Evaluation du prix des véhicules d'occasion sur le marché français

Jean-Louis Delebecque

Mémoire sur l'Automobile, la Data Science, le Marketing et les Ventes

Campus, Date : Pôle léonard de Vinci, 21/06/2021

Tuteur : Christophe Bourgoin

Compilation Similiratity Index : 1 %

Document confidentiel : OUI

Déclaration d'originalité de l'auteur

Je certifie par la présente que je suis le seul auteur de cette thèse et qu'elle présente principalement mes idées, mes analyses et mon évaluation des résultats de mes recherches.

Je certifie qu'à ma connaissance, ce document n'enfreint pas le droit d'auteur de quiconque et que toutes les idées, méthodologies, citations ou tout autre matériel sont protégés.

droits d'auteur de quiconque et que toutes les idées, méthodologies, citations ou tout autre matériau des travaux d'autres personnes ont été entièrement et correctement référencés dans le corps du travail et dans les de l'ouvrage et dans les références et/ou la bibliographie. Toutes les citations utilisent le système de référencement qui a été adopté par l'école.

Toutes les citations directes de données écrites ou verbales sont indiquées entre guillemets, et sont référencées de manière appropriée.

Toute partie de ce document qui a été utilisée dans des évaluations académiques précédentes au cours de mon programme d'études a été identifiée et référencée conformément aux règles de l'APA.

Je confirme qu'il n'y a pas eu d'aide non autorisée d'autres étudiants ou amis pendant la production de ce document. Aucune partie de ce travail ne provient d'un site Internet de soutien. *

Je déclare que ce travail est une copie conforme de mon propre travail et j'accepte qu'il soit soumis

J'accepte qu'il soit soumis au logiciel anti-plagiat utilisé par l'école pour confirmation de ce fait.

Date : 21/06/2021

Signature : Jean-Louis Delebecque

Remerciements

Je voudrais dans un premier temps remercier, mon directeur de mémoire M.BOURGOIN, pour sa disponibilité et son aide précieuse.
Je voudrais également remercier ma famille pour avoir partager mon sondage à leurs connaissances.

Résumé

Dans ce mémoire nous allons étudier comment évaluer le prix d'un véhicule d'occasion grâce au Machine Learning. Pour se faire, nous analyserons les techniques déjà existantes proposées sur le marché ainsi que des études recueillant des données sur les ventes et annonces de voitures. L'étude est focalisée sur le territoire français, mais concerne tous les constructeurs automobiles. C'est un problème de Machine Learning supervisé, dans un premier temps, nous ferons de la régression pour essayer de prédire le prix du véhicule le plus précisément possible, puis nous effectuerons un sondage sur le comportement des particuliers dans l'achat d'un véhicule d'occasion . Dans cette étude, vous verrez donc beaucoup de modèles différents : régression linéaire simple et multiple, des arbres de décision ou des réseaux de neurones pour le problème de régression. Nous évaluerons ensuite les résultats de ces modèles sur des données scrappées sur des sites de voitures d'occasion ou avec des datasets Kaggle s'ils sont suffisamment pertinents.

Mots-clés :

Prédiction , Machine Learning, automobile, marché de l'occasion, France

Abstract :

In this paper we will study how to evaluate the price of a used car using Machine Learning. To do so, we will analyze existing techniques proposed on the market as well as studies collecting data on car sales and ads. The study is focused on the French territory, but concerns all car manufacturers. It is a supervised Machine Learning problem, in a first step, we will do regression to try to predict the price of the vehicle as accurately as possible, then we will do a survey on the behavior of individuals in the purchase of a used car. In this study, you will see many different models: simple and multiple linear regression, decision trees or neural networks for the regression problem. We will then evaluate the results of these models on scrapped data from used car websites or with Kaggle datasets if they are relevant enough.

Key-words :

Prediction , Machine Learning, automotive, second hand market, France

Table des matières

Introduction

L'automobile a été créé en 1885 et depuis son utilisation est de plus en plus répandue. En effet 86 % des ménages français possèdent un ou plusieurs véhicules contre 79 % en 1998 (chiffres de 2018). Cette augmentation montre la croissance et la démocratisation de l'automobile en France où en 2018 on a recensé 34,19 millions de véhicules.

C'est donc un immense marché que représente aujourd'hui l'automobile français, en particulier le milieu de l'occasion. Car en France pour 1 véhicule neuf acheté 3.34 véhicules d'occasion seraient vendus, soit un ratio de 1 voiture neuve pour 3 d'occasion. Si le prix des véhicules neufs est étudié et régulé par les fabricants et leurs concurrents, le prix d'une voiture d'occasion peut lui varier selon différents facteurs et est souvent déterminé par rapport à des moyennes ou des cotes.

L'étude de la tarification des véhicules d'occasion n'est pas un problème nouveau, mais le nombre de ces véhicules ne peut qu'augmenter, l'achat de voitures neuves est aussi biaisé par certains facteurs comme les véhicules de fonction qui représenterait la moitié des véhicules vendus neufs en France.

Le marché de l'occasion lui est très intéressant, car il touche bien plus de véhicules, les tarifs sont parfois choisis par des particuliers qui ne font que s'adapter au prix du marché, la concurrence est bien plus large puisqu'on retrouve tout type de véhicules de n'importe quelle année. L'arrivée des véhicules électriques ou hybrides dans ce secteur est aussi quelque chose de nouveau. Comment se comporte le marché de l'occasion sur ces véhicules et quel impact cela a-t-il sur les véhicules thermiques par exemple ? Les prix des voitures de collection sont eux aussi durs à prédire, la cote de certains vieux modèles de BMW remontent aujourd'hui beaucoup car la demande a augmenté mais l'offre est restée la même puisque ces véhicules ne sont plus produits. Certains événements culturels (film, sport automobile) peuvent aussi influencer le prix de revente d'un véhicule d'occasion. On peut aussi penser aux voitures de collection dont le prix de revente peut dépasser le prix neuf, idem pour certains véhicules fabriqués en édition limitée, si on pense à l'offre et la demande certains indicateurs pourraient permettre de prédire cette montée de prix. Comme avec la Porsche 993, avec le recul cela parait évident car étant la dernière Porsche a refroidissement par air ce modèle a beaucoup

monté en cote avec les années. Pourrait-on avoir le même phénomène avec les derniers V12 Ferrari, la mode actuelle étant à la réduction de la cylindrée, certains moteurs à grosse cylindrée se faisant de plus en plus rare par exemple.

Si les événements culturels sont pour le moins prévisibles et très compréhensibles, d'autres facteurs sont eux pour le coup plus méconnus et nécessitent l'analyse de données.

C'est l'objet de cette étude : utiliser les Data Science pour établir un ou plusieurs modèles capables de déterminer le prix de revente d'un véhicule d'occasion en France. Ce modèle devra être basé sur un maximum de critères (features) caractérisant chaque véhicule. Le choix de l'utilisation de technique de Machine Learning est très pertinent dans ce problème. L'évolution des prix n'est observable qu'en fonction du temps, donc en entrainant un modèle sur des données de véhicules déjà vendus (on connaitrait alors déjà le prix), puis en faisant des prédictions sur le prix des véhicules en vente. Une fois ces véhicules vendus nous pourrons vérifier réellement la précision de notre modèle et l'optimiser de plus en plus.

Ce mémoire continuera avec la revue de littérature où nous étudierons les solutions déjà existantes, puis la méthodologie de recherche où nous présenterons précisément d'où viennent nos données et quels techniques et modèles nous allons utiliser. Après quoi nous entrerons dans la présentation des résultats et les premiers commentaires. Enfin nous donnerons notre conclusion sur ces résultats et nos recommandations.

Revue de littérature

Des études sur la prédiction du prix de revente d'un véhicule de seconde main existent déjà ; elles se distinguent entre elles par des axes de recherche spécifique. Il existe aussi des entreprises proposant d'évaluer le prix de votre véhicule d'occasion, comme L'ARGUS qui est une référence sur le territoire français.

Nous étudierons en première partie les solutions déjà proposées sur le marché par ces entreprises. Puis dans un second temps nous examinerons les axes étudiés et les conclusions des études déjà réalisées sur ce sujet. Enfin nous analyserons les modèles déjà existants et leurs résultats.

1. Les cotes automobiles en France : acteurs et services

1.1 La référence la plus réputée : L'Argus

L'ARGUS est un magazine historique de l'automobile français, crée en 1927 il couvre le sujet de l'automobile de manière professionnelle. De telle manière que lorsque les véhicules se sont démocratisés sur le territoire français, le magazine est devenu une revue de conseils sur l'achat et la revente de véhicules allant jusqu'à gagner en 1993 le trophée de l'utilitaire de l'année. Mais surtout il se sont imposés comme une référence, la « cote à l'argus » est réputée et utilisée par tous les garages français lorsqu'ils vendent ou achètent des véhicules d'occasion.
Il existe aujourd'hui d'autres entreprises proposant des cotes sur les véhicules mais L'ARGUS est la première entreprise à l'avoir fait en France ce qui leur donne une forte crédibilité et comme ils le précisent sur leur site leur estimation est utilisée par le fisc, l'administration, les douanes ou encore les notaires.

Toujours sur leur site, ils répondent à la question de comment est calculée leur cote, la réponse purement technique n'est pas dévoilée par question de confidentialité et de concurrence. Mais globalement ils utilisent une équipe d'experts qui recueillent beaucoup d'informations sur le terrain (professionnels ou particuliers), ils utilisent aussi leur longue expérience dans ce domaine. D'après l'article de A.Fruchard, les experts se baseraient aussi sur les variations du marché et les chiffres réels des transactions des modèles d'occasion. C'est donc leur ancienneté et donc une bonne connexion avec les professionnels de ce milieu

qui leur permet d'avoir toutes les informations nécessaires à l'évaluation du prix de revente des véhicules.

Si nous avons parlé de la cote à l'ARGUS, le journal propose en fait deux types de cote le « cours-moyen » : qui est donc le prix d'un véhicule d'occasion pour un état standard (aucune information à part le nom et l'année du modèle). Et la cote personnalisée prenant en compte les options comme le kilométrage. Cette deuxième cote est plus intéressante dans le cadre de notre étude nous réaliserons des côtes personnalisé grâce aux Data Science.

1.2 La Centrale : de la vente à la prédiction

Une autre référence dans le marché de l'occasion français : La Centrale. Ce site de véhicule d'occasion n'a pas forcément le plus grand nombre d'annonces avec environ 300 000 annonces. Mais contrairement à son plus grand concurrent Le Bon Coin (850 000 annonces), La Centrale est spécialisée uniquement dans l'automobile. C'est pourquoi elle propose de nombreux services liés à l'automobile comme proposer la cote des véhicules gratuitement y compris pour des cotes personnalisées grâce au fait qu'ils utilisent les données de leurs annonces pour la déterminer. Leurs estimations seraient donc principalement basées sur leurs historiques de ventes, précis pour des modèles communs mais plus compliqué pour des véhicules rares. Lors de notre étude nous effectuerons une démarche légèrement similaire (analyse de véhicules vendus et non de simples annonces) mais nous utiliserons plus de données.

1.3 Les autres services et entreprises du marché de la cote automobile

Alors que les précédentes entreprises citées ont bien sûr un site internet, il existe aujourd'hui plusieurs pages de recherche Google de sites offrant d'évaluer le prix de votre voiture et souvent gratuitement. Il y a Turbo, historiquement une émission française automobile qui a diversifié son activité en proposant eux aussi des cotes automobiles. Turbo et quelques autres sites (Auto Plus, Ouestfrance) sont donc reconnus pour leur expérience et souvent considérés comme une bonne référence.

Finalement d'autres sites proposent aussi ce service , ce qui les différencie principalement est leur expérience et leur crédibilité auprès des particuliers et des professionnels. Les techniques concrètes de ces entreprises ne sont bien sur jamais dévoilées, on peut juste connaitre les critères sur lesquels ils se basent où connaitre l'origine de leurs données. Dans cette étude notre objectif sera de déterminer la meilleure façon d'évaluer le prix de ces véhicules grâce au

Machine Learning. Etudions donc les recherches qui ont déjà été effectuées sur ce sujet

2. La Science de l'analyse de données et de la prédiction : Le Machine Learning

Le Machine Learning est un cas d'utilisation particulier de l'intelligence artificielle. L'utilité du Machine Learning est de pouvoir développer un modèle, basé sur des algorithmes, capable d'apprendre tous seuls grâce à des données fournies. L'apprentissage va se faire par la diversité des données recueillies et la détection d'un pattern (relation qui se répète) dans ces données. Dans le cas simple de la régression linéaire (prédiction de valeur numérique), le pattern en question est une droite linéaire par exemple. Il existe aujourd'hui de multiples applications aux Machine Learning : pour la prédiction de risques (crédit, aléas naturelles), de maladies, de prix ou pour la reconnaissance de texte, d'image permettant aujourd'hui le développement des véhicules autonomes.

2.1 Définition

Toutes ces applications sont concrètes et déjà utilisées dans l'industrie apportant une nouvelle manière de répondre aux problèmes classiques, selon M. Mitchell (1997) on peut diviser ces problèmes en 3 catégories : « Data Mining », « Difficult-to program applications » et « Customized software applications ».

La première catégorie « Data Mining » correspond au fait d'utiliser d'anciennes données pour en prédire de nouvelles, améliorant ainsi la prise de décision. Le risque de crédit est un problème de ce type où on analyse d'anciens comptes clients en sachant s'ils ont été solvables ou non (ont-ils remboursé leurs emprunts), le modèle va alors analyser les caractéristiques des clients pour dans le futur être capable de reconnaitre un client qui ne remboursera pas son prêt.

La deuxième catégorie « Difficult-to program applications », ici le Machine Learning est une solution au programme trop difficile à coder manuellement, comme pour la reconnaissance d'image ou de texte. En effet le Machine Learning a l'avantage de pouvoir apprendre sur des millions des données très rapidement, il existe aujourd'hui des modèles de reconnaissance de texte qui ont été entrainé sur des milliers de textes , leur permettant aujourd'hui de comprendre une langue entière et de faire des regroupements de mots par sujet.

Le dernier groupe « Customized software applications » fait référence aux applications de recommandations et leur capacité de configurations personnelles. Par exemple les navigateurs internet aujourd'hui connaissent vos préférences grâce aux données récoltées sur vos recherches (cookies), le but étant par exemple de faire de la publicité ciblée. Le modèle dresse alors un profil d'une personne et est capable de prédire quelles informations ou articles sont fait pour vous.

Plus classiquement on sépare le Machine Learning en deux grandes catégories : le supervisé et le non-supervisé. La différence majeure vient des données utilisées, dans le cas du supervisé les données contiennent la variable à prédire, le but est donc de retrouver une variable connue. Dans le cas du non-supervisé l'algorithme va essayer de deviner les réponses d'une certaine façon puisqu'il ne les verra jamais.

Dernière distinction importante dans le Machine Learning les types de problèmes, pour le cas de données supervisées on distingue la régression et la classification. La régression est le fait de vouloir prédire une variable continue (souvent numérique), et la classification consiste à prédire une catégorie (parmi 2 à plusieurs catégories). Le cas le plus simple de classification est la binaire : prédire deux catégories : oui ou non. Pour le Machine Learning non-supervisé on retrouve le clustering, il s'agit de la classification mais le modèle ne connait pas les catégories à prédire il va donc les créer lui-même par analyse des similarités des individus.

Le Machine Learning a donc beaucoup de différentes applications et peut servir dans de nombreux contextes, pour ce mémoire nous allons nous focaliser sur une tâche bien précise : la prédiction de biens, un problème supervisé de « Data Mining ».

2.2 Le Machine Learning pour évaluer la valeur de biens

La prédiction de valeurs pour des biens est un secteur en pleine croissance, des milieux très anciens commencent à utiliser ces techniques comme l'immobilier ou la vente d'actions. Il s'agit souvent de régression, le but étant de prédire un chiffre exact : le prix. Mais on peut aussi aborder le cas de la prédiction de valeurs par la classification avec minimum deux catégories : prix montant ou descendant. Nous allons voir dans cette partie comment le Machine est déjà utilisé ou a été étudié dans différents secteurs et comment les solutions ont été implémentées.

2.2.1 Dans l'immobilier

L'immobilier est un secteur propice au Machine Learning, pour plusieurs raisons. Les données se prêtent très facilement au « Data Mining » par les caractéristiques d'un bien immobilier qui sont souvent les mêmes, il est donc facile de comparer deux biens immobiliers. Deuxièmement car l'estimation de ces biens est très importante que ce soit pour les particuliers ou les professionnels, il s'agit d'un problème courant qui n'est pas prêt de disparaitre. Dernièrement l'immobilier est très important pour l'économie d'un pays, les demandes en logement augmentent constamment et c'est une nécessité vitale pour chaque personne. Nous analyserons donc les modèles utilisés ainsi que les conclusions de différents articles.

Zulkifley, Abdul Rahman, Ubaidullah et Ibrahim (2020) ont décidé d'analyser plusieurs études déjà réalisées sur l'analyse de prix dans l'immobilier dans le cadre d'un problème de régression (estimer le prix exact du bien). Leur but a été de comparer l'efficacité des modèles implémentés dans ces études, ainsi que les « features » ayant le plus d'impact sur l'estimation. Les « features » sont les caractéristiques du bien immobilier dans ce cas (nombres de m², voisinages, commerces à proximité, numéro d'étages, …). Dans leurs études ils séparent d'ailleurs les « features » en 4 catégories : localisation, structurel, voisinage (socio-économique) et économique (revenu). Ils ont révélé que les variables économiques et de voisinage était difficiles à évaluer et que donc les études déjà réalisées ne se basaient pas sur ces caractéristiques.

Le choix des « features » est une étape cruciale pour le développement d'un modèle, comme on vient de le voir si des données sont trop dures à évaluer il vaut mieux ne pas les prendre en compte car elles seront soit peu fiables (ne traduisent pas bien l'information souhaitée) ou dures à comparer car trop spécifiques. La sélection des features se fait toujours au début de manière souple (en premier lieu on ne veut pas supprimer trop d'informations) et au moment de l'optimisation du modèle (on supprime les features les moins corrélées).

Une étude de Singh, Sharme et Dubey (2019) nous montre justement (figure 1) comment ils optimisent leurs features lors de l'optimisation d'un de leurs meilleurs modèles : le gradient boosting.

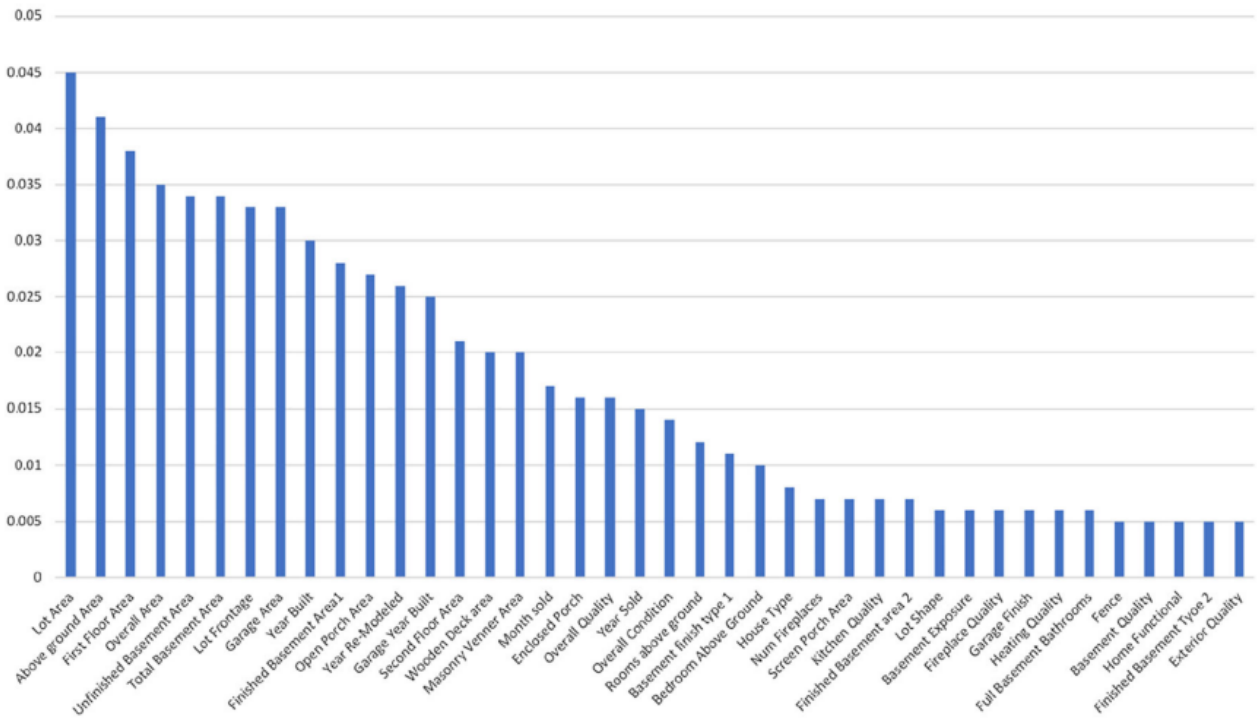


Figure 1: Selection des "features" en utilisant le gradient

On observe en ordonnées la corrélation qui correspond à l'importance des variables en abscisses. On peut voir par exemple que les variables de surface (totale, jardin, premier étage) du bien immobilier sont les caractéristiques qui impactent le plus le prix du bien correspondant à la catégorie structurelle. Ce schéma est souvent généré par l'analyse des corrélations ou des méthodes de sélection de « features » forward (ajouter un à un) ou backward (retirer un à un) visant à compiler le modèle avec 1 feature puis 2, 3 et ainsi de suite. La performance d'un modèle est souvent calculée comme le ratio de bonnes prédictions sur les mauvaises, c'est ce qui permet de comparer les modèles entre eux pour un même problème. Quand est-il des modèles les plus performants pour la prédiction du prix de l'immobilier ?

Dans l'étude de Zulkifley, Abdul Rahman, Ubaidullah et Ibrahim (2020), les chercheurs ont comparé les résultats de différentes recherches amenant aux résultats suivants (figure 2).

Models	Attributes	RMSE	Previous Study
Multiple Linear Regression	Structural attributes	0.1261	[38]
	Locational and structural attributes	-	[3]
	Locational and structural attributes	0.267	[39]
	Locational attributes	-	[40]
	Locational attributes	-	[10]
	Economic attributes	-	[41]
	Locational, structural and neighborhood attribute	-	[11]
	Locational and structural attributes	0.3079	[7]
Support Vector Regression	Locational and structural attributes	0.2362	[7]
	Locational and structural attributes	-	[22]
	Locational attributes	0.0047	[23]
Artificial Neural Network	Locational and structural attributes	0.5155	[7]
	Locational attributes	-	[32]
	Locational attribute	-	[40]
	Locational attributes	0.0581	[23]
XGBoost	Structural attribute	0.1212	[38]

Figure 2: Performance (RMSE) de différents modèles pour la prédiction du prix de biens immobiliers

Pour comprendre ce tableau, définissons le RMSE qui est un indicateur de performance des modèles, plus précisément le RMSE correspond à l'erreur quadratique moyenne. Il s'agit de la moyenne des carrés des écarts entre les prédictions et les observations, c'est donc un indicateur du taux d'erreur ; plus il est faible plus le modèle est précis donc performant. On peut donc voir ici que 3 modèles présentent de meilleurs résultats que les autres : Le SVR (Support Vector Regression), le ANN (Artificiale Neural Network) et le XGBoost (variante de l'arbre de décision). L'étude de Zulkifley, Abdul Rahman, Ubaidullah et Ibrahim (2020) révélait que les modèles de Random Forest (variante de l'arbre de décision) et le GBM (Gradient Boost Machine) avaient des RMSE proches de 0.1. Le GBM et le XGBoost étant des modèles très proches on peut déterminer que ces modèles sont efficaces pour ce genre de prédiction, nous les étudierons donc par la suite ainsi que l'Artificial Neural Network.

2.2.2 Autres secteurs

D'autres études ont été réalisées sur la prédiction de prix de revente dans différents secteurs, Fathalla, Salah, Li, Li et Francesco ont étudiés les objets d'occasion dans le e-commerce. Pour eux le marché de l'occasion est un secteur en plein essor, aujourd'hui les sites de ventes d'occasions sont nombreux et l'estimation du prix des objets est un problème

crucial pour les particuliers qui ont du mal à estimer la valeur de leurs biens. Dans cette étude ils ont mis au point deux modèles : le premier basé uniquement sur la description et l'analyse des objets est un modèle hybride : le CNN-LSTM (Long Short-Term Memory Networks). Le modèle présente de bons résultats avec un MAE (Erreur Absolue Moyenne) de 0.07 ce qui est faible et se traduit par très peu d'erreurs dans les prédictions. Un deuxième modèle extradant le prix maximal et minimal de l'objet. La combinaison des deux modèles (prix moyen et prix déterminé grâce aux caractéristiques de l'objet) est très pertinente et montre même de bons résultats. Leur implémentation du LSTM combinant ces deux modèles a montré de très bons résultats avec un MAE de 0.02. Cette étude est très intéressante puisqu'elle combine deux aspects du Machine Learning pour améliorer la performance finale, cette séparation du problème en un problème de « Data Mining » (évaluer le prix moyen avec le min et le max) et de « Difficult-to program applications » (reconnaissance d'image et de texte) est très pertinente. En effet il serait dur voire impossible de demander à un seul modèle d'effectuer toutes ces prédictions, la répartition des tâches est donc ici indispensable. L'utilisation de reconnaissance d'image pour l'évaluation de biens est aussi innovante et leur permet de traiter plus de données, attention cependant à la pertinence de celle-ci car les photos peuvent être trompeuses.

D'autres secteurs d'activités sont sujets à la prédiction au Machine Learning comme la prédiction du prix d'une action (étude de Long, Lu et Ciu (2019) , d'une crypto monnaie ou encore du prix de certaines ressources comme l'essence ou des huiles. Dès lors qu'il existe un historique de données où l'on veut prédire une variable (ici un prix), on peut utiliser le Machine Learning de manière pertinente. Dans l'étude de Chen, Li et Sun (2019) les chercheurs ont réussi à prédire le prix du bitcoin grâce à des modèles de Machine Learning (SVM, RNN). Cette technique s'est révélée bien plus efficace que les méthodes traditionnelles d'analyses statistiques lorsque les données sont à « haute fréquence ». C'est-à-dire que les données utilisées par le modèle sont rafraîchies (ajout de nouvelles données) bien plus souvent, augmentant alors énormément le nombres de données. C'est en effet l'avantage du Machine Learning il peut apprendre sur un nombre infini de données tant qu'elles gardent les mêmes features. Détail intéressant, les chercheurs ont établi une prédiction toutes les 5 minutes, l'échelle du temps étant très importante pour les crypto-monnaies. Ce détail moins pertinent pour l'automobile reste une bonne question à se poser lors de la conception du modèle et la collecte des données.

2.2.3 Dans l'automobile

Le cas de l'automobile a été sujet à plusieurs études qui différent souvent par la localisation de l'étude et des techniques employées. Nous allons voir ici quels modèles ont été les plus performants et quels ont été les conclusions sur l'importance des features d'un véhicule.

Une étude réalisée par Monburinon, Rungpheung, … (2018) sur le marché de l'occasion allemand a révélé une très bonne performance du modèle GBM avec une erreur absolue moyenne de 0.28 suivi d'un Random Forest avec 0.35. Encore une fois le modèle GBM a fourni les meilleures estimations de prix, nous détaillerons ce modèle dans la partie suivante. Dans leur étude les chercheurs n'ont malheureusement pas cherché à connaitre les features les plus corrélés, nous savons cependant que leurs données contenaient plusieurs informations comme le prix, le modèle ou la marque et d'autres caractéristiques techniques du véhicules (type de boite, puissance, type de carburant, …)

Dans un article de Pudaruth (2014) c'est le marché de Maurice (l'île Maurice et l'île Rodrigues à l'est de Madagascar). Dans cette étude l'auteur traite donc de l'estimation du prix des voitures d'occasion dans ce marché local en testant 4 modèles . C'est un marché spécifique car ils ne produisent pas de voitures, les véhicules d'occasion sont quasiment tous importés d'autres pays et reconditionnés. Une limitation d'une telle étude est le nombre de données disponibles, leur datasets n'étaient constitués que d'un peu plus de 400 véhicules amenant donc peu de diversité. Mais cette étude est très intéressante notamment par sa démarche : la classification. C'est la première étude qui privilégie la classification (figure 3) à la régression sûrement à cause du faible nombre de données.

#	Minimum	Maximum	Category
1	95000	168000	A
2	180000	245000	B
3	260000	325000	C
4	335000	410000	D
5	425000	450000	E
6	550000	625000	F

Figure 3: Groupe créé pour la classification

Comme le présente le tableau, le chercheur a décidé de classer les véhicules en 6 groupes par rapport au prix, chaque groupe correspondant donc à une tranche de tarification. Le but de la prédiction est donc de savoir à quel groupe appartient le véhicule afin d'avoir une estimation du prix. Certes cette technique est en définitive moins précise puisqu'on obtient plusieurs tranches de valeurs, mais c'est un bon moyen de contourner le fait qu'il y a peu de données. Autres astuces pour s'adapter à ce problème : séparer les données selon une caractéristique, ici la marque du véhicule. L'étude ne montre pas de modèle particulièrement plus efficace l'un que l'autre, mais nous apprend beaucoup sur comment surmonter des problèmes de manque de données ou de reformulation du problème de régression en classification.

Une autre étude réalisée en Bosnie et Herzégovine par Gegic (2019) a montré de très bons résultats à la suite d'une recherche approfondie et beaucoup de tests. Il traitait le problème par classification (tranches de prix) puisque leurs données ne contenaient pas de prix exact mais uniquement des fourchettes, la régression était alors impossible. Leur classification contenait 13 groupes de « prix ». Après avoir utilisé les modèles classiques (Random Forest, ANN et SVM) un à un ils ont déterminé que seul un modèle ne pourrait pas être assez efficace (précision inférieure à 50%). Ils ont donc combiné deux modèles : le Random Forest pour reclassifier les véhicules en 3 groupes seulement (cheap, moderate et expensive) correspondant à des fourchettes plus grandes. Le Random Forest a donc servi à s'assurer de la bonne classification des données pour ces 3 groupes, puis ils ont appliqué le SVM et le ANN sur ces données. Les résultats sont très corrects puisqu'ils ont obtenu 90.48 % de précision avec le SVM et 85.71 % avec le ANN. Cette étude est la plus pertinente puisqu'elle montre de bons résultats et une démarche intéressante où les chercheurs ont encore une fois su décomposer le problème de régression ce coup-ci en deux tâches de classification.

2.3 Les différents modèles utilisés dans la prédiction de prix

Dans la partie précédente nous avons vu de nombreux modèles, dans cette partie nous allons détailler les plus performants.

Le GBM (Gradient Boosting Machine) peut servir pour la régression et la classification, variante de l'arbre de décision. Ce modèle se différencie par l'implémentation de plusieurs arbres dépendants ; les arbres sont créés un à un et chaque nouvel arbre doit corriger le précèdent.

L'ANN (Artificial Neural Network) est un réseau de neurones classiques, ces modèles ont été copiés sur le fonctionnement des neurones humains. Il fait partie des modèles de Deep learning souvent utilisés dans la reconnaissance d'image. L'ANN est le cas le plus basique du Deep Learning et se constitue de 3 couches : entrées, sorties et le « caché » qui réalise le traitement de la donnée. D'autres réseaux de neurones utilisent plusieurs couches « cachées ». Ces types de modèles sont principalement pour la classification.

Le SVM (Machine à Vecteurs de Support) permet de traiter aussi bien la régression que la classification. Son fonctionnement se fait par délimitation, son but étant de séparer les données en classes, en établissant des frontières. C'est un modèle souvent performant. Il existe le SVR pour les cas de régression.

Nous avons vu rapidement 3 des modèles les plus efficaces dans les articles étudiés, il en existe d'autres que nous testeront dans la partie empirique de ce mémoire.

3. L'automobile : un secteur influençable

Comme énoncé en introduction, l'automobile est sujet à pleins d'événements extérieurs qui viennent modifier le comportement du marché. Nous allons voir dans cette partie quels sont les différents secteurs qui peuvent impacter le marché de l'automobile.

3.1 Par les événements culturels

Que ce soit sportif ou cinématographique, les marques de voitures ont toujours essayé de se faire de la publicité. On peut penser au développement des 4 roues motrices, qui inventé début 1980 par Audi, n'était pas si démocratisé. C'est grâce au rallye notamment au fameux groupe B, qui était tellement spectaculaire et si dangereux qu'il a été banni en 1986 (4 ans après son commencement), qu'Audi a fait connaitre sa technologie au monde entier. Si performant en rallye, les 4 roues motrices ont été adoptées par tous les concurrents au fil du temps. Cet événement à fait connaitre à jamais l'Audi Quattro. Toujours dans le rallye on peut parler de Lancia qui jouissait d'une réputation prestigieuse grâce à leurs performances en rallye, c'était l'âge d'or de Lancia. Aujourd'hui Lancia est une marque éteinte qui produit peu de véhicules. Dans le cinéma maintenant, la saga « Fast and Furious », dans chacun des épisodes, la licence avait des partenariats pour tous les types de véhicules (Dodge a fourni les 4x4 et voitures de police dans le 4ème volet de la série). Ou encore dans le deuxième film, la voiture héroïne, une Mitsubishi Lancer, la voiture n'était même pas encore sortie, faisant alors exploser les précommandes. Pleins d'autres films (« 60 secondes chrono », « Bullit ») ou

sport (Mans, Formule 1) ont servi à promouvoir des marques automobiles, mais ce sont ces marques qui ont pensé à ces stratégies la plupart du temps ou se sont juste servis de ces évènements pour promouvoir leurs dernières avancées.

3.2 Les différents marchés à travers le monde

Comme nous l'avons vu dans les précédentes études, chaque pays a un marché particulier. Ils sont souvent différents par leur taille, le nombres de marques présentes ou encore selon le niveau de développement du pays. Les exportations automobiles ne sont pas des cas généraux et concernent souvent des passionnées ou des professionnels. Les entreprises nationales impactent aussi énormément le marché, en France par exemple plus de 50 % des véhicules sont fabriqués par des entreprises françaises. Une autre grande différence est sur les législations.

3.3 Quand la législation conditionne le futur automobile

Que ce soit dans pays à un autre, ou d'une époque à une autre les législations automobiles changent souvent. Ces lois impactent souvent le marché que ce soit des véhicules neufs ou d'occasion. Par exemple le diesel a longtemps été promu par l'état pour ses vertus écologiques, aujourd'hui l'idée serait plutôt de ne plus avoir de véhicules diesel pour les mêmes raisons. Cela va forcément faire baisser le prix de revente des voitures diesel, on n'en est pas sûr aujourd'hui, mais il existe déjà des études sur la prolongation de vie de ces véhicules grâce au rétrofit électrique. L'avenir du diesel n'est donc pas assuré. Les malus écologiques, c'est une norme aujourd'hui en France : plus votre véhicule pollue plus vous allez payer de malus, l'inverse étant vrai pour des véhicules hybrides ou électriques : vous pouvez gagner de l'argent. Cette norme incite clairement les particuliers à s'orienter vers des véhicules les moins polluants possibles, à contrario si vous souhaitez une grosse cylindrée il va falloir l'assumer financièrement, ce qui peut causer une perte de valeur chez ces véhicules. Autre législation importante : l'importation ! Suivant les pays cela diffère complétement, en France on paye une taxe (carte grise) qui augmente en fonction de la cylindrée du véhicule. Il y a aussi le cas très particulier de la Turquie, où l'importation de voitures d'occasion est tout simplement interdite, le marché étant donc local et fermé, les voitures d'occasions coûtent alors aussi cher que les neuves. Les voitures électriques sont plus susceptibles de garder de la valeur c'est ce que prouve Richardson (2009) dans leurs recherches.

III) Recherche méthodologique

1. Approche méthodologique

Pour mener à bien ce mémoire nous allons suivre une démarche de Data Scientist puisqu'il s'agit d'un problème de Machine Learning. Il s'agit généralement de 10 étapes, nous en suivrons principalement 9 : Comprendre le problème et traduire en problème de Machine Learning, analyser l'approche, établir quelles données sont nécessaires, collecter les données, comprendre les données, préparer et nettoyer les données, créer les modèles, évaluer les modèles et résultats et enfin conclure et faire les recommandations. Chacune des ces étapes est nécessaire au bon déroulement de ce mémoire afin de s'assurer de la qualité et de la pertinence du résultat.

La compréhension du problème et sa traduction est une étape très préliminaire, elle est partiellement effectuée lors de la sélection du sujet, en effet en choisissant la problématique de comment évaluer le prix de revente d'un véhicule d'occasion, nous savions déjà que la réponse se trouverait dans les annonces et historiques de ventes. Il nous faudrait aussi développer une technique permettant d'absorber les fiches techniques d'une voiture, ou du moins un maximum d'informations et de prédire une des caractéristiques de la voiture : son prix de vente. Le but ici est donc de trouver quelles caractéristiques de la voiture influencent le prix de revente de celle-ci, il peut aussi s'agir de facteurs annexes (culturels, législatifs) et d'être capable de faire une évaluation la plus précise possible.

L'analyse de l'approche est complémentaire à la première étape. Le problème étant l'analyse de données automobiles et la prédiction, nous avons donc choisi le Machine Learning supervisé, solution assez naturelle pour ce genre de problème. En effet la fonction principale du Machine Learning est de créer et entrainer des modèles mathématiques pour prédire une donnée. Il s'agirait soit de régression (prédire le prix exact) soit de classification (1 ou 0) si la cote est montante ou descendante. Le Machine Learning supervisé est le fait d'utiliser des données comportant la variable que l'on veut prédire, l'entrainement du modèle se fait donc sans les réponses puis lors de l'évaluation des performances on compare les prédictions aux réponses (ici prix des véhicules).

Pour ce qui est de choisir les données nous savons déjà qu'il nous faut des annonces de véhicules d'occasion ; plus les annonces seront détaillées, meilleur sera l'analyse. Le prix est indispensable ainsi que le modèle et la marque du véhicule. Les données devront être françaises car l'étude ne concerne que le territoire français. Si possible les données devront être étalées sur plusieurs années et concerner majoritairement des véhicules qui ont déjà été vendus. La collection des données se fera donc par le « scrapping » de sites spécialisés comme La Centrale ou Leboncoin, nous scrapperons à intervalles régulières afin de connaitre les ventes. Si ces données sont trop protégées alors nous utiliserons des « datasets » libres de droits issus de compétition Kaggle ou de banques de données.

La préparation des données est une étape cruciale et la plus chronophage, il s'agit ici de nettoyer toutes les données : regarder les valeurs manquantes et les remplacer, chercher les valeurs aberrantes. Mais aussi de faire du feature engineering : créer de nouvelles données ou en supprimer. Observer les corrélations entre les variables pour sélectionner les meilleures features. L'objectif est d'obtenir un dataset « propre » directement utilisable pour l'élaboration du modèle.

Modeling : il va s'agir ici d'implémenter plusieurs modèles sur les données, on peut alors utiliser la cross validation pour tester plusieurs modèles rapidement. On va régler les hyperparamètres (paramètres d'un modèle) au mieux.

Avant-dernière étape complémentaire au modeling, l'évaluation des modèles se fait par rapport aux prédictions et aux observations (valeurs justes fournies par le dataset). Il existe plusieurs indicateurs : la précision, le recall, l'AUC (Area Under the Curve) ou encore les moyennes d'erreurs (MAE et RMSE). Tous ces indicateurs sont pertinents, certains le sont plus dans certains cas comme l'AUC pour les cas de classification binaire.

Dernière partie la conclusion et les recommandations, on va ici valider quel modèle est le plus efficace et pourquoi, analyser l'importance des features dans la prédiction afin de savoir quelles caractéristiques de la voiture influencent le plus le prix.

Si nous pouvons et si c'est pertinent nous déploierons le modèle sur une API par exemple pour qu'il soit accessible et interrogeable en ligne.

2. Présentation du contexte de la question

2.1 Rappel du contexte

Pour rappel le sujet de ce mémoire est la prédiction de prix de véhicules d'occasion. L'objectif final est d'arriver à isoler des features (caractéristiques du véhicule) qui influencent le prix de revente de ce dernier. Dans un second temps nous développerons un modèle de Machine Learning capable de prédire avec une bonne précision les prix du véhicule avec comme données en entrée comme l'année, le nombre de kilomètres , la puissance ou encore la marque et le modèle …. Puis en complémentarité du modèle nous ferons une API (local) qui nous permettra de tester le modèle sur des annonces récentes et d'établir une comparaison entre le prix prédit et attendu.

L'enjeux de ce travail de recherche est à la fois économique, puisque nous cherchons à déterminer des facteurs influençant un prix de revente mais aussi industrielle puisque le marché automobile de l'occasion suit les besoins des clients. Nous pourrions ainsi découvrir sur quels critères les acheteurs particuliers font leur sélection.

Si la portée de cette étude est destinée à la France vous verrez plus tard que certaines données ont dû être récupérés en Allemagne, car c'était le marché le plus « voisin » et il y avait beaucoup de données disponibles . En effet beaucoup des sites d'annonces d'occasion protègent leurs données et en France il était impossible de récupérer des données. L'étude sera donc étendue au marché Français/Allemand mais nous privilégierons une analyse du marché Français. Pour éviter tout biais nous mettrons en avant les différences entre le marché Français et Allemand comme le fait qu'il y a des routes sans limite de vitesse en Allemagne favorisant l'achat de grosses cylindrées (BMW, Mercedes, Porsche). D'un autre côté la France elle, privilégie plus l'écologie par exemple, Renault a très récemment annoncé qu'il allait mettre une limite de vitesse de 180 km/h sur toute la gamme Renault et Dacia.
Ce comportement n'est pas encore expliqué clairement, mais on le sait : réduire la vitesse est synonyme de réduction de la cylindrée du véhicule et donc de la consommation. De plus Renault est très fortement soutenu par l'état Français financièrement et la marque doit parfois suivre les indications de l'état, comme une obligation d'investir fortement dans l'électrique depuis maintenant quelques années. Qui d'ailleurs se remarque par les ventes de véhicules électriques de Renault : Zoé est actuellement première des ventes en Europe !

En conclusion le contexte global de ce mémoire est le marché automobile Français et Allemand. On s'intéressera aux caractéristiques techniques du véhicule mais aussi des notions plus globales ou personnelles comme la marque ou l'esthétique de la voiture.

2.2 Présentation des données et sources

Pour réaliser ce mémoire nous avons collecté deux types de données bien différentes : des données de « web scrapping » et un sondage. Le web scrapping consiste à émettre des requêtes automatiquement (comme lorsque l'on fait une recherche dans un navigateur internet) et récupérer le contenu des pages le plus souvent sous forme de tableau. Cette technique demande des connaissances dans la structuration des pages web et comment y sont rangées les données. Cette discipline très courante, a malheureusement quelques défauts sur lesquelles nous reviendrons.

2.2.1 Le sondage

Le sondage a été réalisé sur un public français diversifié et de préférence sur des personnes concernées (ayant le permis et ayant acheté un véhicule). Il s'agit d'une étude surtout quantitative avec 150 répondants issus de mon réseau personnel ou du partage du sondage sur internet. Le sondage avait la forme d'un Google doc facilement partageable. La plupart des questions étaient fermées et le questionnaire durait environ 5 minutes. Le sondage visait à connaître le comportement et les préférences des particuliers par rapport à l'achat d'un véhicule. Dans la constitution de mon panel j'ai souhaité ne viser que des adultes et de préférence supérieur à 25 ans (plus la personne est âgée plus elle a de connaissances et un avis mâture sur la question d'achat d'un véhicule) , ce qui n'étaient pas très facile puisque la grosse partie de mon réseau et des personnes les plus réactives à ce genre de requête sont les jeunes. Mais je remercie mon réseau familial qui a pu partager le sondage a beaucoup de leurs connaissances alors bien plus âgées et pertinentes pour le sondage. La répartition des âges chez les répondants a été plutôt réussi par rapport à l'objectif d'avoir plus de personnes au-dessus de 25 ans. Comme on peut le voir dans la figure 4 ci-dessous : 67 % des répondants ont plus de 25 ans et on a aussi 14.3 % des répondants qui ont plus de 60 ans et donc toute une vie d'expérience dans l'achat de véhicules.

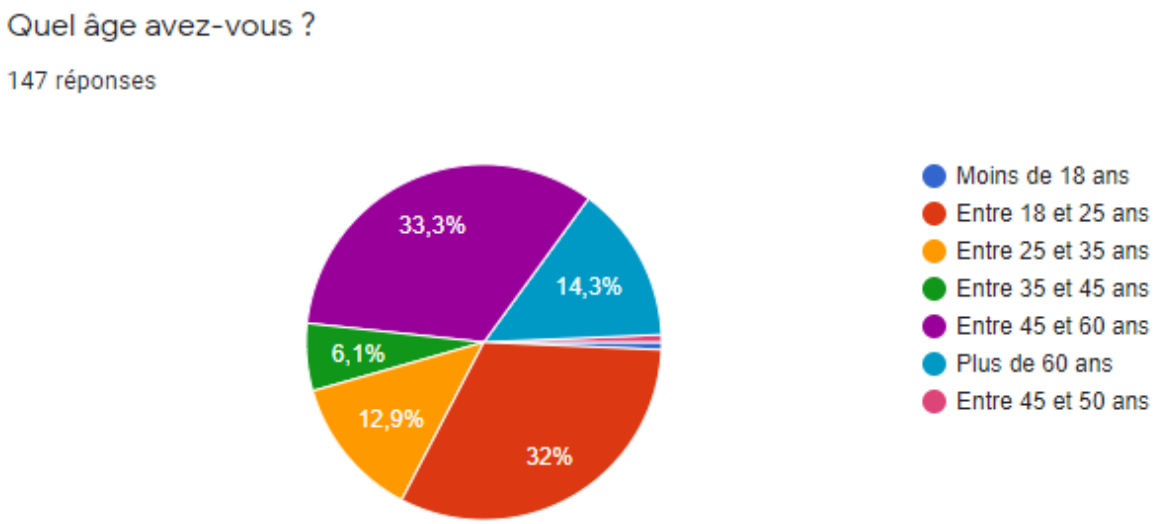


Figure 4: Répartition des âges chez les répondants

En complémentarité j'ai aussi partagé mon sondage sur des groupes Facebook spécialisés dans l'achat de véhicules d'occasion, dans des commentaires de vidéos YouTube de chaine de passionnés d'automobile, dans des clubs privés de voitures (à leurs membres) ainsi que chez des personnes du milieu de l'automobile travaillant chez Renault.
Le panel interrogé constitue un échantillon très pertinent de la population (voir figure 5 ci-dessous), en effet 50 % des répondants sont intéressés ou même passionnés d'automobile. Il y a aussi 23 % de passionnés ce qui représente 30 personnes, ce qui est très satisfaisant pour une étude sur l'automobile. Plus les personnes sont intéressées, plus leur avis va être nuancé par leurs expériences et multiples connaissances accumulées sur le sujet

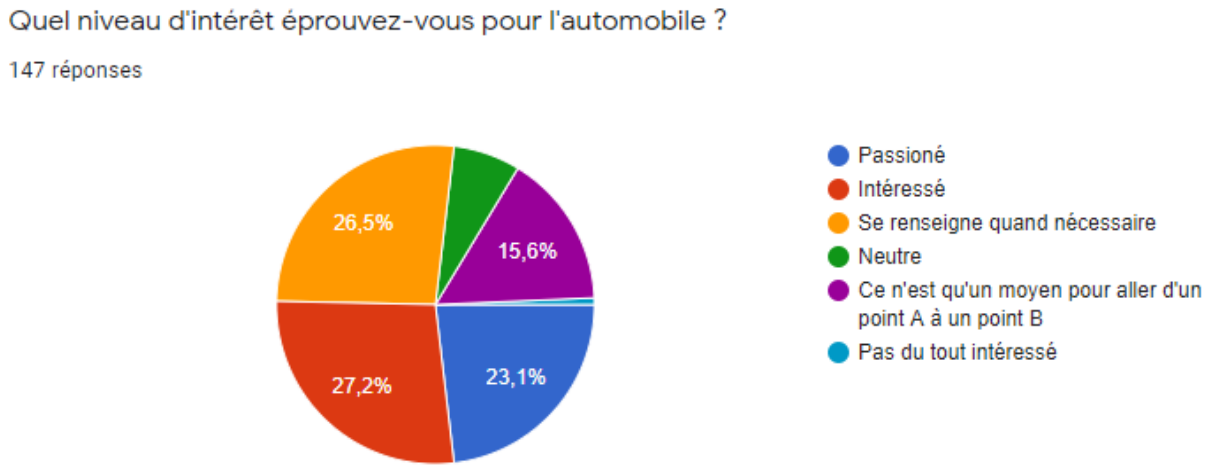


Figure 5: Intérêt porté à l'automobile chez les répondants

Cibler la population est très important dans la création d'un sondage comme le souligne Lugen, M dans son guide sur les enquêtes. Dans ce texte l'auteur rappelle deux éléments clés dans la sélection de la population :
La possibilité que l'individu apporte de l'information, dans mon cas n'importe quel adulte ayant le permis (91 % des répondants) est concerné par l'achat d'un véhicule sauf ceux qui n'en n'ont pas (19 % des répondants seulement).
Et est-ce qu'il est nécessaire que l'individu ai un bon niveau d'informations, dans mon cas non car je ne vise pas les professionnels de l'occasion mais bien des particuliers. Mais plus le particulier a acheté de véhicule (76,2 % des répondants ont déjà) plus son avis est intéressant, de même s'il éprouve de l'intérêt pour l'automobile.

Quelques redressements ont été effectués sur les questions mais peu, quand un individu répondait dans la catégorie 'autre' une réponse pertinente nous les avons rajoutés au choix des questions. Comme le carburant hydrogène ou certains facteurs extérieurs pouvant influencer leur décision (guide YouTube par exemple).

En conclusion nous allons voir ensemble les grands types de questions qui ont été posées lors du sondage : 7 questions fermées sur le profil : âge, sexe, a le permis ou non, intérêt pour l'automobile, possède un véhicule, nombre de véhicules neufs et d'occasions achetés). Puis 7 autres questions sur leurs comportements et préférences dans l'achat d'un véhicule : leur critère primordial, leurs 4 plus gros critère, préfèrent-ils acheter neuf ou d'occasion et où (concession, garage, internet, …) ? , leur préférence pour ce qui est de la motorisation (type), l'importance de la marque du véhicule et enfin si certains facteurs extérieurs peuvent influencer leur avis. Le sondage a donc pour but premier de dresser un ordre d'importance parmi les différents facteurs qui rentrent en compte dans l'achat d'une voiture. Vous pourrez regarder tous les résultats du sondage dans la première annexe.

2.2.2 : Les données issues du web scrapping

Ces données ont été récoltées pour réaliser les tâches de Machine Learning, plus précisément de prédiction de prix. Dans un premier temps je me suis dirigé vers les sites de revente Français détaillé dans la revue de littérature (La Centrale, L'Argus et Le bon coin).

Pour L'ARGUS ils n'ont pas vraiment de sites avec des annonces de véhicules, mais ils ont une API qui est capable de donner leur cote, tous les modèles existants à ce jour et leurs caractéristiques. Malheureusement j'ai effectué une demande sur cette API par mail, car

il faut une clé d'accès mais je n'ai jamais eu de réponse et de toute manière nous avons besoin d'annonces avec un prix. Leur cote aurait pu me servir de « features » supplémentaires ou de référence.

La Centrale était la meilleure option, il y a beaucoup d'annonces et elles sont bien détaillées et souvent pertinentes par rapport à Leboncoin ou sont postées toutes sortes d'annonces peu intéressantes (véhicules endommagés voir inutilisables). En effet La Centrale est une plateforme plus « sérieuse » et professionnelle et surtout elle est dédiée aux voitures. Cependant après recherche sur le web et plusieurs essais de récupération de données sur leur site je me suis aperçu qu'ils renvoient de fausses données. C'est-à-dire que les données que j'ai récupérées ne correspondent pas avec les annonces que je vois dans mon navigateur, j'ai pu retrouver cette conclusion sur plusieurs forums. Le site est donc protégé contre le web scrapping par le fait de changer les données renvoyées aux utilisateurs repérés comme un programme automatique, car dans le scrapping on simule beaucoup de requêtes, ce qui n'est pas humain comme comportement (équivalent de faire 50 recherches Google en quelques secondes). Il existe des parades sur lesquelles nous reviendrons dans la fin de cette partie.

Leboncoin, tout comme La Centrale est protégé mais ici c'est par une entreprise de protection des données : Datadome. Ce service identifie 100 % des bots et les rejette, le résultat de mon scrapping fut un message d'erreur disant que le contenu était protégé.

Alors comment contourner ces protections ? Sur les forums les réponses sont divisées : certains n'y arrivent pas , d'autres gardent secrètement leurs techniques car plus elles seront partagées plus Datadome va la repérer et fermer la faille. Certains utilisent des solutions qui ne marchent déjà plus et donc ne marcheront plus jamais. Les seules solutions viables sont l'utilisation de service de scrapping payant, un service de VPN amélioré (aussi payant) qui va changer votre IP afin que toutes vos requêtes aient l'air de venir d'ordinateurs différents et quelques logiciels de scrapping (moins automatisé puisque on sélectionne les pages à la main). Le fonctionnement de ces services de scrapping est simple : charger des pages, récupérer les résultats des requêtes dans le « network de la page » et le puis le convertir en json. Cette solution a marché pour Leboncoin mais est beaucoup trop longue et fastidieuse si l'on veut beaucoup de données.

Alors qu'avons-nous fait ? Aucun des principaux sites en France de revente étaient protégés et de plus en voyant toutes ces protections nous nous sommes demandé si le scrapping de ces sites était légal ? Car ces entreprises vivent aujourd'hui de leurs données, leurs estimations et leurs cotes sont basées dessus. Elles donnent aussi une crédibilité auprès de leurs clients. De plus il est facilement pensable qu'ils fassent de l'analyse de données dessus, leurs résultats leur permettant sûrement d'augmenter leur bénéfice par rapport à leurs estimations. Ainsi que de proposer de nouveaux services ou renforcer d'autres afin d'attirer de nouveaux clients ou rester compétitif. De plus récupérer leurs données permettrait de proposer la même qualité de service qu'eux et de pouvoir copier leurs algorithmes d'estimations.

Nous avons donc renoncé au web scrapping que ce soit à cause des protections qui nous ont empêché de web scrapper et puis à cause de notre réflexion sur la légalité de cette action. Afin d'éviter tous problèmes nous nous sommes tournés vers des datasets déjà existants. Dans notre recherche nous n'avons pas trouvé de jeux de données avec des annonces de ventes de véhicules en France mais seulement aux Etats-Unis (où les données sont plus libres d'accès) et un en Allemagne. Pour des raisons de cohérence j'ai privilégié l'Allemagne.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
exploration_date	371528	280500	2016-03-24 14:49:47	7	NaN	NaN	NaN	NaN	NaN	NaN	NaN
name	371528	233531	Ford_Fiesta	657	NaN	NaN	NaN	NaN	NaN	NaN	NaN
seller	371528	2	privat	371525	NaN	NaN	NaN	NaN	NaN	NaN	NaN
offer_type	371528	2	Angebot	371516	NaN	NaN	NaN	NaN	NaN	NaN	NaN
price	371528.0	NaN	NaN	NaN	17295.141865	3587953.74441	0.0	1150.0	2950.0	7200.0	2147483647.0
ab_test	371528	2	test	192585	NaN	NaN	NaN	NaN	NaN	NaN	NaN
vehicle_type	333659	8	limousine	95894	NaN	NaN	NaN	NaN	NaN	NaN	NaN
registration_year	371528.0	NaN	NaN	NaN	2004.577997	92.866598	1000.0	1999.0	2003.0	2008.0	9999.0
gearbox	351319	2	manuell	274214	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Power_HP	371528.0	NaN	NaN	NaN	115.549477	192.139578	0.0	70.0	105.0	150.0	20000.0
model	351044	251	golf	30070	NaN	NaN	NaN	NaN	NaN	NaN	NaN
kilometer	371528.0	NaN	NaN	NaN	125618.688228	40112.337051	5000.0	125000.0	150000.0	150000.0	150000.0
registration_month	371528.0	NaN	NaN	NaN	5.734445	3.712412	0.0	3.0	6.0	9.0	12.0
fuel_type	338142	7	benzin	223857	NaN	NaN	NaN	NaN	NaN	NaN	NaN
brand	371528	40	volkswagen	79640	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Damaged	299468	2	nein	263182	NaN	NaN	NaN	NaN	NaN	NaN	NaN
created_date	371528	114	2016-04-03 00:00:00	14450	NaN	NaN	NaN	NaN	NaN	NaN	NaN
nb_pictures	371528.0	NaN	NaN	NaN	0.0	0.0	0.0	0.0	0.0	0.0	0.0
postal_code	371528.0	NaN	NaN	NaN	50820.66764	25799.08247	1067.0	30459.0	49610.0	71546.0	99998.0
last_seen	371528	182806	2016-04-06 13:45:54	17	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 6: Description des colonnes (features) du dataset

Le dataset allemand était présent sur Kaggle (il a été supprimé de la plateforme récemment). Il contient 370 000 lignes d'annonces provenant de EBAY en Allemagne. Les données ont été récupérées en 2016 et comportent 20 colonnes (figure 6, ci-dessus). Parmi ces features on retrouve entre autres le modèle, la marque, le kilométrage, la puissance, le type de carburant ou encore la classe du véhicule (citadine, berline, …). Vous pouvez retrouver ce dataset sur le site : https://data.world/data-society/used-cars-data . Ces données proviennent du scrapping mais ont déjà été utilisées dans une compétition Kaggle (site référence de compétition de Machine Learning). Nous reviendrons plus en détail sur l'analyse de ces données dans la prochaine partie.

3. Procédure de l'analyse des données

Nous allons voir maintenant comment nous allons analyser et faire les pré-traitements sur nos données pour le sondage et notre dataset qui est par ailleurs notre source principale de données dans ce travail de recherche.

3.1 Sondage : méthode d'analyse

Grâce au Google form toutes les réponses de notre sondage sont disponibles sous forme de graphique et aussi en format Excel (table où chaque ligne est la réponse d'un individu). Les graphiques nous permettent déjà d'analyser la répartition des différentes réponses, notre but étant pour les questions sur les critères de sélection, de déterminer quels sont les facteurs qui reviennent le plus. Les profils des répondants sont aussi répartis en pourcentage comme leur âge ou s'ils ont le permis ou non. Les graphiques fournissent par Google sont suffisants à l'analyse des réponses (comme nous le verrons dans la présentation des résultats). Une autre méthode plus poussée aurait été de filtrer les réponses par exemple en fonction de l'âge, ainsi nous aurions pu avoir les préférences de chaque tranche d'âge. Mais dans notre cas nous n'avons que 150 répondants, ce qui ferait de petite population à chaque fois et en plus d'autres filtres pourraient être appliqués comme leur niveau d'intérêt pour l'automobile. Voulant rester général et garder le plus grand échantillon possible nous analyserons les résultats du sondage sans filtrer les réponses.

3.2 Jeux de données (datasets) : méthode d'analyse

Dans le Machine Learning le but est de créer un modèle à partir d'une grande table de données. Mais la tâche la plus importante et chronophage n'est pas la création des modèles ou encore l'analyse des résultats, c'est la préparation des données. En effet les données « non-propres » feront obtenir des résultats biaisés ou erronés, de plus beaucoup de modèles surtout dans le cas d'une régression n'acceptent pas de données autres que des chiffres décimaux. C'est ce que souligne Jason W. Osborne dans son article sur le nettoyage des données, en effet selon lui lors des études quantitatives, on a tendance à considérer trop vite les données comme pertinentes et alors ces analyses sont prises comme référence alors qu'elles sont basées sur des données peut-être erronées.

3.2.1 Compréhension des données

Dans un premier temps nous allons observer quelles features (colonnes) sont présentes dans le dataset et nous allons déterminer si elles sont pertinentes ou non (figure 7, page d'après).

dateCrawled	Date à laquelle l'annonce a été récupérée
name	Nom de l'annonce : contient des informations comme le modèle, la cylindrée. Mais tous les noms ne sont pas régularisés (pattern différent) ! Chaque utilisateur a pu rentrer ce qu'il voulait et tous les noms n'ont ni le même format ni les mêmes informations.
seller	Type de vendeur : particulier ou professionnel
offerType	Type d'offre : offre ou demande
price	Prix : variable à prédire (cible)
abtest	Variable de EBAY
vehicleType	Type du véhicule (coupe, SUV, citadine, berline, cabriolet, camionnette, bus)
yearOfRegistration	Année à laquelle le véhicule a été immatriculé
gearbox	Type de transmission (automatique ou manuelle)
powerPS	Puissance en chevaux
model	Modèle
kilometer	Kilométrage
monthOfRegistration	Mois auquel le véhicule a été immatriculé
fuelType	Type de carburant (essence, diesel, hybride, électrique, CNG, LPG)
brand	Marque
notRepairedDamage	Est-ce que le véhicule a déjà été endommagé
dateCreated	Date à laquelle l'annonce a été créée
nrOfPictures	Nombres de photos disponibles
postalCode	Code postal
lastSeenOnline	Dernière date à laquelle l'annonce a été vue en ligne

Figure 7: Description et explication des colonnes

Nous avons donc 20 colonnes (features), la plupart sont très intéressantes pour notre étude, il est a noté qu'avoir beaucoup de features est souvent bien mais nous allons voir dans notre cas qu'ici 20 features est amplement suffisant.

3.2.2 Nettoyage du jeu de données et imputation des données

L'enjeux du nettoyage de données est crucial pour la suite, car bon nombre de modèles n'acceptent pas les valeurs manquantes et les valeurs aberrantes pourraient venir fausser nos prédictions. Nous allons décrire ici étape par étape quelles modifications nous avons effectué et pourquoi.

En premier lieu nous avons renommé les colonnes pour une meilleure compréhension :
Tous les noms avec des majuscules au milieu ont été transformés comme suit :
- offerType est devenu offer_type, idem pour (postalCode, vehicleType, fuelType, …)
- yearOfRegistration est devenu registration_year, idem pour registration_month
- powerPS est devenu Power_HP

Au final 13 colonnes ont été renommées de la sorte.

Dans un deuxième temps, nous nous sommes penchés sur les features n'ayant qu'une valeur possible (aucune signification) ou deux mais une des deux valeurs est trop fréquente. C'est le cas de :
- offer_type : 371 516 lignes sur 371 528 avaient la valeur « particulier », donc seulement 8 annonces n'étaient pas des offres.
- seller : 371 525 lignes sur 371 528 avaient la valeur « privé », donc seulement 3 annonces provenaient de professionnels.
- nb_pictures : Toutes les annonces avaient un nombre de photos de 0.

Nous avons donc supprimé ces 3 features car elles ne comportaient aucune information.

Ensuite nous avons renommé toutes les valeurs allemandes ('ja', 'nein', …) en anglais afin que la visualisation des données soit plus claire, changé le type des 3 features avec des dates en format 'datetime' afin de pouvoir faire des analyses et calculs dessus (nous y reviendrons).

Puis nous avons commencé à regarder les valeurs manquantes (figure 8)

```
exploration_date          0
name                      0
price                     0
ab_test                   0
vehicle_type          37869
registration_year         0
gearbox               20209
Power_HP                  0
model                 20484
kilometer                 0
registration_month        0
fuel_type             33386
brand                     0
Damaged               72060
created_date              0
postal_code               0
last_seen                 0
dtype: int64
```

Figure 8: Liste des valeurs manquantes

Comme on peut le voir certaines colonnes pouvaient avoir jusqu'à 72 000 valeurs manquantes, ce qui représente 20 % du jeux de données et ceci concerne 5 colonnes.
Nous avons donc cherché à les remplacer et puis les supprimer en plusieurs étapes :

Premièrement : nous avons supprimé toutes les lignes où les valeurs étaient manquantes dans les 5 colonnes : suppression de 5 000 lignes. Supprimer les lignes est la dernière chose à faire mais si ces lignes ont 5 valeurs manquantes il vaut mieux s'en séparer car elles portent moins d'informations que les autres lignes.

Deuxièmement : la recherche, méthode spécifique à ce dataset car le 'name' des annonces comportait souvent des informations cachées (modèle du véhicule, type de carburant, …). Nous avons donc pour les lignes avec des valeurs manquantes dans modèle et type de carburant, chercher dans le name si l'information manquante y était.
Exemple : Dans le 'name' : « A5_Sportback_2.7_Tdi » on a non seulement le modèle (A_5) mais aussi le carburant car Tdi signifie que c'est un diesel.
Ce sont ce genre d'informations cachées dans le 'name' qui nous ont permis de retrouver :
- 3 000 lignes du type de carburant
- 4 000 lignes de modèle (très important car impossible à remplacer autrement)

Troisièmement : les valeurs aberrantes du prix. Le prix étant notre cible, aucun remplacement des valeurs n'est autorisé car cela viendrait fausser les estimations. Nous avons donc supprimé 10 000 lignes dans lesquelles le prix était à 0.
Puis en regardant un résumé des valeurs de la colonne prix (figure 9), nous avons remarqué que le min était 1 et le max 2 milliards. Nous avons donc réduit aux prix compris entre 500 et 150 000 euros, ce qui a supprimé 24 000 lignes. Ces valeurs étaient trop extrêmes, et s'il existe des prix entre 0 et 500 c'est parce que EBAY est un site d'enchères et que si personne n'enchérit alors le prix peut être très bas.

```
count    3.590940e+05
mean     1.740345e+04
std      3.645440e+06
min      1.000000e+00
25%      1.250000e+03
50%      3.000000e+03
75%      7.499000e+03
max      2.147484e+09
Name: price, dtype: float64
```

Figure 9: Résumé de la colonne prix

Quatrièmement : comme pour le prix, pour l'année d'immatriculation nous avons supprimé toutes les lignes où les dates étaient supérieures à 2016 car ce n'est pas possible (le dataset date de 2016) et les dates inférieures à 1900 car cela paraissait fortement erroné.
20 000 lignes ont encore été supprimées.

Cinquièmement : les doublons : ce n'est pas courant mais certaines lignes étaient en double (25 000 lignes). Nous les avons supprimées pour deux raisons : car si ces données se retrouvaient dans le set de test et d'entrainement à la fois cela biaiserait fortement nos prédictions. De plus ces lignes tendraient à moyenner nos prédictions puisque plusieurs individus auraient les mêmes caractéristiques.

Sixièmement : le remplacement des dernières valeurs manquantes par une valeur moyenne. Nous avons décidé de remplacer les valeurs manquantes des colonnes (type de carburant, type de transmission et type de véhicule) avec la moyenne de la valeur dans cette colonne par modèle de véhicule.
Par exemple si sur une ligne le type de carburant était manquant on a regardé pour ce modèle quel type de carburant était le plus courant et on a mis cette valeur. Cette technique nous a permis de combler 23 000 valeurs manquantes.
Pour 'notRepairedDamage' nous avons remplacé les 43 000 valeurs manquantes par la valeur inconnue, car le fait de ne pas connaitre l'historique du véhicule est également une valeur possible.

- Puis nous avons fini par retirer toutes les lignes où subsistaient des valeurs manquantes (8000 lignes)

Finalement le jeu de données « propre » contient 281 091 lignes, nous avons donc retiré quasiment 100 000 lignes (un tiers du dataset) mais ces valeurs étaient soit erronées, soit manquantes soit en double.

3.2.3 Implémentation de nouvelles features

Méthode courante : on crée de nouvelles features, si elles sont pertinentes. Cette technique va nous permettre de créer de l'information avec les features existantes. Nous pourrons plus tard avoir le choix entre garder ou les retirer si elles sont problématiques (les dates ne sont pas prises en compte par certains modèles de Machine Learning, les transformer en chiffres numériques leur faire perdre du sens).
Nous avons créé la colonne 'day_in' comme cela : elle correspond aux nombres de jours durant lesquels l'annonce a été affichée, il s'agit de la soustraction entre la date de création et la dernière date à laquelle l'annonce a été vue. La nouvelle colonne a 253 296 valeurs différentes de 0 et une moyenne de 9, avec beaucoup de valeurs différentes donc nous la gardons.

3.2.4 Préparation à la modélisation

Comme nos variables ne sont pas toutes numériques (8 colonnes avec des variables textuelles) nous devons trouver un moyen de les transformer de manière pertinente.
Les trois features avec des dates : ‘created_date’, ‘exploration_date’ et ‘last_seen’ ont été converties à l’aide de la fonction ‘datetime’.
Pour ‘Damaged ‘nous avons effectué un mapping, c’est-à-dire que si la valeur est **non** on la remplace par 0, si c’est un **oui** par 1 et **inconnu** par 2.

Nous avons supprimé ‘ab_test’ et ‘name’, ‘ab_test’ est une notion de EBAY qui n’est pas pertinente dans notre analyse puisqu’elle n’est pas liée au véhicule. Quant à ’name’ nous avons extrait de cette feature des noms de modèles et type de carburant, mais la feature ‘name’ est remplie de champs de caractères allemands sans signification que l’on pourrait étudier ou d’informations déjà présentes dans les autres colonnes. De plus c’est cette feature qui a le plus grand nombre de différentes valeurs (171 000 valeurs différentes : il est inutile de les mapper ou de les one-hot encoder).

L’encoding, le mapping et le regroupement (pour faire des moyennes) sont 3 méthodes utilisées pour se défaire des variables textuelles non-interprétables par les algorithmes de régression.
Nous avons donc déjà effectué un mapping, maintenant faisons un regroupement sur ‘model’. Chaque valeur de model a été remplacé par la moyenne de prix de ce modèle, puis nous avons séparé ces moyennes en plusieurs intervalles, ces différents intervalles ont été converties en classes. Ainsi un modèle dont le prix moyen aurait été entre 0 et 1000 est devenu la valeur 1 (classe 1) et ainsi de suite jusqu’à la classe 7 qui correspond à plus de 60 000 euros. Le mapping aurait été possible même s’il avait conduit à 274 valeurs différentes. Cette feature aurait alors perdu beaucoup de son sens alors qu’elle est primordiale.

Puis sur les 4 dernières colonnes ‘gearbox’, ‘brand’ , vehicle_type’ et ‘fuel_type’ nous avons effectué un encoding. Cela consiste à remplacer une valeur d’une colonne par une colonne entière dans le dataset. Par exemple si une ligne du jeu de données est de marque BMW, nous créons une colonne pour cette marque (avec 1 dedans pour cette ligne) et nous avons 0 dans les colonnes des marques différentes de BMW. Dans son article J. Brownlee nous rappelle que les données de types catégories peuvent être gérées par des modèles comme l’arbre de décision (nous pourrons donc tester ce modèle sur notre jeu de données sans le encoding). Mais J. Brownlee nous rappelle aussi que beaucoup de modèles ont besoin de données exclusivement numériques et nous décrit les différents types d’encoding (integer et one_hot). L’encoding one-hot est similaire a celui que nous avons effectué et l’integer correspond à ce que nous avons fait avec le mapping.

En conclusion, nos données sont prêtes pour le modeling, nous avons suivi la plupart des techniques de nettoyage des données décrites dans l’article web de E. Rençberoğlu : imputation des valeurs manquantes, suppression des valeurs aberrantes (outliers), regroupement en classe (binning) sur les modèles de voitures, l’encoding, l’extraction de date et création de nouvelle feature avec. Dans son article l’auteur parle aussi de log transformation et de scaling que nous allons voir dans la description de la méthodologie du modeling (implémentation du modèle).

4. Machine Learning : méthode

Comme précisé dans la partie approche méthodologique nous avions le choix entre un problème de régression ou de classification. Nous avons choisi la régression car il est plus pertinent d’essayer de prédire un prix. De plus la classification binaire (0 et 1) ne convient pas dans ce cas, et la classification multi-classe reviendrait à prédire notre nouvelle variable modèle. La variable modèle qui représente déjà le prix moyen réparti en classes.

Nous allons détailler rapidement les étapes et méthodes clés que nous allons utiliser afin de créer notre modèle. Une fois le jeu de données « propre » et constitué uniquement de variables numériques nous commençons par un split du jeu de données. C'est très important pour la suite, nous mettons 70 % des données dans le jeu d'entrainement (training set) et 30 % dans le jeu de test (test set). Ainsi nous pourrons entrainer nos modèles sur des données et les tester sur d'autres. Ainsi le modèle n'aura jamais vu les réponses aux prédictions, et nous assure plus de crédibilité des prédictions.

Le modeling : une fois nos deux jeux de données prêts, nous allons tester un maximum de modèles avec la cross_validation. Cette méthode permet de tester rapidement un modèle sur l'ensemble de nos données . En effet pas besoin de séparer le jeu de données, l'algorithme de cross_validation va lui-même tester plusieurs séparations et pour chaque essai retourner une précision (indice de performance que nous verrons après). Cette méthode est d'ailleurs bien décrite par M. Sanjay dans son article sur le cross_validation, où il décrit les avantages de cette méthode comme le fait de pouvoir tester plusieurs fois le modèle sur différentes portions du jeu de données, ce qui réduit les chances d'avoir des prédictions baisées.
Avec cette méthode nous testerons différents algorithmes : régression linéaire, régression Ridge, Lasso et LassoLARS, SVR, arbre de décision et random_forest.

L'évaluation des prédictions est très important. Pour la régression nous avons donc différents indicateurs :

- RMSE ou Root Mean Square Error : racine carrée de l'erreur quadratique moyenne, qui est la somme des carrés des distances entre les prédictions et les vraies valeurs attendues.
Plus cet indicateur est grand plus les prédictions sont éloignées des vraies valeurs.
- R^2 ou coefficient de détermination : la formule nous explique très bien cet indicateur :
1 – RSE ou RSE (erreur carré relative) est compris entre 0 et 1. Le RSE étant la somme des erreurs (différence entre la prédiction et ce que l'on attend) au carré, divisé par la somme des erreurs moyennées (valeurs attendues moins la moyenne des ces données).
En interprétation générale le R^2 est donc compris entre 0 et 1, 0 ce qui signifie des prédictions erronées. Si le R^2 vaut 1 ce qui est utopique, cela signifierait que le modèle justifie toutes les variations dans les données. Donc nous essayerons d'avoir un R^2 le plus proche de 1 possible.

Avec la cross_validation nous sélectionnerons le modèle avec le meilleur R^2 et puis nous optimiserons ce modèle grâce à la méthode GridSearch qui permet de tester un modèle en faisant varier les hyperparamètres que nous aurons choisis. Nous opterons également pour deux types de scaler (mise à l'échelle) : le standard et le MinMax.

IV) Présentation des résultats et discussion

1. Résultats du sondage

Première information révélée par le sondage : 75.8 % des répondants ont déjà achetés un véhicule ce qui représente 111 individus :
Parmi lesquels 108 ont déjà acheté un véhicule d'occasion et seulement 71 ont déjà acheté un véhicule neuf. Donc l'achat d'un véhicule est plus souvent tourné vers l'occasion.
De plus 77.2 % (115 individus) des répondants préfèrent acheter un véhicule d'occasion contre seulement 18.8 % (28 individus) qui préférèrent l'acheter neuf. Cette grosse différence montre bien qu'acheter un véhicule neuf est moins courant.

Rentrons maintenant dans les critères de sélection du véhicule :
Parmi 5 critères globaux (prix, récent et peu kilométré, l'utilité, la marque et l'esthétique) les répondants ont privilégié le prix (34 %) mais aussi le fait que le véhicule soit récent et peu kilométré (30.2%) et son utilité (24.2%). Les deux dernières catégories ont été faiblement représentées (figure 10).

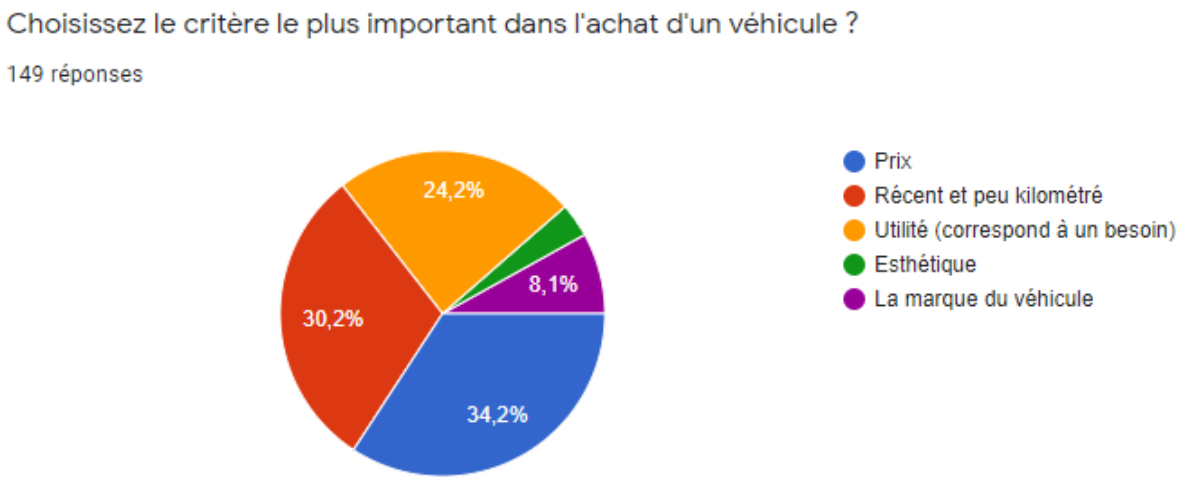


Figure 10: Répartition des répondants sur leur critère principal de sélection

Rien d'étonnant pour le prix mais il est intéressant de constater que l'esthétique et la marque du véhicule sont des critères plus secondaires et que beaucoup d'individus attachent de l'importance à l'âge et au kilométrage du véhicule.

Lorsque nous avons demandé leurs 4 critères préférés parmi un longue liste de critères : Nous avons retrouvé le prix en critère numéro 1 (81.9 %), suivi du kilométrage (61.7 %) et de la marque du véhicule (49.7%).
Il est à noter que récent était proposé comme critère et n'a reçu que 10.1 % de sélection ce qui nous indique que le kilométrage est bien plus important que l'âge du véhicule. Ensuite 3 réponses arrivent en quatrième place quasiment exæquo :
- Les équipements (options) : 38.3 %
- L'esthétique : 34.2 %
- L'historique du véhicule : 31.5 %

Nous pouvons voir que ces 3 critères sont secondaires (surtout l'esthétique qui n'est jamais placée en critère numéro 1).
Finissons avec les critères tertiaires :
- La puissance : 22.8%
- L'écologie : 18.1 %
- La sécurité : 16.8 %

Enfin il est à noter qu'utilité n'a été choisie que dans 12.8 % des cas ce qui est très faible compte tenu du fait que c'était le 3ème critère principal. Nous ne tiendrons pas compte de cette information car lorsqu'ils n'ont eu qu'un seul choix les répondants ont préféré l'utilité à la marque ou même l'esthétique.
Les critères de facilité d'achat ou de garantie et de payement ont eu été très peu sélectionné. On en déduit qu'ils préfèrent attendre et faire plus d'efforts pour l'achat d'un véhicule.

Au niveau motorisation on retrouve en premier l'essence (44.3%), suivi de l'hybride (26.2%) et enfin du diesel (19.5%). L'électrique arrive juste après mais très faiblement représentée (6.7%). L'information à retenir ici est que l'hybride est plus demandé que le diesel, l'hybride est donc assez attractif, ce qui s'explique par les bonus/malus écologique offerts en France. Mais les gens ne sont pas encore prêts pour le 100 % électrique.

Pour ce qui est de la marque, les chiffres dévoilent que les individus y apportent de l'importance assez souvent (87.9%). Parmi eux 48.3 % tiendraient quand même à éviter certaines marques alors que les 39.6 % restants se disent prêt à regarder la concurrence en fonction des offres. Très peu de répondants dans les deux extrêmes, la marque n'est donc pas un critère primordial mais secondaire.

Enfin à la question : Certains facteurs extérieurs pourraient-ils vous influencer dans l'achat d'un véhicule ?

Est revenu le plus souvent : une offre intéressante (59.1%) , conseil d'un ami (44.3%) et conseil d'un guide d'achat (33.6%).

Les facteurs culturels ne sont pas souvent sélectionnés (voiture de film) (7.4%) et personne n'a mis de sport automobile dans l'option autre.

2. Résultats du Machine Learning

Tous les codes et graphiques seront disponible sur un GitHub dont le lien est donné au début de l'annexe.

2.1 Data visualisation

Comme vous pouvez le voir sur la figure 11 (annexes) , les prix sont principalement compris entre 0 et 20 000 euros.

Avec les graphiques on observe que les kilométrages sont déjà regroupés par valeurs (figure 12 en annexes). Il est à noter que le maximum est 150 000 ce qui laisse supposer une limite. Le type de carburant le plus représenté est l'essence et on a 75 % environ de véhicule à boite manuelle.

Les corrélations, permettent de savoir quelles colonnes sont les plus corrélées avec le prix. Avant l'encoding et les changements de variables textuelles vers numériques nous avons regardé les corrélations des features (figure 13)

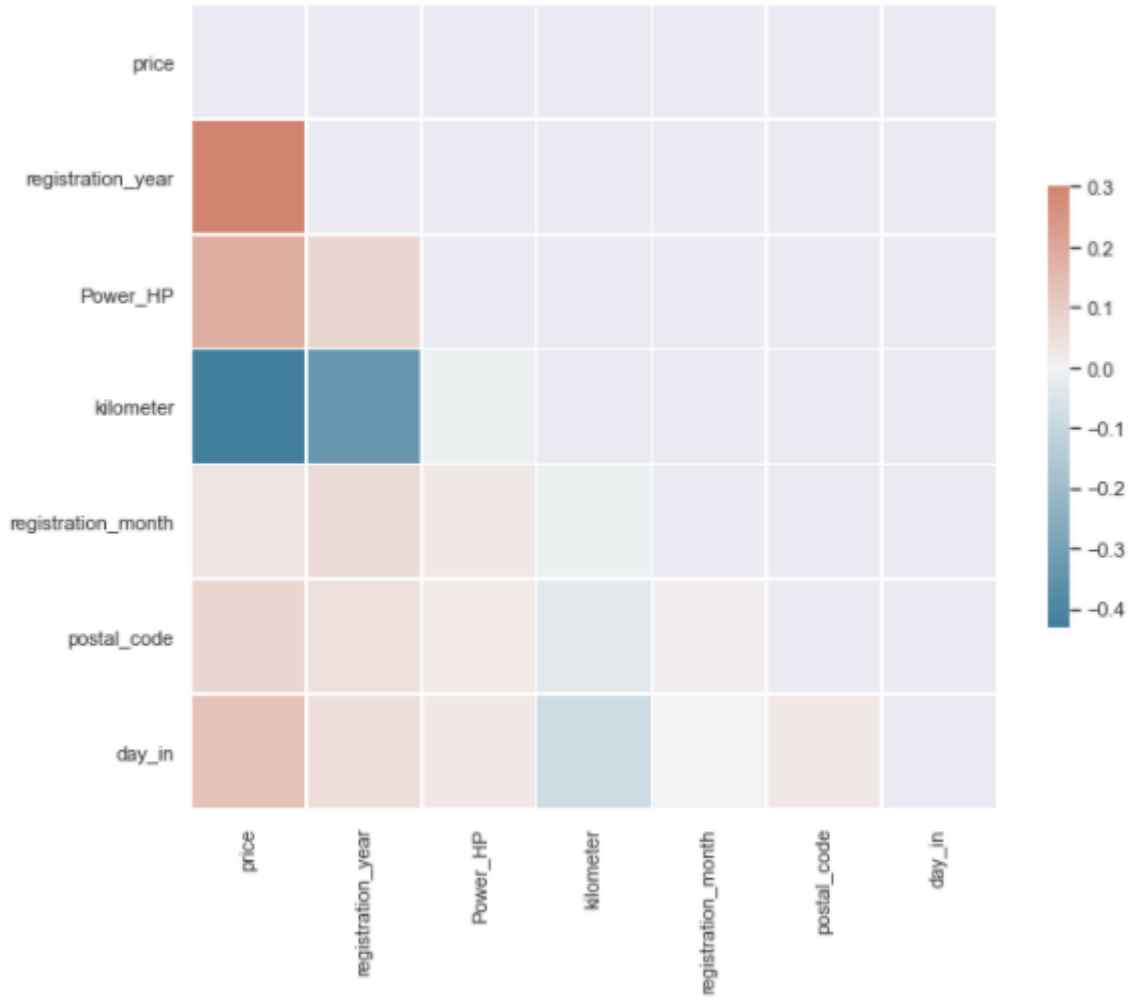


Figure 13: Matrice de corrélation des premières features

Les variables les plus corrélées sont donc les kilomètres, l'année d'immatriculation (récent ou pas) et la puissance.

Après nos changements nous avons maintenant 68 colonnes. Il est donc impossible de les afficher, mais nous mettrons en annexe notre liste des corrélations. En résumé les variables les plus corrélées sont :

- Modèle, année d'immatriculation, les marques allemandes (one-hot encoder), les kilomètres, le nombre de jour de diffusion de l'annonce, certains types de carburant (essence, diesel, hybride), la puissance, certains types de véhicules (SUV, coupé et cabriolet, citadine), et les deux types de transmission.

2.2 Résultats des modèles

La cross validation a fourni les résultats suivants (figure 13) :

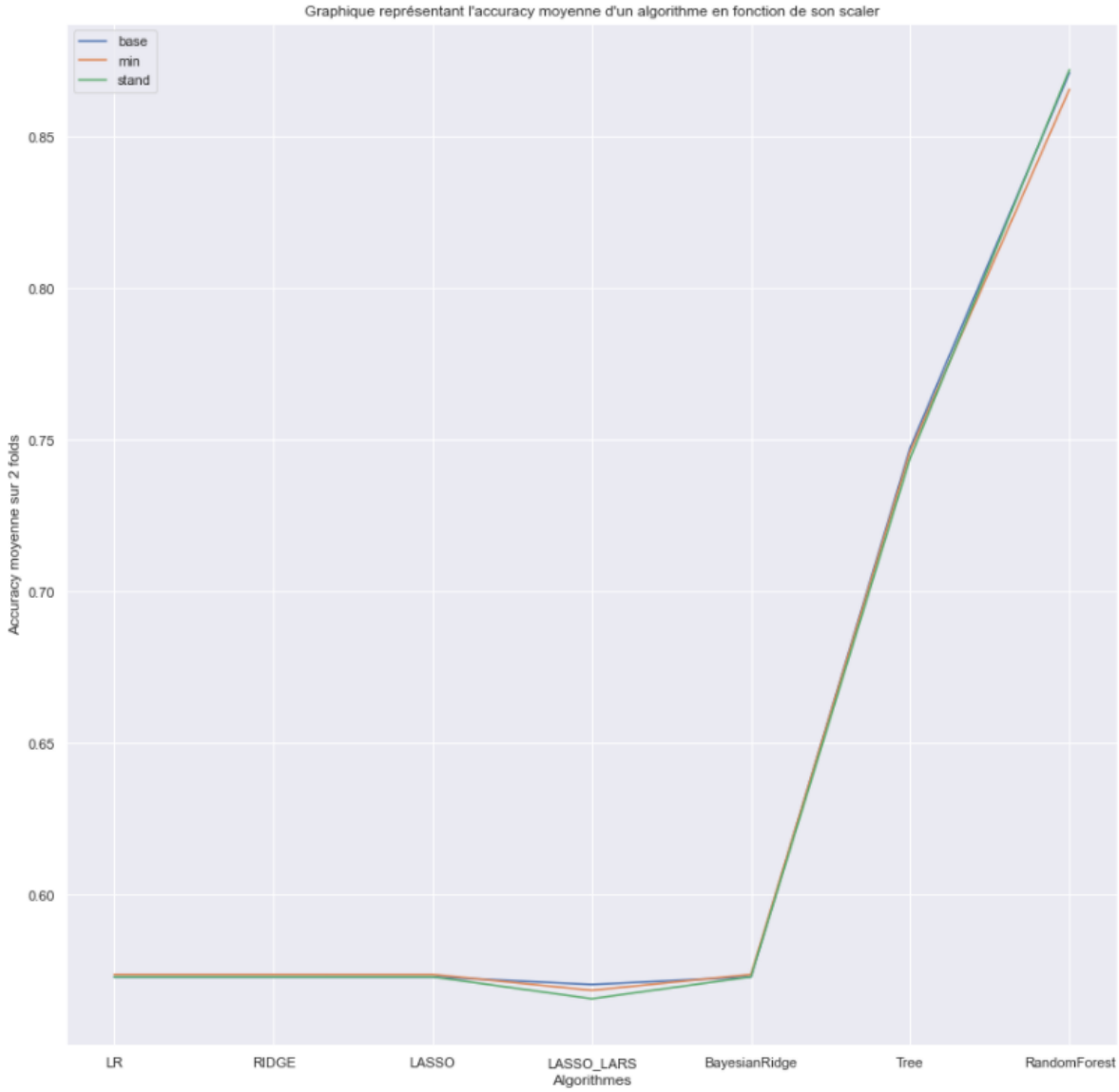


Figure 13: Résultats de la cross_validation

Les deux meilleurs modèles sont l'arbre de décision et le Random Forest avec respectivement un R^2 de 0.75 et 0.88.
Nous avons donc choisi de rester avec ces deux modèles. Les autres ayant un R^2 inférieur à 0.60.
Les type de scaler n'ont pas trop affecté les résultats, malgré les 68 colonnes numériques.

En implémentant ces deux modèles sans cross validation nous avons retrouver nos score R^2 similaires.

Ensuite nous avons donc appliqué un GridSearch sur le Random Forest Regressor en faisant varier deux paramètres :

- Max_depth : profondeur de l'arbre maximum : 25 ou 50
- Max_features : qui correspond aux nombres de colonnes à prendre.

Les meilleurs paramètres sont : max_depth = 50 et max_features = auto (nombre de feature = toutes les colonnes)

En implémentant le modèle avec ces paramètres nous avons obtenu un R^2 de 0.86 (aucune amélioration). Comme le montre la figure 14 la plupart des prédictions sont sur ou près de la ligne de régression (prédictions 100% égales aux valeurs attendues)

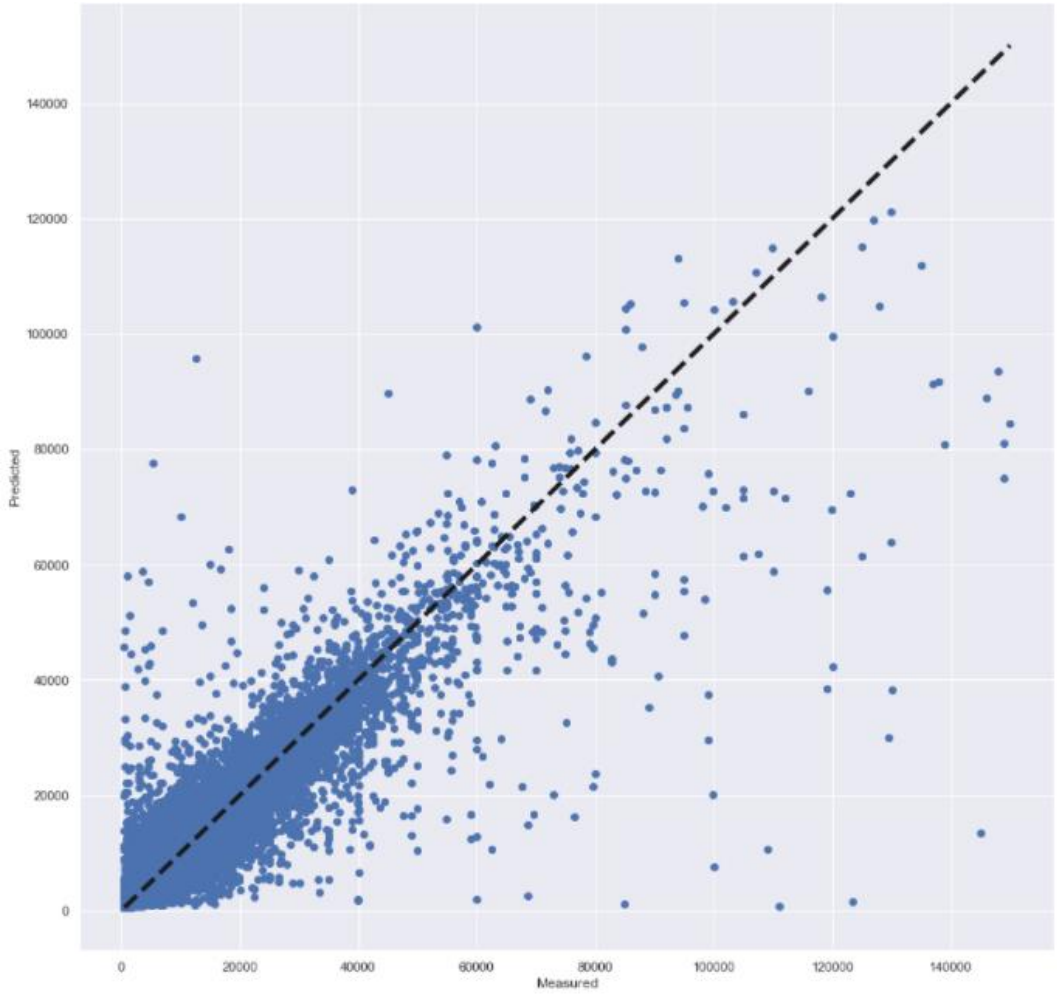


Figure 14: Graphiques des predictions de notre modèle

V) Conclusion et recommandations

Après avoir généré notre meilleur modèle avec les meilleurs paramètres nous avons un modèle avec un très bonne précision :
Nous avons exporté ce modèle sous forme d'API (dans le git) mais nous ne l'avons pas déployé faute de moyens.

1. Interprétation des résultats

Le sondage a révélé plusieurs informations pertinentes :
Les particuliers préfèrent acheter des véhicules d'occasion.
On distingue trois classes de facteurs distingués par leur importance :

- Classe 1 : le prix, le kilométrage, marque du véhicule
- Classe 2 : l'utilité (besoins de l'individu), les équipements, l'historique du véhicule et l'esthétique.
- Classe 3 : la puissance, l'écologie (consommation) et la sécurité.

Il est à noter que le fait que le véhicule soit récent importe peu tant qu'il est peu kilométré.
D'autre part un quart des personnes sondés sont intéressées par la motorisation hybride.

Avec le Machine Learning : nous avons pu regarder les corrélations et nous en déduisons que pour le marché national, une voiture est de cette nationalité aura un prix plus stable. De plus encore une fois la motorisation hybride se démarque et est fortement corrélée au prix ce qui est remarquable pour des données de 2016. on peut alors penser que c'est encore plus vrai aujourd'hui.
Le kilométrage, modèle et historique du véhicule sont eux aussi des facteurs justificatifs du prix.

Enfin notre modèle ayant de bons résultats, on peut en déduire que le prix d'un véhicule peut s'expliquer et se prédire avec une bonne précision. Cette prédiction est basée sur toutes les features que nous avons lister juste avant.

2. Comparaison

Nous l'avions vu dans la revue de littérature beaucoup de modèles ont déjà été utilisés dans la prédiction du prix de véhicule d'occasion, le Random Forest avait déjà montré de bons résultats dans cette tâche il est donc logique que ce soit ce modèle qui soit ressorti.
Bien que certaines études aient transformé le problème de régression en problème de classification par manque de données ou par des résultats insuffisants, nous avons pu rester sur une tâche de régression.

Notre étude se démarque des autres par l'important travail de préparation des données. Nous avons pu garder toutes les features ce qui rend notre modèle plus robuste et lui permettra d'analyser plus de critères par la suite. Le fait de bien veiller à retirer les valeurs aberrantes nous assure que le modèle n'est pas biaisé.

En effet l'étude de Monburinon, Rungpheung, … était sur le même jeu de données (c'est un des seuls complet) à l'origine mais nous avons nettoyé le jeu de données en profondeur et notre méthode pour convertir le texte en nombre nous a permis de garder toutes les informations présentes au départ. Tout notre travail de préparation de données a été bien différen et nous avons pu avoir de meilleurs résultats tout en sélectionnant mieux nos données.

Comme les autres études de Machine Learning nous avons réussi à prédire notre cible, toutes les étapes ont été décisives pour que ce modèle soit le plus proche le plus possible de la réalité.

3. Conclusion et recommandations

En conclusion notre étude nous a permis de déceler les critères expliquant le prix d'un véhicule d'occasion, nous permettant alors de connaitre le comportement des acheteurs mais aussi leur préférence. C'était le but de cette étude : arriver à prédire le prix d'un véhicule d'occasion et isoler les critères dont le prix dépend le plus.

Notre démarche a été celle du Data Scientist, nous avons beaucoup travaillé nos données et leur traitement a été fait avec pertinence. En nous basant sur des annonces de véhicules nous savions que nous pourrions extraire des features avec les corrélations mais le véritable objectif était d'implémenter un modèle précis basé sur des données nettoyées.
Notre meilleur modèle est donc un Random Forest (multiple arbre de décision aléatoire, permettant d'avoir le meilleur arbre parmi plusieurs arbres générés de façon aléatoire.

La précision obtenue est très satisfaisante compte tenu du fait qu'il s'agit d'un problème de régression et donc qu'il ne suffisait pas de prédire une classe ou deux mais des milliers de valeurs possibles.

En complément de ce travail de Machine Learning, nous avons également réalisé un sondage pour connaitre mieux le comportement des Français par rapport à l'achat de véhicule d'occasion, ce qui a permis de tirer des conclusions sur l'importance de différents critères. Ce sondage nous a permis de renforcer nos conclusions émisse par l'analyse des données de l'EBAY Allemand.

Les recommandations et faits à retenir concernant ce mémoire sont les suivants : les plus gros facteurs expliquant le prix d'un véhicule sont le kilométrage, la marque du véhicule et d'autant plus si cette marque est produite dans le pays de vente, l'historique du véhicule, la motorisation (essence et diesel) mais aussi l'hybride qui sera un des prochains critères importants. On a pu aussi déterminer des classes de facteurs (1,2,3) grâce aux réponses du sondage. En autre il est à retenir que les problèmes de régression posent des soucis de type de données, dans notre cas il était impossible de vraiment décrire les caractéristiques d'un véhicule seulement avec des chiffres mais nous avons surmonter ce problème par des techniques d'encoding. Enfin le Random Forest est un modèle efficace dans la prédiction de prix de revente de véhicule d'occasion.

Pour continuer nous pourrions essayer de discuter avec des entreprises comme La Centrale afin qu'ils nous donnent accès à leurs données, nous compterions sur tous nos travaux pour montrer notre implication mais cela reste utopique. L'objectif serait de travailler sur encore plus de features. Nous pourrons aussi déployer notre API et permettre au gens de l'interroger même si aujourd'hui ce service est déjà répandu. Enfin il serait très pertinent d'accéder à des données toujours plus récentes et regarder l'évolution des véhicules électriques ainsi que la consommation en fonction des nouvelles législations.

Bibliographie

- M. Mitchell, T. (1997). Does Machine Learning Really Work? *AI Magazine*, *18*, 1–10. https://ojs.aaai.org/index.php/aimagazine/article/view/1303
- Zulkifley, N.-H., Abdul Rahman, S., Ubaidullah, N.-H., & Ibrahim, I. (2020). House Price Prediction using a Machine Learning Model: A Survey of Literature. *I.J. Modern Education and Computer Science*, 1–9. https://doi.org/10.5815/ijmecs.2020.06.04
- Singh, A., Sharma, A., & Dubey, G. (2019). Big data analytics predicting real estate prices. *The Society for Reliability Engineering, Quality and Operations Management (SREQOM)*, 1–12. https://doi.org/10.1007/s13198-020-00946-3
- Fathalla, A., Salah, A., Li, K., Li, K., & Francesco, P. (2019). Deep end-to-end learning for price prediction of second-hand items. *Knowledge and Information Systems*, 1–28. https://doi.org/10.1007/s10115-020-01495-8
- Chen, Z., Li, C., & Sun, W. (2019). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, 1–13. https://doi.org/10.1016/j.cam.2019.112395
- Monburinon, N., Rungpheung, S., Chertchom, P., Buya, S., Kaewkiriya, T., & Boonpou, P. (2018). Prediction of Prices for Used Car by Using Regression Models. *International Conference on Business and Industrial Research (ICBIR)*, *5*, 1–5. https://doi.org/10.1109/ICBIR.2018.8391177
- Pudaruth, S. (2014). Predicting the Price of Used Cars using Machine Learning Techniques. *International Journal of Information & Computation Technology*, *4*, 1–11. http://ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf
- Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car Price Prediction using Machine Learning Techniques. *TEM Journal*, *8*, 2–7. https://doi.org/10.18421/TEM81-16
- Long, W., Lu, Z., & Cui, L. (2019). Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems*, *164*, 1. https://doi.org/10.1016/j.knosys.2018.10.034
- Richarson, M. (2009). DETERMINANTS OF USED CAR RESALE VALUE. *The Faculty of the Department of Economics and Business*, 1. https://digitalccbeta.coloradocollege.edu/pid/coccc:1346/datastream/OBJ

- W. Osborne, J. (2013). Is data cleaning and the testing of assumptions relevant in the 21st century ? *Psychol*, 1. https://doi.org/10.3389/fpsyg.2013.00370/full
-

Références

1. Ressources web :

- Fruchard, A. (2020, July 30). *Tout savoir sur la cote argus*. Reassurez-Moi. https://reassurez-moi.fr/guide/assurance-auto/argus
- Brownlee, J. (2017, 28 juillet). *Why One-Hot Encode Data in Machine Learning?* machinelearningmastery. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/
- Rençberoğlu, E. (2019, 1 avril). *Fundamental Techniques of Feature Engineering for Machine Learning*. Towardsdatascience. https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114
- Sanjay, M. (2018, 13 novembre). *Why and how to Cross Validate a Model?* Towardsdatascience. https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

2. Ressources non scientifiques :

- Lugen, M. (2017). *Petit_guide_de_méthodologie_de_l_enquête*. https://igeat.ulb.ac.be/fileadmin/media/publications/Enseignement/Petit_guide_de_me%CC%81thodologie_de_l_enque%CC%82te.pdf

Annexes

- Lien du github : https://github.com/jdelebec/Memoire

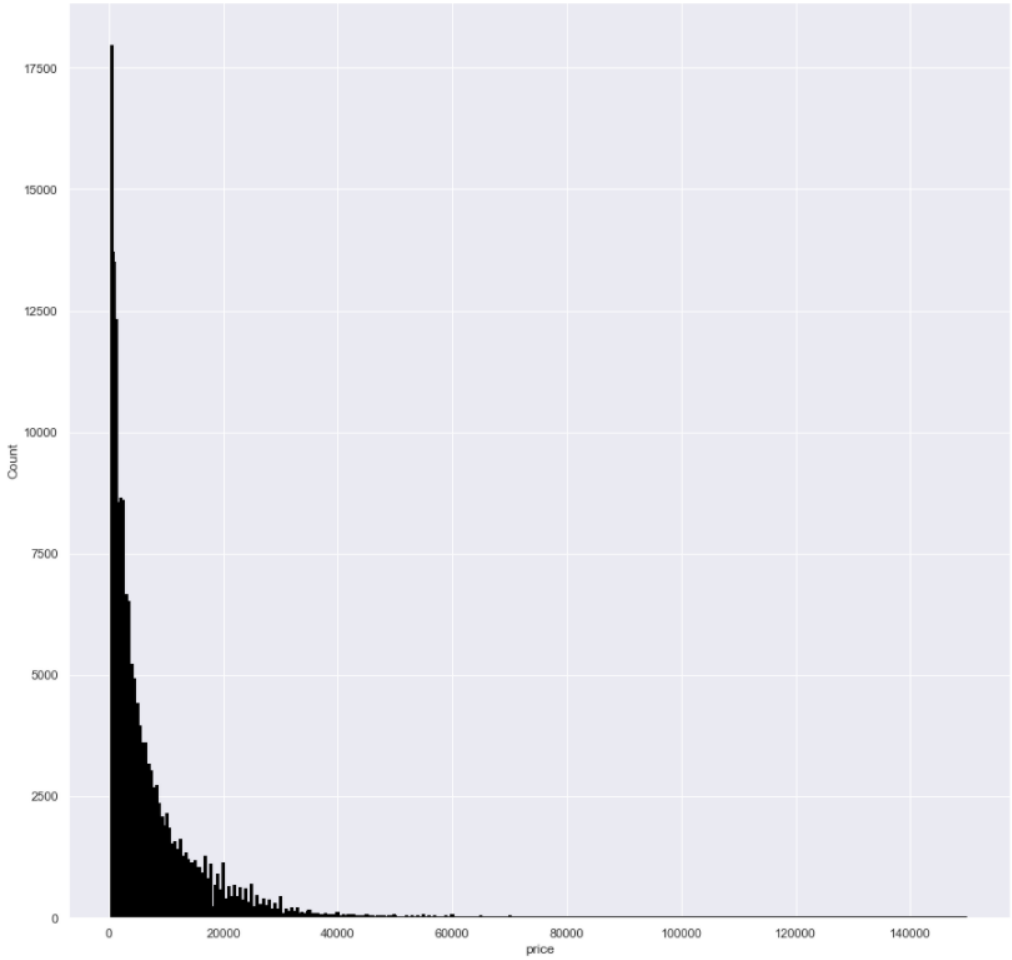


Figure 11: En abscisse les prix, en ordonnée le nombre d'occurences de chacun

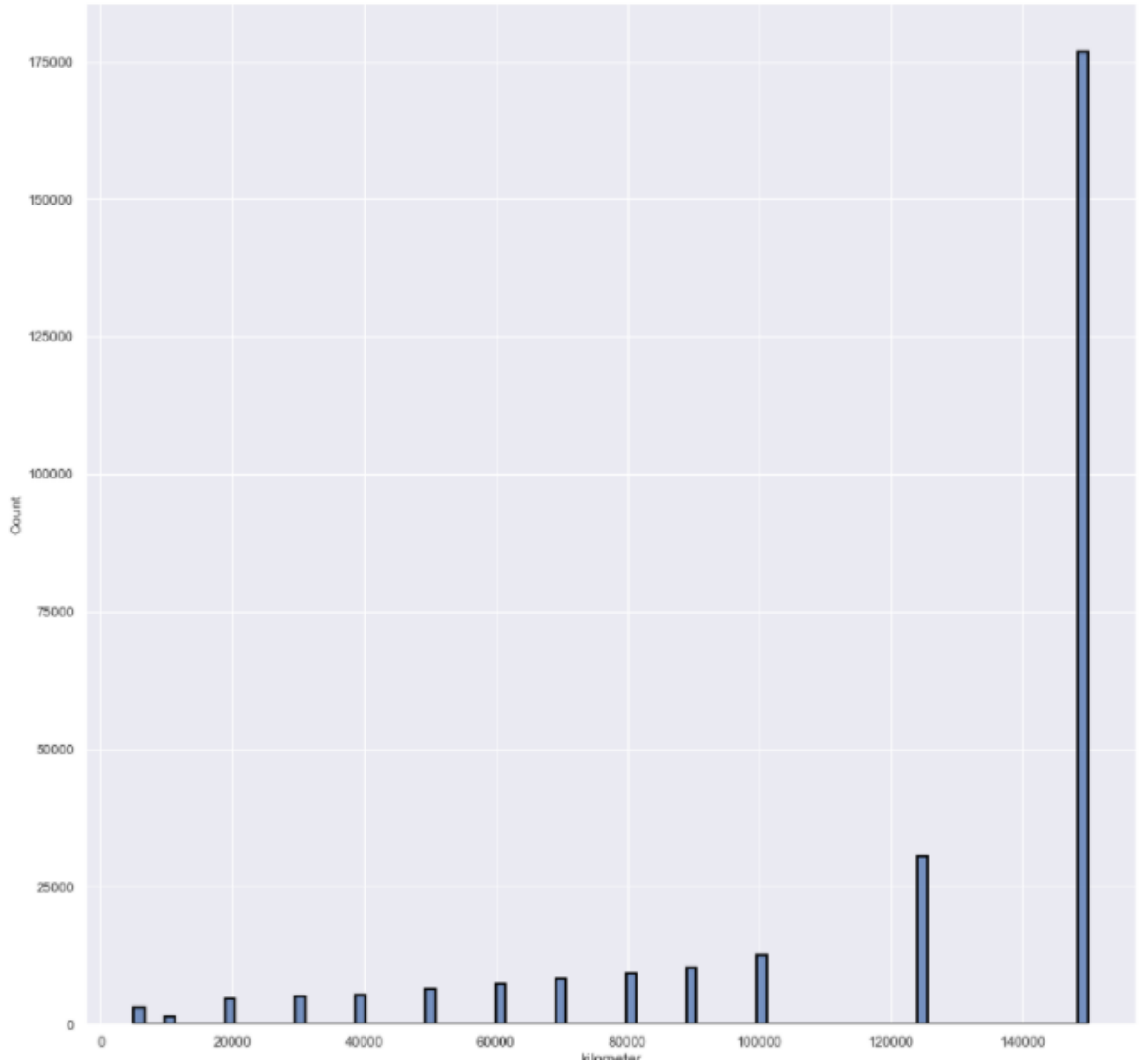
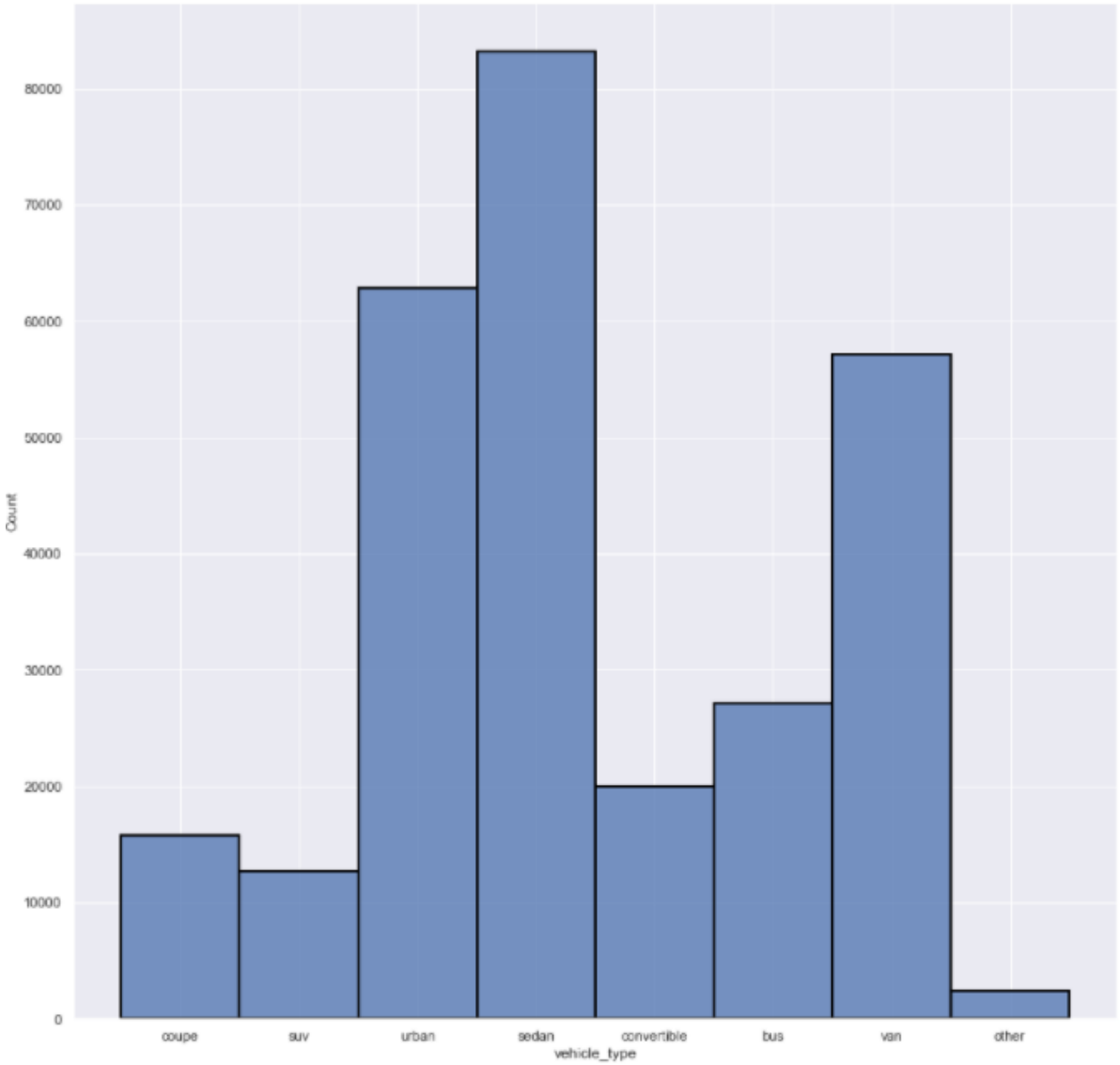


Figure 12: En abscisse les kilomètres, en ordonnée le nombre d'occurences de chacun

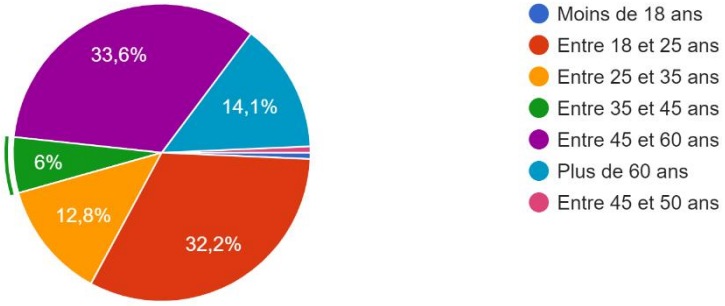


Graphiques annexes 1: Repartition des type de véhicules

Sondage (Toutes les réponses)

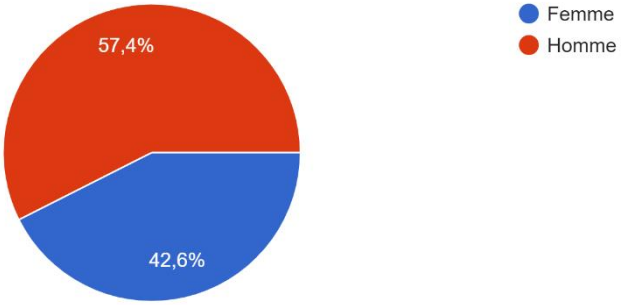
Quel âge avez-vous ?

149 réponses



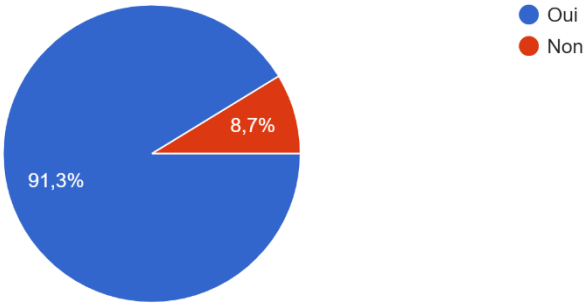
Quel est votre sexe ?

148 réponses



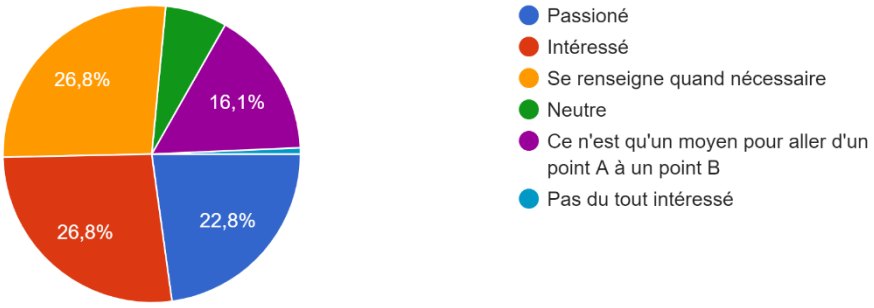
Avez-vous le permis ?

149 réponses



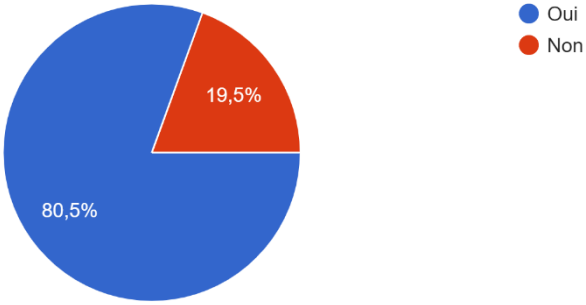
Quel niveau d'intérêt éprouvez-vous pour l'automobile ?

149 réponses



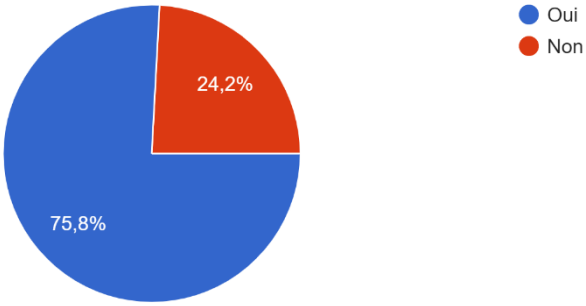
Avez-vous un véhicule ?

149 réponses



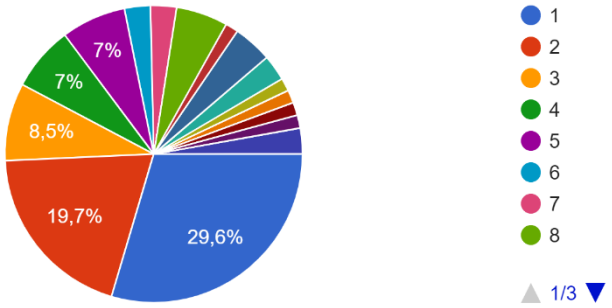
Avez-vous déjà acheté un véhicule ?

149 réponses



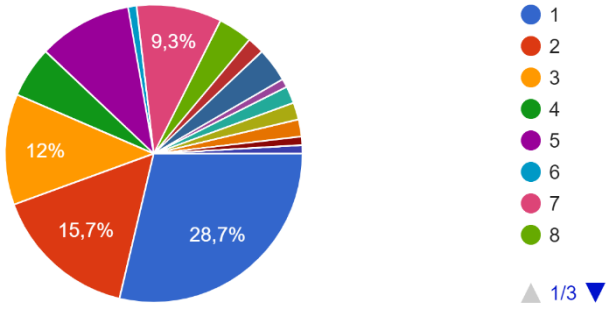
Si oui, combien de véhicules neufs avez-vous achetés ?

71 réponses



Si oui, combien de véhicules d'occasion avez-vous achetés ?

108 réponses

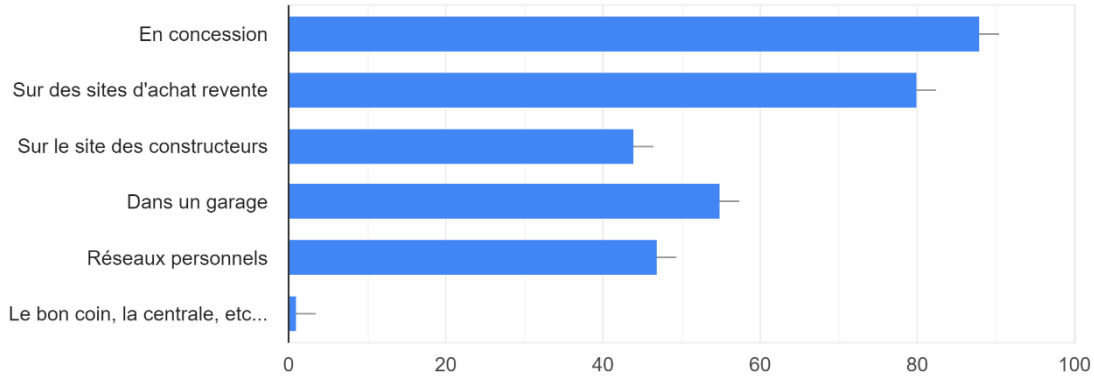


Si demain vous vouliez acheter un véhicule, le prendriez-vous neuf ou d'occasion ?

149 réponses

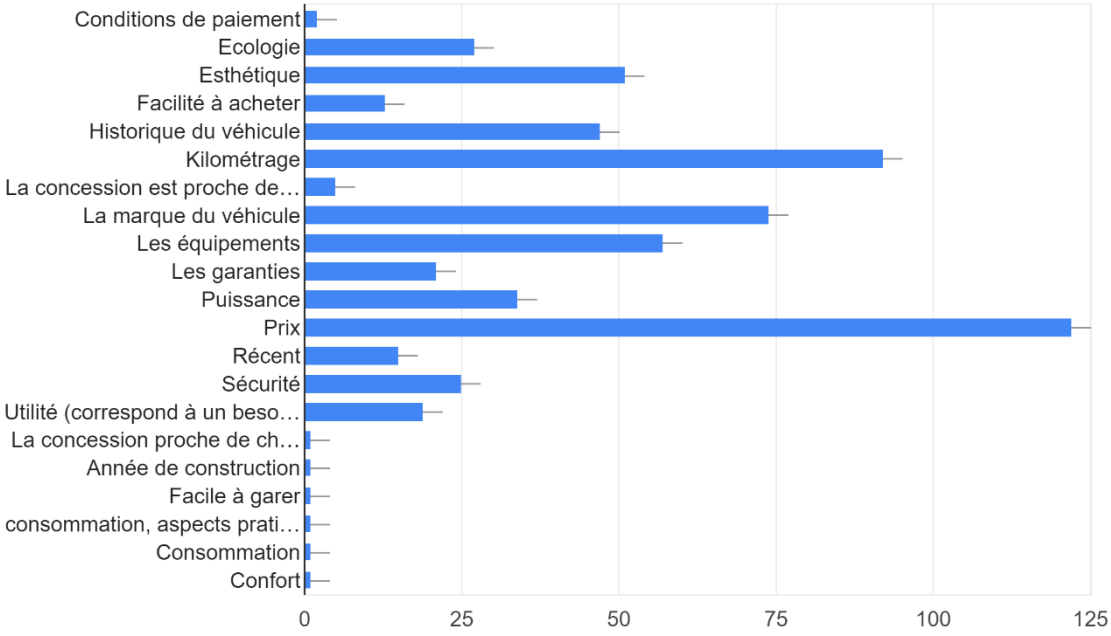
Où rechercheriez-vous ? (plusieurs choix possibles)

149 réponses



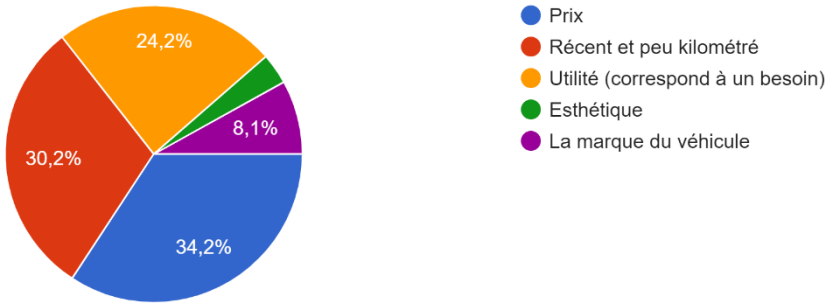
Quels sont vos critères de recherche pour un véhicule ? (maximum 4 choix possibles)

149 réponses



Choisissez le critère le plus important dans l'achat d'un véhicule ?

149 réponses

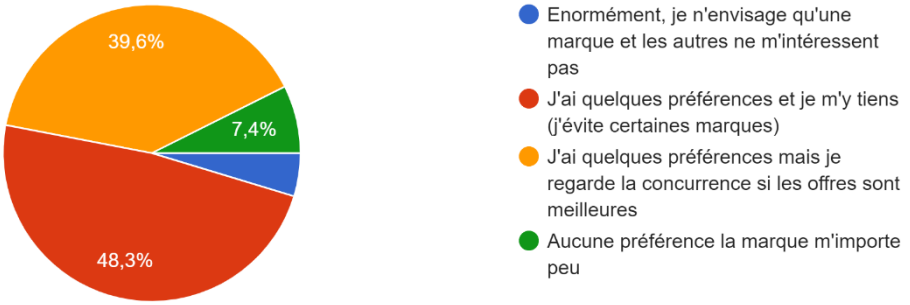


Quelle motorisation choisiriez-vous ?

149 réponses

Dans l'achat d'un véhicule : la marque du véhicule vous importe-t-elle ?

149 réponses



Certains facteurs extérieurs pourraient-ils vous influencer dans l'achat d'un véhicule ? (plusieurs choix possibles)

149 réponses