

Projet - Prédicteur de prix

Théo Lemaire - Aymane Ichou

April 25, 2024



Contents

1	Introduction	2
2	Scraper	2
2.1	Bibliothèques utilisées	2
2.2	Fonctionnement du scraper	2
3	Corpus	3
4	Modèle d'IA	4
5	Phase de test	4

1 Introduction

Dans le cadre du cours d'ingénierie des langues, nous avons implémentés en langage Python un modèle de prédiction de prix de voitures d'après le marché de l'occasion. De la réalisation de ce projet a découlé plusieurs sous-projets que nous allons détailler.

2 Scraper

Dans un premier temps, il a fallu réaliser un scraper qui récolte les données souhaitées avec le minimum de bruit possible. Nous avons sélectionné le site internet <https://www.autosphere.fr/>, car celui-ci permettait un scrapping efficace. D'une part, avec un nombre d'annonces correct (environ 15,000 annonces). Et d'autre part, un code HTML ordonné, facilitant le parcours des différentes balises des annonces sur chaque page.

2.1 Bibliothèques utilisées

Pour mener à bien ce projet, nous avons utilisés les librairies : Pandas, Sklearn, xgboost, datetime, math, scrapy

2.2 Fonctionnement du scraper

Le fichier "quotes spider.py" situé dans le sous-répertoire de "car price predictor/spiders/" est le fichier qui se charge du scrapping. Par défaut lors de son exécution, il récupérera l'ensemble des véhicules présent sur le site et les stockera dans le fichier "scrapped/". Pour chaque page de recherche sur site, nous récupérerons les liens d'annonces, et pour chaque annonce, nous récupérerons les attributs suivant d'un véhicule :

- marque
- modele
- couleur
- kilométrage
- boite de vitesse
- annee
- prix
- puissance fiscale
- carburant
- nombre de portes
- nombre de places

Nous avons sélectionné ces caractéristiques car ce sont celles qui impactent le plus le prix d'un véhicule d'occasion. A titre d'exemple, la puissance d'un véhicule impacte plus son prix que les options proposées sur celui-ci sur le marché de l'occasion.

A chaque annonce scrappé, nous récupérons les caractéristiques du véhicule sous la forme d'un fichier .json, qui sera ensuite placé dans le sous-repertoire "scrapped".

```
{
  "annee": 2023,
  "kilometrage": 19000,
  "places": 5,
  "portes": 5,
  "marque": "peugeot",
  "modele": "408",
  "couleur": "rouge",
  "boite_de_vitesse": "automatique",
  "puissance_fiscale": 7,
  "carburant": "essence"
}
```

Nous avons choisis le format .json afin d'effectuer plus facilement la serialisation de notre objet en Python. En effet, comme Json est un format très utilisé dans le monde du développement, des bibliothèques proposent directement des fonctions qui prennent en entrée un fichier Json, et retournent un objet Python crée à partir des caractéristiques présentes dans le fichier (fonction json.load).

3 Corpus

Une fois nos données textuelles créée, il a fallut créer un dataset pour le fournir à notre futur modèle. Pour ceci, nous avons choisis le format csv. Grâce à la bibliothèque "csv" de Python, nous avons pu récupérer l'ensemble des fichiers .json et les insérer dans notre "dataset.csv". Cependant, cela n'était pas suffisant. En effet, il a fallut traiter d'avantage nos données afin de retirer un maximum de "bruit" dans notre dataset.

```
// Pour le champ kilométrage
Ex: 12,000Km --> 12000.0
// Pour le nombre de portes / places
Ex: 5 portes, 5 places --> 5 5
```

Une fois le bruit enlevé, nous obtenons des données très satisfaisantes, avec à peu près la moitié de variables d'entrées catégorielles et l'autre moitié, quantitatives (Illustration partiellement représentative de la totalité des champs).

marque	modele	couleur	kilometrage	boite_de_vitesse	annee	prix_ttc
peugeot	308sw	grisartense(m)	38543.0	automatique	2023.0	27499.0
volkswagen	id.5	bleucrépusculemétallisée/toitnoir	9017.0	automatique	2022.0	47990.0
mercedes	classea	gris montagnemétallisé	27112.0	automatique	2021.0	35499.0
peugeot		2008 grisartense	79144.0	manuelle	2018.0	13499.0
peugeot		2008 grisartense	80482.0	manuelle	2017.0	11980.0
renault	clio	bleu	2810.0	manuelle	2022.0	18999.0
suzuki	across	platinumwhitepearlmetallic	5000.0	automatique	2023.0	49990.0
nissan	qashqai	noirmétallisé	71238.0	automatique	2019.0	21990.0

4 Modèle d'IA

Dans un premier temps, nous avons essayé de créer une phrase avec toutes les caractéristiques d'un véhicule. Pour ensuite utiliser un modèle de langue pré-entraîné, et l'ajuster pour notre utilisation. Cependant deux faits nous ont écartés de cette piste : premièrement, l'ajustement de ce modèle est extrêmement coûteux en ressources. Deuxièmement, nous avons un bon nombre de variables quantitatives comme le kilométrage, l'année, le nombre de porte/places. Ce type de modèle n'était donc pas adapté car il est plus performant avec des variables non catégorielles.

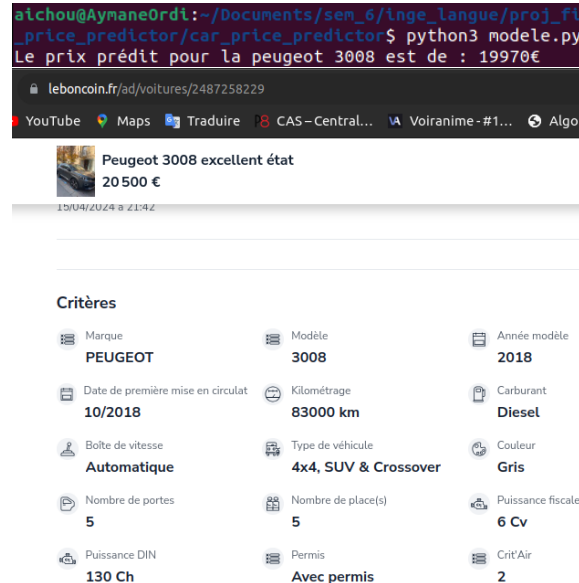
Pour notre modèle, nous avons donc fait le choix d'implémenter un modèle de régression XG-Boost, sur les conseils de notre professeur Mr Louis Falissard. Ce modèle est particulièrement bien adapté sur les données dites "tabulaires", ou quand il s'agit d'entraîner un modèle avec des objets avec des propriétés similaires. Ce type de modèle est beaucoup moins coûteux en ressources que la solution précédente.

Nous avons donc séparé nos variables d'entrée (catégorielles/quantitatives) grâce à un pré-processeur, puis nous avons instancié un "XGBRegressor()", et entraîné notre modèle. Nous ne prenons pas en compte l'erreur quadratique de notre modèle lors de son entraînement car lorsqu'on fait de la prédiction des données de cet ordre, même de petites erreurs relatives dans la prédiction des prix des voitures peuvent conduire à des erreurs quadratiques importantes.

5 Phase de test

Une fois notre modèle entraîné, nous avons pu le tester sur plusieurs modèles en comparant directement les prédictions de celui-ci aux prix du marché de l'occasion en France.

Prédiction pour un Peugeot 3008, on obtient une différence de 0.44 %



Terminal output:

```
aichou@Aymane0rdi:~/Documents/sen_6/inge_langue/proj_fl
_price_predictor/car_price_predictor$ python3 modele.py
Le prix prédit pour la peugeot 3008 est de : 19970€
```

Car listing details:

leboncoin.fr/ad/voitures/2487258229

Peugeot 3008 excellent état
20 500 €

12/04/2024 11:42


Critères		
Marque	Modèle	Année modèle
PEUGEOT	3008	2018
Date de première mise en circulation	Kilométrage	Carburant
10/2018	83000 km	Diesel
Boîte de vitesse	Type de véhicule	Couleur
Automatique	4x4, SUV & Crossover	Gris
Nombre de portes	Nombre de place(s)	Puissance fiscale
5	5	6 Cv
Puissance DIN	Permis	Crit'Air
130 Ch	Avec permis	2

Prédiction pour un Peugeot 408, on obtient une différence de 1.39 %

```
_price_predictor/car_price_predictor$ python3 modele.py
Le prix prédit pour la peugeot 408 est de : 30416€
```

leboncoin.fr/ad/voitures/2609011118

YouTube Maps Traduire CAS - Central... Voiranime - #1... Algorithmic

 Peugeot 408 1.2 PureTech 130ch S&S Allure Pack EAT8
29 999 €

Critères


Marque PEUGEOT	Modèle 408	Année modèle 2023
Date de première mise en circulat 04/2023	Kilométrage 7280 km	Carburant Essence
Boîte de vitesse Automatique	Type de véhicule Berline	Couleur Gris
Nombre de portes 5	Nombre de place(s) 5	Puissance fiscale 7 Cv
Permis Avec permis	Référence 3504-0973-40_MD1-5	Crit'Air 1
Durée de disponibilité des pièces Non renseignée		

Prédiction pour une Audi A3, on obtient une différence de 2.59 %

```
_price_predictor/car_price_predictor$ python3 modele.py
Le prix prédit pour la audi a3 est de : 23597€
```

leboncoin.fr/ad/voitures/2463151828

YouTube Maps Traduire CAS - Central... Voiranime - #1... Algori

 Vente de voiture
23 700 €

Critères

Marque AUDI	Modèle A3	Année modèle 2019
Date de première mise en circulat 12/2019	Kilométrage 53400 km	Carburant Essence
Boîte de vitesse Automatique	Type de véhicule Berline	Couleur Noir
Nombre de portes 5	Nombre de place(s) 5	Puissance fiscale 7 Cv
Puissance DIN 150 Ch	Permis Avec permis	