Système de Recommandation

Apprentissage chez Junglebike

Komlan Jean-Marie DANTODJI

Université Paris 8, LIASD
Encadrante : Mme Rakia JAZIRI
Tutrice : Mme Alice Battarel

12 septembre 2022

JungleBike

- Start Up de E-Commerce de B2B et B2C.
- Spécialisée dans le secteur du Vélo.
- Mise en relation des clients avec les réparateurs.

Contexte RH

- Equipe Data
- Formé de trois data scientistes
- Une responsable de l'équipe data

Contexte technique

- Intégration de données
- Construction des algoritmes de catégorisation et d'extration de données
- Outils : DataIku, DBeaver
- Langages et librairies : Python, Scikit Learn, Keras, Tensorflow, PyTorch

Problème

- Recommandation des produits basées sur le vote et avis des clients.

Problematique des données

- L'information sur l'avis client.
- Biais de popularité : non diversité des produits recommandés
- Manque d'avis clients en comparaison au nombre d'article à recommander
- Données issues du scrapping des sites des fournisseurs.

Données

	product_id	product_name	brand	user	city	age	activity	level	vote	avis	description
0	0	Pneu Route Continental GP 5000 700 mm Tubetype...	CONTINENTAL	Br74	Annecy	45-54	VTT - XC	Eclairé	3	Très déçu par le poids réel	Connaissant très bien ce pneu car utilisé en 7...
1	0	Pneu Route Continental GP 5000 700 mm Tubetype...	CONTINENTAL	StM21	Dijon	45-54	Route - Cyclosportive	Eclairé	5	Très satisfait	Après plus de 15000 km parcourus avec ces pneu...
2	0	Pneu Route Continental GP 5000 700 mm Tubetype...	CONTINENTAL	boddishiva	barcares	45-54	Route - Cyclosportive	Amateur	2	deçu peut etre un default	j ai acheté ses pneu l an dernier j ai pas par...
3	0	Pneu Route Continental GP 5000 700 mm Tubetype...	CONTINENTAL	Conti2021	None	35-44	Route - Cyclosportive	Eclairé	4	Une fissure après 1200 km	J'avais fait bcp de bornes avec le GP5000 en 2...
4	0	Pneu Route Continental GP 5000 700 mm Tubetype...	CONTINENTAL	Thibj	Strasbourg, France	25-34	Route - Cyclosportive	Eclairé	1	Mauvaise usure	À peine une dizaine de sortie (courte) et l'on...
...	...	...	...	...	...	...	...	...	...	...	...
30907	9221	Chaussures de Trail Femme Asics Gel Sonoma 6 G...	ASICS	None	None	None	None	None	5	Au top !	[Cet avis a été recueilli en réponse à une off...
30908	9221	Chaussures de Trail Femme Asics Gel Sonoma 6 G...	ASICS	None	None	None	None	None	5	Souple et confortable	[Cet avis a été recueilli en réponse à une off...
30909	9221	Chaussures de Trail Femme Asics Gel Sonoma 6 G...	ASICS	None	None	None	None	None	4	Confort	[Cet avis a été recueilli en réponse à une off...
30910	9221	Chaussures de Trail Femme Asics Gel Sonoma 6 G...	ASICS	None	None	None	None	None	4	chaussure confortable	[Cet avis a été recueilli en réponse à une off...
30911	9221	Chaussures de Trail Femme Asics Gel Sonoma 6 G...	ASICS	None	None	None	None	None	5	Asics gel sonoma trade 6 G TX	[Cet avis a été recueilli en réponse à une off...

30912 rows × 11 columns

FIG. 1 : Jeu de données issue du Scrapping des sites des

Détail des colonnes


```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   product_id    30912 non-null  int64
 1   product_name  30909 non-null  object
 2   brand         29285 non-null  object
 3   user          26452 non-null  object
 4   city          23821 non-null  object
 5   age           26277 non-null  object
 6   activity      26277 non-null  object
 7   level         26277 non-null  object
 8   vote          27312 non-null  object
 9   avis          27312 non-null  object
 10  description   27312 non-null  object
dtypes: int64(1), object(10)
memory usage: 2.6+ MB
```

FIG. 2 : Détail des colonnes

Validation de modèle

MSE : Mean Squared Error :

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

RMSE : Root Mean Squared Error

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

MAE : Mean Absolute Error :

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Méthodes élémentaires

- Recommandation aléatoire

Recommandation Objet

- Caracteristiques des produits
- Projection des produits dans un repere

Methode du Cosin Similarity

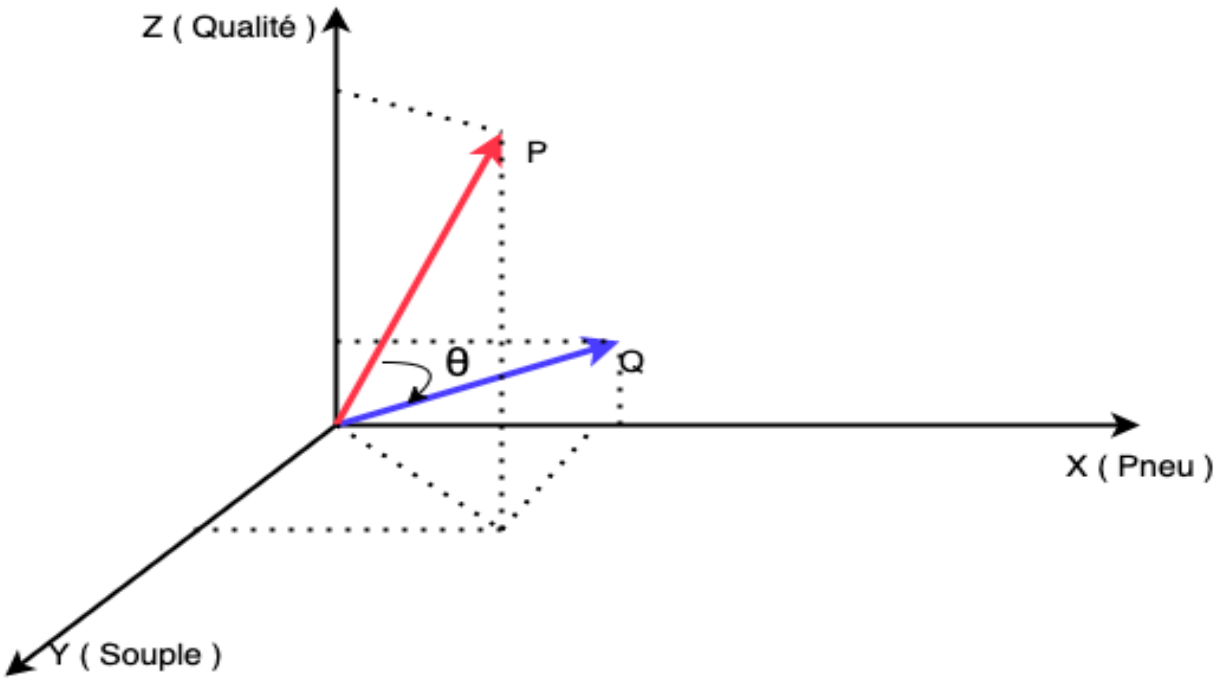

FIG. 3 : Projection des produits

La Matrice de Factorisation

- Décomposer la matrice de votes en deux

$$M = UxI$$

Décomposition

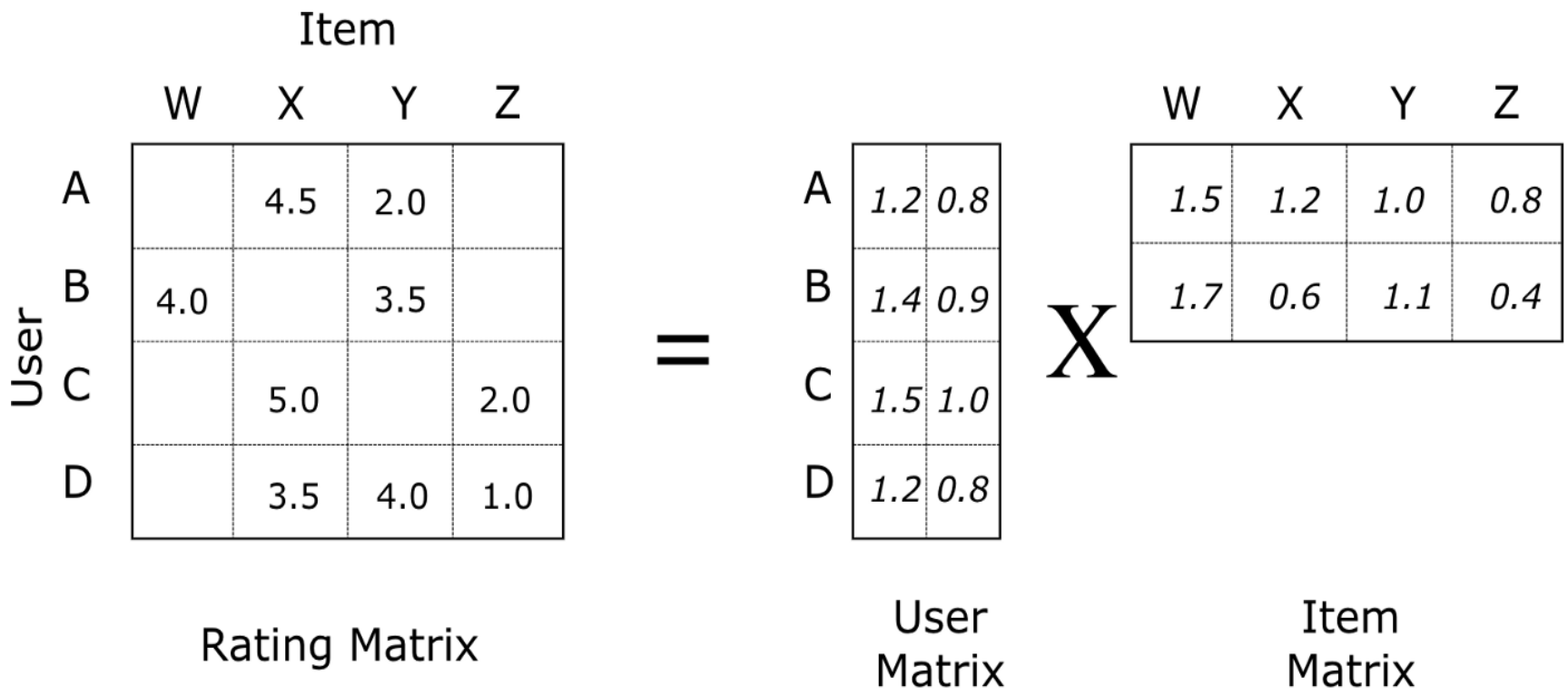

FIG. 4 : Décomposition de la matrice

Matrice similaire

3.16	1.92	2.08	1.28
3.63	2.22	2.39	1.48
3.95	2.4	2.6	1.6
3.16	1.92	2.08	1.28

FIG. 5 : Décomposition de la matrice

Neural Collaborative Filtering (NFC)

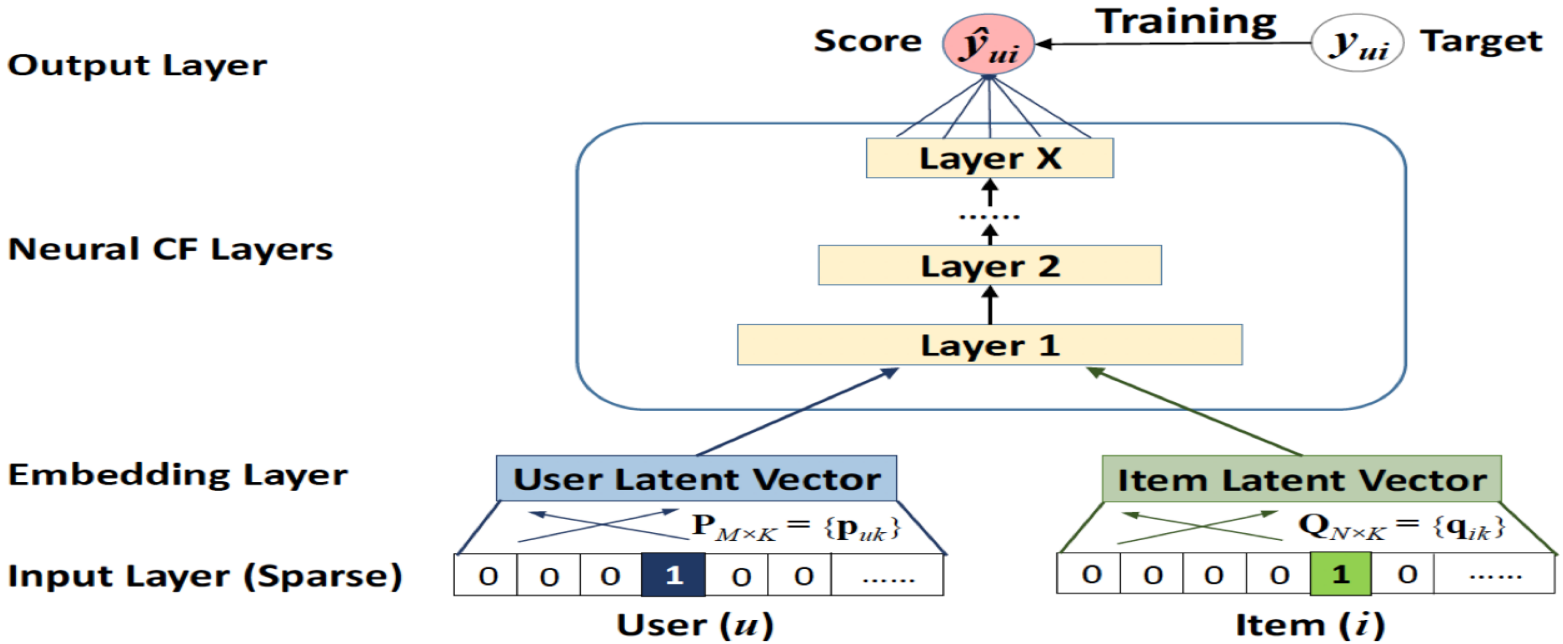

FIG. 6 : Neural Colaborative Filtering,
https ://towardsdatascience.com/neural-collaborative-filtering-96cef1009401

Généralisation du NFC

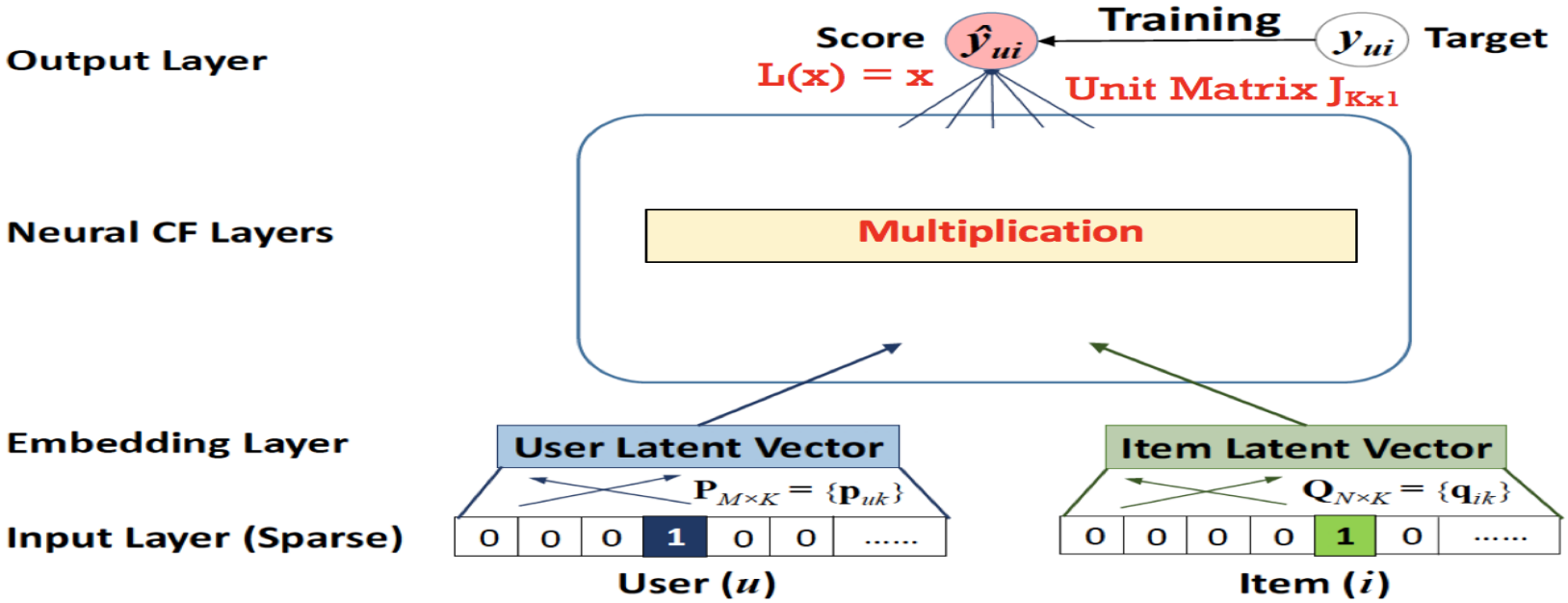

FIG. 7 : Généralisation du NFC,
https ://towardsdatascience.com/neural-collaborative-filtering-96cef1009401

Conditions de généralisation

- initialiser le poids de la couche de sortie à une matrice J dont toutes les valeurs sont égales à 1.
- Considérer une fonction d'activation L linéaire :

$$L(x) = x$$

Généralisation du NFC

$$\hat{y}_{ui} = L(p_u \odot q_i \times J_{Kx1})$$

$$\hat{y}_{ui} = L(p_u^T \cdot q_i)$$

$$\hat{y}_{ui} = p_u^T \cdot q_i$$

FIG. 8 : Généralisation du NFC

Combinaison des méthodes

- Combinaison du Collaborative Filtering au Content Based.

Combinaison des méthodes

- Combinaison du Collaborative Filtering au Content Based.

LSTM

LSTM : Long Short Term Memory

- Famille de Réseau de Neurone Récurrent :

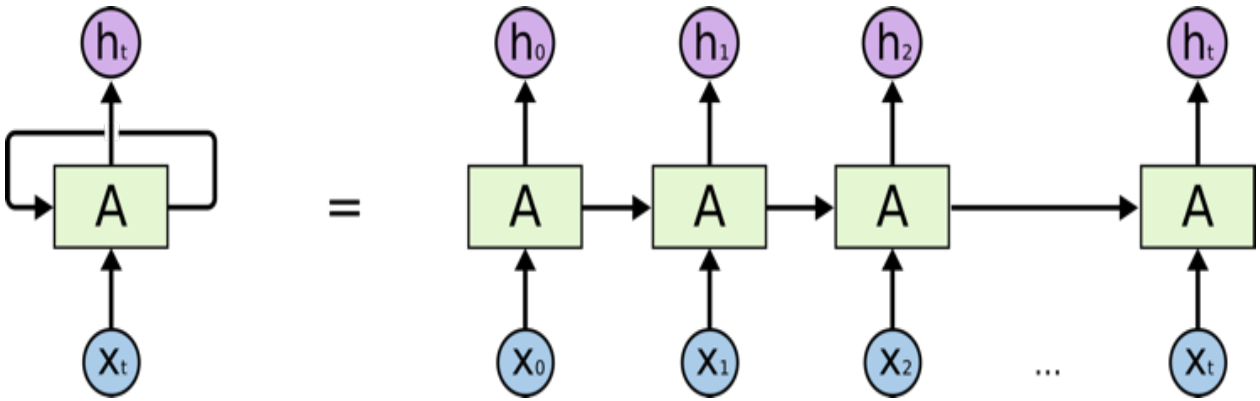

FIG. 9 : Etats du RNN : https ://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47

LSTM

- Cas particulier du LSTM :

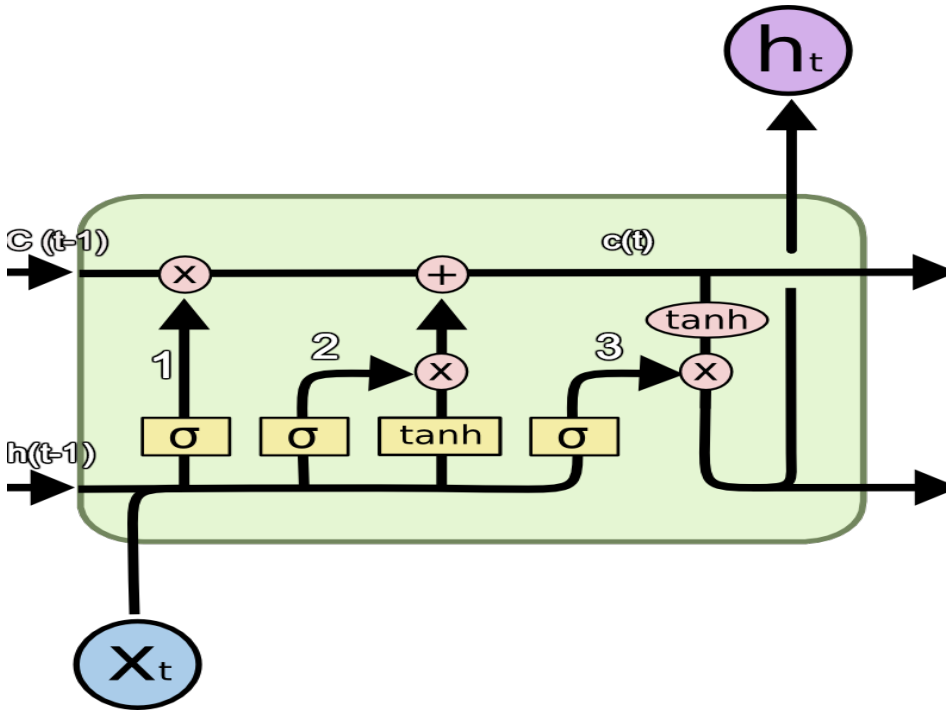

FIG. 10 : Couche du LSTM :
https ://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47

LSTM

- Les différentes portes du LSTM :
  - Input Gate (Couche de Sigmoide σ) :
    Contrôle de quel vecteur entre en mémoire $c(t)$
  - Forget Gate (Couche de Sigmoide σ) :
    Contrôle de quel information supprimer de la mémoire $c(t)$
  - Candidate Gate (Couche de Sigmoide $tan(h)$) :
    Détermine quelle information écrire dans la mémoire $c(t)$.
  - Output Gate (Couche de Sigmoide σ) :
    Détermine quelle information sort en sortie de la l'état caché.

Aperçu du jeu de donnée

	item	product_name	user_name	user	rating
0	2864	Pneu Route Continental GP 5000 700 mm Tubetype...	Br74	9600	3.0
1	2864	Pneu Route Continental GP 5000 700 mm Tubetype...	StM21	3666	5.0
2	2864	Pneu Route Continental GP 5000 700 mm Tubetype...	boddishiva	1098	2.0
3	2864	Pneu Route Continental GP 5000 700 mm Tubetype...	Conti2021	4601	4.0
4	2864	Pneu Route Continental GP 5000 700 mm Tubetype...	Thibj	11470	1.0
...	...	...	...	...	...
60110	2180	COUPE VENT VELO VTT FEMME BLEU MARINE	Gerard	4739	5.0
60111	2180	COUPE VENT VELO VTT FEMME BLEU MARINE	Caroline	6430	5.0
60112	2180	COUPE VENT VELO VTT FEMME BLEU MARINE	CLAUDINE	78	5.0
60113	2180	COUPE VENT VELO VTT FEMME BLEU MARINE	MICHELE	11439	4.0
60114	2180	COUPE VENT VELO VTT FEMME BLEU MARINE	ERIC	2853	5.0

35789 rows × 5 columns

Statistiques sur la dataset :

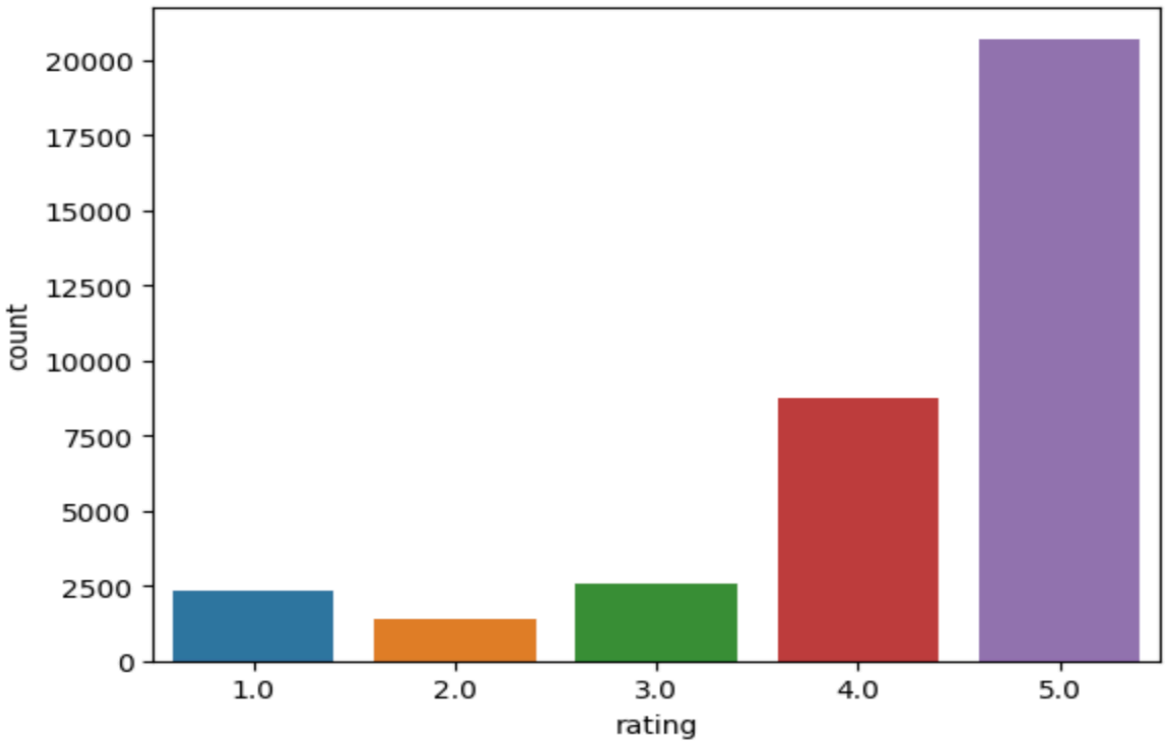

FIG. 12 : Nombre de produits votés en fonction du score

Statistiques sur la dataset :

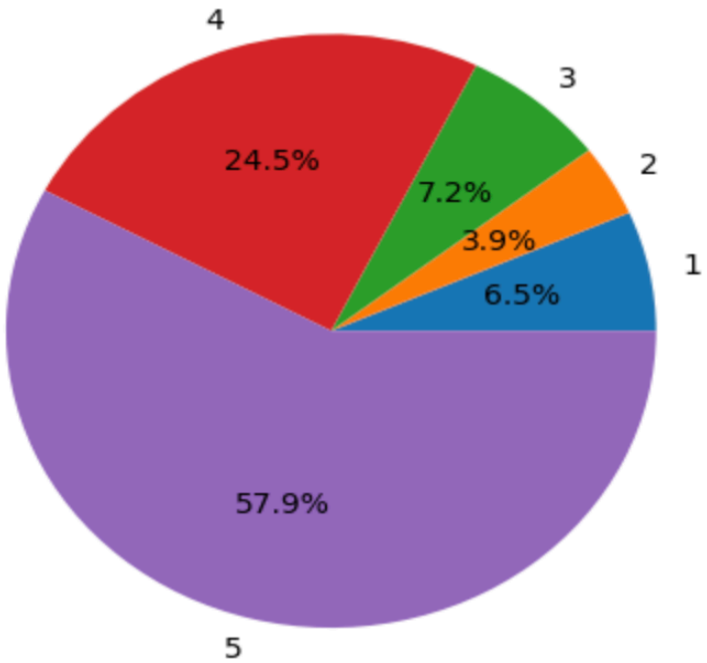

FIG. 13 : Proportion du nombre de produits votés en fonction du score

Différents modèles testés

- Matrice de Factorisation :
- Matrice de Factorisation et Réseau de Neurone :
- Matrice de Factorisation et Multilayer perceptron :
- LSTM : Long Short Term Memory :

Performances des modèles

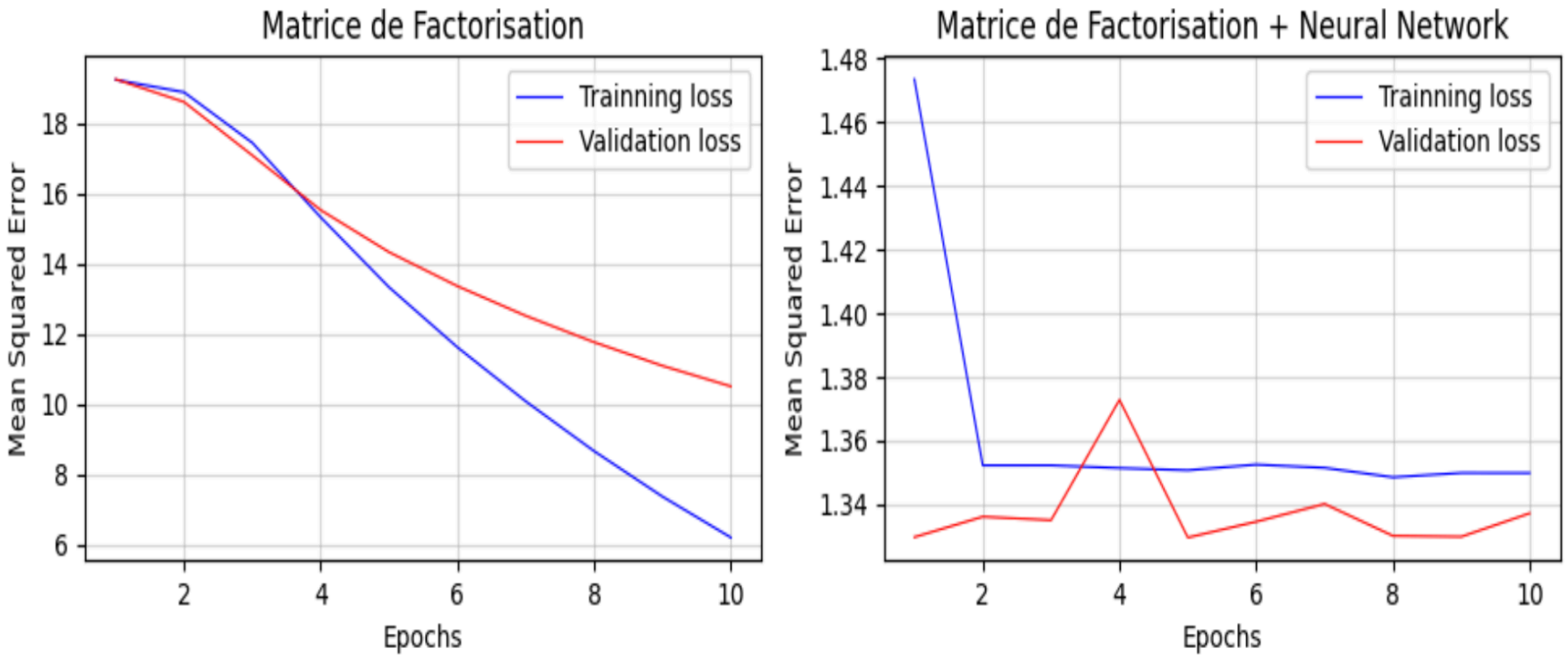

FIG. 14 : Etude de la validation des modèles

Performances des modèles

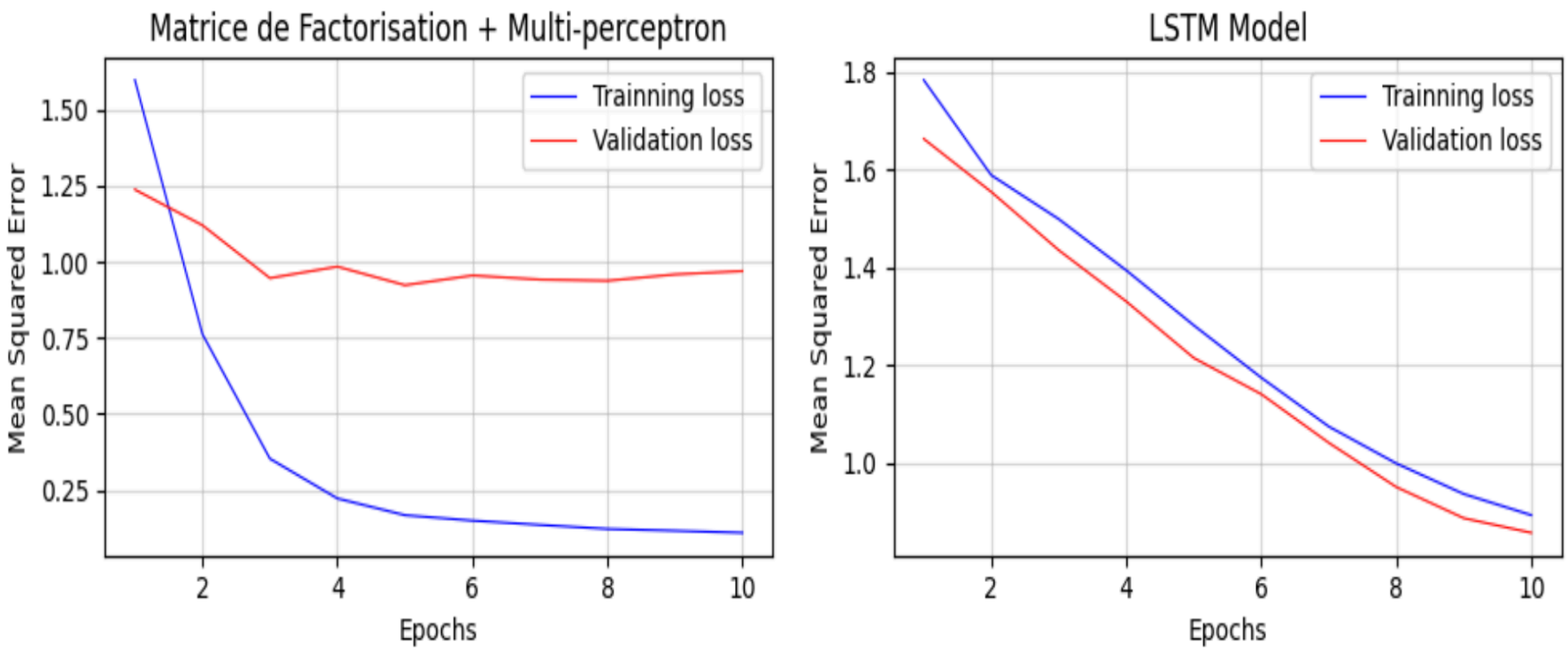

FIG. 15 : Etude de la validation des modèles

Performances des modèles

	MSE Training	MSE Testing	Epoques	Durée
Matrice de Factorisation	6.21	10.51	10	41 s
Matrice de Factorisation et Réseau de Neurone	1.34	1.33	10	61 s
Matrice de Factorisation et Multilayer perceptron	0.10	0.96	10	108 s
LSTM: Long Short Term Memory	0.89	0.86	10	106 s

FIG. 16 : Statistiques de performances des modèles

Conclusion

- La méthode de Deep Learning généralise au mieux la Matrice de Factorisation.
- Succès du modèle issu de la combinaison Matrice de Factorisation et Multilayer-Perceptron.
- Ajout de nouvelles données pour améliorer la performances des modèles.
- Combiner les votes aux avis dans la recommandation.

Références

- Sumit Sidana. Recommendation systems for online advertising. Computers and Society [cs.CY]. Université Grenoble Alpes, 2018. English. ffNNT : 2018GREAM061ff. fftel-02060436ff
- D Gunawan et al. The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. 2018 J. Phys. : Conf. Ser. 978 012120

Références

- Chakrabarti S, van den Berg M, Dom B 1999 Focused crawling : a new approach to topic-specific Web resource discovery Comput. Networks 31 11–16 pp 1623–1640

Merci pour votre attention