Projet de Scoring : Identification des clients fragiles

Sabine Lohier
sb.lohier@gmail.com

DATAGONG

Objectifs :

- Sélectionner les variables pertinentes en lien avec la fragilité des clients
- Bâtir une modélisation prédictive permettant :
  - d'attribuer un score de churn aux clients
  - de gagner en compréhension quant aux profils des clients à cibler
  - de les prioriser pour les différents campagnes marketing

- **Métrique utilisée :** Area Under the Curve (AUC)
- **2 jeux de données contenant** des indicateurs socio-démographiques, les différents services auxquels ils ont souscrit, les données relatives à leur fidélité et leur facturation
  - données *train* **labellisées**
  - données *eval* **non labellisées**

Données labellisées et y_i discrets ⟶ **Modèle de classification binaire - Apprentissage supervisé**

- **Données *train* :**
  - ➢ 5158 observations
  - ➢ Pas de doublon, pas de valeur manquante
  - ➢ Séparation en données d'entraînement (80%) et en données de test (20%)

- **Données *eval* :**
  - ➢ 1885 observations
  - ➢ Pas de doublon, pas de valeur manquante

- ★ 17 variables booléennes
- ★ 5 variables catégorielles
- ★ 12 variables numériques

Analyse Exploratoire // **Présentation des variables**

Socio / Démographie

- `genre` `age` `under_30` `senior_citizen`
- `married` `dependents` `number_of_dependents`

Contrat / Mode de paiement

- Type d'abonnement `contract` *['Month-to-Month' 'One Year' 'Two Year']*
- Offre marketing `offer` *['None' 'Offer A' 'Offer B' 'Offer C' 'Offer D' 'Offer E']*
- Mode de paiement et support de facturation `payment_method` `paperless_billing`

Options / Services suppl.

- `device_protection_plan` `phone_service` `multiple_lines` `internet_service` `online_security` `online_backup` `premium_tech_support` `unlimited_data`
- Le type de service Internet fixe : `Internet_type` *[DSL None Fiber Optic Cable]*

Charges/ Totaux.

- `monthly_charge`
- `avg_monthly_long_distance_charges` `total_long_distance_charges`
- `total_charges` `total_extra_data_charges` `total_refunds` `total_revenue`

Flux

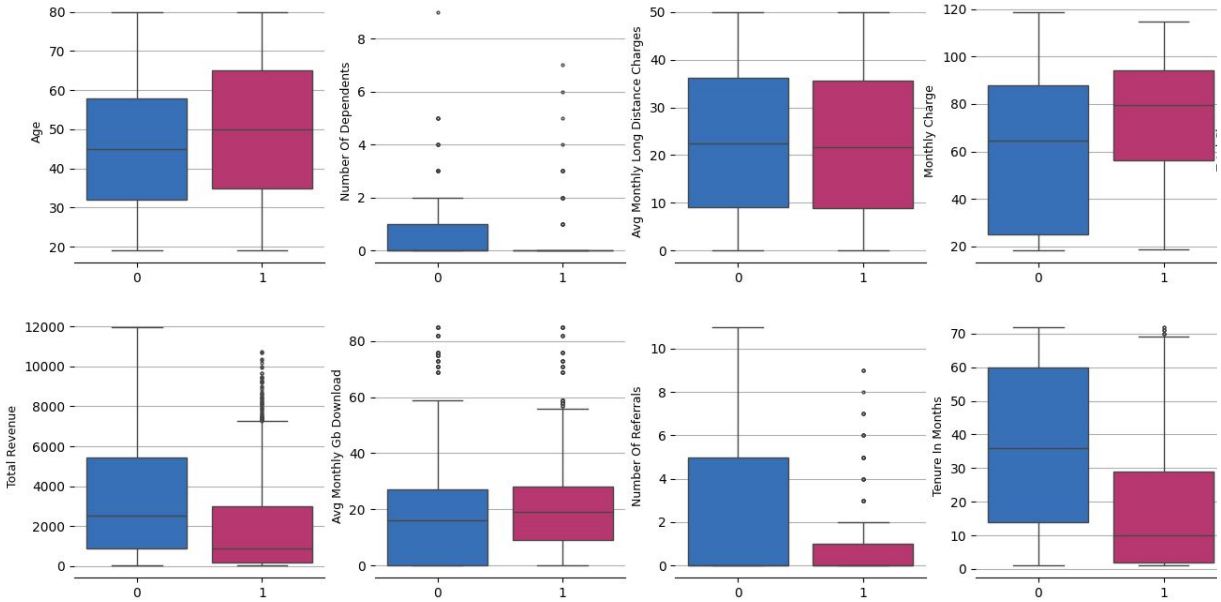
- Volume de data téléchargées `avg_monthly_gb_download`
- `streaming_movies` `streaming_music` `streaming_tv`

Fidélité

- Ancienneté du client `tenure_in_months`
- Parrainage `referred_a_friend` `number_of_referrals`

Churn :

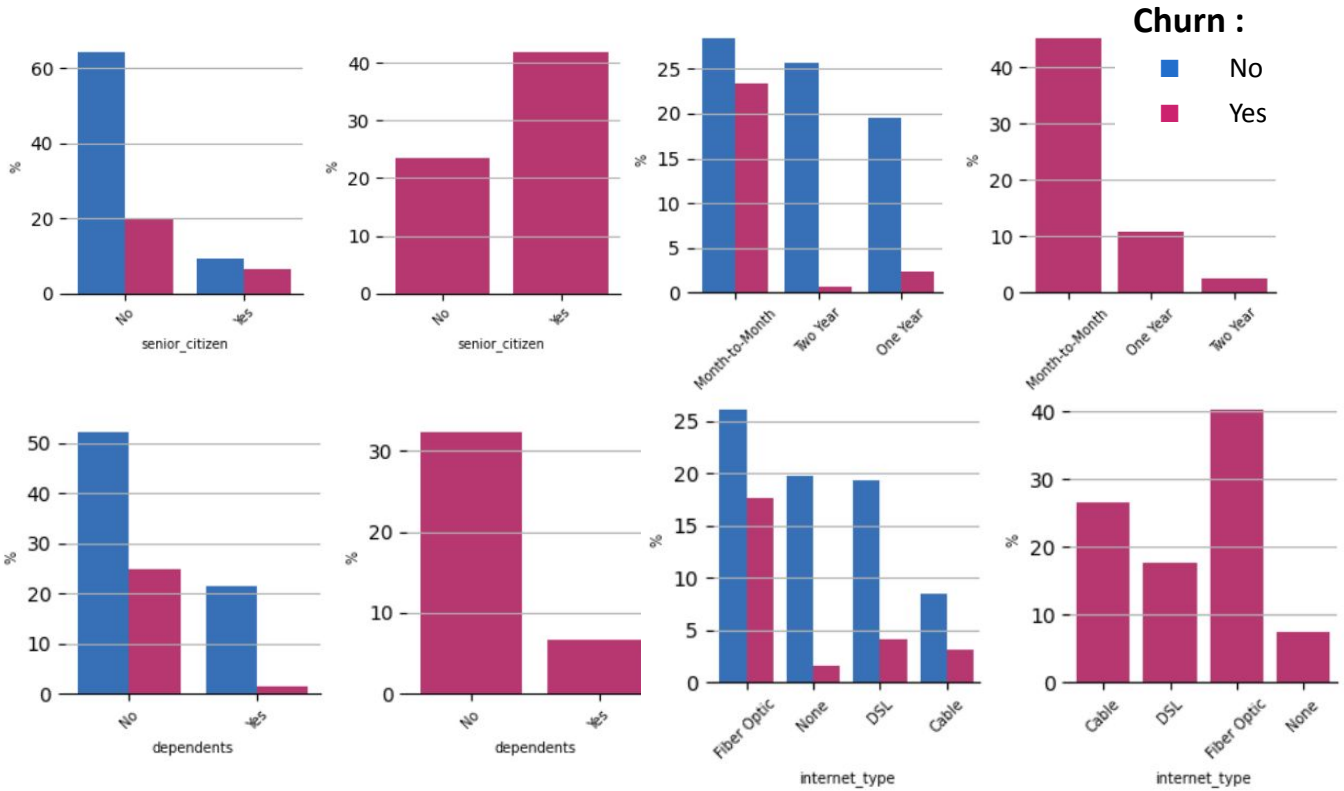
- No : 74 %
- Yes : 26 %



Distribution des clients en fonction des variables numériques, les clients fragiles sont représentés en rouge.

Tendances de profils des clients fragiles :

- ➢ Âge médian de 50 ans (contre 45 ans)
- ➢ Nouveaux clients : ancienneté de 10 mois en valeur médiane (contre 3 ans)
- ➢ Factures mensuelles plus élevées (80 euros contre 65 euros)



Répartition des clients fragiles en fonction de 4 variables catégorielles et/ou booléennes, les clients fragiles sont représentés en rouge. A droite les pourcentages sont normalisé par modalité.

Tendances de profils des clients fragiles :

- ➢ Ils ont souscrit à un abonnement mensuel
- ➢ Ils ont la fibre optique
- ➢ Sans enfant à charge
- ➢ Séniors (> 65 ans)

Data Management // **Discrétisation des variables numériques**

Socio / Démographie	■	`genre` `age` `under_30` `senior_citizen`
	■	`married` `dependents` `number_of_dependents`
Contrat / Mode de paiement	■	Type d'abonnement `contract`
	■	Offre marketing `offer`
	■	Mode de paiement et support de facturation `payment_meth`
Options / Services suppl.	■	`device_protection_plan` `phone_service` `multiple_li` `online_backup` `premium_tech_support` `unlimited_data`
	■	Le type de service Internet fixe : `Internet_type`
Charges/ Totaux.	■	`monthly_charge`
	■	`avg_monthly_long_distance_charges` `total_long_distance_charges`
	■	`total_charges` `total_extra_data_charges` `total_refunds` `total_revenue`
Flux	■	Volume de data téléchargées `avg_monthly_gb_download`
	■	`streaming_movies` `streaming_music` `streaming_tv`
Fidélité	■	Ancienneté du client `tenure_in_months`
	■	Parrainage `referred_a_friend` `number_of_referrals`

Discrétisation : `age_bins`

- âge < 30 **-> 0**
- [30 - 40 [**-> 1**
- [40 - 50 [**-> 2**
- [50 - 65 [**-> 3**
- âge > 65 **-> 4**

Socio / Démographie	■	`genre` `age_bins`
	■	`married` `dependents` `number_of_dependents`
Contrat / Mode de paiement	■	Type d'abonnement `contract`
	■	Offre marketing `offer`
	■	Mode de paiement et support de facturation `payment_meth`
Options / Services suppl.	■	`device_protection_plan` `phone_service` `multiple_li` `online_backup` `premium_tech_support` `unlimited_data`
	■	Le type de service Internet fixe : `Internet_type`
Charges/ Totaux.	■	`monthly_charge`
	■	`avg_monthly_long_distance_charges` `total_long_distance_charges`
	■	`total_charges` `total_extra_data_charges` `total_refunds` `total_revenue`
Flux	■	Volume de data téléchargées `avg_monthly_gb_download`
	■	`streaming_movies` `streaming_music` `streaming_tv`
Fidélité	■	Ancienneté du client `tenure_in_months`
	■	Parrainage `referred_a_friend` `number_of_referrals`

- **Création de la variable définissant la taille de la famille** : `family_size`
- **Discrétisation :** `family_size_bins`
  - Personnes vivant seules **-> 0**
  - Familles composées de 2 pers. **-> 1**
  - 2 < pers. =< 4 **-> 2**
  - pers. > 4. **-> 3**

Data Management // **Discrétisation des variables numériques**

Socio / Démographie	■ `genre` `age_bins` ■ `married` `family_size_bins`
Contrat / Mode de paiement	■ Type d'abonnement `contract` ■ Offre marketing `offer` ■ Mode de paiement et support de facturation `payment_method` `paperless_billing`
Options / Services suppl.	■ `device_protection_plan` `phone_service` `multiple_lines` `internet_service` `online_security` `online_backup` `premium_tech_support` `unlimited_data` ■ Le type de service Internet fixe : `Internet_type`
Charges/ Totaux.	■ `monthly_charge` ■ `avg_monthly_long_distance_charges` `total_long_dista` ■ `total_charges` `total_extra_data_charges` `total_refur`
Flux	■ Volume de data téléchargées `avg_monthly_gb_dow`**`nload`** ■ `streaming_movies` `streaming_music` `streamin`**`g_tv`**
Fidélité	■ Ancienneté du client `tenure_in_months` ■ Parrainage `referred_a_friend` `number_of_referrals`

Ancienneté du client :
`tenure_in_months_bins`

- x < 6 mois **-> 0**
- 6 mois <= x < 1 an **-> 1**
- 1 an <= x < 2 ans **-> 2**
- … **…**
- x > 5 ans **-> 6**

Data Management // **Discrétisation des variables numériques**

Socio / démographie	■ `genre` `age_bins` ■ `married` `family_size_bins`
Contrat / mode de paiement	■ Type d'abonnement `contract` ■ Offre marketing `offer` ■ Mode de paiement et support de facturation `payment_method` `paperless_billing`
Options / services suppl.	■ `device_protection_plan` `phone_service` `multiple_lines` `internet_service` `online_security` `online_backup` `premium_tech_support` `unlimited_data` ■ Le type de service Internet fixe : `Internet_type`
Charges/ totaux.	■ `monthly_charge` ■ `avg_monthly_long_distance_charges` `total_long_distance_charges` ■ `total_charges` `total_extra_data_charges` `total_refur`
Flux	■ Volume de data téléchargées `avg_monthly_gb_download` ■ `streaming_movies` `streaming_music` `streaming_tv`
Fidélité	■ Ancienneté du client `tenure_in_months_bins` ■ Parrainage `referred_a_friend` `number_of_rererrals`

Parrainages : `number_of_referrals_bins`

- nb de parrainages = 0 **-> 0**
- nb de parrainages = 1 **-> 1**
- 2 <= nb de parrainages <5 **-> 2**
- Nb de parrainages >= 5 **-> 3**

Création de 2 indicateurs supplémentaires :

- Nombre d'options / services suppl. d'adhésion du client : `total_of_services_add`

- Augmentation de la facture `increased_monthly_charge` :
  - Calcul du ratio = `monthly_charge` / (`total_charges` / `tenure_in_months`)
  - `increased_monthly_charge` = 1 si ratio > 1
  - `increased_monthly_charge` = 0 sinon

Data Management // **Discrétisation des variables numériques**

Socio / Démographie	■ `genre` `age_bins` ■ `married` `family_size_bins`
Contrat / Mode de paiement	■ Type d'abonnement `contract` ■ Offre marketing `offer` ■ Mode de paiement et support de facturation `payment_method` `paperless_billing`
Options / Services suppl.	■ `device_protection_plan` `phone_service` `multiple_lines` `internet_service` `online_security` `online_backup` `premium_tech_support` `unlimited_data` `total_of_services_add` ■ Le type de service Internet fixe : `Internet_type`
Charges/ Totaux.	■ `monthly_charge` `increased_monthly_charge` ■ `avg_monthly_long_distance_charges` `total_long_distance_charges` ■ `total_charges` `total_extra_data_charges` `total_refunds` `total_revenue`
Flux	■ Volume de data téléchargées `avg_monthly_gb_download` ■ `streaming_movies` `streaming_music` `streaming_tv`
Fidélité	■ Ancienneté du client `tenure_in_months_bins` ■ Parrainage `number_of_referrals_bins`

Présélection des variables // Tri univarié // **V de Cramer**

	features	cramer_v	p-value
1	contract	0.31	0.00
9	tenure_in_months_bins	0.25	0.00
7	number_of_referrals_bins	0.23	0.00
8	total_of_services_add	0.22	0.00
4	internet_type	0.21	0.00
2	offer	0.18	0.00
3	payment_method	0.17	0.00
6	family_size_bins	0.17	0.00
15	internet_service	0.16	0.00
11	paperless_billing	0.14	0.00
16	online_security	0.12	0.00
18	premium_tech_support	0.12	0.00
19	unlimited_data	0.11	0.00
5	age_bins	0.11	0.00
10	married	0.09	0.00
17	online_backup	0.06	0.00
22	streaming_tv	0.05	0.00
12	device_protection_plan	0.05	0.00
20	streaming_movies	0.04	0.00
23	increased_monthly_charge	0.04	0.00
14	multiple_lines	0.03	0.01
21	streaming_music	0.03	0.00
0	gender	0.02	0.12
13	phone_service	0.01	0.41

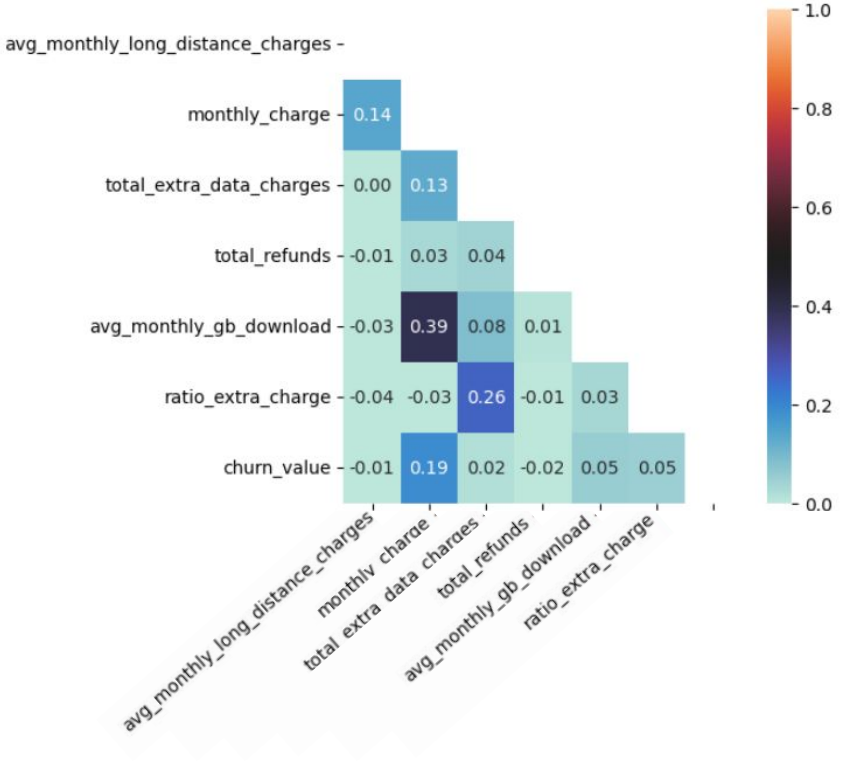
Liaison de Cramer et critère de significativité des variables catégorielles et booléennes avec la target

- Pas de corrélation aberrante entre variables explicatives et target (V de Cramer max de 0.31)
- Pas de redondance entre variables explicatives
- **Critère de sélection** : V de Cramer >= 0.09

→ **9 variables supprimées**

Présélection des variables // Tri univarié // **Corrélation de Pearson**

Matrice de corrélation



Liaison de Pearson et p-value des variables numériques avec la target

	features	coef_pearson	abs_coef_pearson	p_value
5	monthly_charge	0.190	0.190	0.0000
4	avg_monthly_gb_download	0.052	0.052	0.0008
3	ratio_extra_charge	0.048	0.048	0.0019
0	total_refunds	-0.025	0.025	0.1106
2	total_extra_data_charges	0.017	0.017	0.2653
1	avg_monthly_long_distance_charges	-0.008	0.008	0.5898

- Pas de liaison aberrante entre variables explicatives (corrélation de Pearson max de 0.39)
- **Critère de sélection** : Corrélation de Pearson >= 0.09

  ⟶ **5 variables supprimées**

Présélection des variables // Tri univarié

16 variables explicatives restantes

Socio / démographie	■ `genre` `age_bins` ■ `married` `family_size_bins`
Contrat / mode de paiement	■ Type d'abonnement `contract` ■ Offre marketing `offer` ■ Mode de paiement et support de facturation `payment_method` `paperless_billing`
Options / services suppl.	■ `device_protection_plan` `phone_service` `multiple_lines` `internet_service` `online_security` `online_backup` `premium_tech_support` `unlimited_data` `total_of_services_add` ■ Le type de service Internet fixe : `Internet_type`
Charges/ totaux.	■ `monthly_charge` `increased_monthly_charge` ■ `avg_monthly_long_distance_charges` ■ `total_extra_data_charges` `total_refunds`
Flux	■ Volume de data téléchargées `avg_monthly_gb_download` ■ `streaming_movies` `streaming_music` `streaming_tv`
Fidélité	■ Ancienneté du client `tenure_in_months_bins` ■ Parrainage `number_of_referrals_bins`

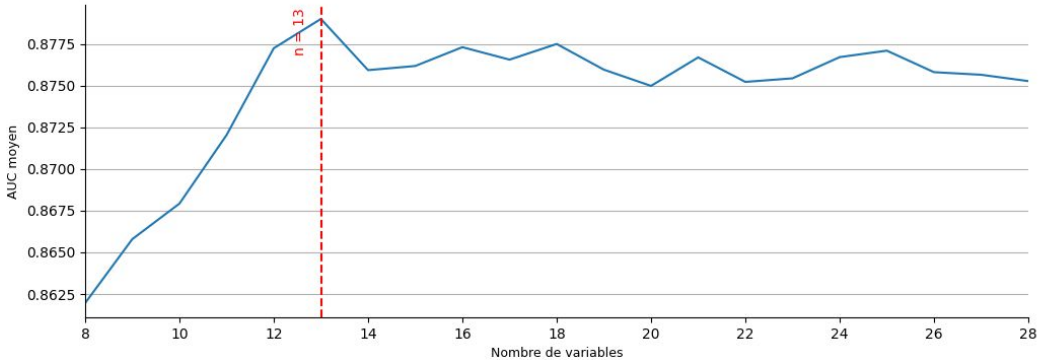
Préprocessing des données

- Encodage manuel des variables booléennes ("No"--> 0, "Yes" --> 1)
- Encodage one-hot avec *pg_dummies* des variables catégorielles textuelles
- Normalisation des variables quantitatives (*StandardScaler*)

} 28 Variables

Recursive Feature Elimination

- Validation croisée : **5 échantillons**
- Estimateur : **RandomForest**
- Nb minimum de variables à sélectionner : **8**
- Nb de variables à éliminer à chaque tour : **1**
- Métrique d'évaluation : **AUC**



Evolution de l'AUC moyen en fonction du nombre de variables durant les itérations de la RFE

→ **12 features sélectionnées** (soit 13 features issues du préprocessing)

Méthode Boruta

➢ Estimateur : **RandomForest sans hyperparamétrage spécifique**

⟶ **14 features retenues incluant les 12 features retenues par la RFE**

12 Features pertinentes retenues

1 Variable numérique :

- `monthly_charge`

5 Variables catégorielles ordinales :

- `age_bins`
- `family_size_bins`
- `number_of_referrals_bins`
- `tenure_in_months_bins`
- `total_of_services_add`

4 Variables catégorielles textuelles :

- `contract`
- `internet_type`
- `offer`
- `payment_method`

2 Variables booléennes :

- `online_security`
- `paperless_billing`

1 Variable numérique :

- `monthly_charge`

➢ **Valeurs manquantes :** Imputation de la valeur moyenne

➢ **Normalisation** via *StandardScaler()*

5 Variables catégorielles ordinales :

- `age_bins`
- `family_size_bins`
- `number_of_referrals_bins`
- `tenure_in_months_bins`
- `total_of_services_add`

➢ **Valeurs manquantes :** Imputation du label le plus fréquent

4 Variables catégorielles textuelles :

- `contract`
- `internet_type`
- `offer`
- `payment_method`

➢ **Valeurs manquantes :** Imputation du label le plus fréquent

➢ **Encodage** via *One-Hot-Encoder()*

2 Variables booléennes :

- `online_security`
- `paperless_billing`

➢ **Valeurs manquantes :** Imputation du label le plus fréquent

1. Evaluation des performances initiales des modèles candidats par validation croisée à 5 blocs

2. Fine-tuning des modèles candidats par l'algorithme RandomizedSearchCV avec cv = 5

Modèle candidats	Distribution des hyper-paramètres

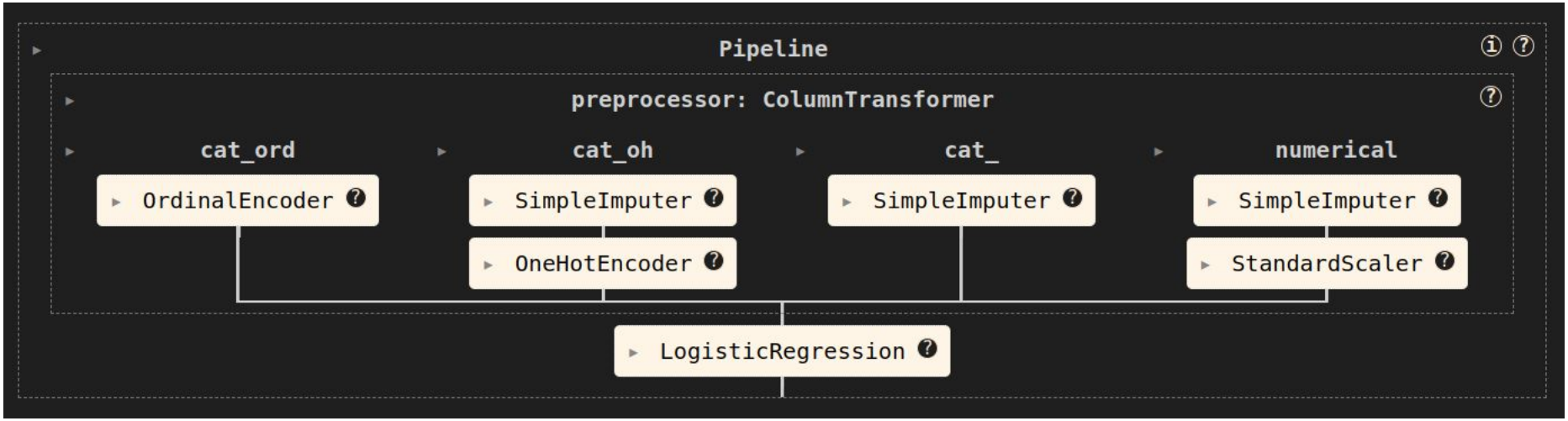
Régression logistique	`solver : ["lbfgs", "liblinear"]`
	`C : loguniform(0.1, 10)`
	`tol : loguniform(0.0001,0.3)`

SVM (kernel linéaire)	`C : loguniform(0.1, 10)` `tol : loguniform(0.0001,0.3)`

XGBoost	`n_estimators : range(100,600,50)`
	`max_depth : range(3,7,1)`
	`eta : loguniform(0.01,0.3),`
	`lambda : loguniform(0.1,10)`

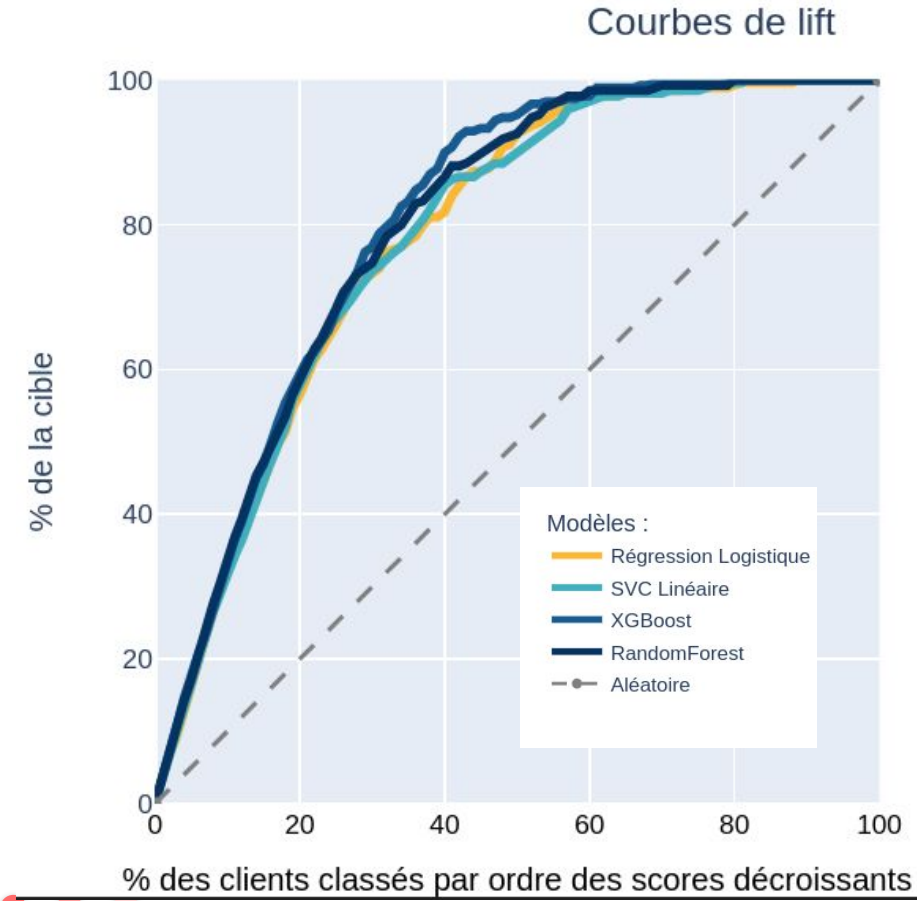
RandomForest	`max_depth : range(3,7,1),` `n_estimators : range(100,600,50)`

Pipeline de transformation



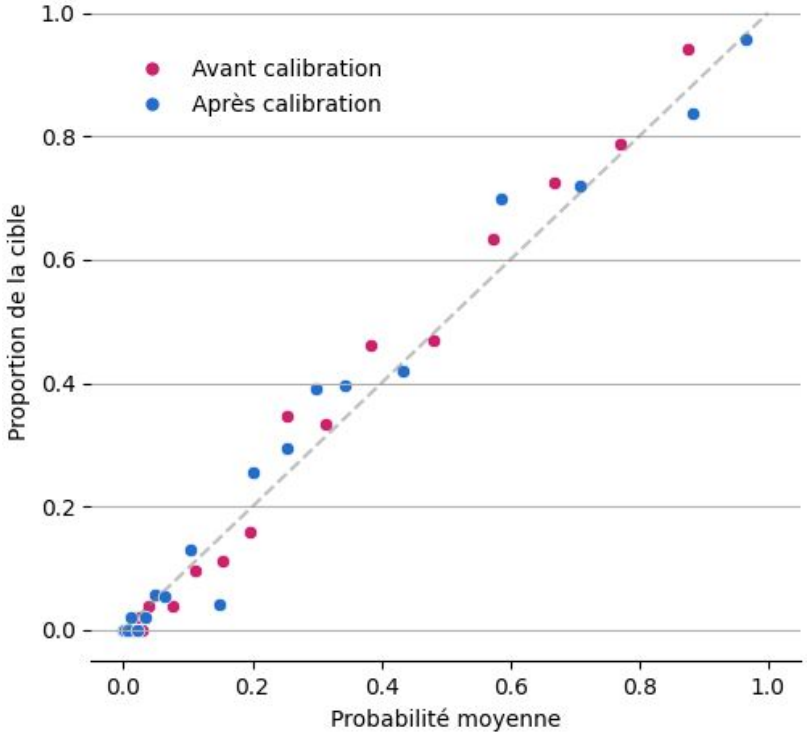
➢ **Métrique :** Area Under the Curve

Modèle	**AUC moyen initial** *(cv sur train)*	**Meilleur AUC moyen** *(RandomizedSearchCV sur train)*	**Hyperparamètres retenus**	**AUC final** *(train)*	**AUC final** *(test)*	**Score de Spiegelhalter** *(test)*
Régression logistique	0.8689	0.8689	C : 1.1 solver : liblinear tol : 0.00135	0.8719	0.8856	**-0.68**
SVM (kernel linéaire)	0.8639	0.8649	C : 3.5 tol : 0.00049	0.8671	0.885	-2.17
XGBoost	**0.8759**	**0.8990**	eta : 0.0367 lambda : 9.2 max_depth: 3, n_estimators : 200	**0.9232**	**0.9126**	-1.59
RandomForest	0.8738	0.8877	n_estimators : 150 max_depth': 6	0.9053	0.903	-4.67



- **Les courbes de lift sont très proches pour les 4 modèles**
- **Tendance des modèles XGBoost et RandomForest à être meilleurs sur les premiers 20% des clients.**
- **XGBoost : en ciblant 10% des clients, on parvient à identifier 33.1% de la cible, soit 3.3 fois plus que pour un ciblage aléatoire.**

➢ **modèle XGBoost retenu**



Courbe de calibration, avant calibration du modèle (en rouge) et après calibration (en bleu).

Faible volumétrie des observations dans le jeu de test (~ 1000 observations) :

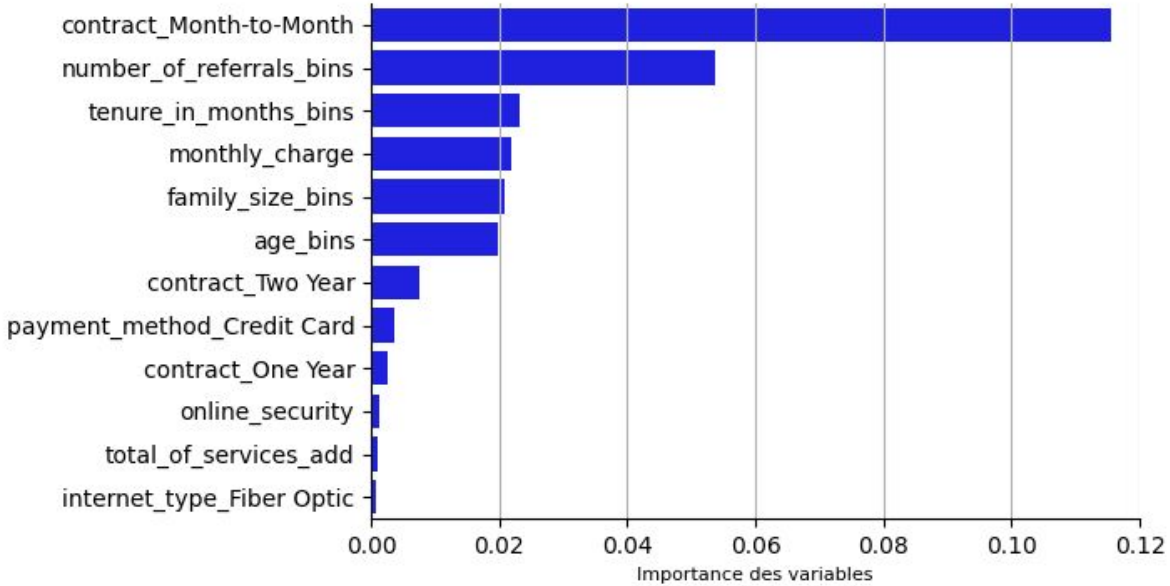
- ➢ Réentraînement du modèle en le calibrant.
- ➢ Stratégie de calibration : régression isotonique.

Score de Spiegelhalter *(test)* :

- ➢ Avant calibration : -1.59
- ➢ Après calibration : 0.56

- **Nette amélioration du score de Spiegelhalter**
- **Amélioration mineure sur la courbe de calibration**

Permutation Feature importance

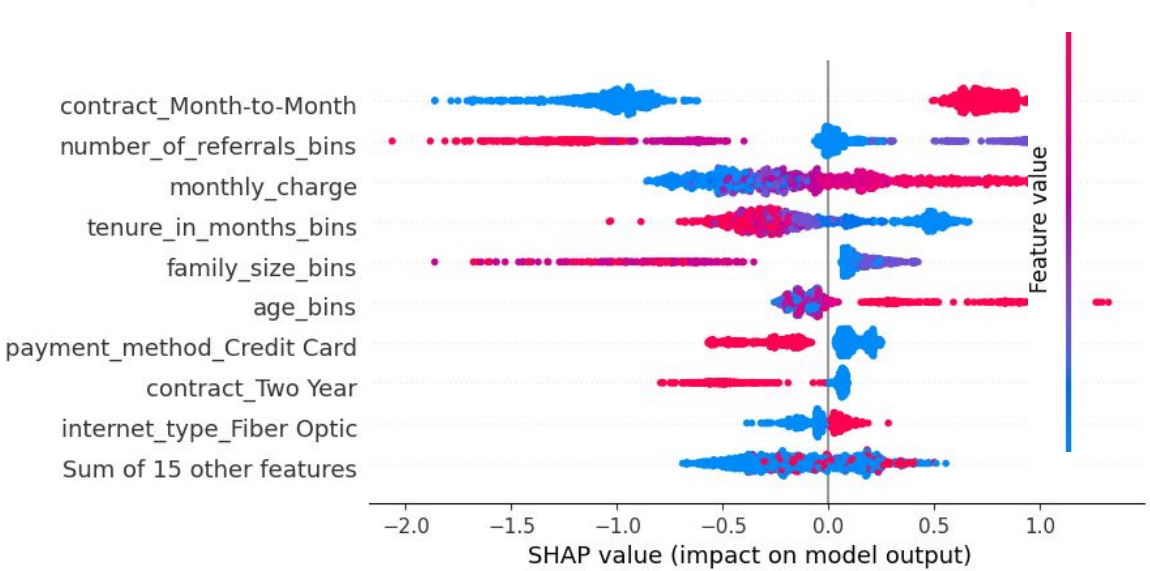


Importance des variables calculée selon la technique 'Permutation Feature Importance'

Pour la prédiction du churn des clients :

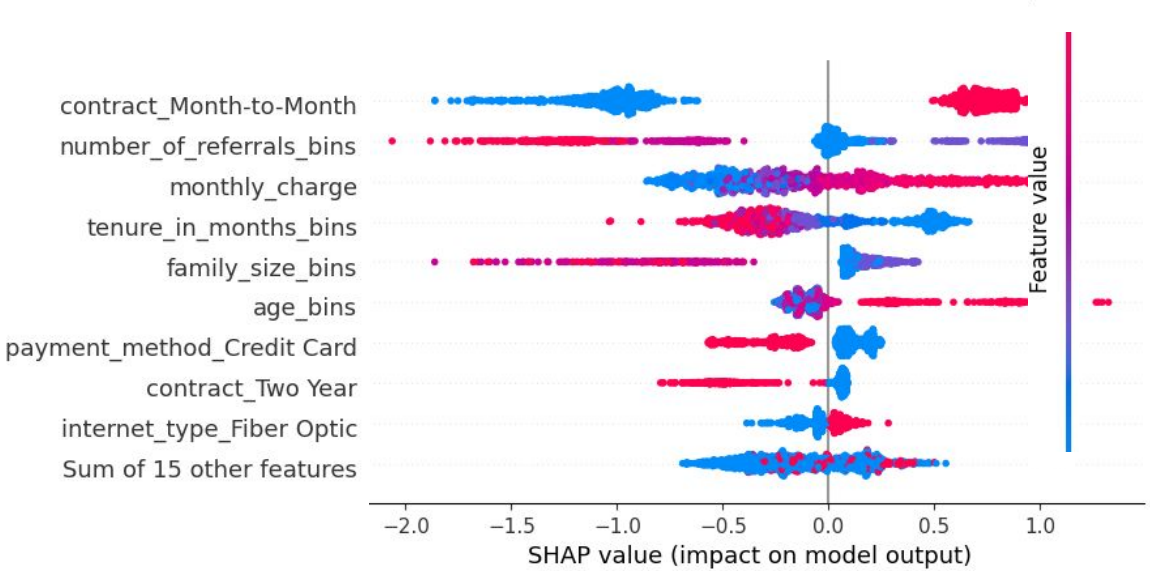
- **3 variables liées à leur engagement sont importantes :**
  - **Abonnement au mois**
  - **Le nombre de parrainages**
  - **Leur ancienneté**
- **Leur facture mensuelle actuelle**
- **La taille de leur famille et leur catégorie d'âge sont également importantes**
- **impact des autres variables mineur**

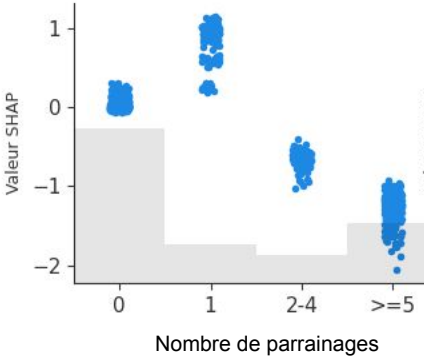
Modélisation & Évaluation // Interprétabilité // **SHAP**



- **Un contrat mensuel sans engagement influence positivement la fragilité des clients**

Modélisation & Évaluation // Interprétabilité // **SHAP**





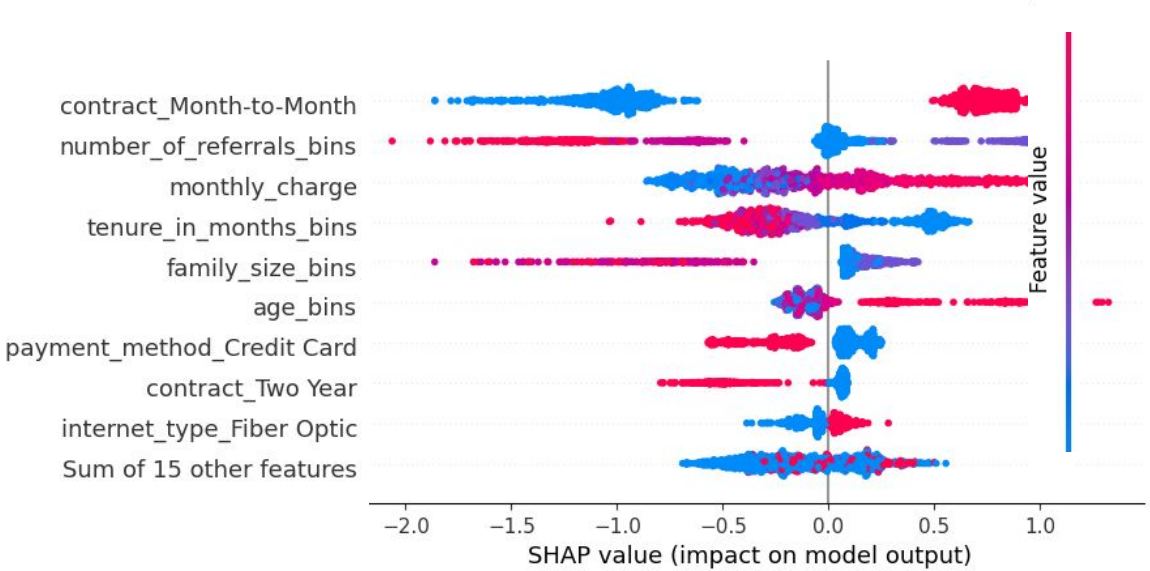
- **Un contrat mensuel sans engagement influence positivement la fragilité des clients**
- **Parrainages :**

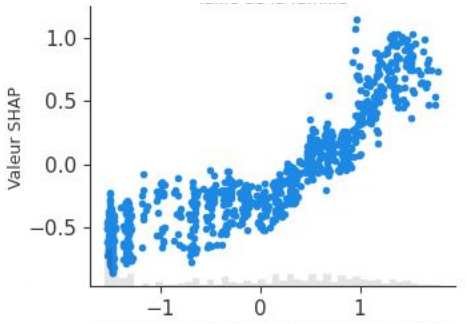
➢ **1 < nb parrainages : la fidélité du client est renforcée**

➢ **Aucun parrainage : peu d'influence sur la fragilité du client**

➢ **1 parrainage tend à favoriser le churn.**

Modélisation & Évaluation // Interprétabilité // **SHAP**





Montant de la facture mensuelle actuelle (normalisé)

- **Un contrat mensuel sans engagement influence positivement la fragilité des clients**
- **Parrainages :**

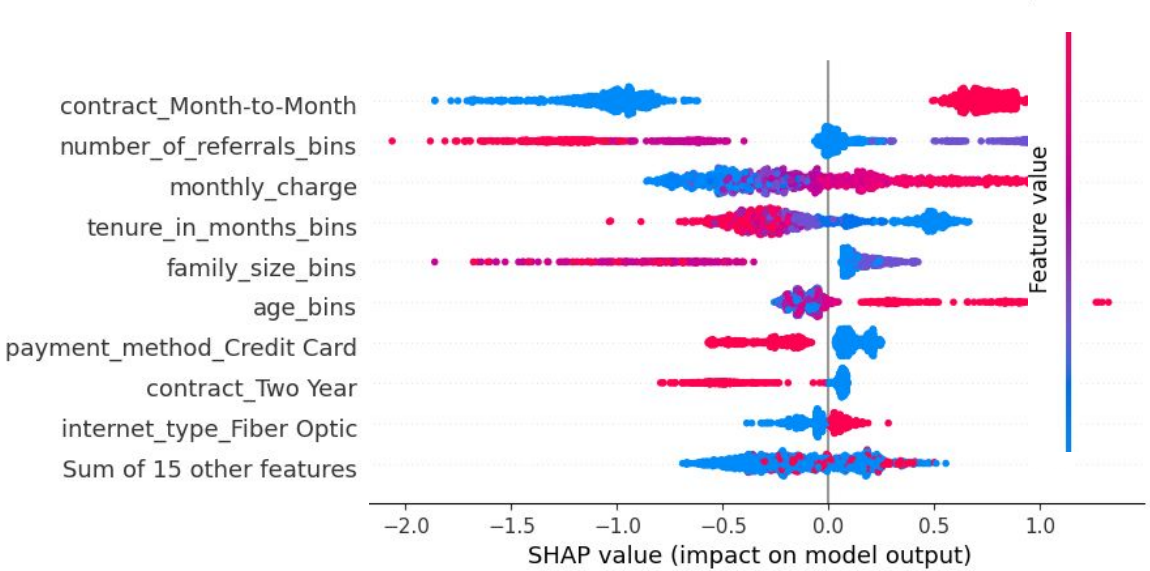
➢ **1 < nb parrainages : la fidélité du client est renforcée**

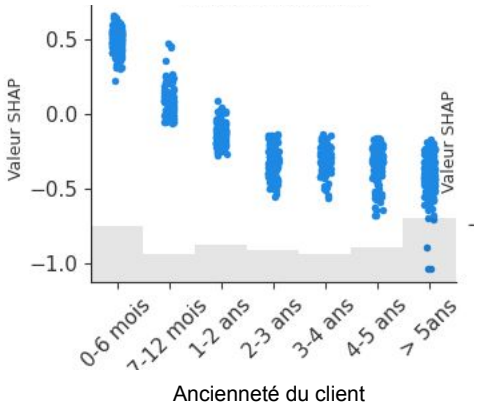
➢ **Aucun parrainage : peu d'influence sur la fragilité du client**

➢ **1 parrainage tend à favoriser le churn.**

- **Les valeurs SHAP sont proportionnelles aux charges mensuelles payées**

Modélisation & Évaluation // Interprétabilité // **SHAP**





- **Un contrat mensuel sans engagement influence positivement la fragilité des clients**
- **Parrainages :**

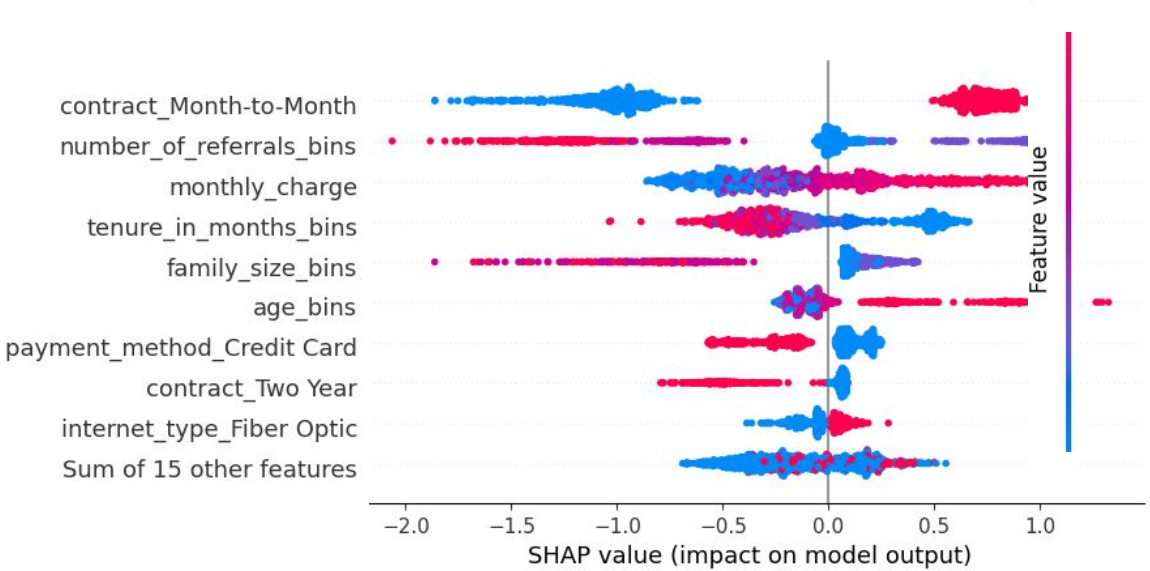
➢ **1 < nb parrainages : la fidélité du client est renforcée**

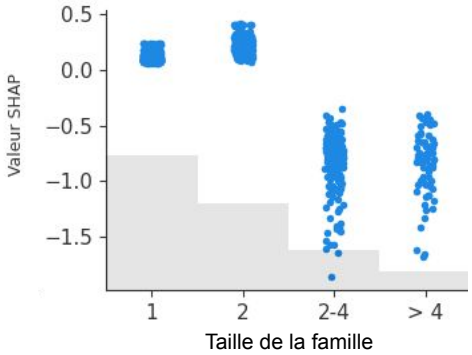
➢ **Aucun parrainage : peu d'influence sur la fragilité du client**

➢ **1 parrainage tend à favoriser le churn.**

- **Les valeurs SHAP sont proportionnelles aux charges mensuelles payées**
- **Les clients récents dont l'inscription date de moins d'un an sont plus disposés à résilier leur contrat.**

Modélisation & Évaluation // Interprétabilité // **SHAP**





- **Un contrat mensuel sans engagement influence positivement la fragilité des clients**
- **Parrainages :**

➢ **1 < parrainages : la fidélité du client est renforcée**

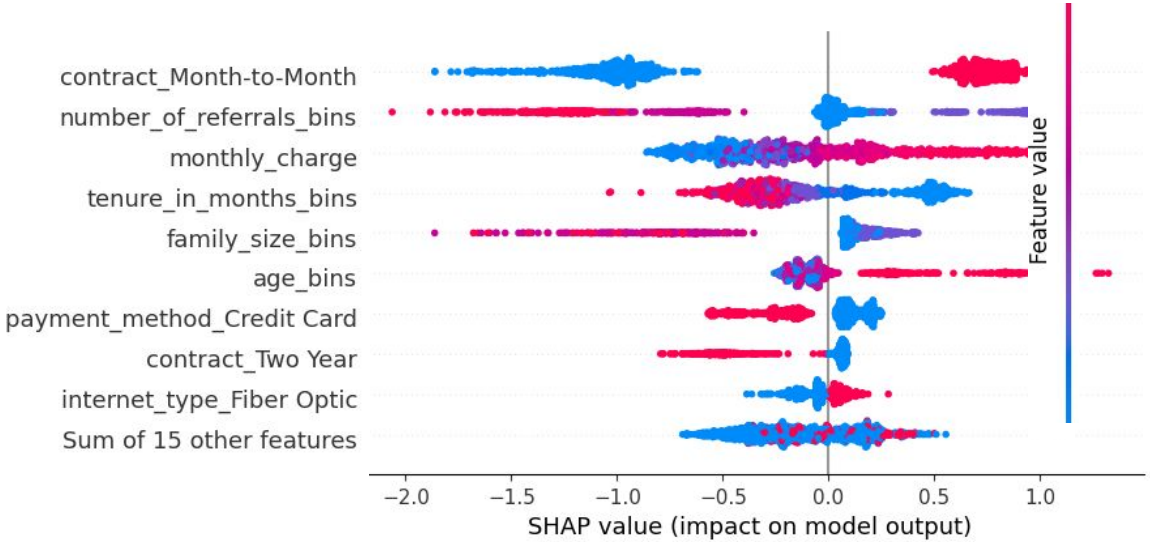
➢ **Aucun parrainage : peu d'influence sur la fragilité du client**

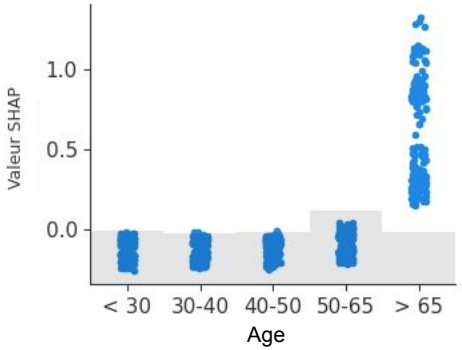
➢ **1 parrainage tend à favoriser le churn.**

- **Les valeurs SHAP sont proportionnelles aux charges mensuelles payées**
- **Les clients récents dont l'inscription date de moins d'un an sont plus disposés à résilier leur contrat.**
- **Taille de la famille :**

➢ **Les personnes seules ou à deux sont considérées comme plus fragiles**

Modélisation & Évaluation // Interprétabilité // **SHAP**





- **Un contrat mensuel sans engagement influence positivement la fragilité des clients**
- **Parrainages :**

➢ **1 < nb parrainages : la fidélité du client est renforcée**

➢ **Aucun parrainage : peu d'influence sur la fragilité du client**

➢ **1 parrainage tend à favoriser le churn.**

- **L'importance des valeurs SHAP est proportionnelle aux charges mensuelles payées**

- **Les clients dont l'ancienneté est inférieure à un an sont plus disposées à résilier leur contrat.**
- **Taille de la famille :**

➢ **Les personnes seules ou à deux sont considérées comme plus fragiles**

- **Seule la catégorie des séniors (> 65 ans) influence positivement les valeurs de SHAP**

Conclusion

L'analyse exploratoire de la base de données des clients de l'entreprise de télécommunication nous a permis de réduire le nombre de variables explicatives pertinentes de 35 à 12.

Après entraînement, fine-tuning et calibration d'un modèle ML XGBoost :

- Bonnes performances : AUC = 0.9232 sur la base d'entraînement
- Une bonne robustesse et pas de surapprentissage : AUC de 0.9126 sur la base de test
- Une calibration correcte : Score de Spiegelhalter de 0.56
- Courbe lift : en ciblant 10% des clients, on parvient à identifier 33.1% des clients possiblement fragiles.

Le degré de fragilité des clients de l'entreprise dépend :

- de leur **fidélité** et de leur **engagement** vis à vis de l'entreprise :
  les clients ayant souscrit à un contrat **mensuel**, **n'ayant parrainé qu'une seule personne,** ayant une **ancienneté de moins d'un an** sont plus susceptibles de résilier leur contrat.
- du montant de leur facture mensuelle
- des paramètres socio-démographiques : les clients séniors ainsi que les clients vivants seuls ou à deux sont possiblement plus fragiles.