Julien OLLIVIER Hugo MAURINIER

Visualization of Massive Data

Bibliothèques externes

Pour ce projet, nous avons eu recours à différentes bibliothèques Python :
- Numpy pour la génération de datasets
- Pandas pour la lecture et l'écriture de datasets dans des fichiers cdv
- matplotlib, seaborn et plotly pour pour la création de graphiques
- scikit-learn pour l'apprentissage automatique et la construction du modèle de régression linéaire

Génération d'un jeu de données

Tout d'abord, nous avons généré un jeu de données à l'aide de la bibliothèque plotly.
Il se compose de 300 lignes qui comprennent les colonnes suivantes :
- 4 colonnes suivants des lois normales aux moyennes respectives de 2.5, 78, -67 et 457, et suivant la matrice de covariance suivante :

1	0	-75	-10
0	1	42	2345
-75	42	1	689
-10	2345	689	1

- 1 colonne comprenant des entiers compris entre 0 et 79
- 1 colonne suivant une loi normale de moyenne 56 et d'écart-type 2.5

Analyse

Pour la suite du projet, nous avons choisi un dataset comprenant un total de 72435 véhicules vendus au Royaume-Uni définis par un id, le nom du modèle, l'année de production, le prix de vente, le type de transmission, le kilométrage (en miles), le type de carburant, la taxe routière, la consommation (en galon par mile), la cylindrée et la marque. Notre but était d'utiliser ces critères pour estimer le prix d'un véhicule.

Tout d'abord, les histogrammes des variables quantitatives montrent la présence de nombreuses valeurs aberrantes pour le prix, le kilométrage, la taxe routière et la consommation des véhicules. Ainsi, à l'aide de la règle interquartile, et en ignorant le millième quantile de kilométrage, nous excluons ces lignes du jeu de données. De la même façon, nous remarquons qu'il existe des cylindrées de 0cm3, nous excluons donc également ces véhicules car ils ne peuvent pas exister.

Ensuite, en listant le volume des données catégoriques (type de transmission, type de carburant et marque) on remarque les présences de types de transmission et de carburant « Autres ». Nous les excluons également du jeu de données.
Nous pouvons également remarquer que le prix semble dépendre de la marque du véhicule mais également de son type de transmission.

Enfin, la matrice de corrélation des variables quantitatives montre, entre autres, les corrélations positives entre le prix et l'année de production mais également la cylindrée, et corrélations négatives entre le prix et le kilométrage ainsi que la consommation.

Visualisation

Nous avons ensuite choisi de représenter ces données sous la forme d'un graphique de coordonnées parallèles représentant le prix en fonction du kilométrage, de la consommation, de l'année de production, de la cylindrée et de la taxe routière, et d'un diagramme de dispersion représentant le prix en fonction du kilométrage du véhicule et dont la couleur des points représente la marque et leur taille la cylindrée, afin de mieux représenter les corrélations entre ces critères.
Nous pouvons choisir de représenter les valeurs dont l'id est compris entre deux valeurs minimum et maximum, et/ou un échantillon aléatoire de taille définie.

Nous pouvons ainsi visualiser clairement les fortes corrélations entre le prix et le kilométrage, la cylindrée et l'année de production, mais également entre le prix et la marque.

Régression linéaire

Tout d'abord, nous préparons le jeu de données entraîner notre modèle.
Nous remplaçons les données catégoriques par des indicateurs numériques et nous mettons les valeurs numériques de toutes les colonnes à l'échelle.
Nous prélevons ensuite 30% des valeurs de notre ensemble d'apprentissage afin de former un ensemble de test.

Une fois que tout est prêt, nous construisons notre modèle de régression linéaire grâce à scikit_learn et nous en déduisons les prix de notre ensemble de test avant d'évaluer la qualité des résultats obtenus.

Nous obtenons tout d'abord un coefficient de détermination de 0.8397769774989262, ce qui signifie que notre modèle explique quasiment 84% des valeurs.
Nous observons également une erreur absolue moyenne de 2117.653813580215, la moyenne des prix de notre ensemble de test s'élevant à 15828.435695815082, nous obtenons une erreur absolue moyenne de 13.3787940532%. Notre modèle permettrait donc d'obtenir, en moyenne, une valeur aux alentours de 13.4% de la valeur réelle.
Nous observons également un écart quadratique moyen de 7620718.443822252, ce qui donne une erreur quadratique moyenne de 2760.56487767. Nous obtenons donc cette fois une erreur quadratique moyenne de 17.4405413821%, ce qui indique que nous sommes capable d'estimer une valeur aux alentours de 17.4% de la valeur réelle.

Par conséquent, nous pouvons estimer le prix d'un véhicule d'occasion de façon acceptable en fonction de ces paramètres, mais il existe probablement d'autres paramètres qui rentrent en jeu lors d'une transaction de ce type.