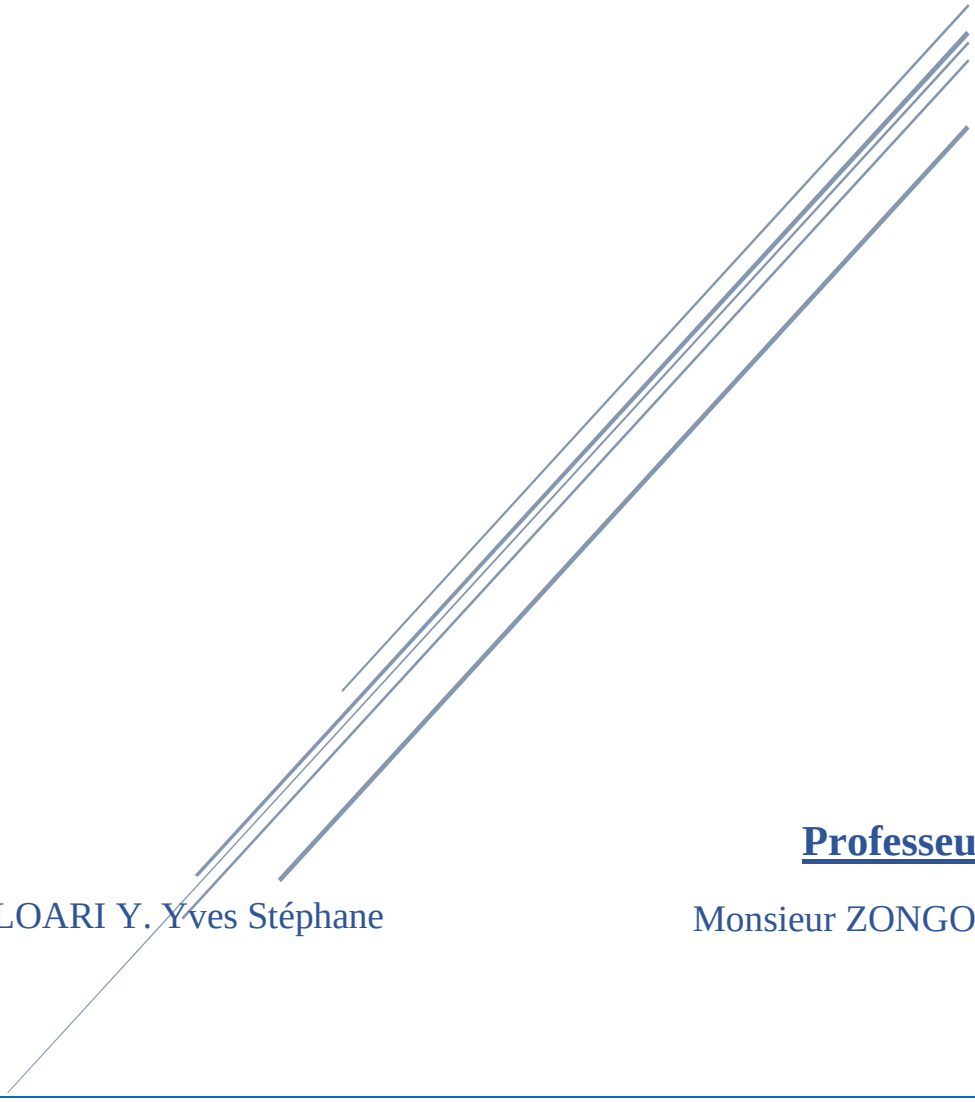
RAPPORT DU PROJET DE IA

Thème : SYSTEME DE RECOMMANDATION



LOARI Y. Yves Stéphane

Professeur

Monsieur ZONGO Sylvain

SOMMAIRE

Introduction

L’intelligence artificielle (IA) est une technologie qui bouleverse le monde dans de nombreux domaines tels que l’économie, la robotique, la médecine, l’agriculture, etc…

Elle fait référence à des systèmes imitant l’intelligence humaine pour effectuer des tâches et qui peuvent s’améliorer en fonction des informations collectées grâce à l’itération. Les moteurs ou systèmes de recommandation sont l’une des nombres formes de l’intelligence artificielle. Un système de recommandation permet de comparer le profil d'un utilisateur à certaines caractéristiques de référence, et cherche à prédire l’avis que donnerait un utilisateur.

1) Exploration des données

Dans cette partie, nous avons réalisé une exploration rapide avant et après le traitement des données. Cette exploration a permis de générer deux rapports détaillants tout notre jeu de données. Nous joindrons ces deux rapports à notre rapport.

Le traitement des données signifie manipuler notre jeu de données afin d’obtenir des informations plus précises qui seront utilisées pour travailler.

Nous avons constaté que la colonne "GroupPrice" ne faisait pas partir de notre jeu de données et nous avons dû le rajouter.

Nous sommes passés à l’étape suivante "la vérification des données manquantes".

	Colonne	Nbre de valeurs manquantes	% de Valeurs manquantes:
0	CustomerID	138727	25.599686
1	Description	7489	1.381966
2	StockCode	6035	1.113656
3	Quantity	6035	1.113656
4	InvoiceDate	6035	1.113656
5	UnitPrice	6035	1.113656
6	Country	6035	1.113656
7	GroupPrice	6035	1.113656
8	InvoiceNo	0	0.000000

Image 1 : vérification des données manquantes

Apres vérification, nous avons constaté qu’il manquait plus de données dans la colonne « CustomerID ».

Nous avons réalisé un histogramme de nos données manquantes.

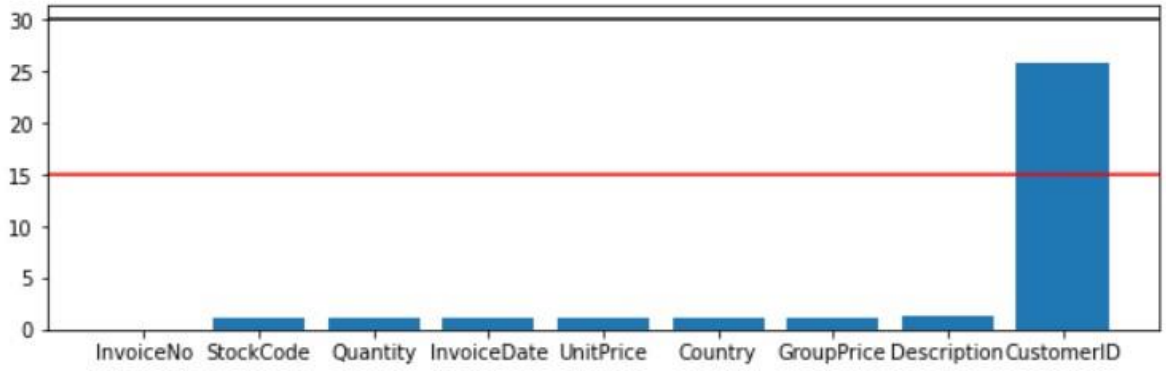


Image 2 : histogramme des données manquantes

Nous avons visualisé nos données manquantes avant traitement.

```
les dimensions du jeu de données sont après ajout de la colonne GroupPrice:  (541909, 9)
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	GroupPrice
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6.0	12/1/2010 8:26	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6.0	12/1/2010 8:26	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8.0	12/1/2010 8:26	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6.0	12/1/2010 8:26	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6.0	12/1/2010 8:26	3.39	17850.0	United Kingdom	20.34
...	...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12.0	12/9/2011 12:50	0.85	12680.0	France	10.20
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6.0	12/9/2011 12:50	2.10	12680.0	France	12.60
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4.0	12/9/2011 12:50	4.15	12680.0	France	16.60
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4.0	12/9/2011 12:50	4.15	12680.0	France	16.60
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3.0	12/9/2011 12:50	4.95	12680.0	France	14.85

541909 rows × 9 columns

Image 3 : visualisation des données avant traitement

Nous avons visualisé nos données manquantes après traitement.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	GroupPrice
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6.0	12/1/2010 8:26	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6.0	12/1/2010 8:26	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8.0	12/1/2010 8:26	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6.0	12/1/2010 8:26	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6.0	12/1/2010 8:26	3.39	17850.0	United Kingdom	20.34
...	...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12.0	12/9/2011 12:50	0.85	12680.0	France	10.20
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6.0	12/9/2011 12:50	2.10	12680.0	France	12.60
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4.0	12/9/2011 12:50	4.15	12680.0	France	16.60
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4.0	12/9/2011 12:50	4.15	12680.0	France	16.60
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3.0	12/9/2011 12:50	4.95	12680.0	France	14.85

398003 rows × 9 columns

Image 4 : visualisation des données après traitement

2) Division du jeu de données en 3 datasets :

Nous avons divisé notre jeu de données en trois datasets :

- Le dataset d'entrainement est le jeu de données utilisé pour entrainer le model. Il vaut 40% de notre jeu de données.
- Le dataset de test est le jeu de données utilisé pour tester le modèle et il vaut 30% de notre jeu de données.
- Le dataset de validation est le jeu de données utilisé pour valider le modèle et il vaut 30% de notre jeu de données.

3) Implémentation du système de recommandation de produits aux clients (avec un algorithme de notre choix)

Fp Growth est un modèle de Data Mining basé sur des règles d'association.

Ce modèle permet, à partir d'un historique des transactions, de déterminer l'ensemble des règles d'association les plus fréquentes dans le jeu de données. Pour cela, il a besoin comme paramètre d'entrée de l'ensemble des transactions composé des paniers de produits que les clients ont déjà achetés.

Étant donné un ensemble de données de transactions, la première étape de la croissance FP consiste à calculer les fréquences des articles et à identifier les articles fréquents.

La deuxième étape de la croissance FP utilise une structure arborescente de suffixes (FP-tree) pour coder les transactions sans générer explicitement des ensembles candidats, qui sont généralement coûteux à générer. Après la deuxième étape, les itemsets fréquents peuvent être extraits de l'arbre FP et le modèle renvoie un ensemble de règles d'association de produits comme dans l'exemple ci-dessous :

{Produit A + Produit B} --> {Produit C} avec 60 % de probabilité

{Produit B + Produit C} --> {Produit A + Produit D} avec une probabilité de 78 %

{Produit C} --> {Produit B + Produit D} avec 67 % de probabilité

Pour établir ce tableau, il faut munir le modèle de 2 hyperparamètres :

- minSupRatio : support minimum pour qu'un itemset soit identifié comme fréquent. Par exemple, si un élément apparaît 3 transactions sur 5, il a un support de 3/5=0,6.
- minConf :confiance minimale pour générer la règle d'association. La confiance est une indication de la fréquence à laquelle une règle d'association s'est avérée vraie. Par exemple, si dans l'itemset des transactions X apparaît 4 fois, X et Y ne coexistent que 2 fois, la confiance pour la règle X => Y est alors 2/4 = 0,5. Le paramètre n'affectera pas l'exploration des ensembles d'éléments fréquents, mais spécifiera la confiance

minimale pour générer des règles d'association à partir d'ensembles d'éléments fréquents.

Une fois les règles d'association calculées, il ne vous reste plus qu'à les appliquer aux paniers produits des clients.

4) Evaluation du modèle à l'aide des métriques vues en cours

L'algorithme Fp Growth utilise déjà ses propres métriques. Nous avons donc utilisé une autre méthode pour tester les métriques vues en classe sur notre jeu de données.

Les métriques de classification et de matrix de confusion (voir code).

```
#affichage de la métrique de classification
print(classification_report(y_test, y_predict, target_names=class_names))
```

```
              precision    recall  f1-score   support

     class_0       0.95      0.90      0.92        20
     class_1       0.95      0.87      0.91        23
     class_2       0.71      0.91      0.80        11

    accuracy                           0.89        54
   macro avg       0.87      0.89      0.88        54
weighted avg       0.90      0.89      0.89        54
```

```
#affichage de la métrique de matrix de confusion
print(confusion_matrix(y_test, y_predict))
```

```
[[18  1  1]
 [ 0 20  3]
 [ 1  0 10]]
```

Image 5 : métriques de classification et de matrix de confusion

5) Interprétation des résultats obtenus

La précision, le rappel et le score f1 sont des mesures très populaires dans l'évaluation d'un algorithme de classification.

Pour rappel, la précision est le nombre de vrais positifs divisé par le nombre total de prédictions positives. En d'autres termes, la précision découvre quelle fraction des positifs prédits est réellement positive.

Dans notre cas, la précision (accuracy en anglais) est de 0.89, ce qui est très élevée. Donc le pourcentage de prédictions positives est de 89%.

Le rappel (recall) est le vrai positif divisé par le vrai positif et le faux négatif. En d'autres termes, le rappel mesure la capacité du modèle à prédire les positifs. Donc le pourcentage de cette capacité du modèle est de 89%.

Le score F1 (f1-score) est la moyenne harmonique de la précision et du rappel. Juste comme une mise en garde, ce n'est pas la moyenne arithmétique. Si la précision est de 0 et le rappel de 1, le score f1 sera de 0 et non de 0,5.

La macro-précision (macro avg) moyenne est la moyenne arithmétique simple de la précision de toutes les étiquettes. Elle est de 87% à 89%.

6) Proposition d'une approche d'amélioration du modèle

Pour améliorer ce modèle,

Conclusion

Comme conclusion nous avons fait l'exploration des données c'est-à-dire avant et après traitement, ensuite nous avons divisés les données en 3 dataset et nous avons et nous avons implémenter un algorithme de système de recommandation de donnée de produits aux clients et un autre algorithme pou le test de métriques vues en classe.