

Concevez une application au service de la santé publique

Data & Analytics
Eric Blanvillain - 16-11-2021





Problématique

Mon rôle :

- L'agence "[Santé publique France](#)" a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation. Je souhaite y participer et proposer une idée d'application



Les points à aborder :

1. Présentation de la donnée de Santé Publique France
2. Comment traiter la donnée (cleaning/feature engineering) -> création du dataset de test
3. Présentation des indicateurs clefs (nutri-score et éco-score)
4. Présentation de mon application
5. Etude du modèle de prédiction "potentiel" pour évaluer l'éco-score
6. Pour aller plus loin

Présentation de la donnée

Présentation par Santé publique France :

“Open Food Facts est une base de données sur les produits alimentaires, faite par tout le monde, pour tout le monde. Elle permet de faire des choix plus informés, et comme les données sont ouvertes (open data), tout le monde peut les utiliser pour tout usage.”

Liste des variables* :

# general information	-> 10
# tags	-> 24
# ingredients	-> 3
# misc. data	-> 16
# nutrition facts	-> 89
	<u>Total</u>
	142

2 indicateurs clefs :

- eco-score (carbon footprint)
- nutrition-score

nombre de produits : 290517
nombre de catégories : 63359

-> 4GB de donnée

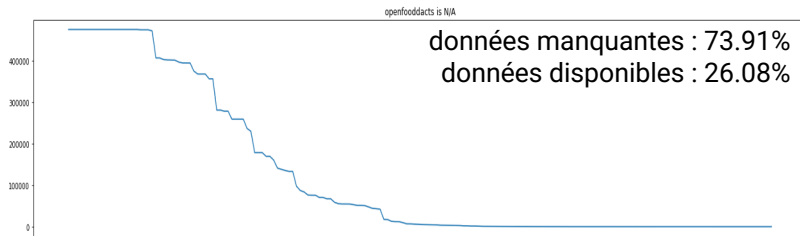


*source: <https://world.openfoodfacts.org/data/data-fields.txt>

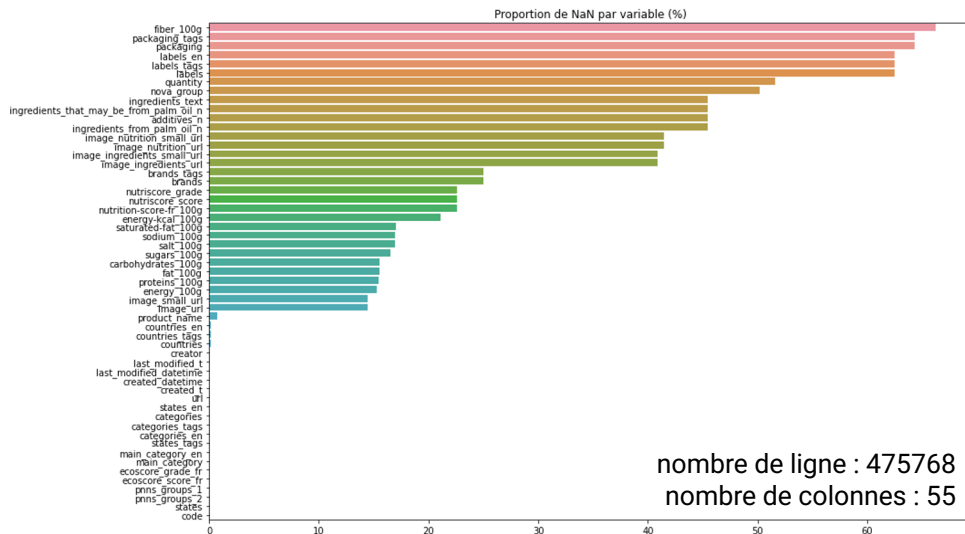


Nettoyage de la donnée

1 - Evaluation des donnée manquantes



2 - Suppression des colonnes vides (+70%)



3 - Evaluation des données restantes

- nombre de features catégoriques : 39
- nombre de features numériques : 16

4 - Remplir les données manquantes restantes

- features catégoriques -> 'No information available'
- features numériques -> moyenne par catégories (ou moyenne globale le cas échéant)

→ Création du Dataset de test



Présentation du Dataset de test

Dataset statistics

Number of variables	17
Number of observations	363956
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	47.2 MiB
Average record size in memory	136.0 B

Variable types

Numeric	16
Categorical	1

'main_categories reduced' -> preproc_shape = (327560, 3560)

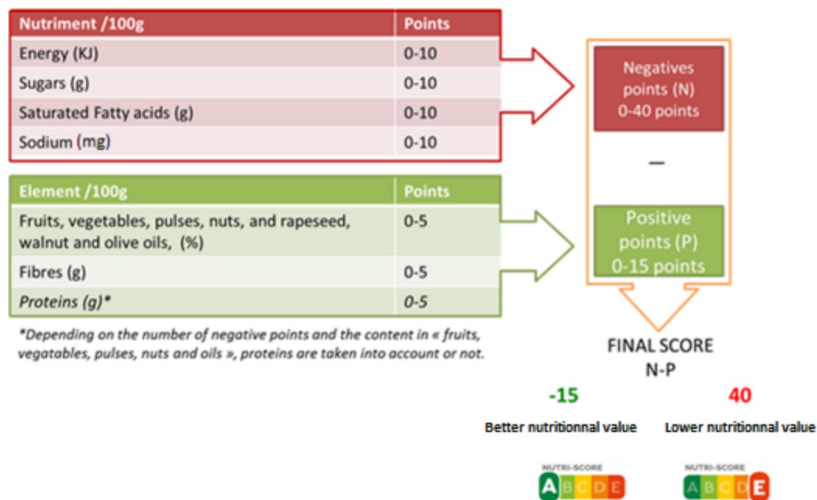
Et maintenant ?

- > Etude de la donnée/indicateurs
- > Présentation de l'application
- > Modélisation

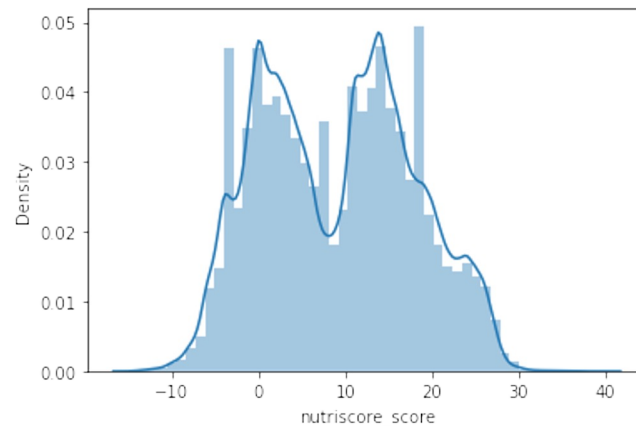
Qu'est ce que le nutri-score ?

Définition par Santé publique France :

“Le Nutri-Score est un logo qui indique la qualité nutritionnelle des aliments avec des notes allant de A à E. Avec le Nutri-Score, les produits peuvent être facilement et rapidement comparés.”



Dans notre dataset :



Qu'est ce que l'éco-score ?

Définition par Santé publique France :

“L'Eco-Score est conçu sur un modèle similaire au Nutri-Score : c'est une note de A à E qui synthétise 15 impacts environnementaux. La note Eco-Score est matérialisée par un logo de couleur en forme de feuille avec une lettre de A (très faible impact) à E (impact très important).”



Détails du calcul de l'Eco-score »

Score de référence de la catégorie du produit

🔗 Analyse de cycle de vie (ACV)

Catégorie Agribalyse : **Boisson à base d'avoine, nature**

Score environnemental PEF : 0.05 (plus le score est bas, plus l'impact est faible)
- dont impact sur le changement climatique : 0.36 kg CO2 eq/kg de produit

Détail des impacts par étapes du cycle de vie

Etape	Impact
🌾 Agriculture	9.3%
🏭 Transformation	2.1%
📦 Emballage	32.4%
🚚 Transport	31.8%
🛒 Distribution	19.1%
👤 Consommation	5.3%

Score ACV sur 100 : 98

Bonus et malus complémentaires

🌱 Mode de production

Label Bio européen : +15

🌍 Origine des ingrédients

France: 100%

Politique environnementale : +4

Transport : +15

🐾 Espèces menacées

Aucun ingrédient dont la culture menace des espèces n'a été détecté.

📦 Emballage

Brique (ratio : 1) - Carton (score : 91.25)

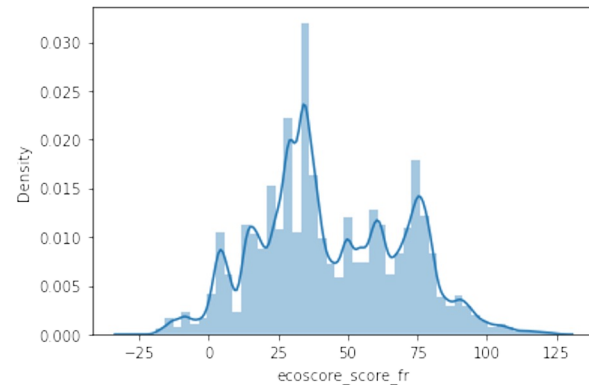
Score de tous les composants : 91.25

Emballage : -0.875

Score final

Score Éco-score : 131 - Note Éco-score : A

Dans notre dataset :



Présentation de mon application







1 - Informations nutritionnelles :

Note nutritionnelle de couleur NutriScore 

NUTRI-SCORE



Repères nutritionnels pour 100 g 

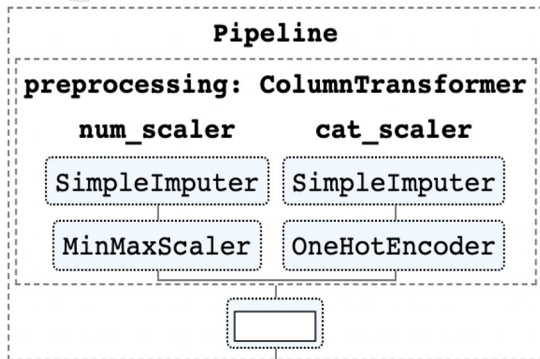
-  30.9 g **Matières grasses / Lipides** en quantité élevée
-  10.6 g **Acides gras saturés** en quantité élevée
-  56.3 g **Sucres** en quantité élevée
-  0.107 g **Sel** en faible quantité

2 - Recommandations :

- > Ce produit est **moins nutritif** que la moyenne des produits de cette catégorie
- > Ce produit est **moins écologique** que la moyenne des produits de cette catégorie
- > Cherchez un produit avec moins de **matières grasses**
- > Exemple de produit dans ce magasin :
 - pâte à tartiner Michel et Augustin
 - nutri-score C, eco-score B

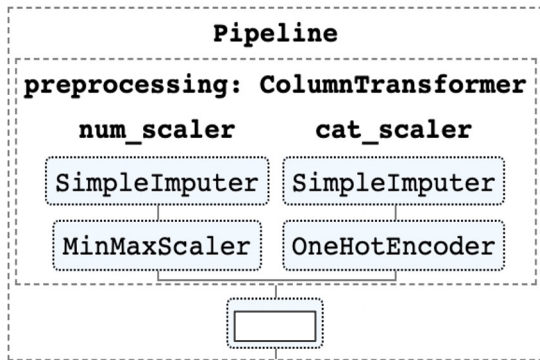
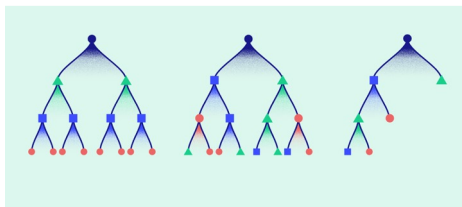
que faire si l'éco-score est absent sur le produit ? (32% du dataset)

Types de modèles de prédiction possibles ?



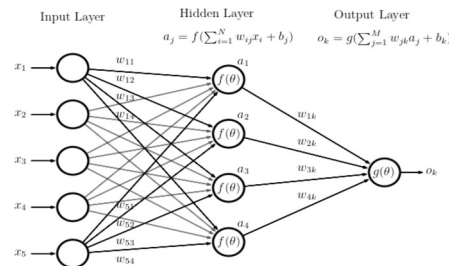
arbres de décisions boostés

XGBoost

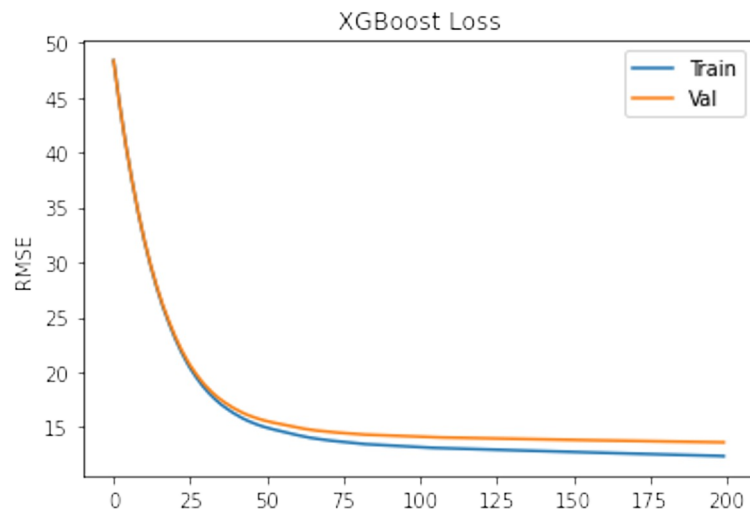
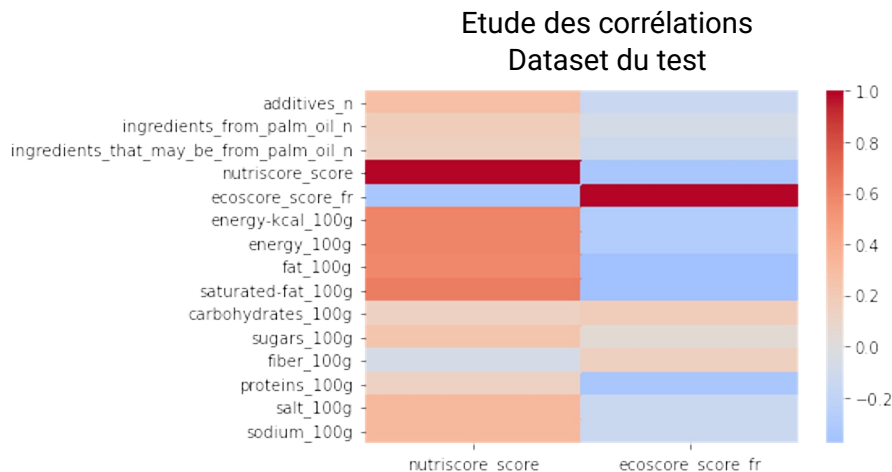


réseaux de neurones

TensorFlow



Est-ce que le machine learning est performant pour prédire l'éco-score à partir de l'information disponible au consommateur ?



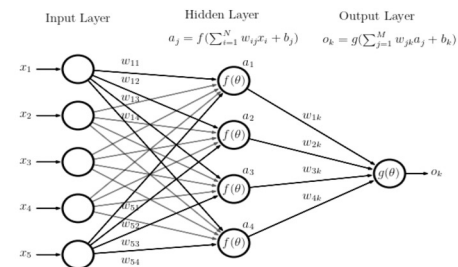
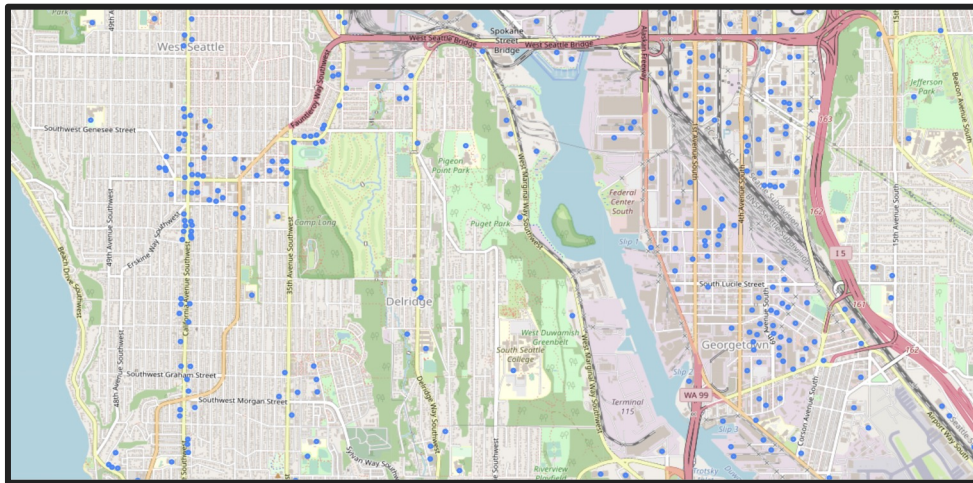
+ "main_categories" (reduced)

-> Forte corrélation :
passage du BVS de 17 à 11.50
et du SB de 34 à 16

RMSE
Best Validation Score = 11.50
Score Baseline = 16.61

Pour aller plus loin

Géolocalisation:



- permettrait d'affiner le calcul de l'éco-score (distance de transport)
- permettrait d'affiner les recommandations

+ Prix (pour 100g)

Data & Analytics Project
Eric Blanvillain

