Cas d'Usage
Taxis de New York

Ceci est un exercice simple sur la régression linéaire, permettant une première prise en main de librairies usuelles de gestion de Data et de Machine Learning.

Objectif :

En utilisant une base de données historiques des trajets de taxi observés pendant un mois, on veut prédire, en fonction de l'heure de départ, et de la longueur du trajet, la durée prévisionnelle du trajet (variante : on pourrait aussi prédire le prix du trajet)

Rappel sur la régression linéaire :

Étant donné une suite de N observations (x_i, y_i) on cherche à construire un modèle linéaire approximant l'observation, c'est-à-dire, les paramètres optimaux a, b pour le modèle : $y_i \approx a x_i + b$.

La régression linéaire permet de faire cela.

Si les observations sont (x_i, y_i) et les valeurs estimées par la régression linéaire sont $(x_i, \hat{y}_i)$, avec $\hat{y}_i = a x_i + b$, l'algorithme de régression linéaire cherche à minimiser la distance moyenne entre les observations réelles et les grandeurs estimées, c'est à dire que les paramètres a, b minimisent la grandeur :

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

Une autre manière d'interpréter la régression linéaire est d'affirmer qu'elle maximise la grandeur :

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sigma^2}$$

où $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ est la moyenne des observations, et $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$ leur variance.

Une analyse rapide de R^2, appelé score de la régression linéaire, permet de constater qu'il peut prendre toutes les valeurs possibles, positives ou négatives. Cependant, si la régression est fiable, R^2 est compris entre 0 et 1, et plus il est proche de 1, plus le modèle est fiable. En d'autres termes : un modèle qui remplacerait toutes les observations par leur moyenne, aurait un score nul. Un modèle qui prédirait exactement les observations aurait un score égal à 1. On peut résumer cette grandeur en disant que le score R^2 mesure la manière dont notre modèle prédit les oscillations de y autour de sa moyenne.

Régression linéaire : les Taxis de New York

Cette section décrit pas-à-pas les étapes à réaliser pour modéliser les données historiques des trajets des taxis de New York, dans le but de construire une régression linéaire permettant de prédire la durée d'un trajet en fonction de la distance à parcourir.

1. Sur le site de la ville de New York, récupérer les données historiques des trajets de taxis lors d'un des mois de votre choix, sur l'année 2022 (pour éviter la période Covid) : en suivant ce lien.
   Note : Au bas de la page, vous avez des informations complémentaire vous permettant de comprendre les données (codes divers, …).
   Remarque : Le fichier est au format PARQUET (Apache) dont vous pouvez obtenir une description et les avantages caractéristiques le le Web.

2. Créer un Notebook JupyterLab, puis, en utilisant la bibliothèque Pandas, charger ce fichier dans un dataframe. Attention à tenir compte de son format.

3. Explorer le dataframe, afficher les valeurs statistiques intéressantes, courbes (en utilisant la bibliothèque matplotlib), …

4. Repérer les champs (colonnes) qui nous intéressent et les mettre au format qui nous intéresse, en ajoutant éventuellement des colonnes.

5. En utilisant la régression linéaire de la bibliothèque sklearn, effectuer une modélisation des trajets de taxi, et afficher le score (R^2) de cette modélisation. Que constatez-vous ?

6. Reprendre l'étape 5, en faisant une régression linéaire pour chaque créneau horaire. En d'autres termes, pour chaque heure H comprise entre 0 et 23 heures, on analyse les trajets qui ont eu lieu entre H et H+1, et on fait une régression linéaire uniquement sur ces trajets, pour pouvoir avoir un modèle pour chaque heure de la journée. Est-ce que cela améliore les résultats ?

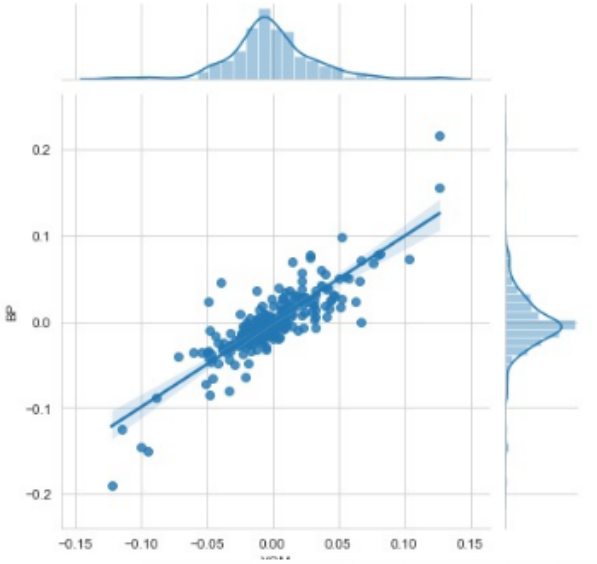
7. Visiblement, le fichier de données fourni par la ville de New-York comporte des données de mauvaise qualité. On propose de le « nettoyer » :
   - Extraire du dataframe initial un dataframe ne contenant que les informations qui nous intéressent : heure de départ, distance, durée, en créant si nécessaire de nouvelles colonnes.
   - Enlever les lignes comportant des données vides (NaN)
   - Retirer les lignes aberrantes : durée supérieure à D heures ou distance supérieure à X km (D et X à choisir)

8. Reprendre l'étape 6. Que constatez vous ?

9. Pour chaque créneau horaire, afficher les paramètres obtenus pour la régression linéaire (coefficient : reg.coef_ et ordonnée à l'origine : reg.intercept_ , si reg est l'objet obtenu après exécution de la régression).

10. Afficher en utilisant matplotlib une courbe qui donne le coefficient R^2 obtenu pour chaque créneau horaire.

11. En utilisant la bibliothèque seaborn, afficher pour chaque créneau horaire un graphique de ce type :



Dans ce type de courbe,

- la figure centrale illustre la droite de régression en comparaison avec les observations
- la figure au dessus donne la distribution des observations de la grandeur en abscisse,
- la figure à droite donne la distribution des observations de la grandeur en ordonnée

12. Comment pourrait-on améliorer encore les résultats ?