

Prédiction de recherches Google



Problématique

- ☐ Saturation des messages marketings reçus par chacun d'entre nous
 - Mail / SMS / sites web / pubs télé et radio / panneaux / vitrines / etc
- ☐ Multiplication des promotions en tous genres
 - Soldes / ventes privées / black Friday / french days / etc
- ☐ Pour se démarquer, un annonceur a besoin de faire passer le bon message, au bon endroit et au bon moment
 - « Marketing du contexte »

Problématique

- ☐ Pour arriver à ce marketing du contexte, il est nécessaire de connaître l'appétence du consommateur à l'instant t
- ☐ Exemple du barbecue et du climatiseur :
 - On se doute que tous les deux se vendent mieux en été par (très) beau temps
- ☐ Mais peut-on affiner cette perception instinctive, et tenter de connaître à l'avance les moments exacts de l'année les plus propices à la vente ?
 - C'est ce que nous allons essayer de faire ce soir !

Récupération de l'historique google trends

 <https://trends.google.fr> :

- permet de consulter, en libre accès, l'historique de recherche d'un mot ou d'une expression

 Avantages de l'utilisation de Google trends :

- Données non liées à une marque
- Les personnes font la démarche de chercher => réelle appétence

 Inconvénients :

- Recherche ne signifie pas achat
- La robustesse des informations dépend du volume de recherche

Récupération de l'historique google trends

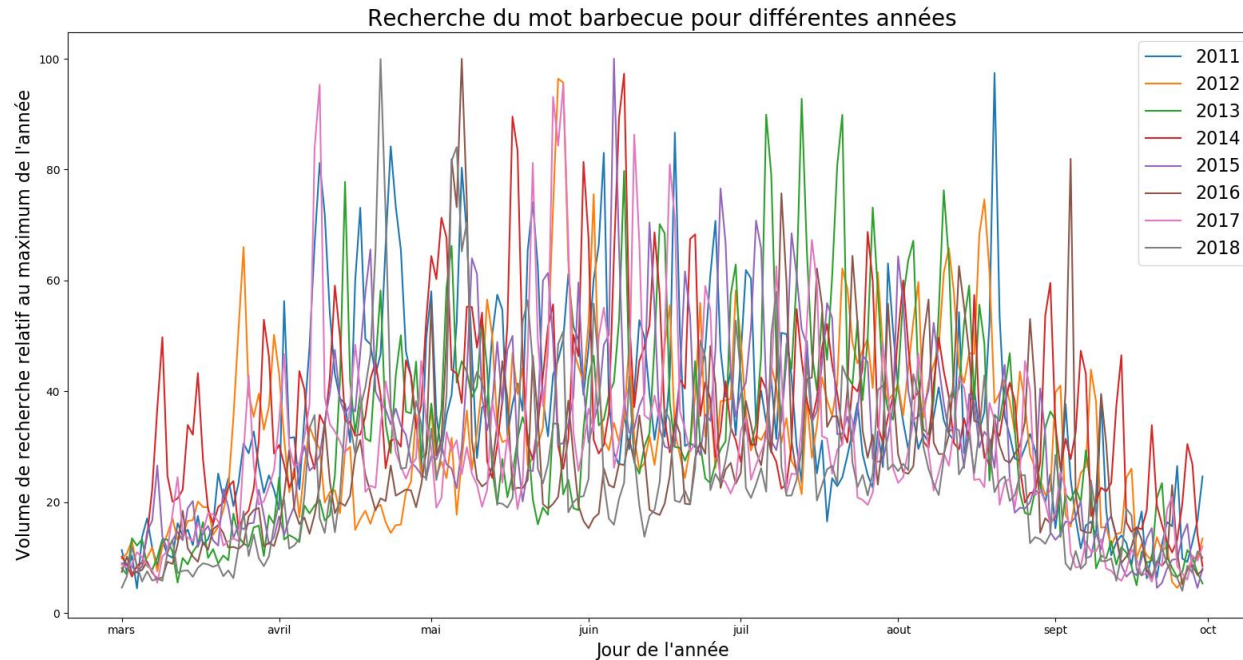
Exemple de données récupérées :



Pour chaque période extraite, les valeurs fournies sont relatives au maximum de cette période, lui-même ramené à 100

Première investigation calendaire

Les années se suivent mais se ressemblent-elles ?

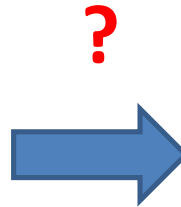


Présence d'une tendance générale, qui se répète d'années en années, mais aussi de pics locaux, qui eux n'apparaissent pas au même moment selon l'année


Introduction à la démarche du machine learning

Construction du tableau des prédicteurs X et de la grandeur d'intérêt y

Index	dayOfWeek	dayOfMonth	dayOfYear
2011-03-01 00:00:00	1	1	60
2011-03-02 00:00:00	2	2	61
2011-03-03 00:00:00	3	3	62
2011-03-04 00:00:00	4	4	63
2011-03-05 00:00:00	5	5	64
2011-03-06 00:00:00	6	6	65
2011-03-07 00:00:00	0	7	66
2011-03-08 00:00:00	1	8	67
2011-03-09 00:00:00	2	9	68
2011-03-10 00:00:00	3	10	69
2011-03-11 00:00:00	4	11	70
2011-03-12 00:00:00	5	12	71
2011-03-13 00:00:00	6	13	72
2011-03-14 00:00:00	0	14	73



Index	Recherche
2011-03-01 00:00:00	6.04
2011-03-02 00:00:00	7.4
2011-03-03 00:00:00	5.28
2011-03-04 00:00:00	0
2011-03-05 00:00:00	14.2
2011-03-06 00:00:00	12.44
2011-03-07 00:00:00	4.4
2011-03-08 00:00:00	3.8
2011-03-09 00:00:00	0
2011-03-10 00:00:00	8.28
2011-03-11 00:00:00	5.8
2011-03-12 00:00:00	8.64
2011-03-13 00:00:00	7.6
2011-03-14 00:00:00	12.8

 Objectif : construire la relation $y=f(X)$ entre la grandeur d'intérêt et le tableau des prédicteurs (on parle ici d'apprentissage supervisé)

Introduction à la démarche du machine learning

Quelques méthodes parmi les plus classiques :

- Régression Lasso, Ridge
 - Machines à vecteurs supports
 - Krigage
 - Arbres de décision, RandomForest, GradientBoosting
 - Réseaux de neurones
-
- Et bien d'autres !

Introduction à la démarche du machine learning

Un piège à éviter : le surapprentissage

- Séparation en jeu d'entraînement et jeu de généralisation (très important !)
- Ici, on va entraîner les modèles sur les années 2011 à 2017
- Et les tester sur l'année 2018

Introduction à la démarche du machine learning

Validation croisée pour le réglage des hyperparamètres

- Durant la phase d'apprentissage, on va optimiser les performances des modèles en jouant sur leurs hyperparamètres
- On va ajouter une nouvelle couche de séparation en jeu d'entraînement et de test en multipliant les scénarii ou chaque année d'entraînement va successivement devenir une année de test

Introduction à la démarche du machine learning

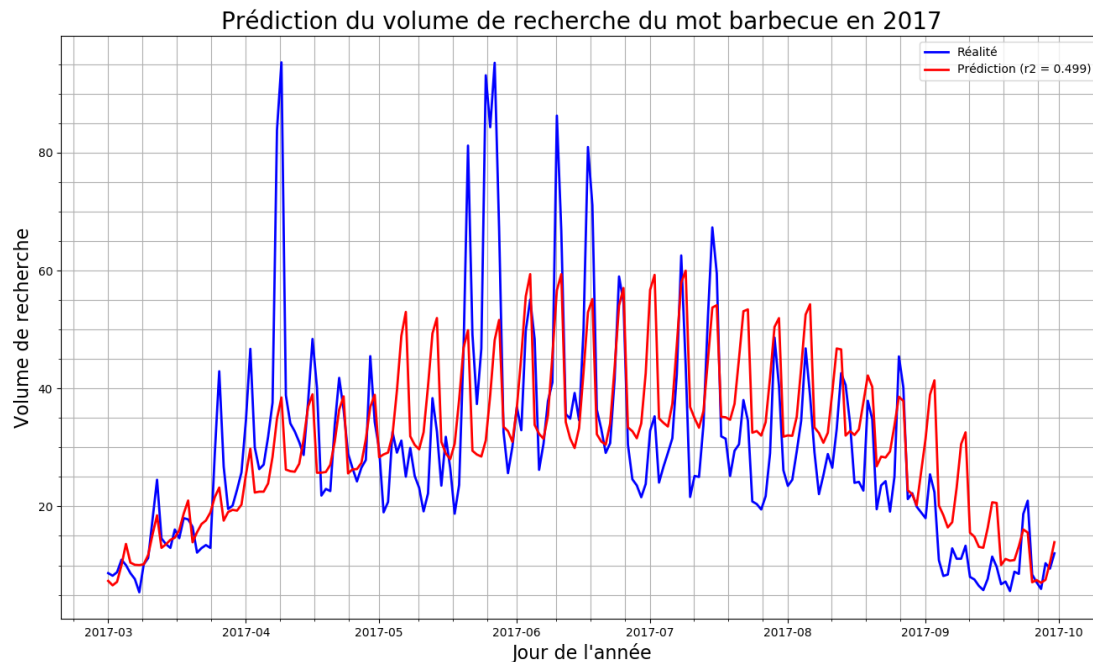
Comment évaluer la performance de mon modèle prédictif ?

- D'une part, on obtient une liste de score de prédiction issue de la validation croisée au sein du jeu d'entraînement
 - Cette liste nous permet d'évaluer une performance moyenne ainsi qu'une variabilité de cette performance
- D'autre part, on obtient le score de prédiction sur l'année de généralisation
- On observe alors ces deux sources d'information pour valider, ou non, notre modèle prédictif

Premières tentatives de prédiction



Prédictions faites uniquement sur la base des infos calendaires



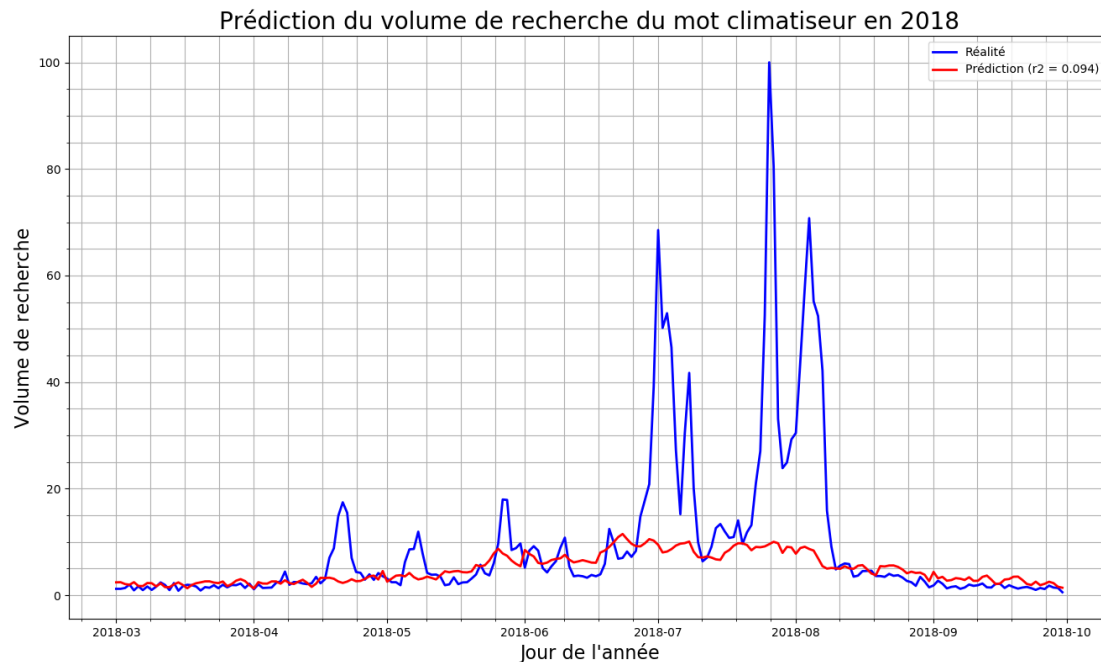
La prédiction nous donne la tendance saisonnière et l'alternance semaine / week-end mais rate les pics et creux locaux

➤ On a bien capté ce qui se répète d'années en années, mais pas les phénomènes non calendaires

Premières tentatives de prédiction




Prédictions faites uniquement sur la base des infos calendaires



Impossible de trouver un motif calendaire

- Le modèle ne prédit quasiment rien d'intéressant
- On ne s'avoue pas vaincu pour autant !

Ajout de nouveaux prédicteurs

-  On propose d'ajouter des données météo
- Base Météo France SYNOP (données observées et/ou mesurées)
 - ~40 villes dont Orly, Clermont-Ferrand, Lyon, Toulouse, Lille, etc
 - On récupère la température en °C, la pluie en mm et la nébulosité en %
 - Données disponibles toutes les 3H mais ramenées à la journée

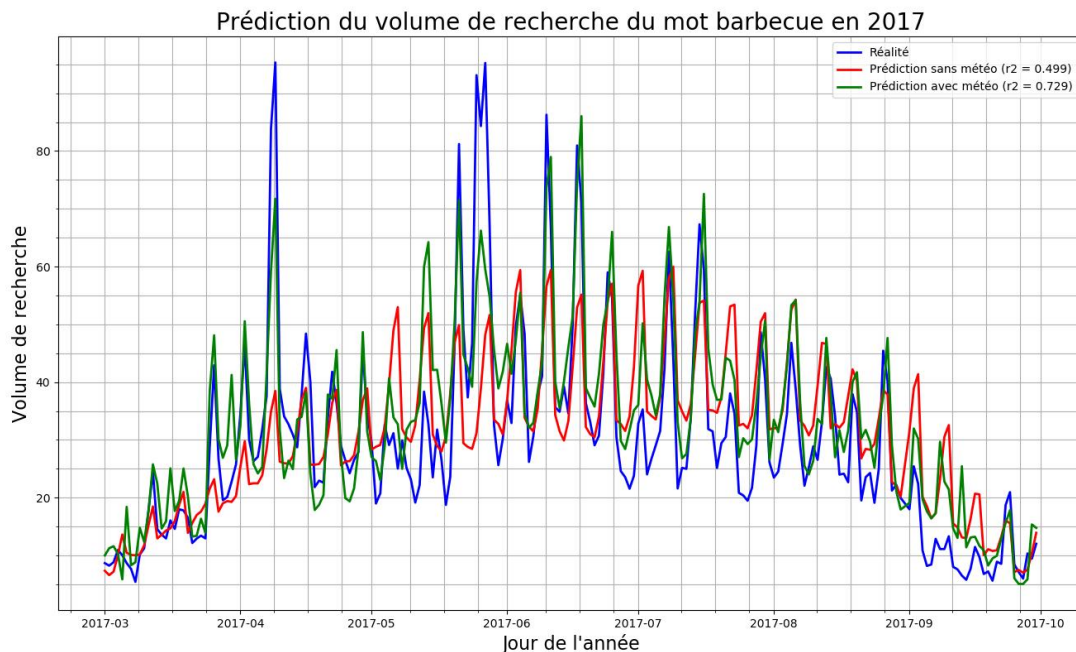
Ajout de nouveaux prédicteurs

Nouveau tableau des prédicteurs

Index	met_pluie	met_tempe	met_nebu	met_Trel	dayOfWeek	dayOfMonth	dayOfYear
2011-03-01 00:00:00	0	4.2625	1.25	-1.90875	1	1	60
2011-03-02 00:00:00	0	2.15	11.875	-4.3075	2	2	61
2011-03-03 00:00:00	0	1.9125	66.875	-4.9525	3	3	62
2011-03-04 00:00:00	0	3.9125	70.3125	-3.00625	4	4	63
2011-03-05 00:00:00	0	4.3625	95.3125	-1.08625	5	5	64
2011-03-06 00:00:00	0	2.925	88.75	-2.19125	6	6	65
2011-03-07 00:00:00	0	4.9125	96.875	-0.8375	0	7	66
2011-03-08 00:00:00	0	8.0875	97.5	0.755	1	8	67
2011-03-09 00:00:00	0	7.675	80	0.42125	2	9	68
2011-03-10 00:00:00	0	8.55	85	1.0925	3	10	69
2011-03-11 00:00:00	0	7.6125	47.5	-0.65125	4	11	70
2011-03-12 00:00:00	0.125	9.325	8.75	1.3	5	12	71
2011-03-13 00:00:00	2	7.7875	10	0.65375	6	13	72
2011-03-14 00:00:00	0	9.45	29.375	1.56	0	14	73

Nouvelles prédictions

Prédictions faites sur la base des infos calendaires et météo



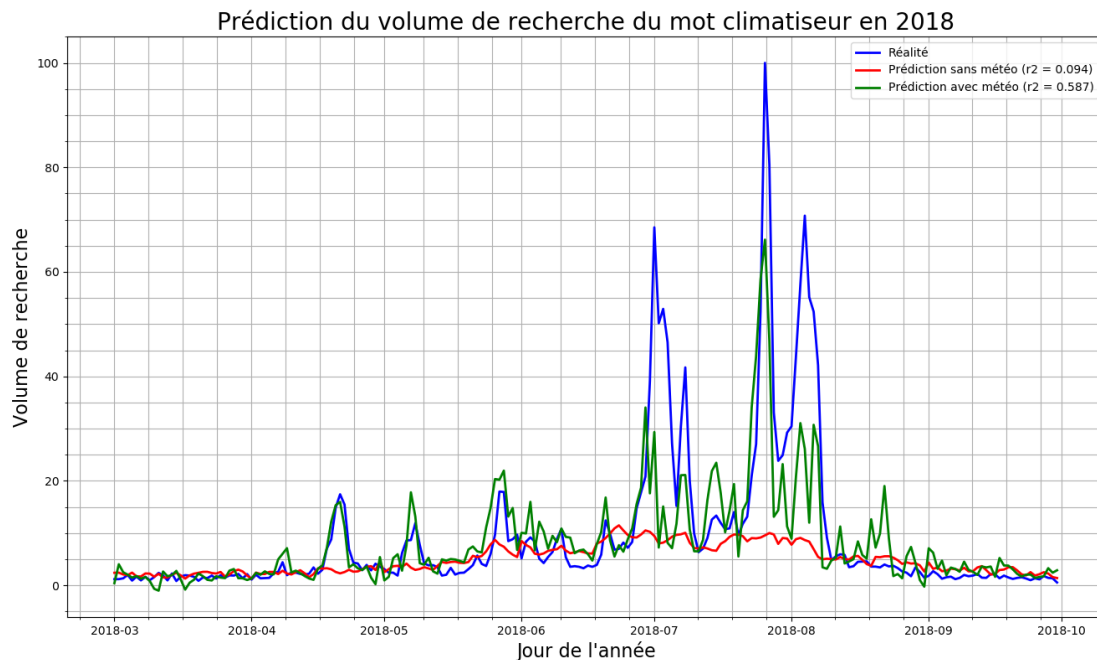
Bien meilleure prédiction de notre variable d'intérêt

On garde la tendance saisonnière et l'alternance semaine / week-end

Auxquels viennent s'ajouter des phénomènes locaux induits par la météo

Nouvelles prédictions

Prédictions faites sur la base des infos calendaires et météo



Bien meilleure prédiction de
notre variable d'intérêt

On voit apparaitre des
phénomènes locaux induits par
la météo

Hiérarchisation des prédicteurs

☐ Selon le modèle utilisé, il est plus ou moins aisé de remonter à l'influence de chaque prédicteur

- Par exemple, le random forest nous donne le poids de chaque prédicteurs dans sa décision finale

➤ barbecue

Index	0
dayOfYear	0.320777
met_tempe	0.231171
dayOfWeek	0.199145
met_nebu	0.120038
met_Trel	0.0682883
dayOfMonth	0.0310728
met_pluie	0.0295084

➤ climatiseur

Index	0
met_tempe	0.654065
dayOfYear	0.110774
dayOfMonth	0.0685515
met_nebu	0.0600982
met_Trel	0.0534243
met_pluie	0.0323346
dayOfWeek	0.0207517

Comment aller plus loin ?

Ajouter d'autres prédicteurs ?

- Attention au fléau de la dimension et au surapprentissage

Évaluer d'autres algorithmes ?

- Algorithmes génétiques, autres ?

Évaluer la pertinence d'un modèle de classification ?

- Prédire si tel jour sera défavorable / normal / favorable

La question reste ouverte !

Limites de cette méthode

La prédiction ne sera jamais parfaite

- Variabilité intrinsèque des comportements humains
- Bruit contenu dans les données google trends
- Historique disponible pas si long que ça
- etc

Dans le futur, le contexte, la réglementation, la mode, ou d'autres facteurs peuvent changer

- Notre modèle sera-t-il encore valide si demain le barbecue devient ringard ? Ou si les climatiseurs sont limités en puissance pour des questions d'économie d'énergie ?

Conclusions

- Nous avons vu deux apports de la data science :
 - Prédire une grandeur d'intérêt
 - Hiérarchiser les facteurs influents de cette grandeur
- Nécessité pour cela d'un historique significatif, d'autant plus volumineux que le nombre de prédicteurs est important
 - ~ quelques dizaines X le nombre de prédicteurs
- Il n'y a pas de recette miracle (« no free lunch »)
 - Aucun algorithme n'est plus puissant que les autres à tous les coups
 - Chaque cas est unique et demande une phase d'investigations