RAPPORT FINAL — PROJET

Cursus Data Analyst, Formation continue, Promotion avril 2023

Sujet

Prédiction du succès d'une campagne Marketing d'une banque

Participants

Nour Becam

Ludovic Durand

Stephane Kobenan

Simon Martinez

Table des matières

I. Cadrage du projet

A. Contexte du projet

Nous allons nous imaginer data analyst pour une banque qui souhaite améliorer sa stratégie de marketing pour les dépôts à terme. La banque propose des dépôts à terme aux clients, leur offrant des intérêts attractifs sur une période déterminée.

L'objectif global est d'optimiser les revenus et de renforcer la relation avec les clients.

Notre objectif est donc d'identifier les facteurs qui influencent la décision des clients de souscrire à un dépôt à terme afin de cibler plus efficacement les campagnes de marketing et d'augmenter le taux de souscription.

B. Périmètre du projet

Dans ce projet, nous allons analyser un jeu de données contenant les informations sur des clients de la banque et étudier les variables disponibles pour prédire la souscription au dépôt à terme. Nous allons concentrer notre analyse sur les 17 variables de notre dataset qui sont pour analyser les données disponibles pour comprendre les facteurs qui influencent la souscription aux dépôts à terme et à formuler des recommandations pour améliorer l'efficacité de la campagne marketing.

C. Aperçu du Dataset

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	deposit
0	59	admin.	married	secondary	no	2343	yes	no	unknown	5	may	1042	1	-1	0	unknown	yes
1	56	admin.	married	secondary	no	45	no	no	unknown	5	may	1467	1	-1	0	unknown	yes
2	41	technician	married	secondary	no	1270	yes	no	unknown	5	may	1389	1	-1	0	unknown	yes
3	55	services	married	secondary	no	2476	yes	no	unknown	5	may	579	1	-1	0	unknown	yes
4	54	admin.	married	tertiary	no	184	no	no	unknown	5	may	673	2	-1	0	unknown	yes

- **Age :** l'âge de l'individu, généralement mesuré en années.
- **Job :** la categorie de métier ou la profession de l'individu.
- **Marital :** l'état matrimonial de l'individu, tel que marié, célibataire, divorcé, etc.

- **Education :** le niveau d'éducation atteint par l'individu, comme l'école primaire, le collège, le lycée, l'université, etc.
- **Default :** indique si l'individu a déjà fait défaut sur un prêt ou une dette.
- **Balance :** le solde du compte bancaire de l'individu.
- **Housing :** indique si l'individu possède un prêt hypothécaire ou non.
- **Loan :** indique si l'individu a un prêt personnel ou non.
- **Contact :** le mode de contact utilisé pour communiquer avec l'individu, comme le téléphone, le courrier électronique, etc.
- **Day :** le jour du mois où le dernier contact a été établi avec l'individu.
- **Month :** le mois de l'année où le dernier contact a été établi avec l'individu.
- **Duration :** la durée en secondes du dernier contact avec l'individu.
- **Campaign :** le nombre de contacts effectués lors de la campagne publicitaire ou marketing.
- **Pdays :** le nombre de jours écoulés depuis le dernier contact avant la campagne actuelle (-1 signifie que le client n'a pas été contacté auparavant).
- **Previous :** le nombre de contacts effectués avant la campagne actuelle.
- **Poutcome :** le résultat de la campagne marketing précédente.
- **Deposit :** indique si l'individu a souscrit à un dépôt à terme ou non.

D. Objectif commercial

Réduire les ressources de marketing en identifiant les clients qui souscriraient au dépôt à terme et, par le fait même, en faisant du marketing direct.

II. Exploration des données - Observations principales

A. Variable cible et variables explicatives, première approche du jeu de données

“Deposit” est la variable cible. Il s’agit d’une variable binaire. Les autres variables du dataset sont les variables explicatives. Techniquement, il n’y a aucune valeur manquante. Cependant, on verra par la suite que la modalité “unknown” est présente pour plusieurs variables et qu’il nous faudra traiter ces cas.

B. Pertinence des variables explicatives

L’âge

L'âge d'une personne est un facteur déterminant dans ses besoins financiers et ses objectifs à différents stades de la vie. Les personnes plus jeunes peuvent être plus intéressées par des produits tels que les comptes d'épargne, les prêts étudiants ou les produits d'investissement à long terme, tandis que les personnes plus âgées peuvent rechercher des produits de retraite, des assurances ou des prêts hypothécaires. De plus, l'âge peut être lié à la stabilité financière, à l'expérience de gestion des finances personnelles et à la tolérance au risque.

Le métier

Le métier et a posteriori, la catégorie socio-professionnnelle peut fournir des indication sur le revenu, la stabilité financière et la propension à prendre des risques. Certaines professions peuvent être associées à des revenus plus élevés, ce qui peut augmenter la capacité de souscrire à des produits financiers plus avancés. De plus, certaines professions peuvent nécessiter une meilleure gestion financière et une compréhension des produits bancaires, ce qui peut influencer la décision de souscrire.

Le statut marital

Le statut marital peut avoir un impact sur les responsabilités financières d'une personne, sa stabilité économique et ses objectifs financiers. Le statut marital peut également refléter une plus grande maturité financière et une planification à long terme, ce qui peut influencer les décisions de souscription.

Le niveau d'éducation

Le niveau d'éducation peut être un indicateur de la stabilité financière, de la maîtrise des compétences financières et de la compréhension des produits bancaires. Les personnes ayant un niveau d'éducation plus élevé peuvent être plus conscientes des avantages de la planification financière, de l'épargne et des investissements. Elles peuvent également avoir une meilleure compréhension des risques et des avantages liés à certains produits financiers, ce qui peut influencer leur volonté de souscrire à des produits bancaires.

Défaut

Le fait qu'une personne ait déjà fait défaut à rembourser des prêts ou des dettes peut être un indicateur de sa capacité à gérer ses obligations financières.

Le solde bancaire

Le solde bancaire d'une personne peut indiquer sa capacité à épargner et à gérer ses finances.

Prêt immobilier en cours

Le fait de posséder un bien immobilier peut indiquer une certaine stabilité financière et un niveau d'engagement financier plus élevé.

Autre prêt en cours

Le fait d'avoir un crédit en cours peut influencer la capacité d'une personne à prendre de nouveaux engagements financiers. Cela peut montrer que la personne a une bonne expérience dans la gestion de ses obligations financières, dans le cas où il n'aurait pas fait défaut à ses obligations.

Mode de contact

Si une modalité e-mail existait, elle aurait pu être intéressante à étudier et avec une certaine pertinence métier puisqu'on imagine des performances commerciales différentes selon le média utilisé pour la prospection.

Day

Le jour du mois pendant lequel un client est contacté peut influencer la disponibilité, l'état mental et les priorités financières de la personne. On peut imaginer que ces paramètres évoluent au cours du mois.

Mois

Certains mois de l'année sont associés à des événements saisonniers ou des périodes de dépenses spécifiques. Par exemple, les mois de novembre et décembre peuvent être marqués par les fêtes de fin d'année et les dépenses liées aux cadeaux.

Certains mois de l'année peuvent être plus propices à la prise de décisions financières.

Le mois de l'année peut également influencer la disponibilité financière des clients. Certains mois peuvent être plus propices aux rentrées d'argent, tels que les périodes de prime annuelle, les remboursements d'impôts ou les augmentations de salaire.

Durée du contact

La durée de l'appel peut être un indicateur de l'engagement et de l'intérêt du client. Une durée d'appel plus longue peut également indiquer que le client pose des questions, cherche des clarifications et exprime des préoccupations.

Nombre de contacts au cours la campagne

Un nombre d'appels important peut indiquer un engagement plus fort et un intérêt accru de la part du client. Parfois, un client peut ne pas être prêt à souscrire lors du premier contact, mais des appels ultérieurs peuvent rappeler l'existence de l'offre et encourager le client à y réfléchir davantage.Cela peut être perçu positivement par certains clients, montrant qu'ils sont pris en compte et que leurs besoins sont importants pour l'institution.

Jours depuis le dernier contact lors d'une précédente campagne

Si le contact précédent remonte à un laps de temps relativement court, le client peut se souvenir de l'offre précédente et de la discussion qui a eu lieu.

Au contraire, un nouvel appel après une période de temps plus longue peut être l'occasion de présenter des produits ou des solutions mieux adaptés aux besoins actuels du client.

Certains clients peuvent être plus réceptifs à un contact plus fréquent, tandis que d'autres peuvent préférer une approche plus espacée.

Nombre de contacts lors d'une précédente campagne

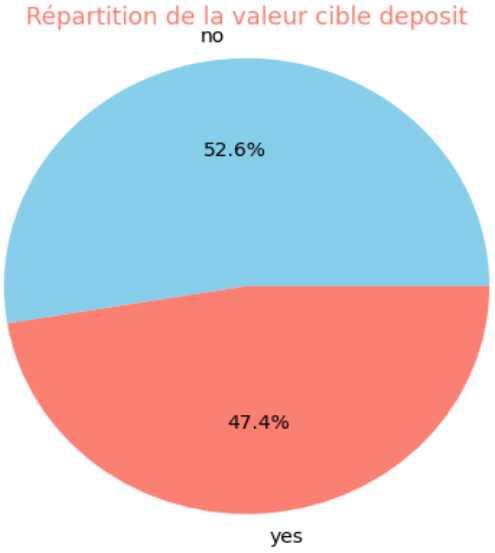
Un nombre élevé de contacts lors d'une précédente campagne peut indiquer que le client a montré une certaine réceptivité et un intérêt initial pour l'offre. Cela peut suggérer que le client est ouvert à la discussion et à l'évaluation des produits et services bancaires proposés. Cependant, il est également important de considérer que trop de contacts répétés peuvent entraîner une saturation ou une lassitude de la part du client. Si le client a été sollicité de manière excessive lors de la précédente campagne, il est possible qu'il devienne moins réceptif aux contacts futurs ou qu'il développe une résistance à la proposition.

Issue de la campagne précédente

Si l'issue de la campagne précédente a été positive et que le client a souscrit au produit ou au service proposé, cela peut indiquer une propension plus élevée à souscrire à de nouveaux produits ou services bancaires.

III. Visualisation des données

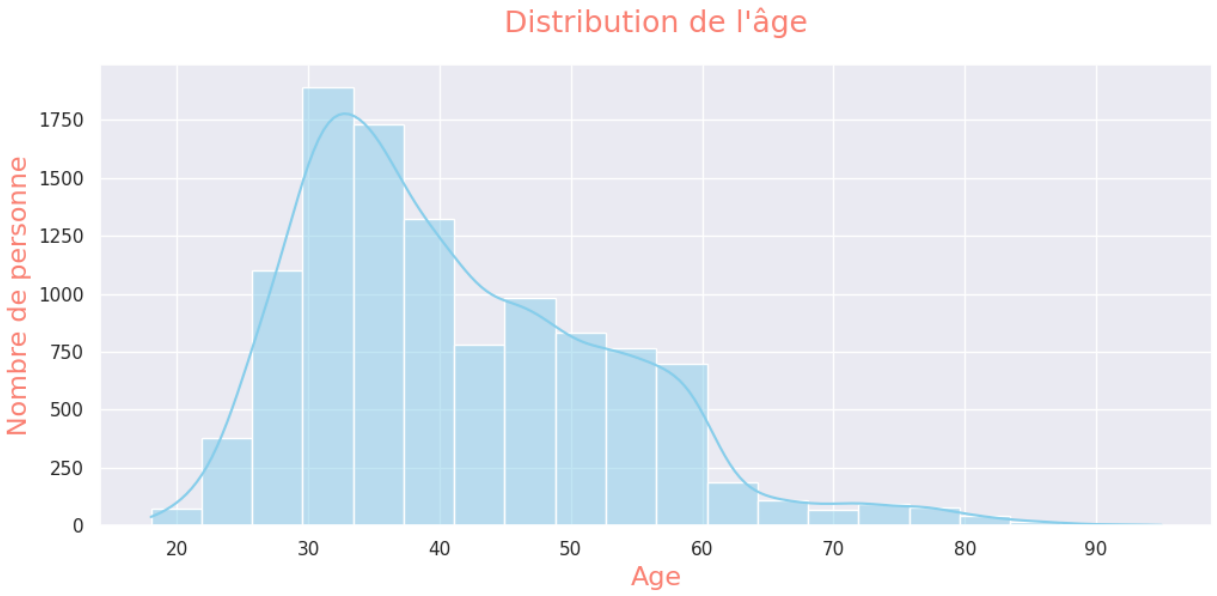
A. Distribution de la valeur cible ‘déposit’



La répartition de la valeur cible nous montre une répartition proche entre le oui et le non avec une légère tendance à 52,6% pour le non.

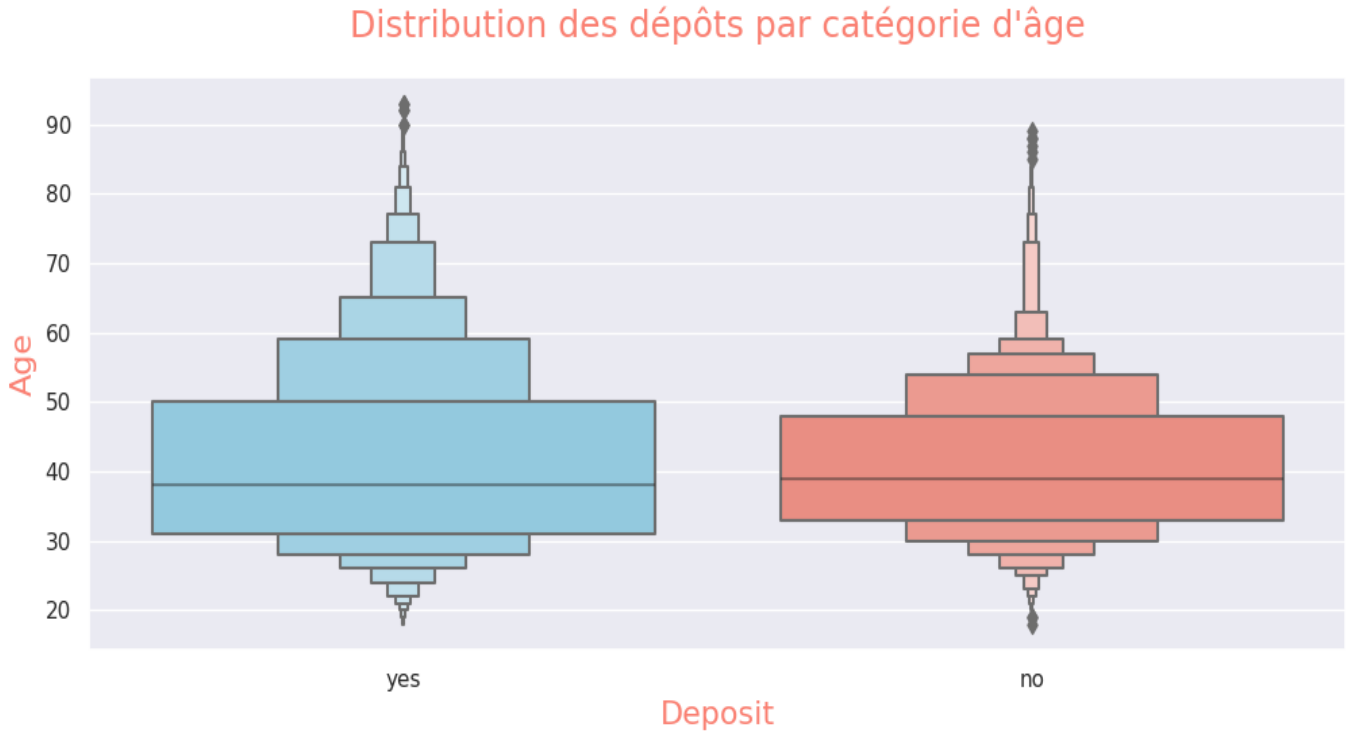
B. Etude de la variable “âge”

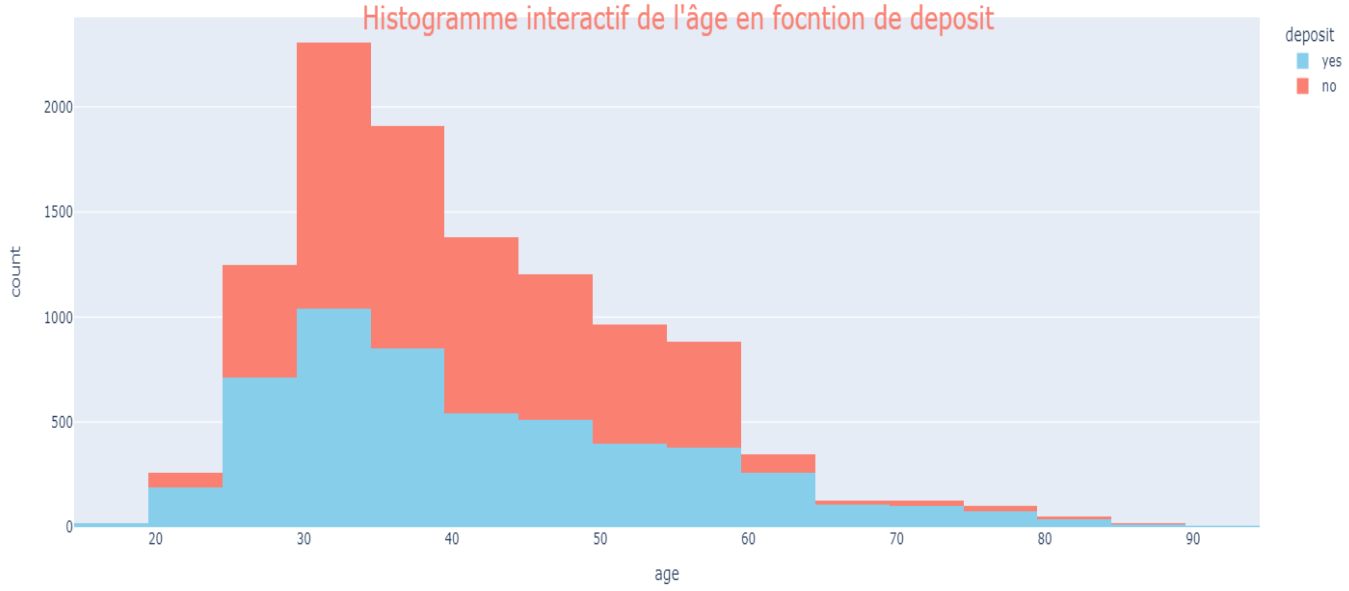
Distribution de la variable “age” dans le dataset



La courbe montre une forte concentration de personnes dans la tranche d'âge de 30 à 40 ans,indiquant que cette catégorie d'âge est particulièrement représentée dans l'échantillon ou la population étudiée. Les autres groupes d'âge semblent être moins fréquents, avec une diminution progressive des effectifs vers les tranches d'âge plus jeunes et plus âgées.

Distribution de l'âge en fonction de la variable cible





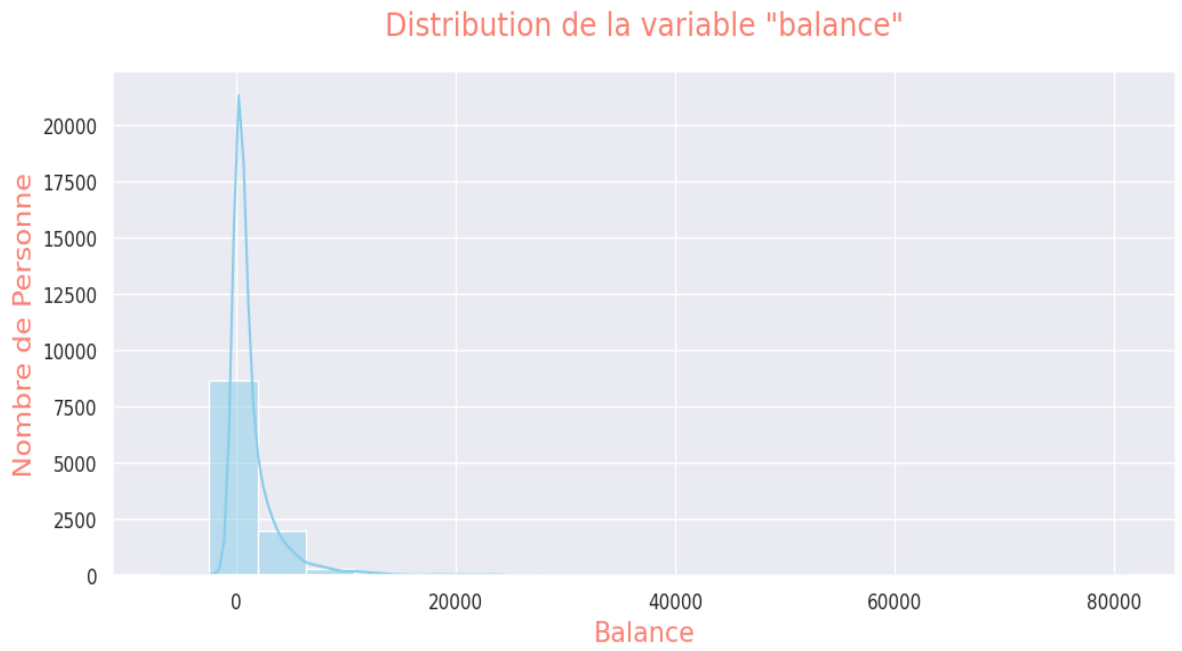
L'analyse du boxenplot entre l'âge et la variable "deposit" révèle que la médiane de l'âge des clients ayant souscrit un dépôt est légèrement inférieure à 40 ans, tandis que la médiane pour les clients n'ayant pas souscrit de dépôt est d'environ 40 ans.En outre le test statistique a permis que la variable âge a une influence significative sur la variable cible.

Cela peut indiquer que les personnes légèrement plus jeunes, âgées de moins de 40 ans, pourraient être plus enclines à souscrire à un dépôt, tandis que les clients d'âge moyen (environ 40 ans) pourraient être moins enclins à le faire. Ces tendances sont mises en avant sur le graphique ci-dessous.

C. Analyse univariée de “Balance”, “Duration” et “Campaign”, identification et traitement de valeurs extrêmes ou absurdes

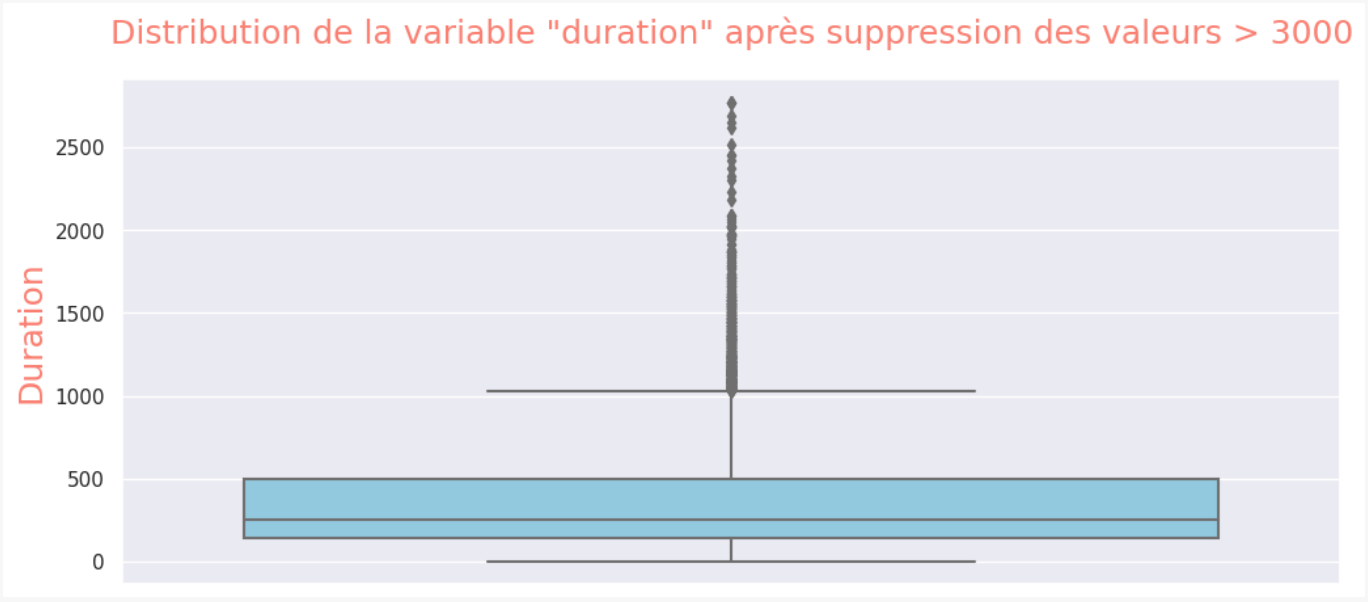
Ces variables représentent une information qui peut-être pertinente pour une étude marketing. Un préalable nécessaire est de visualiser la distribution des ces différentes variables et de traiter les éventuels *outliers* afin que ces données restent réellement pertinentes.

Distribution de la variable Balance



Le graphique montre qu'il y a une très large majorité de clients avec un montant de moins de 20 000 dollars dans la variable "balance", cela peut indiquer que la majorité des clients ont des soldes modestes sur leur compte.

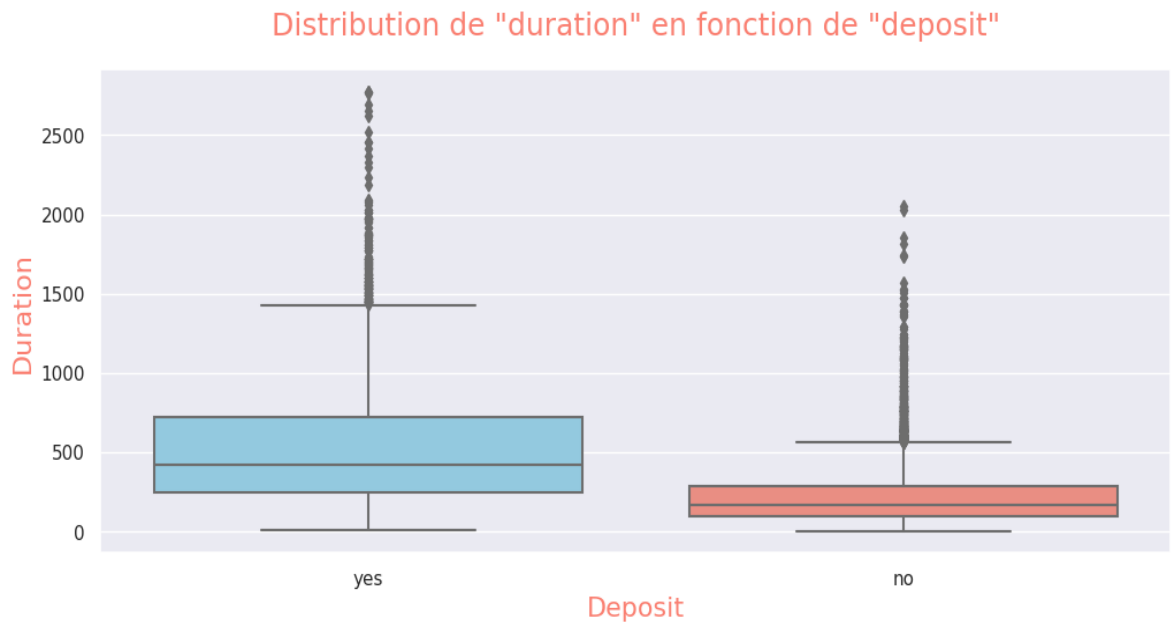
Distribution de duration après suppression des outliers (valeur sup. à 3 000 secondes)



Après le traitement des outliers, la distribution de la variable "duration" est concentrée autour de la médiane de 300, mais il existe des valeurs extrêmes au-delà de 500 qui peuvent potentiellement avoir un impact sur l'analyse ou nécessiter une attention particulière dans le contexte de l'étude.

Il serait judicieux de prendre en compte ces valeurs extrêmes lors de l'interprétation et de l'analyse des données.

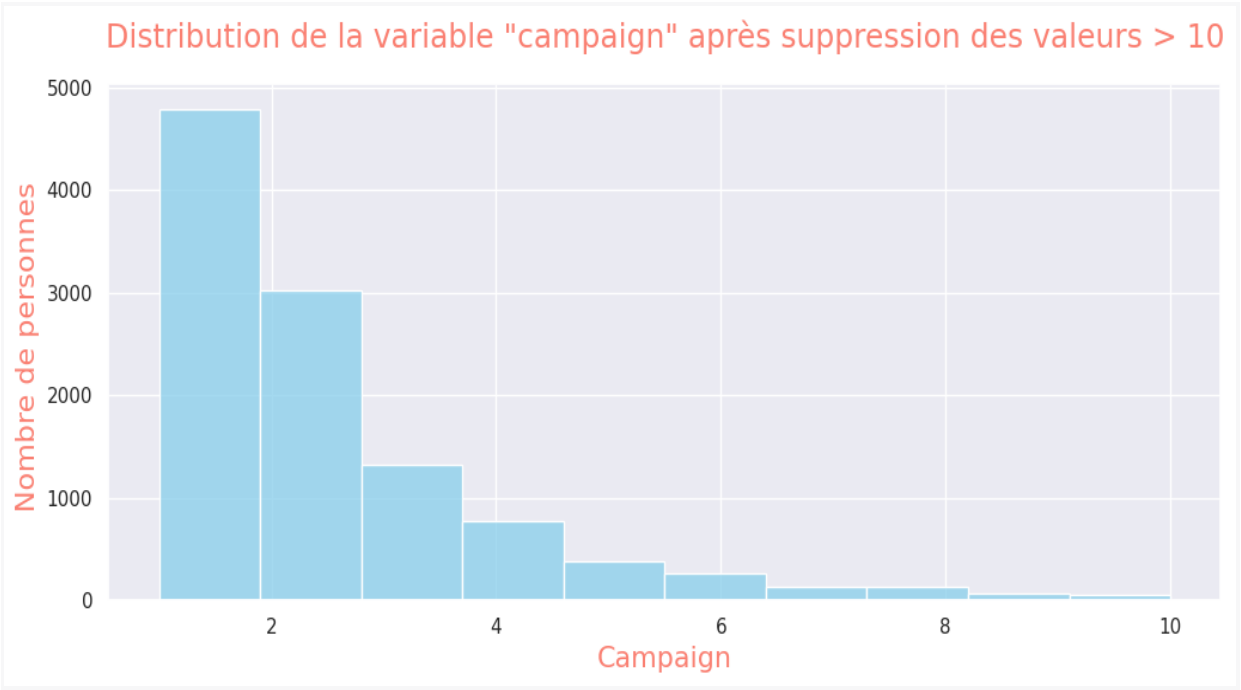
Distribution de la durée en fonction de la variable cible



La durée de contact (duration) semble être plus longue pour les clients ayant souscrit (médiane d'environ 500) par rapport à ceux qui n'ont pas souscrit (médiane d'environ 150).

Cela suggère que la durée d'appel peut être un facteur influençant positivement la décision de souscrire au dépôt.

Distribution de la variable campaign après le traitement des outliers

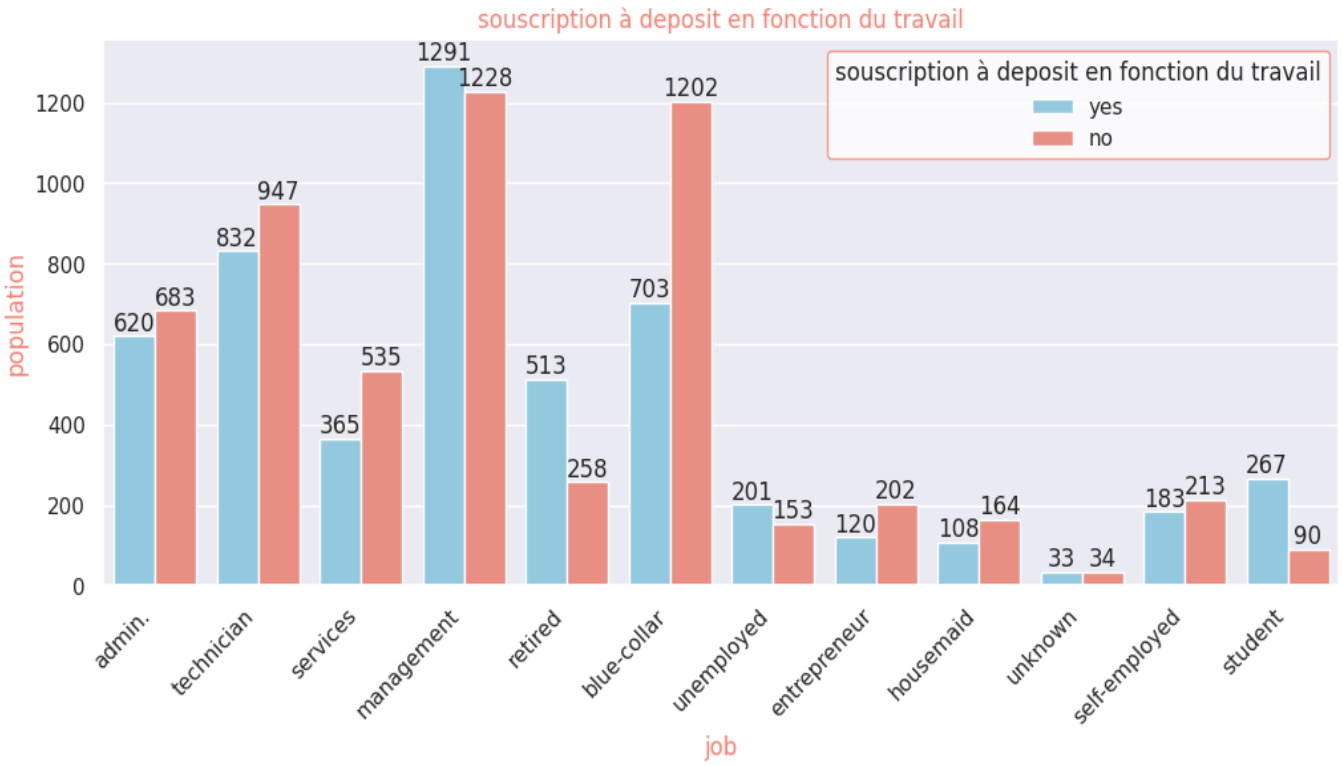


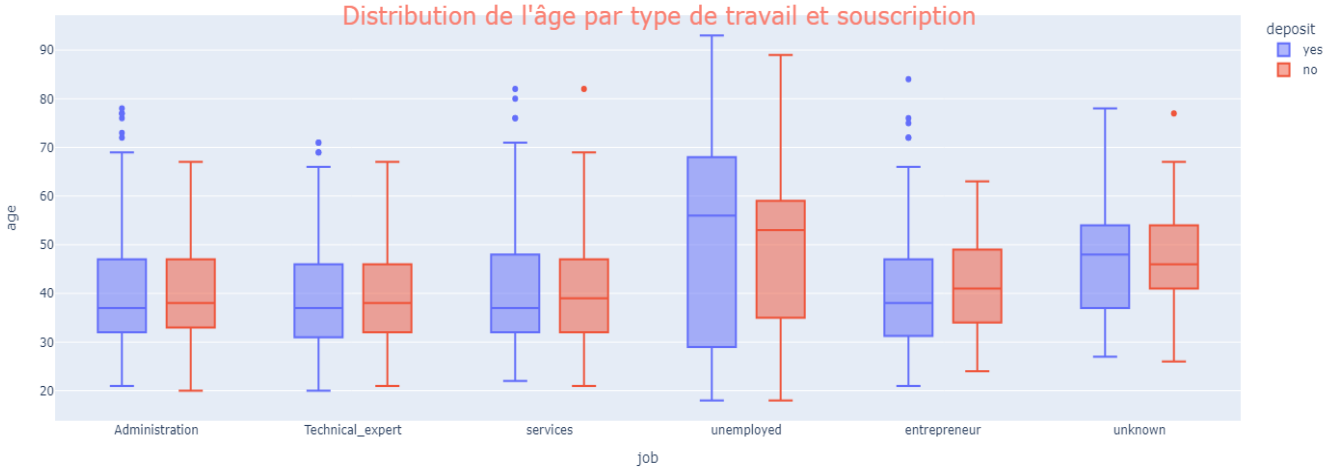
Après la suppression des outliers de la variable "campaign" on note concentration significative des valeurs autour de 1 et 2, ce qui indique que la majorité des clients ont été contactés pour la campagne publicitaire une ou deux fois. Cela suggère que la banque adopte une stratégie de communication plus efficace en limitant le nombre de contacts répétés avec les clients.

D. Première approche d'une relation entre les données personnelles et la variable cible : emploi, d'éducation et d'emprunt en fonction de la valeur cible "deposit"

Nous allons nous intéresser dans un premier temps aux 4 variables contenant des unknown. Nous allons représenter les quatres graphiques montrant la distribution de ces variables en fonction de la valeur cible deposit.

Le premier de ces quatre graphiques représente la souscription à déposit en fonction du travail.





Sur ce premier graphique nous pouvons constater que le plus grand nombre de personnes ayant participé à l'étude sont des personnes travaillant dans le management avec un équilibre quasi parfait de souscription et de non souscription à deposit.

Nous remarquons également une grosse différence dans la catégorie blue collar avec un écart de 40% sur la non souscription à deposit. A l'inverse les catégories 'retired' et 'student' voit leur proportion à avoir souscrit à deposit comme étant plus du double de la non souscription.

Les unknown représentent quant à eux 0.6% des valeurs. Au vu de leurs faibles quantités, nous prenons la décision de les remplacer par le mode ('management') dans le dataset.

Nous avons regroupé par catégorie les différents types de job comme suit :

Administration : "admin",'management'

Technical_expert : "technician","blue-collar"

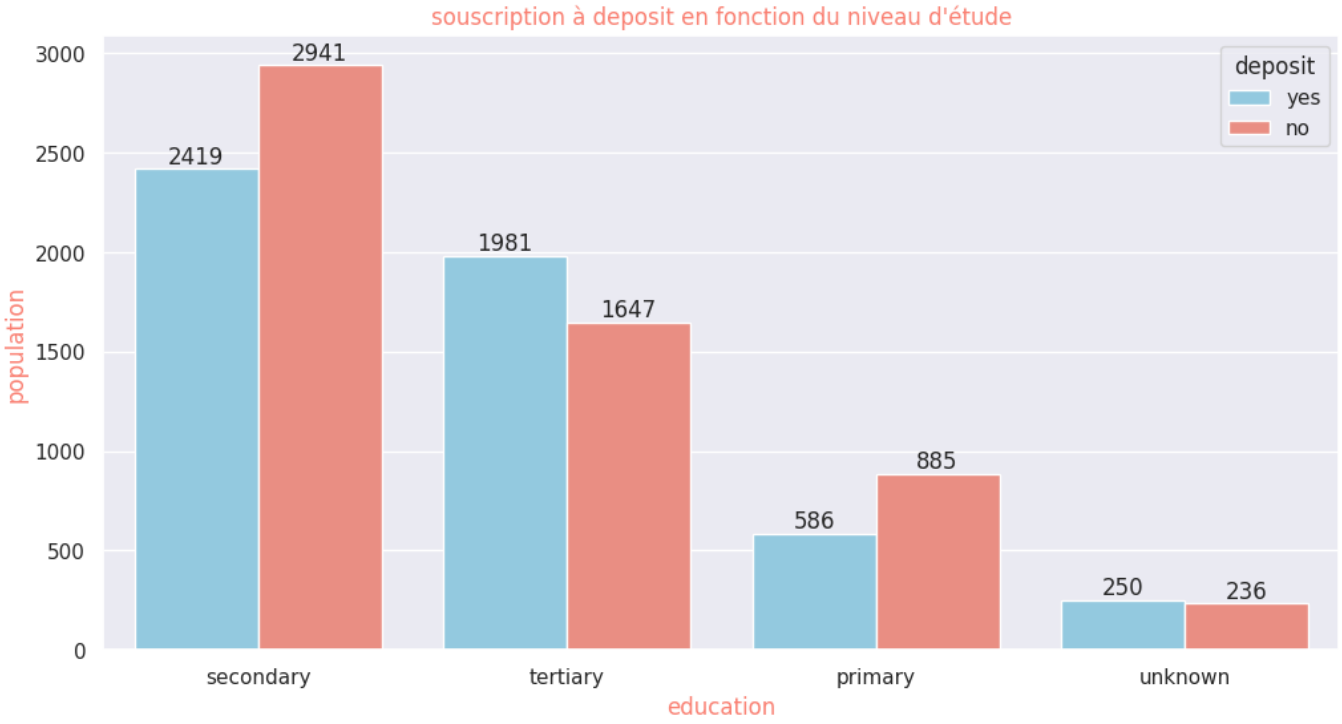
Services : "services","housemaid"

unemployed : "unemployed","student","retired"

Entrepreneurs : "entrepreneur", "self-employed"

Les regroupements non pas modifiés les tendances observées précédemment. La catégorie 'unemployed' contenant les personnes inactives c'est à dire les sans emploies, les retraités et les étudiants ont une tendance forte à avoir souscrit à deposit.

Nous allons maintenant regarder la variable ‘education’ en fonction de deposit.



Niveau d’étude primaire : Apprentissage de base destinées aux enfants de 6 à 11ans.

Niveau secondaire : collège, lycée de 12 à 18 ans

Niveau tertiaire : c’est l’enseignement supérieur ou l’individu se spécialise dans un domaine.

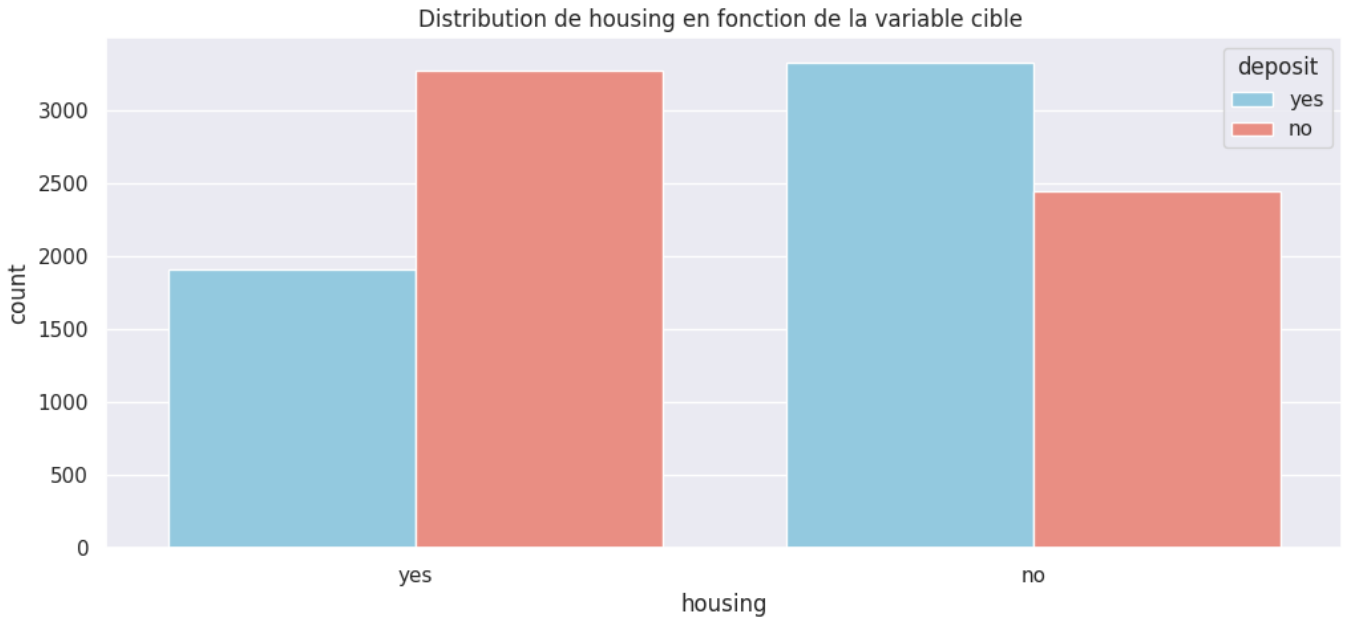
Sur ce graphique nous constatons que le niveau d’étude secondaire est le plus représenté avec une majorité de personnes n’ayant pas souscrit à deposit. Le niveau primaire quant à lui représente la plus faible proportion sur l’étude avec une majorité de personne n’ayant la non plus pas souscrit à deposit.

Nous pouvons remarquer que les personnes ayant fait le plus d'études ont une tendance à avoir souscrit davantage à l’offre commerciale.

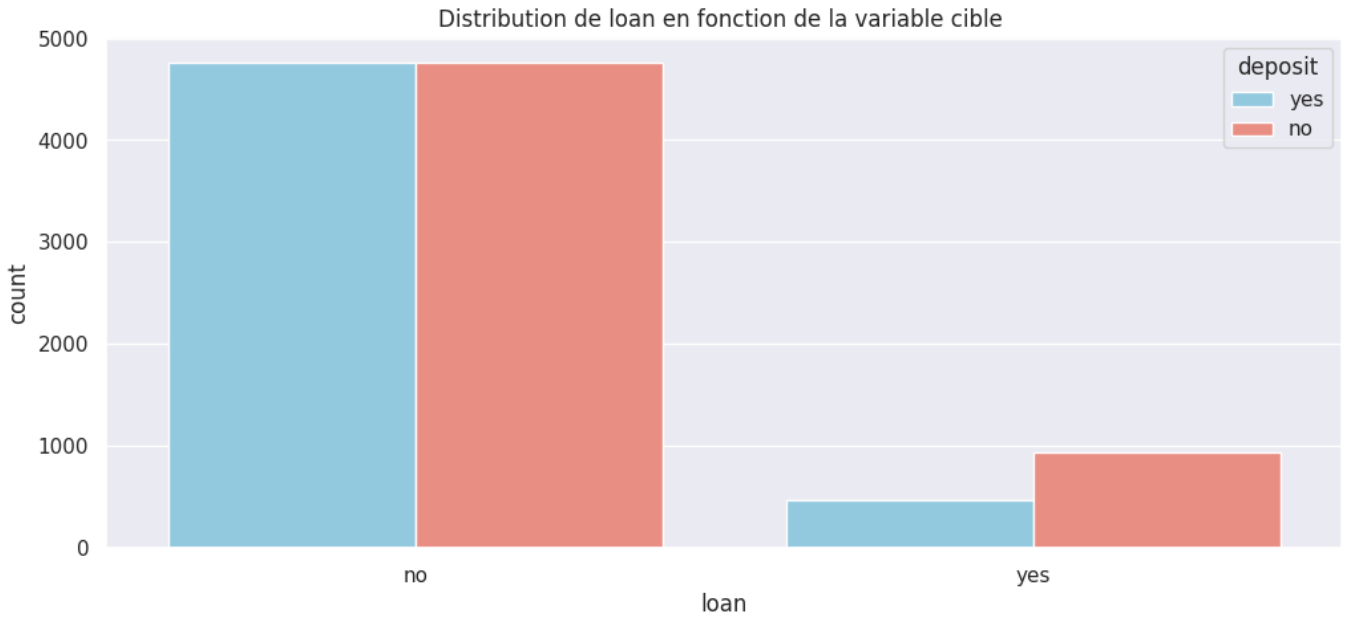
Sur cette variable les ‘unknown’ représentent environ 500 individus soit 4% de la population. Nous pouvons imaginer que ce sont des personnes n’ayant pas voulu donner l’information sur leurs niveaux d’étude. **Pour cette catégorie nous remplaceront les unknown par le mode.**

Emprunts en cours : une limite pour l’épargne des clients

Concernant l’emprunt immobilier, on constate nettement une relation entre l'inexistence d’un crédit et la propension à souscrire au dépôt pour le client.



D'ailleurs, le test du khi² entre ces deux variables retourne une p-valeur très faible : 9.724394114495535e-103.

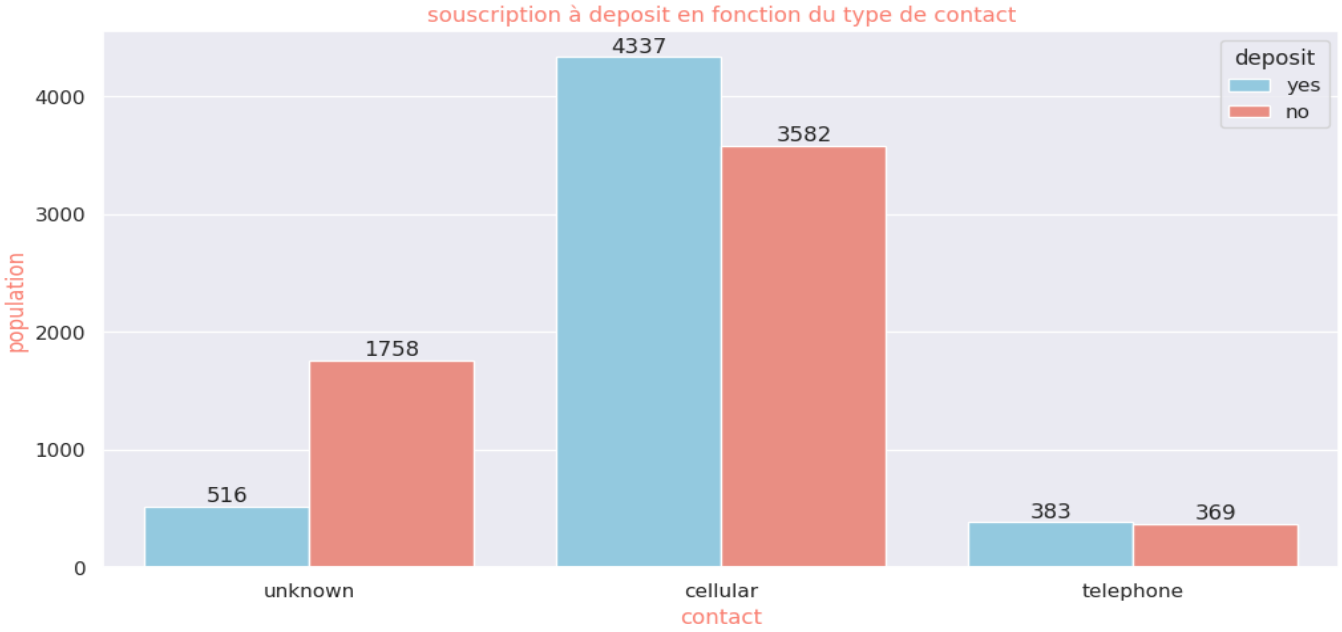


Le constat est similaire pour les autres types d'emprunts même si cette relation est moins évidente que la précédente. La probabilité d'un dépôt est là-aussi plus importante pour les clients n'ayant pas d'emprunt en cours.

La p-valeur pour le test du khi² est également très faible : 2.171286879630289e-31.

E. Modalité de contact et issue de la campagne précédente : analyse multivariée

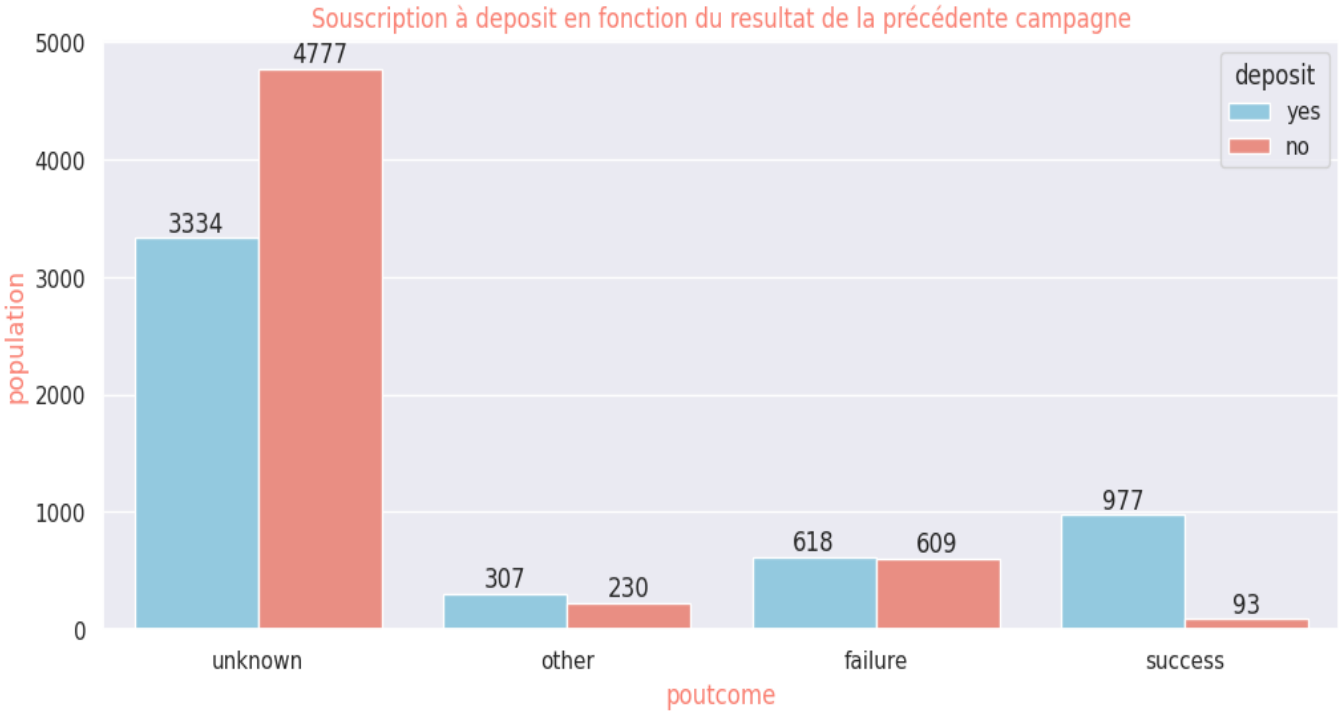
Nous allons regarder maintenant la souscription à deposit en fonction du type de contact.



Cette variable représente la manière dont la personne à été contacté ou à contacter la banque pour répondre à l'offre.

Nous ne retrouvons que des contacts téléphoniques, qu'ils soient par téléphone fixe ou mobile. Les unknown de cette variable représentent 21% de la population de données. **Cette variable ne nous apporte pas d'information sur la souscription. Nous la supprimerons pour la suite de l'étude.**

Dernière catégorie contenant des unknown, la souscription à deposit en fonction du résultat de la précédente campagne.



Le premier constat étant la très forte présence des unknown, représentant 75% des valeurs. **Nous la considérerons comme une catégorie à part entière et la regrouperont avec la catégorie 'other'.**

Nous pouvons remarquer également le fort pourcentage de souscription à deposit quand les personnes ont déjà souscrit à une précédente campagne.

F. Données concernant une campagne précédentes : étude univariée des variables « pdays » et « previous »

Ces deux variables ainsi que poutcome sont liées entre elles car elles communiquent des informations concernant la précédente campagne.

Pour rappel :

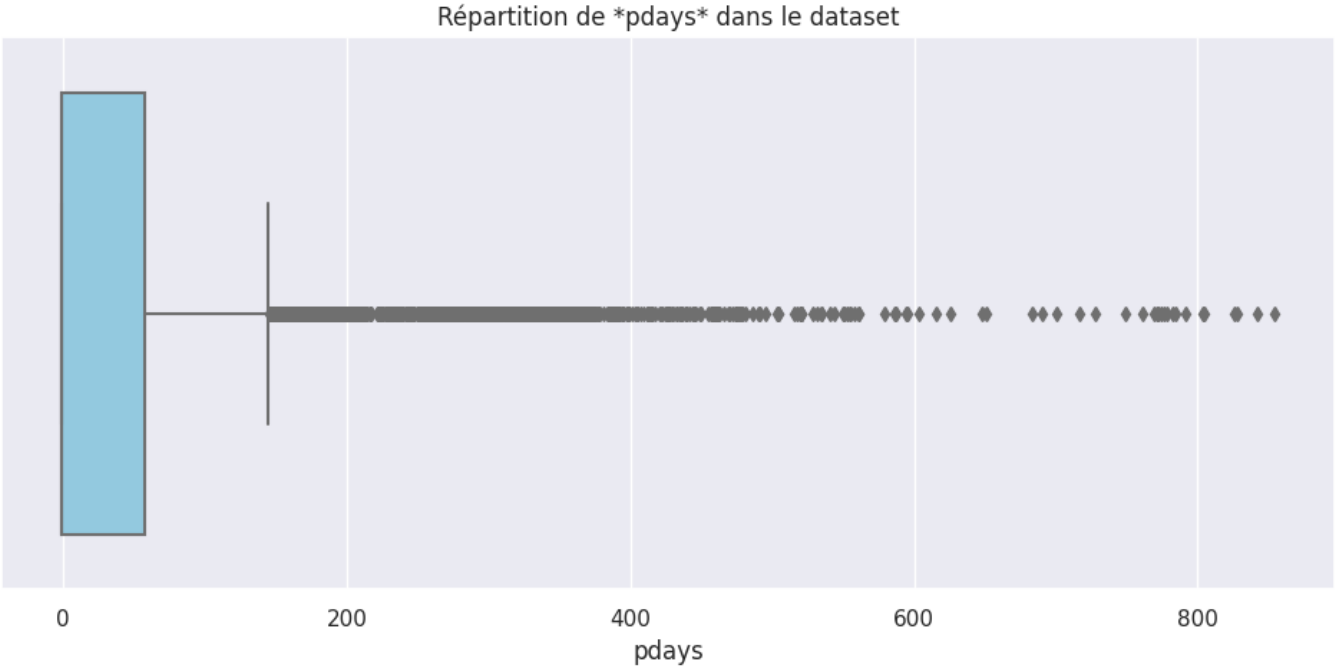
Pdays : Le nombre de jours écoulés depuis le dernier contact avec le client dans le cadre d'une campagne précédente

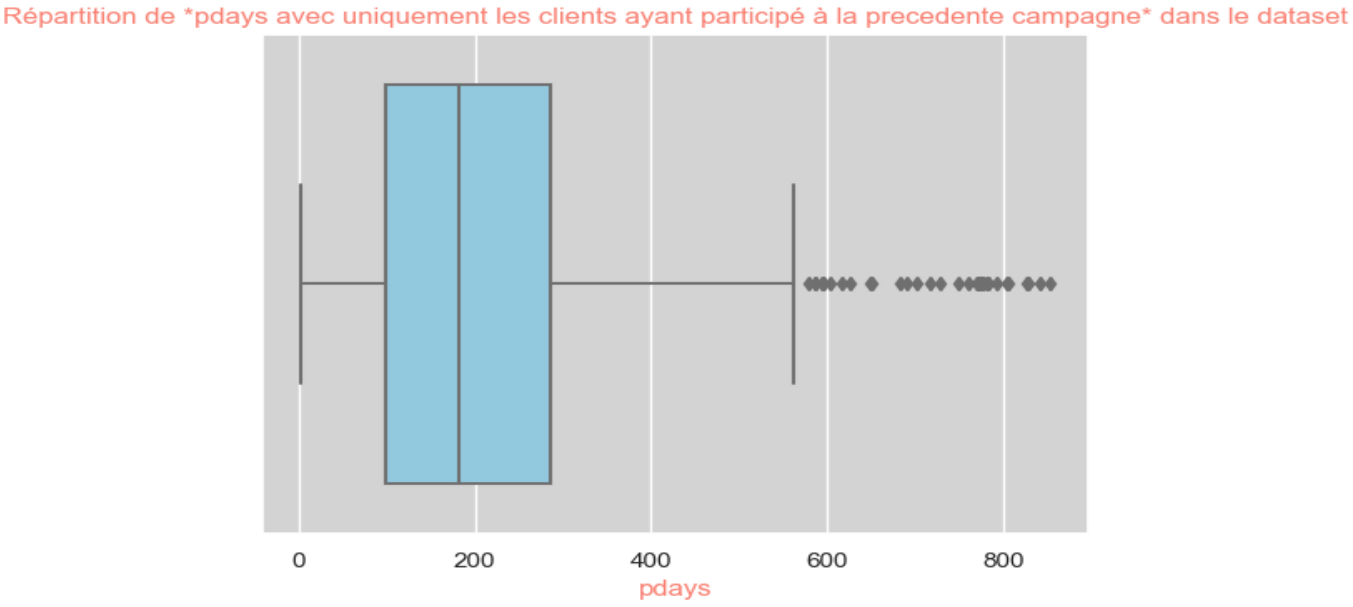
Previous : Le nombre de contacts antérieurs avec le client avant la campagne en cours

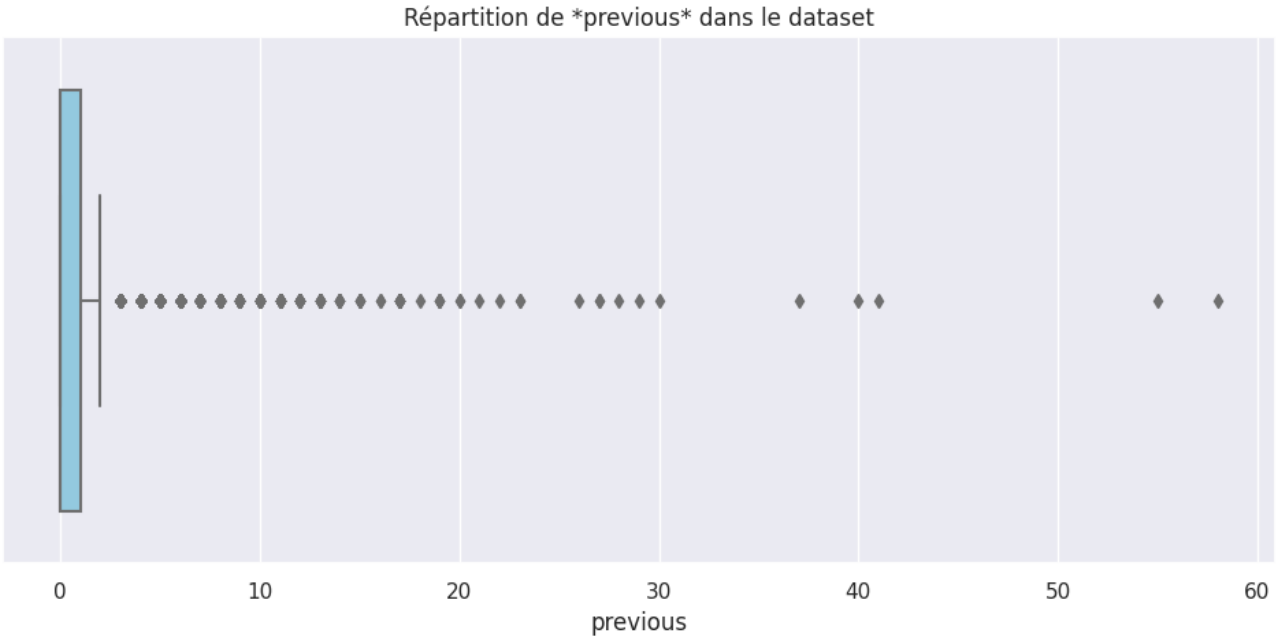
Poucome : Le résultat de la dernière campagne

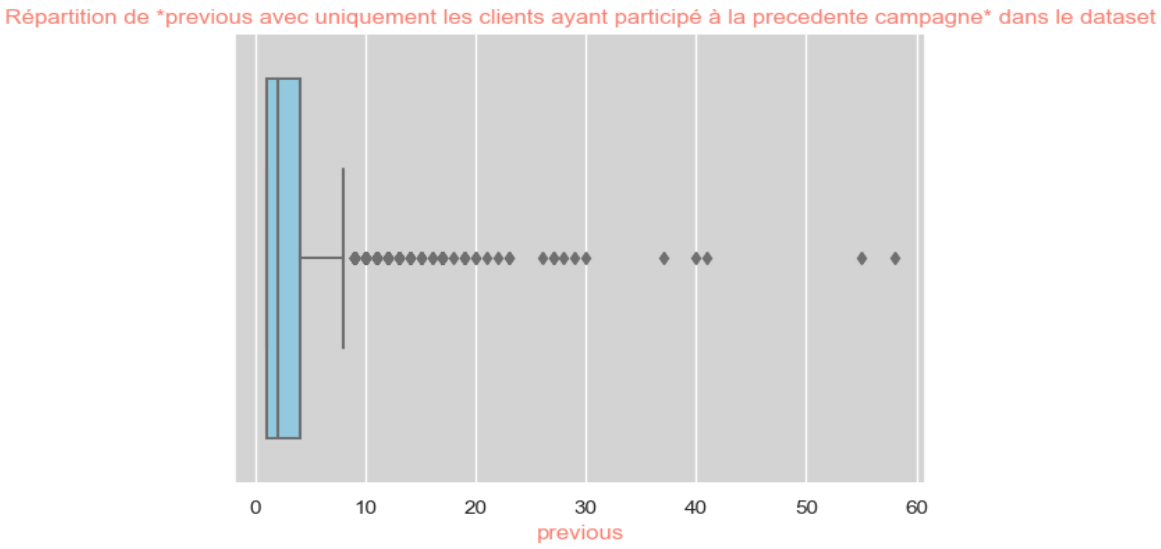
Pour chacune de ces 3 variables une **valeur** nous indique si le client n'a pas participé à la dernière campagne : il s'agit pour **pdays** de la valeur « -1 », **previous** : « 0 » et **poutcome** : « unknown ».

Ci-dessous des graphiques de boites à moustaches montrant la distribution de ces variables sur tous les clients du data frame puis dans un second temps en retirant les clients n'ayant pas participé à la précédente campagne (environ 75% du data set).









On remarque que les valeurs par défaut « -1 » pour pdays et « 0 » pour previous faussent la représentation des graphiques car elles « tirent » les intervalles vers eux.

On remarque notamment pour la variable pdays moins d'outliers du fait que l'interquartile Q3 est par défaut moins élevé car il y a moins de données tirant vers la valeur « -1 ».

Pour la variable previous on constate une concentration plus évidente entre 1 et 5 et un nombre d'outliers à peu près exact au graphique prenant en compte les clients n'ayant pas participé à la précédente campagne. Cela s'explique par des valeurs « cohérentes » assez proches de la valeur par défaut « 0 » attribuée aux clients n'ayant pas participé à la précédente campagne.

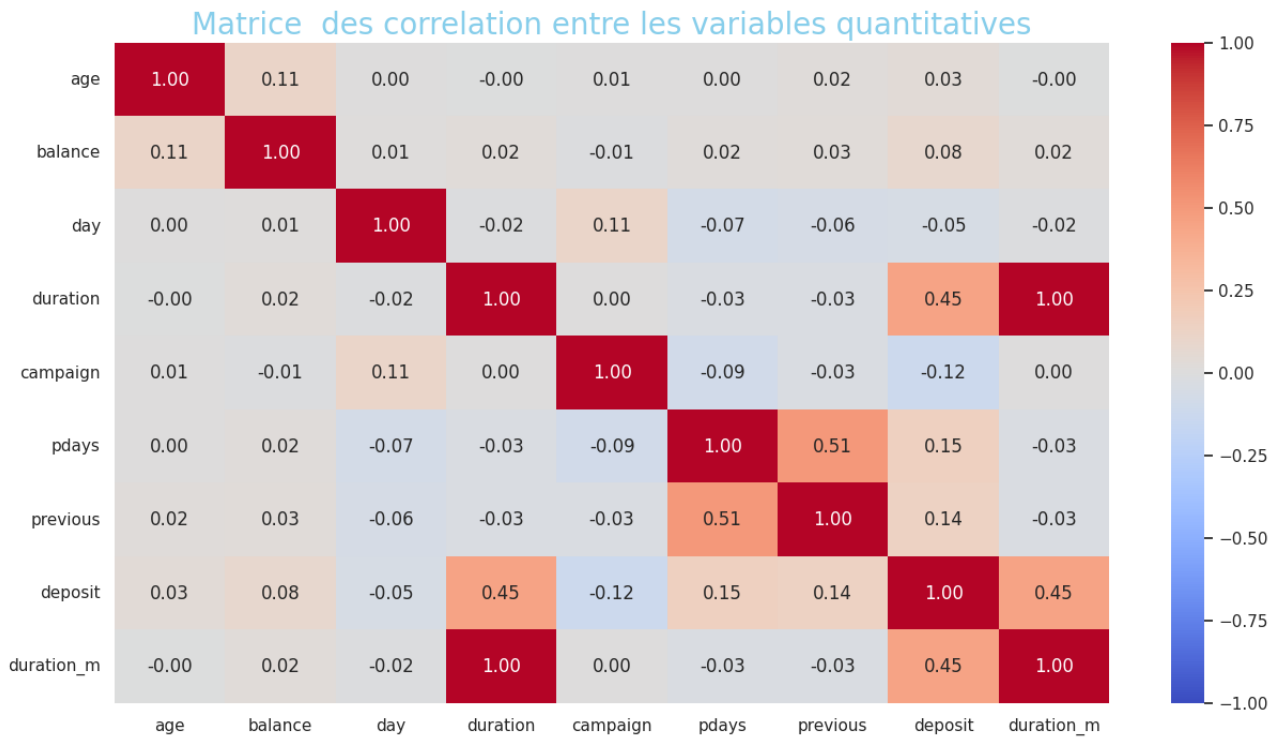
Pour conclure sur les outliers de ces deux variables nous avons fait le choix de garder ceux de la variable pdays car il s'agit de valeurs « possibles ».

Cependant pour la variable previous certains outliers ressemblent bien plus à de fausses informations, **nous avons pris la décision de les remplacer par la moyenne de la variable pour les valeurs supérieures ou égales à 10**.

G. Matrice de corrélation

La matrice de corrélation nous permet d'avoir une idée de la corrélation entre les variables quantitatives du data frame. Plus la valeur se rapproche de 1 et plus les variables sont corrélées entre elles.

Ci-dessous la matrice de corrélation entre les variables quantitatives du data frame :



On remarque rapidement que les deux variables qui ressortent le plus sont « pdays » et « previous » car elles contiennent toutes les deux des informations sur la précédente campagne et notamment une information sur les clients n'ayant justement pas participé à la précédente campagne.

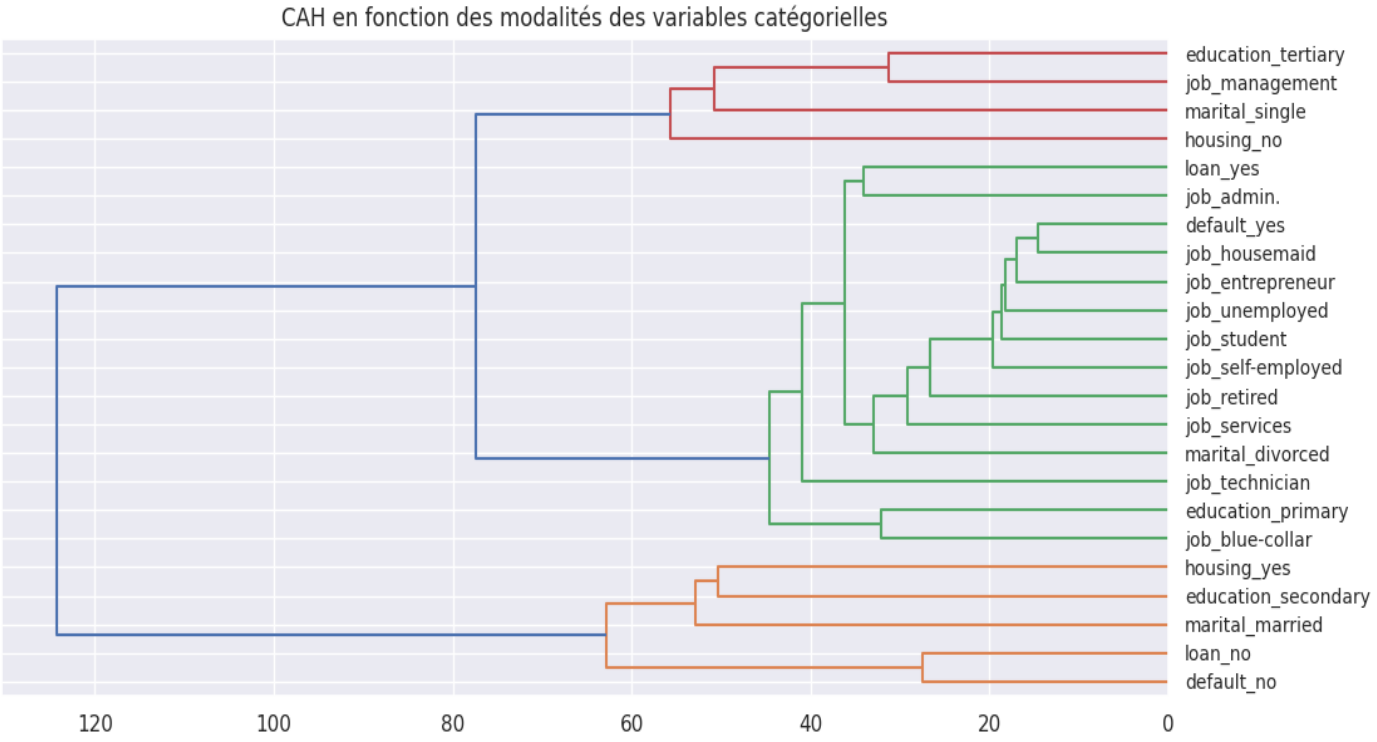
On peut noter également que les variables « age » et « balance » ont une valeur de corrélation assez élevée ainsi que « campaign » et « day ».

H. Clustering à partir des variables catégorielles : étude du profil des clients contactés

Pour conclure notre exploration du dataset, il nous a paru intéressant d'utiliser un modèle de classification pour identifier à première vue, si plusieurs profils se dégagent parmi les clients. Pour ce faire, on a retenu 6 variables catégorielles : job, marital, education, default, housing et loan.

Avec un encodage one hot de ces variables et une fonction dice recensant les distances par paires de colonnes. Cette matrice obtenue étant redondante (par exemple, on y retrouve la distance entre job et marital puis la distance entre marital et job), on a vectorisé celle-ci à l'aide de la fonction squareform.

Le tableau obtenu nous permet d'appliquer une méthode de classification hiérarchique ascendante, à savoir ici, la méthode ward. Le résultat obtenu peut être interprété en affichant son dendrogram.



En paramétrant le seuil à 70, nous pouvons séparer les modalités en 3 groupes. Plus les modalités sont proches, plus leur regroupement est à droite. Les concomitances entre ces modalités nous permettent ici de livrer une première interprétation concernant les clients.

Un premier type de clients (en haut) pourrait être identifié comme "célibataire diplômé". Ils ont plutôt fait des études supérieures, occupent plutôt des postes de managers, sont plutôt célibataires et sans emprunts immobiliers.

Un deuxième type (en bas) pourrait être qualifié comme "stable et en couple". Ils sont plutôt d'un niveau d'éducation du secondaire, marié, avec un emprunt immobilier, plutôt sans autre emprunt et n'ayant plutôt jamais fait défaut.

Le troisième type est plus hétéroclite. Il semble à la fois réunir des situations précaires, des parcours accidentés et une certaine approche d'une classe moyenne. A l'intérieur, on retrouve **un sous-groupe de concomitances très proches décrivant potentiellement une situation précaire :**

- ayant plutôt déjà fait défaut ;
- occupant plutôt soit le métier d'agent de ménage soit une situation d'étudiant, d'auto-entrepreneur, d'entrepreneur ou de sans emploi

Dans le reste du groupe, à des concomitances plus faibles (plus à gauche dans l'arbre), on observe des caractéristiques qu'on pourrait plutôt attribuer à une certaine forme de **classe moyenne :**

- ayant plutôt déjà contracté un crédit auto, une crédit pour un projet ou un crédit à la consommation
- occupant plutôt des métiers d'ouvrier, de technicien, de postes administratifs ou dans les services
- on retrouve aussi des situations de retraité
- on voit des niveaux d'éducation plutôt faibles
- des personnes ayant pu connaitre un divorce

Ces concomitances nous permettent d'identifier des interactions entre les variables et ainsi de dessiner des tendances qu'il peut exister au sein du dataset. Il ne s'agit pas de groupes fermés mais plutôt de stéréotypes qu'on peut ressortir à partir des données.

Cette première approche, en plus de nous permettre de nous approprier ces données catégorielles, est susceptible de nous aider par la suite pour éventuellement interpréter les prédictions que nous obtiendrons par les modèles de machine learning ou peut-être d'optimiser ces derniers.

IV. Définition du problème de machine learning

A. Objectif de la banque

La banque souhaite savoir à l'avance, à partir des données personnelles, si un client va souscrire à un produit faisant partie des dépôts à terme. Il s'agit donc d'un problème de classification. Les dépôts à terme sont des placements financiers où le client dépose une somme d'argent fixe auprès d'une institution financière pour une durée prédéterminée. Pendant cette période, l'argent génère un taux d'intérêt fixe et garanti. À l'échéance, l'investisseur peut retirer le capital initial ainsi que les intérêts générés.

Cette prédiction interprétée à l'échelle de l'individu peut permettre à la banque de comprendre les éléments qui participent à déterminer si un client est susceptible de souscrire au produit ou non. Ce faisant, elle pourrait être en mesure d'opérer une segmentation pertinente afin de concentrer ses efforts sur les prospects ayant le plus de potentiel.

En fonction du poids que représentent les données liées à l'appel en lui-même, l'interprétation des résultats peut aussi aider l'institution financière à identifier ce qui fonctionne concernant les pratiques de télémarketing. Ainsi, elle pourra travailler à augmenter la productivité de la vente en suivant la quantité et la durée des appels. Cela pourrait découler par exemple sur l'adaptation du script ou de la formation des télémarketeurs.trices.

En théorie, il serait intéressant de suivre comment l'application de préconisations en termes de segmentation notamment impacte les KPI de la banque : le taux de conversion, le coût par conversion.

B. Méthodologie déployée

Suite au travail exploratoire, il nous est possible d'opérer le nettoyage et la préparation des données prévus dans le premier rapport, ainsi que la division en un ensemble d'entraînement et un ensemble de test. La taille de l'ensemble de test est définie à 20 %. Nous procédons également à l'encodage one-hot des variables catégorielles ainsi qu'à la normalisation des variables numériques.

L'étape suivante consiste à entraîner sur ces données plusieurs modèles de classification et à comparer les métriques. Un soin doit être apporté au choix des métriques observées afin qu'elles soient pertinentes au regard de la problématique métier.

Les 3 modèles les plus performants font l'objet d'une optimisation en testant plusieurs valeurs pour les hyperparamètres de ces modèles grâce à GridSearchCV. La validation croisée nous permet d'identifier la meilleure combinaison d'hyperparamètres afin d'entraîner les modèles retenus avec ces derniers. L'analyse des évolutions des métriques d'évaluation des modèles optimisés constitue l'aspect final de cet étape.

La dernière étape repose sur l'interprétabilité des modèles. Ce travail permet d'identifier quelles sont les variables les plus importantes pour la prédiction d'un modèle. Cette partie est indispensable pour faire le lien avec les enjeux et les impacts métiers. Elle peut également être utile pour optimiser nos modèles.

C. Résultats visés

Identifier les critères selon lesquels nous appréhendons nos résultats est essentiel pour mesurer l'efficacité de notre approche et sa pertinence par rapport aux objectifs de la banque. Cela reposera autant sur les performances intrinsèques de nos modèles (scores) que notre capacité à les traduire en performance future pour l'entreprise.

- **Les scores de nos modèles de prédictions :** la capacité de nos modèles à prédire avec précision la souscription au produit par un client est incontournable. Les différents scores présentés ci-après doivent s'approcher de 1.
    a. **Accuracy (Exactitude) : Notre objectif principal est d'obtenir une accuracy élevée dans la prédiction du succès d'une campagne de marketing.** Cela signifie que nous visons à maximiser le nombre de prédictions correctes, en minimisant à la fois les faux positifs et les faux négatifs. Une accuracy élevée garantit que notre modèle est capable de classer correctement la grande majorité des clients, que ce soit ceux qui souscriront au produit de dépôt à terme ou ceux qui ne le feront pas.
    b. **Précision (Précision) :** En plus de l'accuracy, nous cherchons à obtenir une précision élevée. La précision mesure la proportion de prédictions positives effectivement correctes parmi toutes les prédictions positives faites par le modèle. Cela signifie que nous souhaitons minimiser le nombre de clients classés à tort comme souscrivant au produit. Une précision élevée garantit que nos actions de marketing sont ciblées de manière plus efficace.
    c. **Rappel (Recall) :** En complément de la précision, nous cherchons également à obtenir un rappel élevé. Le rappel mesure la proportion de clients réellement souscrivant au produit et correctement identifiés parmi tous les clients réellement souscrivant au produit. Cela signifie que nous voulons minimiser le nombre de clients qui souscrivent au produit mais qui sont classés à tort comme ne le faisant

pas. Un rappel élevé garantit que nous identifions efficacement les opportunités de conversion.

d. **F1 Score** : En plus des métriques précédentes, le F1 Score est un indicateur essentiel pour évaluer la performance de nos modèles. Le F1 Score est la mesure de l'harmonie entre la précision et le rappel. Il est particulièrement utile lorsque nous devons trouver un équilibre entre la réduction des faux positifs et des faux négatifs tout en maximisant la précision et le rappel. Un F1 Score élevé indique que notre modèle parvient à bien classer les clients susceptibles de souscrire au produit tout en évitant de classer à tort les clients qui ne le feront pas.

Notre capacité à traduire ces métriques en performances futures pour l'entreprise sera cruciale pour évaluer l'efficacité globale de notre approche."

- **Interprétabilité :** Outre les métriques de performance, nous visons également à rendre nos modèles interprétables. Nous souhaitons identifier les facteurs clés qui influencent la décision de souscrire au produit de dépôt à terme. Cette interprétabilité nous permettra de fournir des informations exploitables à la banque pour améliorer ses campagnes de marketing et ses pratiques de télémarketing.
- **Impact métier :** Notre objectif central est d'aligner notre travail sur les besoins et les objectifs métier de la banque. Nous cherchons à évaluer comment l'application de nos modèles de prédiction peut apporter une réelle valeur ajoutée à l'entreprise.

C’est au regard de ces 3 critères que nous pourrons proposer à la banque un modèle et des solutions pertinentes et performantes.

V. Présentation des modèles mis en place

A. Présentation des modèles testés

Nous avons testé 7 modèles de classification. Cette démarche nous offre une variété d'approches pour résoudre des problèmes de classification, chacun ayant ses avantages et ses inconvénients.

LogisticRegression :

La régression logistique est un modèle de classification linéaire largement utilisé. Elle modélise la probabilité qu'une observation appartienne à une classe particulière. Elle est efficace pour les problèmes de classification binaire et peut être étendue à la classification multiclasse.

DecisionTreeClassifier :

Le modèle d'arbre de décision divise les données en sous-groupes basés sur des règles de décision pour atteindre une décision finale. Il est capable de gérer des données numériques et catégorielles. Cependant, il peut être sujet à l'overfitting si les arbres sont trop profonds.

RandomForestClassifier :

Random Forest est une extension de l'arbre de décision qui combine plusieurs arbres pour améliorer la précision et réduire le surapprentissage. Il crée des sous-ensembles aléatoires de données d'entraînement et d'attributs pour chaque arbre, puis agrège leurs prédictions.

SVM (Support Vector Machine) :

Les machines à vecteurs de support sont des modèles puissants pour la classification binaire. Elles trouvent un hyperplan qui maximise la marge entre les classes, ce qui les rend efficaces même dans des espaces de grande dimension.

KNeighborsClassifier :

Le modèle des k plus proches voisins attribue une classe à une observation en fonction de la majorité des classes de ses voisins les plus proches. Il peut être sensible à l'échelle des caractéristiques et nécessite un choix judicieux du nombre de voisins (k).

GradientBoostingClassifier :

Le gradient boosting est une technique d'ensemble où de nombreux modèles faibles (habituellement des arbres de décision) sont entraînés séquentiellement pour corriger les erreurs des modèles précédents. Il peut nécessiter un ajustement minutieux des hyperparamètres.

AdaBoostClassifier :

AdaBoost est un autre algorithme d'ensemble qui met l'accent sur les exemples mal classés à chaque étape de l'apprentissage. Il combine plusieurs modèles faibles pour former un modèle fort.

B. Les scores utilisés et premiers résultats

Nous avons évalué la performance de sept modèles de classification différents pour prédire le succès de la campagne de marketing bancaire. Nous avons utilisé plusieurs métriques de performance pour évaluer leurs performances, la précision, le rappel, et l'accuracy (score de test), en se concentrant sur l'accuracy puisque comme la répartition de positifs et de négatifs pour la variable cible est équilibrée, ce score pourra nous servir de référence pour la sélection et l'optimisation des modèles.

Support Vector Machine (SVM) :

Le SVM affiche une précision élevée de 0.826415, ce qui indique sa capacité à bien classer les clients susceptibles de souscrire au produit de dépôt à terme. Le rappel de 0.8327 suggère que le SVM peut également bien rappeler les vrais positifs. L'accuracy (score de test) de 0.835541 confirme que le SVM est capable de classer correctement une grande proportion de clients. Le F1 score est de 0.829545 ce qui indique que le modèle classe convenablement les clients susceptibles de souscrire ou pas au produit.

Gradient Boosting :

Le Gradient Boosting présente une précision de 0.818182, montrant sa capacité à bien classer les clients intéressés. Un rappel de 0.855513 indique une bonne capacité à rappeler les vrais positifs. L'accuracy (score de test) de 0.839196 montre que le modèle est globalement performant pour prédire le succès de la campagne. Le F1 score est de 0.836431 ce qui indique que le modèle classe convenablement les clients susceptibles de souscrire ou pas au produit.

Régression Logistique :

La Régression Logistique affiche une précision élevée de 0.825175, montrant sa capacité à bien classer les clients. Un rappel de 0.785171 montre une capacité raisonnable à rappeler les vrais positifs. L'accuracy (score de test) de 0.816811 indique une performance solide dans l'ensemble. Le F1 score est de 0.804676 ce qui indique que le modèle classe convenablement les clients susceptibles de souscrire ou pas au produit.

KNeighborsClassifier :

Le modèle KNeighborsClassifier présente des performances prometteuses en termes de précision (0.767699), mais avec un rappel (0.659696) relativement plus faible. L'accuracy (score de test) de 0.740521 montre que le modèle est capable de classer correctement une proportion significative de clients, bien que l'optimisation des hyperparamètres soit nécessaire pour maximiser son potentiel. Le F1 score est de 0.709611 ce qui indique que ce modèle n'est pas le meilleur pour classer au mieux les clients susceptibles de souscrire ou pas au produit. Il s'agit du modèle avec le F1 score le plus faible.

AdaBoostClassifier :

L'AdaBoostClassifier affiche une performance solide en termes de précision (0.80831), montrant sa capacité à bien classer les clients. Un rappel de 0.813688 suggère une bonne capacité à rappeler les vrais positifs. L'accuracy (score de test) de 0.817725 confirme que le modèle est performant dans l'ensemble. Le F1 score est de 0.810990 ce qui indique que le modèle classe convenablement les clients susceptibles de souscrire ou pas au produit.

RandomForestClassifier :

Le modèle RandomForestClassifier présente une performance robuste avec une précision élevée (0.810406), un rappel élevé (0.873574), et une bonne accuracy (score de test) de 0.841023. Cela indique que le modèle est capable de bien classer les clients et qu'il généralise bien, bien que l'optimisation des hyperparamètres soit nécessaire pour éviter le surapprentissage. Le F1 score est de 0.844096 ce qui indique que le modèle classe convenablement les clients susceptibles de souscrire ou pas au produit. Il s'agit du modèle avec le F1 score le plus élevé.

DecisionTreeClassifier :

Le modèle DecisionTreeClassifier montre une performance acceptable en termes de précision (0.746518) et de rappel (0.764259), avec une accuracy (score de test) de 0.761992. Cependant, il présente un risque de surapprentissage avec un score d'entraînement parfait (1.000000), et il pourrait bénéficier d'une régularisation et d'une sélection de caractéristiques judicieuses pour

améliorer ses performances. Le F1 score est de 0.746070 ce qui indique que ce modèle n'est pas le meilleur pour classer au mieux les clients susceptibles de souscrire ou pas au produit.

Les modèles SVM, Gradient Boosting, Régression Logistique, AdaBoostClassifier, et RandomForestClassifier ont tous montré des performances solides dans différents aspects, ce qui les rend candidats prometteurs pour l'optimisation et l'utilisation future dans cette tâche de classification.

C. Sélection de 3 modèles de classification

La sélection des 3 modèles de classification (SVM, Gradient Boosting, et Régression Logistique) s'est basée sur une approche méthodique, en tenant compte de plusieurs critères essentiels pour répondre à l'objectif de la banque, qui est de prédire la souscription à un produit de dépôt à terme :

- **Performances Globales :** Nous avons commencé par évaluer les performances globales de chaque modèle en utilisant un ensemble de données de validation distinct. Notre objectif était de maximiser l'accuracy, la précision le rappel et le F1- tout en minimisant le risque de surapprentissage. Ces métriques sont cruciales pour garantir que le modèle puisse bien classer les clients susceptibles de souscrire au produit.

- **Complexité des Modèles :** Nous avons également pris en compte la complexité des modèles. Le SVM et la Régression Logistique sont des modèles relativement simples et faciles à interpréter. En revanche, le Gradient Boosting est plus complexe et peut nécessiter un ajustement plus fin des hyperparamètres. Nous avons évalué leur adaptabilité aux données et à l'objectif de la banque.

- **Temps de Calcul :** Le temps de calcul nécessaire pour former et prédire avec chaque modèle a été un facteur important. La Régression Logistique est le modèle le plus rapide, suivie du SVM, du Gradient Boosting. La rapidité d'exécution peut être un atout précieux dans un environnement opérationnel.

En fin de compte, nous avons choisi ces 3 modèles en raison de leurs performances solides, de leur capacité à gérer la classification binaire, de leur adaptabilité aux données spécifiques de la banque, et de leur potentiel pour être interprétés.

VI. L'optimisation des modèles de classification

A. Sélection des hyperparamètres

Pour maximiser les performances de ces trois modèles, nous avons entrepris une étape d'optimisation des hyperparamètres. Chaque modèle a été soumis à une recherche systématique des meilleurs hyperparamètres à l'aide de l'outil GridSearchCV. Voici un aperçu des hyperparamètres que nous avons ajustés pour chaque modèle :

SVM :

- **Le choix du noyau (kernel)** : Nous avons exploré plusieurs types de noyaux, notamment linéaire, gaussien et polynomial, pour déterminer celui qui convient le mieux aux données de la banque.
- **Le paramètre de régularisation C** : Nous avons ajusté C pour contrôler la flexibilité du modèle et éviter le surapprentissage.

Gradient Boosting :

- **Le nombre d'arbres (n_estimators)** : Nous avons recherché le nombre optimal d'arbres à inclure dans l'ensemble pour maximiser la performance.
- **La profondeur maximale des arbres (max_depth) :** Nous avons ajusté la profondeur des arbres de décision pour éviter la complexité excessive.
- **Le taux d'apprentissage (learning_rate) :** Différents taux d'apprentissage ont été testés pour contrôler la convergence du modèle.

Régression Logistique :

- **le coefficient de régularisation (C) :** nous avons recherché à optimiser la force de la régularisation dans la régression logistique ; autrement dit il a été testé plusieurs niveaux de tolérance aux erreurs d'ajustement ; ce paramètre permet de prévenir le surentrainement.
- **L'algorithme de résolution (solver) :** le type d'algorithme le plus adapté aux données a été recherché.

- **Le type de pénalisation (penalty) :** entre “Lasso” et “Ridge”, est recherché la valeur pour une meilleure performance du modèle.

En optimisant ces hyperparamètres, nous visons à améliorer les performances des modèles sélectionnés, en garantissant une meilleure capacité de prédiction tout en évitant le surapprentissage. Cette étape d'optimisation est cruciale pour que les modèles soient ajustés de manière optimale aux données de la banque et puissent fournir des prédictions précises pour la souscription au produit de dépôt à terme.

B. Validation croisée et ajustement des modèles

Pour chacun des 3 modèles sélectionnés, nous avons donc pu instancier des objets GridSearchCV. Nous avons choisi, à chaque de définir “cv” à 5, qui est une valeur courante pour ce paramètre. Ainsi, au sein de l'ensemble d'entraînement, les données seront divisées en 5 sous-ensembles qui seront, à chaque itération, chacuns utilisés comme ensemble de validation, tandis que les autres folds formeront l'ensemble d'entraînement.

Après l'optimisation des hyperparamètres, nous avons procédé à une étape cruciale de validation croisée pour évaluer la performance de chaque modèle avec les hyperparamètres optimisés. Cette approche nous a permis d'obtenir une estimation plus fiable de la performance des modèles sur des données inconnues, réduisant ainsi le risque de surapprentissage.

Nous avons utilisé diverses techniques pour nous assurer que les modèles ne souffrent pas de surapprentissage, notamment :

- Régularisation : Nous avons appliqué des techniques de régularisation pour contrôler la complexité des modèles. Cela a permis d'éviter que les modèles ne s'adaptent trop aux données d'entraînement et conservent leur capacité de généralisation.
- Sélection prudente des hyperparamètres : Les hyperparamètres ont été soigneusement sélectionnés pour optimiser les performances des modèles tout en évitant une adaptation excessive aux données d'entraînement.

Optimisation du modèle de Régression Logistique

Nous n’avons pas réussi à améliorer la performance du modèle de Régression Logistique.

Accuracy :

- Score sur ensemble train 0.8129282777523984

- Score sur ensemble test 0.8163544997715853

Optimisation du modèle SVM

On peut noter que l'amélioration de la performance du modèle sur le jeu de test est significative. L'optimisation ne semble pas avoir engendré un surapprentissage.

Accuracy :

- Score sur ensemble train 0.8573549566011878
- Score sur ensemble test 0.8423937871174052

Optimisation du modèle Gradient Boosting

Le gain de performance est également non négligeable même si on peut noter un léger surentrainement.

Accuracy :

- Score sur ensemble train 0.9090909090909091
- Score sur ensemble test 0.8506167199634537

Lorsque des compromis ont dû être faits entre différentes métriques de performance, nous avons tenu compte des besoins métier. Par exemple, nous avons cherché à optimiser la précision tout en maintenant un bon rappel.

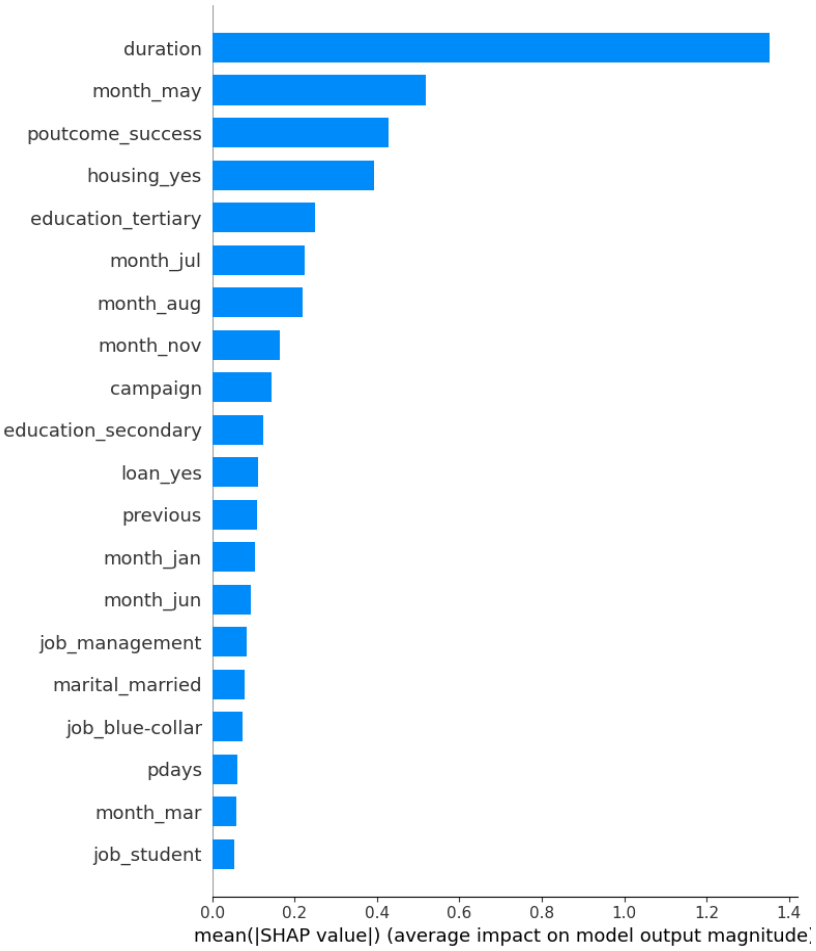
Il est essentiel d'identifier correctement les clients susceptibles de souscrire au produit, tout en minimisant les faux positifs. Cette approche a permis de trouver un équilibre entre la maximisation de la conversion des prospects en clients et la réduction des ressources utilisées pour cibler ces prospects.

VII. Interprétabilité : les facteurs qui influencent la décision

A. Technique d'interprétabilité des algorithmes de machine learning : utilisation de la librairie SHAP

Afin d'être plus performant dans l'interprétation des résultats il est important de connaître l'influence ou le poids des paramètres de notre matrice de départ.

Nous pouvons très bien imaginer l'importance et le gain de temps pour l'utilisateur, que de connaître le ou les paramètres pouvant le plus influencer l'objectif final. Dans notre cas la signature au contrat de la campagne marketing.

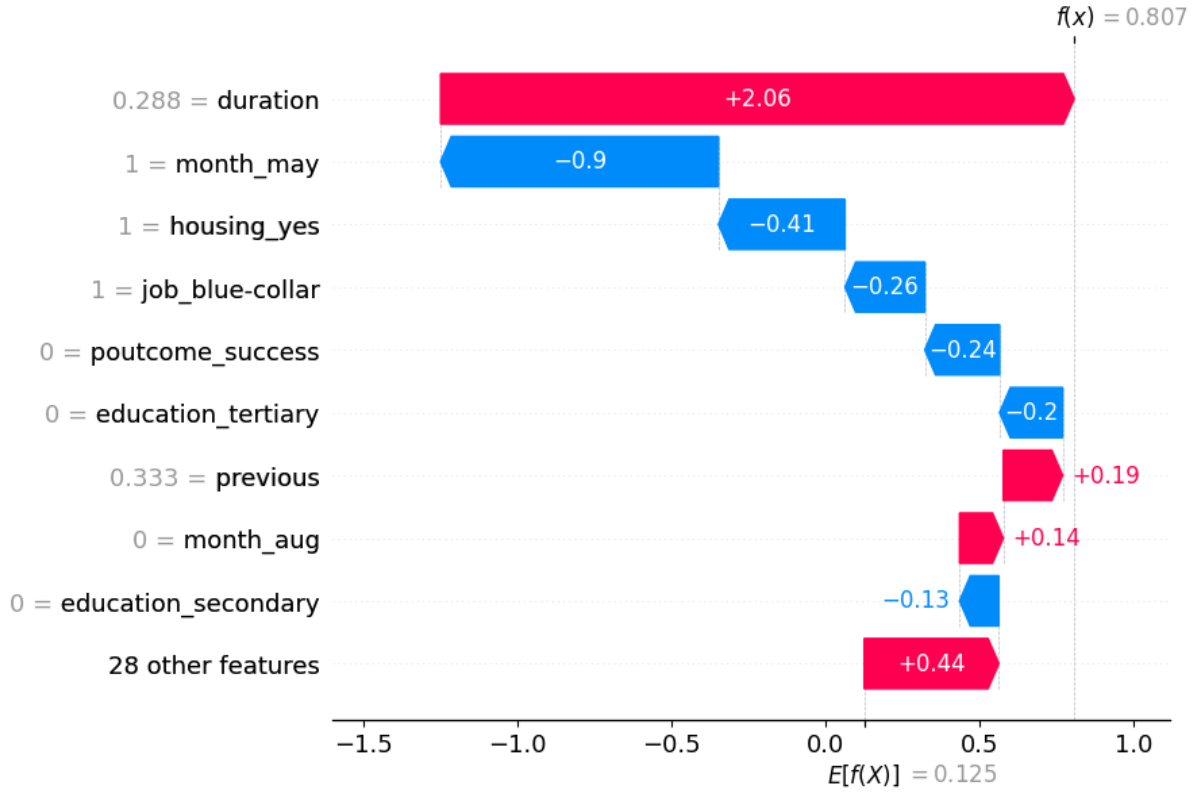


Pour ce faire, la librairie SHAP de Python nous propose un algorithme nous permettant de voir les contributions de chaque caractéristique à chaque prédiction.

Nous l'avons appliqué sur un des trois modèles optimisés, le modèle de régression logistique. Ci dessous le graphique représentant les 20 critères ayant le plus de poids dans la décision finale sur les 37 critères initiaux.

Nous constatons que les quatres critères les plus importants sont : duration, month_may, poutcome_success, housing_yes. Duration restant de loin le critère le plus prépondérant.

Une autre représentation de cette distribution est le diagramme dit de waterfall.



B. Résultats d'interprétation : recommandations aux équipes métier

La durée de l'appel est au centre de la performance de la campagne

Afin d'apporter une expertise au marketing pour optimiser les souscriptions au dépôt, il est intéressant de noter plusieurs éléments :

- La durée d'appel est l'élément le plus important à prendre en compte.

En effet plus cette durée est longue plus le client aura de chance de souscrire au dépôt, ce qui est cohérent avec l'intérêt que ce dernier pourrait apporter. Nous pourrions proposer au marketing de privilégier ce moyen de communication.

L'impact de la durée d'appel est beaucoup plus important que n'importe quelle donnée personnelle. Il sera utile d'insister sur le fait que la capacité du télémarketeur à intéresser le client est un facteur influant sur la souscription au dépôt bien plus que n'importe quelle caractéristique personnelle du client.

Il sera donc pertinent de déployer les techniques que les équipes métiers connaissent bien pour augmenter la durée de l'appel : anticiper les objections, poser des questions ouvertes ou pourquoi pas proposer une rapide démonstration chiffrée de la rentabilité du produit.

L'impact de l'endettement et de l'âge sur la propension à épargner via ce produit

Les personnes ayant un crédit immobilier ont moins de chance de souscrire au produit de dépôt à terme. En mettant de côté la date de l'appel, c'est le second facteur impactant le plus la décision dans le modèle de régression logistique. Optimiser les résultats peut donc passer par le ciblage des personnes n'ayant pas de crédit immobilier en cours.

En revanche, et ce même si c'est dans une plus faible mesure, le fait que le client ait un autre type de crédit impacte positivement la décision.

Ces 2 variables concernant l'endettement peuvent elles-mêmes être corrélées à l'âge. Il peut être intéressant de s'intéresser à l'importance de ce dernier dans la décision.

De manière générale, on sait que les “moins de 30 ans” épargnent peu. Par exemple, en France, moins de 8% d’entre eux épargnent selon le dernier rapport de la Banque de France sur l’épargne des ménages.

Or, ici, l’âge ne se retrouve pas dans les variables les plus importantes et son impact est quasiment neutre. Il est à noter que parmi les clients contactés, très peu ont moins de 30 ans si on se réfère au diagramme de distribution de l’âge. La banque a soit une clientèle bien spécifique, soit une sélection a déjà été faite pour optimiser les résultats de la campagne.

La souscription à un produit lors d’une campagne précédente diminue les chances de succès

A première vue, sans analyse, il peut être difficile de déterminer si, pour un client donné, la souscription à un produit d’épargne lors d’une campagne précédente est un élément plutôt positif ou négatif pour considérer les chances de la banque de compléter la transaction avec lui pour ce produit de dépôt à terme.

D’après les résultats obtenus grâce à la librairie SHAP pour interpréter nos algorithmes, un client ayant souscrit à un produit lors d’une campagne marketing précédente aura moins tendance à souscrire pour le produit qui nous concerne ici. On peut présumer que la diminution de sa capacité d’épargne résiduelle prend le pas sur le fait que ce client ait déjà été réceptif pour un produit bancaire.

Les équipes métiers peuvent en tenir compte dans le ciblage de leur clientèle.

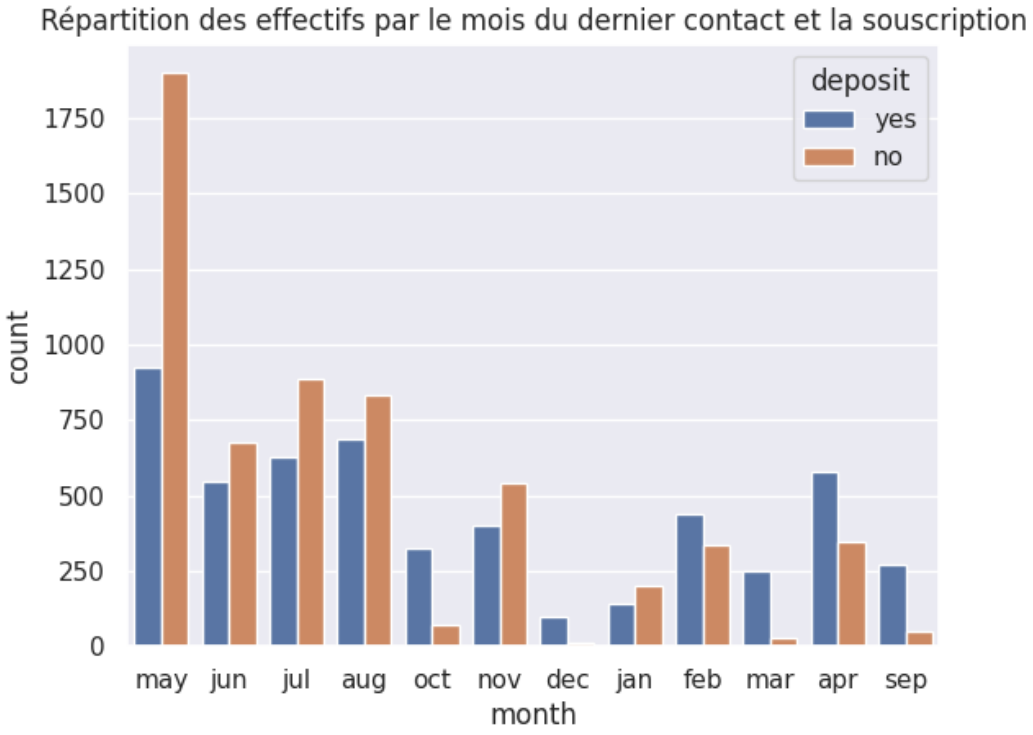
Catégories métiers et niveau d’étude : privilégier les cadres

L’interprétatibilité concernant la variable de niveau d’étude n’est pas évidente puisque des études au niveau secondaire comme au niveau supérieur impactent négativement le résultat.

En revanche, en se référant à la catégorie de métier, les ouvriers-ères ne représentent pas une catégorie prometteuse pour les performances commerciales de la banque pour ce produit alors que la catégorie “management” impacte légèrement mais positivement la décision.

Les périodes de dernier contact : résultats à nuancer

En s'en tenant strictement à l'interprétation, on pourrait dire que le mois de mai est à éviter. Néanmoins, on a vu dans l'exploration que le mois de mai représente plus de 25% des derniers contacts.



De manière générale, on dirait que la campagne s'est terminée entre mai et août puisque c'est sur ces mois qu'on retrouve la plupart des derniers contacts (plus de 60% sur ces 4 mois).

Il semble donc difficile de faire une recommandation en termes de période d'appel sans se renseigner sur l'historique et le fonctionnement de la campagne auprès des équipes métier.

Conclusion

Au moment de la réalisation de cette étude, le service marketing de la banque obtient un taux de succès de 52,8 %. On peut évidemment s'attendre à ce que ce client recherche une amélioration de ce résultat. Nos travaux nous permettent d'apporter plusieurs solutions pour l'accompagner dans ce projet.

L'application de modèles de prédiction optimisés sur de nouvelles données clients peut s'avérer efficace et ce, à des niveaux plutôt intéressants. Deux de nos modèles ont obtenu un score accuracy supérieur à 84 % sur un jeu de test.

Un autre intérêt de notre travail est de pouvoir effectuer des recommandations issues de l'interprétabilité des modèles :

1. augmenter la durée de l'appel
2. cibler en priorité les personnes qui n'ont pas de crédit immobilier en cours
3. prioriser les clients n'ayant pas souscrit lors d'une campagne précédente
4. privilégier les cadres

Il peut également être utile de donner aux équipes métiers une représentation visuelle de leur clientèle avec quelques *plots* bien choisis afin de leur permettre d'appréhender au mieux les tendances qui s'en dégagent. Une visualisation de la distribution de plusieurs variables ou de clusters de clients est également une part non négligeable du service que nous pouvons rendre au service marketing.

Perspectives

Pour aller plus loin, nous pouvons nous poser quelques questions. Il réside des axes d'analyse qu'il est possible d'ouvrir. Afin d'avoir un premier aperçu des directions et du potentiel du travail que nous pourrions accomplir en plus, nous allons exposer ces perspectives.

Durée du dernier appel et dépôt : quel risque de colinéarité ?

Lorsque nous examinons la relation entre la durée de l'appel téléphonique et la souscription à un produit, il est important d'adopter une perspective de bon sens. Dans un contexte de télémarketing, la durée de l'appel est souvent influencée par plusieurs facteurs, y compris l'intérêt du client pour le produit proposé. En général, les clients qui sont véritablement intéressés par un produit ont tendance à poser des

questions, à exprimer leurs préoccupations et à chercher des détails supplémentaires. En conséquence, cela peut entraîner des conversations plus longues avec les télémarketeurs.

Lorsqu'un client montre un intérêt actif envers le produit, les télémarketeurs auront plus d'opportunités pour discuter en profondeur des avantages, des caractéristiques et des modalités de souscription. Dans ce contexte, la durée de l'appel s'allonge naturellement, car il s'agit d'une interaction plus impliquée et informative.

D'un point de vue de causalité, il est plausible que l'intérêt du client pour le produit et sa volonté de souscrire influencent la durée de l'appel. En d'autres termes, la relation entre la durée de l'appel et la souscription est susceptible d'être une relation complexe plutôt qu'une causalité simple et directe où une conversation plus longue déclencherait une souscription.

Comparaison avec d'autres produits similaires

Il est essentiel de confronter nos résultats à ceux d'autres banques proposant des produits similaires. Cette démarche remettrait en question nos conclusions et permettrait de déterminer si les performances de notre campagne se comparent favorablement à celles de nos concurrents. En outre, il serait crucial d'analyser l'impact spécifique des télémarketeurs sur les taux de souscription.

Diversification des moyens de prospection

L'exclusivité d'une campagne téléphonique mérite d'être réévaluée. Envisager d'autres canaux de communication, tels que l'email ou les opportunités en agence, pourrait potentiellement améliorer l'efficacité de la promotion du produit de dépôt. Les emails envoyés par notre banque se révèlent plus influents que la publicité classique, surtout lorsqu'il est question d'économies. Il serait donc judicieux de considérer ces pistes pour optimiser la campagne marketing.

Acquisition des données et qualité : auditer les process des équipes métier

Nous pourrions également indiquer aux équipes travaillant sur la constitution du jeu de données d'améliorer la qualité de ces données afin de disposer de moins de valeurs aberrantes ou "unknown", cela nous permettrait d'optimiser encore plus nos modèles.

Il serait également intéressant d'enrichir le jeu de données avec de nouvelles données pertinentes permettant d'aider les modèles à mieux généraliser.