RAPPORT SUR LE PROJET INFORMATIQUE DÉCISIONNEL EN SAS

MODELISATION MACHINE LEARNING EN PIPELINE POUR LA PREDICTION DE L'ACHAT OU NON D'UN PRODUIT PAR LES CLIENTS SUITE A UNE PROMOTION

Réalisé par :

YABA BILONGO
Michel Davel

Encadré par :

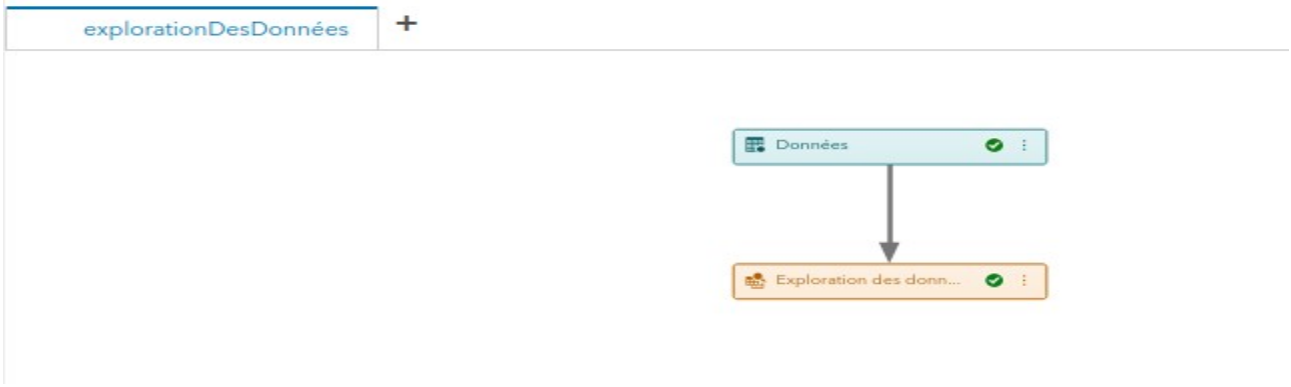
Grégroire
DE LASSENCE

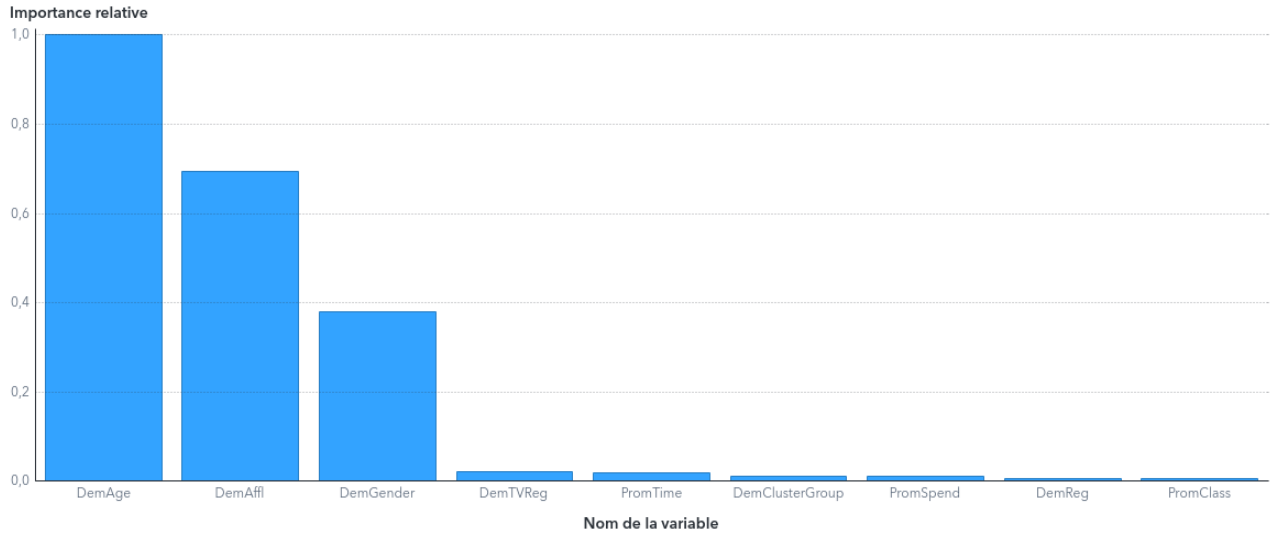
Année académique : 2020 - 2021

Organics est un supermarché britannique souhaitant lancer une nouvelle ligne de produit bio. Les données contenant ces données sont stockées dans une table appelée « Big Organics ». Notre étude, dans le cadre de ce projet, consiste à faire une modélisation Machine Learning en Pipeline afin de prédire l'achat ou non d'un produit par les clients suite à la promotion. Par la suite, nous allons trouver le modèle le plus rentable.
L'idée derrière l'implémentation des modèles Machine Learning en Pipeline est que certains modèles seront trop simples et d'autres trop complexes sur un même dataset.

Initialisation du projet
A la création du projet, on a choisi de partitionner notre dataset avec 70 % des données d'apprentissage et 30 % des données de validation. De ce fait, la performance de ce modèle sera évalué sur la partition de validation. Lors de la création du modèle, les autres paramètres sont restés inchangés.
Notre dataset est composé de 13 colonnes de 111 115 lignes . Parmi ces variables, on a 9 variables qualitatifs à savoir : TargetBuy, TargetAmt, PromClass, id, DemTVReg, DemReg, DemGender, DemClusterGroup, DemCluster et 4 variables quantitatifs à savoir DemAffl, DemAge, PromSpend, PromTime.
Trois (3) variables ont été réjetés. Il s'agit de :
- DemCluster, rejeté automatiquement parce qu'elle dépasse le nombre maximal de coupure qui est de 50.
- Id parce qu'à la base elle a pour rôle ID, d'où nous l'avons donné comme rôle « Rejeter »
- TargetAmt parce qu'elle est une variable de classe au même point que TargetBuy. Or cette analyse ne porte pas sur la quantité de produit acheté.
Après le rejet de ces 3 variables, on se retrouve avec 10 variables sur lesquels on va pouvoir faire notre analyse.

Préparation des données
Exploration des données
L'exploration des données est l'étape préalable dans la préparation car elle nous permet d'extraire les connaissances cachées des variables en utilisant à la fois les méthodes graphiques et numériques.



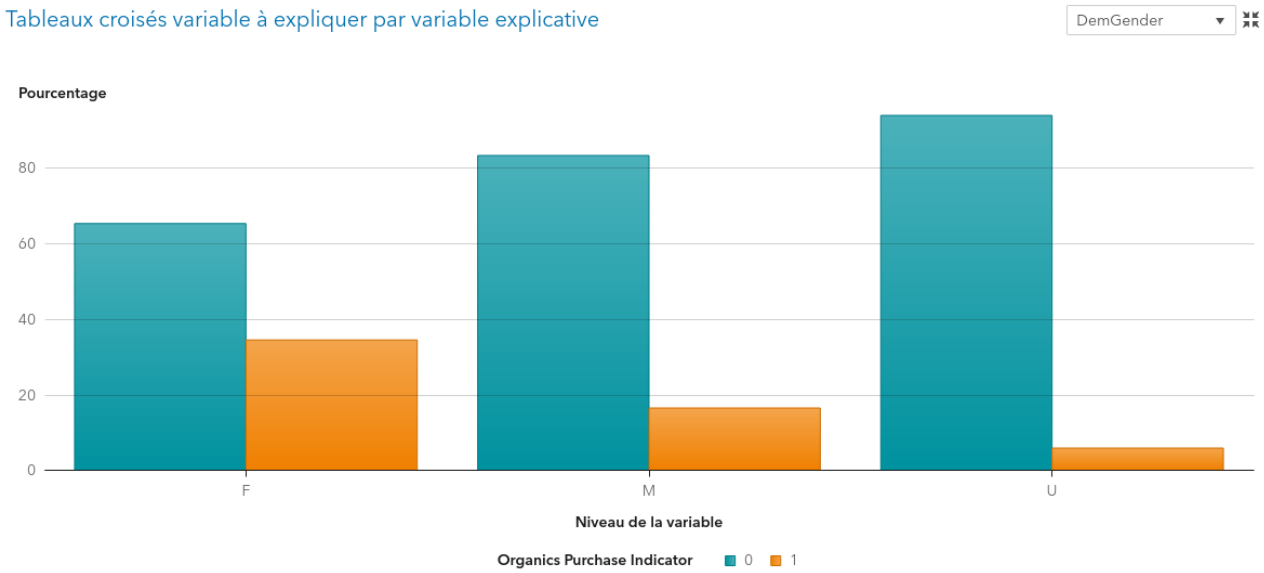


Les variables explicatives qui expliquent au mieux l'achat ou non d'un produit sont :

1- DemAge

2 – DemAffl

3 – DemGender



il y a 65 % des clients femmes qui n'achètent pas le produit bio contre 35 % qui en achètent.

il y a 83% des clients Hommes qui n'achètent pas le produit bio contre 17% qui en achètent.

il y a 94% des clients de sexe inconnu qui n'achètent pas le produit bio contre 6% qui en achètent.

On en déduit qu'il y a plus de femmes qui achètent des produits bio que les hommes ou les personnes de sexe inconnu.

Aussi on a 19 % des clients qui veulent acheter un produit contre 81 % qui n'en veulent pas.

Les méthodes numériques nous permettent d'avoir les statistiques des variables quantitatives :

Moments de variables quantitatives

Nom de la...	Minimum	Maximum	Moyenne	Ecart-type	Skewness	Kurtosis	Variabilité...	Moyenne ...	Moyenne - ...
DemAffl	0	34	8,7119	3,4211	0,8916	2,0962	0,3927	15,5540	1,8698
DemAge	18	79	53,7972	13,2058	-0,0798	-0,8440	0,2455	80,2087	27,3856
PromSpend	0,0100	296 313,850 0	4 420,5900	7 558,9115	8,0368	184,8380	1,7099	19 538,4130	-10 697,2329
PromTime	0	39	6,5647	4,6570	2,2827	8,0759	0,7094	15,8787	-2,7494

On peut constater que les valeurs minimum des variable 'DemAffl' et 'PromTime' sont de 0. D'autre part, le plus petit client a 18 ans et le plus agé en a 79.
On constate aussi que les variables 'PromSpend' et 'PromTime' ont des coefficents d'assymétrie (Skewness) au dessus de la normal soit respectivement 8,0368 et 2,2827. D'où une transformation sera nécessaire.

Cette exploration des données nous a aussi permis d'avoir les statistiques des valeurs manquantes.
On peut remarquer que seules les variables PromClass, TargetBuy et PromSpend n'ont pas de valeurs manquantes.

Valeurs manquantes

Nom de la variable	Nombre de valeurs manquantes	Pourcentage manquant
DemAffl	5425	4,8823
DemAge	7540	6,7858
DemClusterGroup	3370	3,0329
DemGender	12560	11,3036
DemReg	2325	2,0924
DemTVReg	2325	2,0924
PromClass	0	0,0000
PromSpend	0	0,0000
PromTime	1405	1,2645
TargetBuy	0	0,0000

De ce fait, une imputation sera nécessaire avant le traitement des données.

Transformations
Parmi tous les prétraitements possibles, ceux qui seront nécessaires sont : l'imputation, la transformation et la sélection des variables.

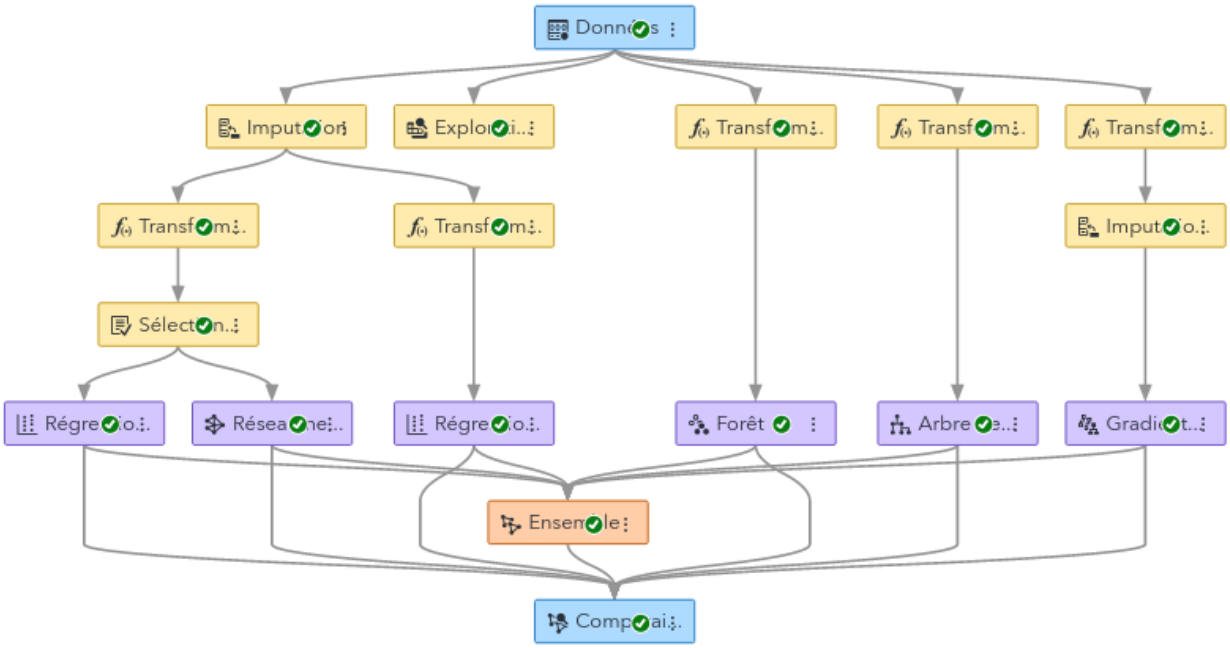
A cause des coefficient d'asymétrie élevés sur « PromSpend » et « PromTime », Nous avons choisi d'appliquer une transformation et c'est celle du type «Log » car il est plus adapté pour traiter les problèmes d'asymétrie, ce qui correspond bien au problème qu'on veut traiter ici.

Imputation

A cause des valeurs manquantes, nous allons appliquer une imputation avant l'implémentation des algorithmes Réseau de neurone, régression logistique et gradient boosting. Elle ne sera pas appliquée sur les algorithmes à base d'arbres.

Pipeline de Machine Learning

Le template utilisé dans ce cadre c'est celui du Modèle avancé sur une variable de classe.



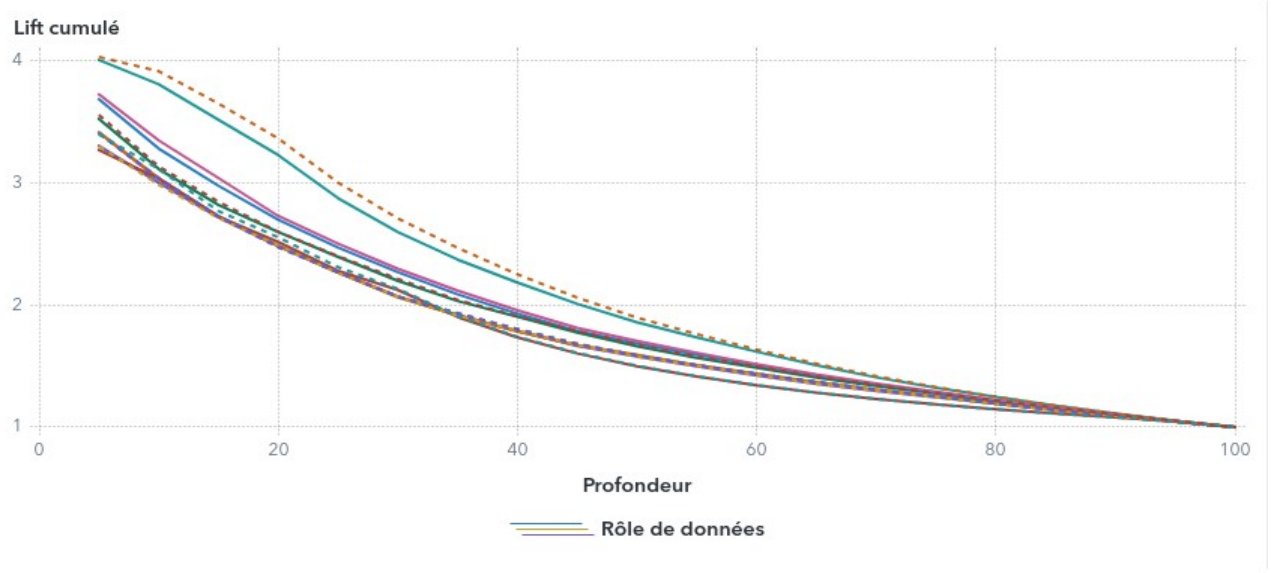
Évaluation de modèle

- **Par l'indice de Youden**

Champi...	Nom	Nom de...	KS (You...	Taux de...	Mauvai...	Racine ...	Erreur q...	Somme...	Perte lo...	Coeffici...	Zone sc
★	Forêt	Forêt	0,6381	0,1322	0,1322	0,3153	0,0994	33 334	0,3258	0,8170	0,90
	Ensemble	Ensemble	0,5069	0,1783	0,1783	0,3599	0,1296	33 334	0,4116	0,6618	0,83
	Gradient Boosting	Gradient Boosting	0,4820	0,1876	0,1876	0,3660	0,1340	33 334	0,4216	0,6342	0,81
	Arbre de décision	Arbre de décision	0,4487	0,1942	0,1942	0,3776	0,1426	33 334	0,4723	0,5216	0,76
	Régression logistique pas à pas	Régression logistique	0,4260	0,1955	0,1955	0,3776	0,1426	33 334	0,4479	0,5749	0,78
	Réseau neuronal	Réseau neuronal	0,4256	0,1984	0,1984	0,3815	0,1455	33 334	0,4556	0,5674	0,78
	Régression logistique ascendante	Régression logistique	0,4237	0,1957	0,1957	0,3776	0,1426	33 334	0,4479	0,5744	0,78

D'après la statistique du KS (Youden), le meilleur modèle dans notre cas est la Forêt.

- **Par la courbe du levier (Lift)**



Sur la courbe du levier (lift), c'est aussi la Forêt.

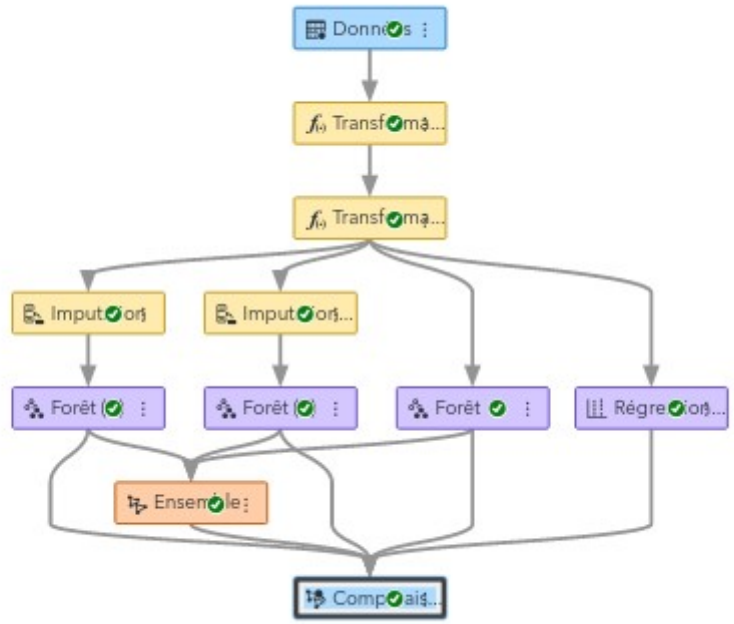
En positionnant la souris sur le point le plus haut, on peut lire que la Forêt a un lift de 4,0063 à 5%. C'est-à-dire que si l'on sélectionne les 5% ayant la probabilité la plus forte d'acheter d'après notre modèle, on peut espérer multiplier le taux de retour de base par 4,0063.

Il y a 19 % des clients qui veulent acheter un produit dans notre base. Donc, si l'on sélectionne les 5% des clients qui veulent acheter un produit, 4,0063 x 19 = 76,1197 % des clients devraient acheter.

Pipeline généré automatiquement

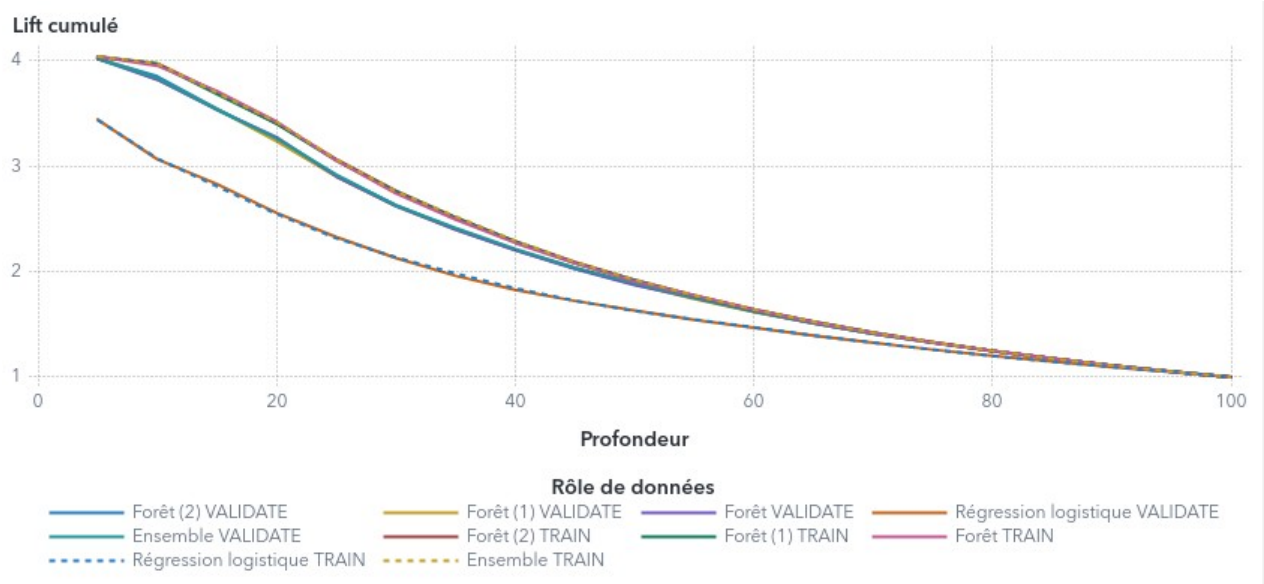
Quand on génère le pipeline automatiquement on obtient les résultats presque semblables mais à quelques différences près. En effet, contrairement au Pipeline que nous avons construit en partant du template choisi, le pipeline généré automatiquement nous propose un ensemble de 3 modèles de type Forêt et un modèle de type régression logistique.

Voici les résultats obtenus :



Comparaison de modèles

Champion	Nom	Nom de l'algorithme	KS (Youden)	Taux de mauvaise classification
★	Ensemble	Ensemble	0,6530	0,1294
	Forêt (2)	Forêt	0,6512	0,1294
	Forêt	Forêt	0,6486	0,1282
	Forêt (1)	Forêt	0,6481	0,1304
	Régression logistique	Régression logistique	0,4520	0,1885



D'après Youden c'est l'ensemble des 3 modèles de Forêt qui évalue au mieux notre dataset. Mais d'après le lévier de Lift c'est le modèle Forêt(1) qui donne les meilleurs résultats soit 4,0233.

Donc, si l'on sélectionne les 5% des clients qui veulent acheter un produit, 4,0063 x 19 = 76,4427 % des clients devraient acheter. Ce qui n'est pas loin des résultats obtenus précédemment.

Comparaison des pipelines

Champion	Nom	Nom de l'algorithme	Nom du pipeline	KS (Youden)	Somme des fréquence
☑	Ensemble	Ensemble	Pipeline 2	0,653	33 334
☐	Forêt	Forêt	Pipeline 1	0,638	33 334

En comparant le premier Pipeline au second, nous constatons que le Pipeline 2 a un meilleur score Youden que le Pipeline 1.

En définitive, ce projet nous a permis de nous familiariser sur la modélisation des données en SAS Viya 3.5. On a travaillé sur le dataset BigOrganics contenant plus de 111115 données pour 13 variables dont 3 rejetés. Nous avons commencé notre modélisation par un travail d'exploration ; ce qui nous a permis de mieux comprendre le dataset. Ce travail d'exploration nous a permis notamment de ressortir les variables les plus importants, les valeurs manquantes de chacune des variables, les statistiques et les distributions de chacune des variables, la prédiction de la variable expliquée en fonction de toutes les autres variables. Par la suite, nous avons fait un prétraitement sur les données. Il s'agit notamment des tâches de transformations et d'imputation sur certains modèles. Une fois ces étapes préalables terminées, nous avons construit un pipeline de modèles en

utilisant un template de modèle avancée sur une variable de classe cible. L'exécution de ce pipeline nous a permis de déduire la Forêt comme le modèle qui s'ajuste au mieux sur ce dataset. Ensuite, nous avons généré un pipeline automatique. En comparant les 2 pipelines, le pipeline généré automatiquement donne un meilleur score. Nous en déduisons donc que le modèle qui s'adapte le mieux à ce dataset c'est un modèle constitué de 3 forêts.