PRÊT À DÉPENSER

Nom du projet : **Réalisez un dashboard et assurez une veille technique**

Présenté par : Nathan FARDIN

Plan

Le besoin

L’entreprise Place de Marché souhaiterait obtenir un un dashboard interactif pour simplifier l’usage du modèle prédictif.
Elle souhaiterait également acquérir des informations quant aux avancés technologiques en data science.

Les données

1050 produits ainsi que diverses informations sur ceux-ci (ID, catégorie, description etc.).

WCAG

Web content accessibility guidelines

Critère 1.1.1

- Fournir des descriptions pour les images et graphiques

Critère 1.4.1

- Fournir des informations autres que la couleur pour décider de l'importance d'une feature dans un graphique

Critère 1.4.3

- Offrir un contraste suffisant (4;5:1) pour assurer une lecture possible pour les personnes malvoyantes

Critère 1.4.4

- Le texte doit pouvoir s'agrandir jusqu'à 200% sans perte de lisibilité

Critère 2.4.2

- Chaque page doit avoir un titre descriptif

Dashboard

simulateur pret

Prediction

Importance

Comparaison

Mise a jour

Application de Prédiction de Capacité de Paiement

Utilisez le menu latéral pour naviguer entre les différentes pages de l'application.

Logo de l'entreprise

Explication des différentes pages

Page 1: permet d'obtenir les prédictions d'accord de prêt pour un identifiant client défini

Page 2:permet d'identifier les données ayant été les plus importantes pour la prédiction

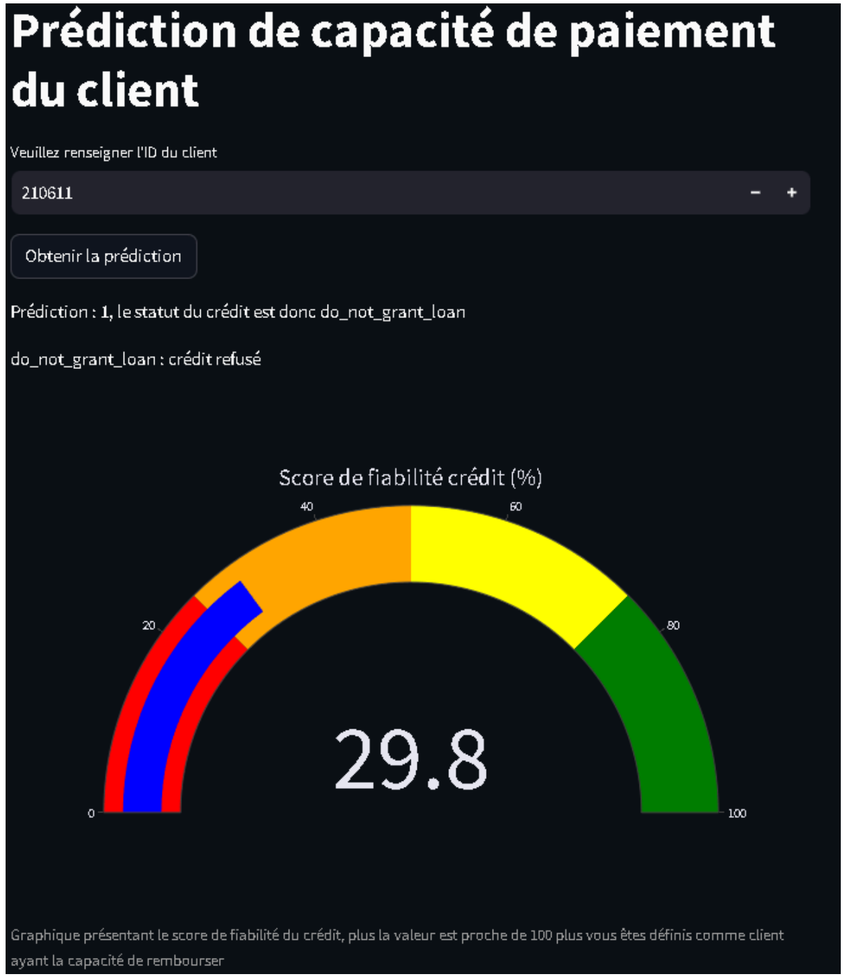
Page 3: permet de comparer les données du client avec celles des autres personnes présentes dans la base de données

Page 4: permet de modifier des données client puis de demander une nouvelle prédiction

Prédiction de capacité de paiement du client

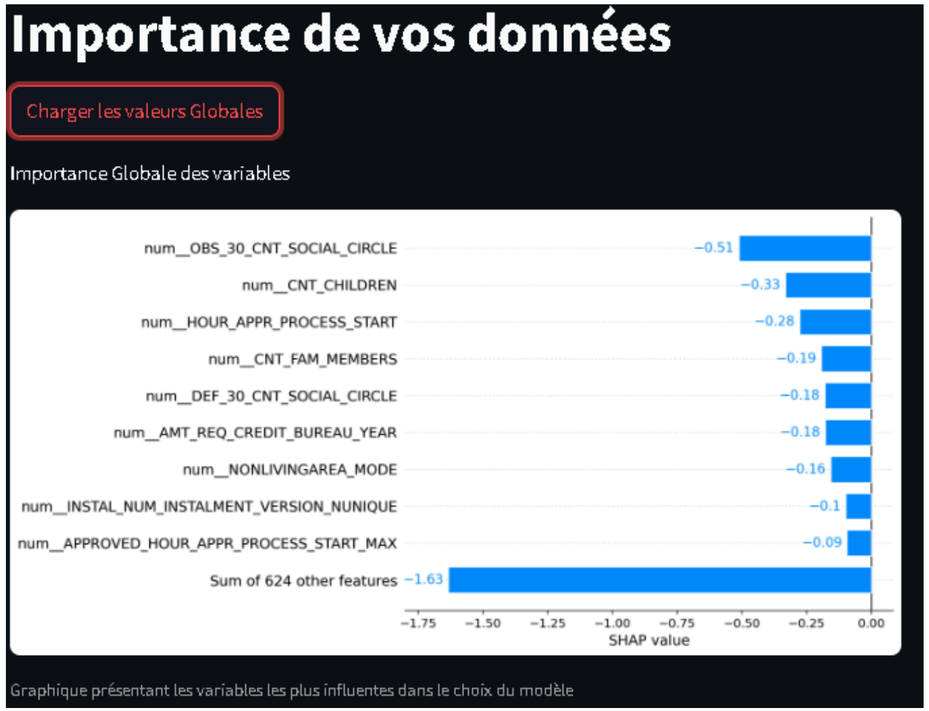
Première page

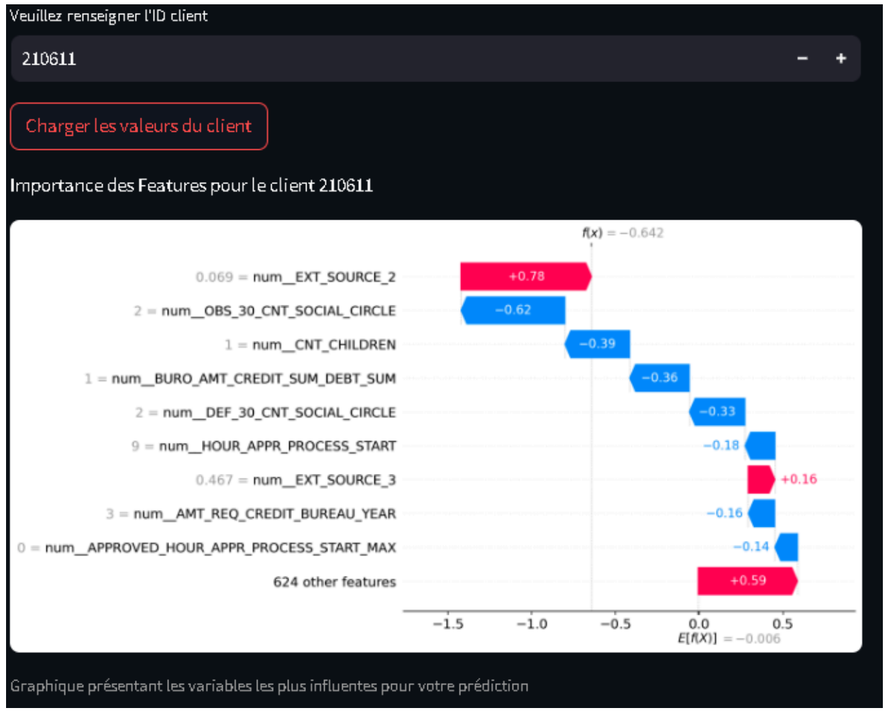
Permet d'obtenir les prédictions d'accord de prêt pour un identifiant client défini ainsi qu'un graphique permettant au client de voir ou il se situe.



Importance de vos données

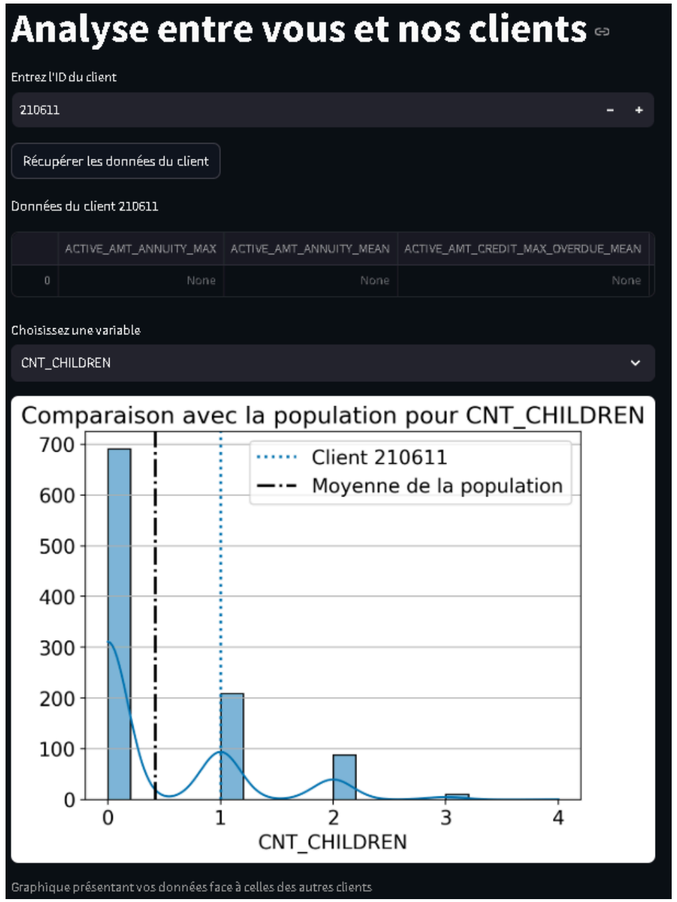
Seconde page

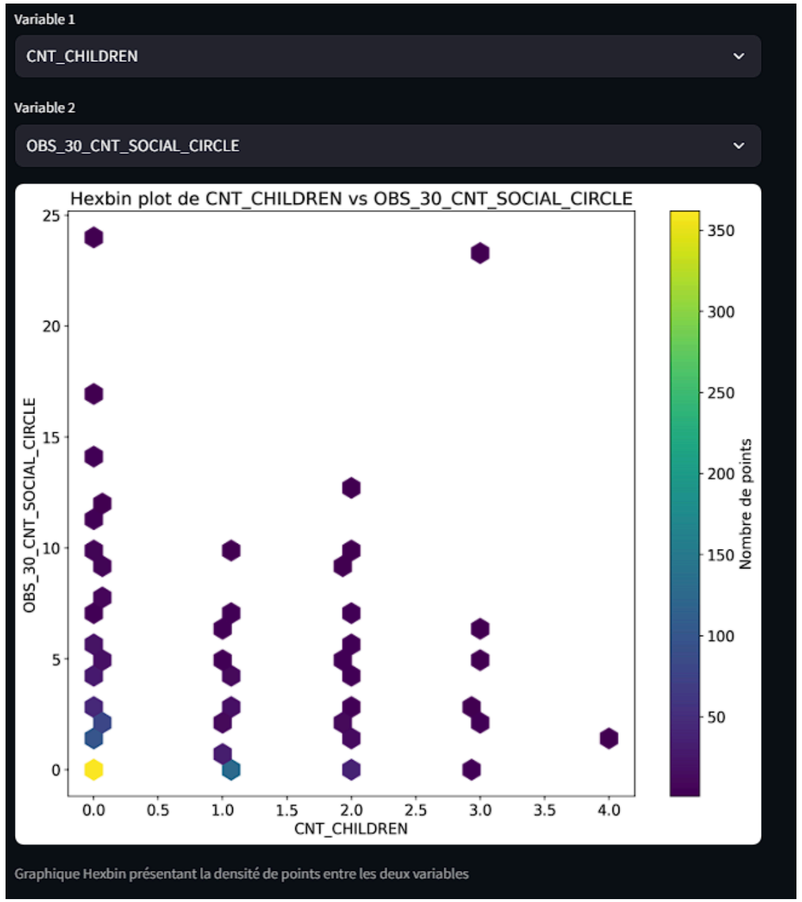




permet au client d'identifier les données ayant été les plus importantes pour la prédiction

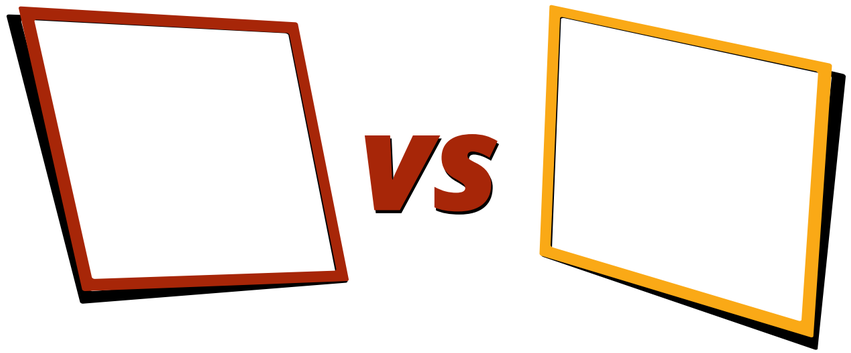
Analyse entre vous et nos clients



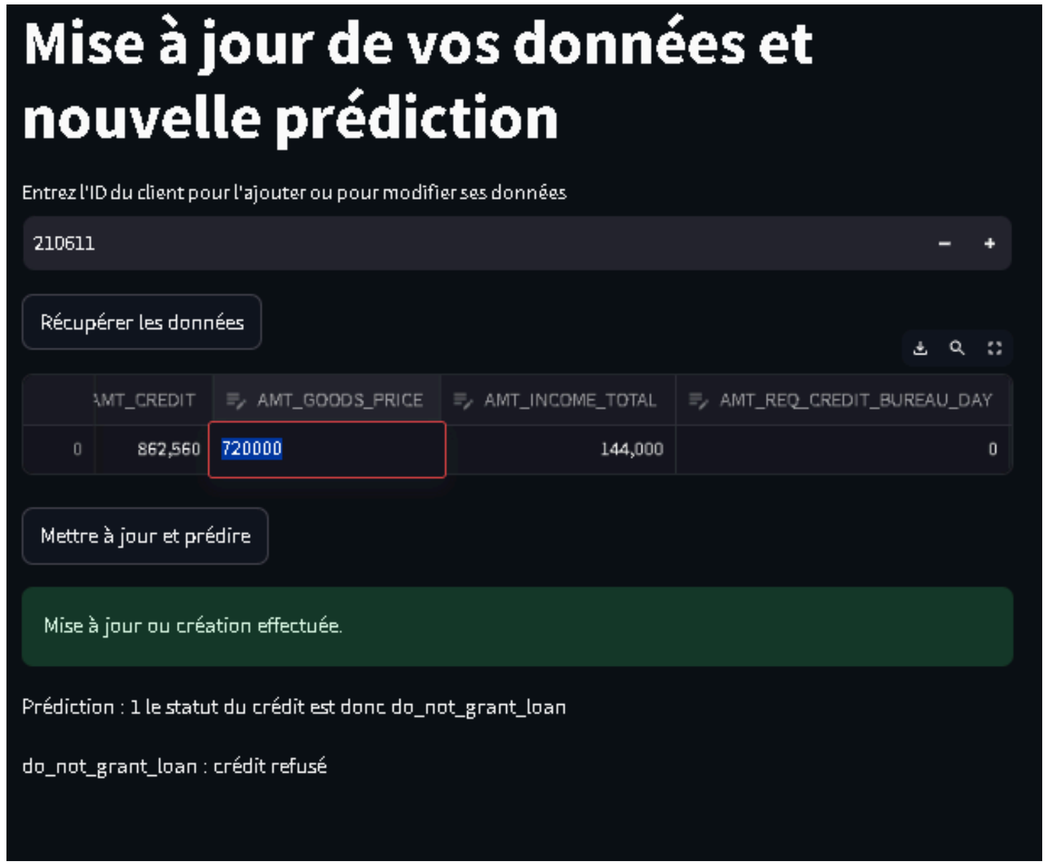


Troisième page

permet de comparer les données du client avec celles des autres personnes présentes dans la base de données ou alors des données entre elles



Mise à jour de vos données et nouvelle prédiction



Quatrième page

permet de modifier des données client puis de demander une nouvelle prédiction

Dataset utilisé et objectif de départ

- Le data set est composée de 1050 produits ainsi que des diverses informations leurs étants rattachés.

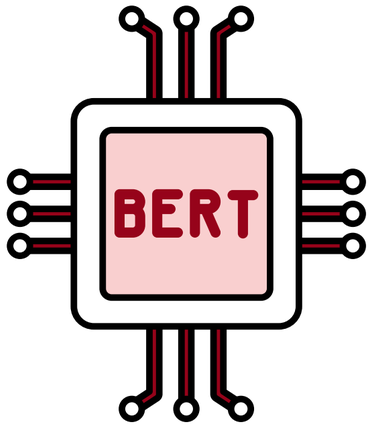
- L’objectif de base était de créer une classification automatique des différents produits

- Nous verrons les performances du modèle dans cette tâche

BERT

Bidirectional encoder representations from transformers

- Un modèle proposé en 2018 par Google basé sur l'architecture encoder-only-transformer (donc qui ne vas pas chercher à produire une séquence de sortie.)
- Le modèle à la capacité de prendre en compte le contexte des mots
- C'est le modèle qui nous servira pour la comparaison



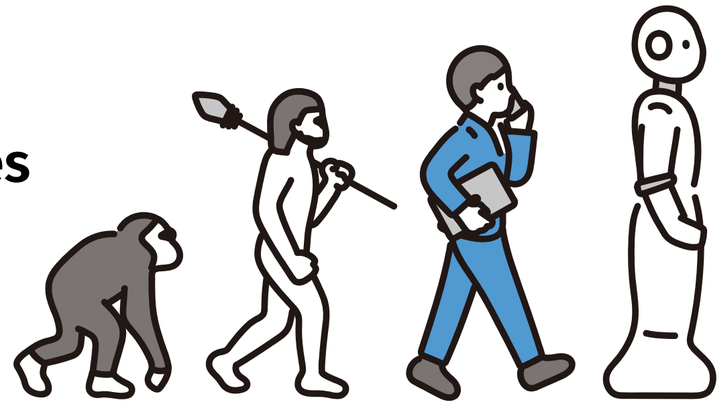
deBERTa

Decoding-enhanced BERT with disentangled attention

- **Modèle ayant pour but d'améliorer les performances de BERT et roBERTa**

- **La principale évolution est le « disentangled attention mechanism » qui représente le contenu et la position du mot dans deux vecteurs différents**

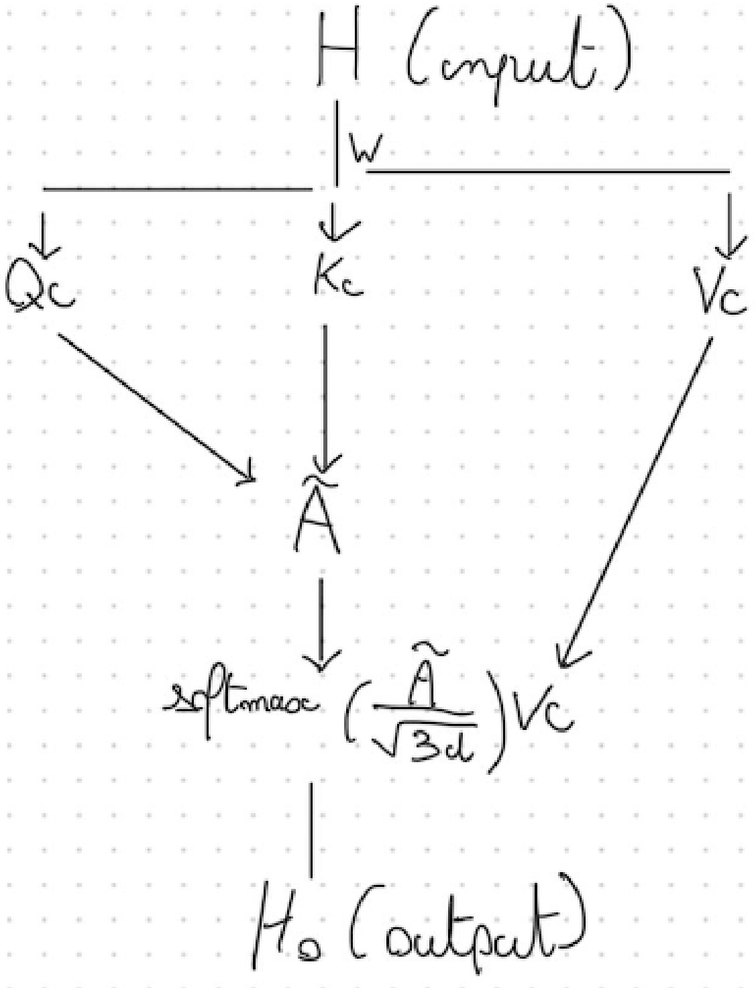
- **L'intérêt est de mieux comprendre les relations et le sens des phrases, sans que la position des mots ne prennent trop d'importance**



Fonctionnement

Attention en trois étapes :

- Content to content : mesure la similarité entre les contenus des tokens via Q et K.
- Content to position : mesure la relation spatiale entre Q et les autres tokens (K)
- Position to content : influence l'importance des tokens K selon leur position à Q et leur contenu.



$$Q_c = HW_{q,c}, K_c = HW_{k,c}, V_c = HW_{v,c}, Q_r = PW_{q,r}, K_r = PW_{k,r}$$

$$\tilde{A}_{i,j} = \underbrace{Q_i^c K_j^{cT}}_{\text{(a) content-to-content}} + \underbrace{Q_i^c K_{\delta(i,j)}^{r\,\mathsf{T}}}_{\text{(b) content-to-position}} + \underbrace{K_j^c Q_{\delta(j,i)}^{r\,\mathsf{T}}}_{\text{(c) position-to-content}}$$

$$H_o = \text{softmax}(\frac{\tilde{A}}{\sqrt{3d}})V_c$$

Modelisation

Choix des données

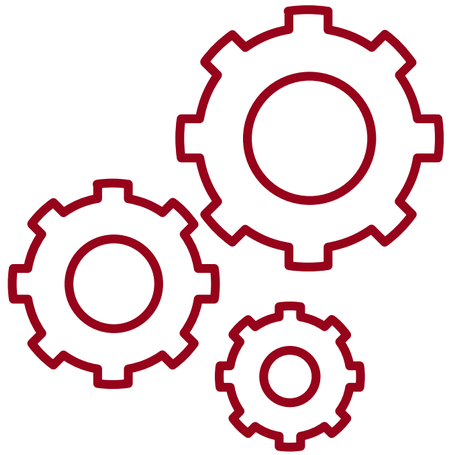
Vérifications de la distribution des catégories

Tokenization

optimisation de max_length

T-SNE et K-Means (ari)

Random forest (accuracy)



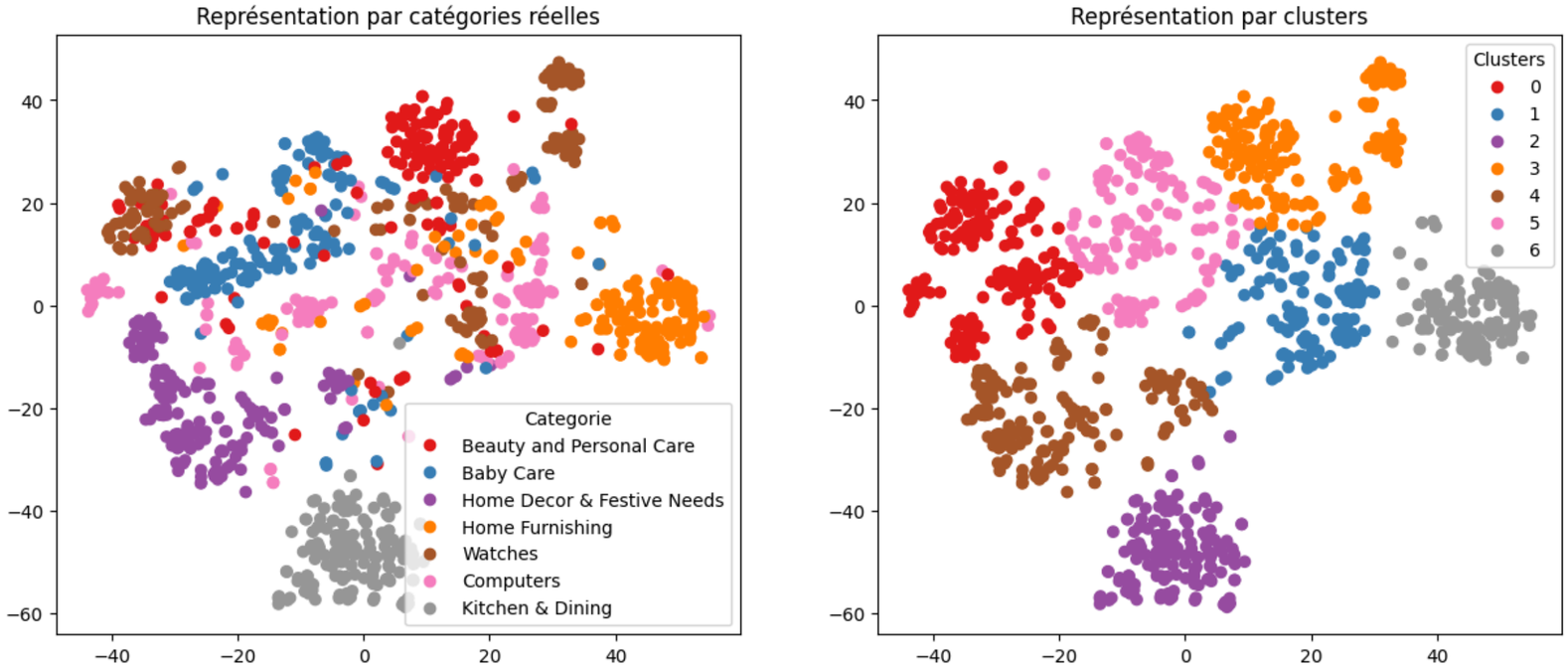
Resultats

BERT

ARI 0.44

Resultats

deBERTa



ARI 0.41

Random Forest

Catégorisation basée sur des arbres de décisions et une agrégation des résultats avec bootstrap (création de sous data set aléatoire).

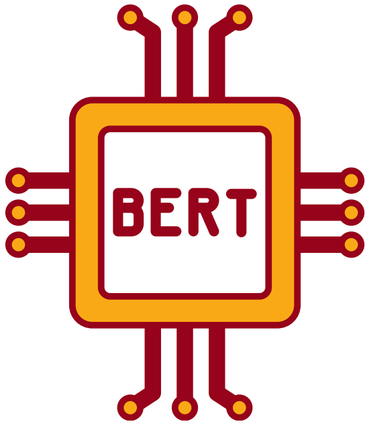
Paramètres

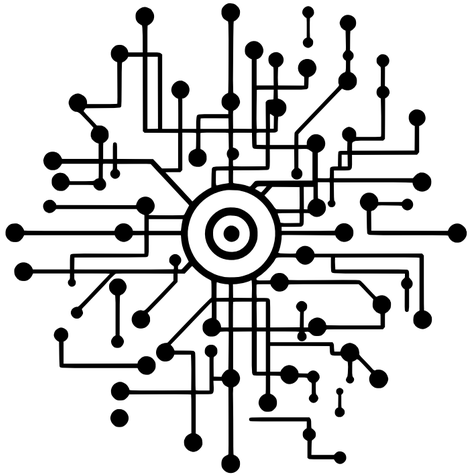
- n_estimators=30
- max_depth=3
- min_samples_split=5
- min_samples_leaf=3
- max_features="log2

Les même paramètres pour les deux modèles afin de les comparer sur une base identique.

Resultats

BERT
Train accuracy: 0.85
Test accuracy: 0.77



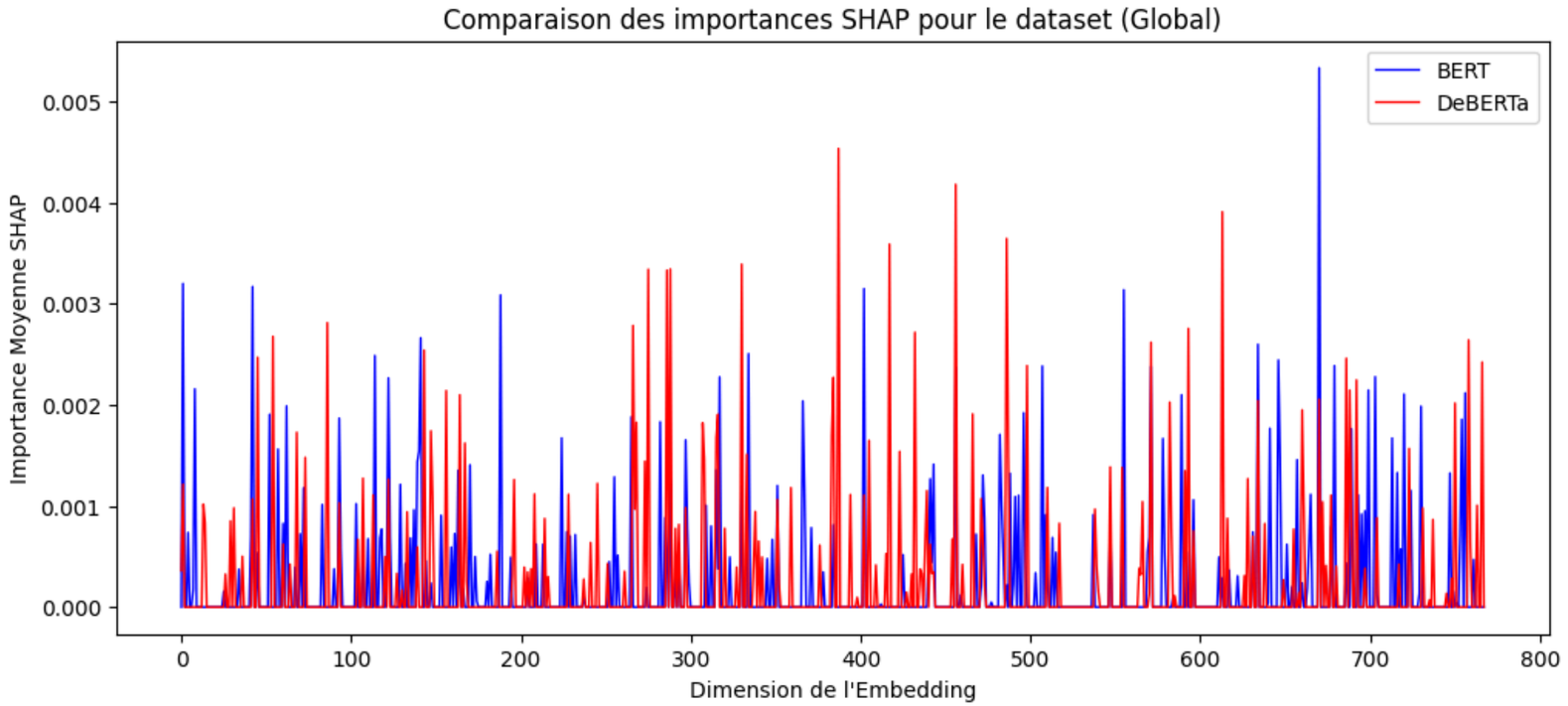


DeBERTa
Train accuracy: 0.83
Test accuracy: 0.80

deBERTa offre ici de meilleures performances ainsi que moins d'écart entre scores de train et de test

Features importance

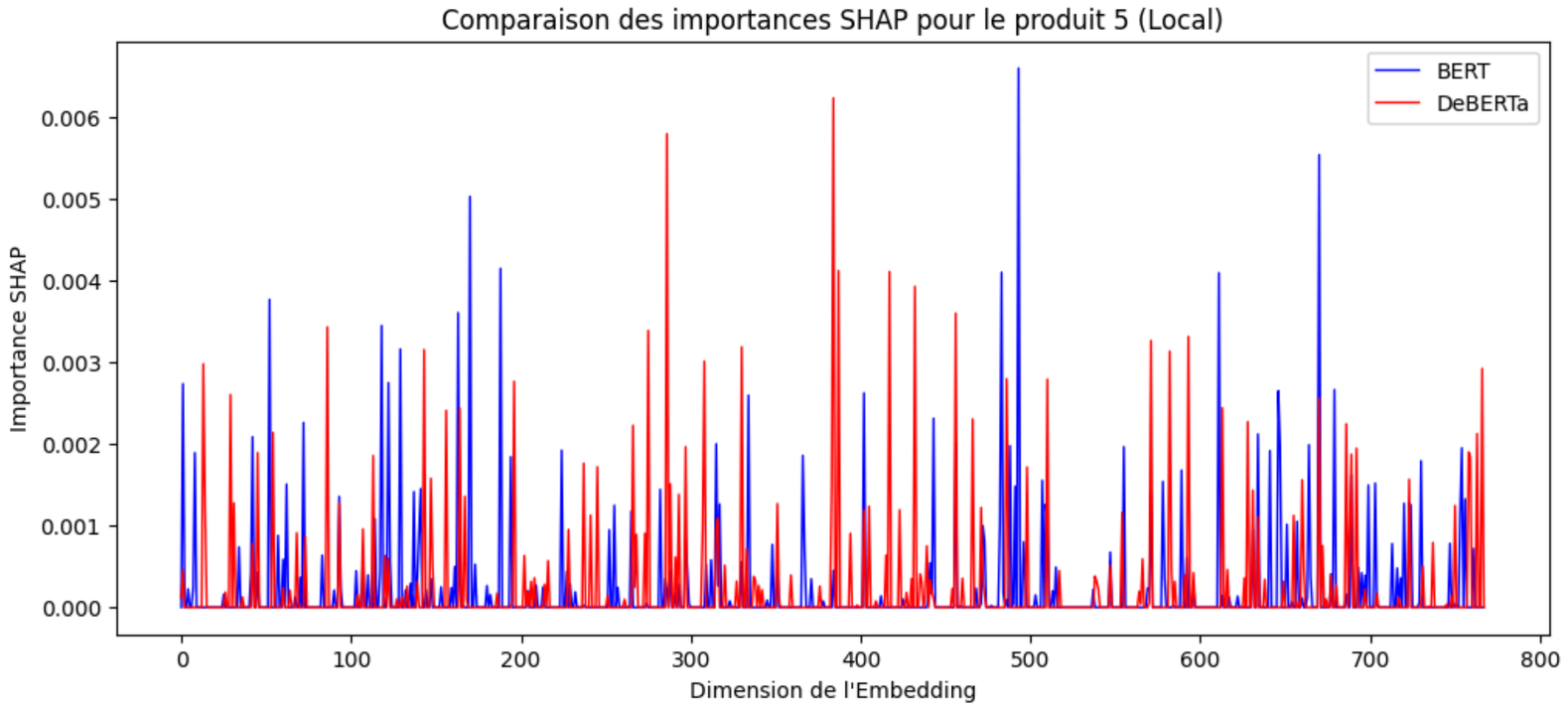
L’embedding ne nous donnant pas directement accès aux différents tokens nous devons changer notre manière d’analyser les résultats.



Les scores de feature importance sont répartis sur l’ensemble des dimensions. Aucune plage de dimension de l’embedding ne porte d’informations fortement discriminantes.

Features importance

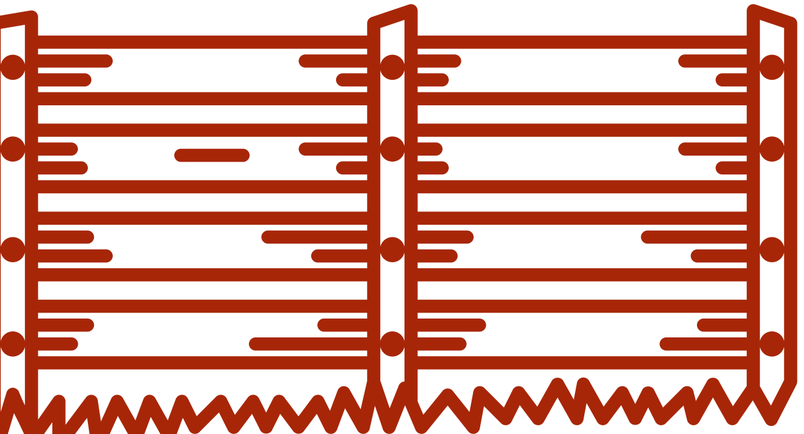
On distingue des pics d'importance suggérant que selon la catégorie certains embedding sont plus impactant.



Les pics d'importances ne sont pas localisé sur les mêmes dimensions mettant en évidence que les deux modèles n'ont pas encoder les données de façon identique.

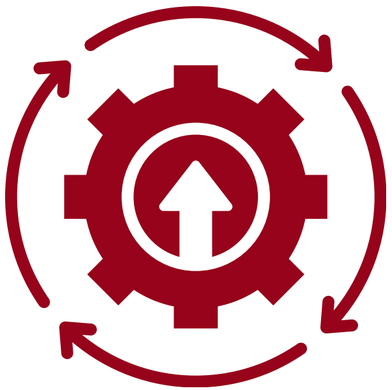
Limites

- **Taille limité du data set**
- **Type de donnée textuelle**
- **Nombre limité de version de deBERTa testé**



Améliorations

- Trouver un data set proposant un plus grand nombre de données
- Trouver d’autres types de contenu textuel à catégoriser plus adaptés au modèle
- Tester des versions différentes quant aux nombres de paramètres d’entrainement, aux mises à jour, par exemple v3-base ou large



Conclusion

Le Dashboard est crée et disponible à l’usage
La veille à été réaliser : les performance de deBERTa sont meilleur que BERT durant le test

Pour aller plus loin

Ajouter au dashboard des recommendations supplementaires du WCAG.
Approfondir le test avec d’autres version ou sur un dataframe plus adapté