Projet Car_prices

Part 3

1. Régression linéaire univariée (ou simple)

Récupérer le jeu de données issu de l'analyse précédente (part 2)

L'objectif, ici, dans un premier temps, est d'établir les relations linéaires existantes entre certaines variable et la cible « price »

Rappel :

Une relation linéaire est une équation du type : $Y = aX + b$

Y est la cible
a est le coefficient directeur ou « slope »
X est la variable choisie
b est l'ordonnée à l'origine ou « intercept »

- Importer la classe *LinearRegression* depuis la librairie *sklearn.linear_model*
- Créer une instance de la classe *LinearRegression*
- Utiliser la méthode *fit*() pour appliquer le modèle à un couple X, Y (X étant une des caractéristiques et Y la cible)
- Récupérer les coefficients de l'équation via les attributs *intercept_* et *coef_* de l'objet *LinearRegression* et écrire l'équation finale obtenue donnant $\hat{Y}$
- Utiliser ensuite la méthode *predict()* pour estimer une valeur de la cible
- Utiliser la méthode *residplot()* de seaborn pour visualiser la distribution des valeurs résiduelles (écarts entre les valeurs réelles et celles prédites par le modèle). Le nuage de points obtenu doit ête uniformément réparti autour d'une ligne horizontale pour confirmer la régression linéaire.
- Répéter les opérations pour d'autres caractéristiques

2. Régression linéaire multivariée (ou multiple) :

L'étude part 2 aura permit de sélectionner les meilleurs prédicteurs parmi les 26 colonnes du dataset.

À titre d'exemple, sélectionner ici, les colonnes Horsepower, Curb-weight, Engine-size et Highway-L/100 en créant un nouveau dataFrame X.

- Appliquer la méthode *fit()* pour entrainer le modèle de régression linéaire sur le jeu de X, Y
- Utiliser ensuite la méthode *predict()* pour tester le modèle
- Extraire les attributs coef_ et intercept_ et écrire l'équation finale obtenue donnant $\hat{Y}$

Comme il est difficile de visualiser la régression obtenue, on a souvent recours à la comparaison entre les valeurs réelles et celles présites.

- Utiliser la méthode distplot() de seaborn pour représenter la distribution de Y et $\hat{Y}$

3. Evaluation et comparaison

Deux mesures très importantes sont souvent utilisées dans les statistiques pour déterminer la précision d'un modèle :
R^2 et l' Erreur quadratique moyenne (MSE)

R carré, également appelé coefficient de détermination, est une mesure indiquant la distance entre les données et la ligne de régression ajustée.
La valeur du R au carré est le pourcentage de variation de la variable de réponse (y) représentée par un modèle linéaire.
Elle est donnée par la méthode *score()* du modèle

Erreur quadratique moyenne (MSE)
L'erreur quadratique moyenne mesure la moyenne des carrés d'erreurs, c'est-à-dire la différence entre la valeur réelle Y et la valeur estimée $\hat{Y}$.
Elle peut être calculée en utilisant la méthode *mean_squared_error()* qui doit être importée de la librairie sklearn.metrics.

Calculer les couples de metrics R^2 / MSE pour les régressions linéaires univariées et comparer la meilleure valeur obtenue au couple obtenu pour la régression multivariée.

À vous de jouer pour trouver la meilleures combinaison …

Bon weekend ☺