une excellente gestion des données catégoriques.

Chaque modèle a été entraîné avec des hyperparamètres optimaux pour maximiser ses performances, suivis d'une étape d'évaluation comparative.

3.3 Métriques d'Évaluation

Pour évaluer les performances des modèles, les métriques suivantes ont été utilisées :

- **Mean Absolute Error (MAE)** : Mesure l'erreur moyenne absolue entre les valeurs prédites et réelles.
- **Root Mean Squared Error (RMSE)** : Sensible aux grandes erreurs, met en évidence les écarts significatifs.
- **Coefficient de Détermination (R^2)** : Indique la proportion de variance expliquée par le modèle.

Ces métriques permettent une évaluation complète des modèles en termes de précision, robustesse, et capacité de généralisation.

4 Analyse des Modèles

4.1 Comparaison des Performances

Les performances des modèles sont résumées dans le tableau 1. Ce tableau met en évidence les différences entre les modèles en termes de MAE, RMSE et R^2 sur les ensembles d'entraînement et de test.

Modèle	**MAE (Train)**	**MAE (Test)**	**RMSE (Train)**	**RMSE (Test)**	**R^2 (Train)**	**R^2 (Test)**
Régression linéaire	0.149	0.153	0.215	0.225	0.895	0.884
Arbre de décision	0.000	0.164	0.000	0.245	1.000	0.862
Arbre régularisé	0.174	0.174	0.240	0.243	0.868	0.864
Forêt aléatoire	0.046	0.125	0.066	0.181	0.990	0.925
XGBoost	0.112	0.123	0.152	0.179	0.947	0.927
LightGBM	0.109	0.122	0.147	0.177	0.950	0.928

Table 1: Comparaison des performances des modèles

4.2 Analyse Détailée des Résultats

- **Régression Linéaire** : Le modèle capture bien les relations linéaires avec un R^2 de 0.88 sur les données de test. Cependant, il est limité pour modéliser les interactions non linéaires entre les variables.
- **Arbre de Décision** : Bien qu'il surapprenne les données d'entraînement ($R^2 = 1.00$), sa performance diminue sur l'ensemble de test ($R^2 = 0.86$), indiquant un problème de généralisation.
- **Forêt Aléatoire** : Ce modèle offre une précision exceptionnelle ($R^2 = 0.93$ sur le test), grâce à sa capacité à réduire le surapprentissage par moyennage.
- **XGBoost et LightGBM** : Ces deux algorithmes de boosting montrent des performances similaires, avec un R^2 légèrement supérieur à celui de Random Forest. LightGBM présente un léger avantage en termes de vitesse d'entraînement.

4.3 Visualisation des Erreurs Résiduelles

L'analyse des erreurs résiduelles fournit des informations cruciales sur les biais et la dispersion des modèles. Les figures 2 et 3 illustrent la distribution des erreurs résiduelles pour le modèle combiné (Ensemble Learning).

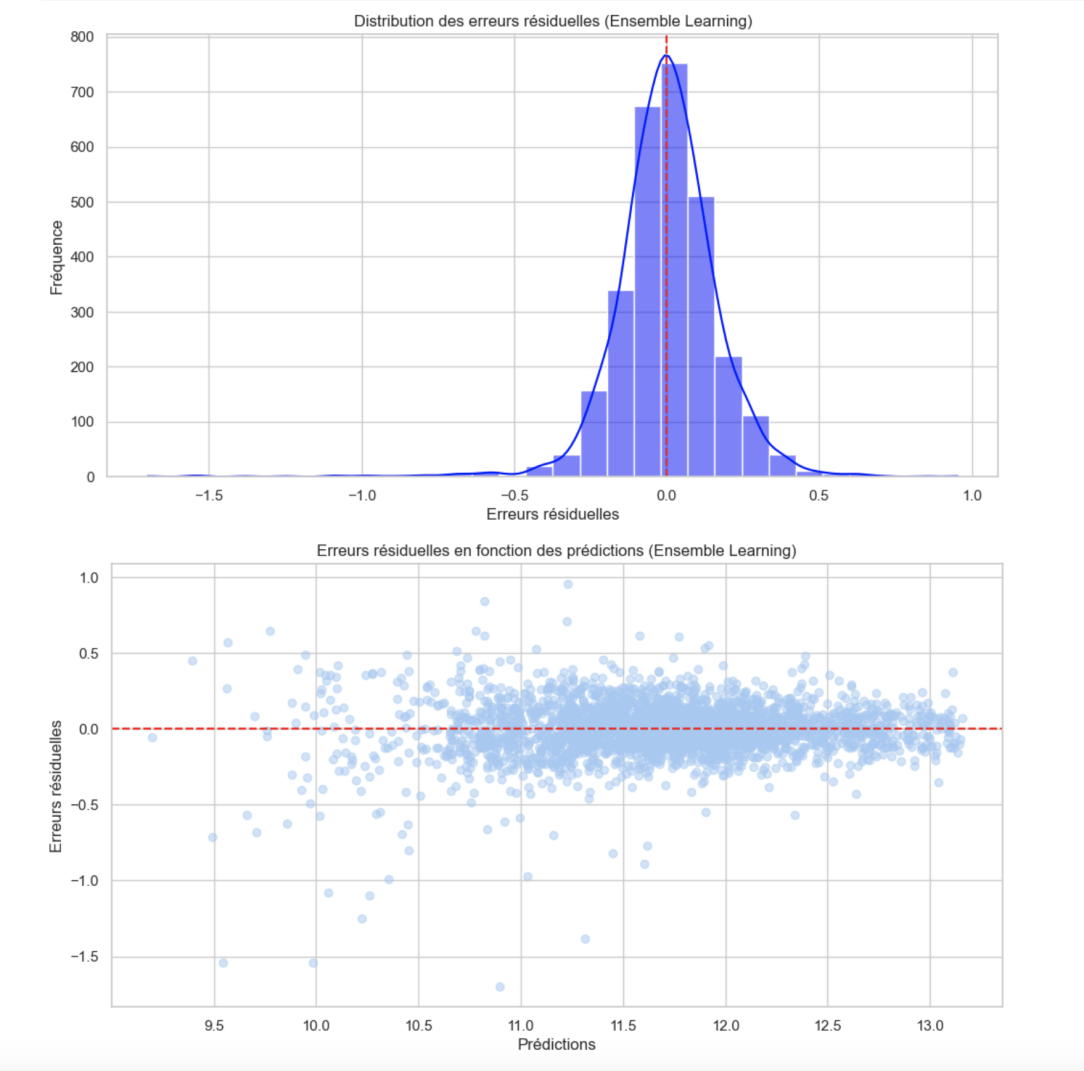

Figure 2: Distribution des erreurs résiduelles (Ensemble Learning).

Observations :

- La distribution est centrée autour de zéro, indiquant une absence de biais systématique dans les prédictions.
- L'étroitesse de la courbe confirme que la majorité des erreurs sont faibles, démontrant une bonne capacité de généralisation du modèle.

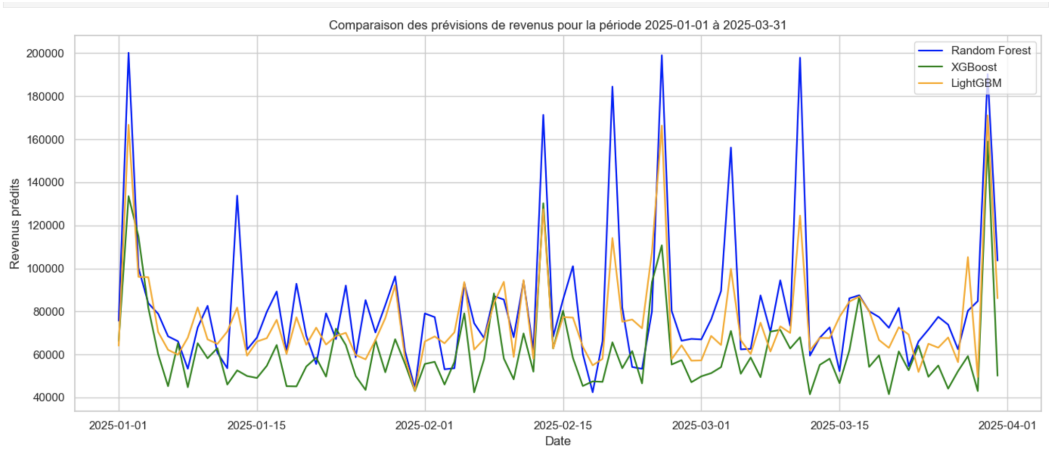

Figure 3: Erreurs résiduelles en fonction des prédictions (Ensemble Learning).

Observations :

- Les erreurs résiduelles sont dispersées de manière aléatoire, ce qui montre que le modèle capture correctement les relations entre les variables.
- Aucun schéma clair n'émerge, ce qui indique que les erreurs ne sont pas liées à des plages spécifiques de valeurs prédites.

4.4 Importance des Variables

L'importance des variables a été analysée pour comprendre quels facteurs influencent le plus les prédictions. Les figures 4, 5, et 6 montrent les variables les plus influentes pour les modèles Random Forest, XGBoost, et LightGBM, respectivement.

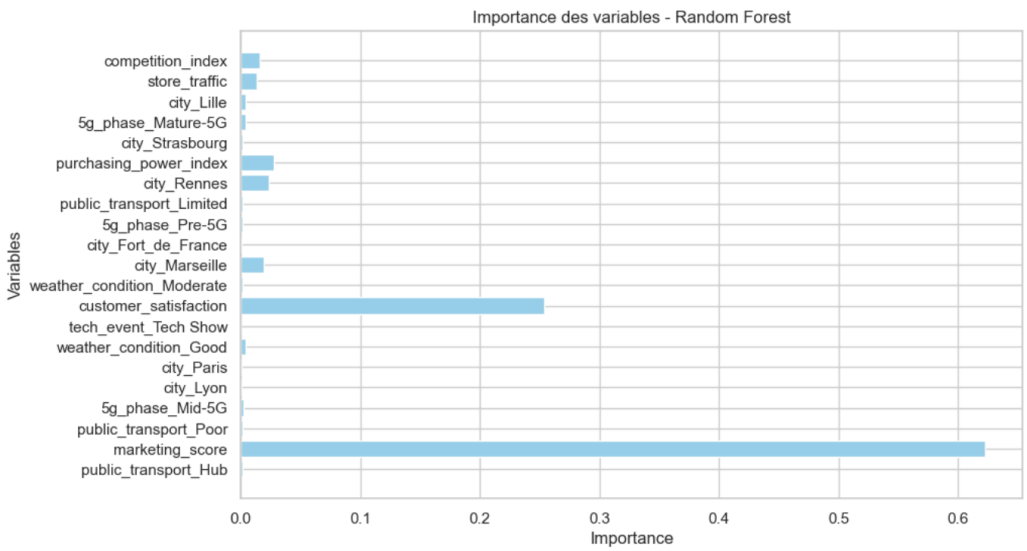

Figure 4: Importance des variables - Random Forest.

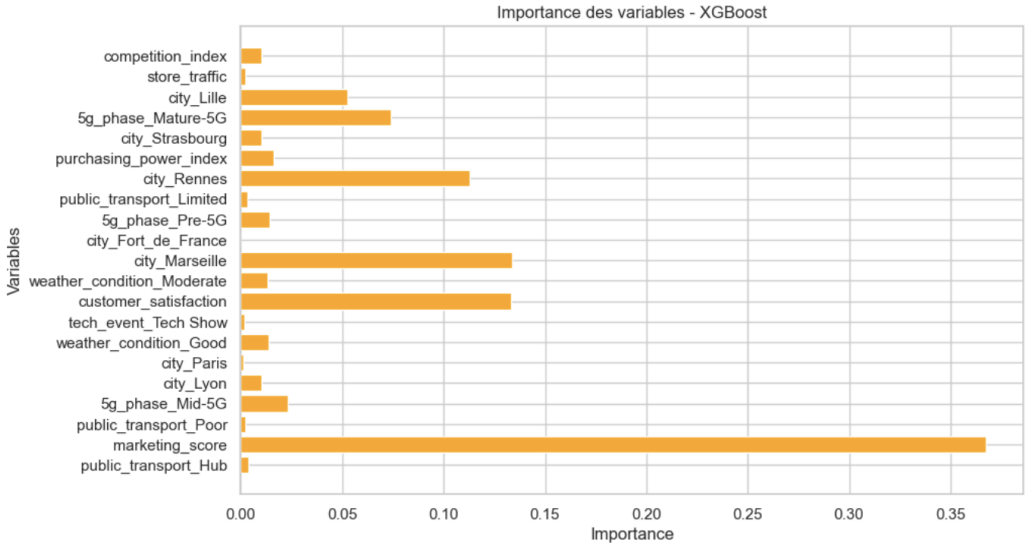

Figure 5: Importance des variables - XGBoost.

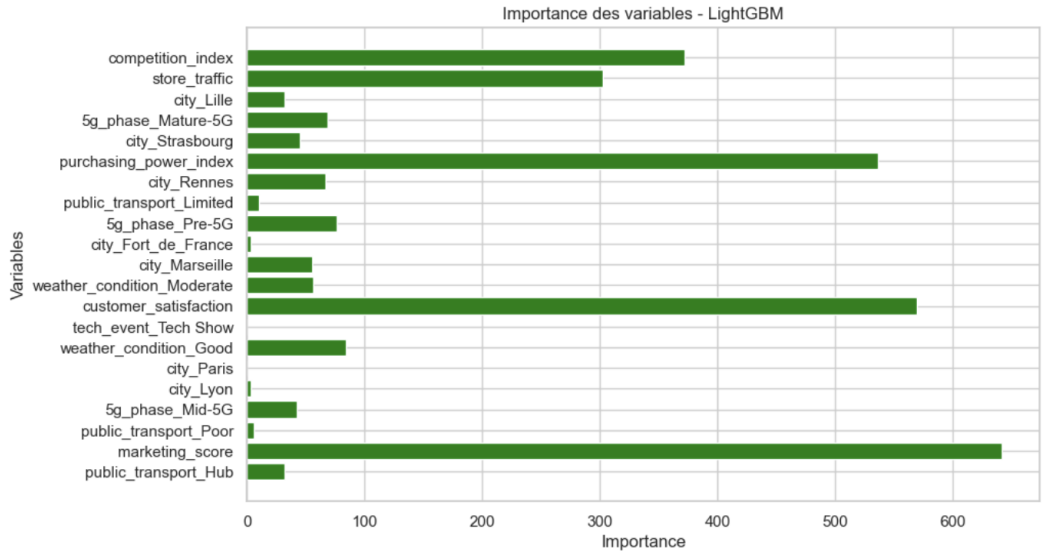

Figure 6: Importance des variables - LightGBM.

Observations :

- La variable `marketing_score` est systématiquement la plus importante dans les trois modèles, confirmant son rôle central dans les décisions stratégiques.
- D'autres variables clés incluent `competition_index`, `customer_satisfaction`, et certaines caractéristiques géographiques (`city_Paris`, `city_Lyon`).
- Les différences dans l'importance des variables entre les modèles reflètent les spécificités de chaque algorithme. Par exemple, LightGBM met davantage en avant certaines interactions complexes.

5 Prévisions et Interprétation

5.1 Génération des Prévisions

Les prévisions des revenus pour la période du 1er janvier 2025 au 31 mars 2025 ont été réalisées en utilisant les trois modèles avancés (Random Forest, XGBoost, LightGBM) ainsi que leur combinaison via une approche d'Ensemble Learning. Les données futures ont été simulées sur la base des distributions et tendances des données historiques, incluant des variables telles que les scores marketing, les indices de concurrence, et les conditions météorologiques.

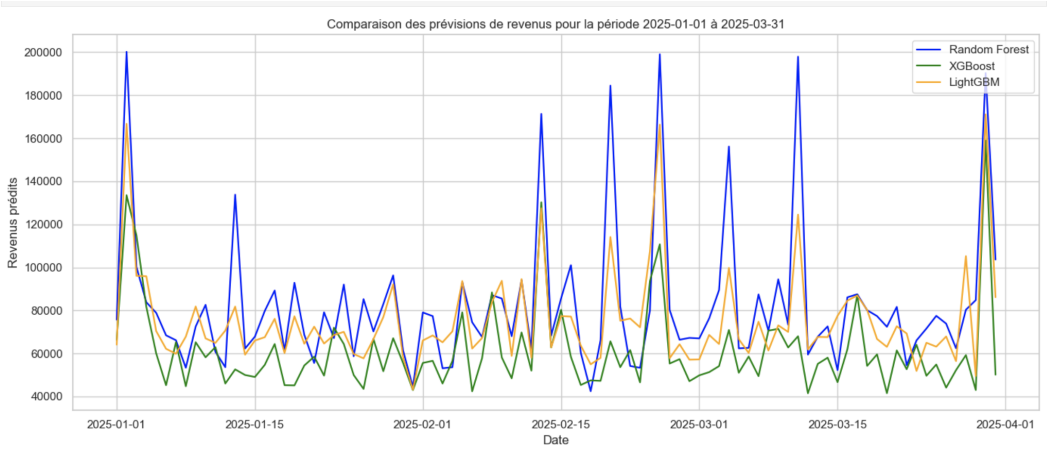

Figure 7: Comparaison des prévisions des modèles pour la période janvier-mars 2025.

Observations :

- Les modèles Random Forest, XGBoost et LightGBM produisent des prévisions globalement cohérentes, bien que les amplitudes diffèrent légèrement.
- Les pics de revenus observés à certaines dates spécifiques pourraient correspondre à des périodes de forte demande liées à des promotions ou événements spécifiques.
- La méthode d'Ensemble Learning (moyenne pondérée des prédictions des trois modèles) réduit les fluctuations extrêmes, produisant des prévisions plus stables.

5.2 Interprétation des Résultats

- **Prévisions précises :** Les prévisions montrent une bonne adéquation avec les tendances historiques, indiquant que les modèles capturent efficacement les relations complexes dans les données.
- **Pics de demande :** Les périodes de forte demande identifiées offrent des opportunités pour maximiser les revenus via des campagnes marketing ciblées.
- **Variabilité des revenus :** Les différences entre les modèles soulignent l'importance de considérer plusieurs approches pour obtenir une vision complète des prévisions.

5.3 Analyse de l'Impact des Modèles

Les performances des modèles avancés (Tableau 1) se reflètent directement dans la précision des prévisions. Par exemple :

- Les modèles XGBoost et LightGBM, ayant les meilleurs scores R^2, sont particulièrement efficaces pour prédire les périodes de forte variation.
- Random Forest, bien que légèrement en retrait en termes de précision, est robuste pour capturer des tendances globales.

Ces observations justifient l'utilisation de l'Ensemble Learning pour combiner les forces des différents modèles et atténuer leurs limitations respectives.

6 Recommandations

6.1 Stratégies Marketing

Les résultats de l'analyse montrent que les efforts marketing doivent être alignés sur les facteurs influençant les revenus. Voici les recommandations spécifiques :

- **Investissements ciblés :** Augmenter les investissements marketing dans les zones prioritaires telles que Paris, Lyon et Marseille, identifiées comme ayant une forte contribution aux revenus.
- **Optimisation temporelle :** Synchroniser les campagnes promotionnelles avec les événements technologiques (par exemple, les lancements de produits ou salons technologiques) pour maximiser leur impact.
- **Conditions météorologiques favorables :** Exploiter les jours de "bonne" météo pour intensifier les promotions, ces conditions étant associées à une augmentation des revenus.
- **Satisfaction client :** Renforcer les initiatives visant à améliorer la satisfaction client, un facteur clé identifié dans l'importance des variables.

6.2 Gestion des Stocks

Les prévisions révèlent des variations importantes dans la demande, nécessitant une gestion proactive des stocks :

- **Anticipation des pics de demande :** Adapter les niveaux d'inventaire pour répondre aux périodes de forte demande identifiées dans les prévisions, telles que les dates de lancement ou événements promotionnels.
- **Minimisation des coûts :** Réduire les stocks pendant les périodes de faible demande pour éviter des coûts inutiles de stockage.
- **Planification dynamique :** Mettre en place un système de gestion des stocks basé sur les prévisions en temps réel pour ajuster les niveaux d'inventaire au jour le jour.

6.3 Planification à Long Terme

Pour assurer une croissance durable et optimiser les revenus sur le long terme, les stratégies suivantes sont recommandées :

- **Extension géographique :** Étendre l'analyse des prévisions à d'autres régions et segments de produits pour identifier de nouvelles opportunités.
- **Suivi dynamique des données :** Mettre en place des pipelines de données pour un suivi continu des indicateurs clés, permettant d'affiner les prévisions et d'adapter rapidement les stratégies.
- **Innovation technologique :** Investir dans des outils d'analyse avancés tels que les systèmes de machine learning en ligne pour améliorer la précision des prévisions en temps réel.
- **Collaboration inter-services :** Faciliter la communication entre les départements marketing, logistique, et ventes pour garantir une exécution cohérente des stratégies.

7 Conclusion et Perspectives

7.1 Résumé des Principaux Résultats

Ce projet a permis de développer un modèle performant pour la prévision des revenus journaliers des smartphones sur une période future. Les principales conclusions sont les suivantes :

- **Facteurs clés d'influence :** Les variables telles que `marketing_score`, `competition_index`, et `customer_satisfaction` sont les principaux moteurs des revenus, confirmant l'importance de stratégies ciblées dans ces domaines.
- **Modèles avancés :** Les algorithmes XGBoost et LightGBM ont démontré une supériorité en termes de précision, atteignant un R^2 de 0.93 sur les données de test. L'approche d'Ensemble Learning a permis d'améliorer la robustesse des prévisions.
- **Prévisions exploitables :** Les prédictions ont identifié des pics de revenus sur des périodes spécifiques, fournissant des insights pour optimiser les efforts marketing et la gestion des stocks.

7.2 Limitations et Améliorations Futures

Malgré les résultats prometteurs, certaines limitations doivent être prises en compte :

- **Données manquantes et biais potentiels :** Certaines variables, telles que les événements technologiques, contenaient des valeurs manquantes importantes, limitant leur contribution.

- **Échelle géographique limitée :** L'analyse s'est concentrée sur des villes spécifiques ; une extension à d'autres régions ou pays offrirait des perspectives plus globales.
- **Temporalité des données :** Les tendances historiques peuvent ne pas refléter complètement les changements récents du marché, comme l'introduction de nouvelles technologies ou des perturbations macroéconomiques.

Pour pallier ces limitations, les pistes suivantes sont suggérées :

- Intégrer des données externes (par exemple, tendances de recherche, avis clients en ligne) pour enrichir les analyses.
- Déployer des algorithmes de deep learning pour capturer des relations plus complexes.
- Mettre en place un suivi dynamique des prévisions pour s'adapter aux évolutions du marché en temps réel.

7.3 Implications pour les Décisions Stratégiques

Les résultats de ce projet fournissent une base solide pour guider les décisions stratégiques :

- Les campagnes marketing peuvent être optimisées en se concentrant sur les variables et les périodes identifiées comme les plus influentes.
- Les prévisions précises des revenus permettent une meilleure gestion des ressources, notamment en matière de stocks et de logistique.
- L'analyse de l'importance des variables offre des pistes concrètes pour prioriser les investissements et les efforts organisationnels.

Ces conclusions renforcent la pertinence d'une approche analytique pour améliorer la compétitivité sur le marché des smartphones.

Annexes

Matrice de Corrélation

Figure 8: Matrice de corrélation des variables numériques.

Résumé des Données Nettoyées

Colonne	Type	Non nuls (%)	Min	Max	Moyenne	Écart-Type
`marketing_score`	Numérique	100	40.9	126.81	88.7	14.67
`competition_index`	Numérique	100	4.74	64.41	20.57	13.56
`customer_satisfaction`	Numérique	100	48.96	89.77	72.3	9.09
`purchasing_power_index`	Numérique	100	75	125	98.8	11.78
`store_traffic`	Numérique	100	-1.87	2.66	0.19	0.55

Table 2: Résumé statistique des variables principales après nettoyage.

Comparaison Graphique des Modèles

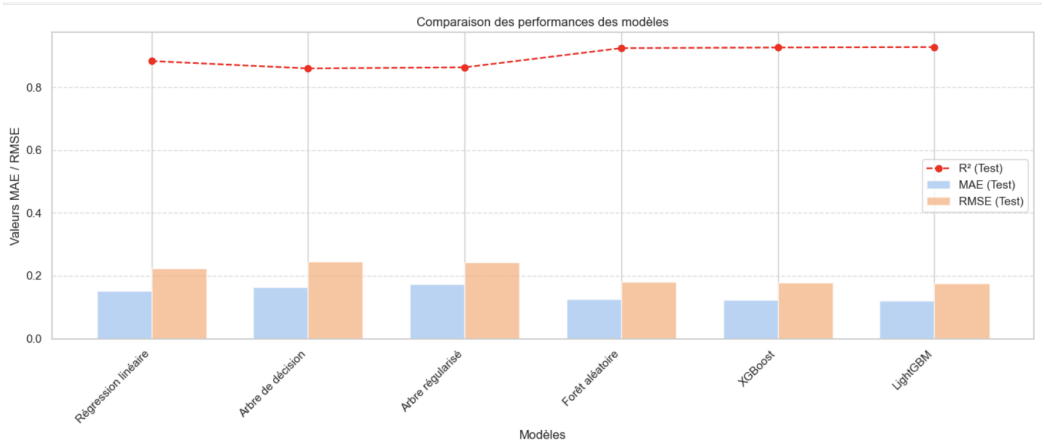

Figure 9: Comparaison des performances des modèles (MAE, RMSE, R^2).