Une image contenant voiture, véhicule, roue, pneu

Description générée automatiquement

Colson Benjamin

Friia Isra

Masseau Luxon

Papelard Charlotte

Saiz Edgar

**Combien vaut votre voiture ?**

Une estimation du prix de revente de votre voiture

Une image contenant voiture, véhicule, roue, pneu

Description générée automatiquement

Table des matières

[Introduction 3](#_Toc152536364)

[1. Présentation du projet 4](#_Toc152536365)

[2. Comment construire un tel outil ? 5](#_Toc152536366)

[2.1. Les données 5](#_Toc152536367)

[2.2. Les variables 5](#_Toc152536368)

[2.3. L’architecture du projet 6](#_Toc152536369)

[3. L’avancée du projet 6](#_Toc152536370)

[3.1. Troisième sprint (du 22/10 au 09/11) 7](#_Toc152536371)

[3.1.1. Interface utilisateur 7](#_Toc152536372)

[3.1.2. Préparation des données 8](#_Toc152536373)

[3.1.3. Algorithme de prévision 8](#_Toc152536374)

[3.2. Quatrième sprint (du 10/11 au 23/11) 9](#_Toc152536375)

[3.2.1. Interface utilisateur 9](#_Toc152536376)

[3.2.2. Préparation des données 9](#_Toc152536377)

[3.2.3. Algorithme de prévision 10](#_Toc152536378)

[3.3. Cinquième sprint (du 24/11 au 7/12) 10](#_Toc152536379)

[3.3.1. Interface utilisateur 10](#_Toc152536380)

[3.3.2. Préparation des données 13](#_Toc152536381)

[3.3.3. Algorithme de prévision 13](#_Toc152536382)

# Introduction

L'industrie automobile en France est un secteur d'activité majeur qui continue de susciter un vif intérêt et d'occuper une place centrale dans la vie quotidienne des Français. Avec plus de 37,8 millions de voitures particulières en circulation, le paysage automobile français est en constante évolution. Cependant, l'achat d'une voiture neuve demeure souvent un investissement financier considérable, ce qui pousse de nombreuses personnes à se tourner vers le marché des véhicules d'occasion. La question de l'achat et de la revente de voitures d'occasion est un enjeu crucial qui concerne des millions de citoyens. Dans ce contexte, notre équipe s'est engagée dans la création d'une application visant à offrir une estimation du prix de revente des véhicules d'occasion, dans le but de simplifier ce processus et de répondre aux besoins des propriétaires de voitures d'occasion en France.

Ainsi, permettez-moi de vous présenter les membres de cette équipe qui ont mis leurs compétences au service de cette application :

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

Notre équipe vous propose un rapport explorant la conception, le développement et le résultat d’une telle application.

# Présentation du projet

L’objectif de ce projet est de concevoir un outil permettant aux automobilistes d’estimer rapidement le prix de leur voiture en renseignant ses caractéristiques. L'utilisateur pourra accéder à notre outil à partir d’un formulaire VBA. Dans celui-ci, il renseignera les caractéristiques de son véhicule. Par ailleurs, l’algorithme créé en parallèle du formulaire ira récupérer les données renseignées par l’utilisateur afin de les traiter et fournir ainsi une estimation du prix du véhicule, ainsi que quelques statistiques complémentaires.

Le but de cette application est de pouvoir fournir une estimation rapide du prix d'un véhicule à partir du renseignement de caractéristiques basiques d’une voiture afin d’éviter le passage de plusieurs heures de recherches. Toutefois, cet outil n'a pas pour but de remplacer une estimation faite par un professionnel de l'automobile mais il doit pouvoir donner un ordre d'idée en comparant les différentes voitures entre elles.

# Comment construire un tel outil ?

## Les données

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

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

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

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

## Les variables

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

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

* **La Marque et le Modèle** : Il s'agit du nom complet de la voiture, comprenant la marque du véhicule et son modèle ;
* **Le Kilométrage** : Le nombre de kilomètres parcourus par le véhicule ;
* **Le Carburant** : Le type de carburant utilisé par la voiture, tel que l'essence, le gazole, l'électrique, etc ;
* **La Transmission** : Le type de boîtier du véhicule, pouvant être automatique ou manuel ;
* **L’Année** : L'année de production du véhicule.

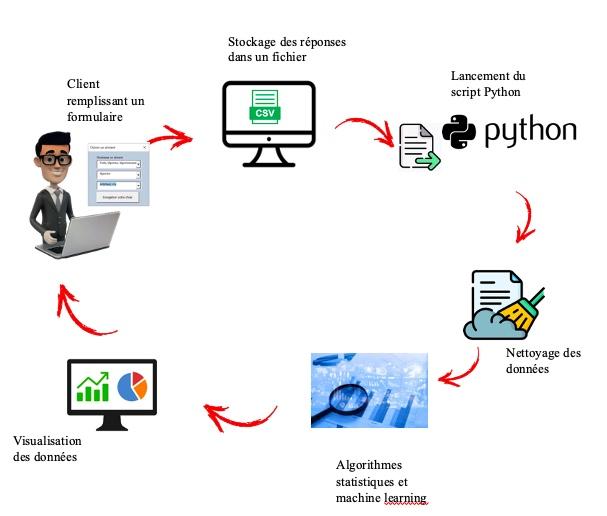
Cette sélection de variables explicatives va permettre de produire une estimation du prix du véhicule en fonction de ces caractéristiques.

Par la suite, nous avons enrichi notre ensemble de variables en ajoutant :

* Cylindrée
* Puissance physique
* Puissance fiscale
* Nombre de places
* Nombre de portes
* Nombre de vitesses

Ces variables additionnelles contribueront à affiner notre modèle d'estimation du prix en tenant compte de davantage de caractéristiques spécifiques du véhicule.

## L’architecture du projet



# L’avancée du projet

## Troisième sprint (du 22/10 au 09/11)

### Interface utilisateur

Interface d’entrée

L’utilisateur remplit le formulaire suivant :

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

La **marque** de la voiture doit être entrée à la main et une liste déroulante apparaît avec les premières lettres inscrites. L’objectif pour le prochain sprint c’est d’améliorer cette entrée en créant une sous partie “modèle”, à partir de la marque. L’utilisateur devrait entrer la marque, puis choisir son modèle de voiture.

Le type de **carburant** et de **transmission** sont renseignés par une liste déroulante. Pas de modifications prévues.

**L’année** de mise en circulation et le nombre de **kilométrage** sont entrés à la main, et seulement les valeurs numériques sont autorisées. Pour les prochains sprints, on pourrait bloquer à 4 le nombre de chiffres pour l’année de mise en circulation afin d’éviter toute erreur.

Interface de sortie

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

Description générée automatiquement

### Préparation des données

Afin d’avoir les données pour réaliser notre projet, nous avons décidé de scrapper les annonces de vente de voitures sur le site Spoticar afin d’avoir une base de données capable de se mettre à jour. À la suite des données récupérés, il a été nécessaire de préparer la base pour pouvoir faire les traitements notamment en séparant bien les marques et les modèles des véhicules. De plus un premier lien entre VBA et Python puis inversement a tenté d’être fait. Par manque de temps nous avons dû reporter cet objectif pour le sprint suivant.

### Algorithme de prévision

Afin de comprendre notre jeu de données, nous avons réalisé des **statistiques descriptives univariées** dans l’objectif d’avoir un regard neuf et de pouvoir s’orienter correctement sur la démarche à suivre sur le choix de la méthode de prédiction. Nous avons couplé ces premières statistiques avec une **analyse bivariée** en relation avec la variable cible qui est le prix et avec quelques regards sur la **corrélation** entre les variables où nous constatons un lien (**0.65**) entre **l’année** et le **kilométrage** ce qui est à priori évident car plus une voiture est ancienne plus le kilométrage risque d’être élevé.

Du fait que notre variable cible soit une variable quantitative, notre première hypothèse fut de partir sur une **régression linéaire multiple** en regardant seulement les variables explicatives quantitatives vu que nous n’avions pas les connaissances des encodeurs. Suite à cette première régression en sélectionnant toutes nos variables, nous avons pu obtenir un résultat avec un R² de **0.79**. Nous aurions pu partir sur une régression logistique mais pour cela une classification de notre prix en tranche de prix aurait été nécessaire cependant cela aurait créé un biais non négligeable et cela n’était pas envisageable aux vues de notre objectif, un client veut savoir une approximation de la valeur de sa voiture la plus précise possible.

## Quatrième sprint (du 10/11 au 23/11)

### Interface utilisateur

Ajout de nouvelles variables via le scraping : catégorie, cylindrée, nombre de places, nombre de portes, nombre de vitesses, puissance physique et fiscale (ces variables sont sélectionnées car explicatives selon AIC, BIC et les critères comme la couleur, le Critair sont enlevés).

Optimisation du questionnaire déjà existant : ajout d’une seconde page au questionnaire pour alléger. Feedback d’Arona Diene : option d’aide à la saisie pour la marque de voiture super. Voir si on peut ajouter une option pour naviguer entre les formulaires. Il préférait également une saisie de l’année via liste.

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

Description générée automatiquementUne image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

### Préparation des données

Une mise à jour du scraping a dû être fait pour avoir toutes les données nécessaires pour estimer correctement le prix d’un véhicule. En conséquence, le script de préparation a dû être mise à jour pour rendre utilisable certaines variables scrapées. À la suite des feedbacks du sprint précédent, nous avons vérifié la syntaxe des codes Python afin d’harmoniser le tout. Et le lien entre VBA et Python a été fait pour envoyer les résultats du formulaire dans notre base de données et pour pouvoir à nouveau re-scraper la base de données pour la mettre à jour.

### Algorithme de prévision

Ayant réalisé un nouveau scrapping, nous avons pu récupérer beaucoup plus de variable comme expliqué précédemment, nous avons donc pu faire des algorithmes de sélection de modèles par les critères de **AIC** et **BIC**, **Ridge**, où ces méthodes nous ont donnée les mêmes variables donc elles sont toutes importantes dans la prédiction future. Niveau qualité de prédiction, nous avons obtenu un R² de **0.83** pour la méthode de **AIC** et **BIC** et **0.88** pour le modèle **Ridge**, en complément, les erreurs prévisionnelles restent très basses. Nous n’avons pas eu le temps de regarder la prédiction par la méthode de pénalisation LASSO. Nous avons également essayé un premier jet de **RandomForest** où les résultats semblaient intéressant R² de **0.93** mais l’erreur prévisionnelle était trop élevée pour pouvoir l’utiliser. Une fois ces méthodes effectuées, nous les avons combinés avec la régression linéaire multiple donnant des résultats convenables. Cependant, il nous fallait vérifier le modèle avec les hypothèses de **normalité**, de **linéarité**, **d’autocorrélation** des erreurs et **d’homoscédasticité**, ces 4 hypothèses n’étaient pas vérifiées. Après quelques transformations de nos variables comme la variable cible en log et mise en log et sqrt sur des variables explicatives nous avons pu atteindre la linéarité avec le test de **Rainbow** et l’autocorrélation des erreurs **Durbin Watson**. Cependant les 2 dernières ne furent pas atteintes.

## Cinquième sprint (du 24/11 au 7/12)

### Interface utilisateur

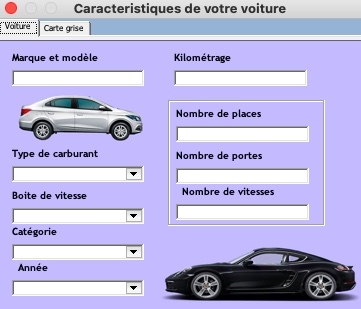
Nous avons introduit une nouvelle page d'accueil pour l’application.

Une image contenant texte, roue, véhicule, Véhicule terrestre

Description générée automatiquement

Interface d’entrée :

Par ailleurs, le formulaire a été repensé suite aux commentaires des utilisateurs.



**Interface d'entrée améliorée : Prise en compte des retours clients**

Suite aux retours des utilisateurs, des ajustements ont été apportés à l'interface d'entrée pour optimiser l'expérience utilisateur. Une **option multipage** a été intégrée, permettant une navigation plus fluide entre les différents formulaires. Les retours suggéraient également une préférence pour la saisie de l'année à travers une liste déroulante.

Par ailleurs, l’aide à la saisie pour la marque et le modèle de la voiture a été très apprécié par les utilisateurs. Cette option sera donc proposée pour le formulaire final.

**Nouveau formulaire multipage : Une réorganisation stratégique**

Le nouveau formulaire a été repensé pour une efficacité accrue, en particulier pour les utilisateurs pouvant fournir des informations de base sur leur voiture sans nécessairement avoir leur carte grise sous la main. La première partie du formulaire se concentre sur ces caractéristiques initiales, tandis que la seconde partie requiert des détails plus spécifiques, impliquant l'utilisation de la carte grise. Cette restructuration vise à offrir une présentation plus claire et ordonnée.

**Saisie facilitée de l'année avec une liste déroulante**

Désormais, la saisie de l'année ne se fait plus à la main mais s'effectue à l'aide d'une liste déroulante, répondant ainsi à la préférence exprimée par les utilisateurs.

**Assistance visuelle intégrée sur la carte grise**

Une nouvelle fonctionnalité a été ajoutée pour aider les utilisateurs à localiser facilement les informations cruciales sur leur carte grise. Lorsque l'utilisateur se trouve sur la case correspondant à la cylindrée, une image de la carte grise s'affiche, indiquant clairement l'emplacement de cette donnée. De même, des aides visuelles sont fournies pour la puissance fiscale et physique, simplifiant ainsi le processus de saisie. Ces améliorations visent à rendre le processus de remplissage du formulaire plus intuitif, tout en offrant un support visuel pour garantir la précision des informations fournies par l'utilisateur.

**Aide à la Saisie Dynamique pour Marque et Modèle**

L'aide à la saisie pour la marque et le modèle est désormais masquée jusqu'à ce que l'utilisateur commence à entrer le nom de sa voiture, offrant une liste de suggestions pour simplifier le processus. Une fois son choix effectué, la liste disparait et la marque et le modèle sont reportés dans la case de départ.

**Logique Intuitive : Adaptation du Formulaire en Temps Réel**

Une logique intuitive a été intégrée, telle que la disparition automatique de la case "nombre de vitesses" lorsque l'utilisateur indique posséder une voiture automatique. Ces ajustements visent à rendre le processus de saisie plus intuitif et agréable.

Ces améliorations globales sont conçues pour optimiser l'expérience de l'utilisateur tout en garantissant la précision des informations collectées.

Interface de sortie :

Une image contenant texte, capture d’écran, Police, Site web

Description générée automatiquement

En ce qui concerne l'interface de sortie, un tableau de bord a été créé, incluant deux graphiques statiques détaillant les caractéristiques de la base de données. Ces graphiques se mettent automatiquement à jour lorsqu'une modification est apportée à la base de données. De plus, un graphique dynamique compare le prix estimé de la voiture avec les moyennes des prix d'autres véhicules, affichant également le prix estimé en haut à gauche. Ces ajustements globaux visent à optimiser l'expérience utilisateur tout en assurant la précision des informations collectées.

### Préparation des données

Le lien entre VBA et Python permet d’envoyer le résultat du formulaire et de renvoyer ensuite un prix rendu dans l’interface de sortie.

Les premiers tests unitaires ont été effectués sur le script de data préparation pour vérifier la solidité et le bon fonctionnement de ces fonctions.

### Algorithme de prévision

Pour ce dernier sprint, nous avons essayer de valider les dernières hypothèses par de nouvelles techniques, en enlevant les **outliers**, en transformant nos variables par la **méthode de Cox** et en utilisant des **régressions robustes et pondérés**. Cependant nous n’avions pu atteindre l’homoscédasticité ni la normalité même si nous l’avons fortement approché avec le Test d’**Anderson Darling**. Heureusement pour nous, les dernières méthodes ont été réalisé. La régression **Lasso** nous donné un R² de **0.88** et une erreur prévisionnelle de **1%** ce qui est très acceptable mais c’est surtout la méthode de **RandomForest** qui a abouti au meilleur résultat, en effet nous avons perfectionné notre algorithme nous donnant un R² de **98%** et une erreur prévisionnelle de **10-6** nous avons couplé notre nouvelle méthode avec la matrice de corrélation pour évité les variables avec un trop gros lien et un encodeur qui pondère le poids de la marque de voiture car les données provenant de **Spoticar**, les voitures étaient essentiellement de provenance française. De plus, nous avons également vérifié les hypothèses même si le modèle par RandomForest n’a pas besoin de les vérifié pour obtenir des résultats fonctionnels mais toutes les hypothèses ont été soient vérifiées soient améliorées !