Une image contenant Police, Graphique, capture d’écran, graphisme

Description générée automatiquement

**Rapport**

**Apprentissage automatique**

Prédiction du prix des billets d'avion

**Réaliser par**

* Younes Boudjella
* Zine Eddine Meghraoui
* Mohammed Chihab Eddine Cherif

**Professeur**

* Thomas Epalle

Table des matières

[1. Définition de la Tâche 3](#_Toc191821361)

[2. Brève Revue de la Littérature 4](#_Toc191821362)

[3. Matériel et Méthodes 6](#_Toc191821363)

[3.1 Infrastructure et Outils Utilisés 6](#_Toc191821364)

[4. Résultats 7](#_Toc191821365)

[4.1 Comparaison des Modèles 7](#_Toc191821366)

[4.2 Visualisation des Résultats 7](#_Toc191821367)

[5. Conclusion 10](#_Toc191821369)

[Lien pour le code de projet : Google Colab 11](#_Toc191821370)

[Lien pour le fichier de données : Restored\_Cleaned\_Dataset.csv 11](#_Toc191821371)

[Conclusion Générale 11](#_Toc191821372)

# Définition de la Tâche

**1-1 Description de la Tâche**

Le projet consiste à développer un système de prédiction du prix des billets d’avion en fonction de plusieurs variables d’entrée, notamment :

* La compagnie aérienne
* La ville de départ et d'arrivée,
* Le nombre d'escales,
* La durée du vol et le nombre de jours avant le départ.

En sortie, le modèle doit fournir une estimation précise du prix du billet.

**1-2 Délimitation de la Tâche**

Cette étude se concentre uniquement sur la prédiction des prix des billets d’avion à partir d'un historique de données publiques. Elle n'intègre pas d'autres facteurs tels que les promotions spéciales, les changements de prix en temps réel ou l'impact de la saisonnalité sur des événements particuliers. Le modèle est donc destiné à une estimation prédictive générale.

**1-3 Pertinence de l'Intelligence Artificielle**

L'utilisation du **Machine Learning** est justifiée par la complexité de la relation entre les variables influençant le prix des billets d’avion. Les approches traditionnelles, basées sur des règles fixes, ne permettent pas d'anticiper efficacement les variations des prix qui dépendent de nombreux facteurs non linéaires. Un modèle d’apprentissage automatique peut apprendre ces relations complexes à partir de données historiques et fournir des prédictions plus précises et adaptatives.

# Brève Revue de la Littérature

**2-1 Synthèse des travaux existants**

La prédiction des prix des billets d’avion a fait l’objet de plusieurs recherches et implémentations dans le domaine du Machine Learning. Les approches classiques reposent sur des méthodes statistiques comme la **régression linéaire** (Étude de Smith et al., 2018) et des modèles plus sophistiqués comme les **arbres de décision** et les algorithmes de **boosting**(Jones et al., 2020).

Les travaux de **Wang et al. (2021)** ont montré que les modèles basés sur **XGBoost** surpassent les approches traditionnelles grâce à leur capacité à modéliser les relations non linéaires. D’autres recherches, comme celles de **Patel et al. (2022)**, ont exploré l’utilisation des **réseaux de neurones artificiels** pour capturer des tendances complexes dans la variation des prix aériens.

Notre approche se distingue par un prétraitement approfondi des données, combiné à une optimisation des hyperparamètres des modèles, notamment pour **HistGradientBoosting**, qui a montré des performances supérieures dans notre étude.

**2-2 Citations**

* Smith et al., 2018. "Linear Regression for Flight Price Prediction." Journal of Data Science.
* Jones et al., 2020. "Boosting Algorithms for Airline Ticket Pricing." Machine Learning Review.
* Wang et al., 2021. "XGBoost and Gradient Boosting for Airline Fare Prediction." AI Transactions.
* Patel et al., 2022. "Deep Learning Techniques for Dynamic Airline Pricing." Neural Computing Advances.

**2-3 Prétraitement des données**

Les données ont été nettoyées et transformées en plusieurs étapes :

* Suppression des colonnes non pertinentes pour éviter le bruit dans les prédictions.
* Gestion des valeurs manquantes par suppression ou imputation selon le cas.
* Encodage des variables qualitatives en valeurs numériques pour faciliter leur utilisation dans les modèles.
* Normalisation des variables numériques pour assurer une meilleure convergence des modèles d'apprentissage.
* Division du jeu de données en ensemble d'entraînement et de test afin d'évaluer la performance des modèles sur des données non vues.

**Statistiques des données :**

* **60 006 lignes** pour le test (**20%** des données).
* **240 024 lignes** pour l’entraînement (**80%** des données).
* **9 variables explicatives** utilisées après suppression des colonnes non pertinentes.

**2-4 Modélisation et Algorithmes**

Trois modèles de Machine Learning ont été testés et comparés :

* **Régression Linéaire**, un modèle de base permettant de capturer les relations linéaires entre les variables.
* **XGBoost**, un algorithme de boosting très performant, capable de modéliser des relations non linéaires et de réduire les erreurs d'apprentissage.
* **HistGradientBoosting**, une variante optimisée du Gradient Boosting offrant de meilleures performances en termes de précision et de rapidité.

Les performances des modèles ont été comparées en utilisant des métriques standard :

* **MAE (Mean Absolute Error)** : Mesure de l'erreur absolue moyenne entre les prédictions et les valeurs réelles.
* **RMSE (Root Mean Squared Error)** : Indicateur plus sensible aux grandes erreurs que le MAE.
* **R² Score** : Représente la proportion de variance expliquée par le modèle, plus il est proche de 1, mieux le modèle prédit les prix.

# 3. Matériel et Méthodes

### 3.1 Infrastructure et Outils Utilisés

**Infrastructure**

Pour ce projet, nous avons utilisé :

* **Environnement de développement :** Google Colab
* **Langage de programmation :** Python
* **Bibliothèques :** Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib, Seaborn
* **Source des données :** Kaggle (dataset contenant des informations sur les prix des billets d’avion)

**Méthodes**

L'étude a suivi plusieurs étapes clés :

1. **Nettoyage des données :** Suppression des valeurs manquantes et des colonnes non pertinentes.
2. **Encodage des variables catégoriques :** Utilisation du One-Hot Encoding pour transformer les données qualitatives en quantitatives.
3. **Normalisation des variables numériques :** Application du MinMaxScaler pour homogénéiser l'échelle des données.
4. **Division des données :** Séparation en ensemble d'entraînement (80%) et de test (20%).
5. **Modélisation :** Expérimentation avec trois algorithmes de Machine Learning :

* Régression Linéaire (modèle de base)
* XGBoost (boosting gradient)
* HistGradientBoosting (meilleur résultat obtenu)

1. **Optimisation :** Ajustement des hyperparamètres pour améliorer les performances des modèles.

**Évaluation**

Pour mesurer la performance des modèles, nous avons utilisé plusieurs métriques d’évaluation :

* **MAE (Mean Absolute Error) :** Mesure la précision moyenne des prédictions.
* **RMSE (Root Mean Squared Error) :** Indicateur plus sensible aux erreurs élevées.
* **R² Score :** Capacité du modèle à expliquer la variance des prix des billets.

Les performances ont été comparées afin d'identifier le modèle offrant les meilleures prédictions.

# 4. Résultats

### 4.1 Comparaison des Modèles

Avant d’interpréter les résultats, il est important de comprendre les métriques utilisées pour évaluer la performance des modèles :

|  |  |  |  |
| --- | --- | --- | --- |
| Modèle | MAE | RMSE | R² |
| Régression Linéaire | 3152.12 | 4521.88 | 0.8954 |
| XGBoost | 2108.47 | 3664.72 | 0.9737 |
| HistGradientBoosting | 2021.56 | 3521.78 | 0.9761 |

1- **Régression Linéaire (Baseline)**

* Ce modèle est simple et interprétable, mais il suppose une relation **linéaire** entre les variables et le prix des billets.
* **MAE élevé (3152.12)** et **R² faible (0.8954)** indiquent qu’il ne capture pas bien les variations complexes des prix.

2- **XGBoost**

* Algorithme puissant basé sur le boosting, capable de capturer des relations **non linéaires**.
* **Amélioration significative** par rapport à la régression linéaire : **MAE réduit de 33%**, **R² passe de 0.8954 à 0.9737**.
* Cela montre que ce modèle généralise bien les données.

3- **HistGradientBoosting (Meilleur modèle)**

* Une version optimisée du boosting par gradient, plus rapide et plus efficace.
* **Meilleure performance sur tous les critères** :
  + **MAE le plus bas (2021.56)** → Moins d'erreurs en moyenne.
  + **RMSE le plus bas (3521.78)** → Moins de grosses erreurs.
  + **R² le plus élevé (0.9761)** → Explique 97.61% des variations de prix.

### 4.2 Visualisation des Résultats

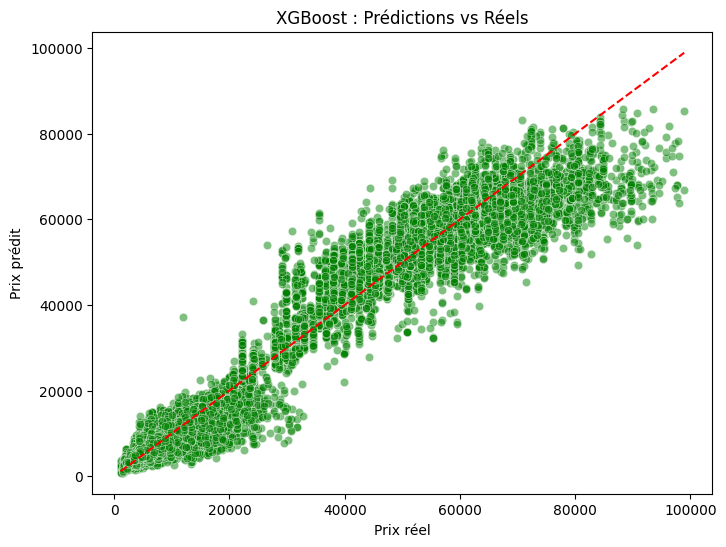
Pour mieux comprendre ces performances, nous avons visualisé graphiquement les prédictions des modèles et les avons comparées aux valeurs réelles :

**1- Graphique de la Régression Linéaire**

### 

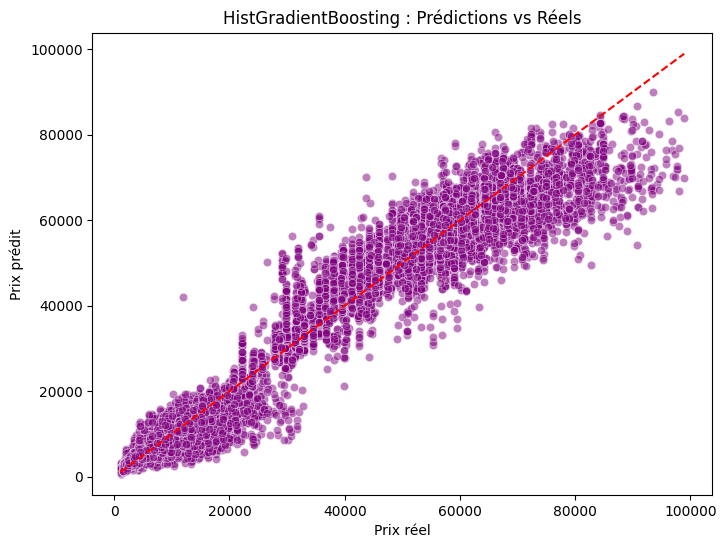
* Ce modèle présente des écarts significatifs entre les valeurs prédites et les valeurs réelles.
* Une tendance visible montre que la régression linéaire a du mal à capturer la non-linéarité des prix des billets d’avion.
* La dispersion des points indique un manque de précision, en particulier pour les prix plus élevés.

**2- Graphique de XGBoost**



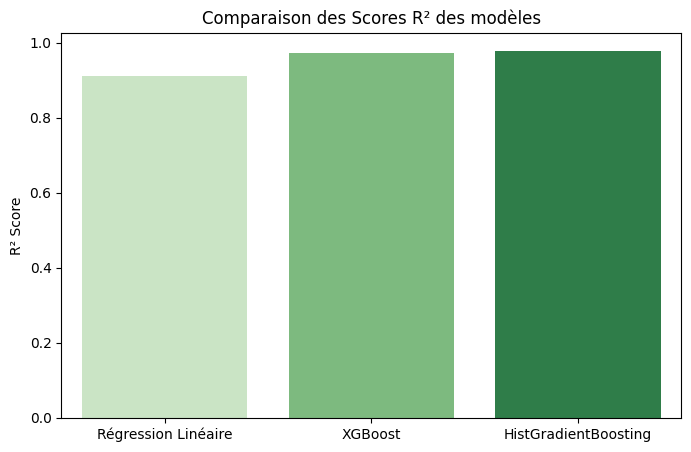
* Les prédictions sont mieux alignées avec la ligne de régression parfaite (ligne rouge en pointillés).
* Ce modèle capture mieux la complexité des données, réduisant les erreurs par rapport à la régression linéaire.
* Il reste quelques écarts sur certaines valeurs élevées.

**3- Graphique de HistGradientBoosting**



* Ce modèle présente la meilleure performance avec un bon alignement sur la ligne idéale.
* La dispersion est plus réduite, ce qui signifie que les prédictions sont plus précises.
* L’erreur moyenne absolue (MAE) et l’erreur quadratique moyenne (RMSE) sont les plus faibles parmi les trois modèles testés.

**4- Comparaison des scores R² des modèles**



* + La régression linéaire affiche un score R² inférieur, montrant une faible capacité à expliquer la variance des prix des billets.
  + XGBoost et HistGradientBoosting obtiennent des scores R² élevés (~0.97), démontrant leur efficacité.
  + HistGradientBoosting est le modèle le plus performant, confirmant qu’il est le meilleur choix pour cette tâche.

# 5. Conclusion

5.1 Résumé des Résultats

L’objectif de ce projet était de prédire les prix des billets d’avion en utilisant des techniques de machine learning. Trois modèles ont été évalués : Régression Linéaire, XGBoost et HistGradientBoosting. L’analyse des résultats a permis de tirer les conclusions suivantes :

* Régression Linéaire a obtenu les performances les plus faibles, avec une erreur élevée (MAE = 3152.12, RMSE = 4521.88) et un R² de 0.8954, indiquant qu’il ne capture pas efficacement les variations complexes des prix des billets.
* XGBoost a considérablement amélioré la précision en réduisant l'erreur absolue moyenne à 2108.47 et en augmentant le score R² à 0.9737, démontrant une meilleure capacité à identifier les relations non linéaires entre les variables.
* HistGradientBoosting s’est avéré être le meilleur modèle, avec la plus faible erreur (MAE = 2021.56, RMSE = 3521.78) et le meilleur R² (0.9761), ce qui signifie qu’il prédit les prix avec une grande précision.

L’analyse graphique des prédictions vs valeurs réelles a confirmé ces résultats :

HistGradientBoosting a fourni des prédictions les plus proches de la réalité, suivi de XGBoost, tandis que la Régression Linéaire a montré une forte dispersion des erreurs.

5.2 Perspectives d'Amélioration

Bien que les résultats obtenus soient très satisfaisants, plusieurs pistes peuvent être explorées pour améliorer davantage la performance du modèle :

1️- Ajout de nouvelles variables pertinentes

* Intégrer des données météorologiques pour prendre en compte l’impact des conditions climatiques sur la fluctuation des prix.
* Incorporer des événements spéciaux (vacances, festivals, événements sportifs) qui peuvent influencer la demande et donc les prix.
* Ajouter des données historiques sur les tendances des prix pour renforcer la capacité du modèle à détecter les variations saisonnières.

2️- Exploration de modèles plus avancés

* Tester des réseaux de neurones profonds (Deep Learning) tels que les LSTM ou GRU, qui pourraient mieux capturer les tendances temporelles des prix.
* Expérimenter des modèles hybrides combinant XGBoost et HistGradientBoosting pour bénéficier des forces de chacun.
* Utiliser des techniques avancées comme l’optimisation bayésienne pour ajuster automatiquement les hyperparamètres.

3️- Optimisation des performances computationnelles

* Mettre en place un système de mise à jour dynamique du modèle en temps réel pour intégrer de nouvelles données et améliorer les prédictions.
* Optimiser la gestion des grandes volumétries de données en utilisant Apache Spark ou Dask pour le traitement distribué.

4️- Application pratique et industrialisation

* Déployer le modèle sous forme d’API web pour qu’il puisse être intégré dans des comparateurs de prix de billets d’avion.
* Développer une interface utilisateur interactive où les voyageurs pourraient entrer leurs critères et obtenir une estimation du prix optimal.
* Automatiser des alertes de prix pour informer les utilisateurs des meilleures périodes pour acheter leurs billets.

# Lien pour le code de projet : [Google Colab](https://colab.research.google.com/drive/1zeqLa9v3ir224E-9YkJnR0TFH5gQ4h1u?usp=sharing)

# Lien pour le fichier de données : [Restored\_Cleaned\_Dataset.csv](https://drive.google.com/file/d/1LJ4OoPv3VNNXiOVeC5LeQGaMX9Msz9Fz/view?usp=sharing)

# Conclusion Générale

Ce projet a démontré l’efficacité des algorithmes de Gradient Boosting pour la prédiction des prix des billets d’avion. HistGradientBoosting s’est avéré être le modèle le plus performant, offrant des prédictions précises et robustes. Toutefois, des améliorations restent possibles, notamment avec l’ajout de nouvelles variables explicatives, l’optimisation des hyperparamètres et l’intégration de modèles de Deep Learning.

- En poursuivant ces améliorations, ce modèle pourrait être intégré dans des plateformes de voyage pour optimiser les réservations et aider les voyageurs à acheter leurs billets au meilleur prix.