Analyse et Modélisation des Prix des Voitures à l'Aide de l'Apprentissage Automatique

Ingénierie Informatique et Réseaux

Réalisé par :

ELKIRAM Malak

ESSAADI Achraf

RAJI Oussama

Encadré et Examiné par :

Mme Hadni Meryem

Contents

List of Figures

Introduction

Ce rapport présente une étude approfondie sur l'analyse et la modélisation des prix des voitures à l'aide d'algorithmes d'apprentissage automatique. L'objectif principal est de développer un modèle capable de prédire les prix des véhicules en se basant sur divers attributs. Nous avons utilisé un ensemble de données riche, comprenant des informations détaillées telles que la marque, le modèle, l'année de fabrication, le kilométrage, la taille du moteur et d'autres caractéristiques techniques. Cette approche permet d'obtenir une compréhension plus précise des facteurs influençant le prix des voitures sur le marché actuel.

Contexte du Projet

Problématique

La prédiction précise des prix des voitures est un défi crucial dans le secteur automobile. Une estimation correcte aide les concessionnaires à fixer des prix compétitifs, les acheteurs à prendre des décisions éclairées et les compagnies d'assurance à évaluer correctement les véhicules.

Objectifs

- Explorer les relations entre les différentes variables et le prix des voitures.
- Développer et évaluer des modèles de régression pour prédire le prix des voitures.
- Comparer les performances des différents modèles afin de sélectionner le plus performant.
- Fournir une analyse interprétable et utile pour les acteurs du marché automobile.

Préparation des Données

Description du Dataset

Le dataset utilisé dans ce projet contient des données relatives à des véhicules, collectées à partir de différentes sources. Il est crucial de bien comprendre ces variables avant de commencer toute modélisation.

- **Brand** : Marque du véhicule (variable catégorielle).
- **Model** : Modèle spécifique du véhicule (variable catégorielle).
- **Year** : Année de fabrication du véhicule (variable numérique).

- **Engine_Size** : Cylindrée du moteur en centimètres cubes ou litres (variable numérique).
- **Fuel_Type** : Type de carburant utilisé (variable catégorielle).
- **Transmission** : Type de transmission (automatique ou manuelle) (variable catégorielle).
- **Mileage** : Kilométrage total parcouru par le véhicule en kilomètres (variable numérique).
- **Doors** : Nombre de portes du véhicule (variable numérique).
- **Owner_Count** : Nombre de propriétaires précédents (variable numérique).
- **Price** : Prix du véhicule, notre variable cible (variable numérique).

Prétraitement des Données

Le prétraitement des données est une étape cruciale pour s'assurer que les données sont propres, structurées et prêtes à être utilisées pour la modélisation. Voici les étapes spécifiques que nous avons suivies :

1. **Gestion des Valeurs Manquantes** : Identification et traitement des valeurs manquantes, soit en les supprimant si le nombre est limité, soit en utilisant des techniques d'imputation (par exemple, la moyenne ou la médiane) si nécessaire.
2. **Conversion des Variables Catégorielles** : Utilisation de techniques d'encodage (One-Hot Encoding ou Label Encoding) pour transformer les variables catégorielles en variables numériques afin qu'elles puissent être utilisées par les algorithmes de machine learning.
3. **Normalisation des Caractéristiques Numériques** : Mise à l'échelle des caractéristiques numériques en utilisant 'StandardScaler' pour que toutes les variables soient sur la même échelle, ce qui aide à la performance des algorithmes.
4. **Séparation des Données** : Division du dataset en deux parties : 70% pour l'entraînement et 30% pour le test afin de vérifier l'efficacité des modèles construits.

Exploration des Données

Analyse Descriptive des Variables

Avant de construire des modèles, il est essentiel d'explorer les données et de comprendre leur distribution et leurs relations.

Distribution des Prix des Voitures

La figure 1 montre la distribution des prix des voitures dans notre dataset.

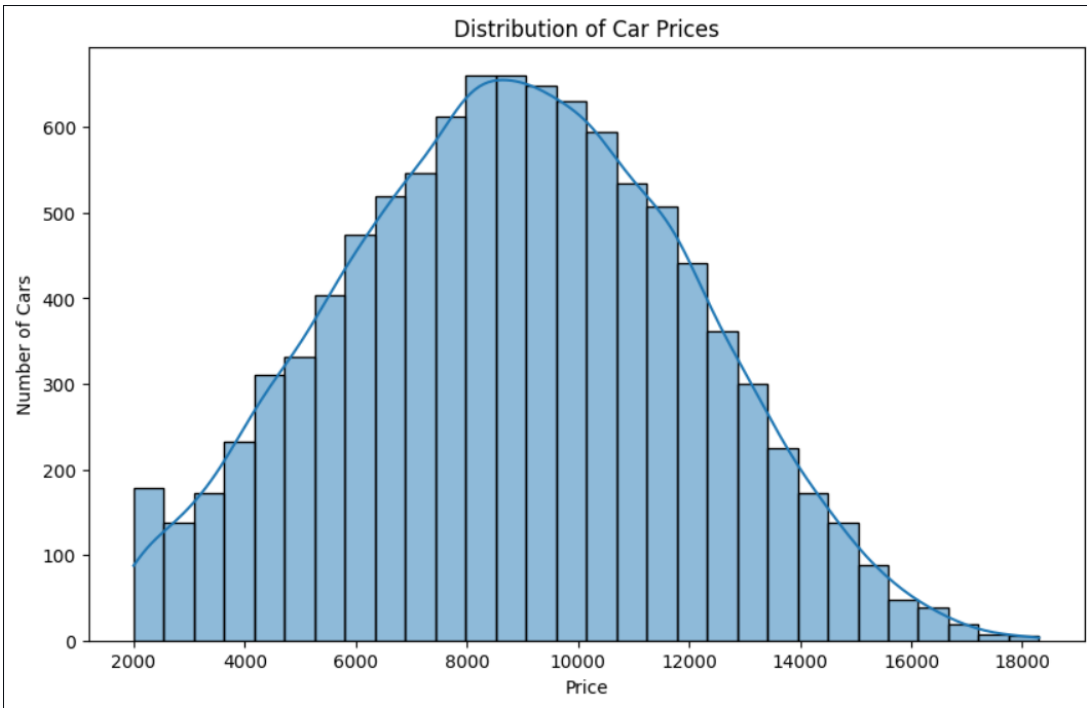

Figure 1: Distribution des prix des voitures

Relation entre l'Année et le Prix

La figure 2 visualise la relation entre l'année de fabrication et le prix des voitures.

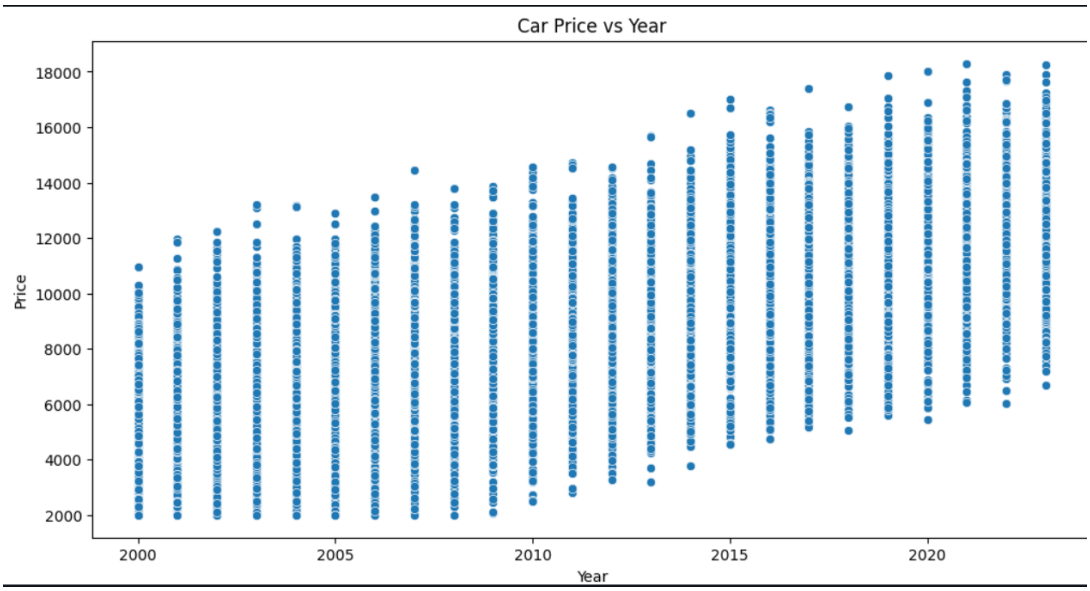

Figure 2: Relation entre le prix et l'année de fabrication

Analyse du Kilométrage

La figure 3 présente la distribution du kilométrage des voitures.

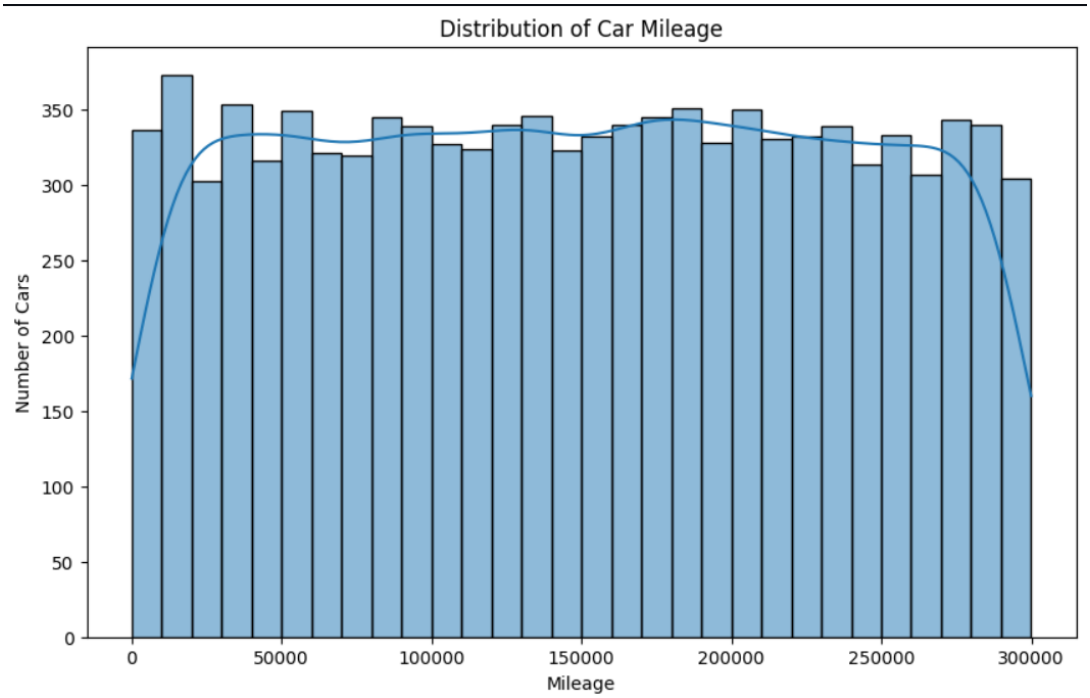

Figure 3: Distribution du kilométrage des voitures

Relation entre l'Année et le Kilométrage

La figure 4 visualise la relation entre l'année de fabrication et le kilométrage des voitures.

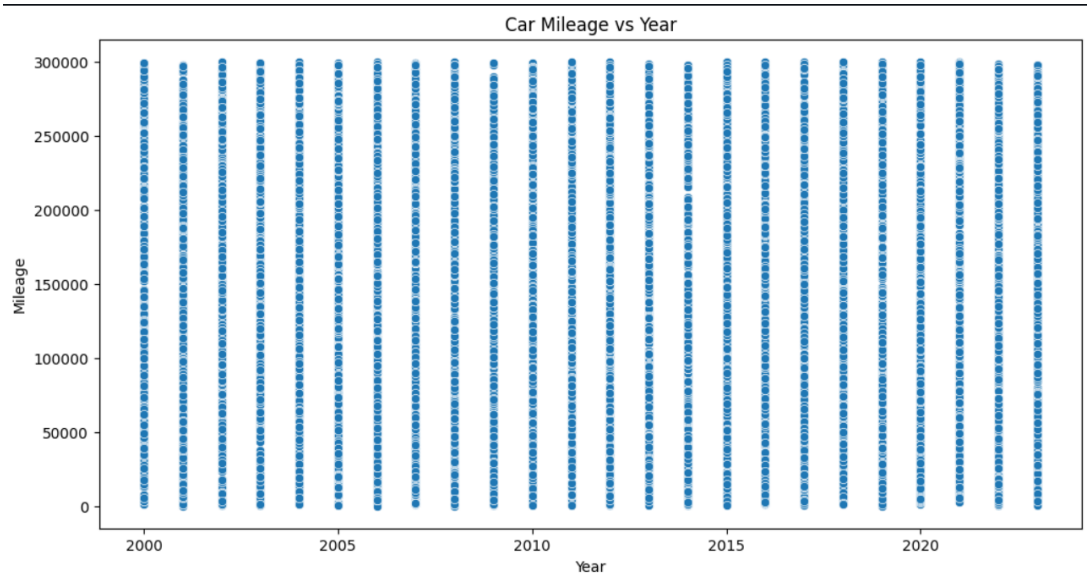

Figure 4: Relation entre le kilométrage et l'année de fabrication

Relation entre la Taille du Moteur et le Prix

La figure 5 montre que les voitures avec un moteur plus grand ont généralement un prix plus élevé, bien que d'autres facteurs influencent cette relation.

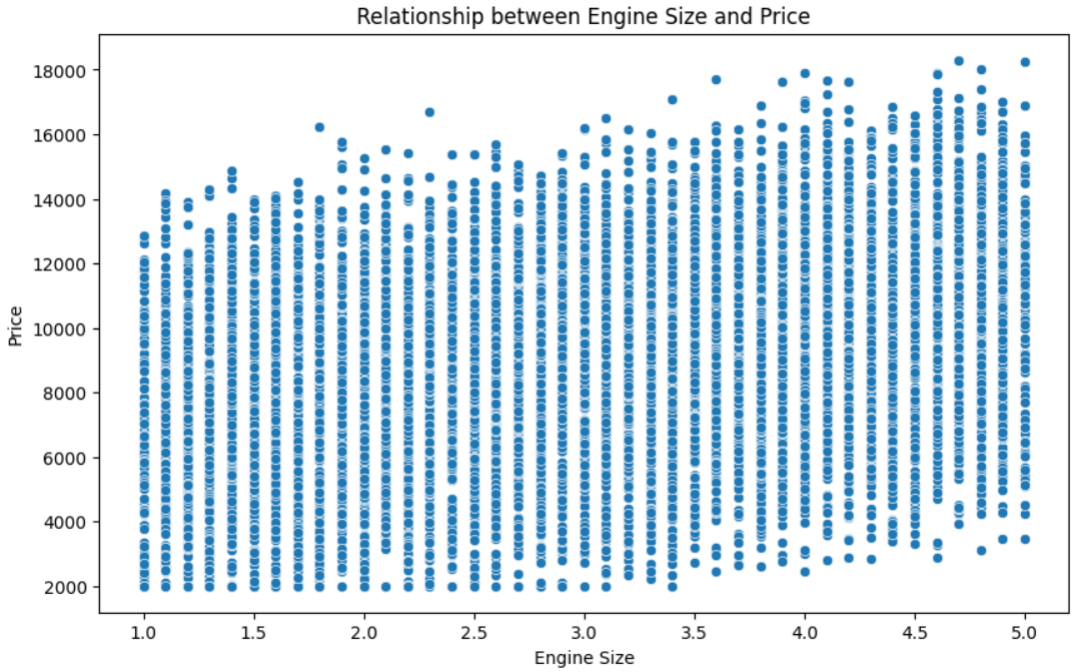

Figure 5: Relation entre la taille du moteur et le prix des voitures

Impact de la Marque sur le Prix

La figure 6 affiche les marques de voitures les plus fréquentes dans notre dataset.

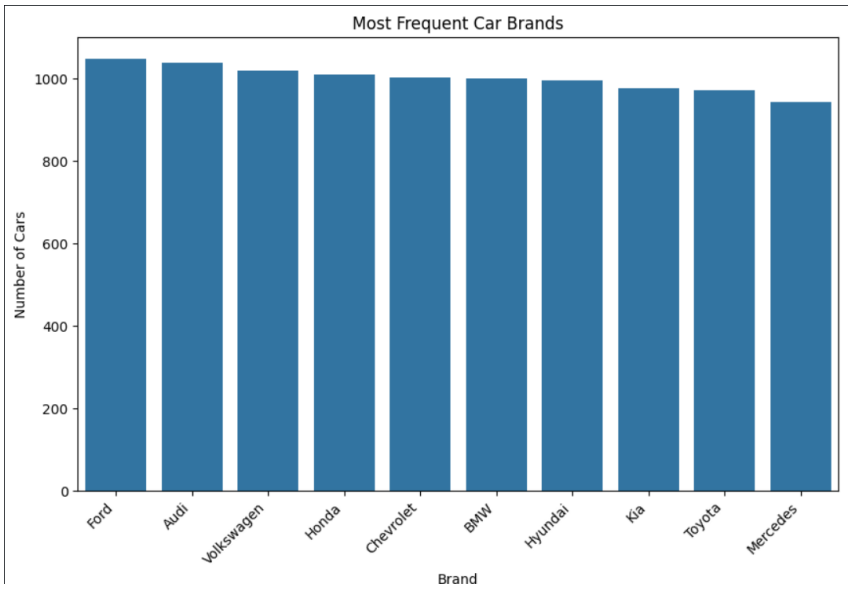

Figure 6: Marques de voitures les plus fréquentes

Prix Moyen par Marque

La figure 7 affiche le prix moyen des voitures par marque.

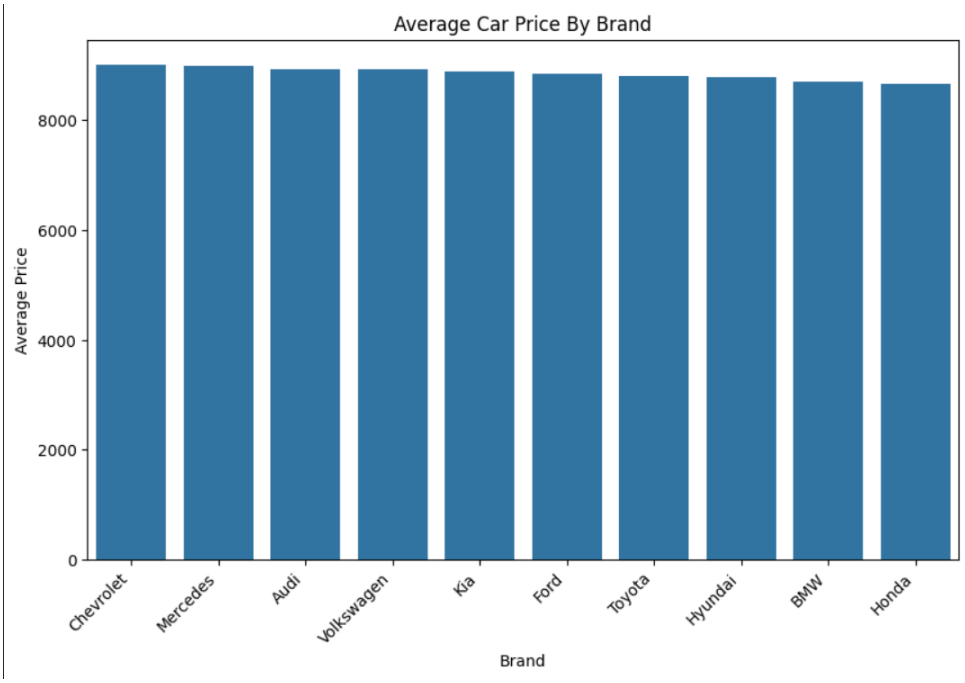

Figure 7: Prix moyen des voitures par marque

Distribution des Types de Carburant

La figure 8 affiche la distribution des types de carburant.

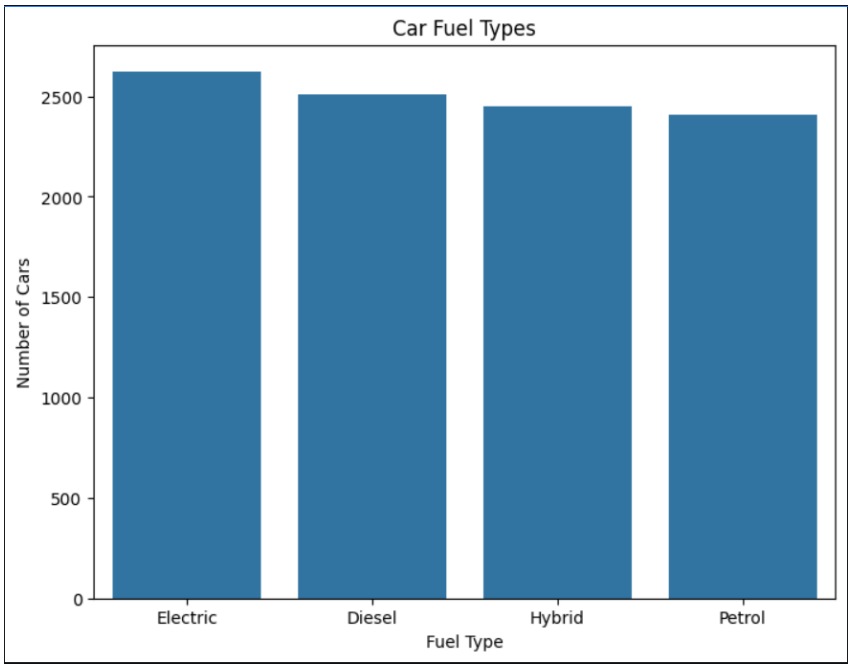

Figure 8: Distribution des types de carburant

Distribution des Types de Transmission

La figure 9 affiche la distribution des types de transmission.

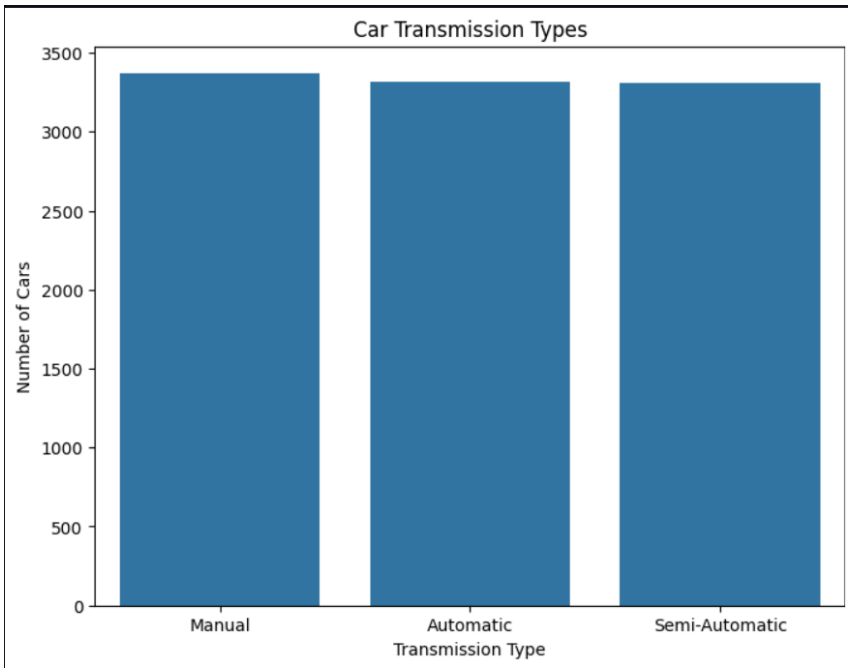

Figure 9: Distribution des types de transmission

Modélisation et Évaluation des Modèles

Modèles Utilisés

Nous avons exploré plusieurs modèles d'apprentissage automatique pour prédire le prix des voitures :

- **Régression Linéaire** : Un modèle de régression de base qui suppose une relation linéaire entre les variables d'entrée et la variable de sortie. Ce modèle est simple et rapide à entraîner, mais il peut ne pas capturer les relations non linéaires complexes. *Précision obtenue : 72%.*
- **Random Forest** : Un modèle d'ensemble qui combine les prédictions de plusieurs arbres de décision pour obtenir une prédiction plus robuste et précise. Random Forest est connu pour sa capacité à gérer des données non linéaires et à éviter le surapprentissage. *Précision obtenue : 87%.*
- **XGBoost (Extreme Gradient Boosting)** : Un algorithme de boosting puissant qui est souvent considéré comme l'un des meilleurs pour les tâches de régression et de classification. XGBoost est connu pour ses performances élevées et sa gestion efficace des données complexes. *Précision obtenue : 90%.*

Évaluation des Performances

Les modèles ont été évalués en utilisant un ensemble de données de test et des métriques de performance appropriées :

- **Régression Linéaire** : Précision de 72%.
- **Random Forest** : Précision de 87%.
- **XGBoost** : Précision de 90%.

Résultats et Interprétation

Analyse des Résultats

L'évaluation des modèles révèle que XGBoost est le modèle le plus performant pour prédire le prix des voitures dans notre dataset.

Importance des Variables

Les principales variables qui influencent le prix des voitures sont :

- **Année de fabrication** : Les véhicules plus récents ont tendance à avoir des prix plus élevés.
- **Kilométrage** : Un kilométrage élevé a généralement un impact négatif sur le prix.
- **Marque** : Certaines marques de voitures conservent mieux leur valeur que d'autres.
- **Type de carburant** : Les véhicules électriques et hybrides ont tendance à avoir des prix plus élevés en raison de leur technologie avancée et de leur aspect écologique.
- **Taille du moteur** : La cylindrée du moteur a un impact direct sur le prix, avec les moteurs plus grands étant généralement plus chers.

Interprétation des Coefficients

Une analyse plus approfondie a été menée sur les modèles pour interpréter les coefficients et évaluer l'impact de chaque variable sur le prix.

Conclusion et Perspectives

Ce rapport présente un modèle prédictif robuste pour estimer les prix des voitures. Les résultats obtenus soulignent l'importance des données dans le secteur automobile et ouvrent des perspectives pour l'optimisation des décisions d'achat et de vente. Plusieurs pistes d'amélioration peuvent être explorées dans de futures études :

- **Collecte de données supplémentaires** : Enrichir le dataset avec plus de variables et de données.
- **Optimisation des hyperparamètres** : Affiner les hyperparamètres des modèles pour maximiser leurs performances.
- **Intégration d'autres variables** : Ajouter des données sur la localisation géographique, les options du véhicule et d'autres détails spécifiques.
- **Comparaison avec des modèles avancés** : Évaluer l'utilisation de modèles de Deep Learning.
- **Analyse des erreurs** : Analyser les erreurs de prédiction pour mieux comprendre les limitations des modèles.