



DataScientest • com

PROJET PYTETRE

Promotion Mai 2021 Continue / Data Analyst

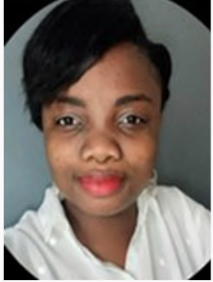




INTRODUCTION

Ce projet est intégré à notre formation de Data Analyst (en mode Formation Continue).
Mené à quatre de Juin 2021 à Décembre 2021 avec comme accompagnateur Jérémy.

NOTRE TEAM ET LE SUJET



Corine

Thierry

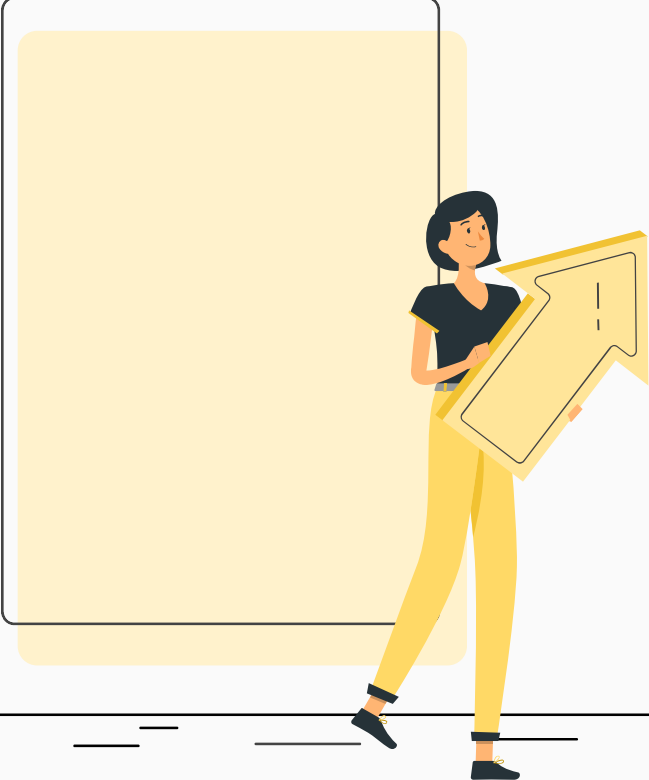


Vivien

Xiyuan



ANALYSE DE LA LOGISTIQUE DE LIVRAISON E-COMMERCE



01
PRESENTATION
DU PROBLEME

02
LES DONNEES

03
MODELES DE
MACHINE LEARNING

04
CONCLUSIONS

01. PRESENTATION DU PROBLEME



PRESENTATION DU PROBLEME

- ❑ Une entreprise spécialisée dans le e-commerce de produits électroniques recherche des solutions afin de pouvoir livrer à temps ses clients
- ❑ Une équipe de Data Analyst a été sollicitée pour utiliser sa base de données clients/commandes et mobiliser les outils de Machine Learning afin de résoudre cette problématique
- ❑ Notre objectif final est de proposer des outils d'aide à la décision mais également de prédiction grâce à un algorithme qui pourra être déployé au sein de cette structure par la suite

02. LES DONNEES



LES DONNEES - PRESENTATION

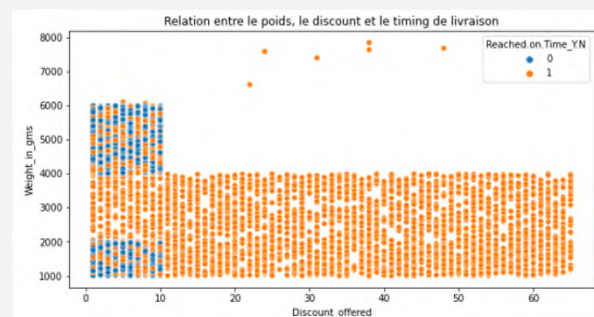
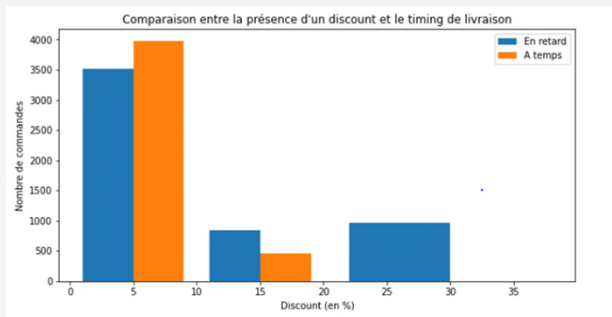
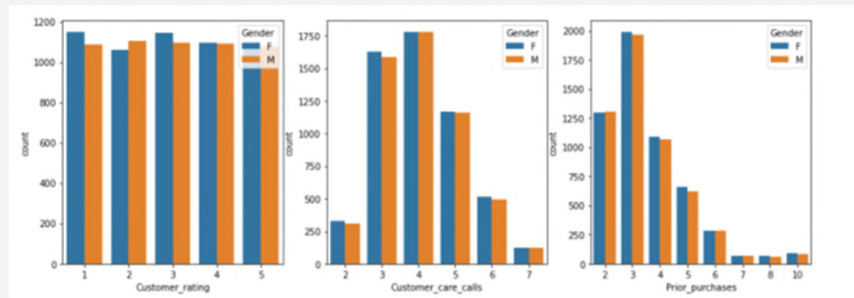
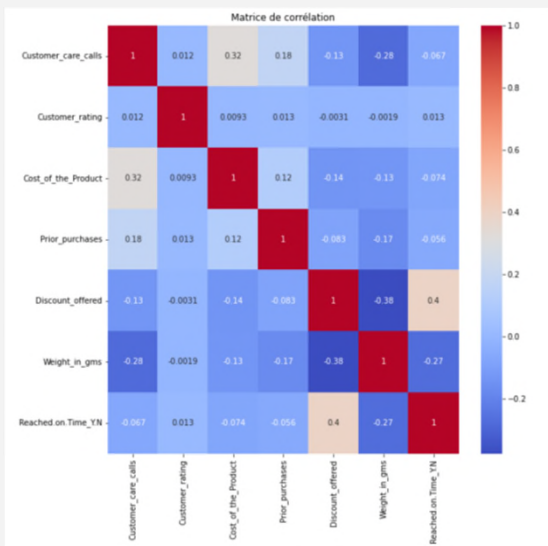
- ❑ Notre jeu de données mis à disposition pour ce projet comporte 12 variables et 10999 enregistrements dont voici un extrait :

ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
1	D	Flight	4	2	177	3	low	F	44	1233	1
2	F	Flight	4	5	216	2	low	M	59	3088	1
3	A	Flight	2	2	183	4	low	M	48	3374	1
4	B	Flight	3	3	176	4	medium	M	10	1177	1
5	C	Flight	2	2	184	3	medium	F	46	2484	1
6	F	Flight	3	1	162	3	medium	F	12	1417	1
7	D	Flight	3	4	250	3	low	F	3	2371	1
8	F	Flight	4	1	233	2	low	F	48	2804	1
9	A	Flight	3	4	150	3	low	F	11	1861	1
10	B	Flight	3	2	164	3	medium	F	29	1187	1

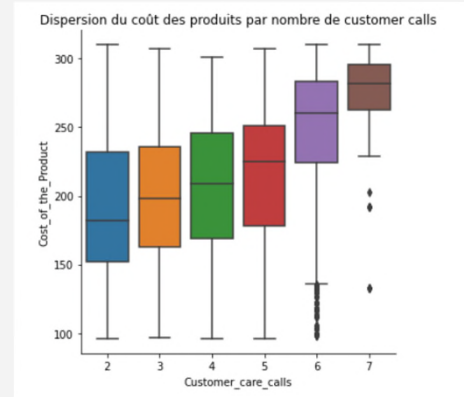
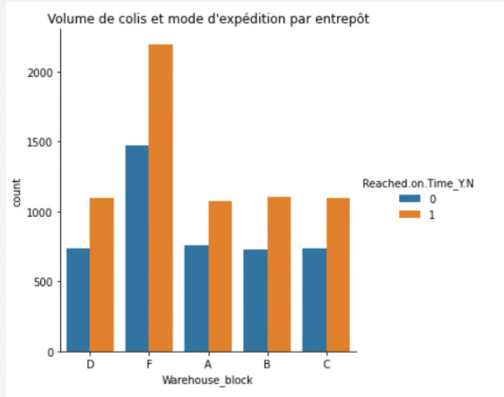
- ❑ La variable cible est Reached.on.Time Y.N prenant pour valeur 1 si la livraison a été effectuée en retard et 0 dans le cas contraire
- ❑ La première variable ID est inutile dans nos analyses et sera supprimée
- ❑ Aucune donnée n'est manquante et/ou nécessite transformation
- ❑ Aucune indication d'autres sources de données disponibles pour analyse complémentaire

LES DONNEES – EXPLORATION I/2

- Durant la première phase du projet nous avons réalisé un grand nombre de visualisations dont nous n'avons conservé que les 6 les plus marquantes, présentées à la fois dans notre rapport technique et dans cette présentation



LES DONNEES – EXPLORATION 2/2



- Ces visualisations ont permis d'identifier que les 3 variables à priori les plus impactantes sur notre variable cible sont :
- le coût du produit
 - le discount offert
 - le poids de l'expédition

03. MODELES DE MACHINE LEARNING



MODELES ML- CLASSIFICATION DU PROBLEME

- ❑ Notre projet s'apparente à un problème de classification (classification nominale binaire).
- ❑ La variable cible (Reached.on.Time Y.N) est qualitative.
- ❑ Le modèle de ML nous aide à prédire à quelle classe chaque commande appartient.

MODELES ML- MISE EN OEUVRE

- ❑ Avant d'appliquer les modèles de machine learning, nous avons suivi toutes les étapes de pré-processing.
- ❑ Nous avons préparé les données en utilisant les pipelines afin d'optimiser le code et de minimiser les bugs.
- ❑ Nous avons utilisé 5 modèles différents : Régression Logistique, K-Nearest Neighbors, Gradient Boosting, Decision Tree et Random Forest.

MODELES ML -DEFINITIONS

01 REGRESSION LOGISTIQUE

La régression logistique est une méthode qui consiste à prédire une valeur de données d'après les observations réelles d'un jeu de données.

02 K-NEAREST NEIGHBORS

La méthode KNN a pour but de classifier des points cibles en fonction de leurs distances par rapport à des points constituant un échantillon d'apprentissage.

03 & 04 DECISION TREE & RANDOM FOREST

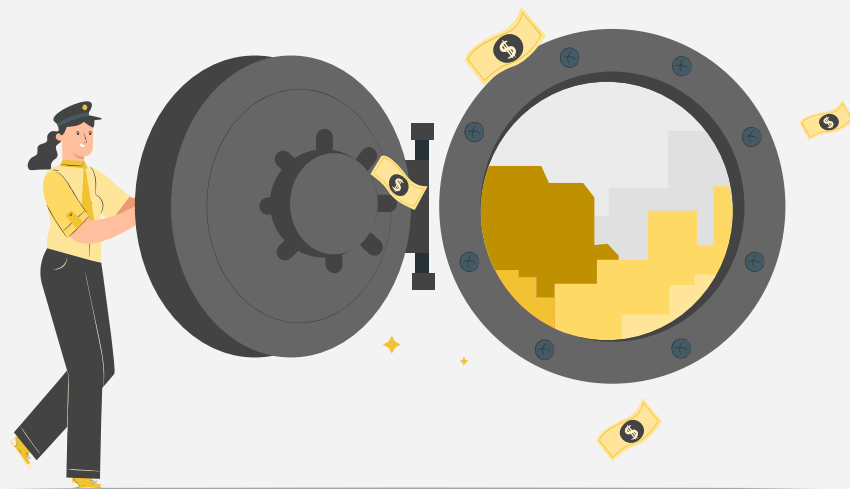
Des méthodes permettant de construire des arbres de décision qui serviront de modèle de classification / prédiction.

05 GRADIENT BOOSTING

Gradient Boosting est une méthode s'appuyant sur des arbres de décision, elle permet de transformer les "apprenants faibles" en "apprenants forts". Chaque arbre s'adapte à la version modifiée du premier ensemble de données.

DEMO MACHINE LEARNING STREAMLIT

04. CONCLUSIONS



CONCLUSIONS – REGARD CRITIQUE

Durant ce projet nous nous sommes rendus compte qu'un certain nombre de données supplémentaires auraient pu nous aider à compléter, affiner nos analyses et modèles.

Par exemple :

- la date de la commande aurait pu nous aider à identifier une éventuelle saisonnalité à la fois de la disponibilité du matériel commandé que des difficultés d'expédition
- La localisation géographique à la fois des entrepôts de stockage que des clients pour valider des hypothèses liées à la distance entre point de stockage et livraison (y compris la notion de continents)

Ces données n'étaient malheureusement ni disponibles dans le jeu ni « inventables ».



CONCLUSIONS

1

Avec peu de données, il est parfois difficile d'être conclusifs. Avoir la possibilité d'accéder à des données complémentaires nous aurait aidé

2

Le meilleur modèle mis en oeuvre (Gradient Boosting) amène à un résultat de maximum 70% de prédictabilité

3

Merci à Jérémie de son support durant ce projet. Guidage, méthodologie et surtout disponibilité ont été les maîtres mots

MERCI DE VOTRE ATTENTION

Session de Questions/Réponses

