

Daniel FERREIRA
Jérémy MALUEKI
Abdelmounaim BOUZERIRA

PROJET DE DATA MINING

2022-2023

I- Définir les objectifs du projet

Décrire ce que les clients peuvent gagner à la fin de l'implémentation du projet

L'objectif de ce projet est d'offrir aux clients une analyse du taux d'attrition (Churn) et donner des éventuelles solutions pour garantir la fidélité des consommateurs afin d'améliorer le rendement général et les intérêts de l'entreprise .

II- Acquisition et exploration des données

A- Décrire la variable à expliquer

Nous nous intéressons ici sur le taux d'attrition (Churn Flag), c'est une variable binaire, qui quantifie si un client a résilié son abonnement (1) ou non (0) dans le service de télécommunication.

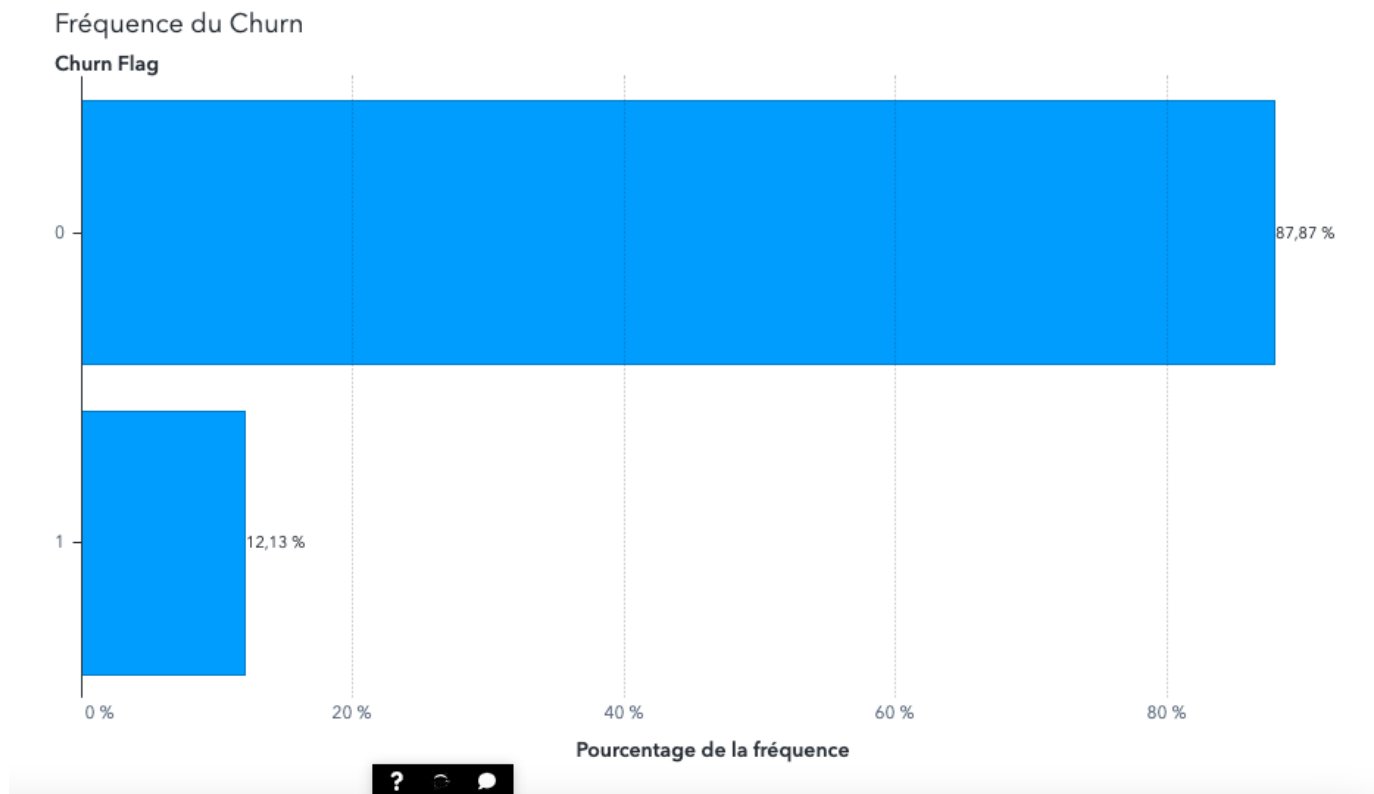
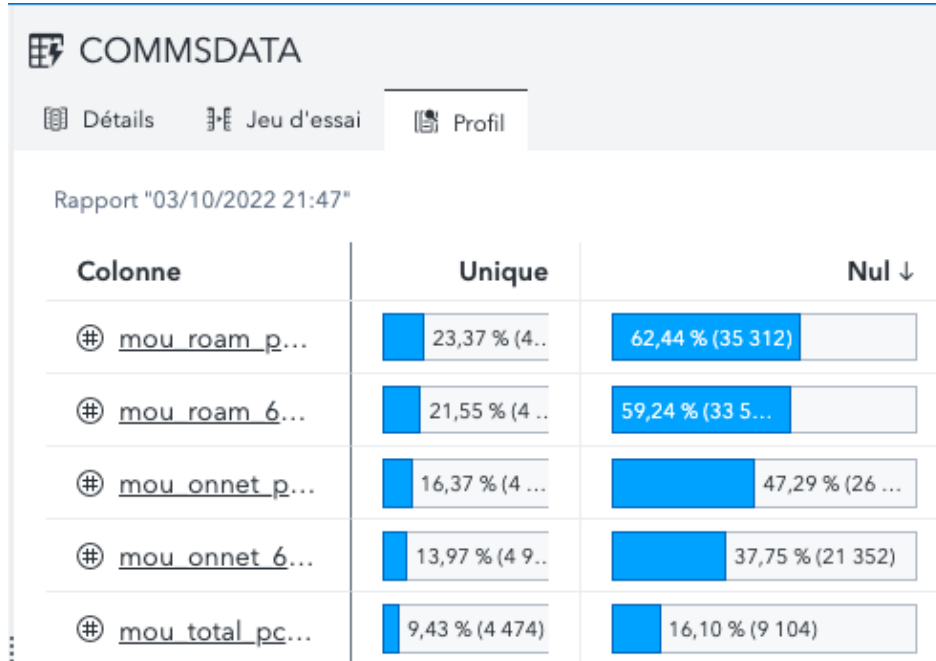


Figure 1: Nous avons 87,87% d'abonnement pour 12,13% de résiliation

B- Les variables à rejeter qui n'apportent pas au projet

Comment expliquer les variables à rejeter ?

- Les variables avec un taux de valeurs manquantes trop élevés, comme la deuxième variable (6M Avg Minutes Roaming Normally Distributed) ayant 59,24% de valeurs manquantes sont à écarter.

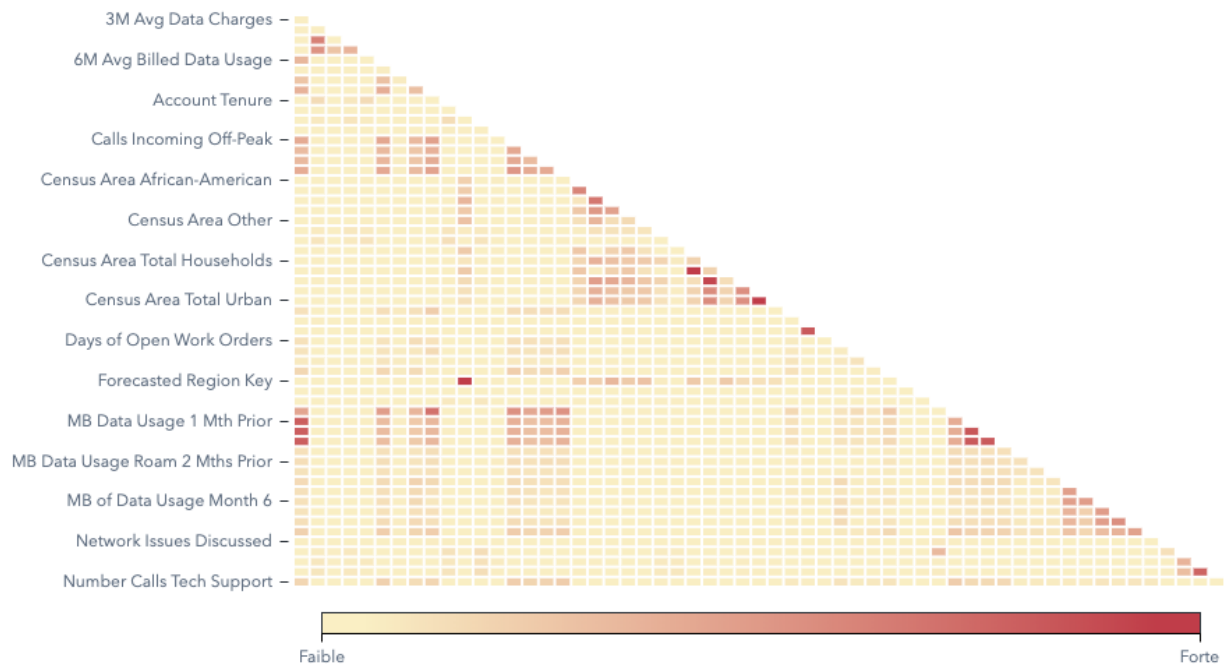


Colonne	Unique	Nul ↓
⊕ mou roam p...	23,37 % (4..)	62,44 % (35 312)
⊕ mou roam 6...	21,55 % (4 ..)	59,24 % (33 5...)
⊕ mou onnet p...	16,37 % (4 ...)	47,29 % (26 ...)
⊕ mou onnet 6...	13,97 % (4 9..)	37,75 % (21 352)
⊕ mou total pc...	9,43 % (4 474)	16,10 % (9 104)

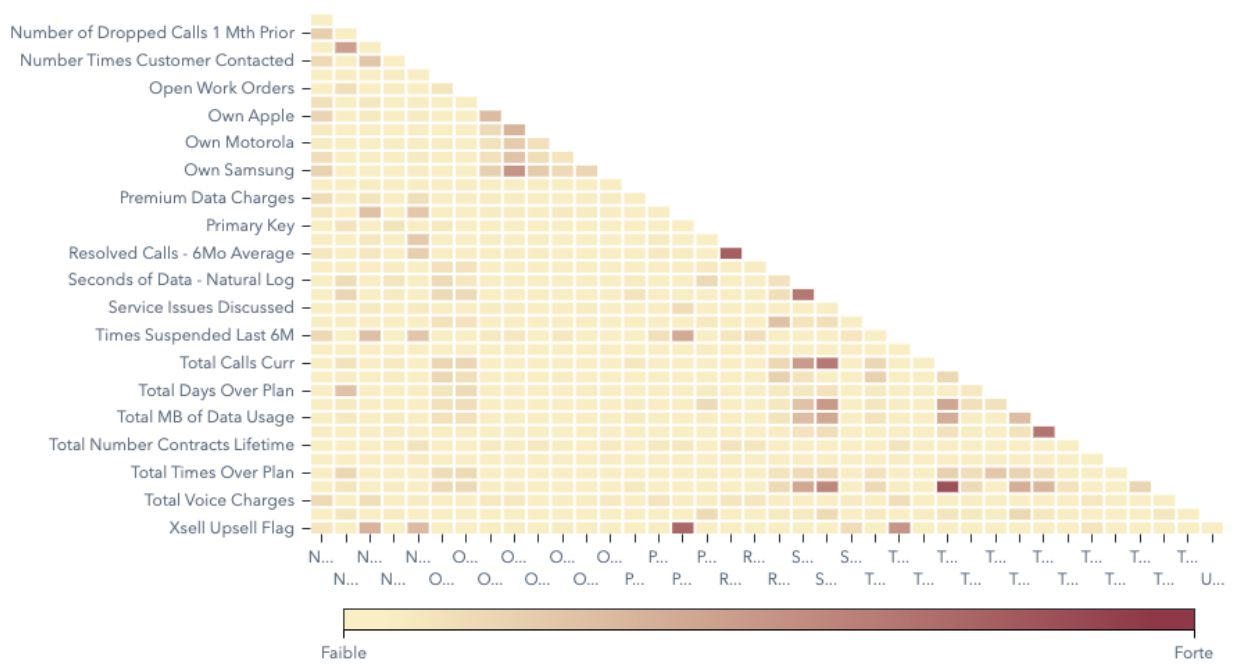
On utilise une matrice de corrélation pour analyser les variables qui peuvent nuire aux résultats lors de l'évaluation de notre modèle :

- Certaines variables peuvent renvoyer la même information (coefficient de corrélation = 1 ou -1), comme savoir si le client est un homme ou une femme, on peut garder une des deux données (comme elles renvoient la même information) pour éviter les redondances dans l'ensemble de données.
- Les valeurs fortement corrélées (> 0.90) peuvent apporter des incohérences, plus le coefficient de corrélation est élevé, plus les variables sont corrélées, donc comme pour le l'exemple d'au-dessus, on peut écarter une variable et garder l'information intacte.

Corrélation des mesures sélectionnées



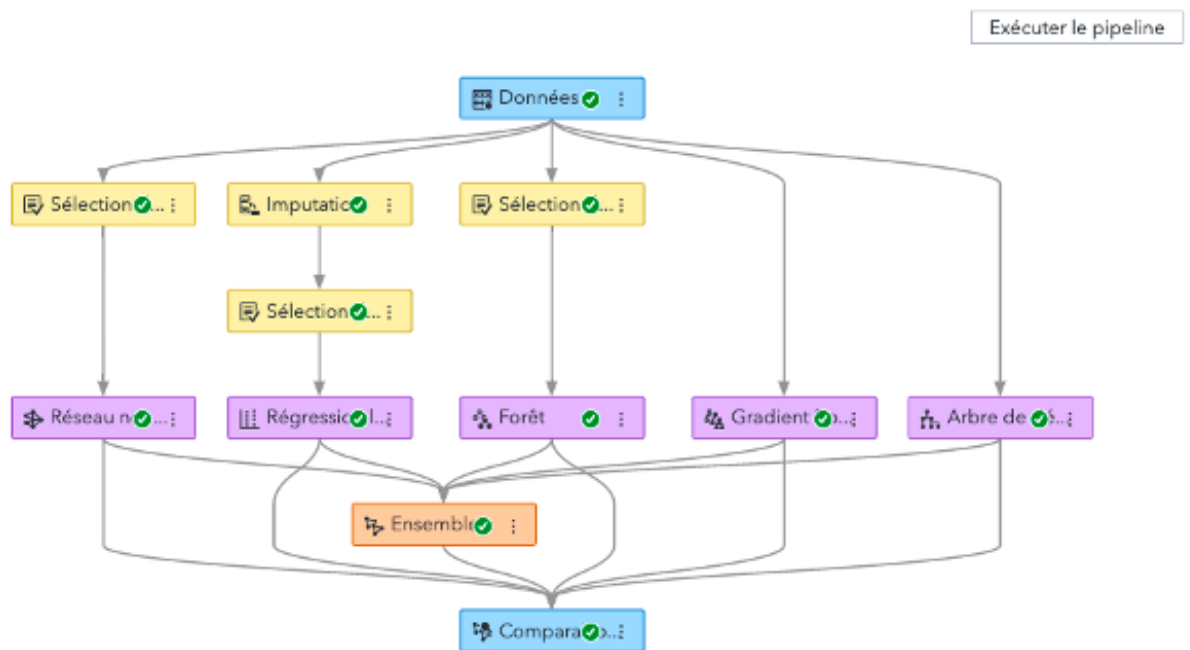
Corrélation des mesures sélectionnées



- Les données trop précises comme les données géographiques (longitude & latitude), n'offrent pas énormément d'avantage dans le cas de notre étude, donc elles sont écartables sans perdre un montant d'informations considérables.

C- Model Data

Création d'un pipeline de base.

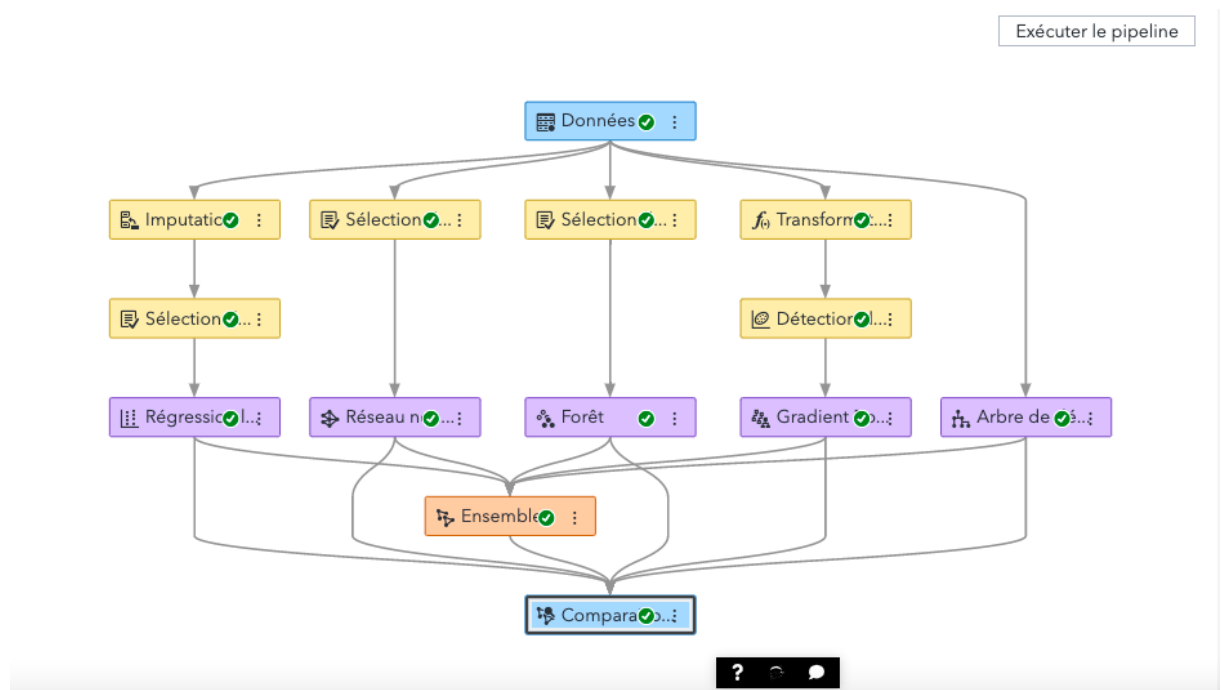


En exécutant le pipeline, on aperçoit que le modèle champion est le Gradient Boosting avec un KS (Youden) égal à 0,59.

Nom	Nom de l'algorithme	KS (Youden)	Taux de mauvaise classification
Gradient Boosting	Gradient Boosting	0,5936	0,0601
Ensemble	Ensemble	0,5869	0,0682
Forêt	Forêt	0,5755	0,0659
Régression logistique (2)	Régression logistique	0,5738	0,0697
Réseau neuronal	Réseau neuronal	0,5571	0,0778
Arbre de décision	Arbre de décision	0,3816	0,0831

Nous allons donc nous concentrer sur ce modèle en effectuant des transformations et en utilisant des outils de data mining sur les données avant l'exécution de notre modèle d'apprentissage.

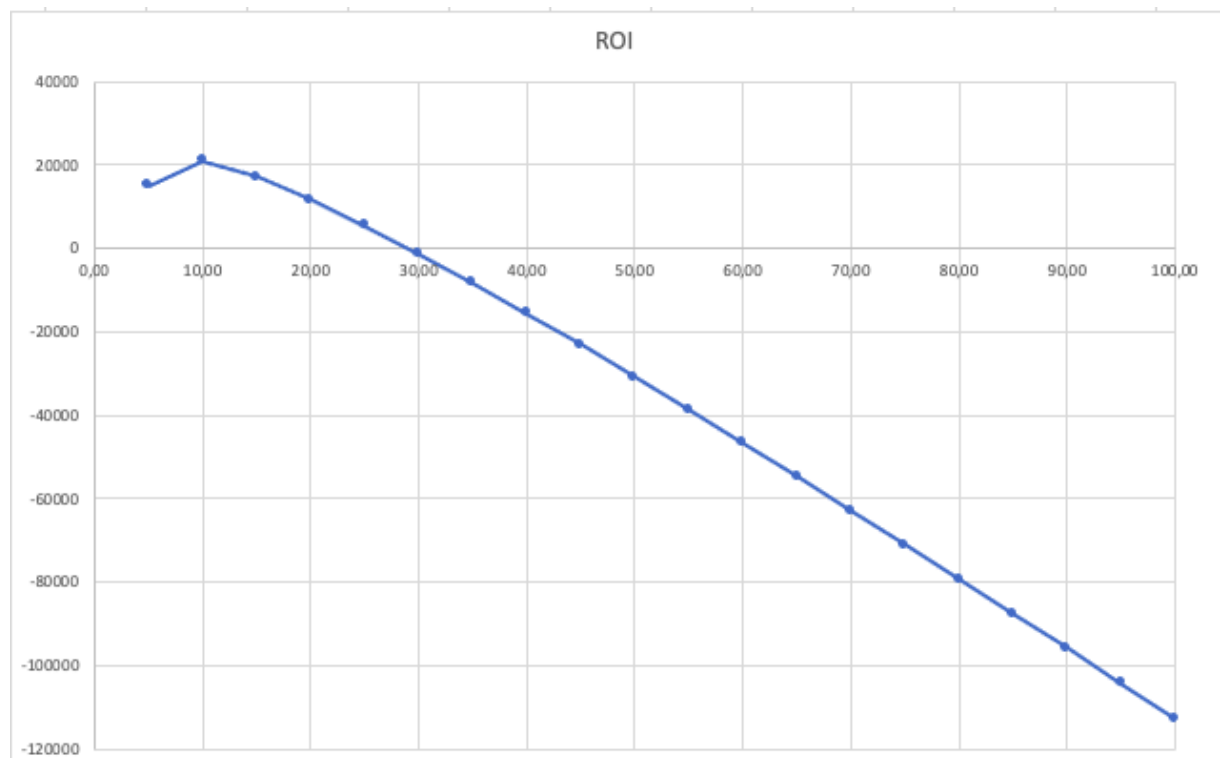
- En optimisant le modèle d'apprentissage, on va appliquer dans un premier temps, une transformation de Fourier sur nos données, puis nous effectuons une détection de potentielles anomalies qui peuvent être contenues dans notre set de données.



Nom	Nom de l'algorithme	KS (Youden)	Taux de mauvaise classification
Gradient Boosting	Gradient Boosting	0,7061	0,0532
Ensemble	Ensemble	0,5843	0,0728
Forêt	Forêt	0,5821	0,0638
Régression logistique (2)	Régression logistique	0,5738	0,0697
Arbre de décision	Arbre de décision	0,4018	0,0815
Réseau neuronal	Réseau neuronal	0	0,1213

Après amélioration de notre modèle et en ajoutant des transformations et des outils de data mining, nous avons réussi à améliorer le modèle d'évaluation de 0,59 à 0,71 KS.

Nous pouvons à présent établir la courbe du ROI grâce au modèle élaboré précédemment, nous avons un coût moyen de communication de 3€ et une marge nette moyenne de 8€.



En utilisant la courbe, on peut en déduire alors qu'il est possible de proposer des avantages à 10% à des clients qui seraient susceptible de se désabonner afin de maximiser le ROI.

On remarque qu'il est possible de se rapprocher d'une valeur max à 15% de percentile en optimisant davantage le modèle.