

# MÉMOIRE

pour obtenir le grade de Master délivré par

**Université Paris 8 Vincennes à Saint-Denis**

Mention *Informatique*

*Parcours MIASHS Big data et fouille de données*

*présenté et soutenu publiquement par*

**Komlan Jean-Marie DANTODJI**

le 12 septembre 2022

## La recommandation des articles

Encadrant universitaire : Rakia JAZIRI

Tuteur de stage : Alice BATTAREL

Stage effectué à : Jungle Bike  
Urban Lab | RIVP Rue René Clair, 75018 Paris

Université Paris 8  
Laboratoire d'Informatique Avancée de Saint Denis  
EA n° 4383 Saint Denis, France

M  
A  
S  
T  
E  
R



# Sommaire

Remerciements	5
Introduction Générale	7
1 Présentation de l'entreprise	9
I Problématique	19
2 Le contexte de résolution du problème	23
II État de l'art	29
3 État de l'art des techniques de recommandation	33
III Système réalisé	43
4 Implémentation du système	47
IV Conclusion	59
Conclusion	63



# Remerciements

Pendant mon parcours de stage, j'ai reçu le soutien moral et technique venant de plusieurs personnes. Ce qui m'a permis d'atteindre les objectifs qui m'ont été assignés.

*J'aimerais remercier en premier lieu, mes très chers parents, qui se sont surpassés tout au long de leur vie pour nous offrir, une éducation exemplaire, un modèle de labeur et de persévérance.*

*Je tiens à exprimer également mon remerciement à ma tutrice Madame Alice BATTAREL, qui m'a accordé sa confiance afin de travailler sur les sujets d'intégration de données et de Deep Learning. Elle m'a assisté tout le temps avec ses pistes qu'elle me donne afin d'avancer.*

*Par ailleurs, j'adresse mes remerciements à mon encadrant Madame Rakia JAZIRI, Maître de Conférences à l'Université Paris 8 pour le temps qu'elle m'a accordé dans le suivi de mon apprentissage.*

*Enfin, je remercie également toute l'équipe de Jungle Bike, et spécialement l'Equipe Data pour leur accueil, leur esprit d'équipe, ainsi que les conditions favorables dans lesquelles j'ai évolué au quotidien.*



# Introduction Génarale

Les systèmes de recommandation aujourd’hui deviennent populaires dans la connaissance des intérêts et préférences des clients qui visitent les sites e-commerce. L’objectif de cette technologie est de comprendre au mieux le client dans ses envies afin de lui proposer encore plus des produits pouvant potentiellement l’intéresser. Les géants de la recommandation aujourd’hui sont Google, Amazon, Netflix,...

L’objectif de ce mémoire est donc de proposer un système de recommandation des d’articles de vélo en ligne sur [junglebike.fr](http://junglebike.fr). Nous découvrirons ensemble différentes méthodes de recommandation basée sur le deep learning et tester leurs performances.



# Chapitre 1

## Présentation de l'entreprise

### Sommaire

---

<b>1.1</b>	<b>L'histoire de JungleBike . . . . .</b>	<b>10</b>
<b>1.2</b>	<b>Chiffres de JungleBike . . . . .</b>	<b>11</b>
<b>1.3</b>	<b>Les Services proposés . . . . .</b>	<b>12</b>
1.3.1	Accès des produits en ligne : . . . . .	12
1.3.2	Enregistrement du vélo . . . . .	12
<b>1.4</b>	<b>Processus d'intégration de donnée et de mise en ligne des produits . . . . .</b>	<b>14</b>
1.4.1	Récupérer le catalogue des produits . . . . .	14
1.4.2	Nettoyage et enrichissement et standardisation : . . . . .	14
1.4.3	Mise en base : . . . . .	15
1.4.4	Mise en ligne . . . . .	16
<b>1.5</b>	<b>JungleBike en interne . . . . .</b>	<b>17</b>
1.5.1	L'équipe d'accueil : . . . . .	17
1.5.2	Organisation de travail : le scrum Agile : . . . . .	17
1.5.3	Outils techniques : . . . . .	17

---

## 1.1 L'histoire de JungleBike

Jungle Bike est une start up spécialisée dans le secteur du Vélo E-commerce dans la vente de matériels de vélos en ligne et aussi proposer à ses clients une facilitation de la réparation de leur vélos.

Elle est fondée en 2020 par Alice Battarel après ses trois années passées auprès de IBM en tant que consultante senior en Innovation et Analytics avancée. Avec un capital de début de 500 €, Jungle Bike lance ses activités officiellement et fixe son siège au 15 Rue des Halles 75001, Paris. En 2020, Jungle Bike lance la mise en place de la plateforme d'aide à la réparation et de personnalisation du vélo. Cette plateforme répertoriera tous les réparateurs disponibles que le client choisira en cas de problème de vélo. Aujourd'hui Jungle Bike dispose de plus d'une vingtaine de partenaires fournisseur de pièces de vélo dont P2R, Comet.

## 1.2 Chiffres de JungleBike



FIG. 1.1: Chiffres actuels sur l'évolution de JungleBike

<b>JUNGLEBIKE</b>		
Société : 892 455 734 <span style="color: green;">Active</span>		
<b>Renseignements juridiques</b>		
Date création entreprise	28-12-2020 <i>il y a 1 an</i>	<a href="#">Statuts constitutifs &gt;</a>
Forme juridique	SASU Société par actions simplifiée à associé unique	<a href="#">Voir l'offre PLUS &gt;</a>
Noms commerciaux	JUNGLEBIKE	
Téléphone	<a href="#">Afficher le numéro</a>	
Adresse postale	<a href="#">15 RUE DES HALLES 75001 PARIS</a>	
<b>Numéros d'identification</b>		
Numéro SIREN	892455734 <a href="#">🔗</a>	
Numéro SIRET (siège)	89245573400012 <a href="#">🔗</a>	
Numéro TVA Intracommunautaire	FR83892455734 <a href="#">🔗</a>	
Numéro RCS	Paris B 892 455 734	

FIG. 1.2: Informations juridique de JungleBike

## 1.3 Les Services proposés

### 1.3.1 Accès des produits en ligne :

Comme dit plus haut Jungle Bike reçoit des produits des fournisseurs en France ou à l'extérieur et les met à disposition des clients. Sur [junglebike.fr](http://junglebike.fr) les clients peuvent rechercher les produits par marque, origine et dimensions du produit. Pour toute commande effectuée, l'entrepôt basé au Mans (France) déclenche une livraison au client dans les 3 à 7 jours. Si le produit n'est pas conforme à l'attente du client, cette dernière peut procéder au retour de l'article dans le but d'être remboursé.: Dans le but d'avoir le stock synchronisé à celui du fournisseur, Jungle Bike dispose des algorithmes de mise à jour du stock. De nouveaux produits y sont régulièrement ajoutés dès qu'ils apparaissent chez le fournisseur.

### 1.3.2 Enregistrement du vélo

L'enregistrement du vélo est un service clé de Jungle Bike permettant d'identifier le modèle et ses différentes pièces que compose le vélo.

Toutes ces données permettront de faciliter la réparation ou la personnalisation du vélo. La plateforme dédiée à ce service permet au client de renseigner la marque, le type d'activité du vélo, période d'achat, taille des différentes parties du vélo. Le deuxième service de cette plateforme est que les clients pourront être mis en relation avec les réparateurs qui connaissent mieux les vélos pour la réparation ou la personnalisation.

## 1.4 Processus d'intégration de donnée et de mise en ligne des produits

Tout d'abord avant qu'on ait accès à la fiche catalogue des produits des fournisseurs, JungleBike passe un contrat de vente des produits des marques du fournisseur..

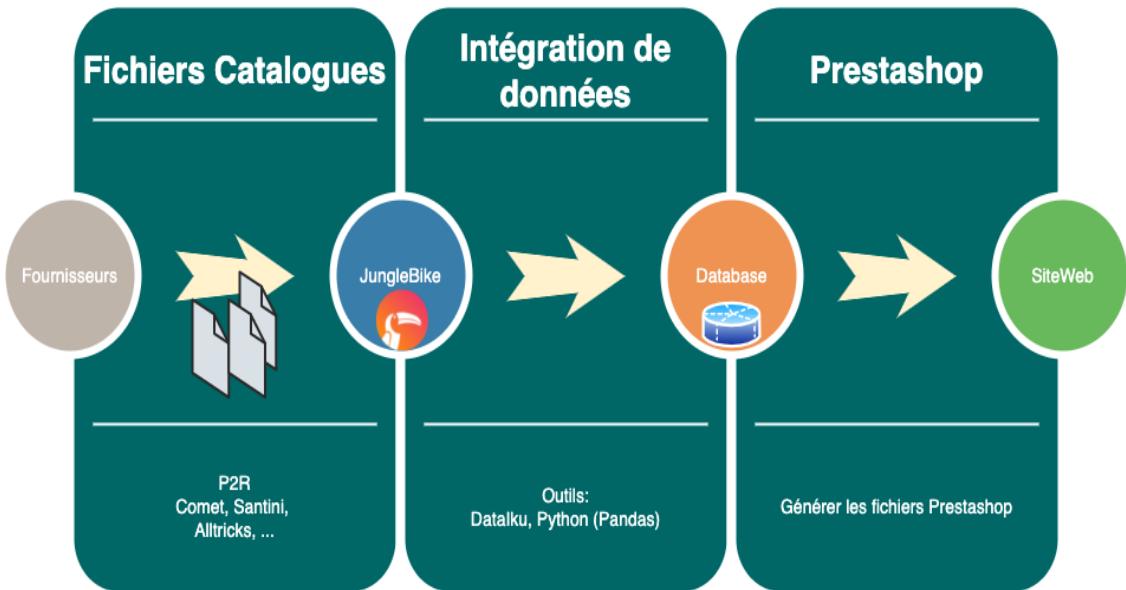


FIG. 1.3: Etapes d'intégration des produits

### 1.4.1 Récupérer le catalogue des produits

Une fois la procédure de vente des produits du fournisseur validée, JungleBike reçoit les produits sous forme de données au format CSV, XLSX, SQL. Extrait du fichier catalogue du fournisseur Mon Zoli Casque

### 1.4.2 Nettoyage et enrichissement et standardisation :

Au début on nettoyait les produits avec Dataiku en créant des pipelines depuis l'extraction des données puis la transformation jusqu'au chargement dans un fichier csv. Ensuite on a décidé de mettre en place un module permettant d'automatiser le processus ETL mais cette fois avec du python. Dans le processus de nettoyage,

ean13	provider	brand	product_name	product_name_decli	cat_label	size	head_size	activity
3770024678253	ankorstore	Mon Zoli Casque	Casque rider	Casque rider bleu marine taille l	Casques vélo l	59-61	ville	
3770024678215	ankorstore	Mon Zoli Casque	Casque rider	Casque rider jaune fluo taille l	Casques vélo l	59-61	ville	
3770024678239	ankorstore	Mon Zoli Casque	Casque rider	Casque rider kaki taille l	Casques vélo l	59-61	ville	
3770024678192	ankorstore	Mon Zoli Casque	Casque rider	Casque rider orange flash taille l	Casques vélo l	59-61	ville	
3770024678246	ankorstore	Mon Zoli Casque	Casque rider	Casque rider bleu marine taille m	Casques vélo m	55-58	ville	
3770024678208	ankorstore	Mon Zoli Casque	Casque rider	Casque rider jaune fluo taille m	Casques vélo m	55-58	ville	
3770024678222	ankorstore	Mon Zoli Casque	Casque rider	Casque rider kaki taille m	Casques vélo m	55-58	ville	
3770024678185	ankorstore	Mon Zoli Casque	Casque rider	Casque rider orange flash taille m	Casques vélo m	55-58	ville	
3770024678178	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling bleu marine taille s	Casques vélo s	52-56	ville	
3770024678116	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling bleu taille s	Casques vélo s	52-56	ville	
3770024678154	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling bubblegum taille s	Casques vélo s	52-56	ville	
3770024678130	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling kaki taille s	Casques vélo s	52-56	ville	
3770024678048	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling menthe glaciale taille s	Casques vélo s	52-56	ville	
3770024678062	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling moutarde taille s	Casques vélo s	52-56	ville	
3770024678093	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling rouge taille s	Casques vélo s	52-56	ville	
3770024678161	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling bleu marine taille xs	Casques vélo xs	46-53	ville	
3770024678109	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling bleu taille xs	Casques vélo xs	46-53	ville	
3770024678147	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling bubblegum taille xs	Casques vélo xs	46-53	ville	
3770024678123	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling kaki taille xs	Casques vélo xs	46-53	ville	
3770024678079	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling menthe glaciale taille xs	Casques vélo xs	46-53	ville	
3770024678055	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling moutarde taille xs	Casques vélo xs	46-53	ville	
3770024678086	ankorstore	Mon Zoli Casque	Casque rolling	Casque rolling rouge taille xs	Casques vélo xs	46-53	ville	
3770024678031	ankorstore	Mon Zoli Casque	Casque baby	Casque baby bleu	Casques vélo xxs	44-48	ville	

FIG. 1.4: Exemple de données fournisseurs

plusieurs étapes suivante sont nécessaire :

1. La catégorisation des produits conformément à toutes les catégories qu'on dispose. Un algorithme de catégorisation est donc appliqué en se basant sur le nom du produit.
2. Le nettoyage du nom du produit : les noms de produits comportant des tailles, couleurs, marques sont enlevés afin d'avoir un nom de produit sans d'autres informations.
3. Extraction des dimensions, couleurs, marque depuis le nom du produit ou depuis la colonne correspondante en suivant le référentiel de chaque colonne.
4. Création des identifiants unique de junglebike pour identifier chaque produit.

#### 1.4.3 Mise en base :

Une fois que les produits ont subi tout le processus de nettoyage et de standardisation, les produits sont chargés dans la base de donnée conçue à cet effet. Les différentes informations de chaque produit sont intégrées dans la table correspondante.

#### LA RECOMMANDATION DES ARTICLES

#### 1.4.4 Mise en ligne

Pour rendre disponible ces produits en ligne, les produits sont chargés depuis la base puis subissent une transformation pour avoir le format spécifique à Prestashop. Cette dernière est uploadé sur le site afin d'être disponible sur le site.

FIG. 1.5: Aperçu des produits qui seront disponible en ligne

## 1.5 JungleBike en interne

### 1.5.1 L'équipe d'accueil :

L'équipe Data est composée de trois Data Scientist dont un intervenant externe. Mon collègue permanent travaille essentiellement sur la mise en place du module de compatibilité entre les produits et du module d'enregistrement des vélos. J'ai travaillé en premier sur le module d'automatisation d'intégration des données en base et sur le site. De plus, je suis amené à mettre en place l'algorithme de recommandation des articles du site en se basant sur les votes.

### 1.5.2 Organisation de travail : le scrum Agile :

Scrum est une méthode agile pour la gestion de projet informatique et a pour objectif d'améliorer la productivité d'une équipe. C'est un cadre de travail au sein duquel les acteurs peuvent aborder des problèmes complexes et adaptatifs, en livrant de manière efficace et créative des produits tout en créant de la valeur ajoutée. Le sprint dure deux semaines et sur cette période chacun travaille sur une ou plusieurs tâches dans le but de fournir un résultat en fin de sprint. Le planificateur de tâche et de travail en équipe principale utilisé est Trello.

### 1.5.3 Outils techniques :

Dans le but de réaliser mes missions au sein de Transaction Connect, cette dernière a mis à ma disposition un ordinateur portable. Comme outils technique ou de communication on dispose entre autres de DBeaver, Pycharm, Slack, Python, DataIku, Scikit Learn, Keras , Pytorch.



# **Première partie**

## **Problématique**



# Sommaire

---

<b>2 Le contexte de résolution du problème</b>	<b>23</b>
2.1 Le problème à résoudre . . . . .	24
2.2 Présentation des données . . . . .	25
2.3 Méthodes de validation des modèles . . . . .	27

---



# Chapitre 2

## Le contexte de résolution du problème

### Sommaire

---

<b>2.1</b>	<b>Le problème à résoudre</b>	<b>24</b>
<b>2.2</b>	<b>Présentation des données</b>	<b>25</b>
2.2.1	Dataset :	25
2.2.2	Détail des colonnes de la dataset :	26
<b>2.3</b>	<b>Méthodes de validation des modèles</b>	<b>27</b>
2.3.1	MSE : Mean Squared Error ou RMSE : Root Mean Squared Error	27
2.3.2	MAE : Mean Absolute Error	27

---

## 2.1 Le problème à résoudre

Bien que la recommandation aujourd’hui soit le moyen efficace d’améliorer non seulement la connaissance client mais aussi bien ciblé ses clients, elle pose des problèmes dans sa mise en place. Les grandes difficultés confrontées dans la mise en place des modèles de recommandation :

1. L’information sur l’avis client :

L’avis du client sur un produit donné peut être explicite soit direct ou implicite. L’avis explicite c’est des formes de likes, votes sur les préférences que le client attribue à un produit après l’avoir testé. Le problème à ce niveau est que peu de clients font des retours d’expérience sur le produits, par conséquent moins d’information pour construire un bon modèle de recommandation.

D’un autre côté on dispose des informations implicites que le client donne en faisant l’analyse de son comportement. C’est l’exemple du temps passé sur une page, les types d’articles qui ont plus de clics... Les avis implicites sont faciles à collecter en masse car ils ne demandent pas d’effort du côté client. Mais, le problème ici aussi est que ces avis implicites ne renseignent pas effectivement l’intérêt ou non du client à un article. La présence des avis implicite des clients en abondance ont permis de faire beaucoup de recherche dans l’amélioration des modèles de recommandation.

2. Manque d’avis clients en comparaison au nombre d’article à recommander :  
On dispose en majorité peu de données sur le vote et les avis clients en rapport avec le nombre produit présent dans la catalogue.

3. Biais de popularité : non diversité des produits recommandés

En effet, l’objectif d’un système de recommandation est d’avoir moins de suggestion en haut de la liste recommandée, et induire plus de diversité dans cette liste recommandée. Par contre certains articles nouveaux par faute de popularité auront moins de chance de faire partie de la liste alors qu’ils pourraient potentiellement intéresser le client.

Le problème à résoudre dans ce mémoire sera d’appliquer des modèles de recommandation au secteur du vélo. Les données sont essentiellement des votes que les clients attribuent à chaque pièce de vélo lors de l’achat.

## 2.2 Présentation des données

### 2.2.1 Dataset :

Les données à étudier contiennent des informations de chaque article avec le vote, l'avis, ... que le client lui a attribué. Ces données sont issues du scrapping des sites des fournisseurs. Elle contient actuellement 30.912 lignes et 11 colonnes.

	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 être un default	j ai acheté ces 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. 2.1: Jeu de donnée

```
#      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. 2.2: Détail des colonnes

**2.2.2 Détail des colonnes de la dataset :**

- product name : Nom du produit que le client a acheté
- brand : la marque du produit
- user : représente le nom du client
- city : la ville où habite le client,
- age : la tranche d'âge du client, elle peut nous aider à identifier pour chaque produit la tranche d'âge d'individus qui s'y intéressent.
- activity : le type d'activité que le client effectue avec son vélo (Compétition, Voyage, ...)
- level : le niveau atteint dans son activité dans la pratique du vélo,
- vote : le score attribué au produit acheté sur le site
- avis : défini le sentiment que porte le client à l'issue de l'achat de l'article,
- description : le commentaire que porte le client sur le produit acheté.

## 2.3 Méthodes de validation des modèles

### 2.3.1 MSE : Mean Squared Error ou RMSE : Root Mean Squared Error

Le Mean Squared Error est une fonction mathématique permettant d'évaluer la performance d'un modèle ayant des valeurs de prédiction continues. C'est la moyenne des erreurs au carré de tous les points de données d'apprentissage ou de validation. Supposons un vecteur de n prédictions noté Y d'un modèle à partir d'une matrice de n donnée X. Le MSE est donné par la formule suivante :

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Le RMSE est donné par la formule suivante :

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

En fonction de l'époque, le modèle apprend en s'améliorant lorsque le MSE converge vers 0.

### 2.3.2 MAE : Mean Absolute Error

Le Mean Absolute Error est une fonction mathématique permettant d'évaluer la performance d'un modèle ayant des valeurs de prédiction continues. C'est la moyenne de la différence absolue entre la prédiction et les valeurs réelles. Sa formule est donnée par :

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

Sa valeur décroît en fonction de l'époque d'apprentissage.

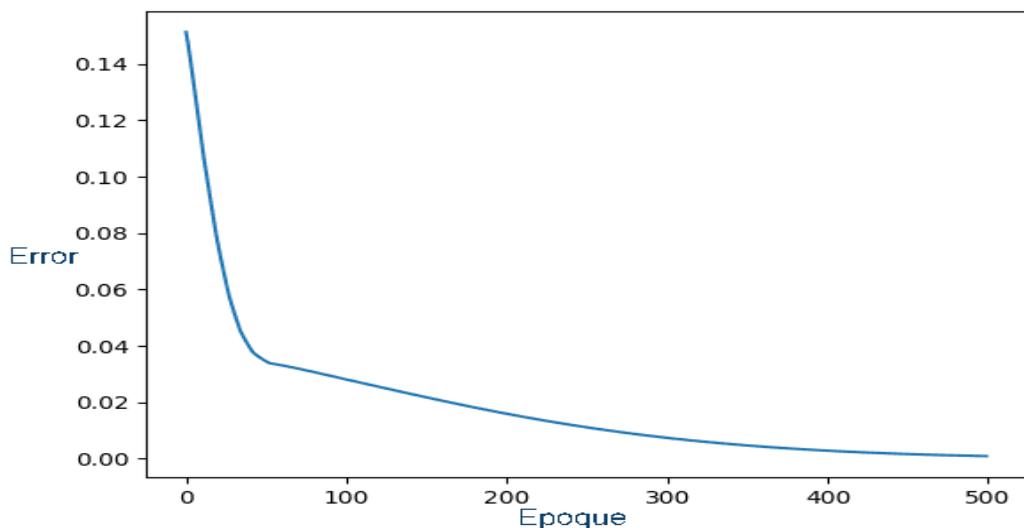


FIG. 2.3: Allure de l'évolution de l'erreur en fonction de l'époque

## **Deuxième partie**

### **État de l'art**



# Sommaire

---

<b>3 État de l'art des techniques de recommandation</b>	<b>33</b>
3.1 Recommandation aléatoire . . . . .	34
3.2 Recommandation Personnalisée . . . . .	34
3.3 Recommandation Objet (Content-Based filtering CB) . . . . .	34
3.4 Recommandation Sociale (Collaborative Filtering CF – Context Aware) . . . . .	35

---



# Chapitre 3

## État de l'art des techniques de recommandation

### Sommaire

---

3.1 Recommandation aléatoire . . . . .	34
3.2 Recommandation Personnalisée . . . . .	34
3.3 Recommandation Objet (Content-Based filtering CB) .	34
3.4 Recommandation Sociale (Collaborative Filtering CF – Context Aware) . . . . .	35
3.4.1 Memory-based CF . . . . .	36
3.4.2 La Matrice de Factorisation . . . . .	36
3.4.3 Neural Collaborative Filtering (NFC) . . . . .	37
3.4.4 LSTM : Long Short Term Memory . . . . .	39

---

### 3.1 Recommandation aléatoire

Bien avant la technologie du machine learning, la recommandation était basée sur des propositions aléatoires des articles au client. Certaines fois les articles les plus populaires sont recommandés. Ce qui pose un problème de personnalisation des produits pour chaque client. Un article peut être populaire mais ne pourra intéresser certains, cela peut devenir contre productif si on se base sur ce type de recommandation. C'est l'exemple d'une recommandation d'articles dépendant du sexe du client.

### 3.2 Recommandation Personnalisée

Cette méthode consiste à recommander un produit sur la base de ses achats précédents, de ses motifs de recherche. Elle vise à proposer au mieux les produits beaucoup plus accessible d'intéresser le client. Mais dans ce cas, si le client n'avait jamais effectué d'achat ou pas assez de commande, il reste difficile de recommander d'autres articles susceptibles de l'intéresser.

### 3.3 Recommandation Objet (Content-Based filtering CB)

Dans ce type de recommandation, on se base sur les caractéristiques que présente le produit et faire la recommandation sur les caractéristiques des produits que le client a déjà choisi (soit recommander des produits similaires à celui choisi dans le panier).

Pour construire le modèle basé sur la méthode du Content Based, il faut tokeniser les caractéristiques des produits ensuite appliquer les méthodes du TF-IDF. Ceci permet d'augmenter l'importance ou la fréquence des mots clés du produits et de réduire les mots inutiles.

Le problème de cette approche de recommandation Objet, est qu'elle nécessite une connaissance profonde des produits à recommander puisqu'elle est basée sur la description ou le nom du produit.

- Exemple : Considérons deux produits qui ont dans leur description les informations suivantes :

P1: "Pneu souple de qualité"

P2: "Pneu de qualité"

Pour connaître la similarité entre ces deux produits on peut appliquer la méthode de similarité basée sur le calcul du cosinus. Tout d'abord, on calcule la fréquence

### 3.4 Recommandation Sociale (Collaborative Filtering CF – Context Aware)

des mots clés que contient chaque description, ensuite on projette chaque produit dans un repère n-dimensionnel. Deux vecteurs P1 et P2 de cet espace ainsi constitués sont similaires si et seulement si le cosinus de leur angle est petit.

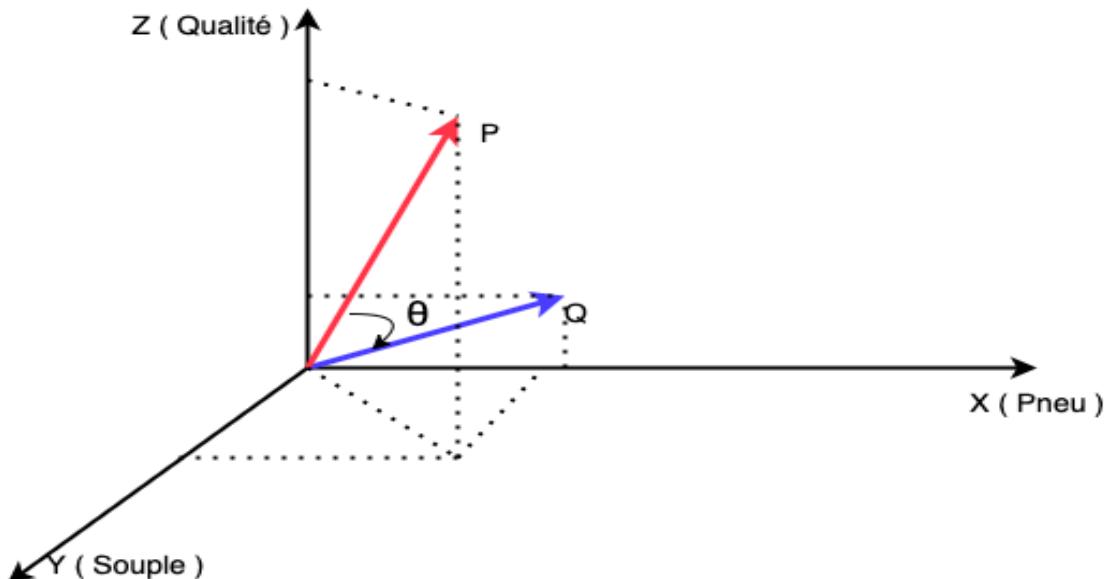


FIG. 3.1: Projection des produits

$$\begin{aligned} \cos(\theta) &= \frac{\mathbf{P} \cdot \mathbf{Q}}{\|\mathbf{P}\| \|\mathbf{Q}\|} \\ &= \frac{\sum_{i=1}^n P_i Q_i}{\sqrt{\sum_{i=1}^n P_i^2} \sqrt{\sum_{i=1}^n Q_i^2}} \end{aligned}$$

## 3.4 Recommandation Sociale (Collaborative Filtering CF – Context Aware)

Basée sur le comportement ou le vote des clients, le modèle du Collaborative Filtering utilise les avis clients sur des produits pour les recommander à d'autres utilisateurs.

On distingue plusieurs approches de collaborative filtering :

### 3.4.1 Memory-based CF

Cette approche se base sur les votes, cliques sur lequel il faut établir une corrélation entre les produits ou entre les utilisateurs afin de recommander un produit quelconque à un utilisateur qui ne l'a jamais vu. Dans ce cas plus précis, les produits recommandés sont ceux achetés par les utilisateurs les plus proches.

### 3.4.2 La Matrice de Factorisation

Cette méthode vise à factoriser la matrice de base obtenue en considérant le vote de chaque client pour chaque article. Cette factorisation permet de simplifier la matrice de base en deux matrices (Client et Article) dont le produit matriciel est similaire à la matrice de base.

$$M = UxI$$

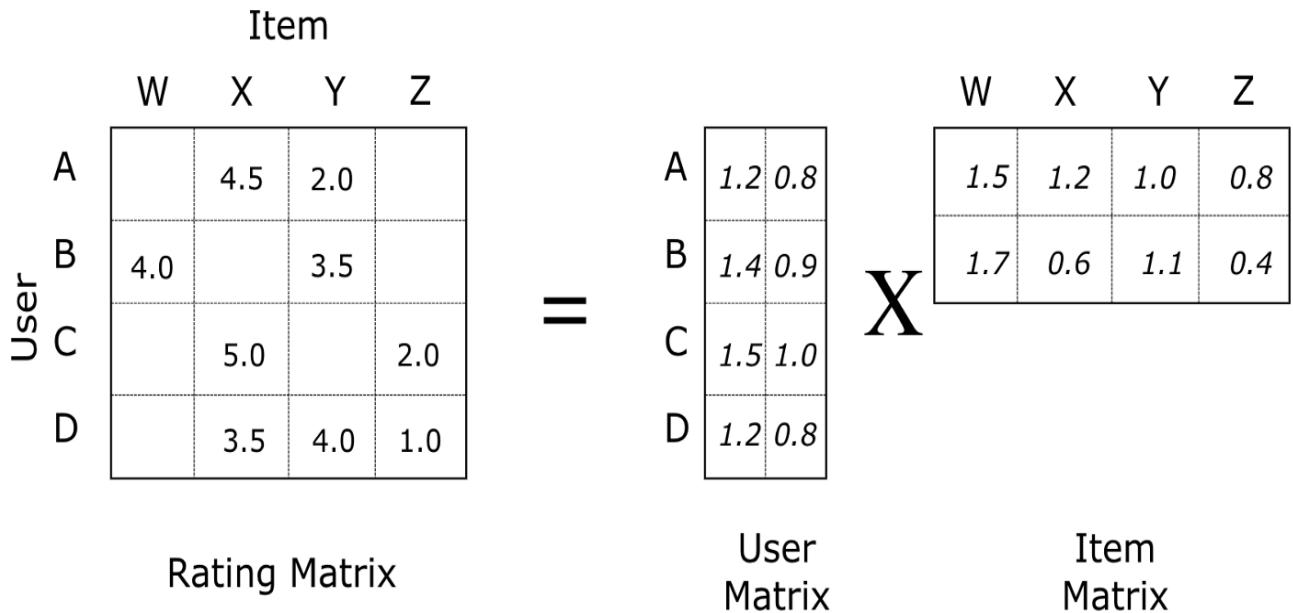


FIG. 3.2: Décomposition de la matrice

<b>3.16</b>	<b>1.92</b>	<b>2.08</b>	<b>1.28</b>
<b>3.63</b>	<b>2.22</b>	<b>2.39</b>	<b>1.48</b>
<b>3.95</b>	<b>2.4</b>	<b>2.6</b>	<b>1.6</b>
<b>3.16</b>	<b>1.92</b>	<b>2.08</b>	<b>1.28</b>

FIG. 3.3: Matrice produit

### 3.4.3 Neural Collaborative Filtering (NFC)

Il est possible d'appliquer le modèle de réseau de neurone au problème de recommandation. A partir de la matrice des clients et des articles, on envoie en entrée du réseau un encodage de vecteur unitaire du client et de l'article. A l'intérieur le vecteur est connecté à plusieurs couches comme par exemple le perceptron multicouche.

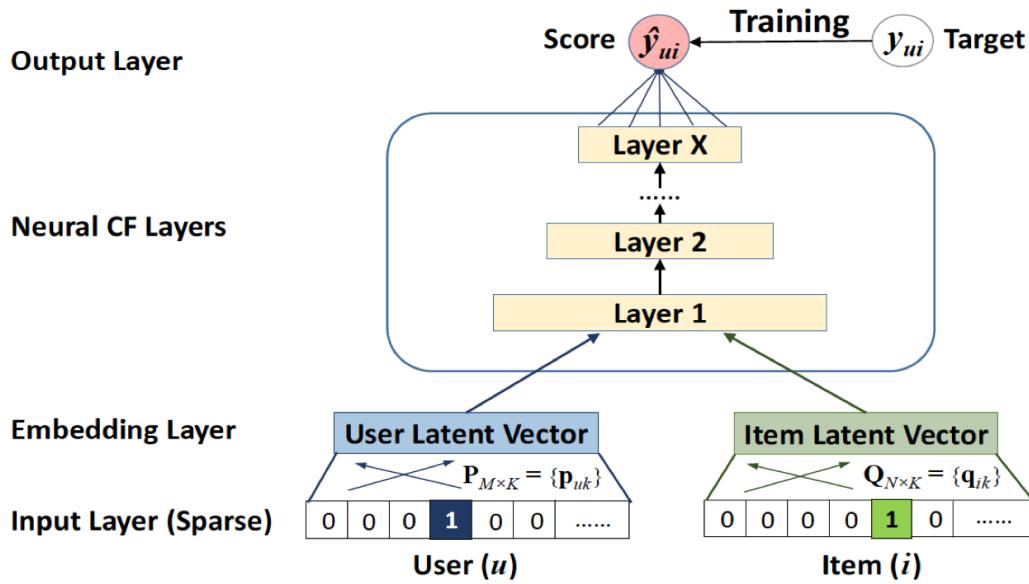


FIG. 3.4: Neural Colaborative Filtering, <https://towardsdatascience.com/neural-collaborative-filtering-96cef1009401>

Cette méthode généralise la méthode de factorisation de matrice. Premièrement, en remplaçant la couche interne avec une unique couche de multiplication, on se retrouve avec le schéma ci-dessous.

Ensuite on initialise le poids de la couche de sortie à une matrice  $J$  dont toutes les valeurs sont égales à 1 et une fonction d'activation linéaire  $L$ .

$$L(x) = x$$

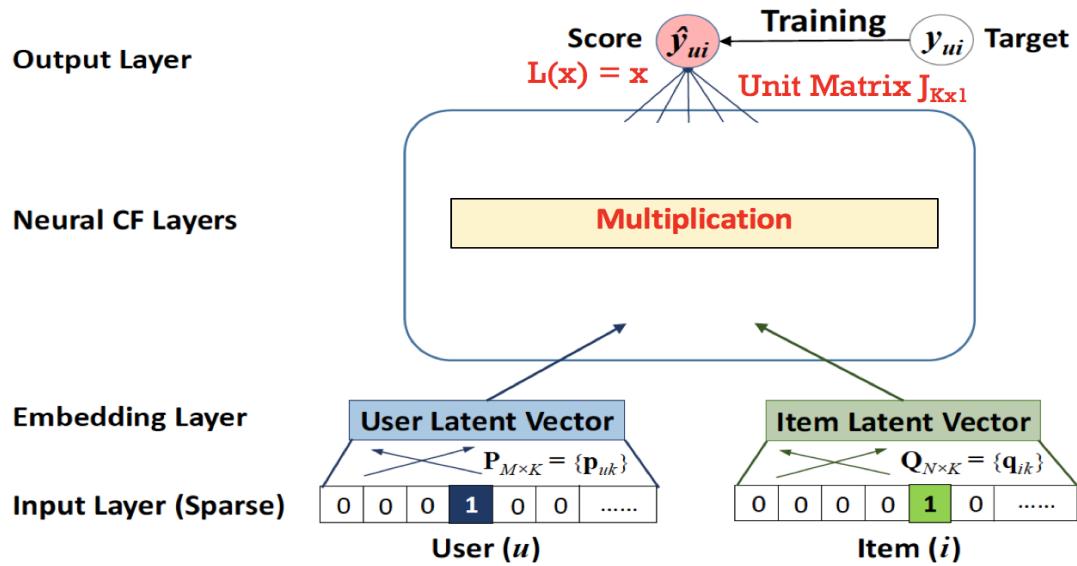


FIG. 3.5: Généralisation du NFC, <https://towardsdatascience.com/neural-collaborative-filtering-96cef1009401>

$$\hat{y}_{ui} = L(p_u \odot q_i \times J_{K \times 1})$$

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

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

FIG. 3.6:

Ce qui revient exactement à une décomposition en un produit de deux matrices. On conclut que la méthode de matrice factorisation est un cas particulier du Neuron Collaborative Filtering.

### 3.4.4 LSTM : Long Short Term Memory

LSTM : Long Short Term Memory est un algorithme de la famille des réseaux de neurone récurrent (RNN)

- Réseau de Neurone récurrent :

Un réseau de neurone récurrent est une succession d'état des neurones qui gagnent des informations du précédent état. Chaque état a une entrée input  $X(t)$  et une sortie output  $h(t)$  définissant la prédiction. La partie A est une couche de neurones dont les informations sont propagées à l'état suivant.

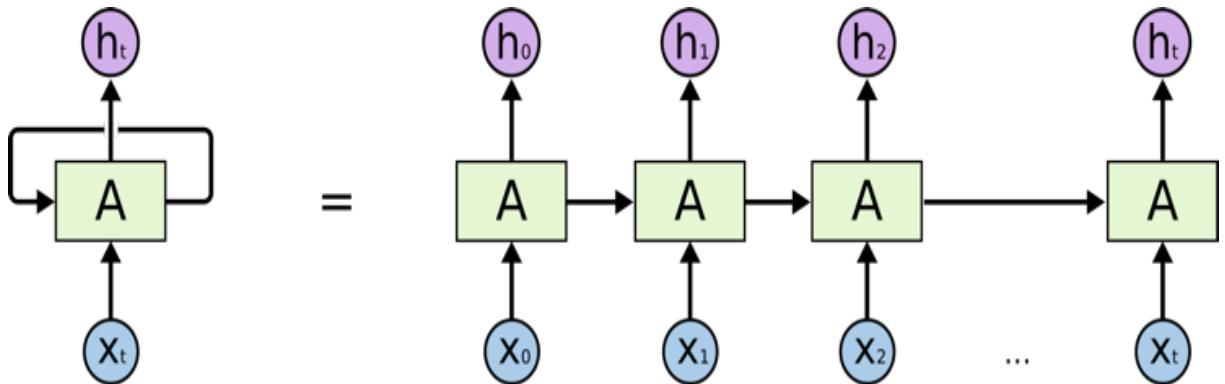


FIG. 3.7: Etats du RNN : <https://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47>

- Cas particulier du LSTM :

Dans chaque couche, il existe quatre portes qui contrôlent le comportement de l'information.

- Input Gate :

Cette porte contrôle s'il faut écrire le vecteur d'entrée dans la mémoire du LSTM  $c(t)$  ou non . Elle comporte une couche de sigmoïde.

- Forget Gate :

Cette porte a une couche de sigmoïde qui déterminent s'il faut supprimer l'information de la mémoire du LSTM  $c(t)$ .

- Candidate Gate :

Cette porte détermine quelle information écrire dans la mémoire  $c(t)$  à partir de la couche de  $\tan(h)$ .

- Output Gate :

Dans cette porte, une fonction sigmoïde aussi détermine quelle information sort en sortie de la l'état caché.

$X$  : Scaling of information

$+$  : Ajout d' information

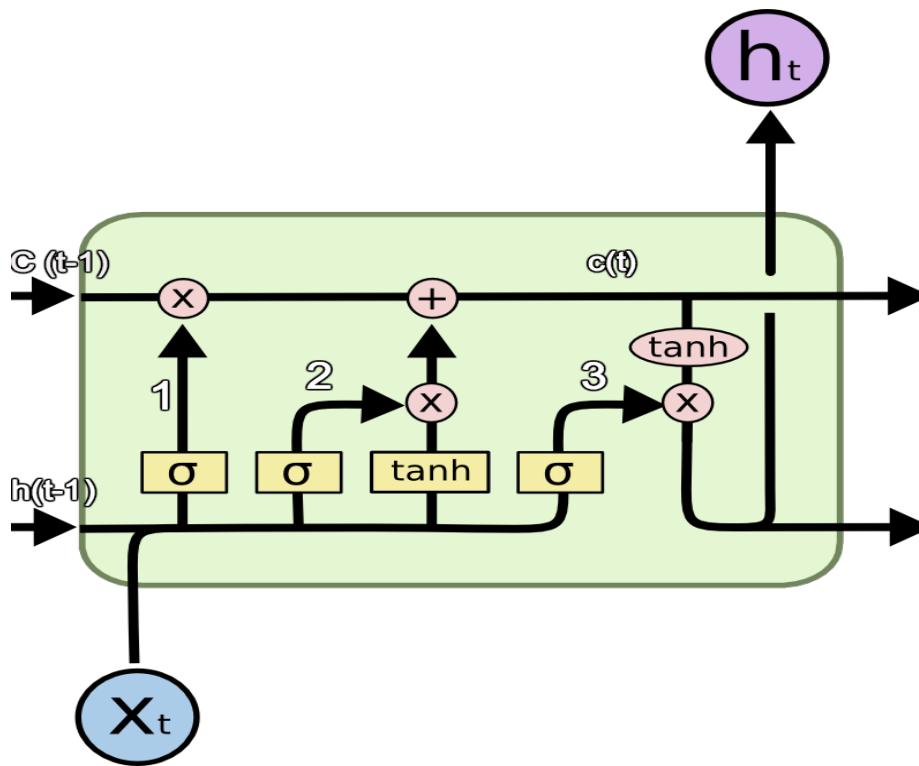


FIG. 3.8: Couche du LSTM : <https://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47>

$\sigma$  : Couche de Sigmoïde (valeurs 0, 1 donc pour faire oublier ou garder l'information)

$\tanh(h)$  : Couche de tangente (pour faire maintenir le gradient non nul plus long-temps)

$h(t - 1)$  : Sortie du précédent LSTM

$c(t - 1)$  : Mémoire du précédent LSTM

$X(t)$  : Vecteur d'entrée en cours

$c(t)$  : Nouvelle mise à jour de la mémoire

$h(t)$  : Sortie en cours



## **Troisième partie**

### **Système réalisé**



# Sommaire

---

<b>4 Implémentation du système</b>	<b>47</b>
4.1 Introduction . . . . .	48
4.2 Analyse de données . . . . .	49
4.3 Application des différents modèles au jeu de données . . . . .	53
4.4 Tests et validation : Analyse des RMSE des différents modèles . . . . .	55
4.5 Conclusion . . . . .	57

---



# Chapitre 4

## Implémentation du système

### Sommaire

---

<b>4.1</b>	<b>Introduction</b>	<b>48</b>
<b>4.2</b>	<b>Analyse de données</b>	<b>49</b>
4.2.1	Aperçu du jeu de donnée :	49
4.2.2	Statistiques sur la dataset :	50
4.2.3	Meilleurs produits et Clients par rapport aux votes :	51
<b>4.3</b>	<b>Application des différents modèles au jeu de données</b>	<b>53</b>
4.3.1	Matrice de Factorisation :	53
4.3.2	Combinaison Matrice de Factorisation et Réseau de Neurone :	53
4.3.3	Combinaison Matrice de Factorisation et Multilayer perceptron :	53
4.3.4	LSTM : Long Short Term Memory	53
<b>4.4</b>	<b>Tests et validation : Analyse des RMSE des différents modèles</b>	<b>55</b>
4.4.1	Statistiques de performance des modèles	55
4.4.2	Evolution d'apprentissage des modèles en fonction du nombre d'époque	55
<b>4.5</b>	<b>Conclusion</b>	<b>57</b>

---

## 4.1 Introduction

Le système de recommandation suivant est construit sur des données récoltées à travers le scraping des sites des fournisseurs de jungle Bike. Après cette étape les données ont subi des séries de nettoyages afin d'avoir la structure idéale pour la construction des modèles de machine learning. Les principaux outils qui interviendront dans la mise place de ce système seront Tensor Flow, Keras et Pytorch. Dans les lignes qui suivront nous allons faire l'exploration de la donnée afin d'avoir des informations statistiques sur les données. Par la suite nous allons appliquer des modèles de deep learning analyser leur performance et choisir le meilleur modèle pour notre jeu de données.

## 4.2 Analyse de données

### 4.2.1 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

FIG. 4.1: Informations sur le produit le client et son vote

#### 4.2.2 Statistiques sur la dataset :

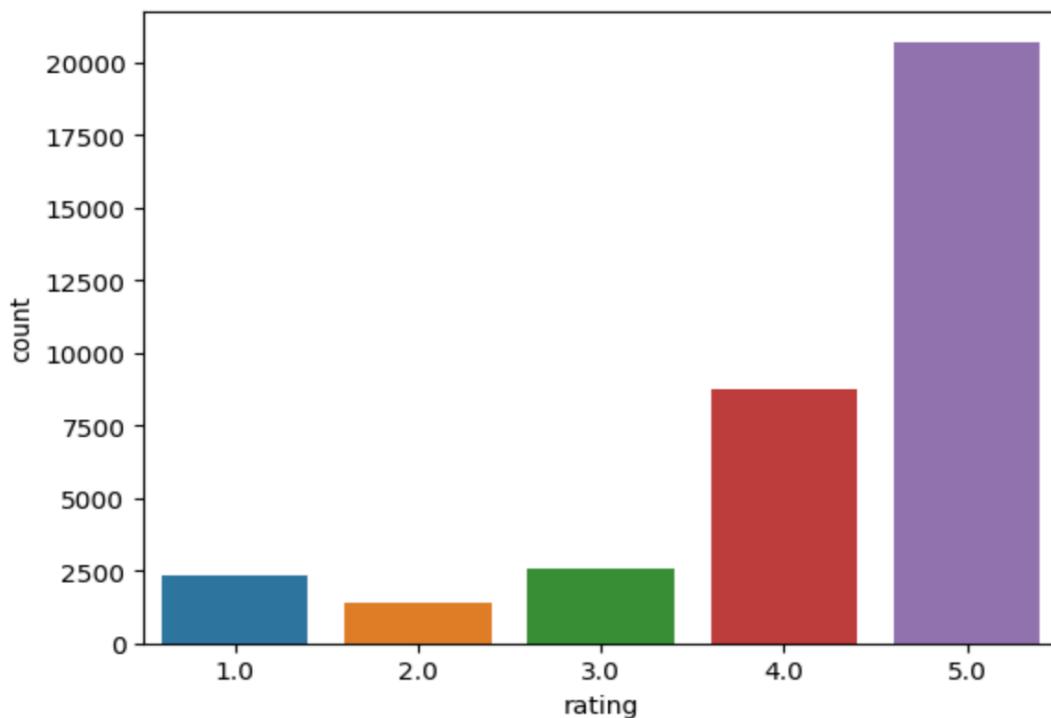


FIG. 4.2: Nombre de produit voté en fonction du score

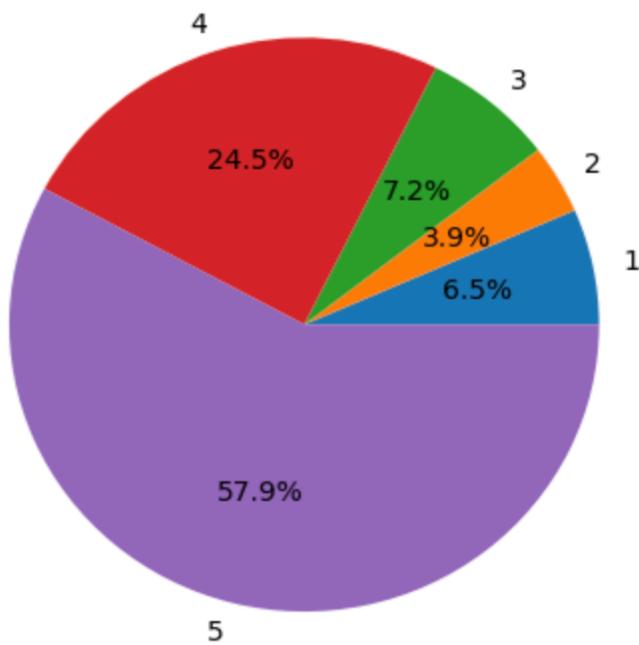


FIG. 4.3: Proportion du nombre de produit voté en fonction du score

#### 4.2.3 Meilleurs produits et Clients par rapport aux votes :

item		product_name	nombre_interaction
<b>1703</b>	1805	VÉLO VTT ÉLECTRIQUE E-ST 900 27,5 PLUS	68
<b>1433</b>	1520	VÉLO VTT ÉLECTRIQUE E-ST STILUS 29"	68
<b>5096</b>	5382	VÉLO TOUT CHEMIN ELECTRIQUE RIVERSIDE 500 E	68
<b>896</b>	955	VÉLO VTT ST 530 MDB 27,5"	67
<b>323</b>	349	VÉLO VTT SEMI RIGIDE ROCKRIDER XC 120 29" SRA...	67
<b>3550</b>	3765	VELO PLIANT A ASSISTANCE ELECTRIQUE TILT 500	65
<b>3131</b>	3318	VTT ENFANT ROCKRIDER ST 120 20 POUCES 6-9 ANS ...	65
<b>4589</b>	4849	VÉLO VTT ST 100 27,5"	64
<b>1769</b>	1876	VÉLO VTT ST 530 27,5"	64
<b>650</b>	700	VÉLO VTT ST 540 V2 BLEU 27,5"	64

FIG. 4.4: Top 10 des meilleurs produits avec plus de la moitié des votes supérieurs au score 3

user		user_name	nb_vote
<b>9326</b>	9327	Nicolas	167
<b>1875</b>	1876	Thierry	156
<b>7821</b>	7822	Julien	147
<b>2180</b>	2181	Christophe	143
<b>9073</b>	9074	Olivier	139
<b>6496</b>	6497	David	133
<b>7182</b>	7183	Laurent	129
<b>1064</b>	1065	Philippe	126
<b>326</b>	327	Pierre	121
<b>6136</b>	6137	Alain	119

FIG. 4.5: Top 10 des Meilleurs clients

## 4.3 Application des différents modèles au jeu de données

### 4.3.1 Matrice de Factorisation :

Tout d'abord on fait passer chacun des vecteurs d'item et user dans une couche d'**Embedding**, Ensuite un produit matriciel est effectué entre le vecteur d'item et de user transformé. Le produit résultant est directement passé au modèle et compilé avec un optimiseur **Adam** avec les métriques **MAE** et **MSE**.

### 4.3.2 Combinaison Matrice de Factorisation et Réseau de Neurone :

On fait passer chacun des vecteurs d'item et user dans une couche d'**Embedding**, Ensuite un produit matriciel est effectué entre le vecteur d'item et de user transformé. Le produit résultant est passé à deux couches **Dense** pour ensuite produire une sortie qui sera passée au modèle. Cette dernière est compilée avec un optimiseur **Adam (learning rate de 0.1)** avec les métriques **MAE** et **MSE**.

### 4.3.3 Combinaison Matrice de Factorisation et Multilayer perceptron :

On fait passer chacun des vecteurs d'item et user dans une couche d'**Embedding**, Ensuite on concatène les vecteurs d'item et de user transformé. Le résultat est passé aux couches **Dense**, **BatchNormalization**, **Dense**, **BatchNormalization**, pour ensuite produire une sortie du multiplayer perceptron. On reprend les vecteurs d'entrée passés en dans les couches d'**Embedding** qui vont subir un produit matriciel. Cette dernière est combinée à la sortie du perceptron multicouche précédent pour passer dans la couche **Dense**. La sortie résultante passe dans le modèle et compilé avec un optimiseur **Adam (learning rate de 0.01)** avec les métriques **MAE** et **MSE**.

### 4.3.4 LSTM : Long Short Term Memory

Pour ce modèle on transforme le jeu de donnée pour avoir chaque user avec les produits interagit et le son score sous la forme : Le modèle est formé d'une couche d'**Embedding** avant de passer dans la couche de **LSTM**. Une couche **Linéaire** est appliquée à la sortie pour produire la sortie définitive. La métrique d'évaluations **Mean Squared Error** est appliquée pour les étapes d'apprentissage et de test.

user		item_vote
<b>11065</b>	11066	([1440], [5.0])
<b>10359</b>	10360	([4885, 2671, 175, 3508, 1733], [1.0, 5.0, 5.0...)
<b>9747</b>	9748	([1110, 1110], [2.0, 5.0])
<b>8257</b>	8258	([5393], [4.0])
<b>8156</b>	8157	([2440], [5.0])

FIG. 4.6: Utilisateurs avec leurs interactions avec les produits

## 4.4 Tests et validation : Analyse des RMSE des différents modèles

### 4.4.1 Statistiques de performance 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. 4.7: Statistiques de performances des modèles

### 4.4.2 Evolution d'apprentissage des modèles en fonction du nombre d'époque

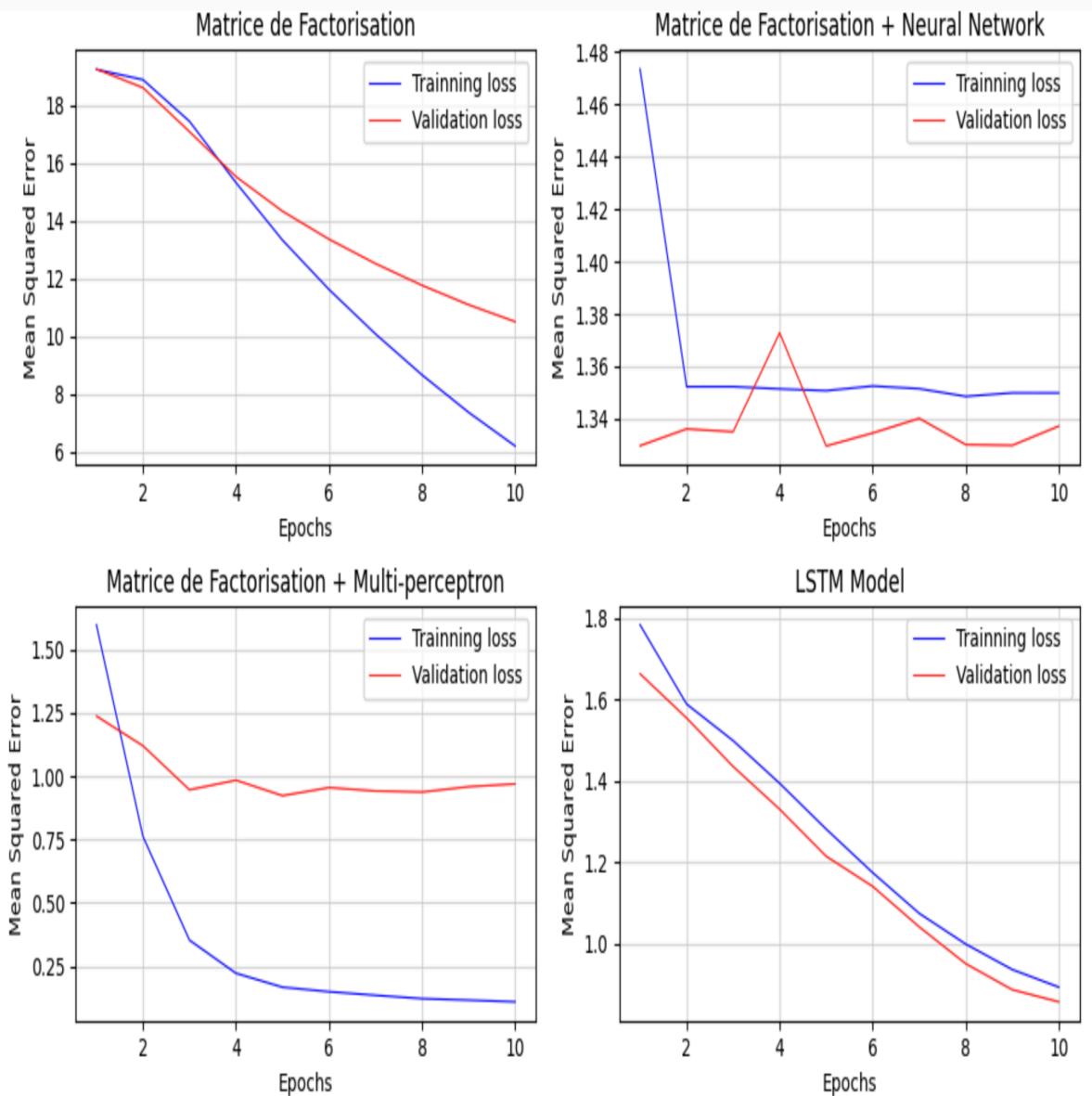


FIG. 4.8: Etude de la validation des modèles

## 4.5 Conclusion

D'après les résultats, on remarque tout d'abord que la matrice de factorisation seule est le modèle le moins performant avec plus d'erreurs. En associant le modèle de matrice de factorisation avec les réseaux de neurones on remarque des améliorations sur les résultats malgré que le temps d'apprentissage augmente. En prenant en compte la courbe de perte d'erreur nous pouvons remarquer que le modèle issu de la combinaison de la matrice de factorisation avec les perceptron multicouche ont enregistré moins d'erreurs à l'apprentissage.



# **Quatrième partie**

## **Conclusion**



# **Sommaire**

---

<b>Conclusion</b>	<b>63</b>
-------------------	-----------

---



# Conclusion

Le système de recommandation est un moyen efficace pour améliorer le rendement des entreprises et est devenu essentiel dans le secteur du commerce en ligne. Depuis longtemps, la recommandation a connu plusieurs transformations, de la recommandation aléatoire au personnalisée puis basée sur le contenu des produits avec leurs similarités. Aujourd’hui, il en sort plusieurs d’autres approches telle que l’application des modèles de Deep Learning. On a essayé de parcourir toutes ces méthodes qui existent et de comprendre leur fonctionnement. On a compris que la méthode basée sur le Deep Learning est plus efficace car elle permet de généraliser la méthode du Matrice de factorisation. Plus tard, il serait possible de combiner la méthode du Content Based et le Collaborative Filtering dans le but de cibler au mieux les préférences des clients.



# Bibliographie

- [1] Sumit Sidana. Recommendation systems for online advertising. Computers and Society [cs.CY]. Université Grenoble Alpes, 2018. English. ffNNT : 2018GREAM061ff. fftel-02060436ff
- [2] D Gunawan et al. The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. 2018 J. Phys. : Conf. Ser. 978 012120
- [3] 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



# Table des figures

1.1	Chiffres actuels sur l'évolution de JungleBike . . . . .	11
1.2	Informations juridique de JungleBike . . . . .	12
1.3	Etapes d'intégration des produits . . . . .	14
1.4	Exemple de données fournisseurs . . . . .	15
1.5	Aperçu des produits qui seront disponible en ligne . . . . .	16
2.1	Jeu de donnée . . . . .	25
2.2	Détail des colonnes . . . . .	25
2.3	Allure de l'evolution de l'erreur en fonction de l'époque . . . . .	28
3.1	Projection des produits . . . . .	35
3.2	Décomposition de la matrice . . . . .	36
3.3	Matrice produit . . . . .	37
3.4	Neural Colaborative Filtering, <a href="https://towardsdatascience.com/neural-collaborative-filtering-96cef1009401">https ://towardsdatascience.com/neural-collaborative-filtering-96cef1009401</a> . . . . .	38
3.5	Généralisation du NFC, <a href="https://towardsdatascience.com/neural-collaborative-filtering-96cef1009401">https ://towardsdatascience.com/neural-collaborative-filtering-96cef1009401</a> . . . . .	39
3.6	. . . . .	39
3.7	Etats du RNN : <a href="https://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47">https ://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47</a> 40	40
3.8	Couche du LSTM : <a href="https://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47">https ://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47</a> 41	41
4.1	Informations sur le produit le client et son vote . . . . .	49
4.2	Nombre de produit voté en fonction du score . . . . .	50
4.3	Proportion du nombre de produit voté en fonction du score . . . . .	51

---

4.4	Top 10 des meilleurs produits avec plus de la moitié des votes supérieurs au score 3 . . . . .	52
4.5	Top 10 des Meilleurs clients . . . . .	52
4.6	Utilisateurs avec leurs interactions avec les produits . . . . .	54
4.7	Statistiques de performances des modèles . . . . .	55
4.8	Etude de la validation des modèles . . . . .	56

# Liste des tableaux



# Table des matières

<b>Remerciements</b>	<b>5</b>
<b>Introduction Générale</b>	<b>7</b>
<b>1 Présentation de l'entreprise</b>	<b>9</b>
1.1 L'histoire de JungleBike . . . . .	10
1.2 Chiffres de JungleBike . . . . .	11
1.3 Les Services proposés . . . . .	12
1.3.1 Accès des produits en ligne : . . . . .	12
1.3.2 Enregistrement du vélo . . . . .	12
1.4 Processus d'intégration de donnée et de mise en ligne des produits .	14
1.4.1 Récupérer le catalogue des produits . . . . .	14
1.4.2 Nettoyage et enrichissement et standardisation : . . . . .	14
1.4.3 Mise en base : . . . . .	15
1.4.4 Mise en ligne . . . . .	16
1.5 JungleBike en interne . . . . .	17
1.5.1 L'équipe d'accueil : . . . . .	17
1.5.2 Organisation de travail : le scrum Agile : . . . . .	17
1.5.3 Outils techniques : . . . . .	17
<b>I Problématique</b>	<b>19</b>
<b>2 Le contexte de résolution du problème</b>	<b>23</b>
2.1 Le problème à résoudre . . . . .	24
2.2 Présentation des données . . . . .	25

2.2.1	Dataset : . . . . .	25
2.2.2	Détail des colonnes de la dataset : . . . . .	26
2.3	Méthodes de validation des modèles . . . . .	27
2.3.1	MSE : Mean Squared Error ou RMSE : Root Mean Squared Error . . . . .	27
2.3.2	MAE : Mean Absolute Error . . . . .	27
<b>II</b>	<b>État de l'art</b>	<b>29</b>
<b>3</b>	<b>État de l'art des techniques de recommandation</b>	<b>33</b>
3.1	Recommandation aléatoire . . . . .	34
3.2	Recommandation Personnalisée . . . . .	34
3.3	Recommandation Objet (Content-Based filtering CB) . . . . .	34
3.4	Recommandation Sociale (Collaborative Filtering CF – Context Aware) . . . . .	35
3.4.1	Memory-based CF . . . . .	36
3.4.2	La Matrice de Factorisation . . . . .	36
3.4.3	Neural Collaborative Filtering (NFC) . . . . .	37
3.4.4	LSTM : Long Short Term Memory . . . . .	39
<b>III</b>	<b>Système réalisé</b>	<b>43</b>
<b>4</b>	<b>Implémentation du système</b>	<b>47</b>
4.1	Introduction . . . . .	48
4.2	Analyse de données . . . . .	49
4.2.1	Aperçu du jeu de donnée : . . . . .	49
4.2.2	Statistiques sur la dataset : . . . . .	50
4.2.3	Meilleurs produits et Clients par rapport aux votes : . . . . .	51
4.3	Application des différents modèles au jeu de données . . . . .	53
4.3.1	Matrice de Factorisation : . . . . .	53
4.3.2	Combinaison Matrice de Factorisation et Réseau de Neurone : . . . . .	53
4.3.3	Combinaison Matrice de Factorisation et Multilayer percep- tron : . . . . .	53
4.3.4	LSTM : Long Short Term Memory . . . . .	53

4.4 Tests et validation : Analyse des RMSE des différents modèles . . . . .	55
4.4.1 Statistiques de performance des modèles . . . . .	55
4.4.2 Evolution d'apprentissage des modèles en fonction du nombre d'époque . . . . .	55
4.5 Conclusion . . . . .	57
 <b>IV Conclusion</b>	 <b>59</b>
 <b>Conclusion</b>	 <b>63</b>