POLYTECHNIQUE MONTRÉAL

UNIVERSITÉ D'INGÉNIERIE

ANALYSE D'UNE BASE DE DONNEES INDUSTRIELLES (RAPPORT)

E-Commerce Shipping Data

Rédigé par :
Thamer Saraei - Christian Njon - Haithem Mzoughi

Professeur :
Bruno Agard

IND6212 - Exploration de données industrielles - Hiver 2021

Sommaire

1. Introduction 2
2. Description des données 2
2.1. La sources de données 2
2.2. Les variables 2
3. Analyse du problème et proposition de solutions 3
3.1. Analyse du problème 3
3.2. Proposition de solutions 3
3.3. Choix de l'outil de travail 4
4. Méthodologie et implémentation 4
4.1. Méthodologie de travail 4
4.2. Implémentation des modèles 5
5. Présentation des résultats (validation du modèle) 5
5.1. Performance sur l'ensemble de TEST 5
5.2. Interprétation des résultats 5
6. Conclusion 6

1. Introduction

La valorisation des données constitue de nos jours un enjeu stratégique pour les entreprises en quête de performance. Elle s'impose comme outil majeur dans l'accélération de la prise de décision et l'amélioration des processus et services. Dans le cadre de ce projet nous aiderons une entreprise internationale spécialisée dans la vente en ligne de produits électroniques, à valoriser ses données de suivi des expéditions des commandes.

Il est en effet question de développer des modèles permettant de prédire si un colis arrivera dans les délais. Et dans la même lancée, établir les relations entre les variables, ainsi que l'influence de celles-ci sur le résultat.

Dans la suite de ce rapport nous présenterons tout d'abord les données à étudier, l'analyse du problème et les solutions proposées, ensuite la méthodologie et l'implémentation des solutions, et enfin les résultats obtenus et les conclusions induites.

2. Description des données

2.1. La sources de données

Les données que nous allons traiter proviennent de la plateforme Kaggle suivant le lien (https://www.kaggle.com/prachi13/customer-analytics). Kaggle est une plateforme web qui organise des compétitions en science de données sur des problématiques réelles et des « données réelles » généralement mises à disposition par des entreprises ou des institutions.

2.2. Les variables

Le jeu de données sur lequel est basé notre étude est constitué de 10999 observations et 12 variables déclinées en une variable de réponse, 4 variables catégoriques d'entrée et 7 variables numériques d'entrée.

Les données contiennent les informations suivantes :

	Type de variable	Nombre de modalité	Description
ID	N/A	N/A	ID du client
Warehouse block	Catégorique (nominale)	**Cinq (05)** A, B, C, D, F	La société dispose d'un grand entrepôt qui est divisé en blocs
Mode of shipment	Catégorique (nominale)	**Trois (03)** Ship, Flight, Road	Les produits sont expédiés de plusieurs façons : par bateau, avion et route
Customer care calls	Numérique	**-**	Le nombre d'appels passés pour des demandes de renseignements sur l'expédition.

Customer rating	Catégorique (ordinale)	**Cinq (05)** 1,2,3,4,5	L'évaluation de l'entreprise donnée par chaque client, 1 est la note la plus basse et 5 est la plus haute.
Cost of the product	Numérique	**-**	Coût du produit en dollars américains
Prior purchases	Numérique	**-**	Le nombre d'achats antérieurs.
Product importance	Catégorique (ordinale)	**Trois (03)** Faible, Moyen, Élevé	Les produits sont classés selon des niveaux d'importance
Sexe	Catégorique (nominale)	**Deux (02)** Homme, Femme	Le genre du client
Discount offered	Numérique	**-**	Valeur de la remise offerte sur ce produit spécifique
Weight in gms	Numérique	**-**	Le poids du colis en grammes
Reached on Time	Catégorique (nominale)	**Deux (02)** 0, 1	**Variable de réponse**, où 1 = délai de livraison NON respecté 0 = délai de livraison respecté

3. Analyse du problème et proposition de solutions

3.1. Analyse du problème

L'interprétation technique de la problématique et la nature de la variable de réponse nous permettent de déduire qu'il s'agit d'un problème de **classification binaire** (à 2 modalités de sortie 1 ou 0).

Cependant, la problématique a une double dimension :

- La prédiction sur le respect du délai de livraison
- La découverte de l'influence des variables d'entrée. Autrement dit, dévoiler les liens cachés (hidden patterns) entre ces variables d'entrée et la réponse, en vue de la mise en place éventuellement d'actions correctives.

3.2. Proposition de solutions

Les solutions aux problèmes de classification binaire en science de données sont multiples, leur mise en œuvre et leurs performances dépendent de l'objectif visé et de la nature de la structure des données.

Ci-dessous les approches possibles :

	Avantages	**Inconvénients**
Les arbres de décision	- Variables d'entrée de type catégorique ou continue - Interprétation claire du système	- Elagage du système (pour éviter le surapprentissage)
Les réseaux de neurone	- Modèle très performant - Variables d'entrée de type catégorique ou continue	- Effet boite noire (impossible de de faire une interprétation du système)
La régression logistique	- Spécialisé dans la classification binaire	- Requiert encodage de toutes les variables catégoriques

Sur la base des données dont nous disposons, et de la spécificité du problème à résoudre, nous porterons notre choix sur **« les arbres de décision » et « la régression logistique »**

3.3. Choix de l'outil de travail

Nous utiliserons le langage de programme Python pour le développement de notre solution. Le choix s'est porté sur ce langage pour les raisons suivantes :

- Il est ouvert et accessible à tous
- Il est adapté aux grandes bases de données
- C'est le langage que nous maitrisons le mieux

4. Méthodologie et implémentation

4.1. Méthodologie de travail (voir plus de détails sur le fichier Python)

4.1.1. Préparation des données

Bien que les jeux de données en provenance de Kaggle soient généralement structurés et prétraités, nous avons néanmoins procédé à un contrôle de la qualité. Il en ressort:

- Aucune valeur manquante (voir fichier Python)
- 122 valeurs aberrantes obtenues par segmentation LOF (Local Outlier Factors) en 3 groupes
- Aucune valeur redondante
- Les données sont statiques dans le temps

Par ailleurs, il a fallu encoder sous Python la variable catégorique ordinale « Product Importance » suivant : Low = 1, Medium = 2, High = 3, afin de définir le niveau d'ordre.

4.1.2. Découpage des données

Nous avons découpé l'ensemble de données en :

- Données d'apprentissage (80%)
- Données de test (20%)

Pour garantir la représentativité des échantillons, nous avons procédé à un découpage assurant la constance de la variance via la fonction train_test_split avec le paramètre Stratify = None

4.1.3. Exploration visuelle

A partir de l'exploration visuelle des données (disponible dans le code) nous voyons que :

- Environ 68% des modes d'expédition sont faites par bateau, les autres étant par vol et route.
- Environ 48% de l'importance du produit est classée comme faible.
- Le sexe des clients semble être réparti de manière égale avec environ 50,4% de femmes.
- Environ 60% des colis ne sont pas livré à temps (On peut dire qu'on a une Balanced dataset)

4.2. Implémentation des modèles

4.2.1. Développement du modèle sur l'ensemble d'apprentissage

Ci-dessous les algorithmes utilisés sur Python:

Régression Logistique	Arbres de décision
LogisticRegression	**DecisionTreeClassifier** *Critère : Entropy* *Elagage : profondeur max_detph = 5*

4.2.2. Performance du modèle sur l'ensemble d'apprentissage (matrice de confusion)

- Algorithme **LogisticRegression**

		Valeurs prédites	
		0	1
Valeurs observées	0	1923	1537
	1	1656	3568

Précision : **63.2%**

- Algorithme **DecisionTreeClassifier**

		Valeurs prédites	
		0	1
Valeurs observées	0	3277	183
	1	2551	2673

Précision : **68.5%**

5. Présentation des résultats (validation du modèle)

5.1. Performance sur l'ensemble de TEST

- Algorithme **LogisticRegression**

		Valeurs prédites	
		0	1
Valeurs observées	0	520	416
	1	349	887

Précision : **64.7%**

- Algorithme **DecisionTreeClassifier**

		Valeurs prédites	
		0	1
Valeurs observées	0	868	68
	1	594	642

Précision : **69.5%**

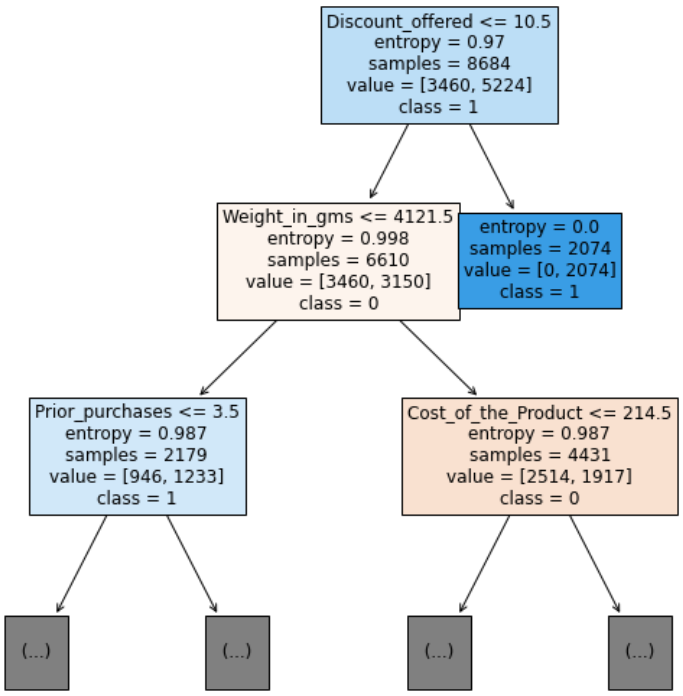
5.2. Interprétation des résultats

Les résultats sur l'ensemble de test donnent une seconde fois l'algorithme DecisionTreeClassifier gagnant avec une performance de 68.5% contre 63.2% pour l'algorithme LogicRegression. De plus,

l'algorithme DecisionTreeClassifier peut nous permettre de lire le comportement de chacune des variables d'entrée sur la variable de sortie et par conséquent aider à conseiller l'entreprise sur sa stratégie. Ce qui nous amène à choisir l'algorithme DecisionTreeClassifier.

En calculant les gains d'entropie de chaque variable sur la base de l'arbre de décision obtenu, nous avons pu classer ces variables selon leur importance. Les trois variables les plus importantes sont : « Discount offered », « Weight in gms » et « Cost of the product »

Arbre de decision:



L'exploitation de cet arbre nous permet de comprendre que les variables Discount_offered et purchases inférieures ou égales au seuil 10.5 et 3.5 respectivement, consolident les chances du dépassement du délai de livraison du colis. Tandis que lorsque les variables Weight (poids) et Cost_of_product sont inférieures à 4121 grs, il y a plus de chance que le colis arrive dans les délais.

6. Conclusion

Nous sommes rendus au terme de cet exercice pratique de valorisation de données. Nous avions pour objectif de développer un modèle d'analyse de données permettant de prédire sur la base de certaines variables qu'une livraison sera faite dans les délais ou non. Mais également de découvrir les liens présents entre les variables de prédiction et la variable de réponse.

Après application de l'algorithme basé sur les arbres de décision, nous avons obtenus des résultats satisfaisants. L'algorithme affiche des performances de l'ordre de 68.5% sur les données d'apprentissage, et de 69.5% sur les données de test (très peu de sous-apprentissage). Au regard des interprétations plus haut, nous pouvons conclure que l'étude a apporté des réponses au problème posé.