Prédiction de prix de voitures d'occasion

Écrit par : Alex Mozerski
Diego Fraile
Damian Spycher

Supervisé par : Dr. Laura Elena Raileanu
Elena Najdenovska
Cédric Campos Carvalho

Date de rendu : 16.06.2024

Résumé

Ce rapport présente les résultats de notre projet de prédiction des prix des voitures d'occasion, mené dans le cadre du cours de Web Mining à la HES-SO Master. Notre objectif principal était de développer un outil logiciel capable de fournir des estimations précises des prix basées sur des caractéristiques spécifiques des véhicules. Nous avons extrait les données nécessaires du site autoscoot24.com, couvrant une variété de paramètres tels que la marque, le modèle, le kilométrage, le type de carburant, entre autres. Nous avons utilisé des techniques avancées de machine learning, notamment des régressions forestières aléatoires et XGBoost, pour construire un modèle robuste et précis. L'outil intégré, avec une interface utilisateur développée en Dash, facilite la navigation des utilisateurs vers des annonces pertinentes, améliorant ainsi l'expérience d'achat de voitures d'occasion.

Introduction

Contexte

Dans le cadre du cours Web Mining de la HES-SO Master, notre équipe s'est penchée sur le développement d'un outil logiciel destiné à prédire les prix des voitures d'occasion en fonction de leurs caractéristiques. Ce projet trouve son origine dans le constat actuel que les consommateurs éprouvent souvent des difficultés à estimer la valeur réelle des véhicules d'occasion sur le marché.

But du projet

L'objectif principal de ce projet est de créer un outil logiciel capable de fournir une estimation précise du prix d'un véhicule d'occasion basée sur des caractéristiques spécifiques telles que la marque, le modèle, l'année de fabrication, le kilométrage, la couleur, etc. Cet outil vise à simplifier le processus de recherche pour les utilisateurs en leur permettant d'obtenir rapidement une estimation de prix sans nécessiter de connaissances techniques approfondies.

En outre, l'outil est conçu pour améliorer l'expérience utilisateur en intégrant une fonction de redirection automatique vers le site autoscoot24. Cette fonction permet à l'utilisateur de naviguer vers les annonces pertinentes sur autoscoot24 sans nécessiter la saisie répétée des informations du véhicule, facilitant ainsi une transition fluide entre l'estimation du prix et la consultation des annonces.

Limites du projet

Ce projet utilise des données extraites une seule fois du site autoscoot24 durant l'année 2024. Son obsolescence est à prévoir en raison de l'évolution rapide du marché des voitures d'occasion.

Données

Les données ont été extraites du site autoscoot24.com et contient des offres provenant de plusieurs pays différents. 15'428 offres de véhicules ont pu être récupérées.

Les différentes caractéristiques extraites sont détaillées ci-dessous :

- **Marque (brand)** : La marque du véhicule, qui est un indicateur important de sa qualité, de sa réputation sur le marché et potentiellement de son prix.
- **Modèle (model)** : Le modèle spécifique du véhicule au sein de la marque. Le modèle peut influencer considérablement le prix en fonction de ses caractéristiques et de sa popularité.
- **Prix (price)** : Le prix demandé pour le véhicule. C'est la variable cible pour le modèle de régression.
- **Kilométrage (mileage)** : Le kilométrage total affiché par le véhicule, qui est un indicateur de son usure et de son état général.
- **Type de carburant (fuel_type)** : Le type de carburant utilisé par le véhicule, tel que l'essence, le diesel, l'hybride, ou l'électrique. Ce paramètre peut affecter la demande et le prix du véhicule.

- **Couleur (color)** : La couleur extérieure du véhicule, un facteur qui peut influencer les préférences des acheteurs.
- **Boîte de vitesses (gearbox)** : Indique si le véhicule a une boîte manuelle ou automatique, un aspect crucial qui peut impacter le prix et la demande.
- **Puissance (power)** : La puissance du moteur en chevaux, qui peut être un facteur de différenciation important pour les acheteurs potentiels.
- **Cylindrée (engine Size)** : La taille du moteur, généralement en litres, qui influence la performance et l'efficacité du véhicule.
- **Vendeur (seller)** : Le type de vendeur, soit un particulier soit un professionnel, ce qui peut influencer la confiance des acheteurs et le prix.
- **Type de carrosserie (body_type)** : La configuration de la carrosserie, telle que berline, SUV, coupé, etc., un aspect qui définit le segment de marché du véhicule.
- **Portes (doors)** : Le nombre de portes du véhicule, qui peut affecter la praticité et l'attractivité pour certaines catégories d'acheteurs.
- **Sièges (seats)** : Le nombre de sièges disponibles, important pour les familles ou les utilisations commerciales.
- **Transmission (drivetrain)** : Le type de transmission (avant, arrière ou intégrale) du véhicule, influençant les performances et les préférences de conduite.
- **Classe d'émission (emission_class)** : Indique la classe d'émission du véhicule, un facteur de plus en plus important dans un contexte de réglementations environnementales strictes.
- **État (condition)** : L'état général du véhicule, évalué comme neuf, excellent, bon, ou juste. Ceci est crucial pour déterminer le prix.
- **Couleur de l'intérieur (upholstery_color)** : La couleur de l'intérieur du véhicule, qui peut jouer un rôle dans l'attrait esthétique global.
- **Année (year)** : L'année de fabrication du véhicule, qui influence directement son prix en raison de l'usure et des technologies embarquées.
- **Pays (country)** : Le pays où le véhicule est vendu, ce qui peut impacter le prix en raison de différentes fiscalités, de la demande, et des préférences régionales.

Etat de l'art

Frameworks de scraping

Dans le cadre de notre projet, nous avons évalué les frameworks de scraping en Python les plus populaires : Selenium, BeautifulSoup et Scrapy. Selenium offre une grande flexibilité mais peut être plus lent que les autres outils. BeautifulSoup est idéal pour des projets simples et de petite envergure mais devient moins performant à mesure que la complexité augmente. Scrapy, quant à lui, bien que nécessitant une configuration initiale plus élaborée, s'avère plus rapide et plus adapté aux projets de grande envergure [1, 2, 3].

Modèle de machine learning

Après l'extraction des données, notre objectif a été de développer un modèle de machine learning pour estimer les prix des voitures, exploitant des données à la fois catégoriques et numériques. Un article de 2021 publié par Springer a exploré divers modèles pour estimer les prix immobiliers en France, utilisant des variables

comparables à celles de notre projet. Les modèles testés incluaient SVR, KNN, Random Forest, MLP et régression linéaire, montrant de bons résultats [4].

Dans le domaine de l'automobile, il existe deux approches principales pour la prédiction des prix. La première utilise des modèles de classification pour catégoriser les prix dans différentes fourchettes, ce qui peut donner d'excellents résultats mais requiert souvent la combinaison de plusieurs modèles pour atteindre une précision élevée. La seconde approche utilise des modèles de régression, qui bien que moins précis en termes de catégorisation, peuvent offrir de bons résultats pour des estimations de prix précises, surtout utilisés dans des contextes où la comparaison directe de prix est essentielle [5, 6].

Conception

Choix des technologies et modèles

Compte tenu de la complexité modérée d'Autoscout24, nous avons choisi d'utiliser Scrapy comme framework de scraping pour sa performance et sa personnalisation étendue. Concernant le modèle de machine learning, nous avons choisi d'utiliser d'en tester deux : un random forest regressor et un XGboost car ils présentent l'avantage de ne pas avoir à normaliser les données et offraient de bonnes performances dans des cas d'utilisation similaires. Concernant l'interface utilisateur, nous avons choisi d'utiliser Dash, un framework Python destiné au développement rapide d'applications web. Dash est particulièrement adapté pour ce projet car il permet une intégration aisée avec les modèles de machine learning et offre une flexibilité considérable pour la création d'une interface intuitive et réactive.

Architecture

L'architecture du logiciel final est décrite ci-dessous. Elle consiste en une interface utilisateur développé en Dash permettant à l'utilisateur de saisir les informations du véhicule pour lequel il souhaite obtenir une estimation. Une fois les informations rentrées, l'interface utilisateur utilise le modèle de régression pour obtenir une estimation du prix du véhicule. Finalement, l'utilisateur peut se rendre sur autoscoot24.com sans avoir à ré-entrer les informations du véhicule.

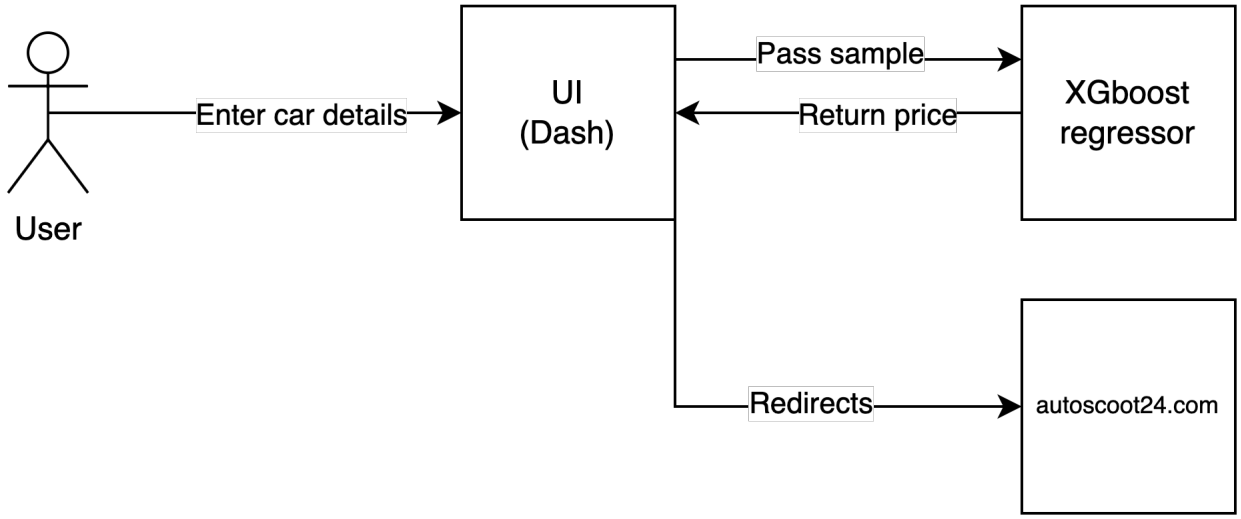


Figure 1 : Architecture du projet

Scraping des données et prétraitement

Comme autoscoot24.com ne permet que d'afficher les 20 premières pages de résultats par recherche de véhicule et que chaque recherche de véhicule affiche 20 offres, nous avons choisi de faire commencer le spider sur la page de recherche de chacune des 40 marques de voiture les plus populaires sans critère additionnel.

Le processus de scraping n'a pas toujours été précis en raison de sélecteurs CSS qui n'étaient pas suffisamment spécifiques, ce qui a entraîné la récupération de données parfois erronées ou incomplètes. Ce manque de précision a rendu indispensable un nettoyage approfondi des données pour garantir leur utilité dans l'analyse et la modélisation.

Traitement des valeurs nulles

Pour les champs catégoriques tels que la couleur, le type de carburant, et le type de boîte de vitesses, les valeurs nulles ont été remplacées par la valeur la plus fréquemment rencontrée pour chaque modèle spécifique de véhicule. Cette méthode préserve la cohérence des données en supposant que les caractéristiques les plus communes sont les plus probables pour un modèle donné. Si une valeur dominante n'était pas disponible pour un modèle spécifique, le remplacement a été effectué en fonction des valeurs dominantes de la marque concernée.

Pour les champs numériques comme le prix, le kilométrage, la puissance et la cylindrée du moteur, la stratégie adoptée a été d'utiliser la moyenne des valeurs disponibles pour le modèle en question, supposant ainsi une uniformité dans les caractéristiques de chaque modèle. En absence de données suffisantes pour un modèle particulier, la moyenne pour la marque a été utilisée. Si les valeurs pour une marque entière étaient insuffisantes ou absentes, la moyenne de l'ensemble de la colonne a été appliquée.

Traitement des valeurs aberrantes

Pour le prix des véhicules, il a été décidé de fixer un seuil maximal de 100’000 euros. Cette décision est fondée sur l'analyse préliminaire des données qui a montré que la grande majorité des véhicules listés sur autoscoot24.com étaient vendus à des prix inférieurs à ce montant. Les offres excédant ce seuil ont été considérées comme non représentatives du marché cible et susceptibles de biaiser le modèle de régression vers des valeurs anormalement élevées.

Analyse statistique

L'analyse statistique présentée a été réalisée dans le but de mieux comprendre les données ainsi que les facteurs les plus corrélés au prix d'un véhicule afin de pouvoir construire un modèle performant.

Corrélation entre les caractéristiques

Le diagramme ci-dessous est une matrice de dispersion (scatter plot matrix) qui illustre les relations entre différentes variables sur les véhicules. Chaque graphique dans la matrice montre la relation entre deux variables, permettant d'observer les tendances, les corrélations potentielles et les distributions des données.

Ces visualisations montrent principalement que des variables comme la puissance (power), la taille du moteur (engine_size) et l'année (year) ont une influence plus significative sur le prix des véhicules. Le kilométrage seul (mileage) semble avoir moins d'impact direct sur le prix, et les relations entre les autres variables sont plus complexes ou faibles.

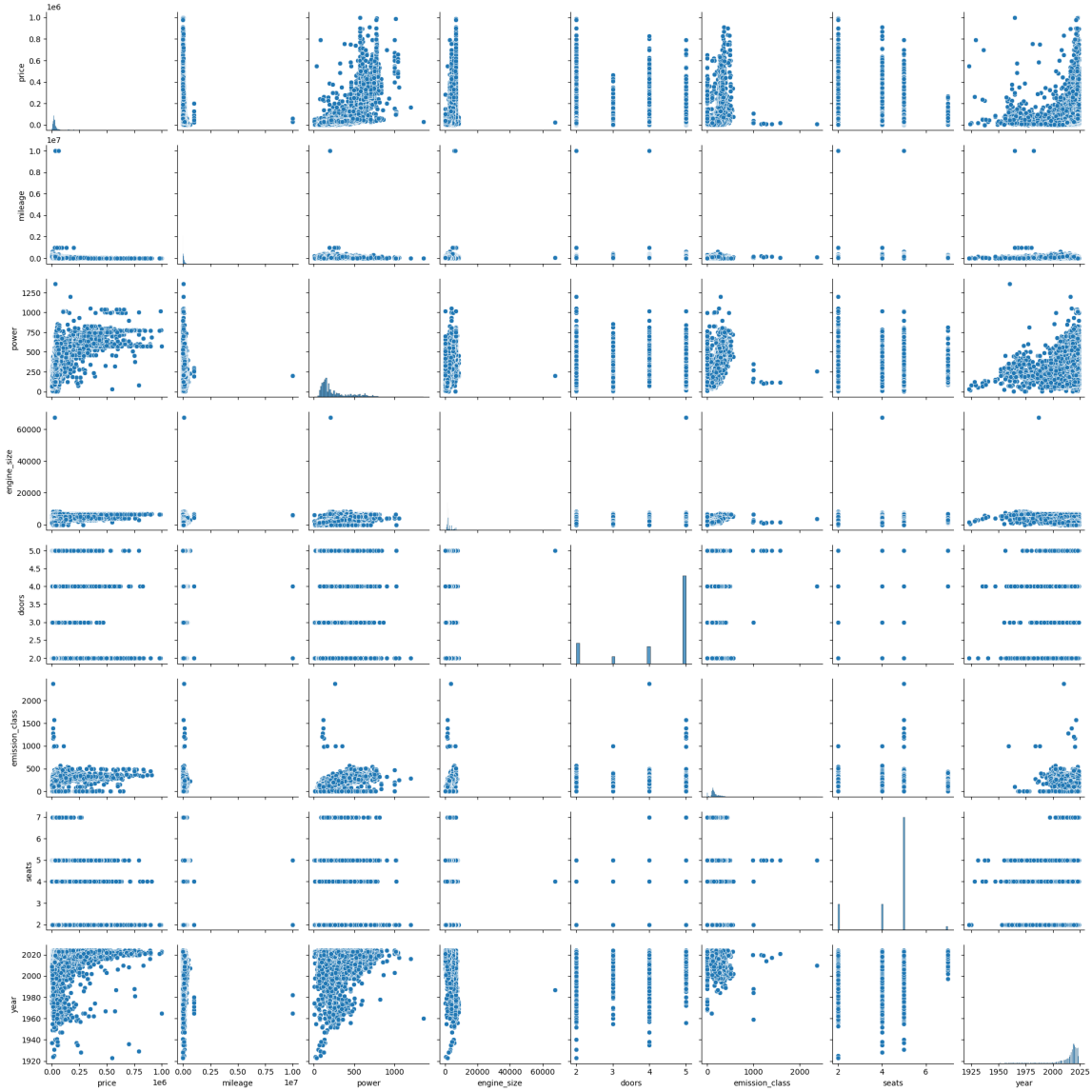


Figure 2 : Scatter plot matrix des caractéristiques numériques

Distribution des prix des voitures par pays

Les graphiques présentés montrent la distribution des prix des voitures dans différents pays européens, ainsi que la médiane des prix par pays.

Les distributions des prix des voitures varient considérablement d'un pays à l'autre. Le Luxembourg et la France se distinguent par des prix médians très élevés, ce qui suggère un marché orienté vers les voitures de luxe ou haut de gamme. À l'inverse, les Pays-Bas et l'Italie présentent des prix médians plus bas, indiquant une concentration sur des voitures plus accessibles. Les différences dans les distributions peuvent être attribuées à divers facteurs économiques, culturels et réglementaires propres à chaque pays.

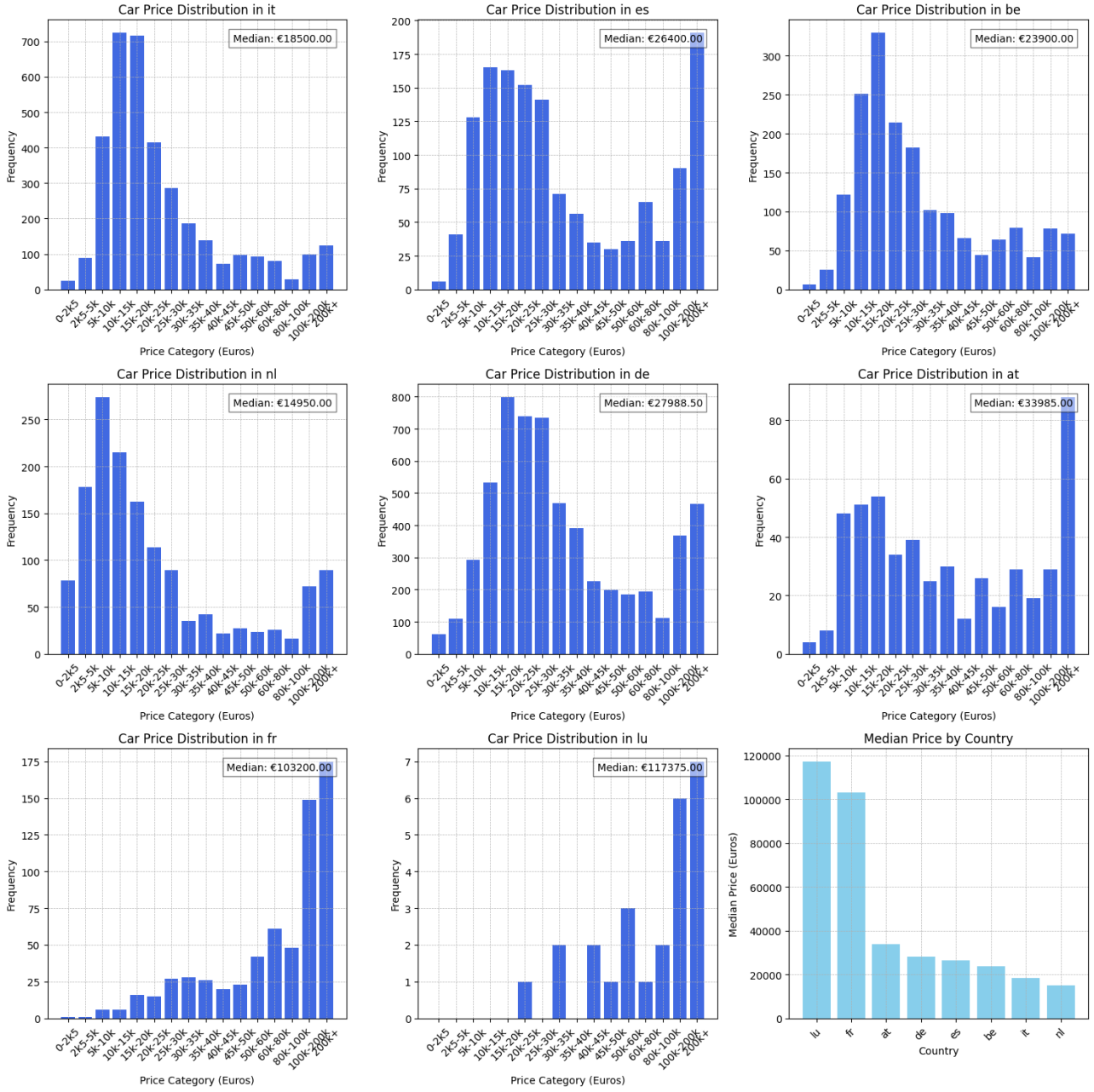


Figure 3 : Distribution des prix des voitures par pays

Analyse des prix par marque

Le graphique montre les prix médians des voitures pour les différentes marques. Sans surprise, les marques de voiture de sport ou de luxe présentent une médiane nettement plus élevée que la majorité restante.

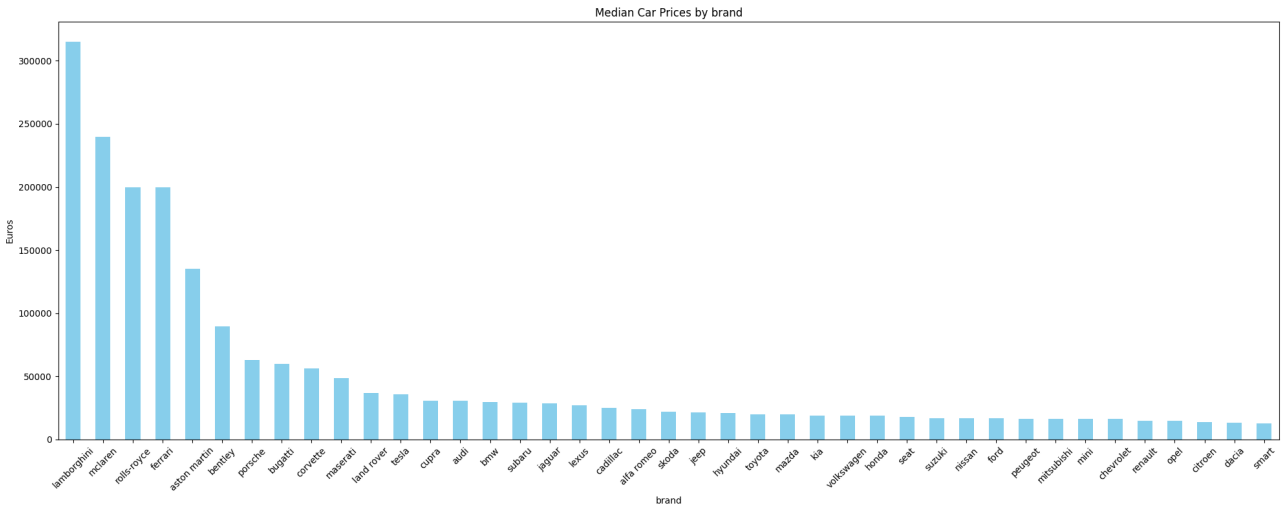


Figure 4 : Prix médian par marque

Analyse des prix par rapport à l'année du véhicule

Le graphique montre les prix médians des voitures en fonction de leur année d'homologation. Il montre également le volume de données disponible pour chaque période. Les voitures très anciennes (voitures de collection) présentent un prix médian nettement supérieur au reste mais beaucoup moins de données sont disponibles.

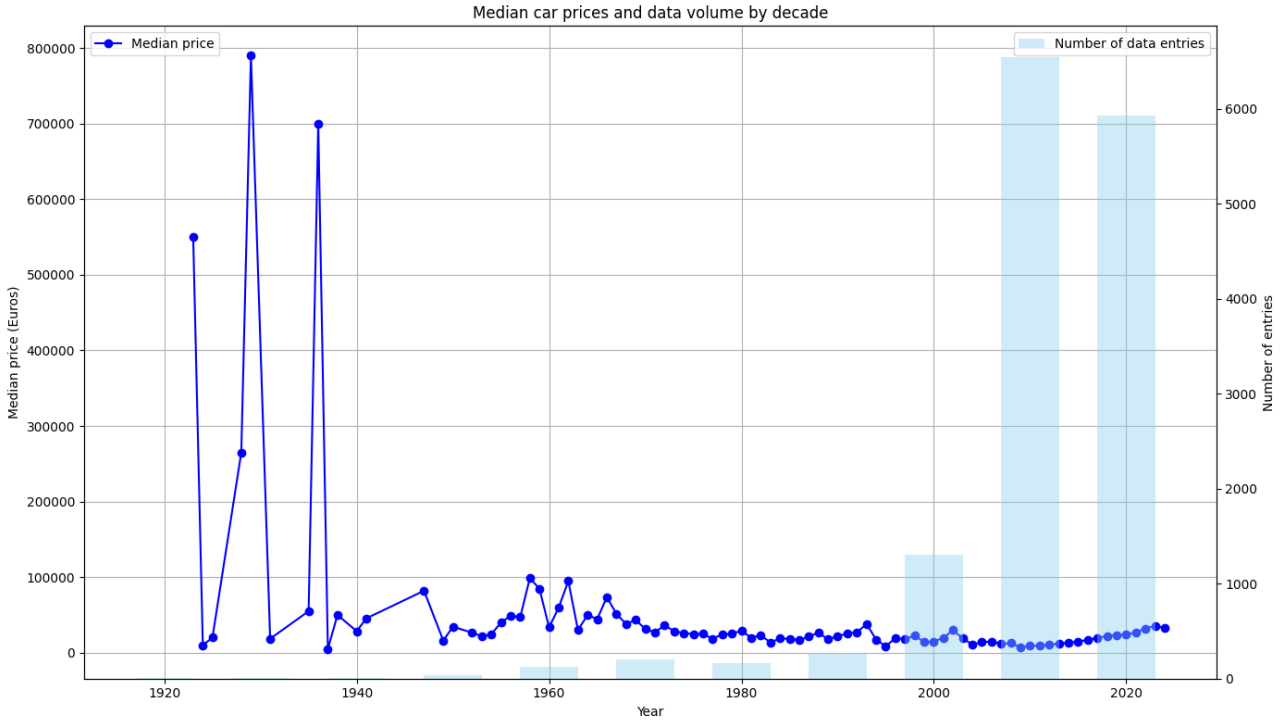


Figure 5 : Prix médian et volume de données par année d'homologation du véhicule

La même visualisation en ne se penchant que sur les véhicules homologués après l'an 2000 a été réalisée. Comme attendu, les véhicules récents sont plus chers. On peut également remarquer que le volume de données ne croît pas de la même façon que le prix. Les voitures sont surtout vendues après 4 ou 5 ans d'utilisation.

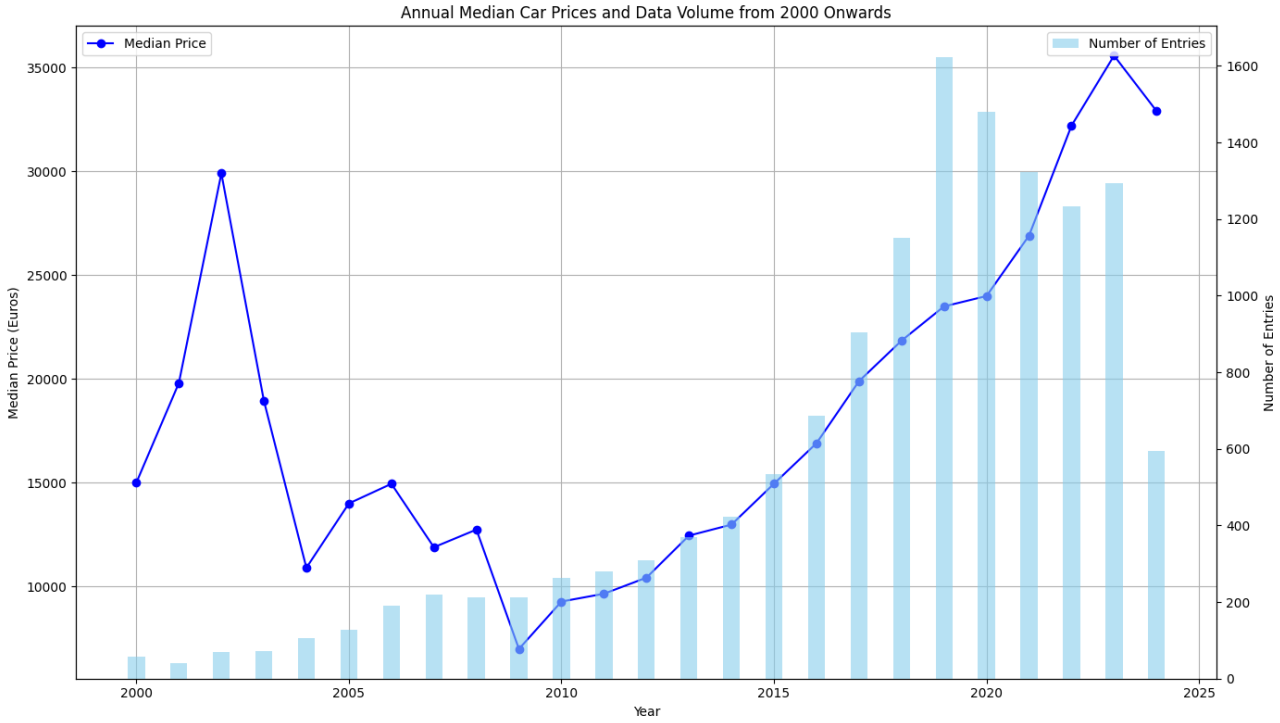


Figure 6 : Prix médian et volume de données par année, à partir de l'an 2000

Analyse des prix en fonction du kilométrage

Le graphique montre la relation entre le prix des voitures (en euros) et leur kilométrage, avec une ligne de tendance de degré 3 en rouge.

Cette visualisation met en évidence l'impact significatif du kilométrage sur le prix des voitures. Les voitures avec un faible kilométrage ont tendance à maintenir des prix plus élevés. Au fur et à mesure que le kilométrage augmente, la valeur des voitures diminue, mais cette diminution ralentit au-delà de certains seuils (environ 100,000 km).

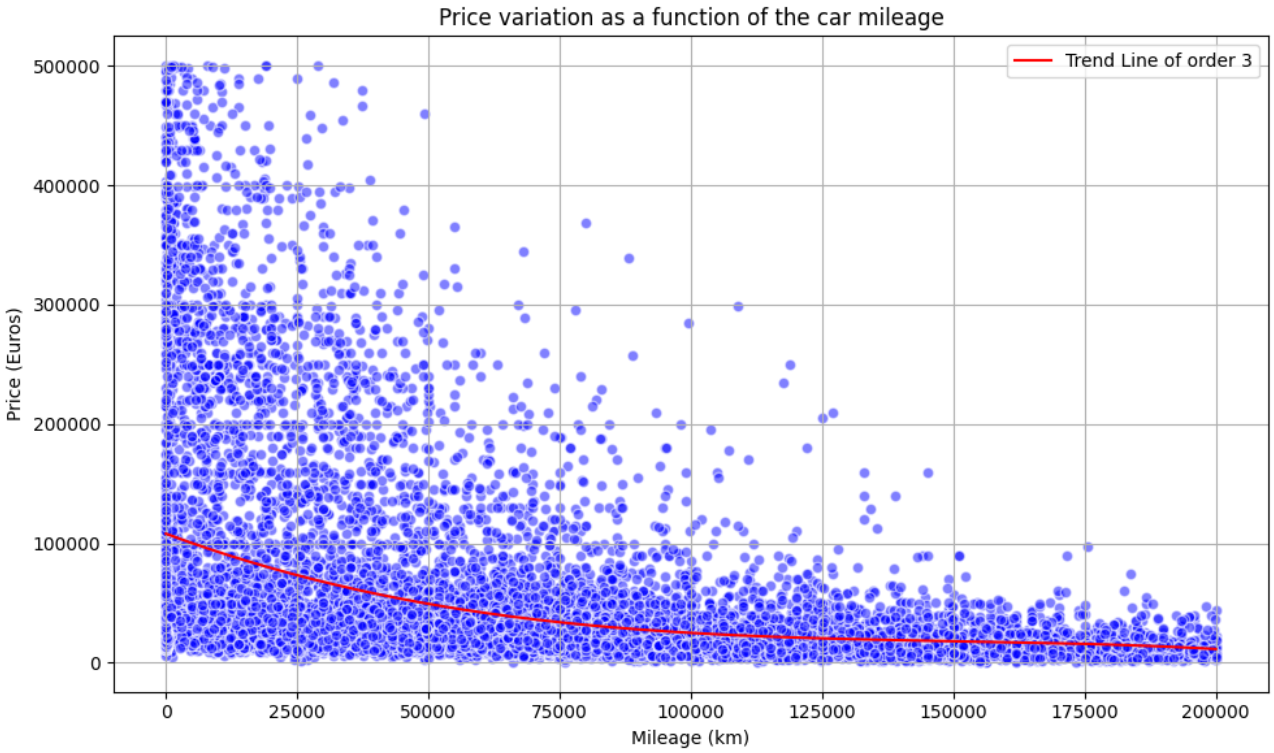


Figure 7 : Variation du prix en fonction du nombre de kilomètres du véhicule

Analyse des prix en fonction du type de carburant

Le graphique montre les prix moyens des voitures en fonction du type de carburant utilisé.

Les prix moyens des voitures varient considérablement en fonction du type de carburant utilisé. Les technologies plus récentes et alternatives comme l'éthanol, l'électricité et les hybrides sont plus coûteuses, tandis que les carburants traditionnels comme l'essence et le diesel sont plus abordables. Les carburants alternatifs comme le CNG et le LPG offrent des options encore plus économiques.

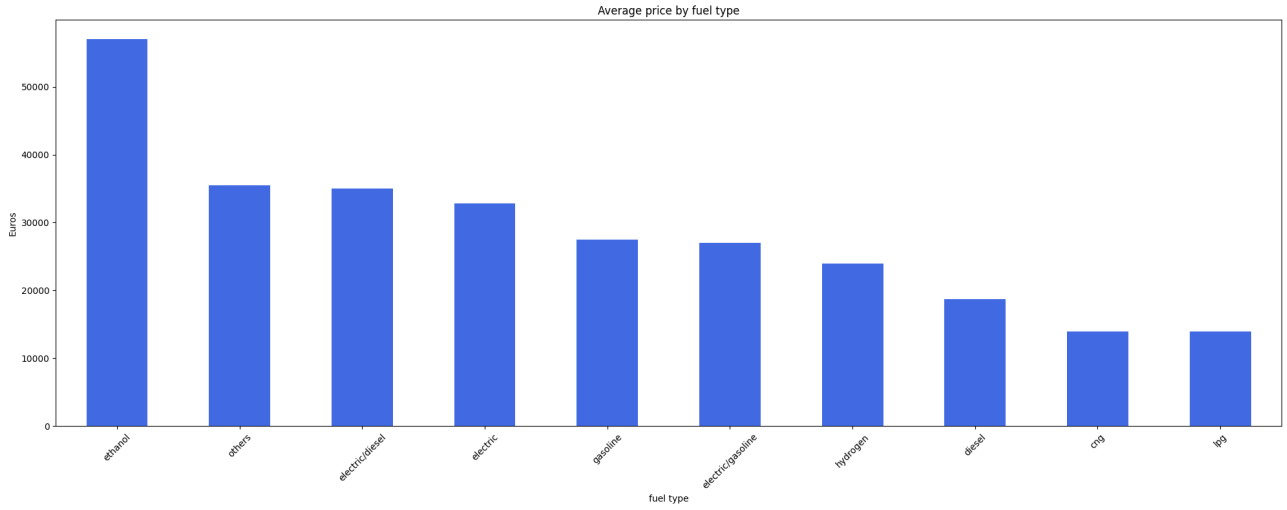


Figure 8 : Prix moyen du véhicule par type de carburant

Analyse des prix en fonction du type de carrosserie

Cette visualisation met en évidence des différences significatives dans les prix médians des voitures selon le type de carrosserie. Les coupés et les cabriolets, souvent associés aux voitures de sport et de luxe, ont les prix les plus élevés. Les types de carrosserie utilitaires comme les vans et les transporteurs ont des prix médians modérés, tandis que les voitures compactes sont les plus abordables.

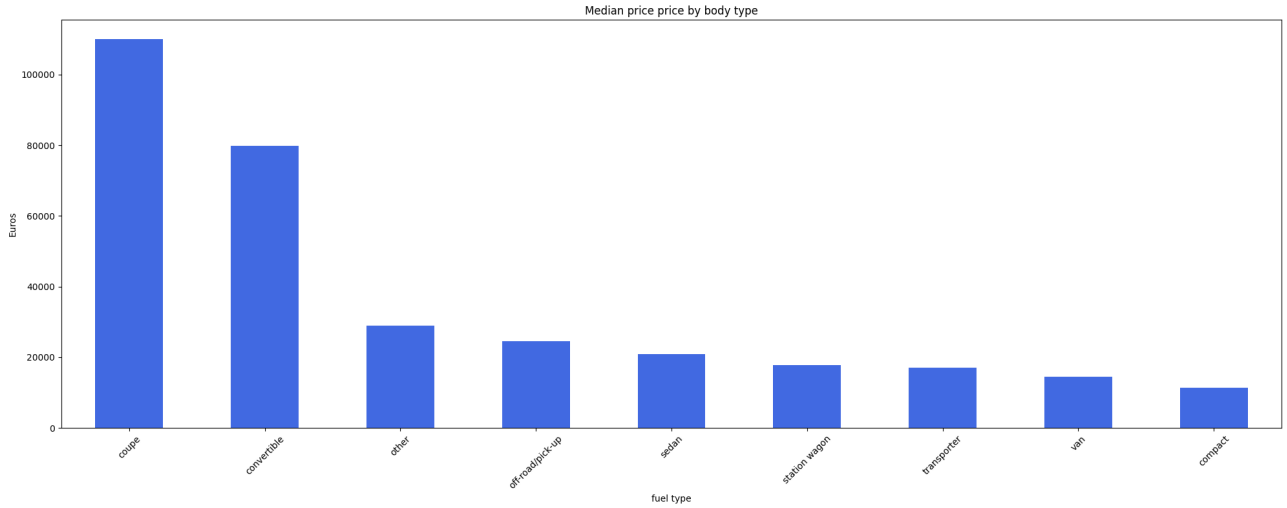


Figure 9 : Prix médian par type de carrosserie

Entraînement et évaluation du modèle

Prétraitement additionnel

Comme certains noms de modèles existent chez différentes marques, la marque et le modèle ont dû être concaténés pour un apprentissage efficace. Un encodage one-hot a dû être appliqué aux caractéristiques catégoriques des données. Finalement, les échantillons de données ont été randomisés.

Entraînement

Le dataset a été séparé en 3 sets distincts :

- entraînement (60%)
- validation (20%)
- test (20%)

Dans un premier temps, un random forest regressor a été testé pour obtenir une intuition quant à la faisabilité d'utiliser un tel modèle pour notre tâche. En obtenant des résultats concluants, nous avons essayé d'utiliser un algorithme XGboost. Ce dernier a fourni des résultats légèrement supérieurs au random forest regressor. Nous avons par conséquent choisi le dernier modèle.

La grille de hyperparamètres a été définie de la façon suivante :

n_estimators :

- **Description** : Nombre d'arbres dans le modèle.
- **Valeurs** : [50, 100, 200]

max_depth :

- **Description** : Profondeur maximale des arbres.
- **Valeurs** : [3, 6, 9]

learning_rate :

- **Description** : Taux d'apprentissage.
- **Valeurs** : [0.1, 0.2, 0.3]

colsample_bytree :

- **Description** : Fraction des colonnes à échantillonner pour chaque arbre.
- **Valeurs** : [0.6, 0.8, 1]

subsample :

- **Description** : Fraction des échantillons à utiliser pour la formation de chaque arbre.
- **Valeurs** : [0.6, 0.8, 1]

gamma :

- **Description** : Gain minimal requis pour effectuer une division supplémentaire sur un nœud de l'arbre.
- **Valeurs** : [0, 1, 5]

L’optimisation des hyperparamètres a été réalisée en utilisant une approche “grid search” avec de la 3-fold cross-validation sur le set d’entraînement. La meilleure configuration trouvée est : {'colsample_bytree': 0.6, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 9, 'n_estimators': 200, 'subsample': 0.6}

Évaluation du modèle

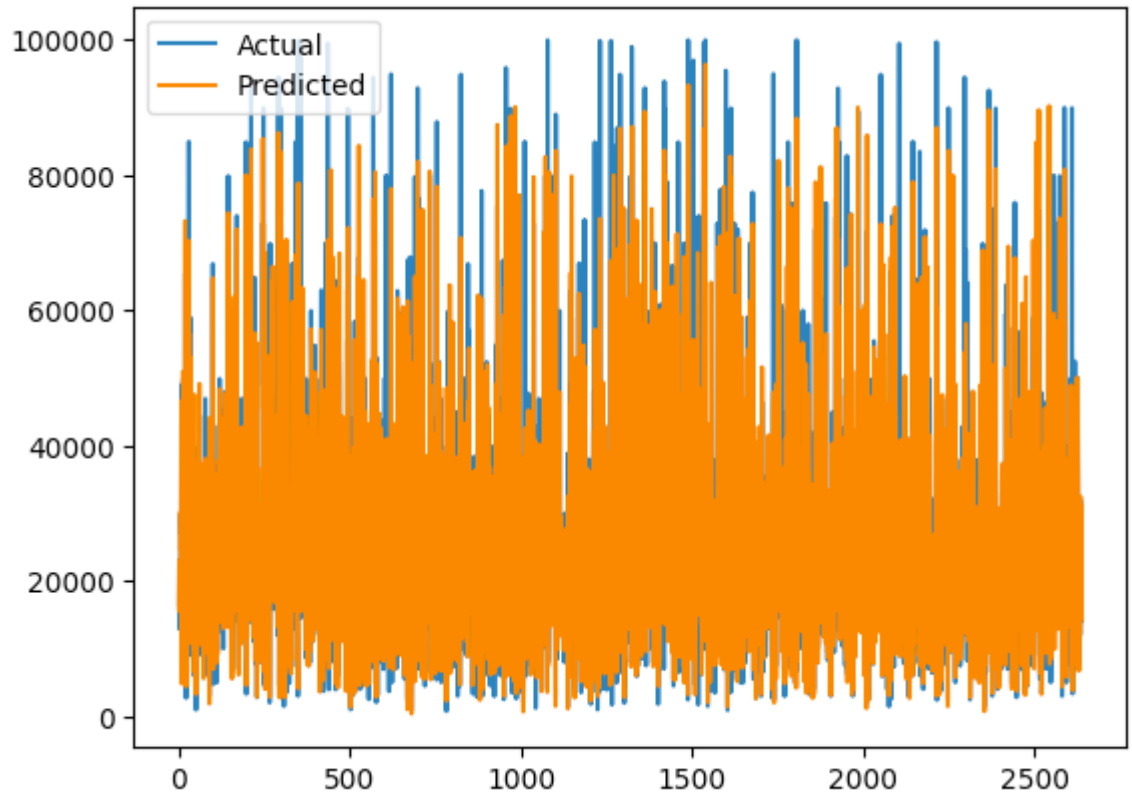
Les performances du modèle d'apprentissage automatique ont été évaluées en utilisant deux métriques principales sur le test set : la Root Mean Squared Error (RMSE) et le coefficient de détermination (R^2).

- **RMSE :** 6425.09
- **R^2 :** 0.8825

Les résultats montrent que le modèle a des performances solides :

- **Précision des prédictions** : Avec une RMSE de 6425.09, le modèle présente une erreur de prédiction moyenne acceptable pour de nombreuses applications.
- **Capacité explicative** : Un R^2 de 0.8825 indique que le modèle explique bien la majorité de la variance des données, montrant une bonne adéquation des caractéristiques utilisées et du modèle choisi.

Le graphique ci-dessous montre en orange les valeurs prédites superposées aux valeurs réelles en bleu.



Conclusion

Ce rapport a détaillé le développement d'un outil de prédiction de prix pour les voitures d'occasion, exploitant des techniques de scraping avancées et des modèles de machine learning sophistiqués comme le Random Forest et XGBoost. Notre système extrait et utilise des données variées, telles que la marque, le modèle, et le kilométrage, pour estimer le prix des véhicules avec une grande précision. L'interface utilisateur conçue facilite également l'accès aux annonces de voitures d'occasion, rendant notre outil pratique et accessible pour les utilisateurs finaux.

Concernant les perspectives futures, nous envisageons d'automatiser l'entraînement du modèle à intervalles réguliers avec les données les plus récentes disponibles sur le marché. Cette amélioration permettra de garder notre modèle à jour avec les fluctuations du marché et d'améliorer continuellement la précision de nos prédictions. L'automatisation de ces mises à jour garantira que notre outil reste pertinent et utile pour les utilisateurs, leur fournissant des estimations de prix fiables en temps réel.

Bibliographie :

[1] Scrapy and Selenium comparaison : https://oxylabs.io/blog/scrapy-vs-selenium

[2] Comparaison des trois framework de scraping Python les plus populaires : https://medium.com/analytics-vidhya/scrapy-vs-selenium-vs-beautiful-soup-for-web-scraping-24008b6c87b8

[3] Selenium Drawbacks for Web Scraping : https://scrapingrobot.com/blog/selenium-web-scraping/#why%20you%20should%20not%20use%20selenium%20to%20scrape%20data

[4] Estimation de prix des maisons en France : (Real estate price estimation in French cities using geocoding - Tchuente & Nyawa) https://link.springer.com/content/pdf/10.1007/s10479-021-03932-5.pdf

[5] Car Price Prediction using Machine Learning Techniques (Gegic & Isakovic) : https://temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf

[6] Old car price prediction with ML (Gajera & Gondaliya) : https://www.irjmets.com/uploadedfiles/paper/volume3/issue_3_march_2021/6681/1628083284.pdf

[7] Autoscout24.com (site ou les datas ont été scrapées) : https://www.autoscout24.com/

Annexes

Planning

										DEADLINE	
										RENDU	PRESENTATION
Tâches	Assigné à	26 avril 2024	3 mai 2024	10 mai 2024	17 mai 2024	24 mai 2024	31 mai 2024	7 juin 2024	14 juin 2024	16 juin 2024	21 juin 2024
Data Cleaning : Lecture des données + Traitement des valeurs manquantes + Création du Clean Dataset	Tout le monde										
Statistique descriptive : analyse de corrélation entre les features et les valeurs cibles	Alex										
Statistique descriptive : analyse de corrélation entre les features	Diego										
Statistique descriptive : analyse des trends (prix par pays, par marque, par carrosserie, par type de carburant, par année, par nombre de kilomètres)	Damian										
Sélection du modèle	Alex/Diego										
Pré-traitement des données : Préparation des datas pour les modèles (sélection des features, scaling, etc)	Alex/Diego										
Optimisation du modèle	Damian/Diego										
Evaluation du modèle	Damian/Diego										
Implémentation du backend	Alex										
Réalisation de l'interface graphique - frontend	Alex										
État de l'art: Recherche de document	Damian										
Mise en place application de base + correction des bugs	Tout le monde										
Réalisation du rapport technique	Tout le monde										
Préparation de la présentation	Tout le monde										