Classification de produits e-commerce Rakuten

Table des matières

[Description du projet : 1](#_Toc173927980)

[Ressources à consulter : 1](#_Toc173927981)

[Conditions de validation du projet : 1](#_Toc173927982)

[Etapes projet 1](#_Toc173927983)

Cursus concerné : Data Scientist

Niveau de difficulté : 08/10

# Description du projet :

*Cataloguer les produits selon des données différentes (textes et images) est important pour les e-commerces puisque cela permet de réaliser des applications diverses telles que la recommandation de produits et la recherche personnalisée. Il s’agit alors de prédire le code type des produits sachant des données textuelles (désignation et description des produits) ainsi que des données images (image du produit).*

# Ressources à consulter :

* **Données :**
  + Ce projet s’inscrit dans le challenge **Rakuten France Multimodal Product Data Classification**, les données et leur description sont disponibles à l’adresse : <https://challengedata.ens.fr/challenges/35>
  + Les données textuelles : **~60 mb**
  + Données images : **~2.2 gb**
  + **99k** données avec plus de **1000 classes**.

# Conditions de validation du projet :

* un rapport d’exploration, de data visualisation et de pre-processing des données ;
* un rapport de modélisation ;
* un rapport final et le GitHub associé.

# Etapes projet

**Étape 0 : Cadrage (Notre première réunion)** :**Semaine du 05/08**

**Étape 1 :** **Exploration des données + DataViz’ : Deadline 16/08**

* Votre première tâche consistera à définir le contexte et le périmètre du projet : j’attends que vous preniez vraiment le temps de bien**comprendre le projet** et de vous renseigner au mieux sur les notions que celui-ci va introduire.
* Il faudra ensuite prendre en main et découvrir votre jeu de données et faire une **analyse** presque exhaustive de celui-ci afin de mettre en lumière la **structure**, les **difficultés**et eventuels **biais**du dataset.

Vous pourrez utiliser ce **template**: [Template - Rapport exploration des données](https://docs.google.com/spreadsheets/d/1T7uEuAvwJZ9lVuPD0TBc1paAcSRIy_P-WLz4nCAjqUQ/edit?usp=sharing)  
J’attendrai également au moins **5 représentations graphiques** construites à partir de votre jeu de données, visuelles et surtout **pertinentes.**Pour chacune d’elle j’attendrai :

* Un commentaire **précis**, qui analyse la figure et apporte un avis “métier”.
* Une validation du constat par des manipulations de données, ou un test statistique.

**Etape 2/ Pre-processing et feature engineering : Deadline 30/08**

* Suite aux constats établis lors de l'étape précédente, vous devrez **nettoyer**votre jeu de données et effectuer les étapes de **pre-processing**.
* Si besoin, il faudra **transformer**et/ou ajouter des features et **enrichir**votre jeu de données.
* L’objectif à la fin de cette étape est d’avoir un **dataset prêt** pour effectuer des analyses approfondies et/ou une modélisation à l’aide de Machine (et Deep) Learning.

**Rendu 1 : rapport d’exploration, de data visualisation et de pre-processing des données.**  
Après le rendu 1, si ce n’est pas déjà fait, votre mentor instanciera un**repo GitHub pour votre groupe**, celui-ci devra **respecter le** [**template**](https://github.com/DataScientest-Studio/Template_DS) **([DataScientest-Studio/Template\_DS (github.com)](https://github.com/DataScientest-Studio/Template_DS)) fourni et être transmis au jury avec le rapport final**.

**,**

**Étape 3 :** **Modélisation : Deadline 20/09**

1. **Step 1 :** **Deadline 06/09**
2. Élaboration des**premières itérations de modélisation**avec des modèles baseline**.**
3. **Analysez**vos premiers résultats, votre approche est-elle pertinente ? Et votre modèle ?
4. **Step 2** : **Deadline 13/09**
5. Utilisez les **métriques**de Machine Learning et de statistiques afin d’exploiter et **comprendre**au mieux les résultats de votre modèle.
6. **Optimisation :**Essayer d’améliorer votre modèle à l’aide d’outils d’**optimisation**, vous pouvez essayer d’autres modèles plus **complexes**et comparer les résultats.
7. **Step 3** : **Deadline 20/09**
8. **Modélisation avancée**: Utilisation de techniques de bagging/boosting et de Deep Learning.
9. **Interprétation**: Utilisez les outils d’interprétabilité pour mieux comprendre les résultats de vos modèles.
10. **Conclusions scientifiques et métiers en fonction du succès ou de l’échec de la modélisation**.

**Rendu 2 : rapport de modélisation.**

**Étape 4/** **Rapport final + codes sur GitHub :** **Deadline 27/09**

* Vous rendrez le **rapport final** : constitué des deux premiers rendus, il concrétise votre vision et votre travail face à votre projet en intégrant une conclusion et une ouverture.
* Vous rendez également votre **code propre et commenté sur GitHub**, si le temps le permet vous pourrez peaufiner votre modèle et votre pipeline de modélisation durant cette étape.

**Étape 5 : Streamlit + Soutenances: 14/10 et 15/10**

* Le document [Soutenance\_Méthodologie](https://docs.google.com/document/d/1hZuFkTLz_KXQ_AL3ge8AABk5RemabxYm6AVQZqjKcJw/edit?usp=sharing" \t "_blank) résume l’organisation de la soutenance.
* La soutenance se déroule de la manière suivante :
* **20 minutes de présentation**
* **10 minutes de questions de la part des membres du jury**

Vous avez la possibilité de réaliser soit :

* une présentation **Powerpoint**+ Démo **Streamlit**
* toute la présentation avec votre application **Streamlit**

Il faudra que votre application soit :

* **esthétique**, en particulier qu’elle contienne plusieurs onglets.
* codée avec soin, **sans ré-entrainer votre modèle**.
* fonctionne **sans bugs**.

Une attention particulière sera donnée à la forme des rapports intermédiaires et du rapport final, ils devront contenir des illustrations, une mise en page, ne pas avoir de fautes d’orthographe etc…  
**Si les rendus ne sont pas à la hauteur et délivrés à temps, votre projet ne pourra pas être validé.**  
Vous l’aurez compris, le projet demande une **implication conséquente** et constitue pour cette raison un pilier de votre formation. **Un bon projet sera le meilleur moyen pour vous de montrer que cette formation a été concluante et que vous êtes opérationnels** ! Je vous propose de faire un premier Meeting sur Zoom  pour que l’on puisse valider ensemble l’étape de **cadrage (etape 0 )**.

**Jeudi 8 aout a 10h30 ?**(Les calls projets ne prennent generalement pas plus d'une heure)Comme vous avez pu le contaster , le call de l'etape 1 a lieu durant vos vacances de DataScientest. Si cela est contraigant pour vous, faites le moi savoir et on trouvera une solution. Réagissez avec un :+1: pour que je puisse savoir que vous avez bien lu ce message.  
Bonne journée et à bientôt,