Informatique décisionnelle

Machine Learning Pipelines sur SAS Model Studio

BigOrganics

Réalisé par :

Aghilas KESSAÏ

Etudiant en M2 EID²

2021-2022

Table des matières

I- Le but du projet

Ce projet sur SAS Model Studio a pour but de retrouver le modèle plus rentable sur la table BigOrganics pour une entreprise qui cherche à se lancer sur une campagne publicitaire ou autre. On cherche alors à trouver le meilleur score KS Youden et donc le meilleur lift qui se traduit par le meilleur ROI (Return Of Investment).

On utilise donc les modèles afin d'ordonner les clients allant de la probabilité la plus forte d'acheter les produits bio à la probabilité la plus faible. Ainsi, ce projet permettrait à l'entreprise de sélectionner un sous-groupe de ses clients afin de cibler leur marketing sur les clients qui seraient les plus susceptibles d'acheter les nouveaux produits.

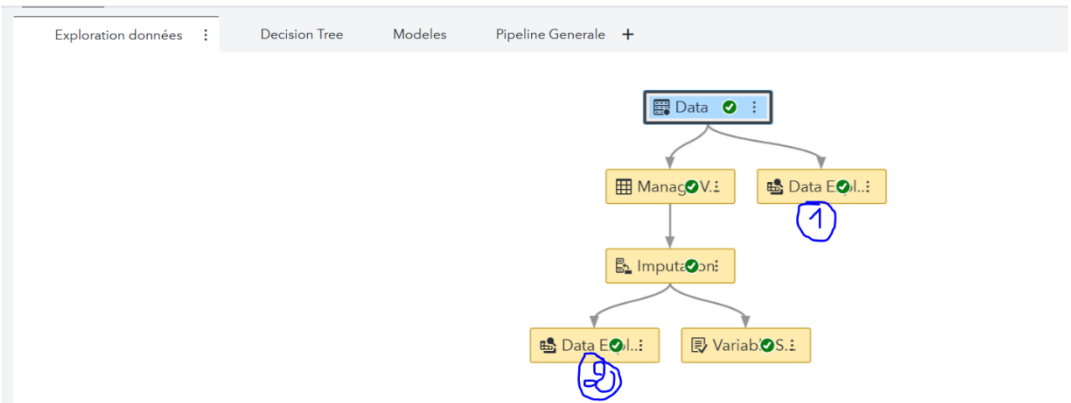
II- Les données de la table BigOrganics

La table BigOrganics disponible sur SAS Model Studio contient 13 colonnes (variables) et 111 115 lignes (clients). On a plusieurs informations allant de sexe du client, âge, … Ainsi, notre **target** variable pour ce projet sera TargetBuy (Organics Purchase Indicator). Pour l'instant nous allons mettre en **rejected** la variable TargetAmt (cette variable ne va pas aider le modèle à apprendre, étant donné que c'est une variable dépendante). Pour les autres variables on les met tous en input. Cependant, nous allons utiliser le node de '**Variable Selection**' afin de voir quelles variables on met en input ou rejected avant d'ajouter un modèle en child node.

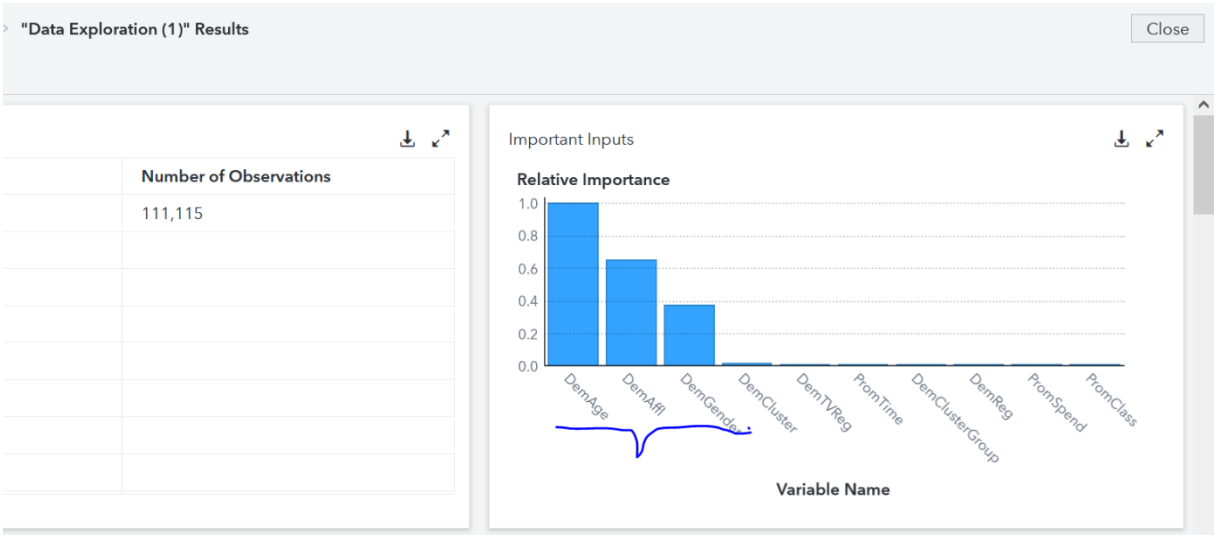
III- Les différents pipelines

D'abord, on fait un premier pipeline simple, avec une exploration de données. On découvre donc les données avec le node 'Data Exploration'. Ainsi, on a différentes informations sur les valeurs manquantes. En général, on essaye d'avoir un node 'Manage Variables' dans les pipelines afin de pouvoir modifier les données dans chaque différent pipeline.

Voici le pipeline de l'exploration de données. On utilise le node 'Data Exploration' avant l'imputation et après afin de voir les différences.

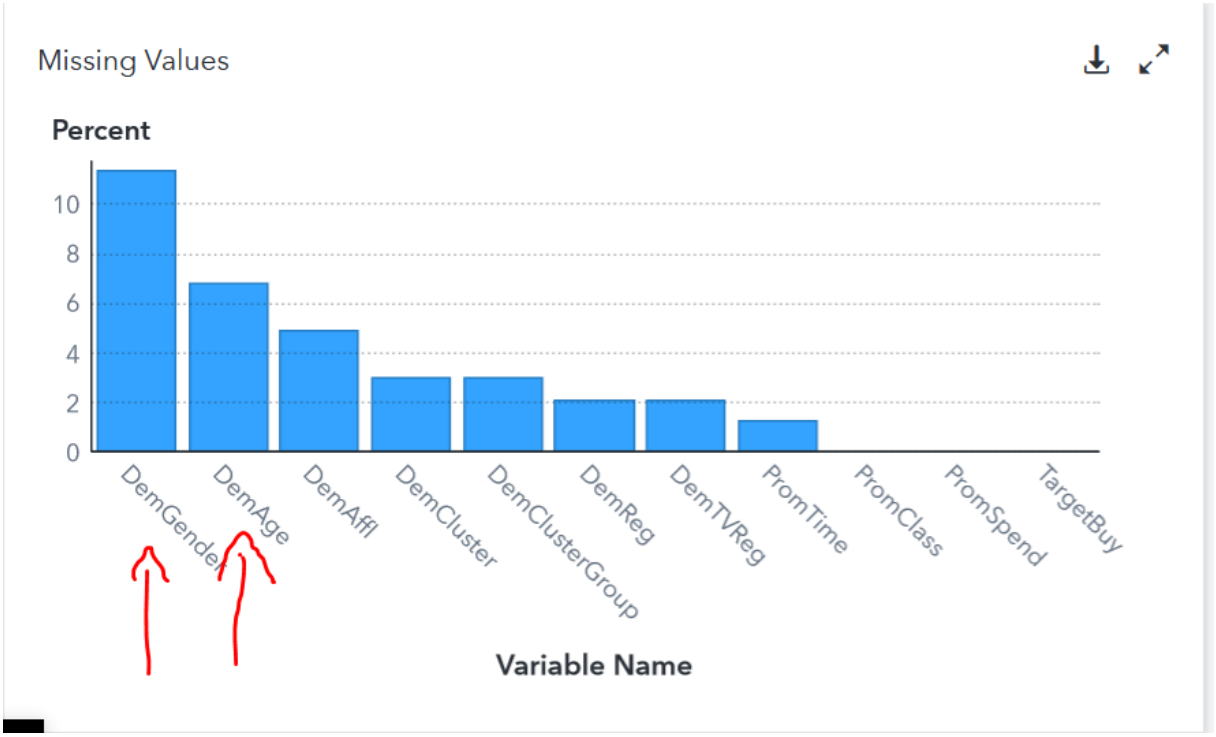


Voici le résultat du node ‘data exploration’ (1) en haut à droite :

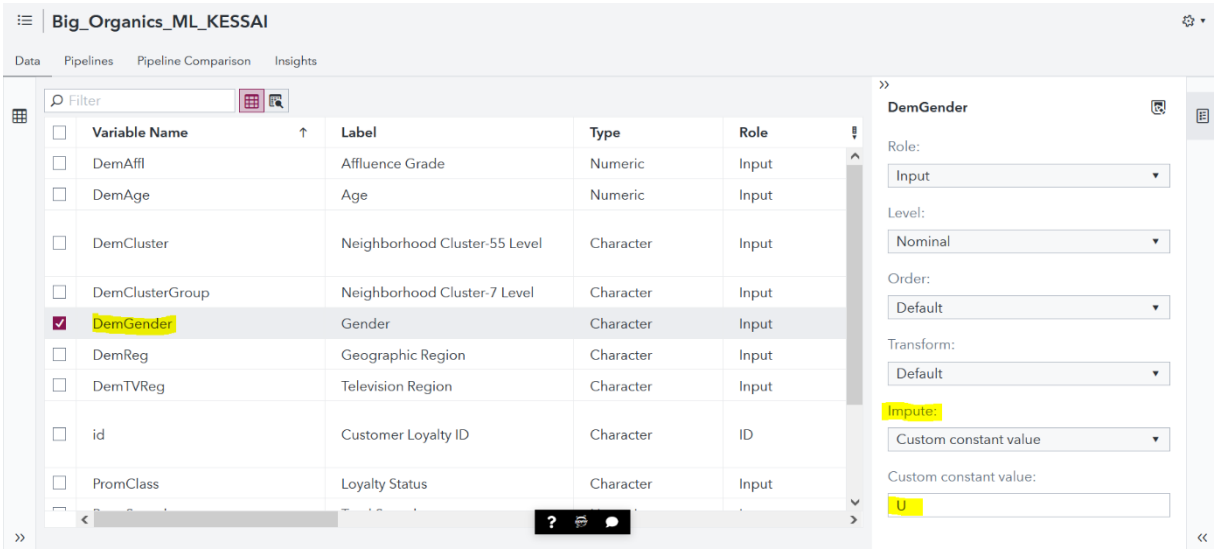


On voit que les variables ‘DemAge’, ‘Dem Affl’, ‘DemGender’ sont importantes.

Or, parmi 2 parmi 3 de ces variables importantes ont un taux de valeurs manquantes assez conséquent. Ainsi, pour le genre on décide de faire quelque chose.

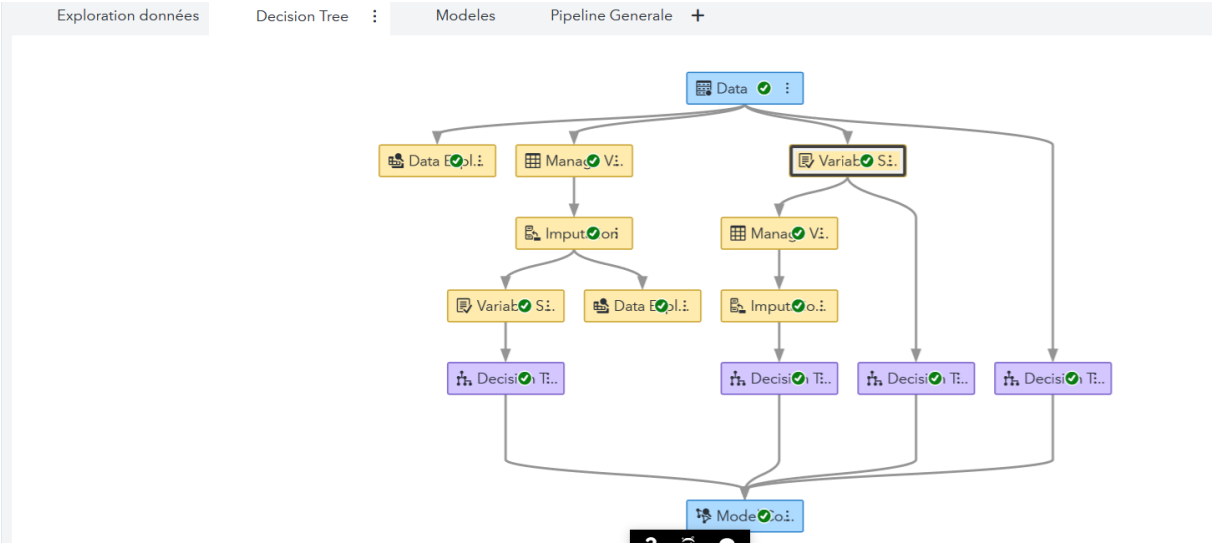


Ainsi, dans l’onglet Data on fixe une méthode d’imputation personnalisée qui va fixer toutes les valeurs manquantes de Gender à U (Unknown) :



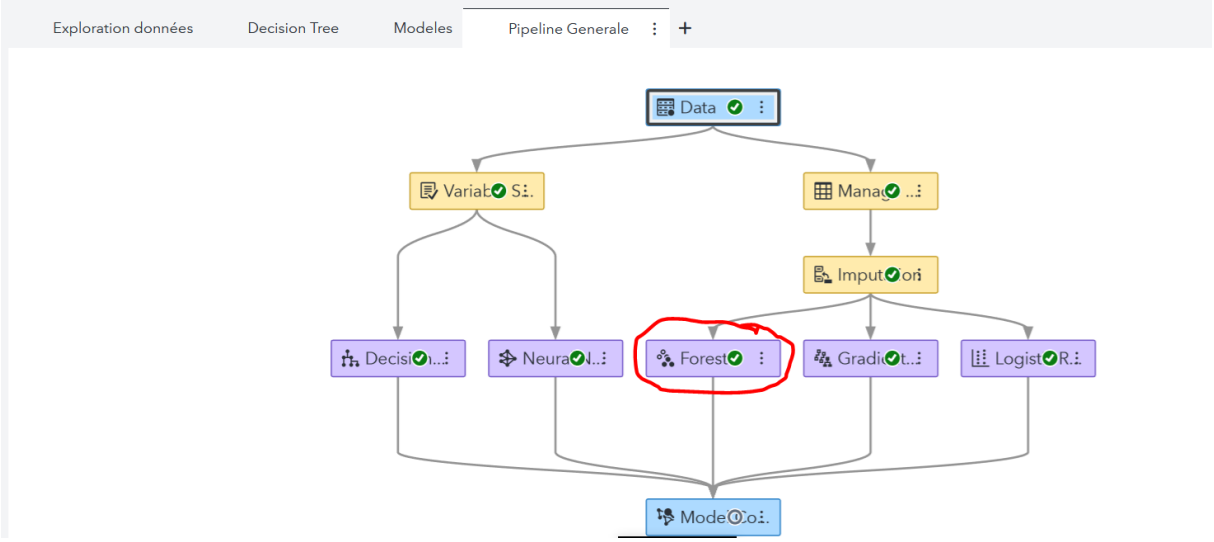
Pour l'âge, on choisit la méthode de winsorized mean mais on choisi cela dans les node de Manage Variables (car on pourrait, dans d'autres pipelines utiliser des méthodes d'imputation différentes).

On va par la suite utiliser différentes méthodes de pré-traitement des données afin de tester les modèles, par exemple :



Ainsi, on essaye de trouver la meilleure méthode de prétraitement (plus haut score KS Youden) , avec imputation, avec sélection de variables…

On regroupe cela dans un pipeline général :

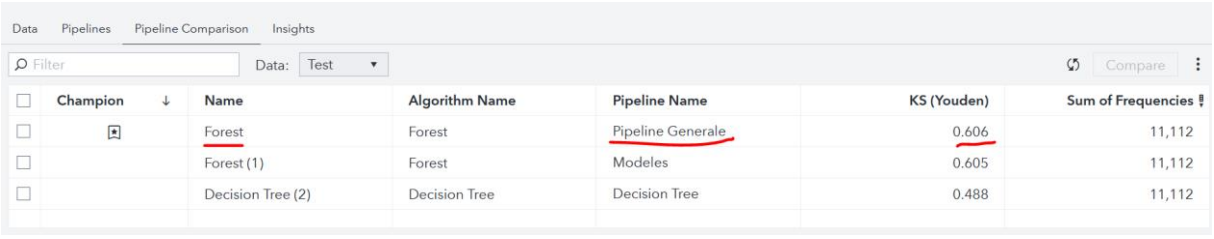


Ainsi, on a le meilleur score KS avec la Forêt entourée en rouge. Voici les résultats liés à ce pipeline :

Model Comparison

Champion	Name	Algorithm Name	KS (Youden)	Misclassification Rate
	Forest	Forest	0.6063	0.1365
	Gradient Boosting	Gradient Boosting	0.5238	0.1760
	Decision Tree	Decision Tree	0.4882	0.1797
	Logistic Regression MV IMpute	Logistic Regression	0.4493	0.1897
	Neural Network	Neural Network	0.0584	0.2478

Le score KS est assez élevé, il est de 0.6063.



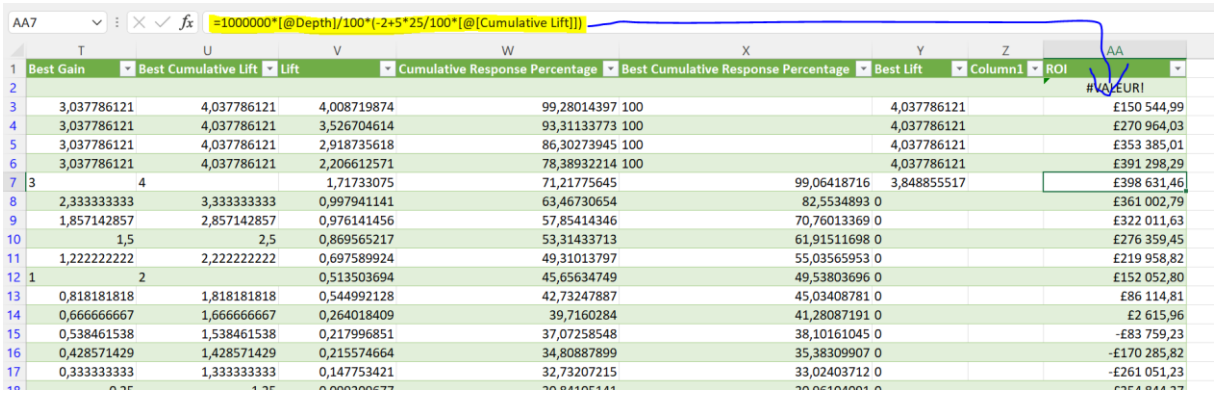
Logiquement, on le retrouve dans la comparaison de pipelines.

On ouvre la page des résultats liés à la forêt, on va dans ‘Assessement’ et on télecharge les données lié au cumulative lift.

IV - Analyse

On ouvre le fichier Excel de ces données là et on ajoute ensuite une colonne du ROI tout en gardant seulement les lignes dont le data rôle est ‘VALIDATE’.

Voici la formule du calcul du ROI :



	T	U	V	W	X	Y	Z	AA
1	Best Gain	Best Cumulative Lift	Lift	Cumulative Response Percentage	Best Cumulative Response Percentage	Best Lift	Column1	ROI
2								#VALEUR!
3	3,037786121	4,037786121	4,008719874	99,28014397	100	4,037786121		£150 544,99
4	3,037786121	4,037786121	3,526704614	93,31133773	100	4,037786121		£270 964,03
5	3,037786121	4,037786121	2,918735618	86,30273945	100	4,037786121		£353 385,01
6	3,037786121	4,037786121	2,206612571	78,38932214	100	4,037786121		£391 298,29
7	3	4	1,71733075	71,21775645	99,06418716	3,848855517		£398 631,46
8	2,333333333	3,333333333	0,997941141	63,46730654	82,5534893	0		£361 002,79
9	1,857142857	2,857142857	0,976141456	57,85414346	70,76013369	0		£322 011,63
10	1,5	2,5	0,869565217	53,31433713	61,91511698	0		£276 359,45
11	1,222222222	2,222222222	0,697589924	49,31013797	55,03565953	0		£219 958,82
12	1	2	0,513503694	45,65634749	49,53803696	0		£152 052,80
13	0,818181818	1,818181818	0,544992128	42,73247887	45,03408781	0		£86 114,81
14	0,666666667	1,666666667	0,264018409	39,7160284	41,28087191	0		£2 615,96
15	0,538461538	1,538461538	0,217996851	37,07258548	38,10161045	0		-£83 759,23
16	0,428571429	1,428571429	0,215574664	34,80887899	35,38309907	0		-£170 285,82
17	0,333333333	1,333333333	0,147753421	32,73207215	33,02403712	0		-£261 051,23

La formule est : 1 000 000 personnes * x_centiles * (-2 + 5*25% * Cumulative LIFT)

Explication : on a 1 000 000 de clients et on sélectionne x_centiles parmi eux (x_centiles = 5% par exemple) et que le coût de la communication (publicité ou autre) est de 2£ et on suppose qu’elle nous rapporte 5£. Le 25% est le pourcentage attendu d’acheteurs de produits bio.

Or, grâce à notre modèle de Forêt, on a un certain LIFT. C’est-à-dire que si on sélectionne un certain pourcentage des clients classés en fonction de la probabilité d’acheter, alors on pourra améliorer ce pourcentage de 25% par 25%*LIFT. Ainsi dans cette population sélectionnée, les clients seraient plus susceptibles d’acheter comparé à la tendance générale.

Le ROI nous donne donc le bénéfice net qu’on pourrait atteindre sur un sous ensemble des 1 000 000 clients.

AA
ROI
#VALEUR!
£150 544,99
£270 964,03
£353 385,01
£391 298,29
£398 631,46
£361 002,79
£322 011,63
£276 359,45
£219 958,82
£152 052,80
£86 114,81
£2 615,96
-£83 759,23
-£170 285,82
-£261 051,23
-£354 844,37
-£451 665,25
-£550 756,93
-£650 756,93

Voici les résultats du ROI :

On voit ici qu’on peut faire un bénéfice de 398 631 £ si on sélectionne les 25 % des clients qui sont le plus susceptibles d’acheter.

Comment cela va se passer ? Grâce au modèle on va pouvoir classer les clients

en fonction de leur probabilité d'acheter un produit bio. Par exemple sur 100 clients :

Client 1 : PROBA 70%

Client 2 : PROBA 67%

Client 3 : PROBA 65%

Client 4 : PROBA 63%

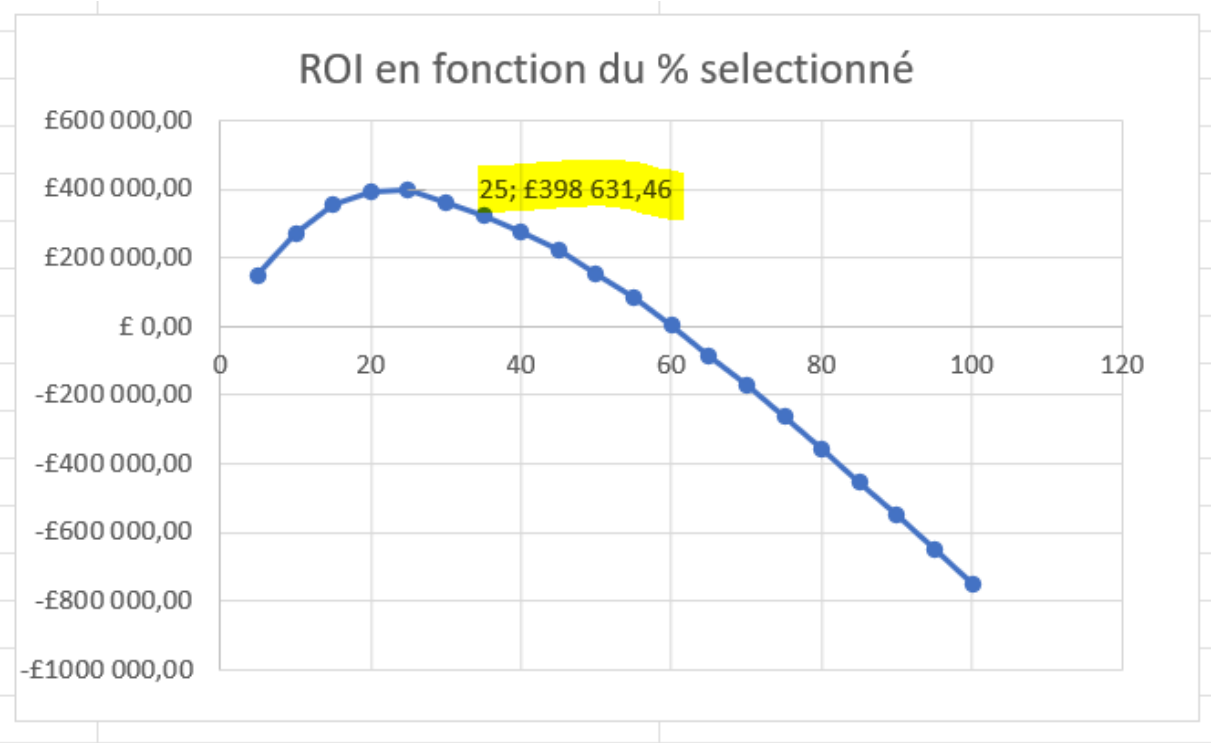
Client 5 : PROBA 60%

Client 6 : PROBA 59,8%

…

Client 100 : PROBA 23%

Ainsi, on va sélectionner les 25 premiers clients de ce classement (25%) afin de maximiser les gains.



Comme le montre ce graphique, on pourra faire un bénéfice net de 398 631£ si on sélectionne les 25% ayant la probabilité la plus forte d'acheter, classés grâce à notre modèle.

V- Conclusion

En choisissant notre modèle réalisé sur SAS Model Studio, qui nous permettrait de classer les clients allant de la probabilité la plus forte d'acheter à la probabilité la plus faible, on pourrait avoir le meilleur ROI de 398 631£. On sélectionnera 25% des meilleurs clients classés. C'est ainsi qu'on pourra maximiser les gains liés à la campagne publicitaire.