Université Paris Nanterre

Rapport de Projet

AutoPredict

Membres du projet :

X. Frédéric
R. Yann
R. Jérémy

26 avril 2025

Table des matières

Chapitre 1

Présentation de AutoPredict

1.1 Problématique

Le choix d'une voiture peut s'avérer complexe pour un acheteur, en particulier lorsque de nombreux critères entrent en jeu : budget, type de motorisation, consommation, puissance, style, marque, etc. Les plateformes existantes n'offrent pas toujours une expérience personnalisée ou intuitive pour explorer l'offre de véhicules selon ses préférences réelles. Par ailleurs, les vendeurs ou les analystes souhaitent mieux comprendre les facteurs qui influencent le prix d'un véhicule, et optimiser leur stratégie de vente.

1.2 Solution

AutoPredict est une plateforme web intelligente permettant d'exploiter une base de données automobile pour proposer deux fonctionnalités principales :

- **Recherche assistée de modèles** : à partir d'un budget donné et de certaines préférences (ex. type de transmission, puissance, taille...), l'utilisateur reçoit une liste de véhicules correspondant à ses besoins.
- **Estimation de prix** : à partir des caractéristiques sélectionnées (ex. année, style, consommation...), l'utilisateur obtient une estimation de la fourchette de prix des véhicules correspondants.

Ces fonctionnalités sont rendues possibles grâce à une architecture combinant analyse de données, machine learning, backend intelligent et une interface frontend intuitive.

1.3 Public cible

Notre projet AutoPredict s'adresse à deux types d'utilisateurs principaux, chacun ayant des besoins distincts que nous avons pris en compte dans la conception de notre moteur de recommandation et dans la structuration de notre analyse.

Les vendeurs, qu'il s'agisse de particuliers ou de petites structures de revente, cherchent à positionner correctement leurs modèles sur le marché. AutoPredict les aide à :

- fixer un prix de vente cohérent et compétitif en fonction des caractéristiques techniques du véhicule (année, style, puissance, segment de marché)
- anticiper la valeur de revente selon la dépréciation attendue, pour optimiser leurs marges
- identifier les catégories ou styles de véhicules les plus demandés selon les tendances observées

Acheteurs de véhicules d'occasion ou neufs

Les acheteurs utilisent AutoPredict comme outil d'aide à la décision pour affiner leur recherche selon leurs préférences et contraintes personnelles. Notre outil leur permet :

- de filtrer les véhicules selon des critères personnalisés (budget, consommation, motorisation, marque, popularité)
- de découvrir des modèles correspondant à leur profil d'usage (familial, urbain, plaisir, économique)
- d'évaluer si un modèle est bien positionné en termes de rapport qualité/prix, à partir des données du marché

Nous explorerons en détail les spécificités de chaque groupe cible, à travers nos personas.

1.4 Personas utilisateurs

Afin de mieux illustrer les usages concrets de notre plateforme AutoPredict, nous avons défini deux *personas* représentant les profils types de nos utilisateurs cibles : un vendeur de véhicule et un acheteur. Ces profils permettent de centrer notre réflexion sur leurs attentes spécifiques et d'adapter les fonctionnalités de notre outil à leurs besoins.

Persona 1 : Julien, vendeur de voitures d'occasion

- **Âge** : 42 ans

- **Profession** : Garagiste indépendant
- **Localisation** : Angers
- **Objectif** : Vendre ses véhicules au bon prix en s'alignant sur les tendances du marché
- **Comportement** :
  - Dispose d'un stock varié de véhicules d'occasion
  - Cherche à évaluer la valeur de chaque voiture en fonction de ses caractéristiques (année, style, puissance, consommation...)
- **Besoins couverts par AutoPredict** :
  - Obtenir une estimation juste du prix de vente d'un modèle spécifique
  - Identifier les caractéristiques qui influencent le plus la valeur résiduelle d'un véhicule
  - Visualiser les tendances de consommation, de prix ou de popularité par segment de marché

Persona 2 : Clara, acheteuse de voiture citadine

- **Âge** : 29 ans
- **Profession** : Infirmière
- **Localisation** : Montpellier
- **Objectif** : Trouver un véhicule fiable, économe et adapté à un usage urbain
- **Comportement** :
  - Recherche un véhicule d'occasion dans un budget de 12 000 à 15 000 euros
  - Privilégie la consommation, la taille et le confort plutôt que la puissance
  - Hésite entre plusieurs modèles et souhaite comparer les options selon ses critères
- **Besoins couverts par AutoPredict** :
  - Accéder à une sélection de modèles filtrés selon ses préférences
  - Comprendre la variation des prix en fonction des caractéristiques sélectionnées

Chapitre 2

Caractéristiques Principales

2.1 Interface Conviviale

L'interface utilisateur est conçue avec ReactJS pour offrir une navigation fluide et interactive. L'accent est mis sur l'ergonomie, la clarté des résultats et la facilité d'utilisation même pour des utilisateurs non-experts.

2.2 Sources Fiables

Les données utilisées proviennent de datasets publics sur les voitures, incluant des caractéristiques techniques (puissance, consommation, taille, style), économiques (prix, popularité), et temporelles (année de sortie).

Chapitre 3

Architecture métier

3.1 Frontend

Le frontend est développé en ReactJS. Il intègre des composants interactifs permettant à l'utilisateur de :

— Rechercher des modèles de voitures correspondant à ses critères et à son budget
— Estimer la valeur d'un véhicule en fonction de caractéristiques spécifiques
— Visualiser graphiquement les résultats obtenus (filtres, comparateurs, graphiques de prix, etc.)

L'interface dialogue avec le backend via une API REST.

3.2 Backend

Le backend est conçu en Python à l'aide du framework Flask. Il gère :

— L'accès à la base NoSQL contenant les données automobiles
— L'exécution des modèles de machine learning pour la recommandation de véhicules et l'estimation des prix
— L'interface avec le frontend via une API structurée

Les requêtes utilisateurs sont traitées dynamiquement pour retourner des résultats adaptés et rapides.

3.3 Base de données

La base de données utilisée est de type NoSQL, permettant une flexibilité dans la gestion des formats de données hétérogènes typiques du domaine automobile.

Chapitre 4

Architecture distribuée

4.1 Application Hosting

Le projet est conteneurisé afin de faciliter le déploiement, la scalabilité et la portabilité. Docker est utilisé pour packager les composants.

4.2 Database Hosting

La base NoSQL est hébergée dans un environnement compatible cloud. Elle stocke les jeux de données enrichis et traités, accessibles via API.

Chapitre 5

Pratiques de Collaboration et de DevOps

5.1 Project Management

Le projet est géré en équipe de trois membres : Frédéric, Yann et Jérémy. Le suivi des tâches se fait de manière collaborative autour d'outils de gestion agile.

5.2 Versionnement

L'ensemble du code source est versionné via Git, avec des dépôts organisés pour le frontend, le backend et les notebooks d'analyse/ML.

5.3 Intégration Continue et Déploiement Continu

Des pipelines CI/CD seront mis en place pour automatiser les tests, le linting, et le déploiement sur l'environnement de développement.

5.4 Maintenabilité du code

L'utilisation de conteneurs, de frameworks standards (Flask, React) et de pratiques de développement modulaire assure la maintenabilité du projet.

5.5 Qualité du code

Le code est documenté, typé et validé avec des outils de linting et des tests unitaires, notamment sur les scripts de preprocessing et les modèles ML.

Chapitre 6

Partie Data Analytique

6.1 Analyse de données

6.1.1 Préparation et nettoyage des données

La phase de préparation a consisté à rendre les données cohérentes, complètes et prêtes à être visualisées. Elle s'est déroulée comme suit :

- **Standardisation** : Uniformisation des noms de colonnes en minuscules avec des underscores pour assurer une manipulation fluide.
- **Suppression des doublons** : Élimination des entrées redondantes basées sur les identifiants véhicule/modèle.
- **Traitement des valeurs manquantes** :
  - Remplacement par des valeurs par défaut ou par la moyenne (ex. nombre de portes ou de cylindres).
  - Suppression des lignes avec des données critiques absentes.
- **Filtrage des transmissions inconnues** : Les entrées comportant 'UNKNOWN' pour la transmission ont été exclues de l'analyse.
- **Création de nouvelles variables** :
  - Quantiles de prix (MSRP) pour catégoriser les véhicules.
  - Plages de puissance moteur pour les regrouper en catégories ('faible', 'moyenne', 'élevée', etc.).
  - Calcul de la consommation moyenne combinée (ville + autoroute).

6.1.2 Présentation des graphiques et interprétation

Dans cette section, nous explorons différentes visualisations de données afin d'extraire des tendances structurelles sur notre flotte de véhicules. Chaque graphique est accompagné d'une interprétation opérationnelle, utile pour guider les recommandations clients.

1. Répartition des types de transmission selon les quantiles de prix

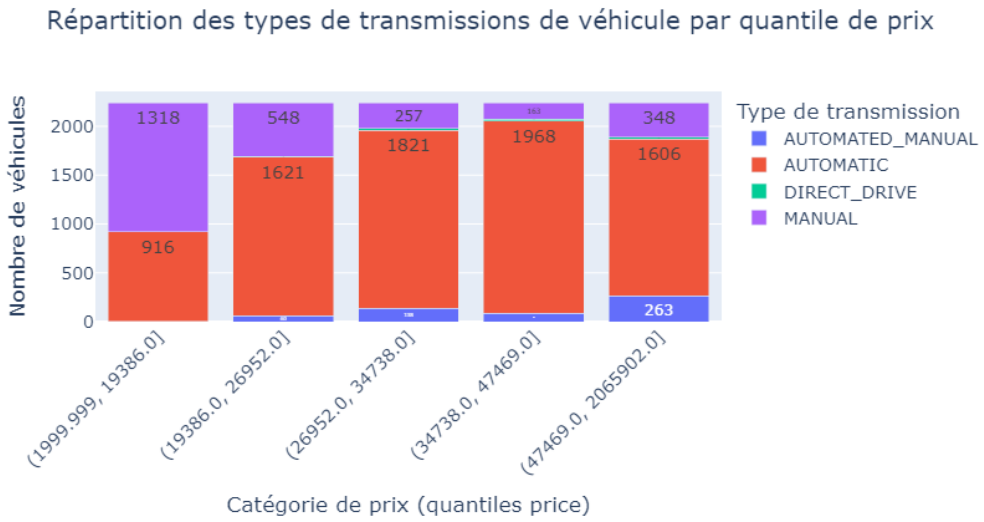

FIGURE 6.1 – Répartition des transmissions par tranche de prix (quantiles MSRP)

Ce graphique à barres empilées montre que :

— Les transmissions **automatiques** dominent très largement dans les gammes de prix *intermédiaires à élevées*, représentant parfois plus de **80 %** des véhicules.

— Les transmissions **manuelles** sont surtout présentes dans les véhicules du *premier quantile de prix*, et disparaissent progressivement à mesure que le prix augmente.

— Les transmissions **automatisées manuelles** sont rares, présentes essentiellement dans des véhicules haut de gamme ou spécifiques.

Cette distribution illustre un lien direct entre la gamme tarifaire et le type de confort/conduite attendu.

2. Répartition des styles de véhicules selon le type de transmission

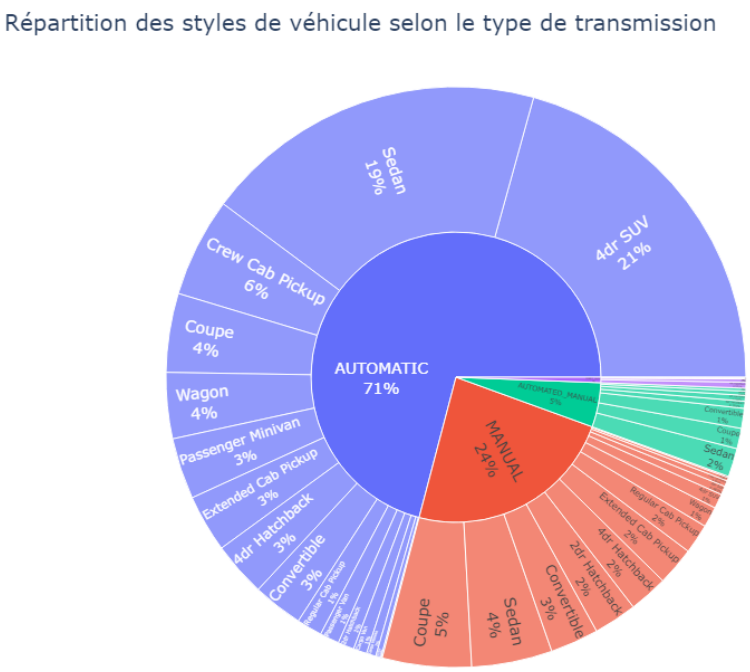

FIGURE 6.2 – Répartition des styles de véhicule selon la transmission

Le diagramme sunburst confirme les tendances précédentes :

— Les **SUV** et **berlines (sedan)** sont principalement associés aux transmissions automatiques, répondant à des besoins de confort et d'usage urbain/familial.

— Les **coupés**, **hatchbacks** ou **pickups** sont souvent en transmission manuelle, adaptés à des usages économiques, sportifs ou professionnels.

— Les transmissions rares comme `DIRECT_DRIVE` restent anecdotiques.

3. Consommation moyenne selon puissance moteur et cylindres

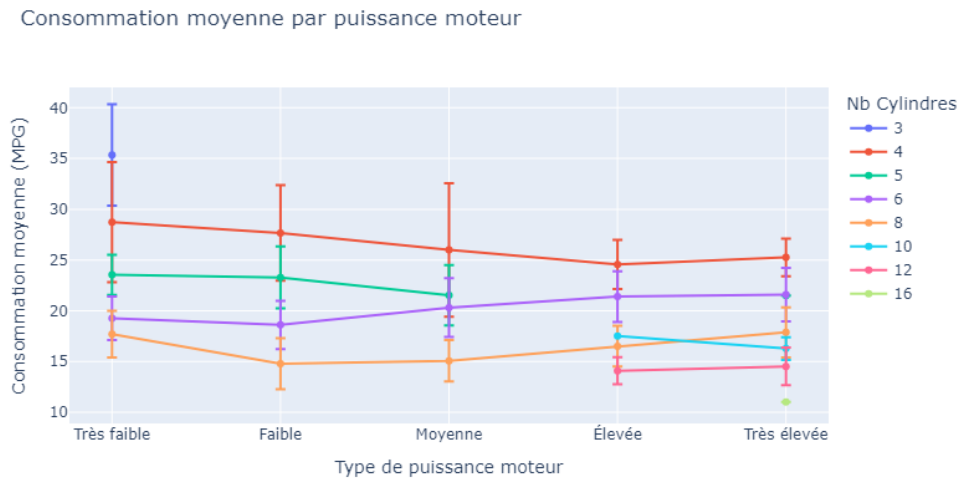

FIGURE 6.3 – Consommation moyenne selon puissance moteur et nombre de cylindres

On observe une relation logique entre la puissance du moteur et la consommation :

— Les véhicules à **puissance très élevée** et **8 cylindres ou plus** affichent une consommation moyenne plus élevée.

— Les modèles à **puissance moyenne à faible** ont des consommations plus stables et optimisées.

— Cette analyse permet de recommander les modèles selon un compromis performance/efficacité.

4. Prix moyen par catégorie de véhicule et puissance moteur

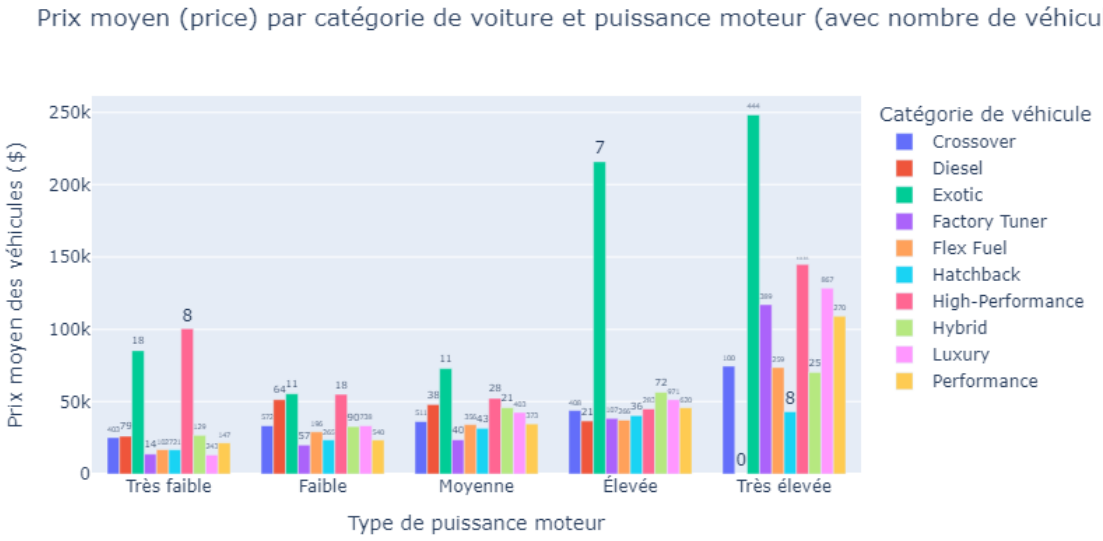

FIGURE 6.4 – Prix moyen par catégorie de voiture et puissance moteur (avec nombre de véhicules)

Cette visualisation en barres groupées révèle :

— Les véhicules **exotiques** dominent dans les plages de puissance élevée avec un prix moyen largement supérieur à 200 000$.

— La catégorie **haut de gamme (Luxury)** est répartie sur toutes les puissances, mais fortement concentrée sur les plages hautes.

— Les **véhicules hybrides, diesel, flex fuel** et **compactes (Hatchback)** occupent les plages basses à moyennes, avec des prix accessibles.

Ce graphique complète la compréhension des segments en croisant l'offre produit avec les puissances moteur disponibles.

5. Profil d'achat et recommandations commerciales

À partir de l'ensemble de ces analyses, plusieurs profils-types émergent :

— Un **client urbain familial**, à la recherche de confort et de fiabilité, sera orienté vers un **SUV automatique** de gamme intermédiaire.

— Un **client professionnel ou rural** peut viser un **pickup manuel**, robuste et économique.

— Un **jeune conducteur ou petit budget** sera conseillé vers un **coupé ou hatchback manuel**, situé dans les premiers quantiles de prix.

— Pour les **amateurs de performance ou de véhicules hybrides**, on oriente vers des modèles à forte puissance ou technologies spécifiques, tout en tenant compte du marché (exotic, performance, luxury).

L'approche analytique appliquée à nos données permet donc d'alimenter directement notre moteur de recommandation personnalisé.

6.1.3 Analyse de la diversification de marque et de modèle

Volume de modèles distincts par constructeur

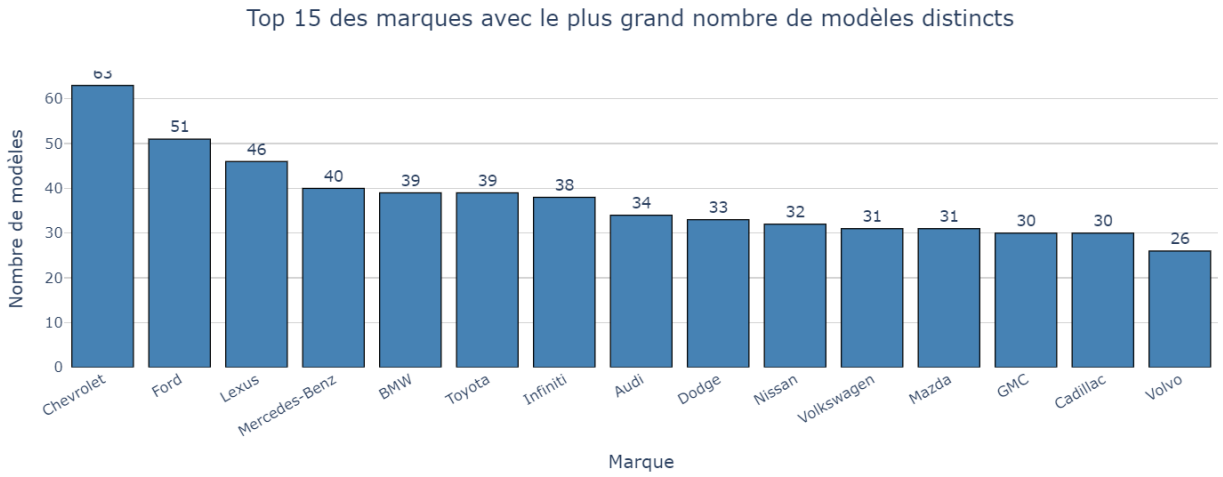

FIGURE 6.5 – Top 15 des marques avec le plus grand nombre de modèles distincts

Ce premier graphique dresse un panorama des marques les plus présentes sur le marché en termes de diversité de modèles proposés. On y observe que **Chevrolet** domine le classement avec 63 modèles, suivi de **Ford** avec 51 modèles. Ces deux constructeurs américains adoptent une stratégie résolument généraliste, en multipliant les déclinaisons afin de toucher une clientèle très large.

Derrière eux, on retrouve **Lexus**, **Mercedes-Benz** et **BMW**, dont les catalogues sont également étoffés mais plus orientés vers le haut de gamme. Cette diversité traduit une volonté de capter des niches plus spécifiques, en multipliant les combinaisons de finitions, de motorisations et de segments.

Ainsi, ce classement met en lumière deux stratégies différentes :

— une logique de volume (Chevrolet, Ford) pour occuper tous les segments du marché,

— une logique de spécialisation (Lexus, BMW, Mercedes-Benz) avec une forte présence dans le premium.

Répartition des segments pour les principales marques

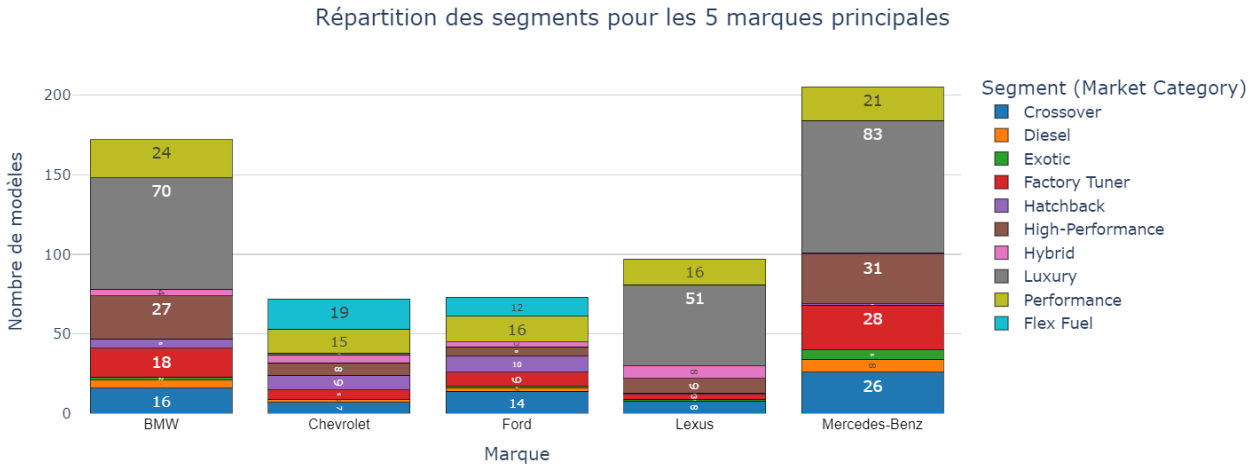

FIGURE 6.6 – Répartition des segments pour les 5 marques principales

Ce second graphique affine la lecture précédente en détaillant la composition des catalogues des cinq marques les plus prolifiques. Plusieurs enseignements clés en ressortent :

- **Mercedes-Benz** affiche une présence remarquable dans la majorité des segments, en particulier dans le *Luxury* (83 modèles), mais aussi dans le *Performance*, le *Crossover* ou encore le *Factory Tuner*. Ce spectre très large traduit une stratégie ambitieuse, capable de séduire aussi bien les amateurs de confort que de performance.
- **BMW** suit une trajectoire similaire, avec une prédominance du segment *Luxury* (70 modèles), mais aussi une forte présence dans les segments dynamiques comme le *High-Performance*.
- **Chevrolet** et **Ford**, bien que moins focalisés sur le haut de gamme, offrent une répartition beaucoup plus homogène entre les différents segments, ce qui illustre leur orientation généraliste. Ils se démarquent notamment par leur présence significative dans les catégories *Crossover*, *Flex Fuel* ou *Hatchback*.
- Enfin, **Lexus** adopte une position intermédiaire : fortement ancrée sur le segment *Luxury* (51 modèles), la marque japonaise complète son offre avec des incursions dans le *Hybrid*, le *Performance* et le *Crossover*, tout en conservant un positionnement qualitatif.

Diversité de segments et popularité moyenne des marques

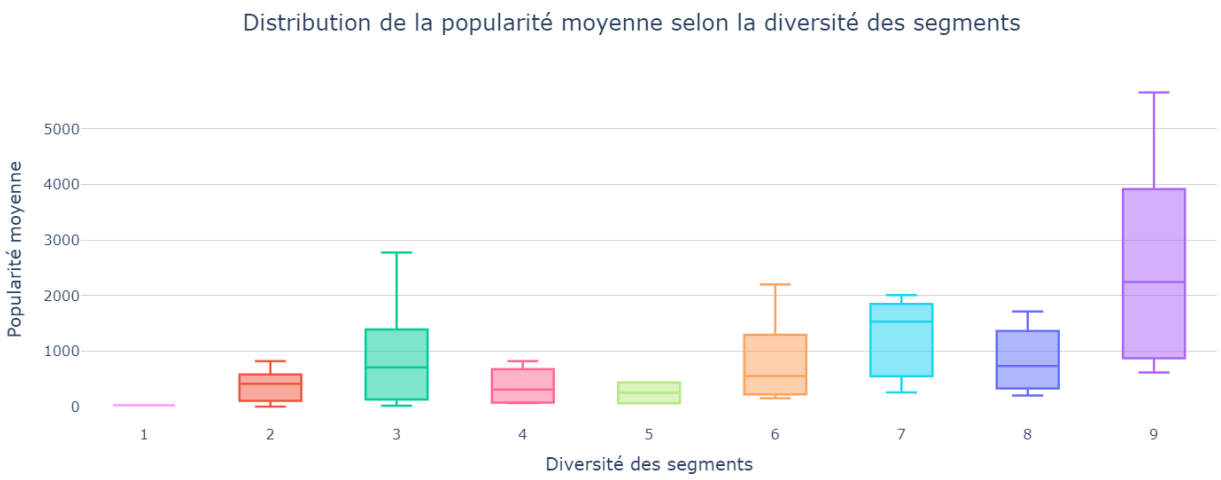

FIGURE 6.7 – Distribution de la popularité moyenne selon la diversité des segments

Ce dernier graphique aborde la question sous un autre angle : la diversité de l'offre est-elle un levier efficace pour accroître la notoriété ou l'attractivité d'une marque ?

On remarque que les marques présentes dans peu de segments (entre 1 et 3) affichent généralement une popularité moyenne faible, avec peu de dispersion. Cela traduit une spécialisation qui, bien que cohérente, reste limitée en termes d'audience.

À partir de 6 segments couverts, on constate une augmentation significative de la popularité moyenne, mais aussi une forte dispersion. Certaines marques diversifiées explosent en popularité, tandis que d'autres, malgré une offre étendue, peinent à s'imposer. Le message est donc nuancé : la diversification est souvent une condition favorable à la visibilité, mais ne garantit pas à elle seule le succès commercial.

Enseignements pour le moteur AutoPredict

Ces différentes visualisations apportent des éléments précieux pour l'algorithme de recommandation :

— Une marque très diversifiée comme Mercedes-Benz peut convenir à des profils très variés. Elle est donc intéressante à proposer dans des contextes où les préférences de l'utilisateur ne sont pas encore clairement établies.

— À l'inverse, une marque spécialisée dans le haut de gamme ou les performances (comme BMW ou Lexus) pourra être suggérée à des utilisateurs plus ciblés, ayant exprimé des attentes spécifiques.

— Enfin, la diversité de segments est un bon indicateur de flexibilité commerciale : intégrer cette variable dans AutoPredict permettra d'anticiper le champ des possibles qu'une marque peut offrir à un utilisateur donné.

Ainsi, cette analyse souligne l'importance d'adapter les suggestions non seulement en fonction des caractéristiques techniques des modèles, mais aussi selon la stratégie commerciale

des constructeurs.

6.1.4 Analyse de la dépréciation

Dépréciation moyenne des véhicules par année

FIGURE 6.8 – Dépréciation moyenne des véhicules par année

Ce graphique représente l'évolution du prix moyen des véhicules en fonction de leur année de fabrication. Il permet d'observer le phénomène de dépréciation sur le long terme, en regroupant aussi bien des véhicules neufs que d'occasion.

Analyse détaillée :

- Les modèles antérieurs à l'an 2000 présentent une valeur résiduelle très faible, majoritairement inférieure à 10 000 dollars. Cette tendance s'explique par une accumulation de kilomètres, une usure plus marquée et une inadéquation technologique avec les normes actuelles.
- Une baisse brutale du prix moyen est visible entre les années 2000 et 2002. Cette rupture peut coïncider avec des politiques environnementales ou fiscales incitatives au renouvellement du parc automobile.
- À partir de 2005, on observe une courbe ascendante continue du prix moyen des véhicules, atteignant un pic autour de 2014. Cela reflète probablement une montée en gamme du marché (meilleur équipement, technologies embarquées, électrification partielle...).

Conclusion : La valeur d'un véhicule chute rapidement durant les 10 à 15 premières années après sa sortie d'usine. Passé ce cap, la décote se stabilise à des niveaux très bas. Ces observations confirment qu'il peut être peu rentable, économiquement, d'acheter un véhicule trop ancien — en particulier en l'absence de fiabilité ou de revalorisation possible.

Dépréciation par style de véhicule

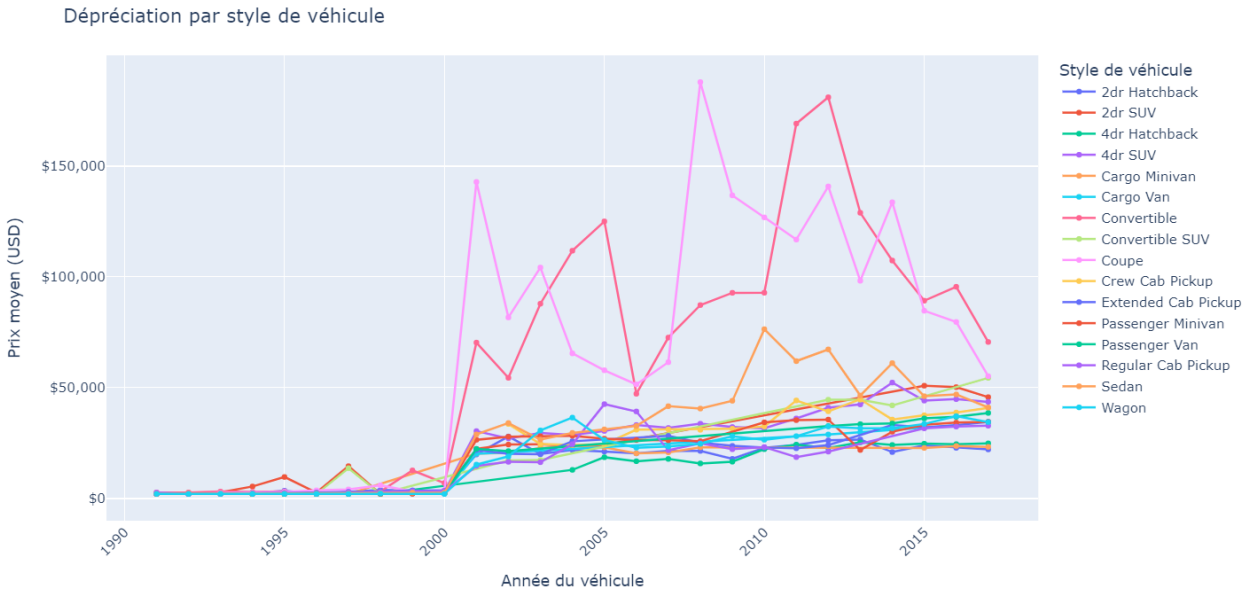

FIGURE 6.9 – Dépréciation par style de véhicule

Ce graphique explore la manière dont le style de carrosserie influence le prix moyen d'un véhicule au fil du temps. Chaque courbe correspond à un type de style (SUV, coupé, van, etc.).

Constats notables :

— Les *SUV 4 portes*, *Convertibles* et *Crew Cab Pickups* conservent une valeur relativement stable, souvent au-dessus de 50 000 dollars. Cette rétention de valeur témoigne de leur attractivité constante sur le marché, même en occasion.

— À l'inverse, les *Cargo Vans*, *2dr Hatchbacks* ou encore les *Sedans* restent globalement en dessous de 20 000 dollars, traduisant une moindre demande ou un usage plus utilitaire.

— Le cas particulier des *Convertible SUV* est à souligner : malgré leur rareté, ces modèles enregistrent des pics de prix dépassant 100 000 dollars. Leur rareté et leur image de niche haut de gamme expliquent cette valorisation atypique.

Interprétation : Les styles qui véhiculent un imaginaire de plaisir (convertible, SUV, pickup) ou de statut conservent mieux leur prix que ceux purement fonctionnels. La perception sociale et l'usage projeté semblent donc peser lourdement dans la dépréciation d'un véhicule.

Dépréciation par catégorie marketing

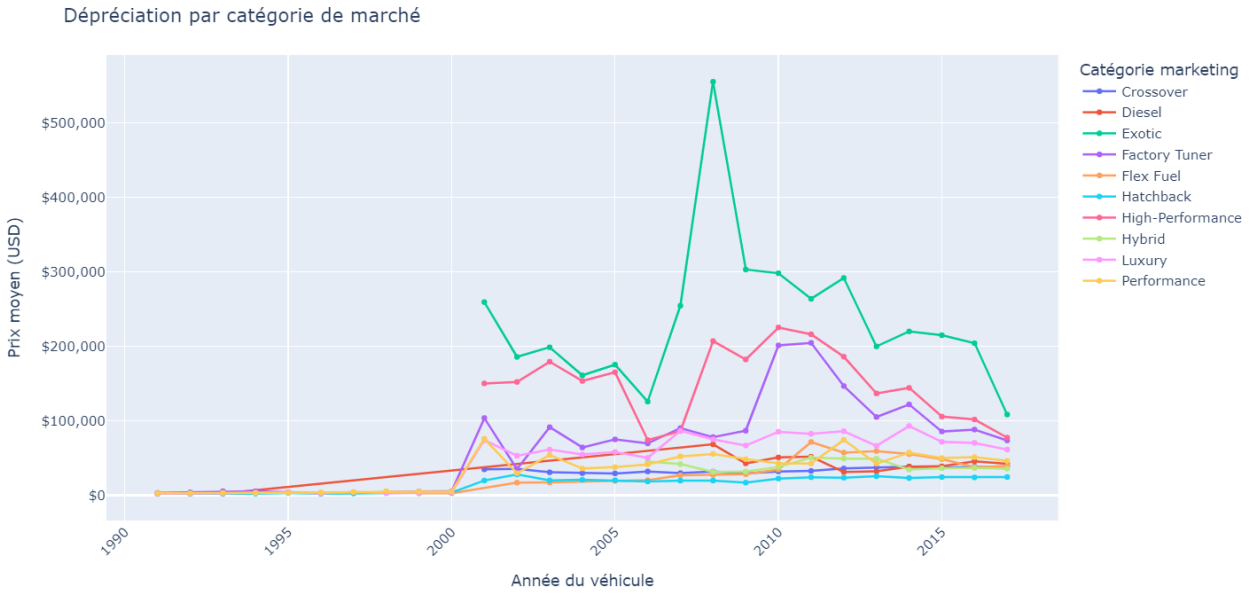

FIGURE 6.10 – Dépréciation par catégorie marketing

Ici, le graphique fait le lien entre la catégorie marketing d'un véhicule (Luxury, Hybrid, Diesel, etc.) et l'évolution de sa valeur au fil du temps.

Tendances principales :

— Les véhicules classés *Exotic*, *High-Performance* ou *Luxury* dominent en termes de prix moyen. Ces segments dépassent largement les 60 000 dollars, et certains modèles de niche (ex : Exotic) peuvent frôler ou dépasser les 500 000 dollars.

— Les véhicules à vocation économique ou utilitaire comme les *Flex Fuel*, *Diesel* ou *Hatchback* restent dans une fourchette basse (en dessous de 30 000 dollars), avec une décroissance régulière.

— Le segment *Crossover* se démarque : il affiche une très bonne stabilité, avec des prix moyens compétitifs et peu de chute marquée. Sa polyvalence et sa modernité en font un style très recherché.

Ce que cela révèle : La dépréciation est fortement corrélée à la manière dont un véhicule est positionné. Plus l'image est prestigieuse, rare ou technologique, plus la valeur est préservée. À l'inverse, un véhicule perçu comme « utilitaire » est souvent plus sujet à l'obsolescence perçue. **Interprétation globale :**

— La catégorie marketing influence fortement la rétention de valeur : les modèles haut de gamme, performants ou hybrides sont mieux valorisés dans le temps.

— À l'inverse, les véhicules à vocation pratique ou économique, bien qu'utiles, perdent plus rapidement de leur valeur.

Ces résultats confirment que la perception du véhicule joue un rôle central dans sa valeur résiduelle. Intégrer cette information dans le moteur de recommandation d'AutoPredict permet d'adapter les suggestions aux attentes des utilisateurs, en tenant compte de leur sensibilité au prestige, à la performance ou à l'économie.

Chapitre 7

Partie Machine Learning

7.1 Objectif et Intégration

La composante *Machine Learning* du projet **AutoPredict** a pour objectif de fournir une fonctionnalité intelligente essentielle :

— **L'estimation du prix** de véhicules d'occasion en fonction de leurs caractéristiques techniques.

Cette estimation s'adresse principalement à un utilisateur se plaçant dans la peau d'un vendeur souhaitant connaître la valeur de revente potentielle de son véhicule. L'intégration de cette composante permet ainsi de proposer un service à forte valeur ajoutée, en facilitant la prise de décision pour les particuliers ou les professionnels. Le code est disponible dans le notebooks `AutoPredict_ML.ipynb`.

7.2 Préparation des Données

7.2.1 Nettoyage

Avant l'entraînement du modèle, un travail de préparation minutieux a été réalisé, avec des étapes spécifiques par rapport à l'analyse des données :

— **Nettoyage des doublons et des valeurs manquantes**. Nous avons décidé de supprimer les colonnes présentant un nombre élevé de valeurs manquantes pour éviter la suppression d'un trop grand nombre de lignes. Par exemple, la colonne "market" comportait 3376 valeurs manquantes sur un total de 11914 lignes et a donc été supprimée. La colonne "popularité" a également été retirée car elle n'était pas pertinente pour la suite.

— **Standardisation des noms de colonnes** pour faciliter leur manipulation, certains noms n'étant pas suffisamment explicites.

— **Regroupement de plusieurs catégories en une**. Dans la colonne du type de carburant, nous avons regroupé les nombreuses catégories en six grandes catégories (essence, flex fuel, électrique, diesel, gaz naturel et autre). Cela sera particulièrement utile pour l'interface utilisateur (UI).

— **Suppression des outliers**. Bien que les outliers soient nécessaires pour l'analyse des

données, ils peuvent nuire à l'entraînement du modèle. Nous avons donc choisi de les supprimer pour obtenir un modèle plus précis. Les véhicules présentant des écarts de prix et de puissance trop importants ont été écartés.

— **Suppression des colonnes non pertinentes pour le prix**. Pour éviter d'entraîner le modèle sur des colonnes inutiles, nous avons supprimé certaines d'entre elles. Nous avons utilisé une matrice de corrélation pour les variables numériques, ainsi qu'une analyse ANOVA et un test du Khi-deux pour les variables catégorielles. Cela nous a permis d'écarter des variables numériques telles que le nombre de portes, la consommation et la consommation urbaine/route. Toutes les variables catégorielles semblent avoir un impact significatif et ont donc été conservées.

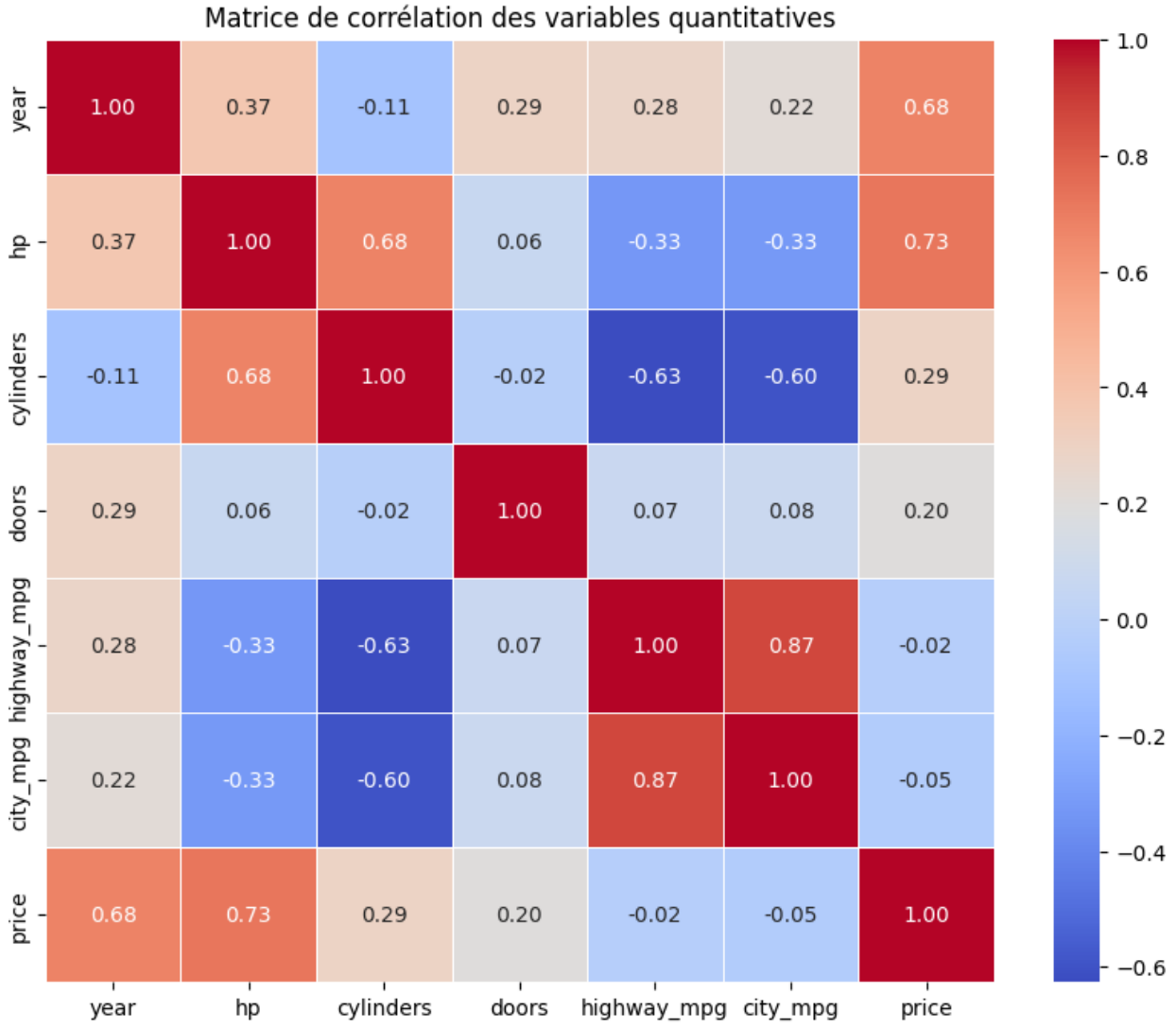

FIGURE 7.1 – Matrice de corrélation

7.2.2 Champs utilisés pour le modèle

Les champs suivants ont été utilisés pour le modèle : *make, model, year, fuel_type, hp, transmission, cylinders, drive, size, style.*

7.3 Modèle et Entraînement

7.3.1 Algorithmes explorés

Dans une démarche d'exploration comparative, trois approches ont été testées pour estimer le prix d'un véhicule à partir de ses caractéristiques techniques : **KNN**, **RNN** et **Random Forest**. Chacune de ces approches présente des avantages spécifiques :

- **K-Nearest Neighbors (KNN)** : Bien que les données soient étiquetées, nous avons souhaité tester ce modèle car il permet de capturer des relations locales entre les caractéristiques des véhicules. Cependant, il peut être sensible à la dimensionnalité des données et nécessite une optimisation du paramètre k. Pour cela, nous avons implémenté une recherche exhaustive similaire à une Grid Search, où nous avons testé chaque valeur de k dans une plage allant de 1 à 30. Cette approche nous a permis de déterminer la valeur optimale de k qui minimise l'erreur quadratique moyenne (RMSE).
- **Random Forest (RF)** : Ce modèle nous a semblé immédiatement pertinent en raison de sa capacité à gérer des données de grande dimension et à capturer des interactions complexes entre les variables. De plus, il est robuste face aux valeurs aberrantes et offre une bonne interprétabilité grâce à l'importance des variables.
- **Recurrent Neural Network (RNN)** : Les RNN sont particulièrement adaptés pour traiter des séquences de données, ce qui peut être intéressant si l'on considère l'historique des prix ou des caractéristiques temporelles. Cependant, leur entraînement peut être plus complexe et nécessite des ressources computationnelles importantes.

7.3.2 Encodage

Target Encoding

Pour les modèles KNN et RNN, nous avons opté pour le *target encoding* sur la variable prix. Cette méthode s'est avérée la plus efficace pour ces algorithmes, car elle permet de transformer les variables catégorielles en valeurs numériques tout en préservant l'information sur la cible.

Label Encoding

Pour le modèle Random Forest (RF), nous avons utilisé le *label encoding*, qui s'est révélé être le plus performant. Cette technique assigne une valeur numérique unique à chaque catégorie, ce qui est bien adapté aux arbres de décision qui composent la forêt aléatoire.

7.3.3 Séparation des données et évaluation

Afin d'évaluer la performance de notre modèle sans le biaiser, nous avons choisi de diviser notre jeu de données en deux sous-ensembles : un ensemble d'entraînement et un ensemble de test. La proportion utilisée pour le test, définie par la variable `test_size`, a été fixée à 30 % des données. Cette séparation permet de mesurer la capacité du modèle à généraliser sur des données inédites, et ainsi de limiter le risque de surapprentissage (overfitting).

7.3.4 Évaluation et Résultats

Validation croisée

Nous avons utilisé la méthode de validation croisée K-Fold pour évaluer les performances des modèles. Cette approche permet de garantir une évaluation robuste en divisant les données en plusieurs sous-ensembles et en entraînant le modèle sur différentes combinaisons de ces sous-ensembles. Nous avons effectué une validation croisée avec 5 folds à chaque fois. Aucun des modèles n'a montré de disparité significative dans les résultats, ce qui indique qu'il n'y a pas d'overfitting. Cela aide à prévenir le surapprentissage et à assurer que le modèle ne se contente pas de mémoriser les données d'entraînement, mais apprend plutôt à identifier et à généraliser à partir de patterns sous-jacents.

K-Nearest Neighbors (KNN)

— **Erreur quadratique moyenne (RMSE)** : $3575.05
— **Erreur quadratique moyenne (RMSE) en pourcentage** : 12.28%
— **Score** R^2 : 0.9463

Random Forest (RF)

— **Erreur quadratique moyenne (RMSE)** : $3237.02
— **Erreur quadratique moyenne (RMSE) en pourcentage** : 11.12%
— **Score** R^2 : 0.9560

Réseau de Neurones (RNN)

— **Erreur quadratique moyenne (RMSE)** : $4055.24
— **Erreur quadratique moyenne (RMSE) en pourcentage** : 13.93%
— **Score** R^2 : 0.9310

7.3.5 Modèle choisi

Nous avons opté pour le modèle **Random Forest (RF)** avec *label encoding*. Cette décision repose sur plusieurs justifications :

— **Performance supérieure** : Le modèle RF a démontré une meilleure performance en termes d'erreur quadratique moyenne (RMSE) et de score R^2 par rapport aux autres modèles testés.

— **Robustesse** : Les forêts aléatoires sont robustes face aux valeurs aberrantes et peuvent gérer efficacement les interactions complexes entre les variables.
— **Interprétabilité** : Le RF permet une meilleure interprétabilité des résultats grâce à l'importance des variables, ce qui est un atout pour comprendre les facteurs influençant le prix des véhicules.

Ces raisons combinées font du Random Forest le choix optimal pour notre projet.

7.4 Interaction avec le Frontend

7.4.1 Variables saisies par l'utilisateur

Nous avons cherché à rendre la saisie des données par l'utilisateur la moins contraignante possible. Pour ce faire, nous avons consulté plusieurs utilisateurs potentiels afin qu'ils nous communiquent les caractéristiques qu'ils jugent les plus importantes lors de l'achat d'une voiture. Les caractéristiques retenues sont les suivantes :

— *make* (marque)
— *model* (modèle)
— *year* (année)
— *fuel_type* (type de carburant)
— *hp* (puissance en chevaux)
— *transmission* (type de transmission)

Ces caractéristiques ont également été identifiées comme significatives dans nos analyses de corrélation, d'ANOVA et de Khi-deux. Nous les avons donc privilégiées dans l'interface utilisateur (UI). Cependant, pour ne pas compromettre les performances du modèle en négligeant les autres variables, nous avons procédé à l'inférence des variables manquantes.

7.4.2 Inférence des variables manquantes

Nous avons mis en place un système d'inférence pour estimer les valeurs des variables manquantes, telles que *style*, *drive*, *cylinders* et *size*. Ce système utilise des filtres successifs basés sur les caractéristiques saisies par l'utilisateur :

1. Année du véhicule
2. Puissance en chevaux (*hp*)
3. Marque (*make*)
4. Type de carburant (*fuel_type*)
5. Type de transmission (*transmission*)

Pour les variables numériques, nous avons utilisé la moyenne des valeurs filtrées. Pour les variables catégorielles, nous avons retenu la catégorie la plus représentée après application des filtres. Si aucun élément n'est retourné par les filtres, la variable sera encodée avec la valeur -1.

7.4.3 Tolérance à la panne

7.4.4 Tolérance aux erreurs de saisie

Pour gérer les cas où l'utilisateur fait des erreurs typographiques, comme écrire "5 series" au lieu de "serie 5" ou omettre un "s", nous avons mis en place un mécanisme de correspondance approximative. Ce mécanisme fonctionne comme suit :

1. **Normalisation du texte** : Les entrées de l'utilisateur et les valeurs connues sont normalisées pour uniformiser leur format. Cela inclut la conversion en minuscules, la suppression des accents et des espaces superflus.
2. **Recherche de correspondances approximatives** : Nous utilisons la bibliothèque `difflib` pour trouver des correspondances approximatives entre le texte normalisé de l'utilisateur et les valeurs connues. Si la similarité entre deux textes dépasse un seuil de 0.6, ils sont considérés comme correspondants.
3. **Sélection de la meilleure correspondance** : Si une correspondance est trouvée, la valeur connue la plus proche est sélectionnée. Sinon, un message d'avertissement est affiché et la valeur est encodée comme -1.
4. **Encodage des valeurs** : Les valeurs saisies par l'utilisateur sont encodées en utilisant les encodeurs spécifiés. Si une valeur ne correspond pas exactement à une valeur connue, la correspondance approximative est utilisée.
5. **Normalisation des valeurs numériques** : Les valeurs numériques sont normalisées en utilisant les valeurs minimales et maximales spécifiées dans les normalisateurs.

Ce mécanisme permet de rendre le système plus tolérant aux erreurs de saisie, améliorant ainsi l'expérience utilisateur.

7.4.5 Affichage des résultats

Les résultats sont affichés dans le frontend sous cette forme :

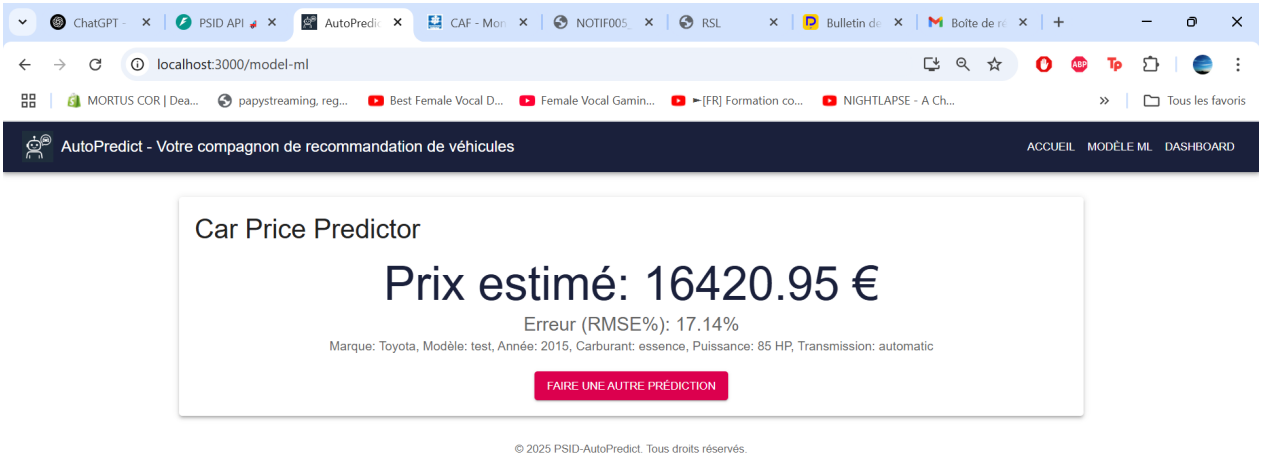

FIGURE 7.2 – Affichage des résultats

Le prix estimé est affiché en dollars puisqu'il s'agit d'un dataset américain. Nous avons également affiché le taux RMSE en pourcentage pour donner une idée de la précision du modèle. Ainsi que les données envoyées à l'API pour que l'utilisateur puisse vérifier les données qu'il a saisies.