ENSAE Paris
MS Data Science
2024 – 2025

– Machine learning avec python –

Prédiction des prix de voitures

Auteurs

Ismaël DEMBELE
Anne-Laure TANOH

Superviseur académique

M. Roberto CASTELLINI

À rendre le 17 Janvier 2025

Contents

1 Introduction

Contexte et justification

La prédiction du prix d'une voiture est un domaine de recherche de grand intérêt, car elle nécessite un effort notable, une connaissance du domaine et un nombre considérable d'attributs distincts pour une prédiction fiable et précise (Gegic et al., 2019). La tâche est particulièrement critique et importante lorsque le véhicule est utilisé et ne sort pas directement de l'usine. D'une part, en raison de l'augmentation de la demande. En effet, en 2021, le marché des voitures d'occasion en France valait environ 9,96 milliards de dollars américains et la taille du marché devrait atteindre près de 12,74 milliards de dollars américains d'ici 2027. [1] Les ventes de véhicules d'occasion en France ont atteint un niveau record en 2021, et s'établissent à environ 5,2 millions de transactions en 2022. D'autre part, la valeur des voitures d'occasion dépend d'un certain nombre de facteurs (âge de la voiture, modèle, kilométrage, puissance, type de carburant, équipements, etc.), mais malheureusement, dans la pratique, la plupart des gens ne connaissent pas exactement toutes ces informations, de sorte que tous ces facteurs ne sont pas toujours disponibles et que l'acheteur doit prendre la décision d'acheter à un certain prix en se basant sur quelques facteurs seulement (Pudaruth, 2014). Les techniques de prédiction de l'apprentissage automatique peuvent être utiles à cet égard.

Problématique
L'objectif de ce projet est de prédire les prix des voitures d'occasion en France.

2 Méthodologie

2.1 Données

a. <u>Collecte des données</u>

La méthode utilisée pour collecter les données de voitures d'occasion issues du site **Leboncoin** est le *scraping web*, une technique permettant d'extraire automatiquement des informations d'un site web. L'objectif initial était ambitieux : scrapper les **845 000 annonces** de véhicules d'occasion disponibles sur la plateforme. Cependant, cette tâche nécessitait une stratégie bien structurée en raison des limitations rencontrées.

Méthodologie de scraping :

- **Ciblage des données :**
  - Les annonces ont été collectées via leurs principales caractéristiques : titre, description, prix, kilométrage, carburant, localisation (département, ville), et d'autres informations importantes pour la prédiction des prix.
  - Un script de scraping utilisant la bibliothèque **Scrapfly** a été implémenté pour automatiser les requêtes, gérer les protections anti-bot du site et optimiser les performances.
- **Organisation des requêtes :**
  - *Pagination limitée à 100 pages :* Nous avons rapidement constaté une limitation importante sur le site, où chaque recherche est limitée aux **100 premières pages**.
  - *Solution adoptée : Scraping par département :*
    - Leboncoin regroupe les annonces par département. Nous avons donc structuré nos requêtes pour cibler les **91 départements métropolitains** et les **DOM-TOM** (Guadeloupe, Martinique, Guyane, Réunion).
    - Cette approche permet de contourner la limite des 100 pages par recherche et de maximiser la couverture des annonces.
- **Hétérogénéité des résultats :**
  - Les données collectées ne sont pas uniformes en termes de temporalité des annonces. Par exemple :
    - Pour *Paris*, les 100 pages de résultats couvrent uniquement les annonces publiées dans les dernières **24 heures**, en raison de la forte densité des annonces dans cette région.
    - À l'inverse, pour des départements moins actifs comme la *Martinique*, les 100 pages contiennent toutes les annonces depuis la création du site.

[1] Statista : Taille du marché des véhicules d'occasion en France en 2021, avec des prévisions entre 2022 et 2027 (en milliards de dollars US)

  - Cette hétérogénéité est un défi à considérer lors de l'analyse, mais elle reflète également la dynamique réelle du marché des véhicules d'occasion.

- **Résultat final :**
  - Bien que les données ne soient pas totalement homogènes, cette approche a permis de constituer une base de données riche et représentative, couvrant plusieurs milliers d'annonces à travers la France. Ces données serviront à construire un modèle prédictif robuste.

b. Traitement et description des données

Le jeu de données d'origine contient **79 214 lignes** et **31 colonnes**. Ce résultat est le fruit d'une décomposition minutieuse des listes et des dictionnaires présents dans les fichiers JSON obtenus via le scraping. Chaque annonce, initialement encapsulée dans des structures imbriquées, a été transformée pour rendre ses informations exploitables dans un format tabulaire.

Parmi les variables conservées au départ, certaines sont particulièrement pertinentes pour la prédiction des prix :

- **prix** : Le prix du véhicule, notre variable cible.
- **kilometrage** : Indique le nombre de kilomètres parcourus, une caractéristique cruciale pour évaluer l'usure d'un véhicule.
- **marque et modele** : Données spécifiques au constructeur et au modèle, permettant de différencier les gammes de véhicules.
- **annee_modele** : Année de fabrication du véhicule, directement corrélée à sa dépréciation.
- **type_vehicule** : Catégorie du véhicule (SUV, Berline, etc.), influençant souvent le prix.
- **critair** : La classe Crit'Air, un indicateur de la pollution, jouant un rôle dans les réglementations et les prix en zones urbaines.

La liste détaillée des variables se trouve en annexe A.

Préparation des données :
Avant d'entamer l'analyse descriptive, nous avons réalisé un nettoyage et une préparation des données rigoureux. Voici quelques étapes clés :

- **Gestion des valeurs manquantes :** Les valeurs manquantes ont été soit imputées, soit supprimées si elles représentaient une fraction négligeable des données. Pour l'imputation, les variables numériques ont été imputées par la médiane, tandis que celles des variables catégoriques ont été remplacées par leur modalité la plus fréquente ou par la mention «non spécifié».
- **Traitement des outliers :** Les annonces avec des prix inférieures à 1 000 euros ou supérieurs à 500 000 euros ont été supprimés.

Ces étapes de préparation ont permis d'obtenir un jeu de données prêt pour l'analyse descriptive et la modélisation, garantissant une qualité et une cohérence maximales.

2.2 Construction du modèle de prédiction des prix

La construction d'un modèle prédictif robuste passe par plusieurs étapes essentielles, allant de l'exploration des données à l'optimisation des performances du modèle. Cette section détaille les différentes phases du processus méthodologique adopté.

a. Connaître ses données et comprendre les variables

L'analyse initiale des données est une étape clé pour poser les bases de la modélisation. Nous avons commencé par une *exploration descriptive* des variables disponibles dans notre jeu de données, qui compte 79 214 lignes et 31 colonnes. Parmi ces variables, certaines étaient directement exploitables, tandis que d'autres nécessitaient des transformations ou des encodages pour être intégrées au modèle.

- Les variables numériques telles que *kilométrage*, *année du modèle*, ou *puissance fiscale* ont été identifiées comme des indicateurs forts du prix.

- Les variables catégoriques telles que *marque*, *type de véhicule*, ou *carburant* apportent une dimension qualitative essentielle, mais nécessitent un encodage approprié.
- Des variables comme *log_prix* ont été introduites pour réduire l'impact des valeurs aberrantes sur la variable cible.

Cette étape a permis de dégager des hypothèses sur les relations entre les variables explicatives et le prix, et d'identifier les caractéristiques potentiellement importantes.

b. Exploration des données (EAD)

L'EAD a été réalisée pour détecter des anomalies, comprendre les distributions des variables et identifier les relations clés. Les principales étapes incluent :

- **Analyse des distributions :** Les distributions des variables numériques ont été examinées pour identifier les asymétries et les valeurs extrêmes. Par exemple, le *kilométrage* montre une forte concentration dans les faibles valeurs, mais avec une longue traîne pour les véhicules très usés.
- **Analyse des corrélations :** Une carte thermique des corrélations a permis de détecter les relations linéaires fortes, comme entre *log_prix* et *kilométrage*.
- **Visualisations croisées :** Des graphiques comme des boîtes à moustache ou des nuages de points ont été utilisés pour explorer l'effet des variables qualitatives sur le prix, comme la marque ou le type de carburant.
- **Création de nouvelles variables :** Par exemple, *prix_par_km*, un indicateur du coût par kilomètre, a été introduit pour capturer l'usure relative des véhicules. Cette variable a été utlisée seulement dans le cadre de l'exploration des données et a été écartée lors de la modélisation puisqu'elle est construite à partir du prix, qui est la variable à prédire.

c. Construction du modèle

L'étape finale consistait à construire et optimiser le modèle prédictif à l'aide d'algorithmes de régression. Plusieurs modèles ont été testés :

- **Régression linéaire :** Utilisée comme baseline pour comparer les performances des modèles plus complexes.
- **Arbre de décision, forêts aléatoires (RandomForest) et CatBoost :** Exploités pour leur capacité à capturer des relations non linéaires et à gérer des interactions complexes entre les variables.
- **LightGBM :** Sélectionné comme modèle principal grâce à ses performances élevées sur des jeux de données de grande taille et ses capacités à gérer des variables catégoriques.

3 Résultats

3.1 Exploration des données

Le jeu de données est composé de 78 313 lignes et 30 colonnes.

Statistic	Value
Moyenne	20 884
Écart-type	20 061
Min	1 000
25%	9 990.00
50% (Median)	16 990
75%	25 987
Max	489 990

Table 1: Description statistique des prix

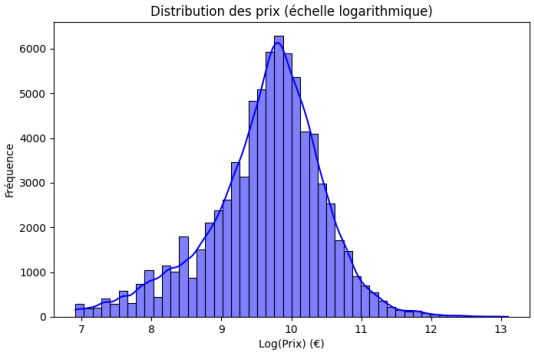

Figure 1: Distribution logarithmique du prix

Les prix vont de 1 000 à 489 990 euros, avec une moyenne de 20 884 euros. Les prix se situent principalement dans la fourchette 1 000 - 30 000 euros. La distribution est en forme de cloche, typique d'une distribution normale. Le pic principal se situe autour de log(10), ce qui correspond à environ 22 000 euros.

Nous étuidions par la suite des variables clés pour la prédiction du prix des voitures, selon la littérature (Pudaruth, 2014) et notre appréciation.

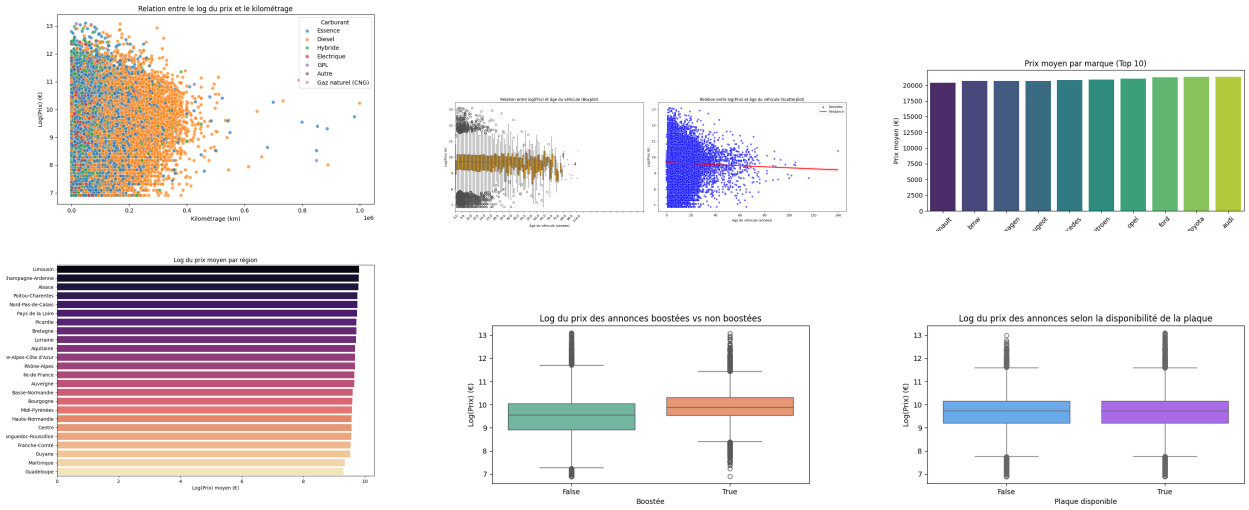

Figure 2: Distribution du prix selon des variables clé

- En général, plus le kilométrage est élevé, plus le prix (logarithmique) est bas (figure en haut à gauche). Toutefois, cette relation semble non linéaire et devient moins évidente lorsque le kilométrage est élevé. Le diesel (orange) domine largement les points de données, ce qui reflète sa popularité sur le marché des véhicules d'occasion. Les véhicules électriques et hybrides (vert et rose) semblent concentrés au niveau des kilométrages les plus faibles et les prix les plus élevés, ce qui indique des modèles récents ou haut de gamme.
- La boîte à moustache (figure du milieu en haut) montre une forte concentration de prix pour les véhicules de moins de 10 ans, avec une variabilité moindre. Les voitures plus anciennes ont tendance à avoir un logarithme de prix réduit, reflétant la dépréciation naturelle des voitures au fil du temps. Le nuage de points confirme la corrélation globale entre l'âge et le logarithme du prix observé, avec la ligne de tendance en rouge. La pente négative de cette ligne confirme que le prix diminue avec l'âge.
- Les marques ont des prix moyens relativement similaires (figure en haut à droite), ce qui indique une fourchette de prix homogène pour les véhicules les plus populaires. Les marques telles que BMW et Mercedes sont légèrement plus chères que les autres, ce qui reflète leur positionnement haut de gamme. Les prix moyens ne varient pas non plus beaucoup d'une région à l'autre (figure en bas à gauche).

- En ce qui concerne les caractéristiques des annonces, les prix des véhicules pour une annonce boostée sont, en moyenne, légèrement plus élevés (figure du milieu en bas). Toutefois, les prix ne varient pas si la plaque d'immatriculation est lisible dans l'annonce (figure en bas à droite).

En résumé, les caractéristiques des véhicules sont plus pertinentes pour prédire les prix que les caractéristiques des annonces. Les variables clés semblent être : le kilométrage, le type d'essence et l'âge du véhicule.

3.2 Résultats du modèle d'apprentissage

Dans cette section, nous présentons les résultats obtenus lors de la construction, de l'entraînement, et de l'évaluation du modèle d'apprentissage automatique visant à prédire le prix des véhicules d'occasion.

a. Comparaison initiale des modèles

Quatre algorithmes d'apprentissage automatique ont été testés avec leurs paramètres par défaut pour une première comparaison des performances. La métrique utilisée était l'erreur absolue moyenne (Mean Absolute Error, MAE). Les résultats obtenus sont résumés ci-dessous :

- **Régression linéaire :** MAE = 8,006.33 €
- **Arbre de décision :** MAE = 10,738.65 €
- **Forêt aléatoire :** MAE = 7,811.35 €
- **LightGBM :** MAE = 7,592.78 €

Le modèle LightGBM s'est révélé le plus performant avec un MAE de 7,592.78 €, et a donc été sélectionné pour une optimisation plus poussée des hyperparamètres.

b. Optimisation des hyperparamètres

Une recherche par grille (GridSearchCV) a été utilisée pour optimiser les hyperparamètres du modèle LightGBM. Les paramètres suivants ont été testés :

- Nombre d'estimateurs (`n_estimators`) : 50, 100, 200, 300
- Taux d'apprentissage (`learning_rate`) : 0.1, 0.05, 0.01
- Nombre de feuilles (`num_leaves`) : 20, 31, 50
- Profondeur maximale (`max_depth`) : 3, 5, 10, None
- Taux d'échantillonnage (`subsample`) : 0.7, 0.8, 1.0

Après optimisation, les meilleurs paramètres trouvés sont :

- `learning_rate` : 0.1
- `max_depth` : 10
- `n_estimators` : 200
- `num_leaves` : 31
- `subsample` : 0.8

Avec ces paramètres, le MAE sur l'ensemble de validation était de 7,587.48 €.

c. Réglage final et résultats sur les données de test

Le modèle optimisé a été réentraîné sur l'ensemble des données d'entraînement et évalué sur les données de test. Les métriques finales sont les suivantes :

- **Train :** RMSE = 15,679.24 €, MAE = 9,657.20 €
- **Test :** RMSE = 16,622.37 €, MAE = 10,214.00 €

d. Analyse des résultats et visualisations

- **Distribution des erreurs absolues :** La majorité des erreurs absolues sont inférieures à 20,000 €, comme illustré dans le graphique ci-dessous :

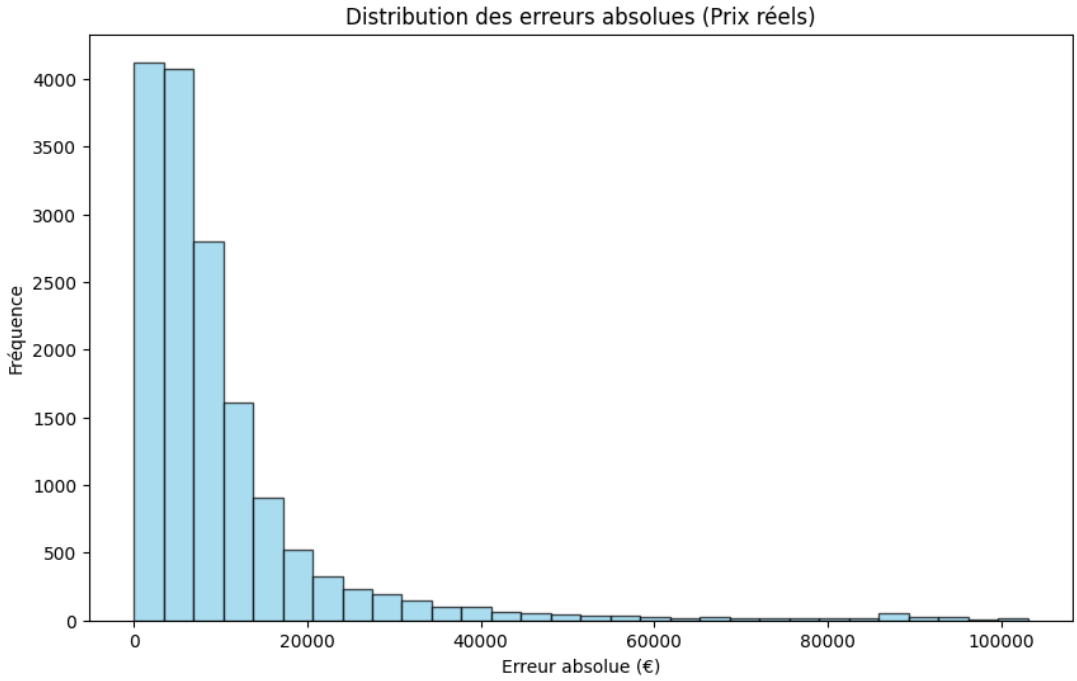

Figure 3: Distribution des erreurs absolues (Prix réels)

- **Comparaison des prédictions avec les valeurs réelles :** Le graphique suivant montre les prédictions par rapport aux valeurs réelles. Bien que les prédictions soient globalement alignées, il existe une certaine dispersion pour les prix élevés :

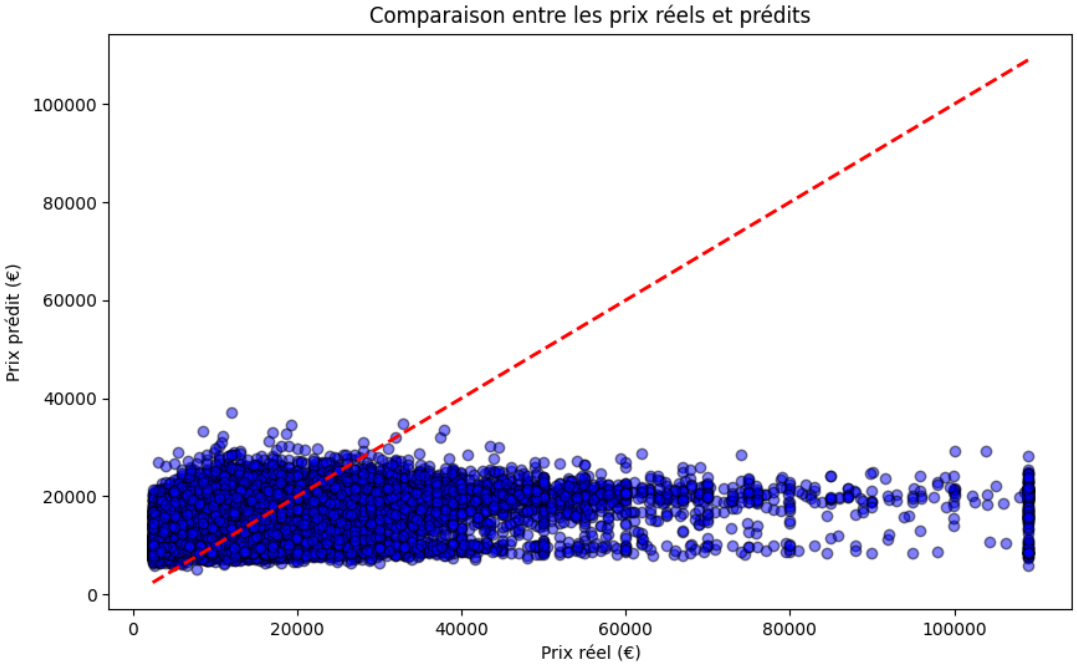

Figure 4: Comparaison entre les prix réels et prédits

- **Importance des variables :** Le graphique des 20 variables les plus importantes montre que les catégories de prix par kilomètre, les caractéristiques du modèle (année), et les informations sur le carburant et la boîte de vitesse jouent un rôle central dans la prédiction :

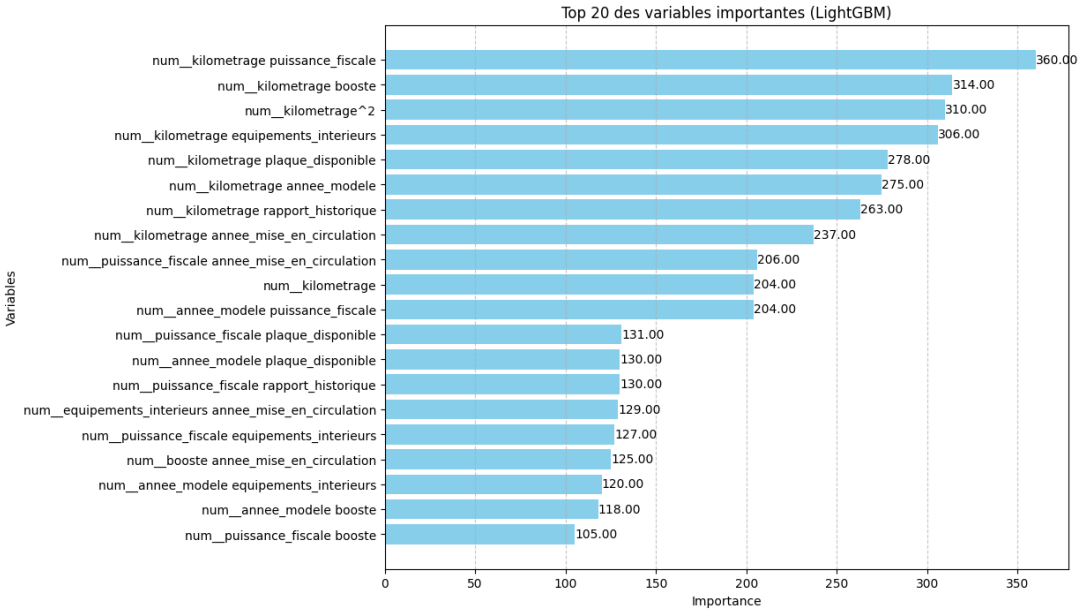

Figure 5: Top 20 des variables importantes (LightGBM)

Conclusion sur les performances : Malgré une amélioration significative grâce à l'optimisation des hyperparamètres et à l'analyse des erreurs, les performances du modèle montrent encore des marges d'erreur importantes, notamment pour les prix élevés. Des ajustements supplémentaires pourraient inclure des transformations spécifiques pour les données de ces segments ou une segmentation supplémentaire des données.

Étape 2: Amélioration du meilleur algorithme

Test 1: Transformation polynomiale de degré 2 des variables numériques pour capter les relations non linéaires (comme évoqué lors de l'exploration des données). Le RMSE est de 18 171, ce qui est une légère amélioration.

Test 2 : Au vue de la variabilité des prix, une classification est faite. Les prix sont prédits par classes.

Classe des prix (euros)	Fréquence (%)	RMSE
1 000 - 5 000	10,5	1 119
5 000 - 10 000	15,5	1 539
10 000 - 20 000	36,0	2 915
20 000 - 30 000	20,4	2 986
30 000 - 50 000	12,8	5 938
50 000+	4,8	47 138

Table 2: Répartition des classes de prix et erreurs de prédiction

En général les RMSE se rapprochent des écart-types (voir annexe B pour la descrption des classes). Les prédictions sont meilleures pour les prix 5 000 à 50 000 euros avec des RMSE de moins de 20% de la valeure moyenne. La prédiciton est moins bonne pour les classes les moins représentées (1 000 à 5 000 euros et 50 000 à 500 000 euros).
En agrégeant les classes, on obtient un RMSE (toutes classes confondues) de 10 764. On constate donc une nette amélioration du RMSE, même s'il représente encore 50% de la valeure moyenne (20 884 euros). Mais le RMSE est largement inférieure à l'écartype du prix (20 061) ce qui signifie que les prédictions sont plutôt correctes par rapport à la variabilité des prix.

4 Limites et perspectives

4.1 Limites

- Taille limitée de l'ensemble de données en raison de contraintes de collecte de données sur leboncoin.fr.
- Qualité de l'information : les sources de données sont des annonces sur un site web et les informations ont été écrites à la main par le vendeur et n'ont pas été directement observées.

- Variables non observées : le prix peut dépendre de la vente ou de la location de la voiture, mais il s'agit d'une variable non observée.

4.2 Perspectives

- Extraction de caractéristiques à partir de la description : peu d'annonces ont une description, mais certaines caractéristiques peuvent être extraites après le traitement du texte.
- Autres types d'analyse sur les annonces : détection des annonces frauduleuses, déterminants d'une vente.

A Annexes

Annexe A: Variables et description

date_publication: Date et heure de la publication de l'annonce
prix: Prix du véhicule en euros **booste**: Indique si l'annonce est mise en avant par le vendeur (booléen)
marque: Marque du véhicule (ex : BMW, Renault)
modele: Modèle spécifique du véhicule (ex : Renault Clio, Volkswagen Golf)
carburant: Type de carburant utilisé (ex : Essence, Diesel, Hybride)
kilometrage: Distance parcourue par le véhicule en kilomètres
annee_modele: Année du modèle du véhicule
boite_vitesse: Type de boîte de vitesses (ex : Manuelle, Automatique
portes: Nombre de portes du véhicule
places: Nombre de places dans le véhicule
controle_technique: Validité ou date de la fin du contrôle technique
date_mise_en_circulation: Date de la première mise en circulation du véhicule
etat_vehicule: État du véhicule (ex : Non endommagé, Endommagé, non spécifié)
type_vehicule: Catégorie du véhicule (ex : Berline, SUV)
sellerie: Type de sellerie intérieure (ex : Cuir, Tissu, non spécifié))
couleur: Couleur extérieure du véhicule
puissance_fiscale: Puissance fiscale théorique du véhicule en chevaux (CV)
puissance_din: Puissance fiscale (réelle/observée) du véhicule en chevaux (Ch)
rapport_historique: Indique si un rapport d'historique est disponible (booléen)
critair: Certificat de la qualité de l'air des véhicules (de 1 pour les véhicules hybrides rechargeables et les véhicules à essence Euro 5, 6; à 5 pour les véhicules conformes à la norme Euro 2, pour les véhicules diesel; non mentionnés)
equipements_interieurs: Indique si des équipements intérieurs sont listés (booléen).
region: Région où le véhicule est mis en vente.
departement: épartement où le véhicule est mis en vente.
ville: Ville où le véhicule est mis en vente
code_postal: Code postal de la ville associée à l'annonce
latitude: Latitude géographique de la localisation
longitude: Longitude géographique de la localisation
age_vehicle: Âge du véhicule en années, calculé à partir de l'année actuelle et de la date de mise en circulation
plaque_disponible: Prix du véhicule divisé par son kilométrage, indicateur du coût par kilomètre

Annexe B: Classes de prix et description

Classe des prix (euros)	Moyenne (euros)	Écart-type
1 000 - 5 000	3 133	1 153
5 000 - 10 000	7 780	1 483
10 000 - 20 000	15 342	2 859
20 000 - 30 000	24 883	2 879
30 000 - 50 000	38 186	5 659
50 000+	81 950	46 477

Table 3: Description des classes de prix

References

[1] PUDARUTH, Sameerchand. Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 2014, vol. 4, no 7, p. 753-764. Available online: https://www.academia.edu/download/54261672/2014_Predicting_the_Price_of_Used_Cars_using_Machine_Learning_Techniques.pdf.

[2] GEGIC, Enis, ISAKOVIC, Becir, KECO, Dino, et al. Car price prediction using machine learning techniques. TEM Journal, 2019, vol. 8, no 1, p. 113. Available online: https://temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf.