Berrandou Assia

https://github.com/Assiab2707/packaging_challenge_brda

Dans cette partie, il s'agissait de prédire le trafic de vélos à Montpellier le vendredi 02 Avril entre 00h01 et 09h.

I Préparation & découverte de données

Nous disposons d'un fichier *.csv* contenant les données journalières depuis le 12 mars 2020 de l'intensité du trafic de vélos à Montpellier. Les données sont recensées une à plusieurs fois par jour. Après les avoir téléchargeées et mises sous forme de dataframe, j'ai effectué différents traitements sur ces dernières.

Dans un premier temps, j'ai nettoyé les données : changement du nom des colonnes de la dataframe /suppression des colonnes « Unamed » et « Remark »/ suppression des lignes dont il manque des valeurs/ formatage de la date (format = AAAA-MM-DD hh:mm:ss) / calcule des données journalières (i.e entre 00h01 et 23h59 de chaque jour).
A la suite de cela, je les ai visualisées de différentes manières(graphiques des données entières/ graphiques des données sans les week-end/ graphique des données en fonction des jours) puis analysées afin de remarquer les différentes tendances. Par exemple, on peut voir que les données datant du premier confinement sont plus faibles que les autres. Ou bien encore que l'intensité du trafic les Samedis et Dimanches est beaucoup moins élevée que celle des autres jours.

Dans un second temps, je me suis intéressée aux données dont l'heure est comprise entre 00h01 et 09h00 (une donnée pour chaque jour i.e on prend le max des heures comprises entre 00h01 et 09h00 pour chaque jour). Je les ai visualisées de la même manière et j'ai analysé les composantes statistiques de chaque jour (i.e moyenne/médiane/min ….). Les conclusions précédentes sont les mêmes, cependant on remarque que nous ne disposons pas d'énormément de données.

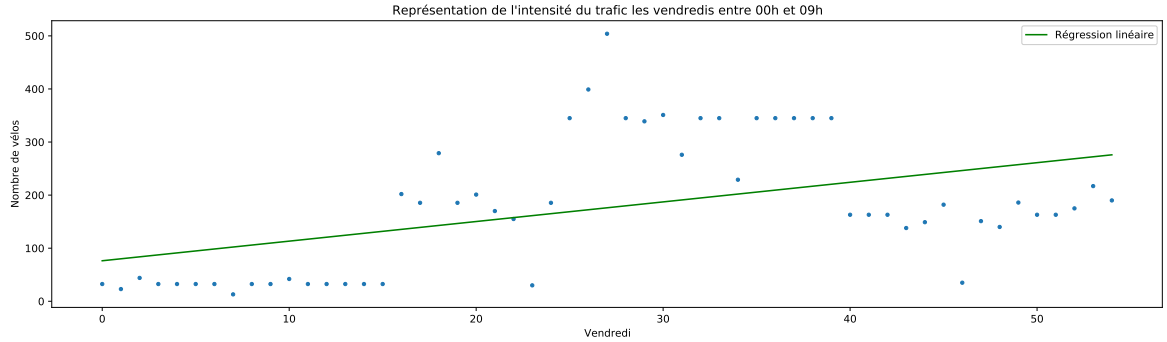
Pour finir, j'ai crée une dataframe qui contient les données des vendredis dont l'heure est comprise entre 00h01 et 09h00 (26 lignes). Parfois, certains vendredis ont leur compteur relevé après 09h, c'est pourquoi elles n'apparaissent pas dans la dataframe. Pour remédier à ce problème et retrouver une continuité temporelle, j'ai ajouté ces données manquantes et leur ai affectées une valeur particulière.
Cette valeur est définie en fonction la date et de la médiane de la dataframe entre 2 intervalles.
Par exemple, s'il manque le vendredi 2 avril, je le rajoute dans la base en lui affectant la médiane de la dataframe à l'intervalle de temps [12 mars ; 1 juillet].

J'applique sur cette dernière le traitement qui va suivre.

III Régression linéaire

Berrandou Assia

J’ai donc décidé de commencer par appliquer une régression linéaire sur la dataframe complète contenant l’intensité du trafic/intensité médiane de chaque vendredi entre 00h01 et 09h00.



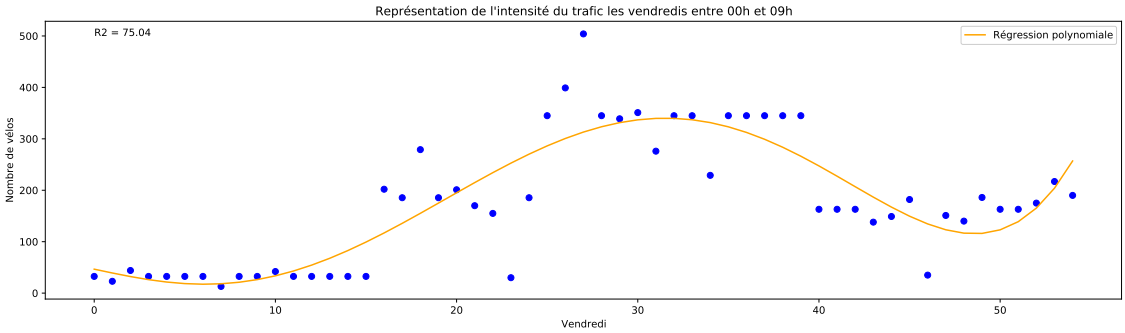
Créer un modèle linéaire à partir de ces données est simple. Cependant, on peut s’apercevoir que nos données ne possèdent pas de relation linéaire i.e il y a une tendance qui n’est pas linéaire et que le R^2 est significativement bas. C’est pourquoi une approche par une régression polynomiale m’a semblé plus judicieuse.

La régression polynomiale est une approche statistique qui est employée pour modéliser une relation de forme non-linéaire entre la réponse y qui est l’intensité du trafic et la variable explicative x qui représente nos vendredis.

On approche donc ces données par une fonction polynomiale de degré ≤ *n* d’équation :

$$y = a_n x^n + \ldots + a_0.$$

Le package **sklearn**, nous aide à calculer cette équation de régression polynomiale de degré n = 5 du nuage de points. On obtient ainsi le graphique ci-dessous :



On s’aperçoit que l’erreur de prédiction est moins élevée que celle d’une régression linéaire simple et que notre droite de régression s’ajuste mieux à nos données. De plus, le R^2 est alors égale à : 75,04 % i.e ce modèle explique 75 % des variations de l’intensité du trafic sont expliquées par le jour. On peut émettre l’hypothèse que les 25 % restant sont dus principalement aux conditions météorologiques/sanitaires. Le modèle semble correct.

On peut donc prédire, à partir de ce modèle, le nombre de vélos qui passeront entre 00h01 et 09h00 le vendredi 02 Avril, ce qui nous donne **326** vélos.