Une image contenant voiture, véhicule, roue, pneu

Description générée automatiquement

Benjamin Colson

Edgar Saiz

Isra Friia

Luxon Masseau

Papelard Charlotte

**Combien vaut votre voiture ?**

Une estimation du prix de revente de votre voiture

Une image contenant voiture, véhicule, roue, pneu

Description générée automatiquement

Table des matières

[Remerciements 3](#_Toc153400509)

[Introduction 4](#_Toc153400510)

[1. Interface utilisateur 6](#_Toc153400511)

[1.1. L’interface d’accueil 6](#_Toc153400512)

[1.2. Le formulaire 6](#_Toc153400513)

[1.3. L’interface de sortie 8](#_Toc153400514)

[2. Préparation des données 10](#_Toc153400515)

[2.1. Scraping 10](#_Toc153400516)

[2.2. Nettoyage des données 10](#_Toc153400517)

[2.3. Lien VBA-Python 10](#_Toc153400518)

[3. Méthode de prédiction 12](#_Toc153400519)

[3.1. Essais de méthode 12](#_Toc153400520)

[3.2. Méthode sélectionnée 17](#_Toc153400521)

[Conclusion 19](#_Toc153400522)

[Annexe 20](#_Toc153400523)

[Présentation de la base 20](#_Toc153400524)

[Architecture du projet 22](#_Toc153400525)

# Remerciements

Nous tenons à adresser nos remerciements particuliers à Morgan Cousin, dont les apports méthodologiques et les outils de travail novateurs ont profondément enrichi notre approche. Ses commentaires avisés ont joué un rôle déterminant dans notre évolution, tant sur le plan de nos compétences verbales que de notre structuration interne.

Nous exprimons également notre gratitude envers Amor Keziou, dont les conseils éclairés sur le choix des modèles de prédiction ont jeté une lumière critique sur nos méthodes.

Enfin, nous saluons Arona Diene pour son expertise précieuse dans la conception et le développement de notre interface.

# Introduction

L'industrie automobile en France demeure un secteur d'activité incontournable, suscitant un intérêt soutenu et occupant une place centrale dans la vie quotidienne des Français. Avec plus de 37,8 millions de voitures particulières en circulation, le paysage automobile français évolue constamment. Malgré cela, l'achat d'une voiture neuve représente souvent un investissement financier considérable, incitant de nombreuses personnes à se tourner vers le marché des véhicules d'occasion. La question de l'achat et de la revente de voitures d'occasion est ainsi cruciale et concerne des millions de citoyens. C'est dans ce contexte que notre équipe s'est engagée dans la création d'une application visant à simplifier le processus d'estimation du prix de revente des véhicules d'occasion, afin de répondre aux besoins des propriétaires de voitures d'occasion en France.

Permettez-moi de vous présenter les membres dévoués de notre équipe qui ont mis leurs compétences au service de cette application :

* **Benjamin Colson (Product Owner)** : Il guide l'équipe en définissant la vision du produit, en priorisant les fonctionnalités et en veillant à ce que l'application réponde aux besoins des utilisateurs. Fort de son expertise en gestion de projets, il est le moteur derrière notre application de prédiction sur le prix des voitures.
* **Edgar Saiz (Scrum Master)** : En tant que Scrum Master, il assure l'application des méthodologies agiles, favorise la collaboration au sein de l'équipe et garantit que les objectifs du projet sont atteints dans les délais.
* **Charlotte Papelard (Développeuse VBA)** : Spécialisée dans le développement de fonctionnalités et d'outils pour l'analyse des données, elle crée des modèles de prédiction en utilisant VBA, automatisant ainsi le traitement des données.
* **Isra Friaa (Data Engineer)** : Son rôle essentiel est la collecte, la préparation et la structuration des données alimentant l'application, assurant ainsi la disponibilité de données de qualité pour des prédictions précises en matière de prix du véhicule.
* **Luxon Masseau (Data Scientist)** : Il se concentre sur l'élaboration de l'algorithme de prédiction au cœur de l'application, développant des modèles analytiques et statistiques pour faire de notre application un outil fiable pour prédire le prix des voitures des automobilistes.

Notre équipe vous propose un rapport explorant la conception, le développement et le résultat d’une telle application. L’objectif de ce projet est de concevoir un outil permettant aux automobilistes d’estimer rapidement le prix de leur voiture en renseignant ses caractéristiques. Cependant, cet outil n'a pas pour but de remplacer une estimation faite par un professionnel de l'automobile, mais il doit fournir une estimation comparative des différentes voitures. Les utilisateurs pourront accéder à notre outil via un formulaire VBA, renseignant les caractéristiques de leur véhicule, tandis que l'algorithme en parallèle récupérera ces données pour fournir une estimation du prix du véhicule, accompagnée de statistiques complémentaires.

# Interface utilisateur

Cette partie explore tout le travail effectué sous Excel et VBA, i.e. qu’elle explore l’interface d’accueil (1.1), le formulaire (1.2.) ainsi que le l’interface de sortie (1.3.).

## L’interface d’accueil

Nous accueillons nos utilisateurs sur une page d’accueil (*cf.* Figure 1) qui rappelle le contexte et le but de l’application ainsi que son fonctionnement. Elle explique le bouton « Formulaire » qui sera la première étape de la prédiction du prix du véhicule ainsi que le bouton « Mise à jour » qui permet de remettre à jour la base de données.

Une image contenant texte, roue, véhicule, Véhicule terrestre

Description générée automatiquement

Figure 1 : Interface d’accueil de l’application

Après une éventuelle mise à jour de l’application, l’utilisateur lance le formulaire.

## Le formulaire

Une fois le formulaire lancé (*cf.* Figure 2), l’utilisateur arrive sur un multipage avec une page « Voiture » et une page « Carte grise ». Cette structuration soulage le formulaire et donne à l’utilisateur une vision intuitive de ce qu’il va devoir faire : d’une part, il va devoir remplir les caractéristiques de sa voiture, d’autre part, il devra aller chercher sa carte grise pour compléter les informations manquantes.

Dans la section « Voiture » :

* Pour trouver la marque et le modèle, une aide à la sélection est proposée. Dès lors, lorsqu’on commence à taper la marque d’une voiture, une liste déroulante apparait et toutes les correspondances avec les premières lettres tapées apparaissent.
* Pour les caractéristiques suivantes (carburant, boite de vitesse, catégorie et année) une liste déroulante incluant toutes les réponses possibles est proposée afin de faciliter et sécuriser la saisie. A noter que si la saisie de la boite de vitesse est « automatique », le nombre de vitesse s’enlève car une voiture automatique n’a pas de vitesse.
* Pour les informations de type numérique comme kilométrage, nombre de places, nombre de portes ou nombre de vitesses, une sécurité empêche de mettre des caractères ou autre symboles spéciaux - seuls les chiffres peuvent être utilisés.

Dans la section « Carte grise » :

* Pour chaque caractéristique à remplir à l’aide de sa carte grise (cylindrée, puissance fiscale et physique) une aide visuelle pour trouver l’information s’affiche. Une sécurité est également mise en place pour forcer l’utilisateur à inscrire des chiffres.
* Le bouton « Valider » enregistre les informations de l’utilisateur.
* Le bouton « Prédire » lance la prédiction.

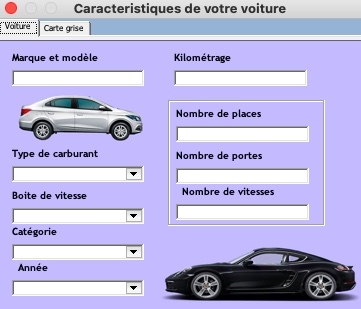


Figure 2 : Formulaire de saisie de l’application

Une fois que le bouton prédire est lancé, l’utilisateur est renvoyé sur l’interface de sortie.

## L’interface de sortie

Sur l’interface de sortie (*cf.* Figure 3), nous pouvons observer plusieurs informations, la première, celle mise en valeur est forcément le prix estimé de votre voiture, nous trouvons à coté le prix moyen de la marque de votre voiture c’est-à-dire que si votre voiture est une Peugeot, nous obtiendrons le prix moyen de toutes les Peugeot de la base de données. De l’autre côté nous avons le nombre de voiture de la même marque que la vôtre. Juste en dessous, 3 graphiques, le premier, un diagramme en donut permettant d’apporter de l’information à l’utilisateur de l’âge des voitures présentes dans notre base. On retrouve également un diagramme de Pareto dans l’objectif de toujours être le plus transparent avec le client exprimant ainsi le nombre de voiture par marque dans la base. Le graphique le plus important est au centre où nous retrouvons le prix moyen des 4 marques de voitures les plus présentes dans la base ainsi que celui de la marque de votre voiture. La marque de votre voiture sera en plus claire pour la distinguer des autres et un indicateur s’y trouvera pour que vous puissiez comparer le prix estimé de votre voiture avec le prix moyen des voitures de votre marque de manière visuelle. Nous retrouvons ensuite le taux de précision de 98% et l’erreur prévisionnelle obtenue dans notre algorithme. Nous rappelons également le kilométrage et le nombre de chevaux physiques de la voiture qui sont les caractéristiques techniques qui influencent le plus le prix de votre voiture et c’est cette dernière information qui est présente par le podium où nous rajoutons l’année de mise en circulation du véhicule comme variable importante.

Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Description générée automatiquement

Figure 3 : Interface de sortie de l’application

# Préparation des données

Une fois le formulaire complété par l’utilisateur, ses données vont être stockées avant d’être envoyées vers un script python qui lancera l’algorithme de prédiction. Toutefois, pour que l’algorithme fonctionne correctement, il lui faut des données sur lesquelles apprendre (2.1.), qu’elles soient propres (2.2.) et que les données de l’utilisateur soient bien envoyées vers ce script (2.3.).

## La base de données est construite avec une méthode de scraping

Le scraping, également connu sous le nom de web scraping, est une technique informatique qui consiste à extraire des données à partir de sites web de manière automatisée. Cette pratique utilise des outils spécifiques pour naviguer à travers les pages web, récupérer le contenu structuré, et extraire les informations recherchées.

Dans la réalisation de ce projet, deux différentes méthodes de scraping ont été appliquées :

* Dans un premier temps, la méthode privilégiée pour récupérer les données sur le site internet Spoticar était selenium. Selenium est une suite d'outils destinée à l'automatisation des navigateurs web. Elle offre une interface permettant aux développeurs de simuler les actions des utilisateurs sur un navigateur, facilitant ainsi le test automatisé des applications web ou l'exécution de tâches répétitives. Ainsi, Selenium nous a permi de récupérer les informations de la première étiquette d’une voiture sur le site (cf. Figure 4) *i.e.* les informations concernant la marque et le modèle, son prix, son type de carburant, la boite de vitesse, son année de mise en circulation, son kilométrage. Au total, c’est 6 variables qui ont été récupérées.
* Dans un second temps, nous avons trouvé que 6 variables, c’était trop peu. Nous voulions que le scraping nous emmène plus loin *i.e.* qu’il fasse en sorte de cliquer sur l’étiquette pour avoir plus d’informations (cf. Figure 5). Nous sommes donc passé par une méthode plus rapide, à savoir aller voir directement l’API du site. Les ressources utilisées sont moindres et le scraping est plus rapide. L’expérience utilisateur est donc améliorée et la prédiction est plus précise.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Figure 4 : Étiquette d’une voiture

Une image contenant texte, véhicule, Véhicule terrestre, voiture

Description générée automatiquement

Figure 5 : Vente d’une voiture sur le site Spoticar

## Nettoyage des données

## Lien VBA-Python

Une fois les données du client récupérées, l'étape suivante consiste à déclencher notre algorithme de prédiction pour obtenir le résultat final. Mais comment établir ce lien ? Dès que nous amorçons la mise à jour, le script du modèle est activé, générant ainsi les coefficients de chaque variable grâce à l'algorithme RandomForest. Ces informations cruciales sont ensuite enregistrées dans un fichier Pickle, lequel est constamment mis à jour à chaque nouvelle opération de scraping. Lorsque nous effectuons des prédictions sur le prix de votre véhicule, nous extrayons ce fichier Pickle pour le connecter aux données fournies par l'utilisateur. Le résultat prédit est ensuite enregistré dans une cellule dédiée d'une feuille de calcul Excel, créant ainsi une traçabilité claire du processus de prédiction.

# Méthode de prédiction

Dans cette dernière partie, nous allons expliquer la mise en œuvre de notre algorithme, des premiers essaies à la sélection finale par RandomForest.

## Essais de méthode

Aux vues de notre variable cible qui est je le rappelle le prix du véhicule, une variable quantitative, nous sommes partis sur une régression linéaire multiple. La régression linéaire multiple est une technique statistique utilisée pour modéliser la relation entre une variable dépendante et plusieurs variables indépendantes. Contrairement à la régression linéaire simple qui se concentre sur une seule variable prédictive, la régression linéaire multiple prend en compte plusieurs prédicteurs, permettant ainsi de mieux capturer la complexité des relations entre les variables. L'objectif principal de cette méthode est de créer un modèle mathématique qui représente au mieux la variation de la variable dépendante en fonction des différentes variables indépendantes. Chaque variable indépendante est associée à un coefficient, indiquant l'ampleur de son impact sur la variable dépendante tout en tenant compte des autres prédicteurs. L'ajustement du modèle est effectué par minimisation des erreurs résiduelles, optimisant ainsi la précision des prédictions. La régression linéaire multiple trouve des applications dans divers domaines, notamment l'économie, la finance et les sciences sociales, offrant une approche puissante pour analyser et modéliser des systèmes complexes avec plusieurs facteurs explicatifs.

Bien sûr ! La formule d'une régression linéaire multiple est généralement exprimée comme suit :

Où :

- est la variable dépendante que vous essayez de prédire.

- est l'ordonnée à l'origine (constante).

- sont les coefficients des variables indépendantes respectivement.

- sont les variables indépendantes.

- est le terme d'erreur, représentant les erreurs non capturées par le modèle.

En termes plus simples, la valeur de la variable dépendante () est estimée comme une combinaison linéaire des valeurs des variables indépendantes, ajustée par des coefficients et une constante. L'objectif de la régression linéaire multiple est de trouver les valeurs optimales des coefficients () qui minimisent les erreurs de prédiction.

Nous avons ensuite créé une sélection de modèle selon plusieurs critères. Le processus de sélection des modèles est une étape cruciale dans la modélisation statistique, visant à identifier le modèle le plus approprié parmi plusieurs candidats. Différents critères, tels que l'AIC (Critère d'Information d'Akaike) et le BIC (Critère d'Information Bayésien), sont souvent utilisés pour évaluer la qualité des modèles. L'AIC prend en compte la qualité du modèle et la complexité, favorisant les modèles qui parviennent à bien ajuster les données tout en restant relativement simples. Le BIC, quant à lui, pénalise davantage la complexité, encourageant une plus grande parcimonie dans la sélection du modèle.

L'AIC (Critère d'Information d'Akaike) et le BIC (Critère d'Information Bayésien) sont des mesures utilisées pour évaluer la qualité d'un modèle statistique, y compris les modèles de régression. Ces critères prennent en compte la performance du modèle tout en pénalisant la complexité pour éviter le surajustement.

1. \*\*AIC (Critère d'Information d'Akaike)\*\* :

- : Le nombre de paramètres dans le modèle.

- : La fonction de vraisemblance maximale du modèle.

L'AIC favorise les modèles qui ajustent bien les données tout en ayant un nombre minimal de paramètres. Un AIC plus bas indique un modèle potentiellement meilleur.

2. \*\*BIC (Critère d'Information Bayésien)\*\* :

- : Le nombre de paramètres dans le modèle.

- : Le nombre d'observations.

- : La fonction de vraisemblance maximale du modèle.

Le BIC pénalise davantage la complexité du modèle que l'AIC. Comme pour l'AIC, un BIC plus bas indique généralement un modèle préférable.

Dans ces formules, représente le logarithme naturel. L'idée est de trouver un équilibre entre l'ajustement du modèle aux données et la complexité du modèle, afin d'éviter le surajustement. Lorsque vous comparez différents modèles, celui avec la valeur d'AIC ou de BIC la plus basse est généralement considéré comme le meilleur, compte tenu de la balance entre ajustement et complexité.

Par ailleurs, les méthodes de régularisation telles que Ridge et Lasso sont couramment employées pour améliorer la performance des modèles de régression en introduisant des termes de pénalité sur les coefficients. Le Ridge ajoute une pénalité quadratique à la somme des carrés des coefficients, tandis que le Lasso introduit une pénalité absolue. Ces approches sont particulièrement utiles pour traiter la multicolinéarité et la sélection automatique des variables, améliorant ainsi la généralisation du modèle. En fin de compte, le choix entre ces méthodes dépend du contexte spécifique du problème et des caractéristiques des données, chaque critère et méthode apportant sa propre perspective pour guider la sélection du modèle optimal.

Les méthodes de régularisation, telles que Ridge et Lasso, sont utilisées pour éviter le surajustement dans les modèles de régression. Les formules pour les termes de pénalité dans Ridge et Lasso sont ajoutées à la fonction de coût de la régression linéaire ordinaire.

1. \*\*Ridge Regression (L2 regularization)\*\* :

Dans Ridge Regression, on ajoute le terme de régularisation L2 à la fonction de coût :

- : Fonction de coût.

- : Somme des carrés des résidus, comme dans la régression linéaire ordinaire.

- : Paramètre de régularisation (lambda dans certaines notations).

- : Terme de régularisation L2, qui pénalise les coefficients en les poussant vers zéro.

2. \*\*Lasso Regression (L1 regularization)\*\* :

Dans Lasso Regression, on ajoute le terme de régularisation L1 à la fonction de coût :

- : Fonction de coût.

- : Somme des carrés des résidus, comme dans la régression linéaire ordinaire.

- : Paramètre de régularisation.

- : Terme de régularisation L1, qui pénalise les coefficients en les poussant vers zéro. La particularité de Lasso est qu'il peut conduire à des coefficients exactement égaux à zéro, réalisant ainsi une sélection de variables.

Dans ces formules, représente les coefficients de régression, est le nombre de prédicteurs, et est le paramètre de régularisation qui contrôle l'intensité de la régularisation. Plus est grand, plus la pénalité est forte, conduisant à des coefficients de régression plus petits.

Ces termes de régularisation sont ajoutés à la fonction de coût afin de trouver une solution qui équilibre la meilleure adéquation aux données et la complexité du modèle.

Lors de la réalisation d'une régression linéaire multiple (RLM), la validation des hypothèses sous-jacentes est essentielle pour garantir la fiabilité des résultats obtenus. Plusieurs hypothèses fondamentales doivent être vérifiées pour s'assurer que le modèle est robuste et que ses conclusions sont interprétables. Tout d'abord, l'homoscédasticité, qui suppose une variance constante des résidus à tous les niveaux des variables indépendantes, doit être évaluée. Des graphiques de résidus ou des tests statistiques peuvent être utilisés pour détecter d'éventuelles violations de cette hypothèse.

Deuxièmement, la normalité des résidus est une hypothèse importante. Idéalement, les résidus devraient suivre une distribution normale pour que les tests d'inférence soient valables. Des graphiques quantiles-quantiles et des tests statistiques comme le test de Shapiro-Wilk peuvent être appliqués pour évaluer la normalité des résidus.

Troisièmement, la linéarité entre les variables indépendantes et la variable dépendante doit être vérifiée. Des graphiques de régression partielle ou des graphiques de résidus peuvent aider à identifier des tendances non linéaires.

En outre, il est crucial de s'assurer que les résidus sont indépendants les uns des autres. La présence de dépendance sérielle peut compromettre l'efficacité des estimations.

Enfin, la multicolinéarité, qui mesure la corrélation entre les variables indépendantes, doit être évaluée. Une forte multicolinéarité peut rendre les estimations des coefficients peu fiables. Des diagnostics tels que le VIF (variance inflation factor) peuvent aider à identifier ce problème.

En résumé, la validation des hypothèses lors d'une RLM est cruciale pour s'assurer de la validité des conclusions tirées du modèle. Elle nécessite une combinaison de diagnostics graphiques et de tests statistiques pour évaluer la conformité aux différentes hypothèses sous-jacentes à la régression linéaire multiple.

Rajouter les graphs

## Méthode sélectionnée

La forêt aléatoire, ou RandomForest en anglais, est une méthode d'apprentissage automatique qui repose sur l'agrégation de multiples arbres de décision pour améliorer la robustesse et la précision des prédictions. Cette approche présente plusieurs avantages, notamment sa capacité à gérer des ensembles de données complexes, non linéaires et comportant de nombreuses variables. Un aspect fondamental de la forêt aléatoire réside dans la diversité des arbres de décision qui la composent, obtenue en introduisant des éléments d'aléatoire lors de la construction de chaque arbre.

Chaque arbre de décision est formé sur un sous-ensemble aléatoire des données d'entraînement et sur un sous-ensemble aléatoire des variables explicatives. Cette diversification réduit le surajustement (overfitting) et améliore la généralisation du modèle aux données inconnues. Lors de la prédiction, chaque arbre contribue à la décision finale, et le résultat est déterminé par un vote majoritaire dans le cas de classifications ou par une moyenne dans le cas de régressions.

La forêt aléatoire est particulièrement efficace pour traiter des problèmes complexes, identifier des interactions non linéaires entre les variables et gérer des jeux de données de grande dimension. Elle est également robuste face au bruit et aux valeurs aberrantes. Cependant, il est important de noter que son interprétation peut être plus complexe par rapport à des modèles plus simples comme la régression linéaire.

En résumé, la forêt aléatoire est une technique d'apprentissage automatique puissante et polyvalente, souvent privilégiée pour sa capacité à fournir des prédictions précises et à gérer des données complexes tout en limitant le surajustement.

Le critère de Gini est une mesure d'impureté utilisée dans la construction d'arbres de décision, y compris ceux utilisés dans l'algorithme RandomForest. Il mesure l'homogénéité des classes dans un nœud donné d'un arbre de décision.

La formule du critère de Gini pour un nœud donné est la suivante :

- : le critère de Gini pour le nœud .

- : le nombre de classes dans le problème de classification.

- : la proportion d'échantillons du nœud qui appartiennent à la classe .

Le critère de Gini est compris entre 0 et 1. Un score de 0 indique une pureté maximale, ce qui signifie que tous les échantillons du nœud appartiennent à la même classe. Un score de 1 indique une impureté maximale, ce qui signifie que les échantillons sont uniformément répartis entre les différentes classes.

Lors de la construction d'un arbre de décision, le critère de Gini est utilisé pour évaluer la qualité de la séparation des données à chaque nœud. Plus la diminution du critère de Gini est importante après une division, plus la division est considérée comme bonne. Cela permet à l'algorithme de prendre des décisions pour maximiser la pureté des nœuds dans l'arbre.

# Conclusion

En conclusion, nous avons réussi à créer une application qui permet de prédire le prix d’un véhicule « de tous les jours ». Notre formulaire n’est en effet pas adapté pour les supercars ou les hypercars. Il permet cependant de remplir sa mission initiale, qui était aux usagers d’avoir un outil simple et intuitif pour prédire le prix de votre véhicule grâce à sa carte grise.

L’autre limite de notre application, c’est le fait qu’elle ne prend pas en compte si le véhicule nécessite ou non des réparations, l’état des pneus et en partant du principe d’un contrôle technique a été fait moins de 6 mois auparavant.

Si nous souhaitons prendre ces paramètres en compte pour pouvoir mieux prédire le prix d’un véhicule, cela nécessiterait des données supplémentaires ou même une base de données différente.

# Annexe

## Présentation de la base

Afin de construire une telle application, nous utiliserons une base de données obtenue par *scraping* sur le site [www.spoticar.fr](http://www.spoticar.fr), leader européen du marché de l’occasion, qui recense plus de 70 000 offres de voitures d’occasion. Le *scraping* possède plusieurs avantages :

**Contrôle** – Il permet d’avoir la main mise sur les données ;

**Actualisation** – Il permet la mise à jour de la base de données en direct, ce qui va permettre de ne pas rendre le produit obsolète vis-à-vis des prix pour ainsi suivre les variations du marché ;

**Personnalisation** – Il permet de faciliter le nettoyage de la base car les données sont récupérées à la source *i.e.* que seules les variables d’intérêt seront récupérées. C’est un vrai avantage car les variables et individus peuvent être augmentées au besoin.

Les variables

La variable que nous cherchons à prédire, *i.e.* notre **variable d’intérêt**, est le prix du véhicule. Les autres variables présentes dans notre base de données seront utilisées pour définir cette variable.

Ainsi, nos **variables explicatives**, qui seront sollicitées auprès du client pour une analyse approfondie de sa voiture seront :

Les données obtenues se présentent de la manière suivante :

|  |  |
| --- | --- |
| Marque | Marque de la voiture (Peugeot, Opel, etc.) |
| Modèle | Modèle de la voiture (308, 107, etc.) |
| Marque et modèle | Fusion des deux premières variables |
| Boite de vitesse | Automatique/Manuel |
| Couleur | Couleur de la voiture |
| Crit’air | Vignette sur la pollution de la voiture |
| Catégorie | Catégorie du véhicule (4x4, berline, etc.) |
| Cylindrée | Nb de cylindrée du véhicule |
| Kilométrage | Nb de kilométrage du véhicule |
| Nb place | Nb de places du véhicule |
| Nb porte | Nb de portes du véhicule |
| Nb vitesse | Nb de vitesse du boitier |
| Puissance fiscal | Nb de chevaux fiscaux |
| Puissance physique | Nb de chevaux physiques |
| Carburant | Type de carburant (essence, diesel, etc.) |
| Année | Année de mise en circulation du véhicule |
| Nb roues motrice | Nb de roues motrices |
| Prix | Prix du véhicule |

Cette base contient initialement 10 000 lignes qu’on nettoiera dans un premier temps avant de déployer notre modèle de sélection de variables puis de prédiction afin d’obtenir le meilleur résultat possible dans le prix du véhicule du client.

## Architecture du projet

Une image contenant texte, capture d’écran, diagramme, conception

Description générée automatiquement