Rapport de Projet

Data Mining

Réalisé par : Groupe 3

NGAM Amadou

SARR Serigne Abdou Lat

KANE Mamadou

KAFANDO Tounwendsida Bertrand

DIALLO Mouhamed Bobo

DIOP Cheikh Ahmadou Bamba

Professeur : M. NELSON BOTERO GIRALDO

Table des matières

I. Description du jeu de données

Notre jeu de données **commsdata** contient des informations sur les clients d'une entreprise de télécommunications et leur utilisation des services de l'entreprise. Les variables d'entrée de ce jeu de données comprennent des données démographiques, des variables décrivant l'utilisation et le type de produits, des données de facturation, ainsi que des informations sur le service client et le centre d'appels.
Le jeu de données est composé de 56557 lignes et de 128 colonnes.

II. Objectifs du projet

L'objectif de ce projet est de prédire le désabonnement des clients à partir des données **commsdata** afin d'apporter plusieurs avantages et bénéfices à l'entreprise. En effet l'entreprise pourrait :

- Prédire si un client risque de résilier son abonnement ou pas.
- Réduire le taux de désabonnement et à conserver plus de clients.
- Renforcer la fidélité des clients existants.
- Augmenter les revenus de l'entreprise.
- Concentrer ses efforts marketing et ses ressources sur certains groupes de clients.

Ces avantages permettent à l'entreprise d'être plus performante et compétitive sur le marché des télécommunications.

III. Acquisition et Exploration des données

1. Variable à prédire

Dans cette étape, nous allons prédire la variable **Churn** qui prend deux valeurs : 0 ou 1. **Churn** prend la valeur 1 si le client résilie son contrat et 0 si le client ne se désabonne pas.

La prédiction de la variable "Churn" revêt une grande importance, car elle permet de mesurer la fidélité des clients, d'anticiper les pertes potentielles de clients, et de prendre des mesures préventives pour les retenir. L'objectif principal est de construire un modèle prédictif capable d'identifier les clients à risque de résilier leur abonnement.

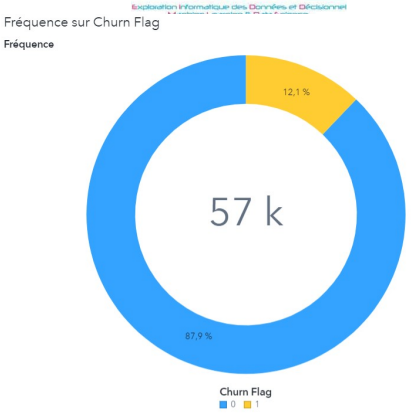


figure 1 : Fréquence du Churn Flag

2. Variables à rejeter

- Toutes les variables qui représentent les coordonnées géographiques du client sont rejetées. Par exemple : **city_lat, city_long, region_lat, region_long** …

En effet ces variables sont spécifiques à un client et ne fournissent pas directement d'informations sur la capacité d'un client à résilier ou non son contrat. Elles sont principalement des données de localisation géographique qui permettent de décrire la localisation des clients, mais elles ne sont pas intrinsèquement liées à la prédiction du churn.

- On a également rejeté la variable **CENSUS AREA TOTAL RURAL.** Cette variable est fortement corrélée négativement avec la variable **CENSUS AREA TOTAL URBAN.**

Nous pouvons l'observer sur la *matrice corrélation de la figure 2* :

figure 2 : matrice de corrélation

- **Elimination des variables après entraînement des données avec du Random forest**

Pour écarter les variables non importantes, nous avons entraîné nos données sur un modèle de random forest. Nous avons obtenu un classement de nos variables en fonction de leur importance.

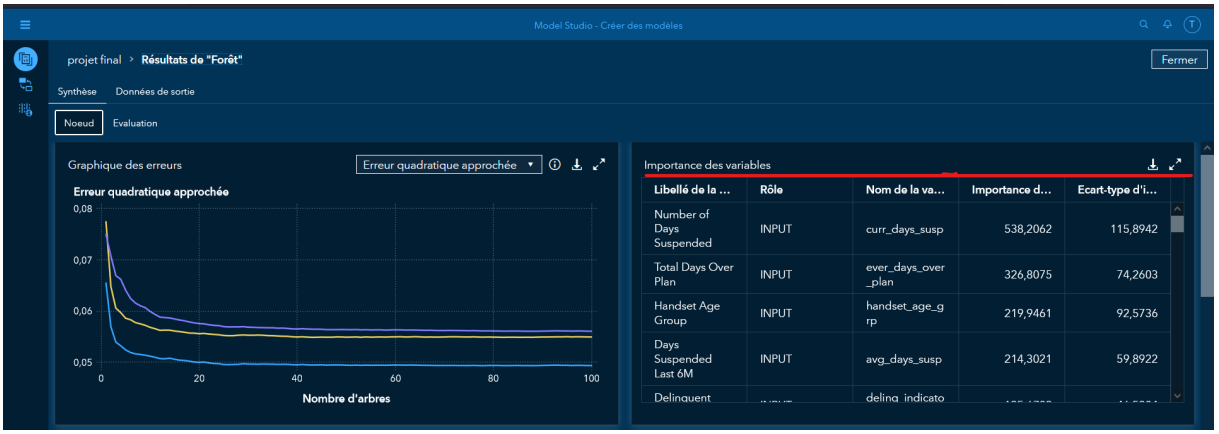


figure 3 : Résultat du Random Forest

Nous avons par la suite généré l'importance cumulée. Ces valeurs indiquent pour le pourcentage de l'information à retenir. Avec ces derniers nous avons faits une série de tests en variant la quantité d'information à retenir : 95% (ligne rouge), 98% (ligne bleue), 100% (ligne jaune).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
55	INPUT	7.235262	3.083132	calls_TS_a	Number Calls Tech Support	0.014977866	2792.966											
56	INPUT	7.23301	1.764189	cs_other	Census Area Other	0.014973204	2800.199											
57	INPUT	7.231398	1.787336	data_devic	Avg Age of Devices on Plan	0.014969868	2807.43											
58	INPUT	7.071306	1.69149	cs_med_h	Census Area Median Home Value In	0.014638458	2814.501											
59	INPUT	6.997335	1.921737	cs_ttl_mal	Census Area Total Males	0.014485329	2821.499											
60	INPUT	6.923562	2.612184	num_tsupc	Tech Support Complaints - LTD	0.01433261	2828.422											
61	INPUT	6.769979	1.866417	nbr_data_(	Number of Data Records	0.014014675	2835.192											
62	INPUT	6.624872	1.757082	cs_ttl_pop	Census Area Total Population	0.013714285	2841.817											
63	INPUT	6.582675	3.834369	unsolv_tsu	Unresolved Tech Support Complaint	0.013626933	2848.4											
64	INPUT	6.54495	1.869976	bill_data_t	Total Billed Data Usage	0.013548838	2854.945											
65	INPUT	6.532051	1.672097	acct_age	Account Tenure	0.013522135	2861.477											
66	INPUT	6.422658	1.612431	handset	Handset Mfg	0.013295678	2867.9											
67	INPUT	6.296315	1.598266	cs_pct_ho	Census Area Percent Home Owner	0.013034134	2874.196											
68	INPUT	5.773927	1.462409	count_of_:	Times Suspended Last 6M	0.011952727	2879.97											
69	INPUT	5.311412	1.917168	forecast_r	Forecasted Region Key	0.010995264	2885.281											
70	INPUT	5.187201	1.593498	avg_data_(	3M Avg Data Charges	0.010738132	2890.468											
71	INPUT	4.742245	1.590618	tot_voice_	Total Voice Charges	0.00981702	2895.211											
72	INPUT	4.588185	1.60215	equip_age	Handset Age	0.009498098	2899.799											
73	INPUT	4.044128	1.831764	sales_chan	Acquisition Channel	0.008371835	2903.843											
74	INPUT	3.395018	1.581	credit_clas	Credit Class	0.007028097	2907.238											
75	INPUT	3.325837	1.429806	cs_ttl_urb	Census Area Total Urban	0.006884885	2910.564											
76	INPUT	3.303017	1.573051	tot_drpd_(	Number of Dropped Calls 1 Mth Pric	0.006837645	2913.867											
77	INPUT	3.169312	2.664253	open_tsup	Open Tech Support Complaints	0.006560858	2917.036											
78	INPUT	2.972823	1.698425	lifestage	Plan Life Stage	0.006154103	2920.009											
79	INPUT	2.965682	1.494357	nbr_conta	Number Times Customer Contacted	0.006139321	2922.975											
80	INPUT	2.87767	1.743573	avg_overa	3M Avg Overage Charges	0.005957124	2925.852											
81	INPUT	2.791366	1.597681	avg_data_(	3M Avg Premium Data Charges	0.005778465	2928.644											
82	INPUT	2.699078	1.506808	calls_care_	Number Calls Care Center	0.005587418	2931.343											
83	INPUT	2.631256	1.608211	nbr_contra	Total Number Contracts Lifetime	0.005447018	2933.974											
84	INPUT	2.409767	1.806352	data_prem	Premium Data Charges	0.004988509	2936.384											
85	INPUT	2.315484	1.854839	last_rep_s	Last Call Satisfaction Rating Given	0.004793333	2938.699											

figure 4 : Importance des variables

Nous avons créé un pipeline sur SAS pour chacun des cas. 98% a donné une meilleure performance et nous l'avons gardé. Nous avons donc supprimé toutes les variables en dessous de la ligne rouge.

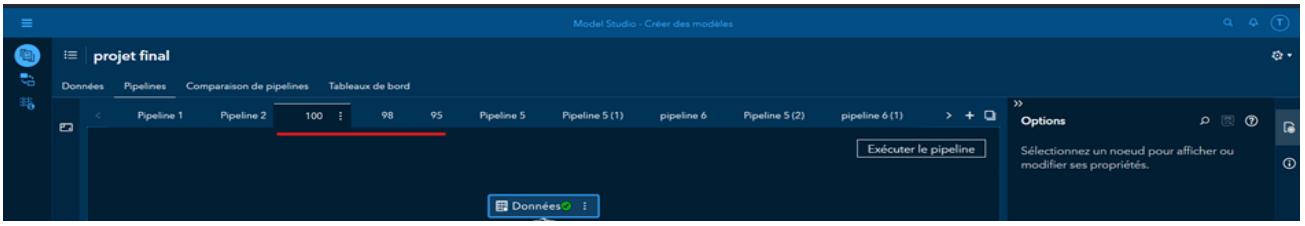


figure 5 : les pipelines pour le choix des modèles

Liste des variables éliminées avec cette méthode.

	A	B	C	D	E	F	G
1	Rôle	Importance d'apprentissa	Ecart-type d'importance	Nom de la variable	Libellé de la variable	Importance relative	IA cumulée
67	INPUT	6.296315493	1.59827	cs_pct_ho	Census Area Percent Home Ownei	0.013034134	2874.2
68	INPUT	5.773927262	1.46241	count_of_	Times Suspended Last 6M	0.011952727	2879.97
69	INPUT	5.311411608	1.91717	forecast_r	Forecasted Region Key	0.010995264	2885.28
70	INPUT	5.187200794	1.5935	avg_data_	3M Avg Data Charges	0.010738132	2890.47
71	INPUT	4.742244857	1.59062	tot_voice_	Total Voice Charges	0.00981702	2895.21
72	INPUT	4.588185289	1.60215	equip_age	Handset Age	0.009498098	2899.8
73	INPUT	4.044128456	1.83176	sales_cha	Acquisition Channel	0.008371835	2903.84
74	INPUT	3.395017555	1.581	credit_cla	Credit Class	0.007028097	2907.24
75	INPUT	3.325837103	1.42981	cs_ttl_urb	Census Area Total Urban	0.006884885	2910.56
76	INPUT	3.303017111	1.57305	tot_drpd_	Number of Dropped Calls 1 Mth F	0.006837645	2913.87
77	INPUT	3.169311666	2.66425	open_tsup	Open Tech Support Complaints	0.006560858	2917.04
78	INPUT	2.972823248	1.69843	lifestage	Plan Life Stage	0.006154103	2920.01
79	INPUT	2.965682402	1.49436	nbr_conta	Number Times Customer Contacte	0.006139321	2922.97
80	INPUT	2.877669891	1.74357	avg_overa	3M Avg Overage Charges	0.005957124	2925.85
81	INPUT	2.791365968	1.59768	avg_data_	3M Avg Premium Data Charges	0.005778465	2928.64
82	INPUT	2.699078392	1.50681	calls_care	Number Calls Care Center	0.005587418	2931.34
83	INPUT	2.631256145	1.60821	nbr_contr	Total Number Contracts Lifetime	0.005447018	2933.97
84	INPUT	2.40976683	1.80635	data_pren	Premium Data Charges	0.004988509	2936.38
85	INPUT	2.315484495	1.85484	last_rep_s	Last Call Satisfaction Rating Give	0.004793333	2938.7
86	INPUT	2.259761339	1.88111	tot_overaj	Total Overage Charges	0.004677979	2940.96
87	INPUT	1.175444186	1.70575	calls_care	Number Calls Care Center 6 Mont	0.002433312	2942.13
88	INPUT	0.936283919	1.86539	price_mer	Price Issues Discussed	0.001938221	2943.07
89	INPUT	0.880212432	1.51791	mfg_apple	Own Apple	0.001822146	2943.95
90	INPUT	0.714244339	1.67085	res_calls_	Resolved Calls - 6Mo Average	0.001478572	2944.67
91	INPUT	0.710611003	1.80411	calls_care	Number Calls Care Center 3 Mont	0.001471051	2945.38
92	INPUT	0.702589038	2.2444	res_calls_	Resolved Calls - 3Mo Average	0.001454444	2946.08
93	INPUT	0.653060444	1.89699	mfg_moto	Own Motorola	0.001351914	2946.73
94	INPUT	0.64058484	1.09111	rp_poolec	Pooled Rate Plan	0.001326088	2947.37
95	INPUT	0.63338186	3.04285	mfg_htc	Own HTC	0.001311177	2948.01
96	INPUT	0.509431201	1.20931	mfg_sams	Own Samsung	0.001054584	2948.51
97	INPUT	0.4503536	1.12178	network_r	Network Issues Discussed	0.000932286	2948.97
98	INPUT	0.447219518	1.3063	service_m	Service Issues Discussed	0.000925798	2949.41
99	INPUT	0.446101527	2.15223	mfg_nokia	Own Nokia	0.000923484	2949.86
100	INPUT	0.354776658	1.30588	mfg_lg	Own LG	0.000734431	2950.21
101	INPUT	0.34104793	1.22417	upsell_xs	Xsell Upsell Flag	0.00070601	2950.55

figure 6 : liste des variables éliminées

- **Exploration des données**

Dans cette partie nous examinerons les diagrammes en barres interactif qui mettent en relation trois variables de notre base de données que nous avons identifiées comme étant importantes. Ces variables incluent la variable "**Churn**" (taux de désabonnement), la variable "**Plan Name**" (nom du plan du client) et la variable "**Final Resolution**" (action de résolution prise par le centre d'appels).

Cette visualisation nous aidera à explorer les relations entre ces variables clés.

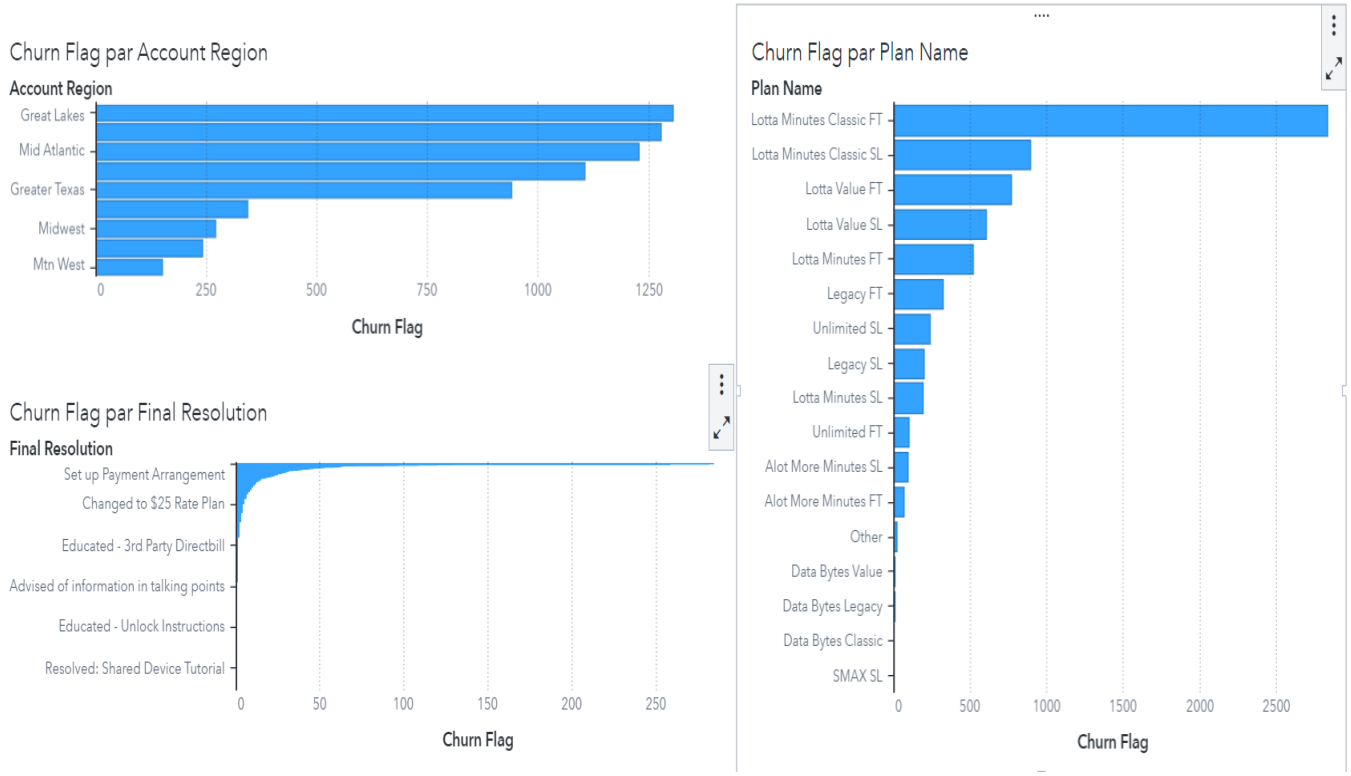


figure 7 : visualisation

Cette phase d'analyse exploratoire des données est essentielle pour cerner les tendances, les corrélations et les points d'intérêt dans nos données, ce qui nous aidera à prendre des décisions éclairées lors de la phase de modélisation à venir.

IV. Implémentation du modèle

1. Pipeline initial

Nous avons élaboré un pipeline pour transformer les données et implémenter divers modèles : ***Gradient Boosting, Réseau Neuronal, Régression Logistique pas à pas, Régression Logistique Ascendante, Arbre de décision et Forêt aléatoire***. Avant ces étapes nous intégrons des étapes d'imputation pour remplacer les données manquantes et une sélection de variables pour retenir les plus pertinentes, essentielles pour optimiser la performance des modèles de réseau de neurone et de régression logistique ensuite nous formons un ensemble qui combine les prédictions de ces modèles pour augmenter la robustesse et la précision. Cette approche consolidée nous permet de comparer efficacement les modèles et désigner les plus performants.

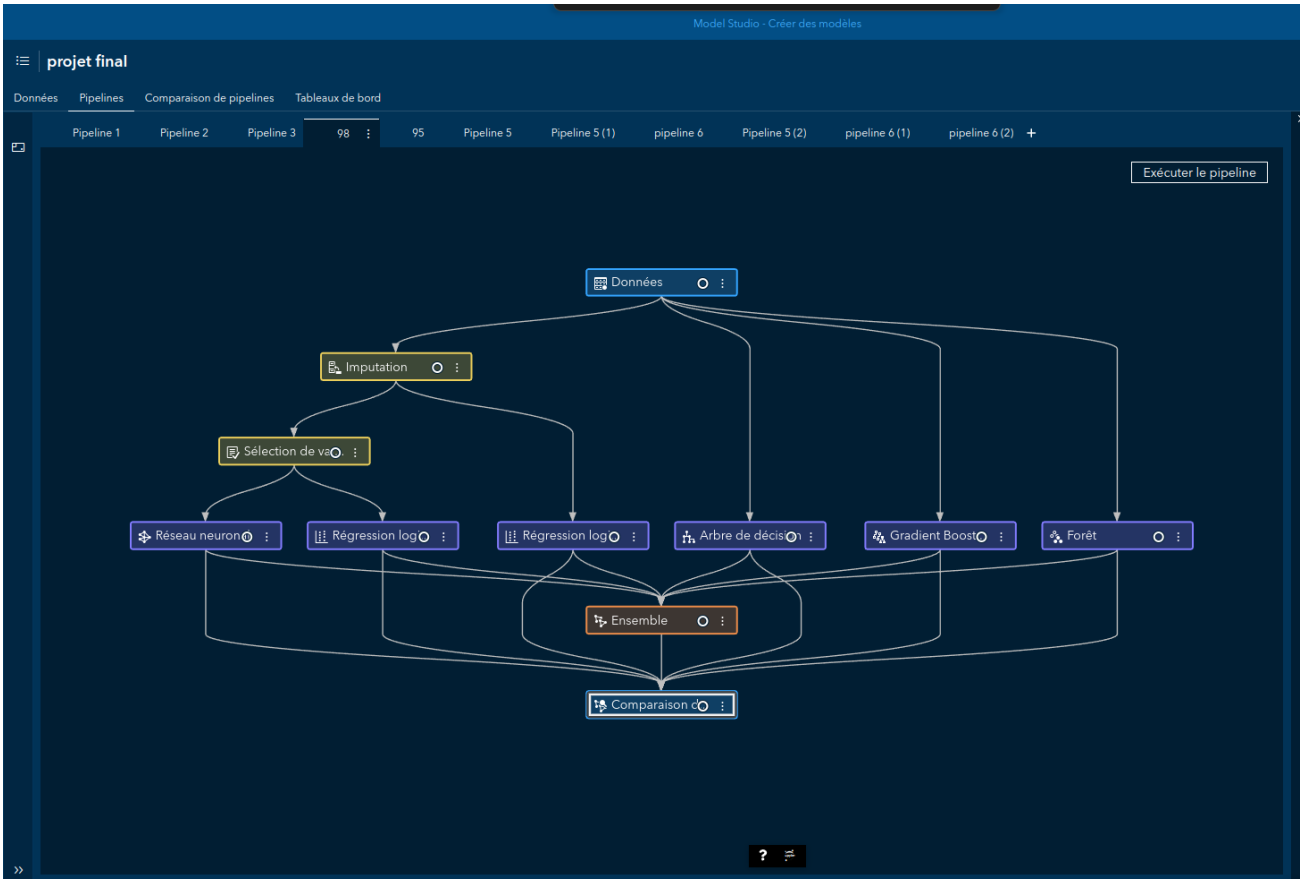


figure 8 : pipeline du modèle initiale

Après exécution du pipeline, nous avons obtenu les résultats de différentes modèles:

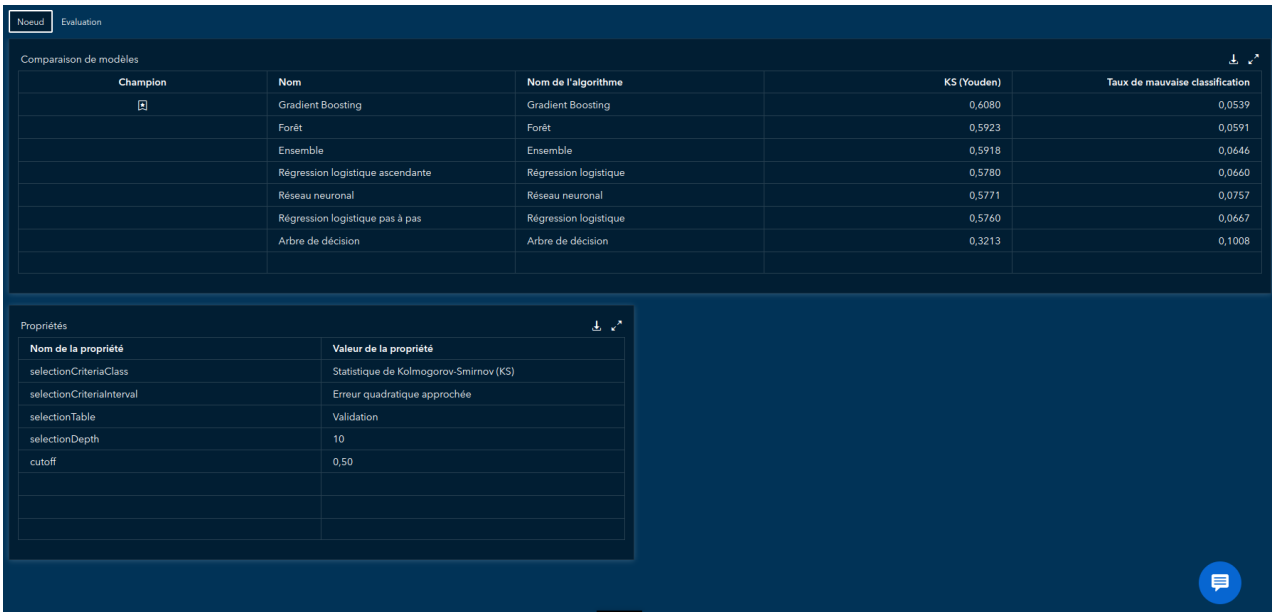


figure 9 : modèle champion

2. Choix du meilleur modèle

On observe que le **Gradient Boosting** est le meilleur modèle avec **0.6080** comme score de KS.Nous allons optimiser le modèle en effectuant des prétraitements sur les données.

3. Optimisation du modèle

Pour avoir un modèle robuste, nous avons effectué une série de traitements et de tests.Premièrement, certaines données présentent des anomalies, nous avons ajouté le nœud "**détection d'anomalies"** dans le pipeline pour éliminer les valeurs aberrantes.

Deuxièmement, Nous avons également ajouté le nœud **"Feature Machine"** qui génère des fonctionnalités qui s'appliquent à une ou plusieurs stratégies de transformations sur les données telles que la cardinalité, missingness, et skewness mais aussi des filtres (coefficient de variation, regrouper les modalités rares) sur les variables explicatives.

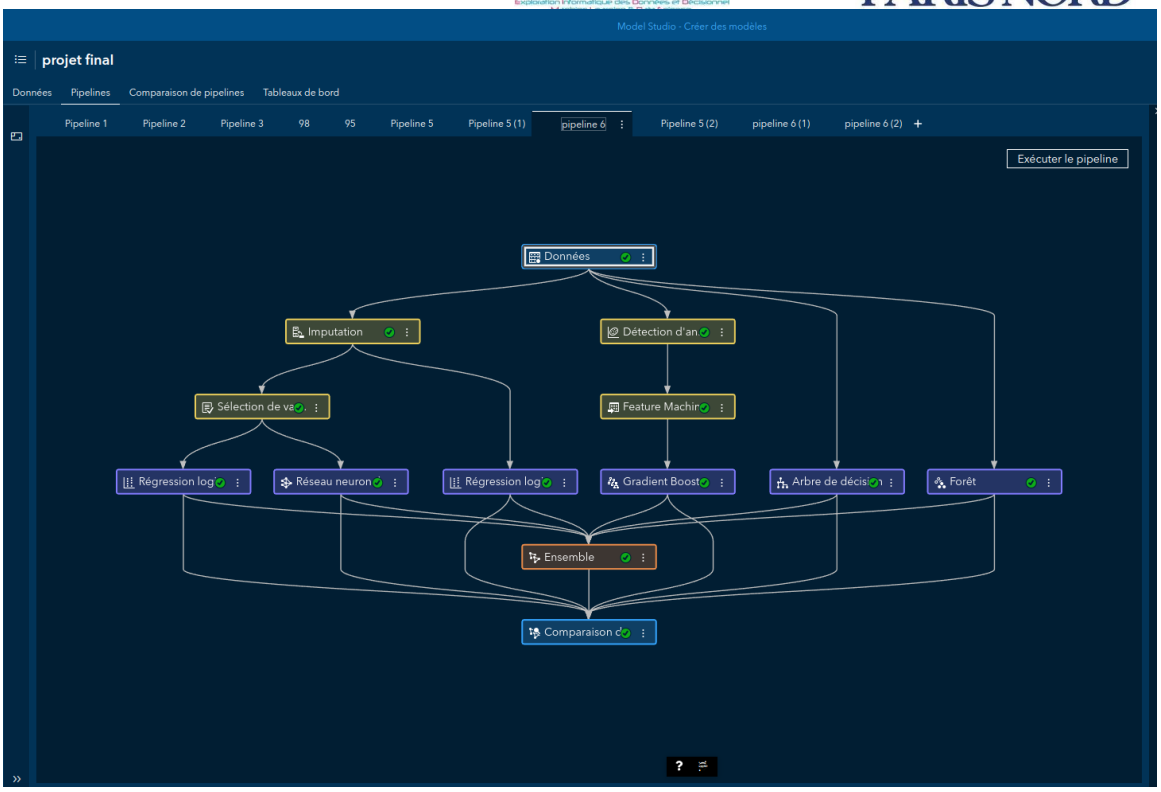
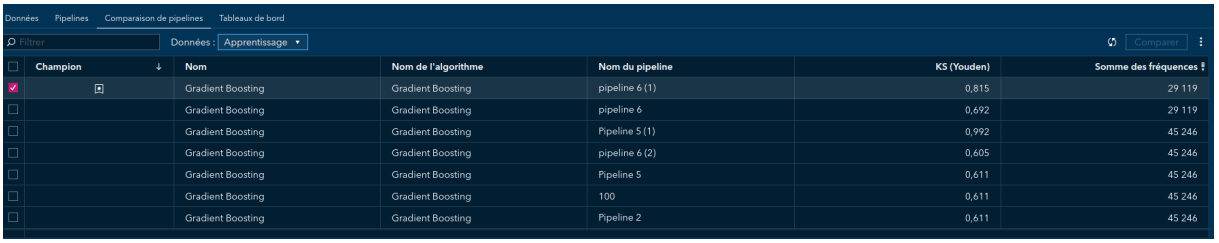


figure 10 : pipeline du modèle optimisé

Ces 2 couches de prétraitement ont boosté considérablement notre modèle.

Enfin, nous avons essayé de jouer sur les hyperparamètres en créant un nouveau pipeline de données. Nous avons par exemple augmenté le nombre d'arbres, la profondeur des arbres. Cette série de tests nous a donné plusieurs pipelines avec des scores différents comme nous l'image ci-dessous .

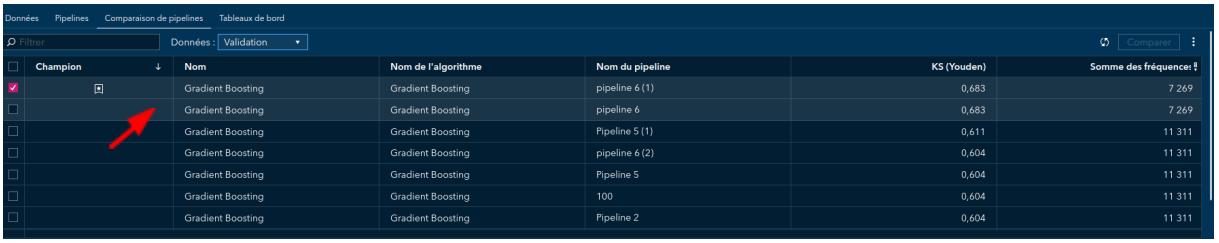
- Comparaison des pipelines sur les **données d'apprentissage**



Champion	Nom	Nom de l'algorithme	Nom du pipeline	KS (Youden)	Somme des fréquences
☑	Gradient Boosting	Gradient Boosting	pipeline 6 (1)	0,815	29 119
☐	Gradient Boosting	Gradient Boosting	pipeline 6	0,692	29 119
☐	Gradient Boosting	Gradient Boosting	Pipeline 5 (1)	0,992	45 246
☐	Gradient Boosting	Gradient Boosting	pipeline 6 (2)	0,605	45 246
☐	Gradient Boosting	Gradient Boosting	Pipeline 5	0,611	45 246
☐	Gradient Boosting	Gradient Boosting	100	0,611	45 246
☐	Gradient Boosting	Gradient Boosting	Pipeline 2	0,611	45 246

figure 11 : résultats des pipelines sur les données d'apprentissage

- Comparaison des pipelines sur les **données de validations**



Champion	Nom	Nom de l'algorithme	Nom du pipeline	KS (Youden)	Somme des fréquences
	Gradient Boosting	Gradient Boosting	pipeline 6 (1)	0,683	7 269
	Gradient Boosting	Gradient Boosting	pipeline 6	0,683	7 269
	Gradient Boosting	Gradient Boosting	Pipeline 5 (1)	0,611	11 311
	Gradient Boosting	Gradient Boosting	pipeline 6 (2)	0,604	11 311
	Gradient Boosting	Gradient Boosting	Pipeline 5	0,604	11 311
	Gradient Boosting	Gradient Boosting	100	0,604	11 311
	Gradient Boosting	Gradient Boosting	Pipeline 2	0,604	11 311

figure 12 : résultats des pipelines sur les données de validation

Nous avons obtenu deux pipelines avec des modèles assez performants .

Pipeline 6(1) : 0,815 pour l'apprentissage et 0,683 pour la validation.

Pipeline 6 : 0,692 pour l'apprentissage et 0,683 pour la validation.

À cause de l'écart entre le coefficient de Ks(youden) de validation et d'apprentissage du modèle Gradient Boosting du **pipeline 6 (1)** nous avons préféré garder le modèle gradient Boosting du pipeline 6. Nous avons fait ce choix parce que nous pensons que l'écart peut potentiellement être synonyme de sur apprentissage.

4. Résultats finaux

Voici les détails de nos hyperparamètres et paramètres de notre modèle final.

Le Système SAS

La procédure GRADBOOST

Informations sur le modÃ¨le	
Nombre d'arbres	100
Taux d'apprentissage	0,1
Taux de sous-échantillonnage	0,5
Nombre de variables par division	89
Nombre de classes	50
Nombre de variables en entrée	89
Nombre maximum de noeuds de l'arbre	31
Nombre minimum de noeuds de l'arbre	25
Nombre maximum de branches	2
Nombre minimum de branches	2
Profondeur maximale	4
Profondeur minimale	4
Nombre maximum de feuilles	16
Nombre minimum de feuilles	13
Taille de feuille maximale	11503
Taille de feuille minimale	5
Valeur initiale	12345
Pénalité de Lasso (L1)	0
Pénalité Ridge (L2)	1
Nombre réel d'arbres	85
Nombre moyen de feuilles	15,8117647
Stagnation de l'arrêt précoce	5
Seuil de l'arrêt précoce	0
Itérations du seuil de l'arrêt précoce	0
Tolérance de l'arrêt précoce	0

	Apprentissage	Validation	Total
nbre d'observations lues	29119	7269	36388
nbre d'observations utilisées	29119	7269	36388

figure 13 : paramètres du modèles

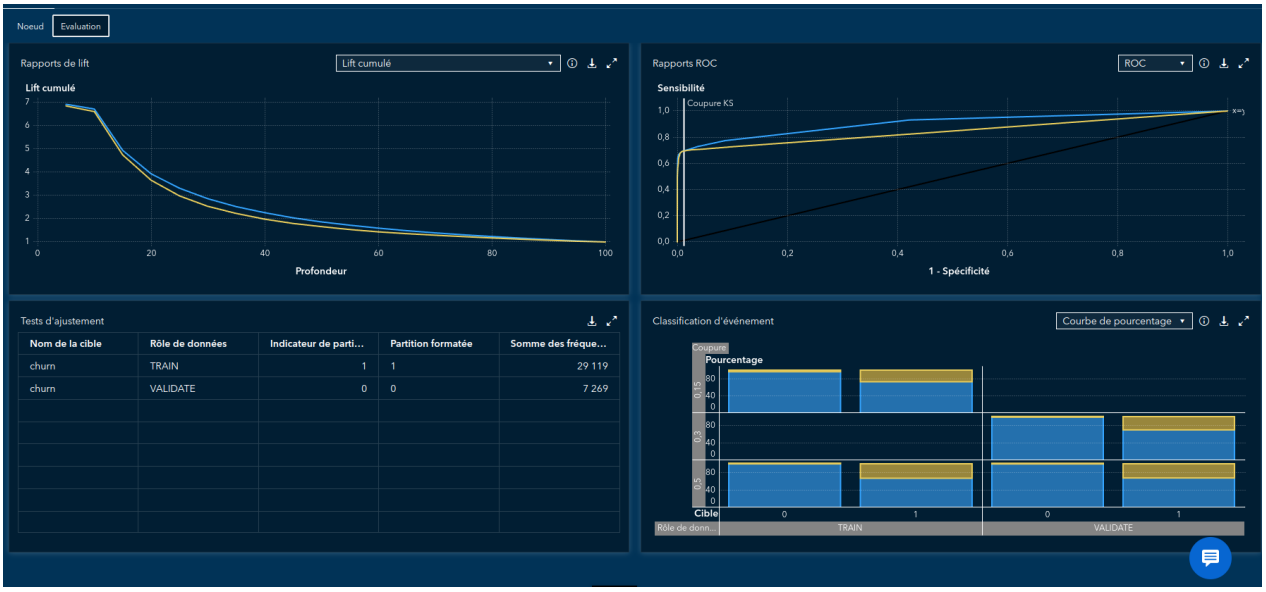


figure 14: courbes d'évaluation du modèle

On observe que la courbe ROC s'approche du coin supérieur gauche, nous pouvons dire que le modèle est capable de distinguer les deux classes.

On observe aussi que la courbe LIFT cumulée diminue lentement. Cela signifie que l'ajout de plus d'exemples positifs à votre ciblage n'améliore que très peu la performance du modèle.

V. Interprétation des résultats

1. Analyse des résultats

Le modèle que nous avons développé constitue un outil décisionnel essentiel pour l'entreprise, permettant une meilleure compréhension de sa clientèle et la mise en place de stratégies de fidélisation. Cette approche s'appuie sur l'analyse des comportements et des caractéristiques des clients. En conséquence, l'entreprise peut cibler de manière plus précise une part optimale de sa clientèle lors de ses campagnes de communication, prévoyant ainsi les désabonnements de services.

Cependant, il convient de noter que bien que le modèle actuel ait obtenu un score accepté (KS=0.68), il est impératif d'envisager une optimisation plus approfondie afin d'améliorer ses performances.

2. Détermination du ROI

On a les données suivantes :

- Coût moyenne d'une communication = 3€
- Marge moyenne = 8€
- Pourcentage des clients qui se désabonnent ~ 12.1% (cf. figure1)

Nous utilisons la formule suivante :

7269*pourcentage*(-3+8*lift_cumulé*(12.1/100))

- 7269 : représente ici les 20% de la population utilisée lors de la validation
- Le pourcentage dans notre fichier fait référence au profondeur qui signifie la proportion de la population ciblée
- Le lift cumulé indique le gain obtenu lors de la campagne

Nous obtenons le résultat suivant sur excel :

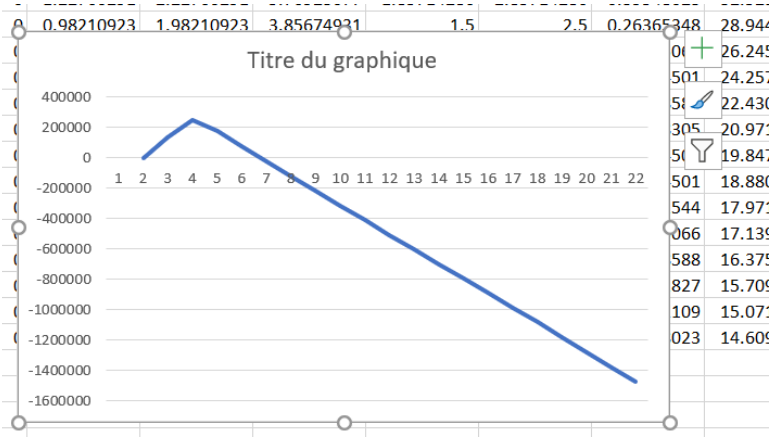
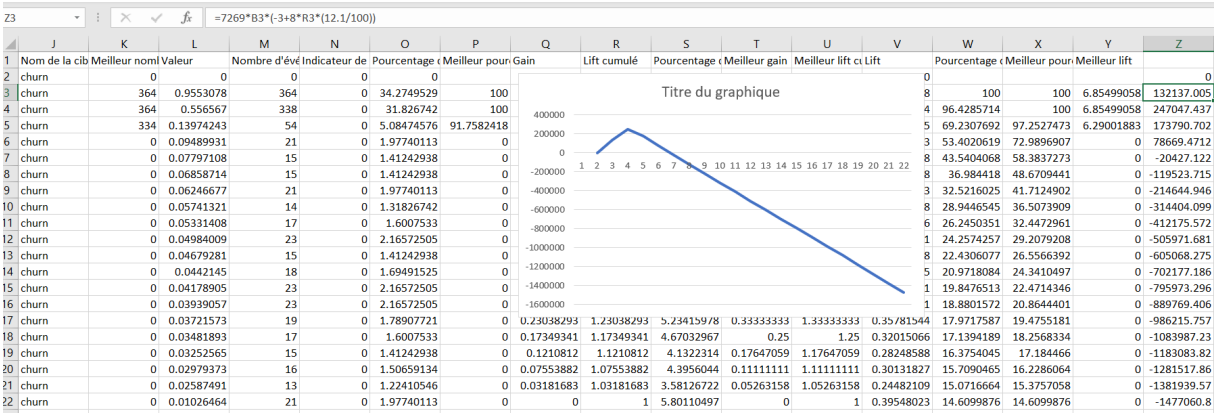


figure 15 : courbe du ROI

En guise d'interprétation nous pouvons dire qu'en ciblant environ *4%* de la population on obtient le profit maximal qui est de *247047$*.



	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Nom de la cib	Meilleur nom	Valeur	Nombre d'évé	Indicateur de	Pourcentage	Meilleur pour	Gain	Lift cumulé	Pourcentage	Meilleur gain	Meilleur lift c	Lift	Pourcentage	Meilleur pour	Meilleur lift	
2	churn	0	0	0	0	0							0				0
3	churn	364	0.9553078	364	0	34.2749529	100						8	100	100	6.85499058	132137.005
4	churn	364	0.556567	338	0	31.826742	100						4	96.4285714	100	6.85499058	247047.437
5	churn	334	0.13974243	54	0	5.08474576	91.7582418						5	69.2307692	97.2527473	6.29001883	173790.702
6	churn	0	0.09489931	21	0	1.97740113	0						3	53.4020619	72.9896907	0	78669.4712
7	churn	0	0.07797108	15	0	1.41242938	0						8	43.5404068	58.3837273	0	-20427.122
8	churn	0	0.06858714	15	0	1.41242938	0						8	36.984418	48.6709441	0	-119523.715
9	churn	0	0.06246677	21	0	1.97740113	0						3	32.5216025	41.7124902	0	-214644.946
10	churn	0	0.05741321	14	0	1.31826742	0						8	28.9446545	36.5073909	0	-314404.099
11	churn	0	0.05331408	17	0	1.6007533	0						6	26.2450351	32.4472961	0	-412175.572
12	churn	0	0.04984009	23	0	2.16572505	0						1	24.2574257	29.2079208	0	-505971.681
13	churn	0	0.04679281	15	0	1.41242938	0						8	22.4306077	26.5566392	0	-605068.275
14	churn	0	0.0442145	18	0	1.69491525	0						5	20.9718084	24.3410497	0	-702177.186
15	churn	0	0.04178905	23	0	2.16572505	0						1	19.8476513	22.4714346	0	-795973.296
16	churn	0	0.03939057	23	0	2.16572505	0						1	18.8801572	20.8644401	0	-889769.406
17	churn	0	0.03721573	19	0	1.78907721	0	0.23038293	1.23038293	5.2341597	0.33333333	1.33333333	0.35781544	17.9717587	19.4755181	0	-986215.757
18	churn	0	0.03481893	17	0	1.6007533	0	0.17349341	1.17349341	4.67032967	0.25	1.25	0.32015066	17.1394189	18.2568334	0	-1083987.23
19	churn	0	0.03252565	15	0	1.41242938	0	0.1210812	1.1210812	4.1322314	0.17647059	1.17647059	0.28248588	16.3754045	17.184466	0	-1183083.82
20	churn	0	0.02979373	16	0	1.50659134	0	0.07553882	1.07553882	4.3956044	0.11111111	1.11111111	0.30131827	15.7090465	16.2286064	0	-1281517.86
21	churn	0	0.02587491	13	0	1.22410546	0	0.03181683	1.03181683	3.58126722	0.05263158	1.05263158	0.24482109	15.0716664	15.3757058	0	-1381939.57
22	churn	0	0.01026464	21	0	1.97740113	0	0	1	5.80110497	0	1	0.39548023	14.6099876	14.6099876	0	-1477060.8

figure 16 : courbe du ROI

VI. Conclusion

Cette plongée approfondie dans les données **commsdata**, visant à décortiquer les facteurs sous-jacents au désabonnement des clients, prend une importance cruciale dans un secteur des télécommunications en perpétuelle mutation. Notre objectif de mettre en lumière les variables déterminantes dans le processus de churn ouvre de nouvelles perspectives aux entreprises du secteur pour anticiper les départs de clients et renforcer leur satisfaction.En effet, cette capacité à anticiper et retenir la clientèle revêt un rôle central dans la stratégie commerciale du domaine. À travers cette analyse approfondie des données, ce projet offre une opportunité précieuse pour mieux appréhender les tendances et les mécanismes du désabonnement, permettant aux entreprises de s'adapter efficacement aux évolutions des besoins des clients tout en maintenant leur compétitivité dans un environnement en constante mutation.