PROJET 5

Segmentez des clients d'un site e-commerce

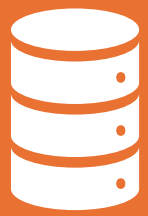
SOMMAIRE

PARTIE 1 – REQUÊTES SQL

PARTIE 2 – ANALYSE EXPLORATOIRE

PARTIE 3 – CONSTRUCTION DU MODÈLE DE CLASSIFICATION

PARTIE 4 – MAINTENANCE DU MODÈLE



Partie 1

SQL

Requêtes SQL

En excluant les commandes annulées, quelles sont les commandes récentes de moins de 3 mois que les clients ont reçues avec au moins 3 jours de retard ?

```sql
WITH
    cte_latest AS(
    SELECT  MAX(order_purchase_timestamp) AS latest_order
    FROM orders
    ),
    cte_age AS(
    SELECT * ,
        ROUND(JULIANDAY(latest_order) - JULIANDAY(order_purchase_timestamp)-1) AS order_age,
        ROUND(JULIANDAY(order_delivered_customer_date) - JULIANDAY(order_estimated_delivery_date)-1) AS delay
    FROM orders,cte_latest
    )


SELECT *
FROM   cte_age a
WHERE
    a.order_status <> "cancelled"
    AND(a.order_age <90)
    AND(a.delay >3);
```

		order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date	latest_order	order_age	delay
232	df8d4ab3	delivered	2018-08-05 14:39:32	2018-08-05 14:50:14	2018-08-15 05:41:00	2018-08-28 21:33:02	2018-08-14 00:00:00	2018-10-17 17:30:18	72.0	14.0
233	d2c9a00ee	delivered	2018-07-25 01:25:29	2018-07-25 01:35:12	2018-08-21 12:14:00	2018-08-22 17:51:41	2018-08-10 00:00:00	2018-10-17 17:30:18	84.0	12.0
234	952dee75e	delivered	2018-08-07 14:07:40	2018-08-07 14:15:21	2018-08-20 12:25:00	2018-08-21 18:07:43	2018-08-14 00:00:00	2018-10-17 17:30:18	70.0	7.0
235	7320036	delivered	2018-08-02 12:06:47	2018-08-02 13:05:56	2018-08-10 14:42:00	2018-08-13 15:50:48	2018-08-08 00:00:00	2018-10-17 17:30:18	75.0	5.0
236	8275807b	delivered	2018-08-10 11:46:09	2018-08-11 02:50:25	2018-08-14 10:09:00	2018-09-03 09:32:31	2018-08-28 00:00:00	2018-10-17 17:30:18	67.0	5.0
237	e4b45f0	delivered	2018-08-14 23:29:21	2018-08-16 03:05:11	2018-08-16 13:28:00	2018-08-28 18:02:52	2018-08-24 00:00:00	2018-10-17 17:30:18	63.0	4.0
238	dba77d8	delivered	2018-07-19 08:37:26	2018-07-21 03:25:17	2018-07-23 15:31:00	2018-08-21 01:12:45	2018-08-10 00:00:00	2018-10-17 17:30:18	89.0	10.0
239	130fe9b5	delivered	2018-08-02 22:46:54	2018-08-02 23:04:06	2018-08-15 17:42:00	2018-08-21 00:03:26	2018-08-16 00:00:00	2018-10-17 17:30:18	75.0	4.0

Requêtes SQL

Qui sont les vendeurs ayant généré un chiffre d'affaires de plus de 100 000 Real sur des commandes livrées via Olist ?

```sql
WITH
    cte_revenue AS (
        SELECT
            seller_id,
            ROUND (SUM(price)) AS seller_revenue
        FROM
            order_items
        GROUP BY
            seller_id
)

SELECT
    seller_id,
    seller_revenue
FROM cte_revenue

WHERE
    seller_revenue > 100000
ORDER by
    seller_Revenue DESC;
```

	seller_id	seller_revenue
1	4869f7a5dfa277a7dca6462dcf3b52b2	229473.0
2	53243585a1d6dc2643021fd1853d8905	222776.0
3	4a3ca9315b744ce9f8e9374361493884	200473.0
4	fa1c13f2614d7b5c4749cbc52fecda94	194042.0
5	7c67e1448b00f6e969d365cea6b010ab	187924.0
6	7e93a43ef30c4f03f38b393420bc753a	176432.0
7	da8622b14eb17ae2831f4ac5b9dab84a	160237.0
8	7a67c85e85bb2ce8582c35f2203ad736	141746.0
9	1025f0e2d44d7041d6cf58b6550e0bfa	138969.0
10	955fee9216a65b617aa5c0531780ce60	135172.0
11	46dc3b2cc0980fb8ec44634e21d2718e	128111.0
12	6560211a19b47992c3666cc44a7e94c0	123305.0
13	620c87c171fb2a6dd6e8bb4dec959fc6	114774.0
14	7d13fca15225358621be4086e1eb0964	113629.0
15	5dceca129747e92ff8ef7a997dc4f8ca	112156.0
16	1f50f920176fa81dab994f9023523100	106939.0
17	cc419e0650a3c5ba77189a1882b7556a	104288.0
18	a1043bafd471dff536d0c462352beb48	101901.0

Requêtes SQL

Qui sont les nouveaux vendeurs (moins de 3 mois d'ancienneté) qui sont déjà très engagés avec la plateforme (ayant déjà vendu plus de 30 produits) ?

```sql
SELECT *
FROM (
    SELECT
        seller_id,
        COUNT(product_id) AS item_count,
        MIN(order_purchase_timestamp) AS min_date_order

    FROM
        order_items oi

    JOIN orders o
        ON o.order_id = oi.order_id

    GROUP BY
        seller_id
    ORDER BY
        item_count desc
)

WHERE
    min_date_order < (SELECT DATE(MAX(order_purchase_timestamp),'-3 month') FROM orders)
    AND (item_count >30);
```

	seller_id	item_count	min_date_order
1	6560211a19b47992c3666cc44a7e94c0	2033	2017-02-17 07:39:19
2	4a3ca9315b744ce9f8e9374361493884	1987	2017-01-08 09:35:14
3	1f50f920176fa81dab994f9023523100	1931	2017-04-03 22:00:31
4	cc419e0650a3c5ba77189a1882b7556a	1775	2017-01-31 17:15:33
5	da8622b14eb17ae2831f4ac5b9dab84a	1551	2017-02-05 21:46:05
6	955fee9216a65b617aa5c0531780ce60	1499	2017-07-24 11:33:53
7	1025f0e2d44d7041d6cf58b6550e0bfa	1428	2017-07-09 11:15:16
8	7c67e1448b00f6e969d365cea6b010ab	1364	2017-01-26 22:44:11
9	ea8482cd71df3c1969d7b9473ff13abc	1203	2017-08-15 12:54:48
10	7a67c85e85bb2ce8582c35f2203ad736	1171	2017-01-27 12:15:07
11	4869f7a5dfa277a7dca6462dcf3b52b2	1156	2017-03-07 12:43:03
12	3d871de0142ce09b7081e2b9d1733cb1	1147	2017-03-05 12:29:19
13	8b321bb669392f5163d04c59e235e066	1018	2017-10-27 15:34:03
14	cca3071e3e9bb7d12640c9fbe2301306	830	2016-10-03 22:06:03
15	620c87c171fb2a6dd6e8bb4dec959fc6	798	2016-10-04 13:15:46
16	a1043bafd471dff536d0c462352beb48	770	2017-02-14 10:44:33
17	e9779976487b77c6d4ac45f75ec7afe9	750	2017-03-01 09:25:31
18	f8db351d8c4c4c22c6835c19a46f01b0	724	2017-01-25 22:46:24
19	d2374cbcbb3ca4ab1086534108cc3ab7	631	2017-02-10 15:36:37
20	391fc6631aebcf3004804e51b40bcf1e	613	2016-10-06 00:06:17
21	fa1c13f2614d7b5c4749cbc52fecda94	586	2017-01-07 20:45:31

Requêtes SQL

Quels sont les 5 codes postaux, enregistrant plus de 30 commandes, avec le pire review score moyen sur les 12 derniers mois ?

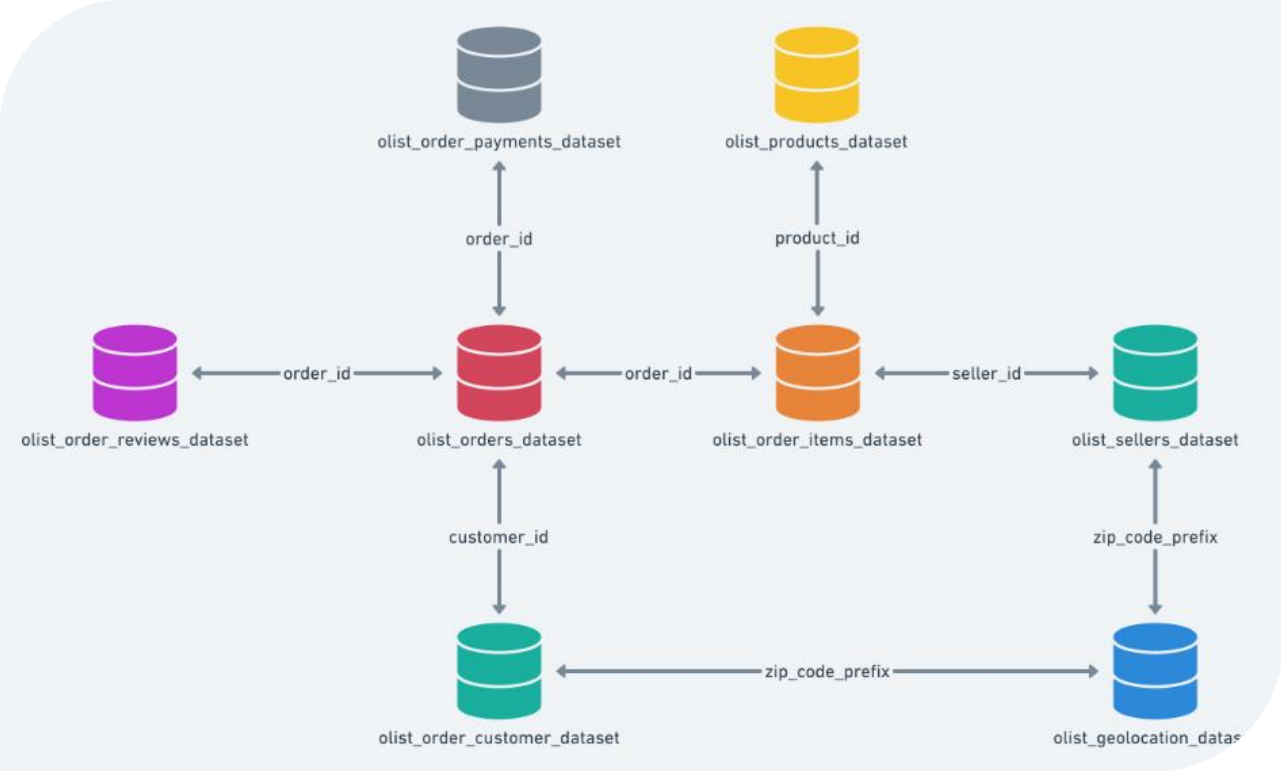
```sql
WITH review AS (
        SELECT
            s.seller_zip_code_prefix,
            COUNT(DISTINCT oi.order_id) AS order_count,
            ROUND(AVG(ro.review_score)) AS avg_score,
            ro.review_creation_date
        FROM
            sellers s
        LEFT JOIN
            order_items oi   ON s.seller_id = oi.seller_id
        LEFT JOIN
            order_reviews ro ON ro.order_id = oi.order_id


        GROUP BY
            s.seller_zip_code_prefix
)


SELECT *
FROM
    review
WHERE
    order_count > 30
    AND review_creation_date BETWEEN
        (SELECT DATE(MAX(review_creation_date), '-12 month') FROM order_reviews)
        AND
        (SELECT MAX(review_creation_date) FROM order_reviews)
ORDER BY
    avg_score LIMIT 5
```

	seller_zip_code_prefix	order_count	avg_score	review_creation_date
1	6506	115	2.0	2017-09-19 00:00:00
2	1512	69	3.0	2017-11-02 00:00:00
3	3017	98	3.0	2017-10-18 00:00:00
4	3273	45	3.0	2018-05-12 00:00:00
5	3476	36	3.0	2017-10-25 00:00:00

Partie 2

ANALYSE EXPLORATOIRE

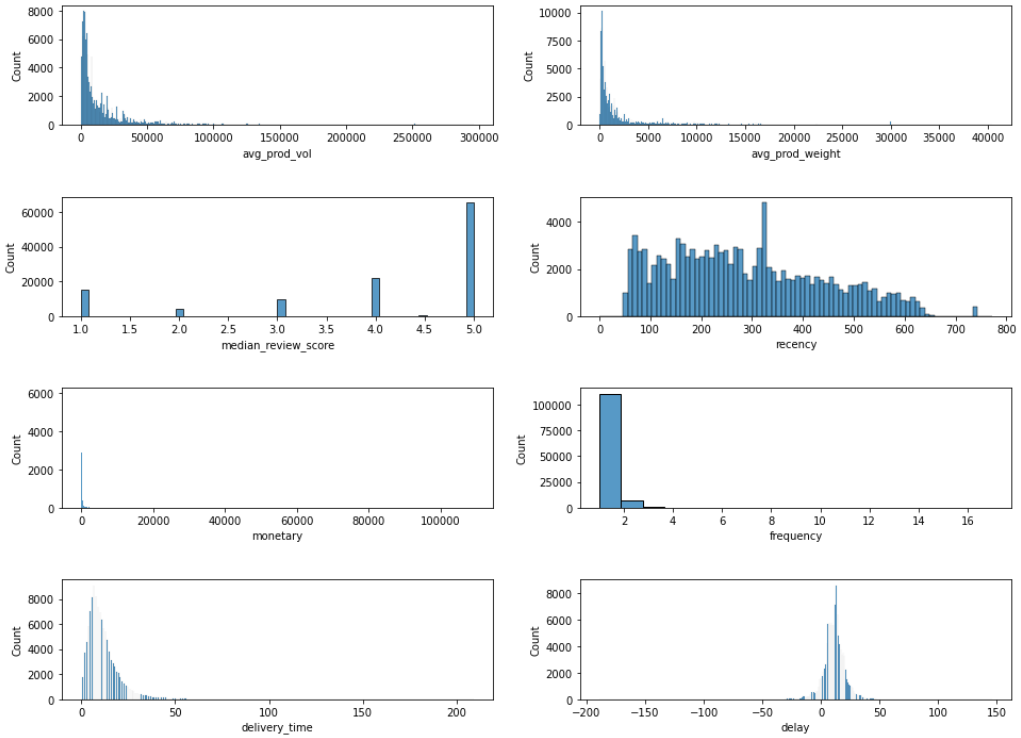
Analyse exploratoire



Nouvelles variables calculées

- recency
- frequency
- monetary
- delay
- avg_prod_vol
- med_rev_score

Analyse exploratoire



Comme annoncé, la part de clients ayant passé plusieurs commande est assez faible.

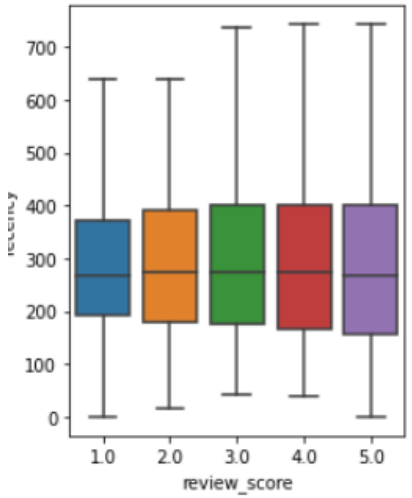
L'ancienneté des commandes est répartie uniformément avec un pic d'activité notable environ 1 an avant l'édition des données.

La distribution des retards de livraison semble suivre une loi normale avec un retard médian aux alentours de 10 jours.

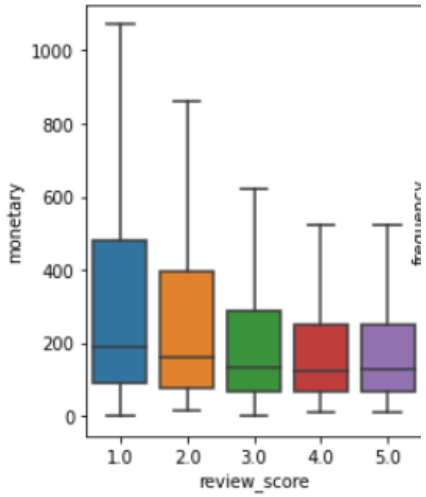
Analyse exploratoire

L’analyse bivariée vis-à-vis des notations clients nous montre clairement que le délai de livraison et le retard sont impactant dans l’évaluation. Les fortes dépenses sont plus présentes dans les mauvaises notes.

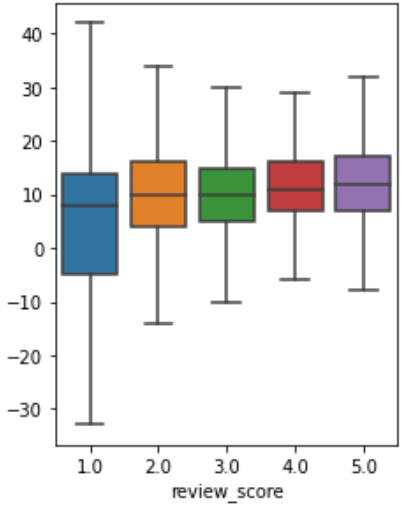
Recency



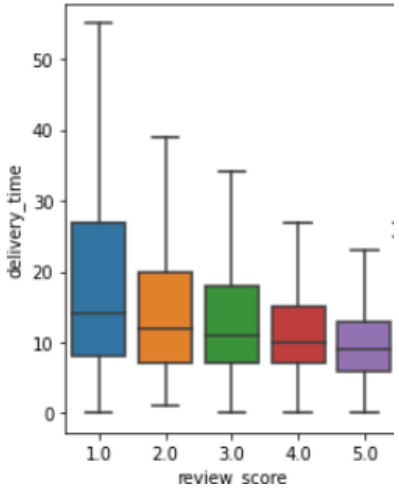
Monetary



Delivery time



Delay



Analyse exploratoire

Normalisation des données

customer_unique_id	avg_prod_vol	avg_prod_weight	median_review_score	recency	monetary	frequency	delivery_time	delay
861eff4711a542e4b93843c6dd7febb0	107136.0	8683.0	4.0	519	146.87	1	8.0	-11.0
290c77bc529b7ac935b93aa66c333dc3	53400.0	10150.						
060e732b5b29e8181a18229c7b0b2b5e	45968.0	8267.						
259dac757896d24d7702b9acbbff3f3c	79968.0	12160.						
345ecd01c38d18a9036ed96c73b8d066	23625.0	5200.						

customer_unique_id	avg_prod_vol	median_review_score	recency	monetary	frequency	delay
861eff4711a542e4b93843c6dd7febb0	3.956612	-0.068053	1.507536	-0.104498	-0.163518	0.085023
290c77bc529b7ac935b93aa66c333dc3	1.644373	0.677346	-0.069863	0.187999	-0.163518	0.385157
060e732b5b29e8181a18229c7b0b2b5e	1.324577	0.677346	-0.891153	-0.087657	-0.163518	1.285561
259dac757896d24d7702b9acbbff3f3c	2.787584	0.677346	-0.454435	-0.063511	-0.163518	-0.115067
345ecd01c38d18a9036ed96c73b8d066	0.363166	0.677346	-1.353944	0.058925	-0.163518	0.585247

Analyse exploratoire

Encodage des variables catégorielles

customer_unique_id	product_category_name
861eff4711a542e4b93843c6dd7febb0	moveis_escritorio
290c77bc529b7ac935b93aa66c333dc3	utilidades_domesticas
060e732b5b29e8181a18229c7b0b2b5e	moveis_escritorio
259dac757896d24d7702b9acbbff3f3c	moveis_escritorio
345ecd01c38d18a9036ed96c73b8d066	casa_conforto

customer_unique_id	moveis_escritorio	utilidades_domesticas	moveis_escritorio	moveis_escritorio	casa_conforto
861eff4711a542e4b93843c6dd7febb0	0	**0**	**0**	**0**	**0**
290c77bc529b7ac935b93aa66c333dc3	1	**0**	**1**	**0**	**0**
060e732b5b29e8181a18229c7b0b2b5e	0	**1**	**0**	**0**	**0**
259dac757896d24d7702b9acbbff3f3c	0	**1**	**1**	**0**	**0**
345ecd01c38d18a9036ed96c73b8d066	0	**0**	**0**	**0**	**0**

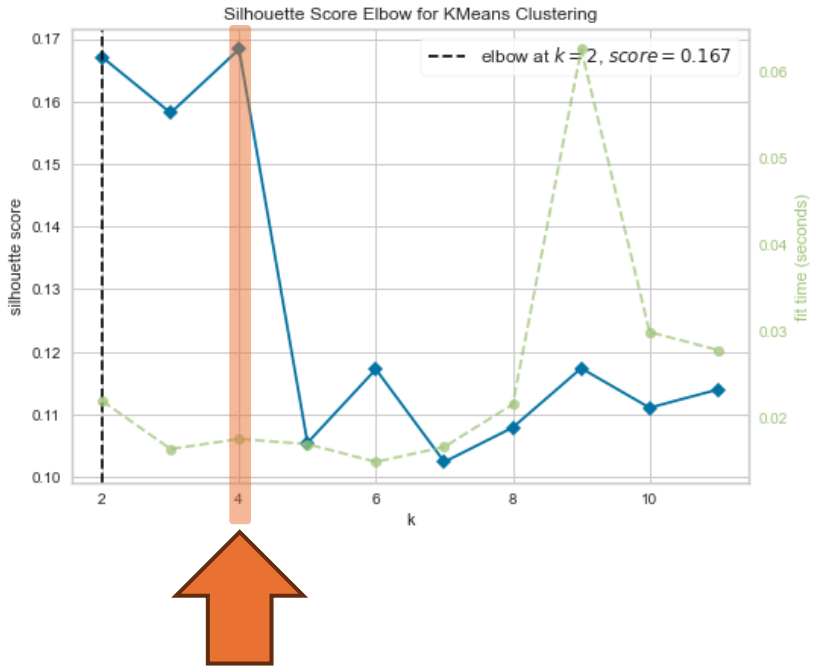
Partie 3

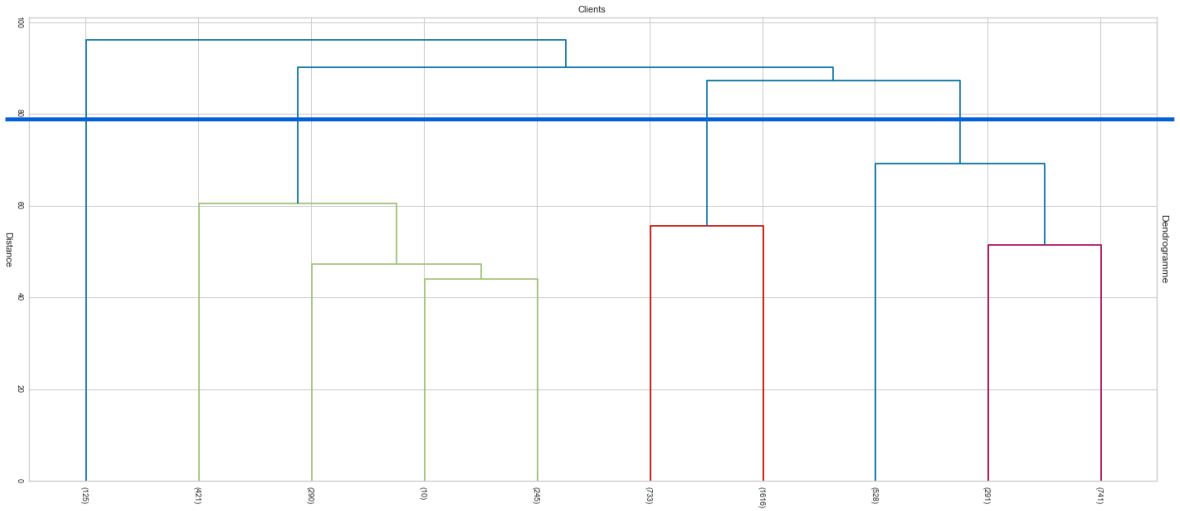
CONSTRUCTION DU MODELE

Construction du modèle

Nous avons mis 2 modèles en concurrence :

- K-Means
- Classification ascendante hiérarchique





Construction du modèle

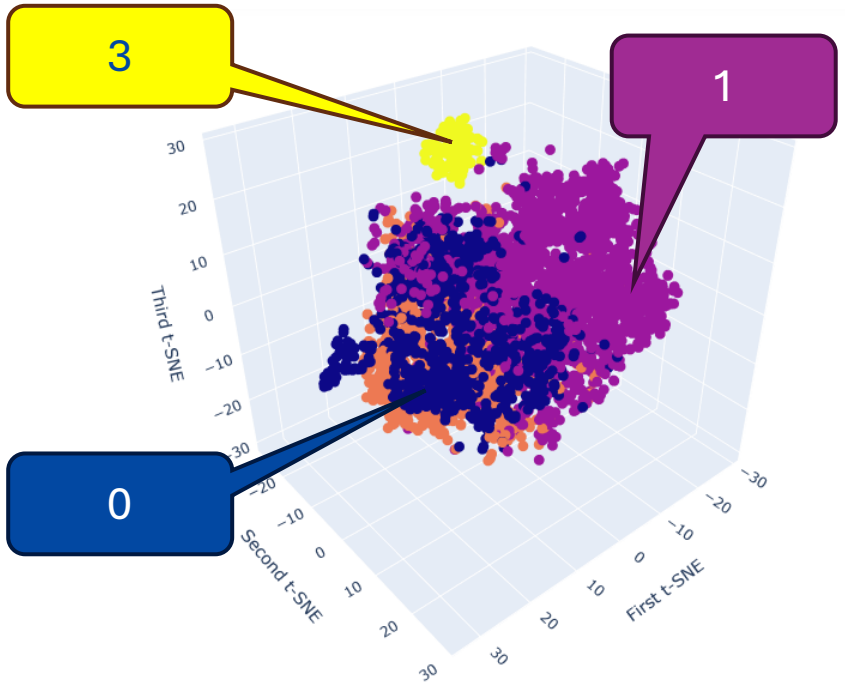
Evaluations des modèles

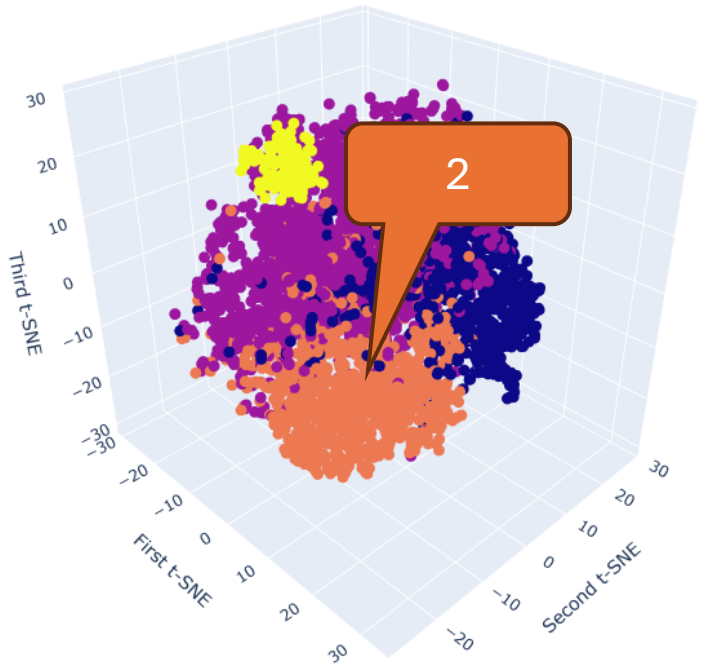
	KMeans	CAH	Kprototype
Silhouette	0.168859	0.115833	0.16804

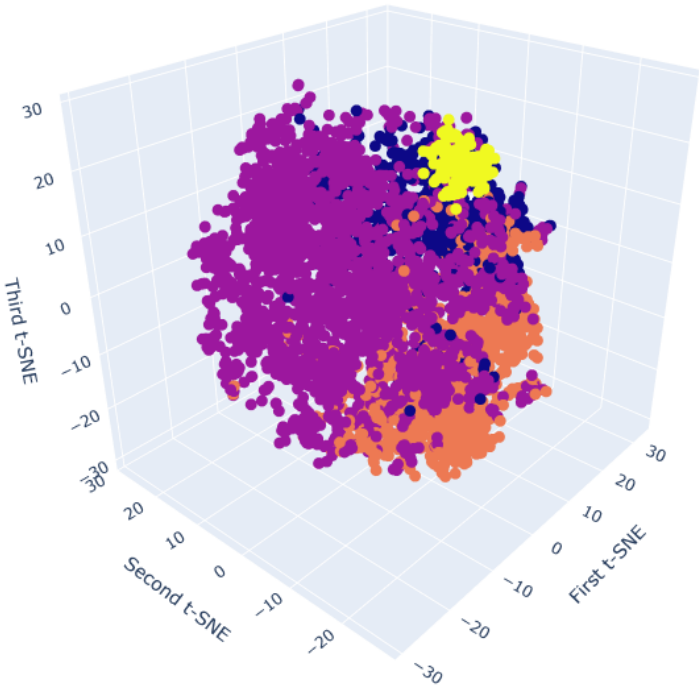
On constate que le Kmeans (ou Kmeans assimilé avec K-Prototype) retourne le meilleur score

Construction du modèle

Nous projetons nos 51 variables retenues dans un espace 3D grâce au **T-SNE**

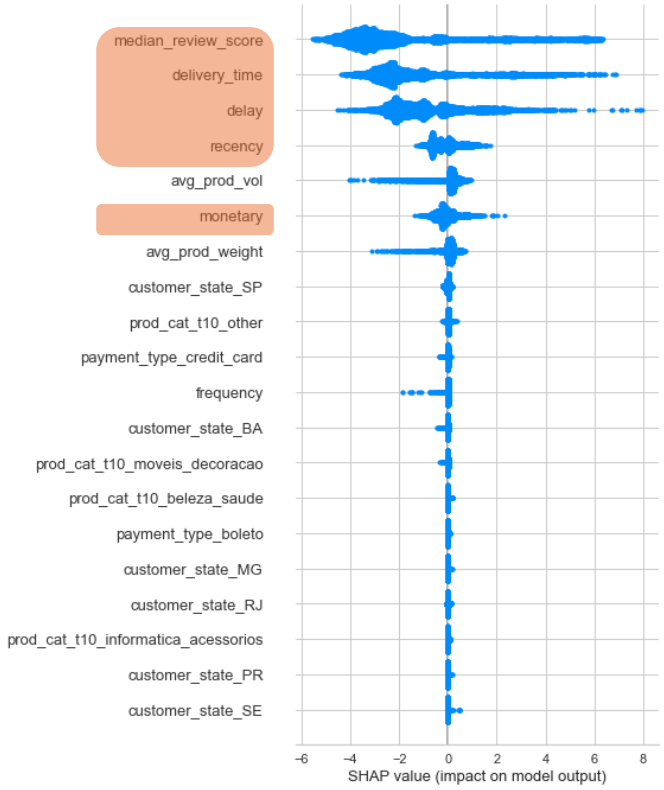


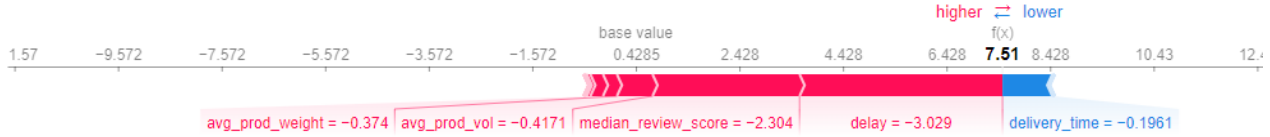




Construction du modèle

cat 0



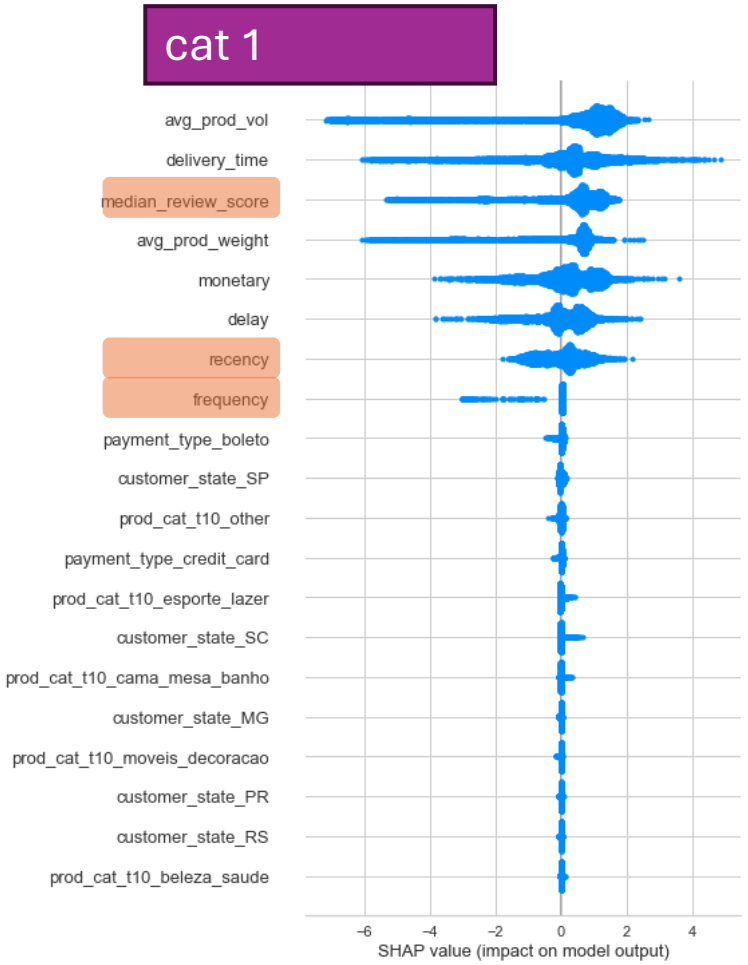


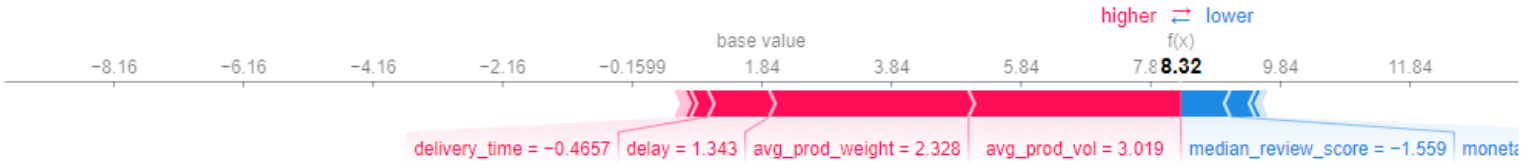
Groupe des clients **déçus** avec:

Notation négative

Délais et retards de livraison

Construction du modèle





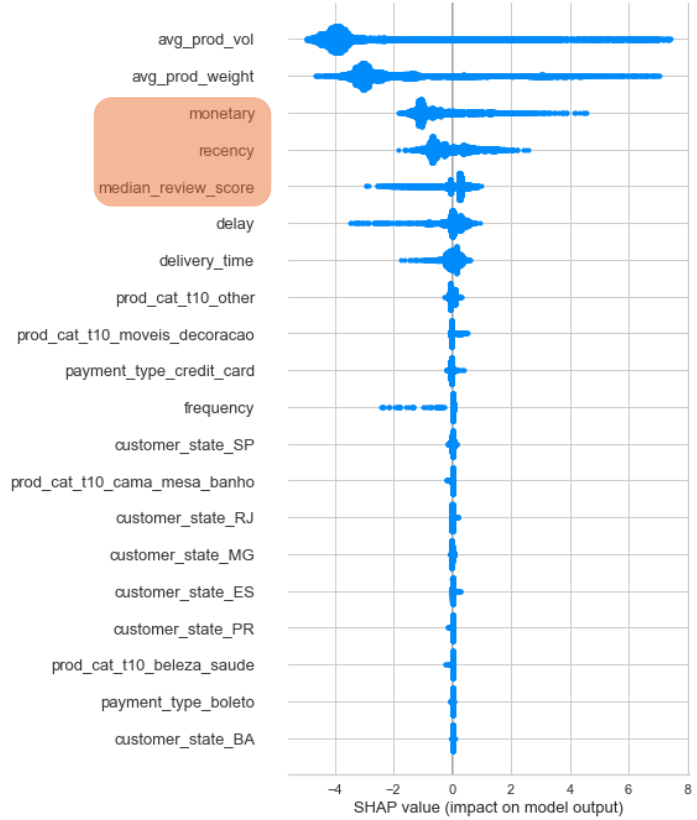
Groupe des **bons** clients avec:

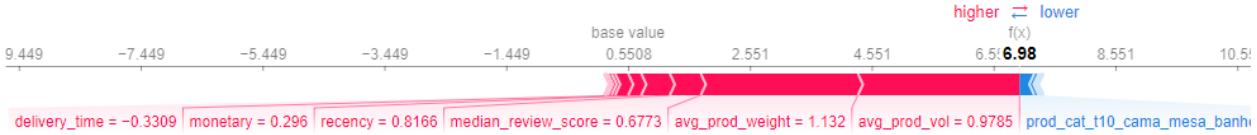
Produits volumineux

Dépenses significatives

Construction du modèle

cat 2



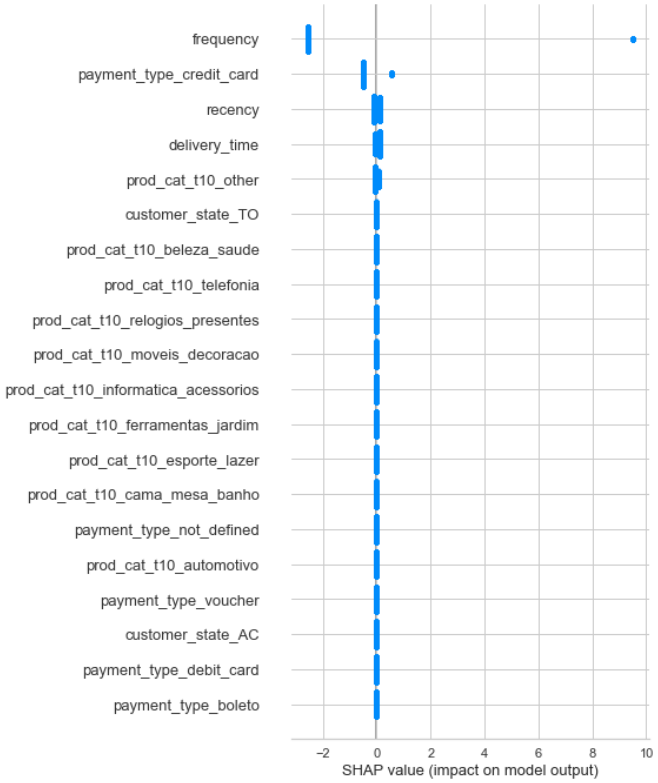


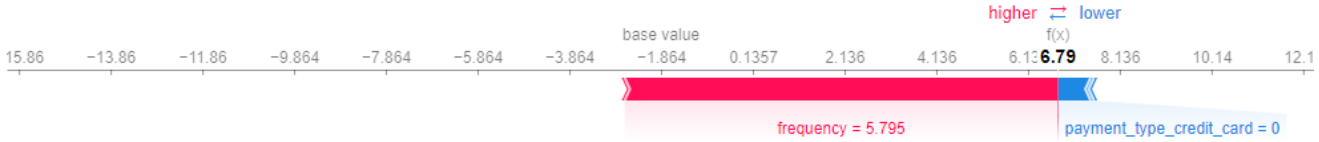
Clients à fidéliser :

Achats récents
Faibles dépenses
Faibles fréquences
Scores positifs

Construction du modèle

cat 3





Clients occasionnels / nouveaux :

Partie 4

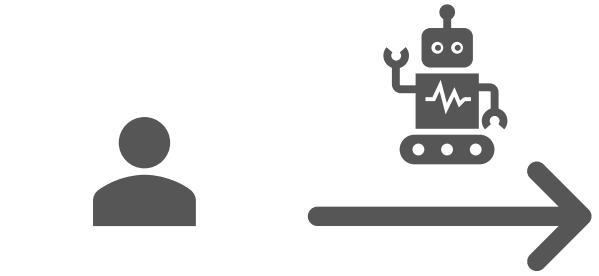
MAINTENANCE DU MODELE

Maintenance du modèle

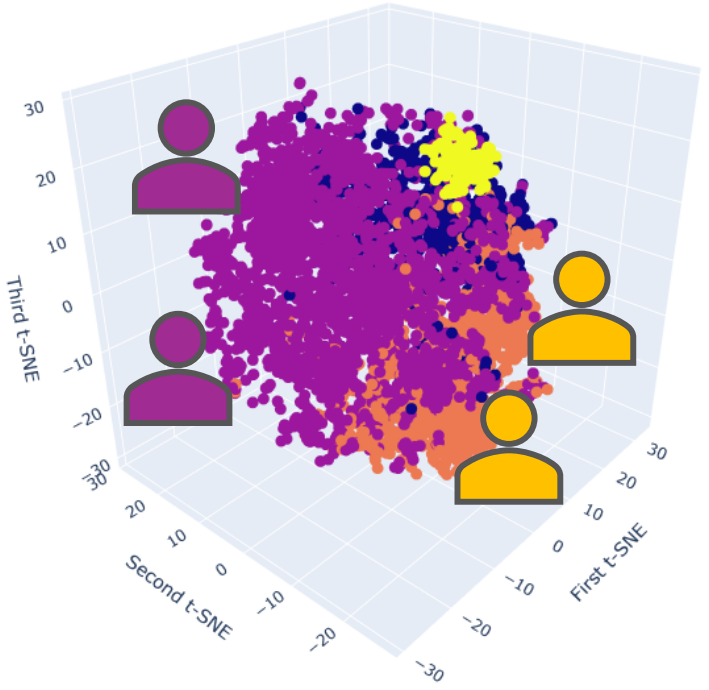
Nous avons entraîné notre modèle sur une période de 18 mois.

Nous avons ensuite simulé l'ajout de nouveaux clients

kmeans_ref.predict(X)



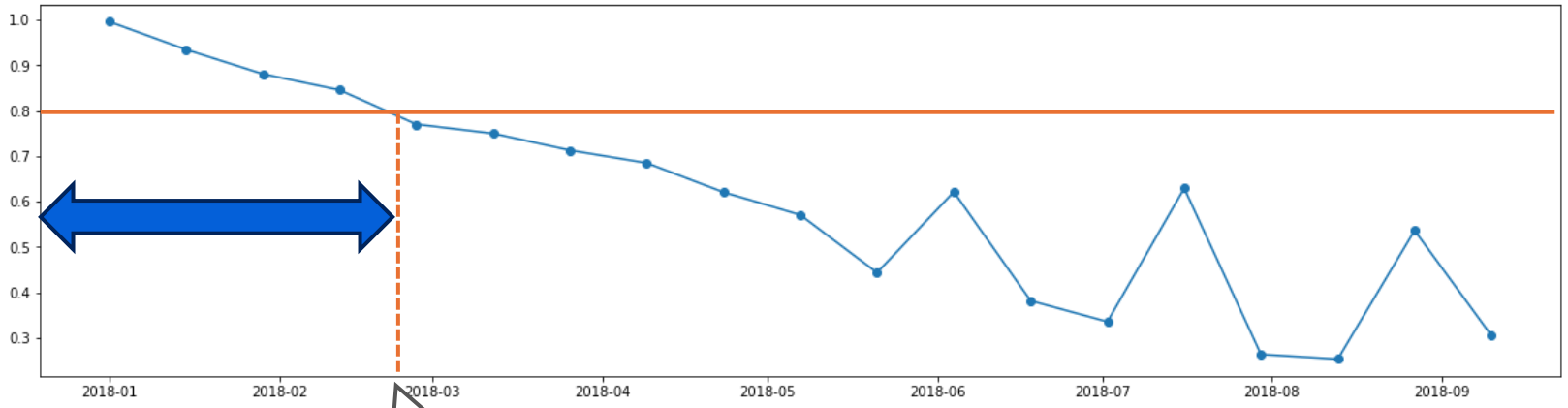
Nouveaux clients



Maintenance du modèle

Les nouveaux clients ajoutés, nous avons évalué notre modèle à une fréquence de 2 semaines.

Il en ressort qu'un réentraînement de notre algorithme sera nécessaire à raison d'une fois toutes les **6 semaines.**



Fréquence de maintenance recommandée = 6 semaines

CONCLUSION

Nous avons élaboré notre modèle sur la base d'attributs existants et d'autres calculés.

Les attributs du type catégorie (encodés) n'ont pas apporté d'informations significatives face aux variables RFM notamment. Nous pouvons donc nous limiter au plus à 8-10 variables numériques.

En nous limitant à 4 classes (voire 3), nous pouvons définir des actions adaptées à chaque profil de client.