École nationale de la statistique et de l'analyse de l'information

Mastère Data Science - Connaissance Client

Projet Machine Learning

Thème

Optimarketing : Modélisation Prédictive pour des Campagnes Marketing Ciblées

---

Réalisé par

DIAKITE GAOUSSOU ADUAYOM MESSAN

Destiné au professeur

CLAUDE PETIT

2022/2023

Table des matières

Résumé

Dans un monde en constante évolution, la compréhension des comportements des clients est vitale pour toute entreprise. Ce projet de machine learning se concentre sur le développement d'un modèle prédictif pour optimiser les campagnes marketing. L'objectif principal est de prédire la probabilité de réponse des clients à une offre. En ciblant les clients les plus susceptibles de répondre positivement, l'efficacité des campagnes peut être améliorée, tout en réduisant les coûts. Ce rapport détaille notre démarche, de l'exploration des données à l'évaluation du modèle.

Chapitre 1

Introduction

Le marketing est un domaine en constante évolution où la capacité à anticiper les comportements des clients joue un rôle essentiel dans le succès des entreprises. Ce projet de machine learning vise à développer un modèle prédictif avancé pour optimiser les campagnes marketing.. L'objectif principal est de prédire la probabilité de réponse des clients à une offre de produit ou de service. En identifiant avec précision les clients les plus susceptibles de répondre positivement, nous cherchons à accroître le taux de réponse tout en réduisant les coûts associés aux campagnes.
L'ensemble de données que nous utilisons provient de sources fiables et contient des informations riches sur les interactions des clients avec diverses campagnes marketing. Chaque entrée de données représente un client unique, incluant des détails tels que les réponses aux campagnes précédentes, le statut marital, le niveau d'éducation, le revenu du ménage, ainsi que les dépenses en divers produits.

Ce rapport présente une méthodologie complète pour atteindre nos objectifs, en commençant par une exploration minutieuse des données, suivie d'une étape de prétraitement pour nettoyer et préparer les données. Nous aborderons ensuite la modélisation de classification, où nous utiliserons des techniques de machine learning pour prédire les réponses des clients. Enfin, nous évaluerons le modèle pour garantir sa fiabilité et son efficacité.

Le résultat de ce projet a le potentiel d'apporter une valeur significative à l'entreprise, en affinant ses campagnes marketing pour qu'elles soient plus ciblées et moins coûteuses. Cette approche basée sur les données est devenue essentielle dans l'ère numérique actuelle.

Chapitre 2

Présentation de la base de données

2.1 Contexte

Un modèle de réponse peut considérablement améliorer l'efficacité d'une campagne marketing en augmentant les réponses ou en réduisant les dépenses. L'objectif de ce projet est de prédire qui répondra à une offre pour un produit ou un service.

2.2 Collecte des données

La base de données utilisée dans ce projet a été collectée à partir de la plateforme Kaggle, une source fiable de données diverses. Elle contient un ensemble de variables qui capturent les interactions des clients avec diverses campagnes marketing, ainsi que des informations démographiques et comportementales.

2.3 Finalité de la base de données

La base de données comprend les éléments suivants :

— **AcceptedCmp1 à AcceptedCmp5 :** Ces variables indiquent si le client a accepté l'offre dans les campagnes marketing 1 à 5 (1 pour accepté, 0 pour non accepté).

— **Response (cible) :** Cette variable cible indique si le client a accepté l'offre dans la dernière campagne (1 pour accepté, 0 pour non accepté).

— **Complain :** Cette variable prend la valeur 1 si le client a formulé une plainte au cours des 2 dernières années.

— **DtCustomer :** La date d'inscription du client à l'entreprise.

— **Education :** Le niveau d'éducation du client.

— **Marital :** Le statut matrimonial du client.

— **Kidhome :** Le nombre d'enfants en bas âge dans le ménage du client.

— **Teenhome :** Le nombre d'adolescents dans le ménage du client.

— **Income :** Le revenu annuel du ménage du client.

— **MntFishProducts, MntMeatProducts, MntFruits, MntSweetProducts, MntWines, MntGoldProds :** Les montants dépensés par le client pour différents produits au cours des 2 dernières années.

— **NumDealsPurchases, NumCatalogPurchases, NumStorePurchases, NumWebPurchases :** Le nombre d'achats effectués avec des remises, à partir d'un catalogue, directement en magasin ou via le site web de l'entreprise.

— **NumWebVisitsMonth :** Le nombre de visites sur le site web de l'entreprise au cours du dernier mois.

— **Recency :** Le nombre de jours depuis le dernier achat du client.

L'objectif principal de cette base de données est de former un modèle prédictif qui permettra à l'entreprise de maximiser le profit de la prochaine campagne marketing. En identifiant les clients les plus susceptibles de répondre favorablement aux offres, l'entreprise pourra cibler ses ressources de manière plus efficace et améliorer les performances de ses campagnes.

Chapitre 3

Exploration de la base de données

Dans ce chapitre, nous allons explorer en détail la base de données sur laquelle repose notre projet. Cette étape est essentielle car elle nous permettra de mieux comprendre les données que nous allons utiliser pour notre modèle de prédiction de taux de réponse à une campagne marketing. Voici ce que nous allons faire dans cette partie :

— **Examen de la structure des données :** Nous allons commencer par observer la structure globale de la base de données, y compris le nombre de lignes et de colonnes, ainsi que les types de données présents.

— **Identification des valeurs manquantes :** Nous allons rechercher les éventuelles valeurs manquantes dans les données. Cette étape est cruciale pour assurer la qualité de notre analyse, car les données manquantes peuvent influencer nos résultats.

— **Renommage des variables :** Afin de rendre les données plus compréhensibles et faciles à traiter, nous avons renommé les variables en utilisant des termes en français. Cette étape vise à clarifier chaque variable pour une meilleure accessibilité.

— **Analyse descriptive :** Nous effectuerons une analyse descriptive des différentes variables. Cela implique de calculer des statistiques de base telles que la moyenne, la médiane, l'écart-type, etc. Cette analyse nous aidera à obtenir un aperçu des tendances et des caractéristiques importantes des données.

Cette exploration initiale est une étape cruciale pour mieux comprendre nos données

et préparer le terrain pour les analyses plus avancées à venir. Elle nous permettra de prendre des décisions éclairées lors de la construction de notre modèle de prédiction de taux de réponse à la campagne marketing

3.1 Taille de la base de données

Après avoir examiné notre base de données, nous avons constaté qu'elle est composée de 2240 lignes et 29 colonnes. Cette taille indique que nous disposons d'un ensemble de données de taille conséquente pour notre analyse, offrant une variété d'informations sur les clients et leurs interactions avec les différentes campagnes marketing.

Lors de l'examen de la structure de la base de données "Data Campagne", nous avons identifié principalement des données numériques (int64 et float64) ainsi que des chaînes de caractères (object). Cette étape nous guide dans le traitement et l'analyse de chaque variable.

3.2 Analyse des valeurs manquantes

Nous avons créé une carte de chaleur pour visualiser les valeurs manquantes dans notre base de données. Cette méthode permet d'identifier rapidement les colonnes et les rangées qui contiennent des données incomplètes. Sur la carte, les zones jaunes représentent les valeurs manquantes. Cela nous aide à comprendre l'étendue et la distribution des données manquantes dans l'ensemble de données.

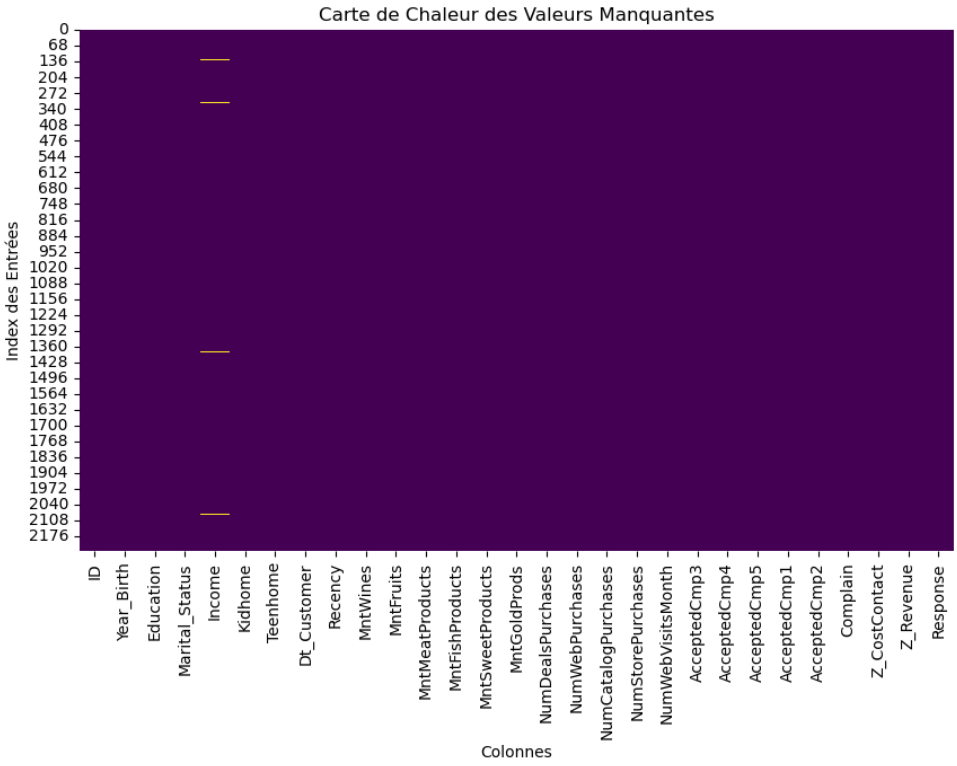

FIGURE 3.1 – Carte de Chaleur des Valeurs Manquantes dans la Base de Données

Il apparaît qu'une seule colonne, *Income*, contient des valeurs manquantes dans notre base de données *Data_ Campagne*. Le nombre total de ces valeurs manquantes est de 24. Il est crucial de connaître ce chiffre pour envisager une méthode d'imputation appropriée. Cette étape est essentielle pour garantir l'intégrité et la qualité de nos données avant de procéder à des analyses plus approfondies.

3.3 Renommage des variables

Afin de faciliter la compréhension et le traitement des données, nous avons renommé les variables de notre base de données. Cette démarche consiste à remplacer les noms de variables en anglais par des termes en français plus descriptifs. Voici le tableau des correspondances entre les anciens et les nouveaux noms de variables :

Nom Original	**Nom en Français**
ID	Identifiant
Year_Birth	Annee_Naissance
Education	Niveau_Education
Marital_Status	Statut_Marital
Income	Revenu
Kidhome	Nb_Enfants
Teenhome	Nb_Adolescents
Dt_Customer	Date_Inscription
Recency	Recence_Achat
MntWines	Depenses_Vins
MntFruits	Depenses_Fruits
MntMeatProducts	Depenses_Viandes
MntFishProducts	Depenses_Poissons
MntSweetProducts	Depenses_Sucreries
MntGoldProds	Depenses_Or
NumDealsPurchases	Achats_Promos
NumWebPurchases	Achats_Web
NumCatalogPurchases	Achats_Catalogue
NumStorePurchases	Achats_Magasin
NumWebVisitsMonth	VisitesWeb_Mois
AcceptedCmp3	Campagne3_Acceptee
AcceptedCmp4	Campagne4_Acceptee
AcceptedCmp5	Campagne5_Acceptee
AcceptedCmp1	Campagne1_Acceptee
AcceptedCmp2	Campagne2_Acceptee
Complain	Plainte
Z_CostContact	Cout_Contact
Z_Revenue	Revenu_Z
Response	Reponse

TABLE 3.1 – Tableau de Renommage des Variables

Cette table de renommage est utilisée tout au long de notre analyse pour assurer une meilleure lisibilité et compréhension des données traitées.

3.4 Analyse Univariée

L'analyse univariée consiste à examiner chaque variable indépendamment afin de résumer et de trouver des modèles dans les données. Dans cette section, nous allons effectuer une analyse univariée pour chacune des variables de notre base de données *Data Campagne*. Cette analyse comprendra l'étude des distributions de fréquences pour les variables catégorielles et des mesures de tendance centrale et de dispersion pour les variables continues.

Pour les variables catégorielles, nous allons utiliser des diagrammes en barres pour visualiser le nombre d'observations pour chaque catégorie. Pour les variables continues, nous allons générer des histogrammes pour observer la distribution des données ainsi que calculer la moyenne, la médiane, l'écart-type et l'asymétrie.

Cette approche nous permettra de comprendre la distribution des variables individuelles, d'identifier les valeurs aberrantes potentielles et de détecter les erreurs de saisie qui pourraient nécessiter un nettoyage supplémentaire. L'analyse univariée est une étape préliminaire cruciale avant de procéder à des analyses bivariées ou multivariées plus complexes.

3.4.1 Analyse des variables catégorielles

Dans cette section, nous examinons les distributions individuelles de plusieurs variables catégorielles de notre base de données.

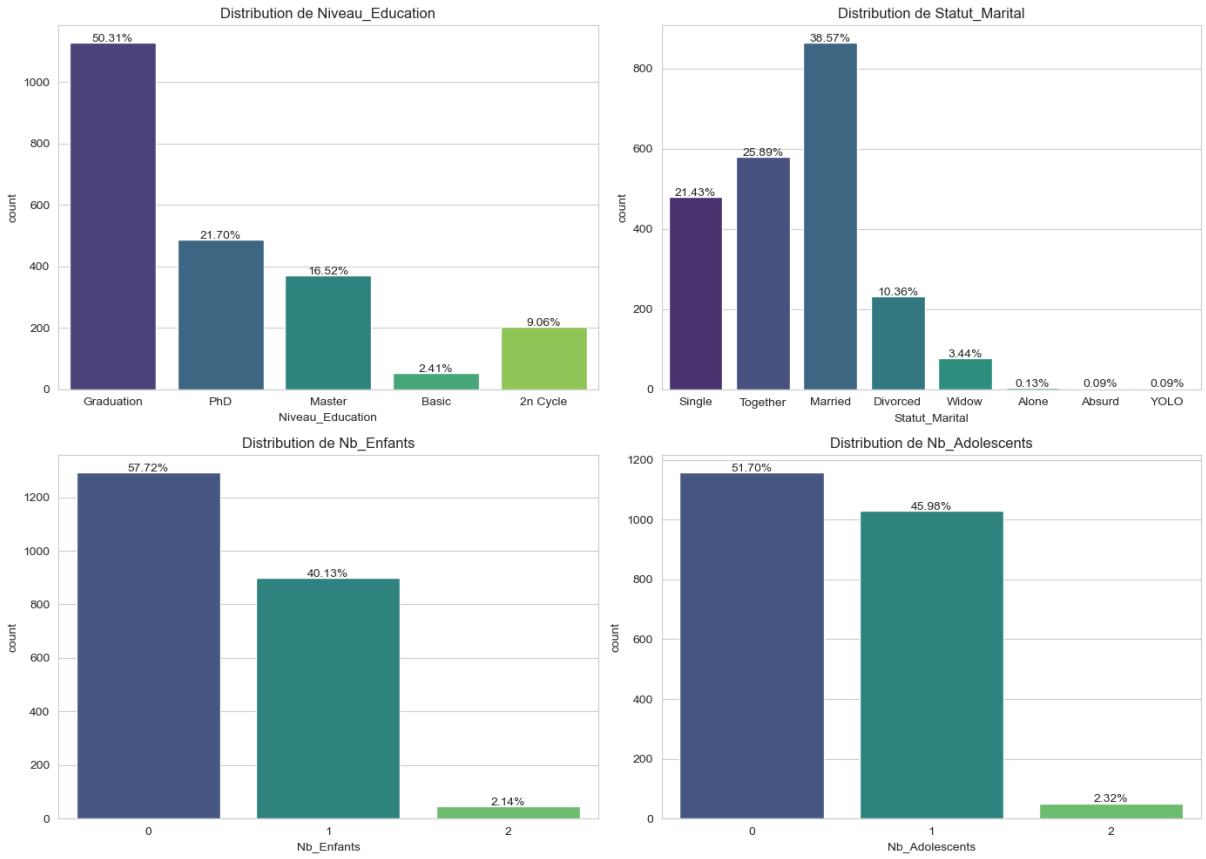

FIGURE 3.2 – Distributions des Variables Niveau_Education, Statut_Marital, Nb_Enfants, et Nb_Adolescents

— **Niveau d'Éducation :** La moitié des clients possèdent un diplôme de type Graduation. Les doctorats et les masters constituent respectivement 21.7% et 16.5% de l'échantillon. Les catégories 'Basic' et '2n Cycle' pourraient être regroupées en vue d'une simplification future.

— **Statut Marital :** Les clients mariés représentent 38.6% de notre échantillon, suivis par ceux en couple (25.9%) et les célibataires (21.4%). Les statuts moins fréquents pourraient être amalgamés en une catégorie unique 'Autre'.

— **Nombre d'Enfants :** La plupart des clients n'ont pas d'enfants, et une minorité a un ou deux enfants. Ceci indique une prédominance de ménages de petite taille.

— **Nombre d'Adolescents :** Plus de la moitié des clients n'ont pas d'adolescents à charge, et très peu en ont deux, soulignant également une tendance vers de plus petits ménages.

Ces observations nous informent sur la démographie des clients et seront prises en

compte lors de l'élaboration des stratégies marketing. De plus, dans le traitement des données, nous envisagerons de regrouper certaines modalités ayant de faibles proportions pour optimiser le modèle prédictif.

Profil de la Clientèle

La clientèle analysée dans notre base de données est majoritairement diplômée, souvent mariée ou en couple, avec une proportion notable de célibataires. L'absence d'enfants dans la majorité des ménages suggère une orientation vers des foyers composés d'adultes ou de couples sans enfants. Cette structure démographique peut influencer les comportements de consommation et sera considérée pour affiner les campagnes marketing, en ciblant des produits et des services appropriés à ces profils.

3.4.2 Analyse des comportements d'achat et visites Web

Nous explorons ici les variables de comptage qui reflètent les comportements d'achat des clients et leur fréquence de visites sur le site web.

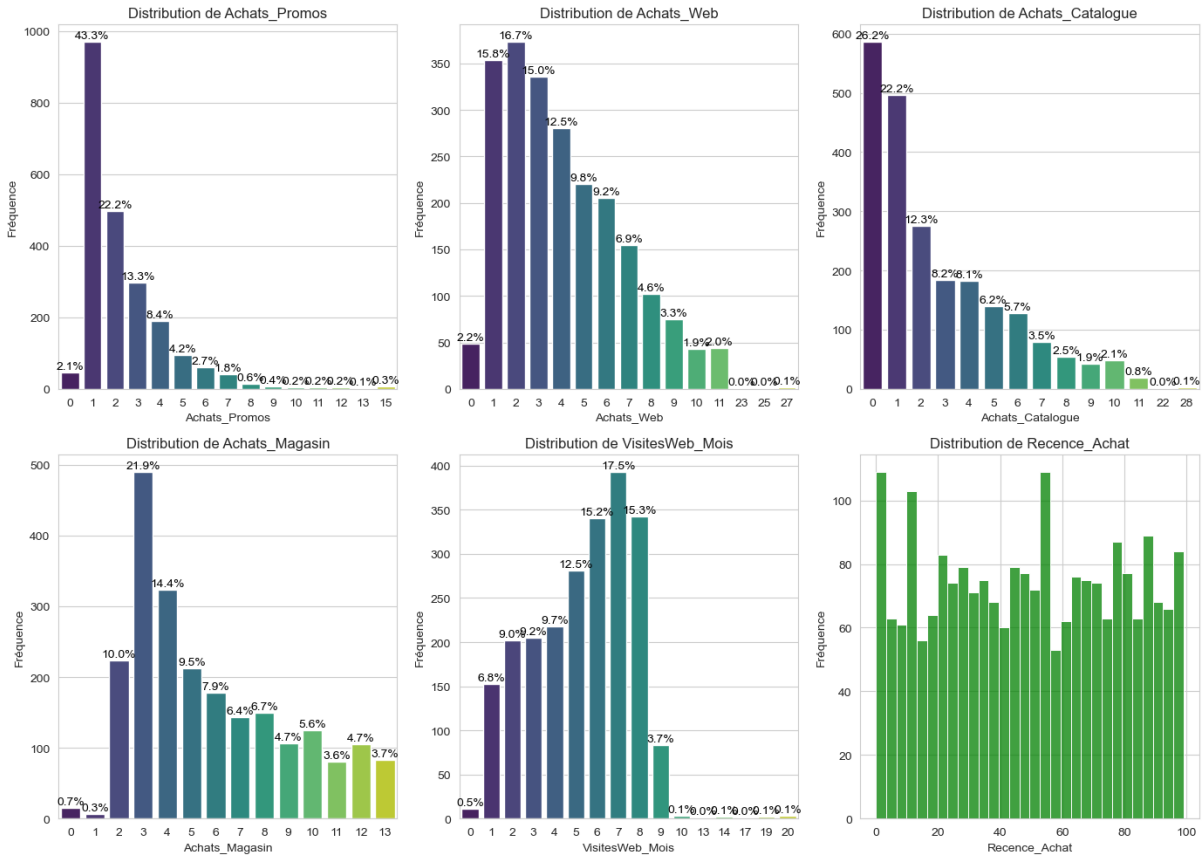

FIGURE 3.3 – Distribution des Achats Promos, Achats Web, Achats Catalogue, Achats Magasin, Visites Web par Mois, et Recence Achat

— **Achats Promos :** Un grand nombre de clients préfèrent réaliser un ou deux achats promotionnels, ce qui montre une inclination à saisir les promotions sans toutefois y recourir de manière excessive.

— **Achats Web :** Les achats en ligne sont répartis de manière variée, indiquant que le site web est un canal d'achat utilisé régulièrement par une portion substantielle des clients.

— **Achats Catalogue :** Une proportion notable de clients utilise le catalogue pour un achat, mais l'utilisation diminue pour des achats multiples, suggérant une préférence pour une utilisation modérée de ce canal.

— **Achats Magasin :** Les achats en magasin tendent à être modérés, avec une majorité des clients qui effectuent entre deux et quatre achats, ce qui peut refléter la fréquence normale de visites en magasin.

— **Visites Web par Mois :** Les visites sur le site web sont fréquentes mais pas

excessives, indiquant un engagement régulier avec la marque en ligne.

— **Recence Achat :** La distribution de la recence d'achat illustre les périodes d'engagement des clients avec l'entreprise, une métrique importante pour évaluer la fidélité et l'activité récente.

Ces indicateurs de comportement d'achat et d'interaction en ligne seront examinés pour regrouper les modalités peu fréquentes et discrétiser les variables continues, afin d'optimiser le modèle prédictif dans la section de traitement des données.

Optimisation des variables pour le modélisation

Les variables de comptage seront regroupées par modalités peu fréquentes et discrétisées pour améliorer la pertinence pour la modélisation. Cette étape de prétraitement vise à simplifier les données et à renforcer la puissance prédictive du modèle que nous chercherons à construire.

3.4.3 Réponses aux Campagnes Marketing et Plaintes

Nous présentons ici l'analyse des réponses aux différentes campagnes marketing et le nombre de plaintes enregistrées.

FIGURE 3.4 – Réponses aux différentes campagnes marketing et distribution des plaintes

Les taux d'acceptation des offres au cours des différentes campagnes marketing varient. Les premières campagnes ont enregistré un taux d'acceptation moins élevé, avec 6.43% pour la première et 1.34% pour la deuxième. Les campagnes ultérieures montrent une légère hausse, se situant autour de 7% pour les troisième, quatrième et cinquième campagnes.

Pour la dernière campagne, la variable "Réponse" indique que 14.91% des clients ont accepté l'offre, tandis que 85.09% ne l'ont pas fait, révélant ainsi la difficulté à élaborer des campagnes qui correspondent aux préférences des clients.

La variable "Plainte" montre que très peu de clients, seulement 0.94%, ont enregistré une plainte, ce qui suggère un niveau de satisfaction général élevé, avec 99.06% des clients

sans plainte officielle.

Ces observations sont importantes pour l'amélioration des futures campagnes marketing, indiquant la nécessité d'une meilleure ciblage et de maintien d'un niveau de satisfaction élevé parmi les clients.

Implications pour les Campagnes Marketing Futures

Ces résultats indiquent un potentiel d'amélioration dans la conception des campagnes marketing pour augmenter les taux de réponse. De plus, le faible nombre de plaintes souligne l'importance de continuer à fournir des services et produits de qualité pour maintenir la satisfaction des clients. L'analyse des plaintes peut également révéler des domaines spécifiques nécessitant une attention pour améliorer l'expérience client.

3.4.4 Statistiques descriptives des dépenses et revenus des clients

Après avoir examiné les distributions, nous résumons les données avec des statistiques descriptives pour chaque catégorie de dépense et les revenus.

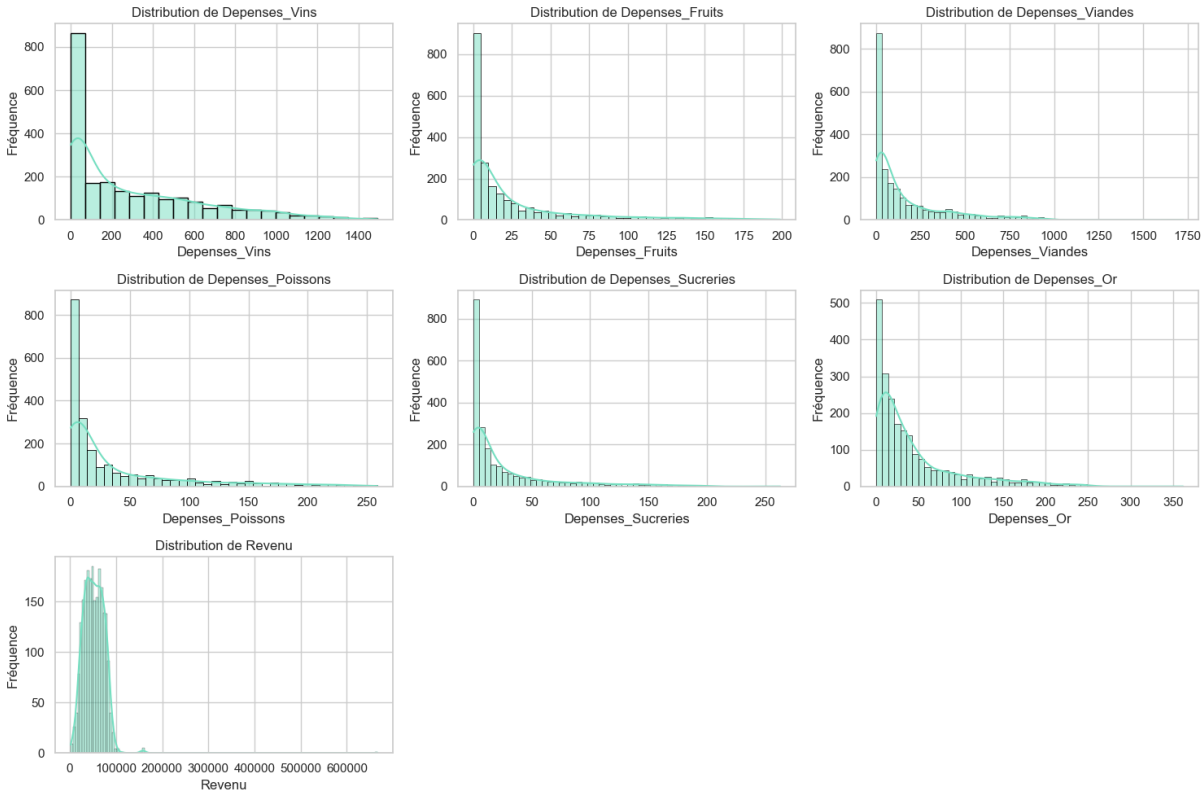

FIGURE 3.5 – Distributions des dépenses des clients sur divers produits et de leurs revenus

Variable	Count	Mean	Std	Min	25%	50%	75%
Max							
Depenses_Vins	2240	303.94	336.60	0	23.75	173.5	504.25
1493							
Depenses_Fruits	2240	26.30	39.77	0	1.00	8.0	33.00
199							
Depenses_Viandes	2240	166.95	225.72	0	16.00	67.0	232.00
1725							
Depenses_Poissons	2240	37.53	54.63	0	3.00	12.0	50.00
259							
Depenses_Sucreries	2240	27.06	41.28	0	1.00	8.0	33.00
263							
Depenses_Or	2240	44.02	52.17	0	9.00	24.0	56.00
362							
Revenu	2216	52247.25	25173.08	1730	35303.00	51381.5	68522.00
666666							

TABLE 3.2 – Statistiques descriptives des dépenses et revenus des clients

Ces statistiques fournissent une vue d'ensemble des tendances centrales et de la dispersion pour chaque variable. La discrétisation de ces variables continues sera effectuée pour la modélisation prédictive. Transformer ces variables en catégories simplifiera le modèle, améliorera son interprétabilité et sa généralisabilité, et aidera à mieux capturer les relations non linéaires potentielles avec la variable cible.

3.4.5 Préparation des variables pour la modélisation

La discrétisation des variables continues permettra de grouper les clients en segments basés sur leurs dépenses et revenu, ce qui est avantageux pour plusieurs raisons :

- Elle réduit la complexité du modèle en diminuant le nombre de valeurs uniques.
- Elle aide à modéliser les tendances importantes sans être influencé par des valeurs extrêmes ou atypiques.
- Elle permet une interprétation plus aisée des résultats du modèle en créant des catégories significatives pour les entreprises.

Cette méthode est essentielle pour personnaliser les campagnes marketing et pour élaborer des stratégies qui résonnent avec différents segments de la clientèle.

Chapitre 4

Traitement de la base de données

Dans ce chapitre, nous abordons le traitement de notre base de données. Le prétraitement est crucial pour assurer la qualité des données qui seront utilisées dans l'analyse et la modélisation.

- **Prétraitement des données**
  - **Imputation des valeurs manquantes :** Les valeurs manquantes, en particulier dans la variable 'Revenu', seront traitées par une méthode d'imputation qui respecte leur distribution.
  - **Discrétisation des variables continues et calcul de l'Âge et de l'Ancienneté :** Les variables continues telles que 'Revenu' et les dépenses par catégorie seront transformées en classes discrètes. L'âge des clients sera déduit de leur année de naissance, et leur ancienneté sera calculée à partir de leur date d'inscription.
  - **Regroupement des modalités :** Pour les variables avec de nombreuses modalités rares, telles que 'Statut Marital' et les différentes mesures d'achat et de réponse aux campagnes, nous regrouperons les modalités pour simplifier l'analyse et améliorer la performance des modèles.

Ces étapes préparatoires sont indispensables pour affiner la qualité de nos données et leur applicabilité dans les modèles prédictifs. Un bon prétraitement est la clé pour obtenir des résultats d'analyse fiables et des prédictions précises.

4.1 Traitement et discrétisation de la variable 'Revenu'

Imputation de la Variable 'Revenu'

La variable 'Revenu' contient des valeurs manquantes qu'il est nécessaire de traiter pour mener à bien nos analyses. Après avoir observé la distribution de cette variable, nous avons décidé d'utiliser la médiane pour l'imputation des valeurs manquantes. Cela est dû à la nature asymétrique de la distribution, où la médiane, moins sensible aux valeurs extrêmes, fournit une meilleure mesure centrale que la moyenne.

	Count	Mean	Std	Min	25%	50%	75%	Max
Revenu	2240.0	52237.98	25037.96	1730.0	35538.75	51381.5	68289.75	666666.0

TABLE 4.1 – Statistiques descriptives de la variable 'Revenu'

Discrétisation de la variable 'Revenu'

Pour la discrétisation, nous avons divisé la variable 'Revenu' en quatre catégories égales basées sur les quartiles. Cela permet de transformer les données continues en une forme catégorielle, facilitant ainsi les analyses et la modélisation.

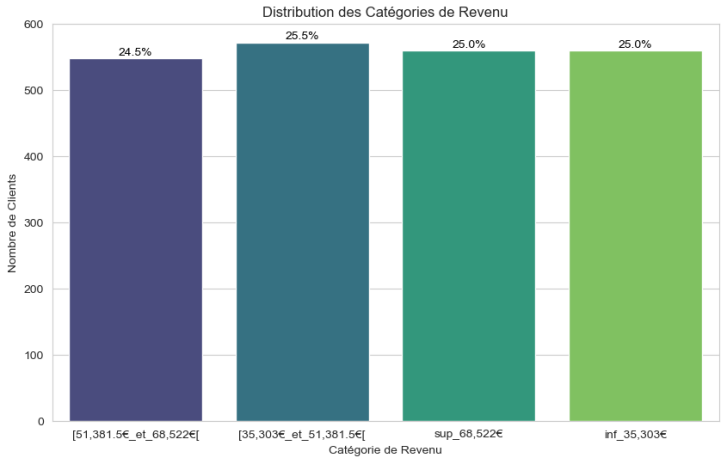

FIGURE 4.1 – Distribution des Catégories de Revenu après Discrétisation

Les catégories sont définies comme suit :

— Revenus inférieurs à 35,303€

— Revenus entre 35,303€ et 51,381.5€

— Revenus entre 51,381.5€ et 68,522€

— Revenus supérieurs à 68,522€

Cette répartition montre que notre base de clients couvre un large spectre de situations économiques, ce qui est bénéfique pour des analyses marketing poussées et la mise en place de stratégies ciblées.

Interprétation des catégories de revenu après discrétisation

La discrétisation en quartiles révèle une distribution équilibrée des revenus parmi notre clientèle. Chaque catégorie représente environ 25% de l'ensemble des clients, reflétant ainsi une diversité économique significative. Cette répartition nous permettra de mieux comprendre et de cibler les différentes tranches de revenus dans nos stratégies marketing.

4.2 Discrétisation des dépenses des clients

Pour approfondir notre compréhension des habitudes de consommation et affiner nos modèles prédictifs, nous appliquons une discrétisation aux variables de dépenses des clients. Cette technique transforme des données continues en catégories ordonnées, ce qui est particulièrement utile pour l'analyse de données et la modélisation statistique.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Depenses_Vins	2240	303.94	336.60	0.0	23.75	173.5	504.25	1493.0
Depenses_Fruits	2240	26.30	39.77	0.0	1.00	8.0	33.00	199.0
Depenses_Viandes	2240	166.95	225.72	0.0	16.00	67.0	232.00	1725.0
Depenses_Poissons	2240	37.53	54.63	0.0	3.00	12.0	50.00	259.0
Depenses_Sucreries	2240	27.06	41.28	0.0	1.00	8.0	33.00	263.0
Depenses_Or	2240	44.02	52.17	0.0	9.00	24.0	56.00	362.0

TABLE 4.2 – Statistiques descriptives des dépenses des clients

Nous segmentons chaque catégorie de dépense selon les quartiles, attribuant ainsi les dépenses à l'une des quatre catégories suivantes : 'Faible', 'Modéré', 'Élevé', et 'Très Élevé'. Cette classification reflète la distribution et le niveau de dépense de chaque client.

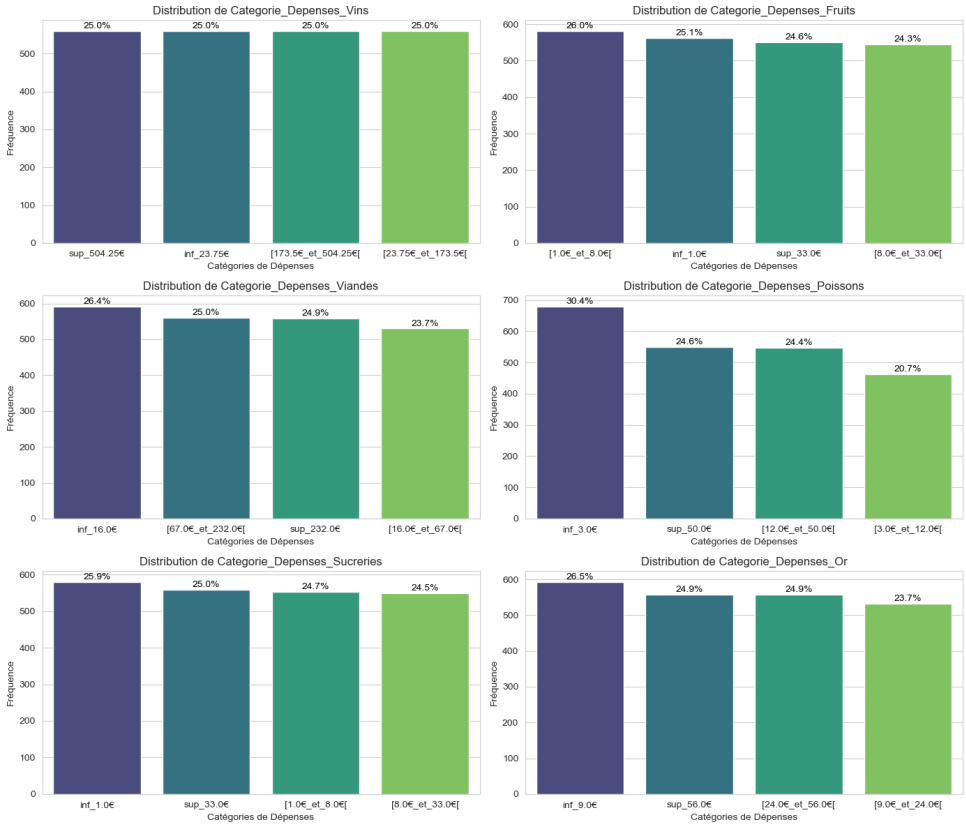

FIGURE 4.2 – Distribution des Catégories de Dépenses après Discrétisation

La discrétisation des données continue en catégories permet une meilleure interprétation des comportements de consommation et facilite l'identification de groupes de clients pour des campagnes marketing ciblées. C'est une étape essentielle pour dériver des insights actionnables et concevoir des stratégies personnalisées adaptées à chaque segment de clientèle.

4.3 Calcul et discrétisation de l'Âge et de l'ancienneté des clients

La connaissance détaillée de nos clients est essentielle pour une stratégie marketing efficace. Cela commence par une compréhension approfondie de l'âge et de l'ancienneté des clients.

Calcul de l'Âge et de l'ancienneté

Nous avons calculé l'âge des clients à partir de leur année de naissance, et leur ancienneté à partir de la date d'inscription, en prenant l'année 2019 comme point de référence. Ces informations nous permettent d'analyser la composition démographique et la fidélité de notre clientèle.

Discrétisation de l'Âge

Pour une analyse plus nuancée, l'âge des clients a été discrétisé en se basant sur les quartiles de la distribution. Les catégories suivantes ont été établies :

- **Jeunes Adultes (moins de 42 ans)** : Cette catégorie inclut nos clients les plus jeunes.
- **Adultes Moyens (entre 42 et 49 ans)** : Cette tranche d'âge représente une part importante de notre base de clients.
- **Adultes Supérieurs (entre 49 et 60 ans)** : Clients d'âge moyen, constituant une proportion significative de notre clientèle.
- **Seniors (plus de 60 ans)** : Une catégorie importante qui peut avoir des besoins et des attentes spécifiques.

La discrétisation aide à aligner nos produits et services avec les attentes et les préférences de ces groupes d'âge distincts.

Analyse des Catégories d'Âge et d'Ancienneté

Nous avons analysé la répartition des clients dans chaque catégorie d'âge et observé une distribution relativement équilibrée, ce qui suggère une clientèle diversifiée. L'ancienneté des clients montre que la majorité sont avec nous depuis 6 ans, soulignant une forte fidélité.

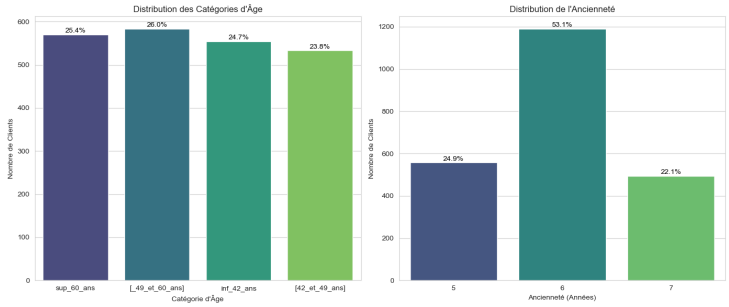

FIGURE 4.3 – Distribution des Catégories d'Âge et d'Ancienneté des Clients

La distribution équilibrée des catégories d'âge et l'ancienneté substantielle des clients soulignent une clientèle diversifiée et fidèle. Ces insights offrent une opportunité précieuse pour élaborer des stratégies de fidélisation et des initiatives de valorisation à long terme. Le regroupement des variables de comptage est donc une étape stratégique dans notre analyse, contribuant à l'efficacité de nos efforts marketing et au développement de relations client solides et durables.

Cette analyse démographique et la compréhension de la fidélité des clients sont cruciales pour concevoir des stratégies marketing ciblées et pour offrir des services personnalisés qui répondent aux besoins de segments spécifiques de notre clientèle.

4.4 Regroupement des variables de comptage

Afin de mieux comprendre les habitudes d'achat et les interactions en ligne de nos clients, nous avons procédé au regroupement des variables de comptage. Cette méthode est essentielle pour simplifier les analyses et renforcer la pertinence des modèles prédictifs, en réduisant la granularité excessive des données et en mettant en évidence les tendances comportementales significatives.

Regroupement des catégories d'achats et de visites

Les variables de comptage, qui incluent les achats et les visites web, ont été regroupées en catégories basées sur leur fréquence. Cette approche permet d'identifier les modèles d'achat et de visite prédominants et de les utiliser pour la segmentation de la clientèle.

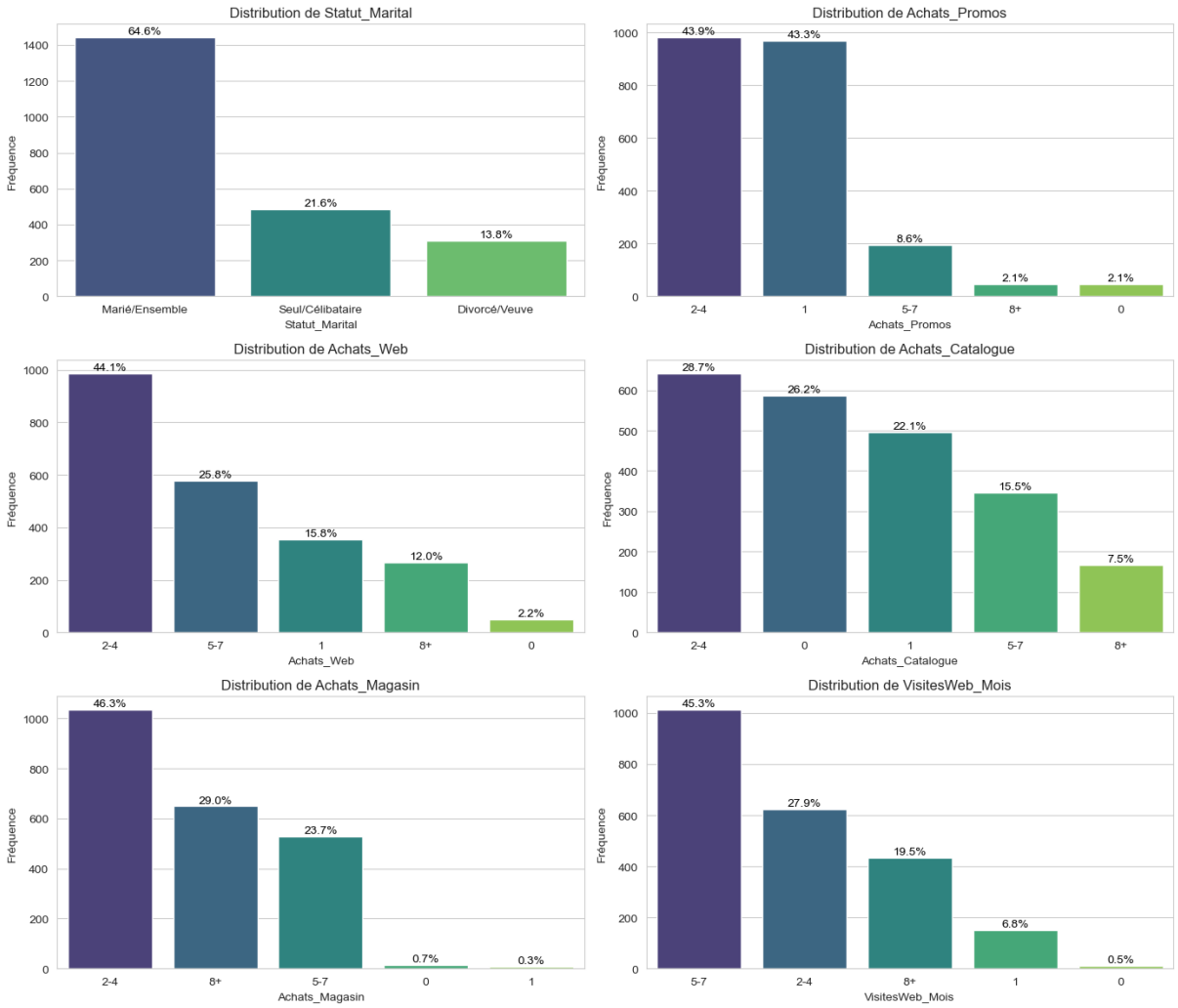

FIGURE 4.4 – Distribution des Variables de Comptage Après Regroupement

Impact sur les Stratégies Marketing

Le regroupement des variables de comptage a des implications directes sur nos stratégies marketing. En identifiant les comportements d'achat et les visites web les plus fréquents, nous pouvons créer des campagnes ciblées qui résonnent avec les activités de nos clients. Cette approche nous permet également de personnaliser les offres et d'optimiser l'engagement client.

4.5 Discrétisation de la récence des achats

La récence des achats est un indicateur clé de l'activité récente des clients. Une compréhension fine de ce paramètre est cruciale pour optimiser nos stratégies de réengagement client.

Signification de la Récence des Achats

La variable 'Recence_Achat' indique le nombre de jours écoulés depuis le dernier achat du client. Cet indicateur est un reflet direct de l'engagement du client avec l'entreprise.

Procédure de Discrétisation

Nous avons discrétisé cette variable en plusieurs catégories basées sur la distribution de la récence des achats pour mieux comprendre et agir sur le comportement d'achat des clients.

Variable	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Recence_Achat	2236	49.15	28.95	0	24	49	74	99

TABLE 4.3 – Statistiques descriptives de la récence des achats

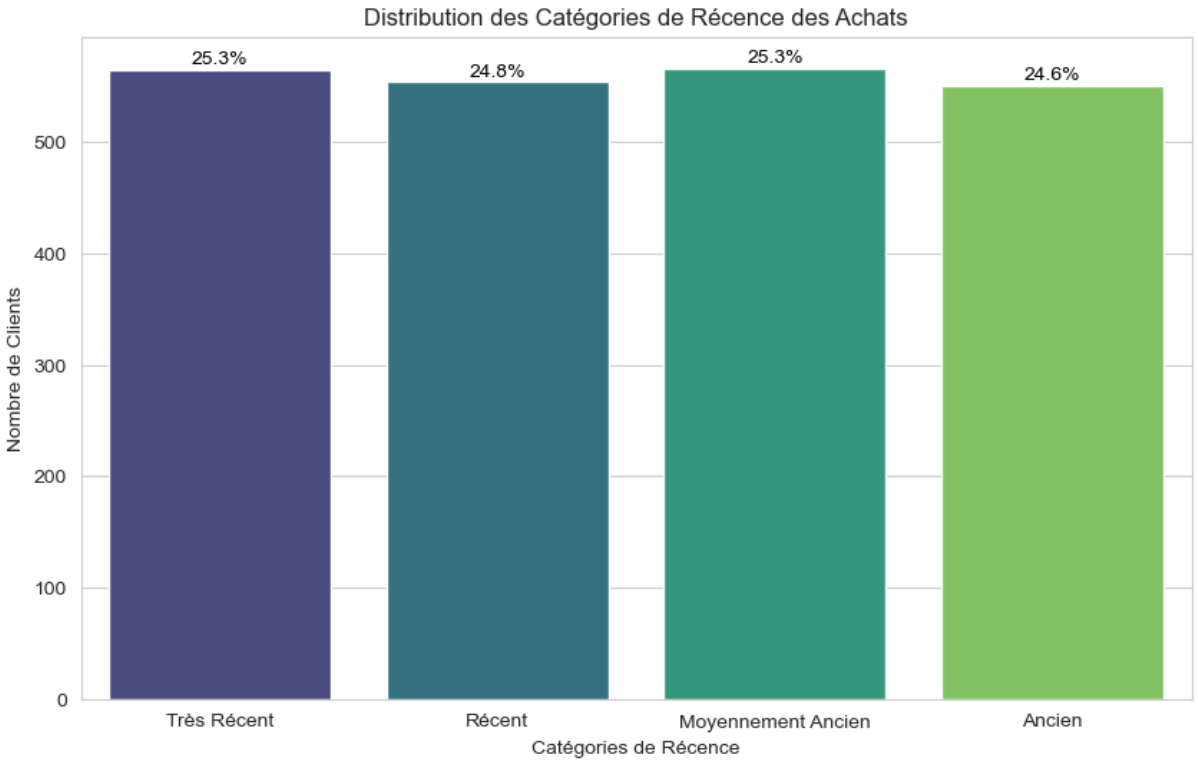

FIGURE 4.5 – Distribution de la Récence des Achats après Discrétisation

Catégorisation de la récence

Les catégories suivantes ont été établies pour refléter les différents niveaux de récence :

— **Très Récent** : clients ayant effectué un achat dans les derniers 24 jours.

— **Récent** : clients ayant effectué un achat entre 25 et 49 jours.

— **Moyennement Ancien** : clients ayant effectué un achat entre 50 et 74 jours.

— **Ancien** : clients n'ayant pas effectué d'achat depuis plus de 74 jours.

Cette classification nous permet d'adapter nos actions marketing à chaque groupe, en mettant l'accent sur le réengagement des clients dans les catégories "Moyennement Ancien" et "Ancien", et en maintenant la dynamique avec ceux classés comme "Très Récent" et "Récent".

La discrétisation de la récence des achats nous aide à délimiter les groupes de clients en fonction de leur interaction récente avec l'entreprise, permettant ainsi de personnaliser les communications et les offres, et d'optimiser les stratégies de fidélisation.

Chapitre 5

Analyse Bivariée et Réponse aux Campagnes

Après avoir préparé nos données à travers le prétraitement et la discrétisation, nous abordons maintenant l'étape de l'analyse bivariée. Cette étape est fondamentale pour comprendre les relations entre les variables deux à deux. Elle est particulièrement éclairante sur la façon dont les variables s'influencent mutuellement et sur leur impact sur les décisions des clients.

Nous nous concentrons sur des variables spécifiques pour examiner leur interaction. L'importance de cette analyse est indéniable, car elle révèle des tendances et des modèles comportementaux qui sont des indicateurs clés pour une interprétation approfondie des données. Ces insights sont cruciaux pour la construction de modèles prédictifs robustes et significatifs.

L'analyse bivariée est également l'occasion d'analyser les différences entre les clients qui ont réagi positivement à nos campagnes et ceux qui ne l'ont pas fait. En identifiant les caractéristiques uniques des clients qui ont montré un intérêt pour nos offres, nous pouvons affiner nos stratégies marketing. Cela nous permet de cibler plus précisément les segments de clients les plus susceptibles de s'engager, améliorant ainsi l'efficacité des campagnes publicitaires et augmentant le taux de conversion.

5.0.1 Réponse en Fonction du Revenu et de la Récence

Relation entre le Revenu et la Réponse

L'analyse de la relation entre le revenu des clients et leur réponse à nos campagnes met en lumière des différences notables. Les clients ayant répondu positivement montrent une distribution bimodale des revenus, avec une tendance à des revenus plus élevés par rapport aux non-répondants.

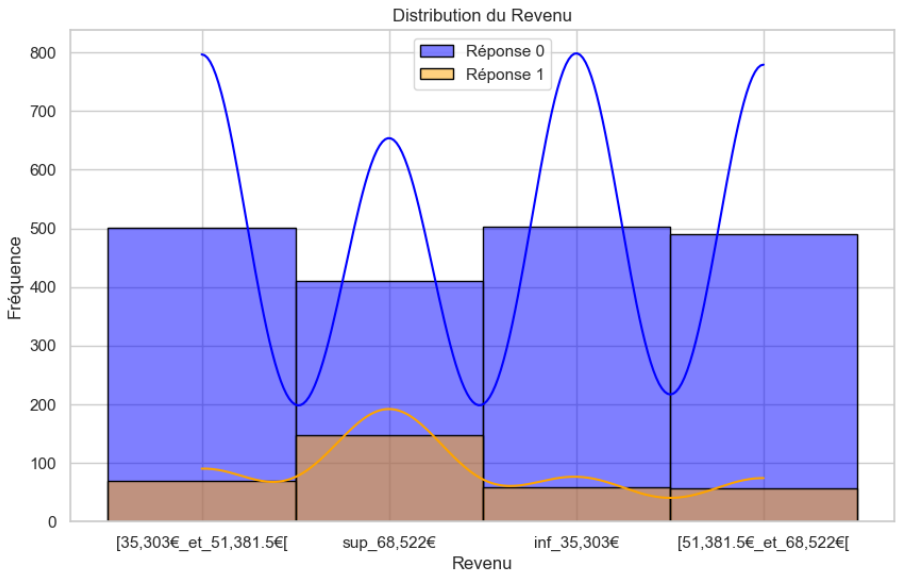

FIGURE 5.1 – Distribution du Revenu par Type de Réponse

Relation entre la Récence d'Achat et la Réponse

L'analyse de la variable "Récence d'Achat" révèle que les clients qui ont répondu positivement ont tendance à avoir réalisé leur dernier achat plus récemment que ceux qui n'ont pas répondu. Cela indique que la fréquence des achats pourrait être un facteur influençant la probabilité de répondre à une campagne.

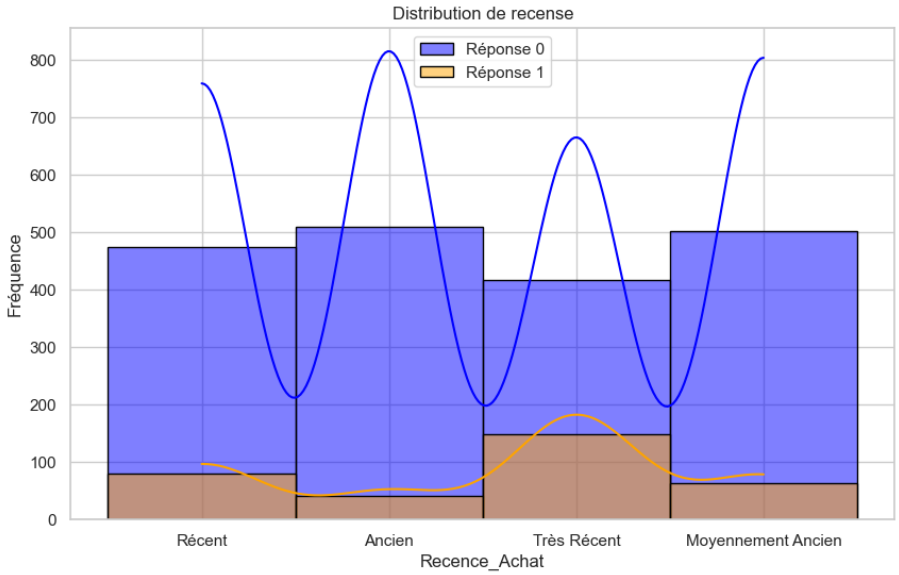

FIGURE 5.2 – Distribution de la Récence d'Achat par Type de Réponse

Ces analyses bivariées nous permettent de comprendre comment les caractéristiques économiques et les comportements d'achat sont liés à la réponse des clients aux campagnes. En explorant ces relations, nous pouvons mieux cibler nos efforts marketing et concevoir des campagnes qui résonnent avec les segments de clients les plus engagés.

5.0.2 Réponse en Fonction du Niveau d'Éducation et du Statut Marital

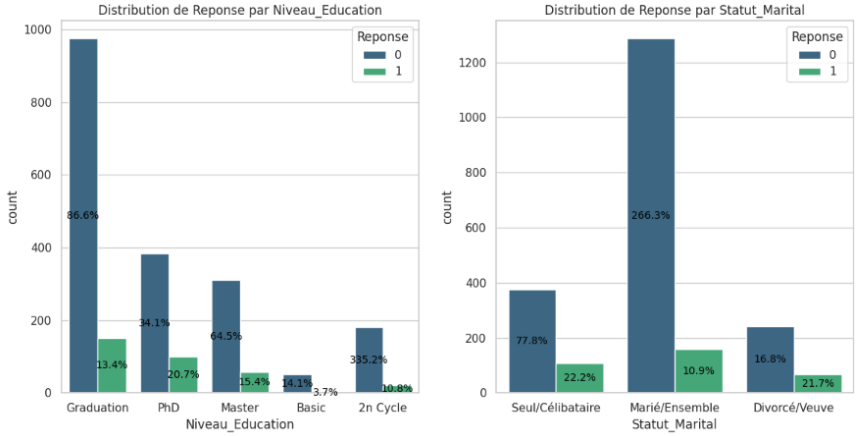

FIGURE 5.3 – À gauche : Distribution de la Réponse par Niveau d'Éducation. À droite : Distribution de la Réponse par Statut Marital

Notre analyse révèle des différences notables dans la réponse aux campagnes en fonction du niveau d'éducation des clients. En particulier, les détenteurs d'un doctorat (PhD) montrent un taux de réponse plus élevé par rapport à d'autres niveaux d'éducation, indiquant une corrélation entre un niveau d'éducation plus élevé et une plus grande réceptivité aux campagnes. À l'inverse, les individus avec un niveau d'éducation '2n Cycle' ont montré le taux de réponse le plus bas.

En ce qui concerne le statut marital, il semble que les individus classés dans les catégories 'Seul/Célibataire' et 'Divorcé/Veuve' sont plus susceptibles de répondre aux campagnes par rapport à ceux qui sont 'Marié/Ensemble'. Cela suggère que les célibataires et les personnes divorcées ou veuves peuvent être des cibles plus réceptives pour certaines campagnes.

Ces observations sont essentielles pour le développement de stratégies marketing ciblées, en permettant de concentrer les efforts sur les segments de la population qui sont les plus susceptibles de répondre positivement aux campagnes.

5.0.3 Réponse en Fonction du Comportement d'Achat

L'analyse des réponses en fonction des comportements d'achat des clients montre des tendances intéressantes. Les clients qui ont effectué un nombre plus élevé d'achats promotionnels et d'achats en ligne sont plus susceptibles de répondre positivement aux campagnes marketing. Ce phénomène pourrait refléter une plus grande réceptivité aux offres promotionnelles et une familiarité avec l'environnement en ligne.

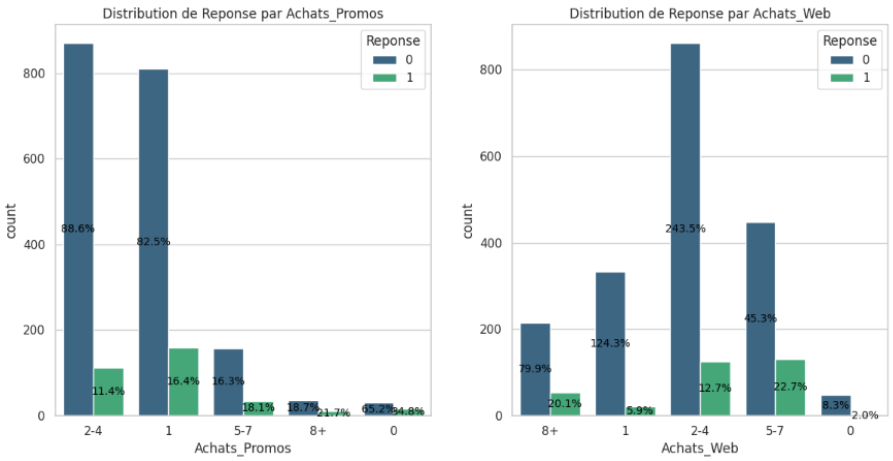

FIGURE 5.4 – À gauche : Distribution de la Réponse par Achats Promotionnels. À droite : Distribution de la Réponse par Achats en Ligne.

Achats Promotionnels

Il apparaît que les clients engagés dans un volume élevé d'achats promotionnels ont un taux de réponse plus important aux campagnes. Cela suggère que les promotions sont un facteur incitatif significatif pour ces clients.

Achats en Ligne

De même, les clients qui réalisent fréquemment des achats en ligne montrent une tendance à répondre plus positivement aux campagnes, ce qui souligne l'importance d'une stratégie marketing numérique bien ciblée.

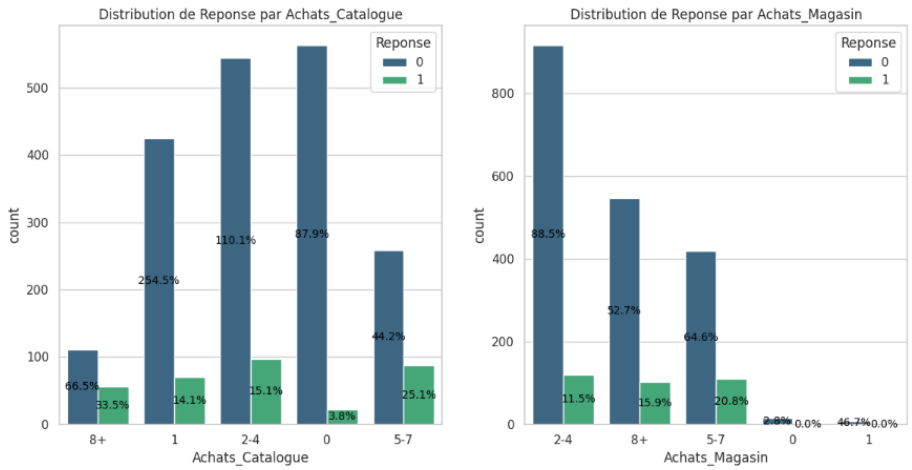

FIGURE 5.5 – À gauche : Distribution de la Réponse par Achats Catalogue. À droite : Distribution de la Réponse par Achats en Magasin.

Achats Catalogue et en Magasin

Les comportements d'achat à travers le catalogue et en magasin révèlent également des schémas de réponse distincts. Les clients qui font des achats moins fréquents via le catalogue ou en magasin ont tendance à moins répondre aux campagnes, indiquant que ces canaux peuvent nécessiter des approches marketing différentes pour stimuler l'engagement.

Cette section révèle que la fréquence et le mode d'achat sont des indicateurs importants de la réponse des clients aux campagnes marketing et doivent être pris en compte lors de la conception de stratégies promotionnelles efficaces.

Réponse en Fonction des Visites sur le Site Web

Les interactions en ligne, mesurées par le nombre de visites sur le site web de l'entreprise, semblent être un bon indicateur de l'intérêt des clients pour les campagnes. Les données montrent que les clients qui visitent le site web fréquemment sont plus susceptibles de répondre aux campagnes, ce qui pourrait indiquer un niveau d'engagement plus élevé avec la marque ou une plus grande réceptivité aux initiatives de marketing numérique.

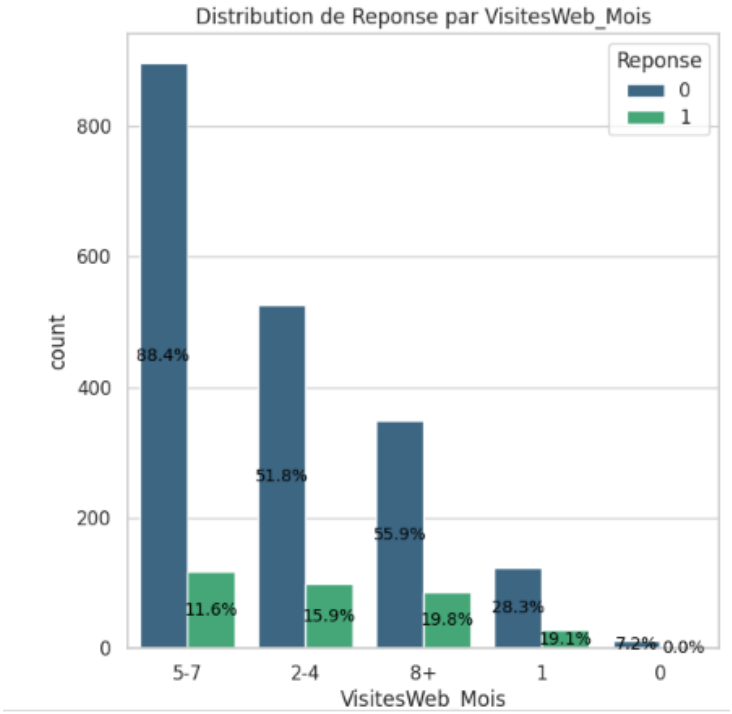

FIGURE 5.6 – Distribution de la Réponse par Visites Web Mensuelles.

Particulièrement remarquable est le taux de réponse plus élevé chez les clients qui visitent le site web plus de 8 fois par mois. Cela suggère que les clients les plus actifs en ligne sont également ceux qui sont le plus engagés et les plus réceptifs aux campagnes marketing. Il est donc crucial pour les entreprises de maintenir une présence en ligne forte et engageante pour encourager ce niveau d'interaction.

5.0.4 Réponse en Fonction des Dépenses

Les graphiques montrent la relation entre les dépenses des clients dans différentes catégories et leur réponse aux campagnes marketing. Les résultats indiquent une tendance où les clients avec des dépenses plus élevées dans certaines catégories ont tendance à répondre plus positivement aux campagnes.

Dépenses en Vins et Fruits

Les clients qui dépensent davantage en vins et fruits semblent plus enclins à répondre aux campagnes, ce qui pourrait refléter un intérêt pour des produits de qualité ou un pouvoir d'achat plus élevé.

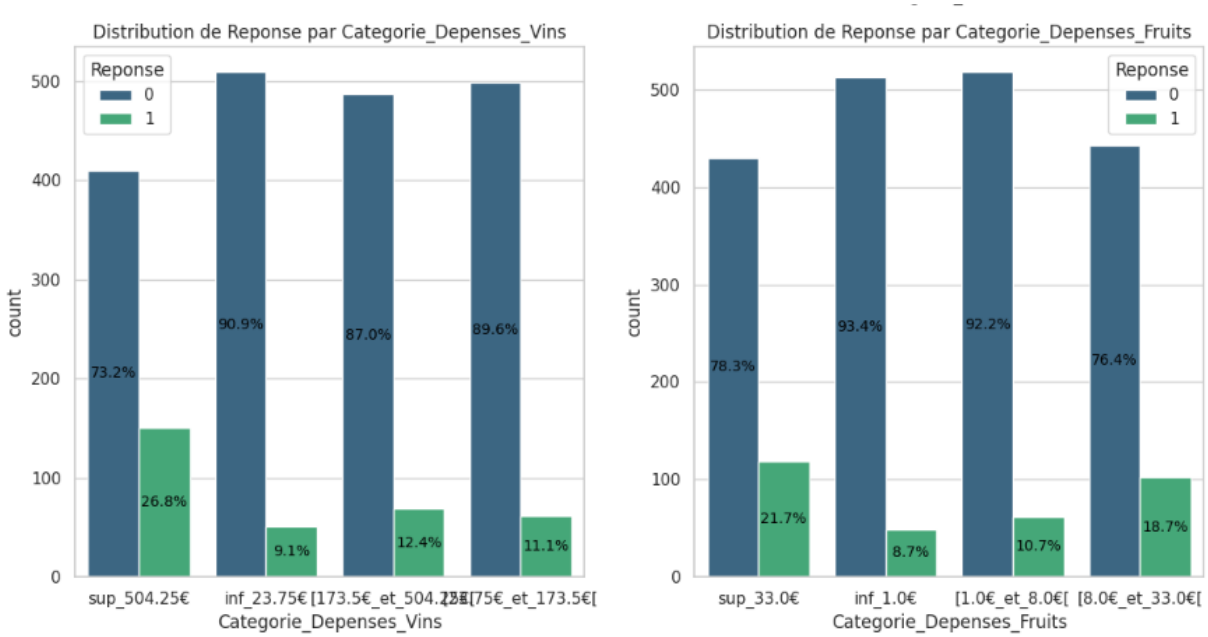

FIGURE 5.7 – Distribution de la réponse par catégorie de dépenses en vins et fruits.

Dépenses en Viandes et Poissons

La même tendance est observée pour les viandes et poissons, avec une réponse positive plus fréquente chez les clients ayant des dépenses modérées à élevées dans ces catégories.

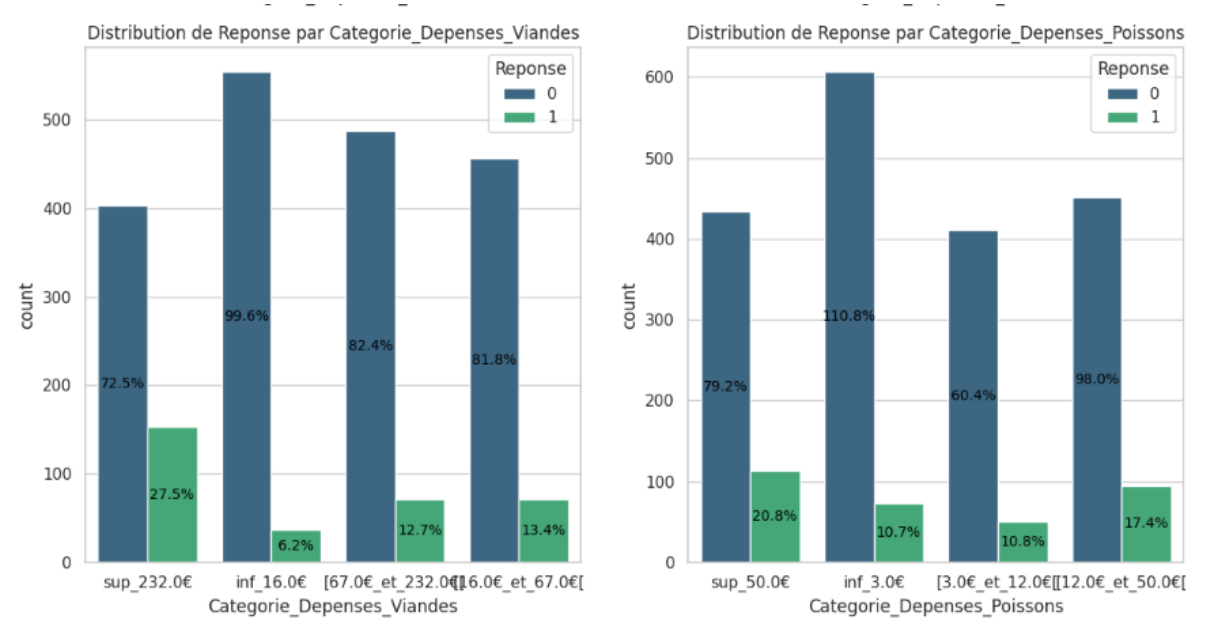

FIGURE 5.8 – Distribution de la réponse par catégorie de dépenses en viandes et poissons.

Dépenses en Sucreries et Or

Enfin, les dépenses en sucreries et en or montrent que les clients qui dépensent plus dans ces catégories ont également un taux de réponse plus élevé, suggérant un goût pour le luxe ou les plaisirs occasionnels.

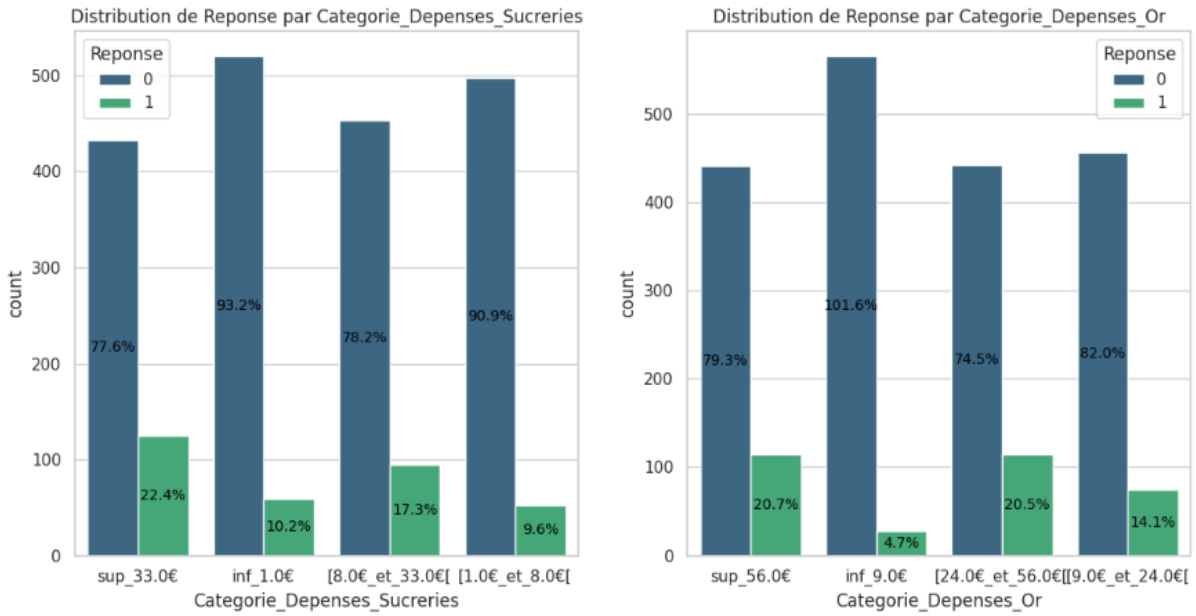

FIGURE 5.9 – Distribution de la réponse par catégorie de dépenses en sucreries et or.

Ces insights permettent de mieux comprendre les habitudes de consommation des clients et peuvent guider les entreprises dans la création de campagnes marketing plus ciblées et efficaces.

Chapitre 6

Modélisation et Evaluation

Dans cette phase cruciale du projet, nous appliquons des méthodes statistiques avancées et des techniques d'apprentissage automatique pour construire des modèles qui vont prédire la réponse des clients aux campagnes marketing. Notre analyse se fonde sur les caractéristiques démographiques, comportementales et transactionnelles des clients pour identifier les déterminants clés des réponses aux campagnes.

- Nous évaluerons plusieurs modèles prédictifs, en cherchant à identifier celui qui présente les meilleures performances.
- Parmi les techniques envisagées figurent la régression logistique, les arbres de décision, les forêts aléatoires, et d'autres modèles plus sophistiqués tels que le boosting et les réseaux de neurones.
- La validation croisée sera employée pour garantir la robustesse de nos modèles, assurant ainsi leur applicabilité à de nouveaux ensembles de données.
- L'interprétabilité des modèles sera un critère important pour permettre une compréhension et une application aisées des résultats par les équipes marketing.

La modélisation est une étape déterminante pour convertir nos analyses en actions et stratégies marketing concrètes. Elle vise à optimiser l'utilisation de nos ressources marketing et à maximiser l'impact de nos actions promotionnelles.

Présentation de la Variable Cible 'Réponse'

Nous nous intéressons maintenant à la variable cible de notre étude, nommée 'Réponse', qui indique si un client a répondu positivement (1) ou non (0) à une campagne marketing. Comme le montre le graphique ci-dessous, nous observons une distribution déséquilibrée entre les deux classes de réponse.

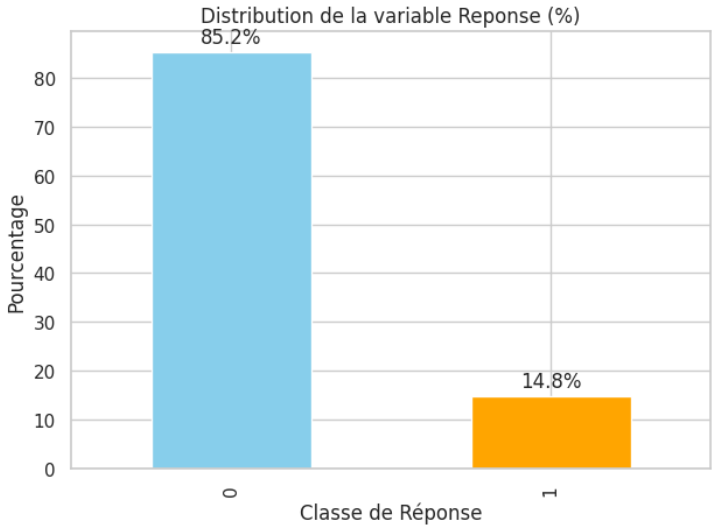

FIGURE 6.1 – Distribution de la variable Réponse en pourcentage.

La majorité des clients, soit 85.2%, n'ont pas réagi positivement à la campagne (Réponse 0), tandis que seulement 14.8% ont répondu favorablement (Réponse 1). Ce déséquilibre de classe peut présenter des défis lors de la modélisation, car les modèles pourraient être biaisés en faveur de la classe majoritaire. Il sera donc nécessaire d'adopter des stratégies spécifiques, telles que le suréchantillonnage, le sous-échantillonnage ou l'utilisation de métriques d'évaluation appropriées comme la précision, le rappel et le score F1, pour s'assurer que notre modèle prédit correctement les deux classes.

6.1 Préparation des Données pour la Modélisation

Dans le cadre de la préparation de notre base de données pour la modélisation, une attention particulière a été portée sur la transformation des variables catégorielles. Nous avons opté pour le codage à chaud, également connu sous le nom de One-Hot Encoding,

une technique de prétraitement des données essentielle pour les modèles de machine learning.

Le codage à chaud permet de transformer les variables catégorielles, qui sont généralement représentées par des chaînes de caractères, en un format numérique qui peut être interprété par les algorithmes. Le processus de transformation est décrit ci-dessous :

1. **Identification des Variables Catégorielles :** Nous avons d'abord identifié toutes les colonnes de type 'object' dans notre dataset, ces dernières représentant les variables catégorielles à encoder.
2. **Application du Codage à Chaud :** Ensuite, nous avons appliqué le codage à chaud à ces variables, transformant chaque catégorie en une nouvelle colonne binaire. Par exemple, si une variable catégorielle contient trois catégories uniques, elle sera convertie en trois colonnes binaires distinctes.
3. **Intégration dans le Dataset :** Après le codage, notre jeu de données s'est enrichi de nouvelles colonnes binaires, augmentant ainsi sa dimensionnalité mais rendant l'information accessible pour l'analyse algorithmique.

Cette étape a été réalisée avec soin, en utilisant la fonction 'get$_d ummies$'$delabibliothquePandasenPy$

Remarque : Le choix de cette technique est guidé par sa capacité à préserver l'intégralité des informations contenues dans les variables catégorielles, tout en les rendant interprétables pour divers modèles prédictifs.

6.2 Analyse de Corrélation avec la Variable Cible 'Réponse'

Notre étude approfondie sur l'impact des différentes caractéristiques des clients sur la variable cible 'Réponse' commence par une analyse de corrélation. Cette analyse est cruciale pour déterminer les facteurs ayant une influence significative sur la propension des clients à répondre positivement aux campagnes marketing.

6.2.1 Méthodologie de l'Analyse de Corrélation

1. Nous avons calculé les coefficients de corrélation entre 'Réponse' et les autres variables pour évaluer la force et la direction de leurs relations linéaires.
2. Les 40 variables avec la plus forte corrélation, qu'elle soit positive ou négative, ont été identifiées pour une investigation plus détaillée.
3. Nous avons ensuite visualisé ces relations à l'aide d'un graphique en barres, facilitant la reconnaissance des variables les plus pertinentes.

Résultats Significatifs

L'analyse a permis de mettre en lumière des variables clés qui méritent une attention particulière dans la conception de nos futures stratégies marketing. En identifiant ces facteurs influents, nous pouvons affiner nos approches pour mieux cibler et engager notre clientèle.

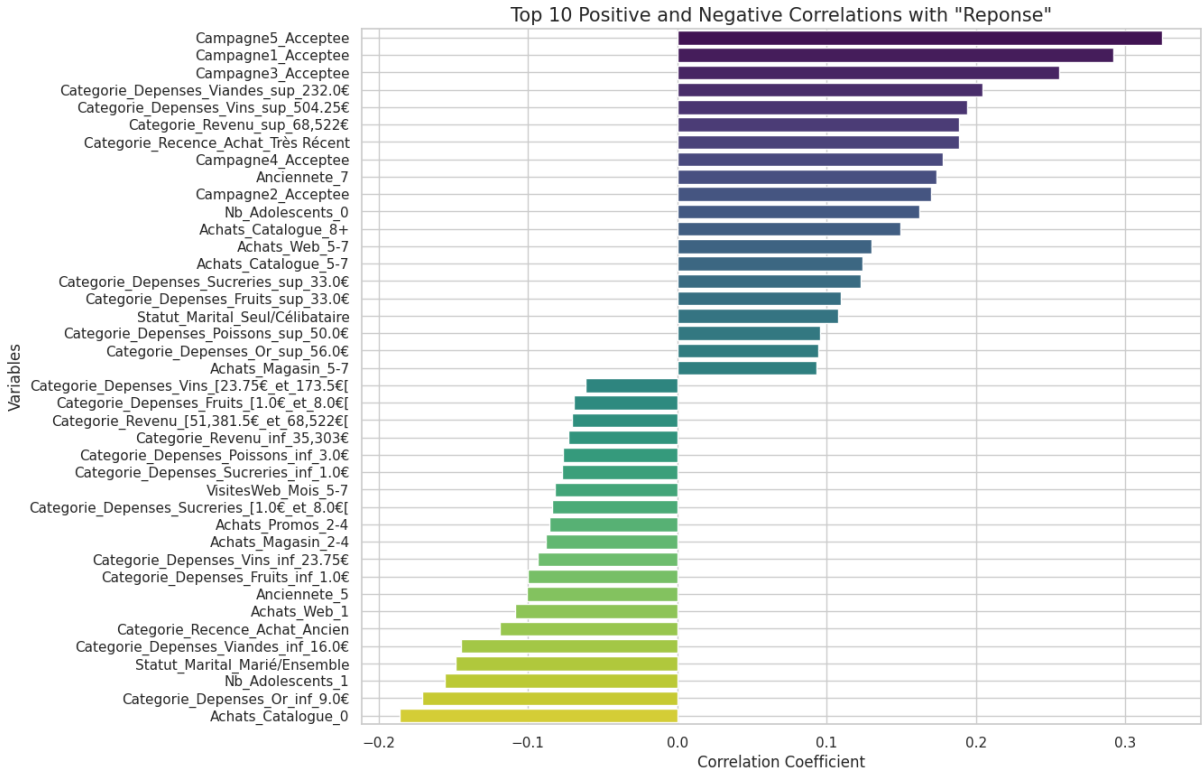

FIGURE 6.2 – Top 40 Variables Corrélées avec 'Réponse'

6.2.2 Interprétation des Variables Clés Corrélées avec 'Reponse'

Notre analyse des corrélations avec la variable 'Reponse' révèle des indicateurs importants pour la réussite des campagnes marketing. Elle met en évidence l'importance de l'expérience des clients avec les campagnes précédentes, leurs dépenses, leurs revenus, la récence de leurs achats, ainsi que leur situation familiale.

- **Historique des Campagnes :** Un historique de réponses positives aux campagnes précédentes augmente la probabilité de réponses futures, soulignant l'importance de l'expérience client antérieure.
- **Dépenses en Viandes et Vins :** Des dépenses élevées dans ces catégories sont souvent le signe d'une réponse positive, suggérant un lien avec un niveau de vie supérieur.
- **Revenu Élevé :** Les individus dans les tranches de revenus supérieures sont plus susceptibles de répondre aux campagnes, ce qui indique le succès potentiel des offres premium pour ce segment de marché.
- **Récence des Achats :** Une récence d'achat réduite est corrélée avec une réponse positive, indiquant une opportunité de ciblage après un achat récent.
- **Situation Familiale :** Être célibataire ou ne pas avoir d'adolescents à la maison prédit une réponse positive, ce qui pointe vers des groupes démographiques spécifiques à cibler.

Inversement, des dépenses plus faibles, un revenu inférieur et être marié ou en couple sont liés à une moindre probabilité de réponse. De même, les clients avec des adolescents ou ceux avec une ancienneté de 5 ans montrent moins d'intérêt pour les campagnes.

Ces insights indiquent que les campagnes ciblant les clients avec des interactions positives précédentes et ceux avec un niveau de vie plus élevé pourraient être plus fructueuses. En revanche, il peut être bénéfique de réévaluer ou d'ajuster les stratégies pour les segments moins engagés. Bien que ces corrélations fournissent des indications précieuses, elles ne doivent pas être interprétées comme des causalités sans analyses supplémentaires telles que des tests A/B.

Cette compréhension approfondie des facteurs qui influencent les réponses des clients peut enrichir significativement notre stratégie marketing, en soulignant les profils de clients

les plus susceptibles de répondre favorablement aux campagnes.

6.3 Division du Jeu de Données en Ensembles d'Entraînement et de Test

La division du jeu de données en ensembles d'entraînement et de test est une procédure standard dans le développement de modèles de machine learning. Elle nous permet de former le modèle avec un ensemble de données (l'entraînement) et de le tester sur un ensemble distinct pour évaluer sa performance.

Équilibrage et Stratification

Nous veillons à ce que l'ensemble de test reflète la distribution de la variable cible 'Réponse'. La technique de stratification est employée pour maintenir une proportion similaire de la réponse dans les ensembles d'entraînement et de test. Cela est crucial lorsque la variable cible est inégale dans sa distribution.

Répartition des Données

En pratique, nous avons alloué 70% de nos données au jeu d'entraînement et les 30% restants au jeu de test. Cette répartition est choisie pour fournir un équilibre entre un entraînement suffisamment robuste et une évaluation précise des capacités prédictives du modèle.

Cette méthode garantit que notre modèle est évalué de manière équitable, nous donnant confiance dans sa capacité à généraliser à de nouvelles données non vues.

6.4 Régression Logistique

La régression logistique est une méthode d'analyse prédictive couramment utilisée en statistique pour modéliser la probabilité d'un événement binaire. Elle est particulièrement utile dans les situations où la variable cible est de nature catégorielle et peut prendre deux valeurs distinctes, souvent codées comme 0 et 1.

6.4.1 Fondements Mathématiques

La régression logistique repose sur l'équation d'un modèle linéaire pour prédire une transformation non linéaire de la probabilité de la variable cible. L'équation générale d'un modèle de régression logistique peut être exprimée comme suit :

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{6.1}$$

où p représente la probabilité que la variable cible prenne la valeur 1, β_0 est l'ordonnée à l'origine (le terme d'interception), $\beta_1, \ldots, \beta_n$ sont les coefficients des variables prédictives $x_1, \ldots, x_n$, et $\frac{p}{1-p}$ est le rapport de cote (odds ratio) de la probabilité de l'événement.

La fonction logistique, également connue sous le nom de fonction sigmoïde, est utilisée pour convertir le modèle linéaire en une probabilité :

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \tag{6.2}$$

Cette équation produit une courbe en forme de S qui limite la sortie entre 0 et 1, ce qui est interprété comme une probabilité.

6.4.2 Estimation des Coefficients

Les coefficients du modèle $(\beta_0, \beta_1, \ldots, \beta_n)$ sont estimés à l'aide de la méthode du maximum de vraisemblance. Cette méthode vise à maximiser la fonction de vraisemblance, qui mesure la probabilité d'observer les données d'échantillon données les paramètres estimés.

6.4.3 Interprétation des Coefficients

Dans un modèle de régression logistique, l'exponentielle d'un coefficient (e^{β_i}) donne le rapport de cote pour une augmentation d'une unité de la variable correspondante, en maintenant constantes toutes les autres variables. Si $e^{\beta_i} > 1$, la probabilité de l'événement augmente avec la variable x_i ; si $e^{\beta_i} < 1$, elle diminue.

6.4.4 Évaluation du Modèle

L'évaluation de la performance du modèle de régression logistique est généralement effectuée à l'aide de la matrice de confusion et des métriques dérivées telles que la précision,

le rappel, le score F1 et l'aire sous la courbe ROC (Receiver Operating Characteristic).

6.4.5 Avantages et Limitations

La régression logistique est appréciée pour sa simplicité et son efficacité, surtout lorsque la relation entre la variable cible et les variables prédictives est effectivement logistique. Cependant, elle a des limites, notamment son incapacité à capturer les relations complexes et non linéaires sans transformation des variables

6.5 Présentation des Résultats

Nous avons testé plusieurs modèles prédictifs pour identifier le plus performant en termes de prédiction des réponses aux campagnes marketing. Notre meilleure performance a été obtenue avec un modèle de régression logistique, qui a surpassé les modèles challengers tels que l'arbre de décision, la forêt aléatoire et le gradient boosting.

6.5.1 Performance du Modèle de Régression Logistique

Le modèle de régression logistique, ajusté pour contrer le déséquilibre des classes, a démontré une capacité remarquable à détecter les réponses positives (classe 1). Voici les métriques de performance obtenues sur l'ensemble de test :

```
Accuracy: 0.8167
Precision (Classe 1): 0.44
Recall (Classe 1): 0.78
F1-Score (Classe 1): 0.56
```

Ces résultats indiquent que le modèle est particulièrement efficace pour identifier les clients susceptibles de répondre favorablement aux campagnes, avec un rappel (recall) de 78% pour la classe 1. Bien que la précision pour cette classe soit plus modeste, notre objectif principal était d'optimiser le rappel afin de capturer le maximum de réponses positives.

6.5.2 Matrice de Confusion et Courbe ROC

La matrice de confusion et la courbe ROC sont des outils d'évaluation essentiels pour comprendre la performance du modèle au-delà de la seule exactitude (accuracy). La matrice de confusion ci-dessous montre le nombre de prédictions correctes et incorrectes pour chaque classe :

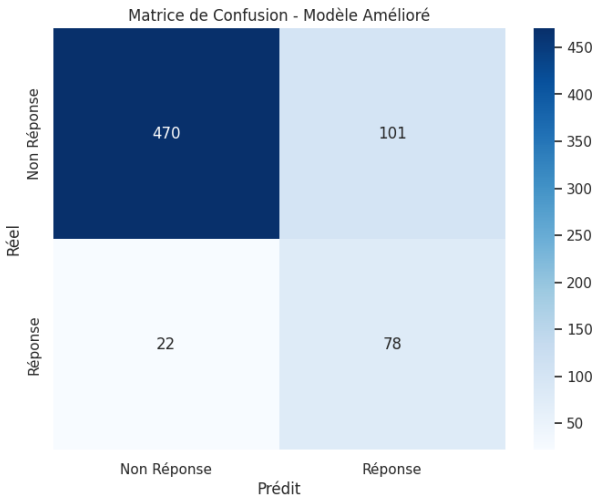

FIGURE 6.3 – Matrice de Confusion du Modèle de Régression Logistique

La courbe ROC, avec une aire sous la courbe (AUC) de 0.90, illustre la capacité du modèle à différencier entre les classes de réponse.

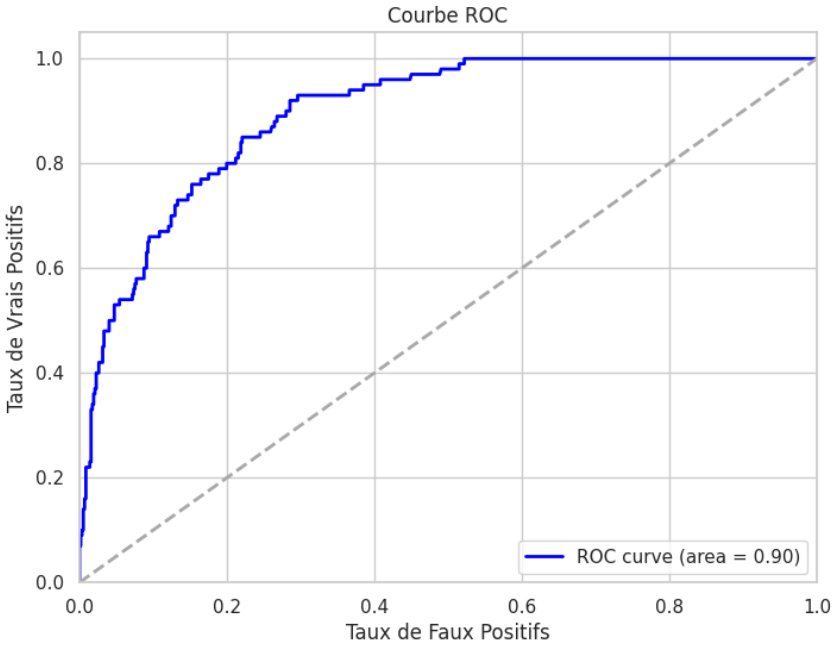

FIGURE 6.4 – Courbe ROC du Modèle de Régression Logistique

6.5.3 Conclusion

La régression logistique s'est avérée être un outil puissant pour notre analyse, permettant de détecter efficacement les clients les plus susceptibles de répondre aux campagnes marketing. La capacité du modèle à identifier correctement un grand nombre de réponses positives est un atout majeur, contribuant à l'élaboration de stratégies de marketing ciblées et à l'optimisation du retour sur investissement.

6.6 Analyse des Coefficients de la Régression Logistique

Cette section décrit les principaux facteurs influençant la probabilité de réponse positive aux campagnes marketing, comme indiqué par notre modèle de régression logistique.

6.6.1 Facteurs Positifs

Les caractéristiques ayant un impact positif indiquent une plus grande chance de réponse favorable. Les antécédents de réponses aux campagnes et les dépenses élevées en produits spécifiques sont particulièrement prédictifs.

6.6.2 Facteurs Négatifs

Les caractéristiques associées à une probabilité réduite de réponse incluent une absence d'activité d'achat dans certaines catégories et des traits démographiques spécifiques.

6.6.3 Tableau des Coefficients

Voici un tableau récapitulatif des dix principaux coefficients positifs et négatifs du modèle :

Variable	Coefficient
Campagne3_Acceptee	2.4236
Campagne1_Acceptee	2.3730
Campagne5_Acceptee	2.2501
Campagne2_Acceptee	2.2427
Niveau_Education_PhD	1.3696
Achats_Web_8+	1.3604
Achats_Magasin_2-4	1.2039
Achats_Catalogue_8+	1.1905
Achats_Web_5-7	1.0614
Anciennete_7	1.6257
Achats_Web_0	-2.9941
Niveau_Education_Basic	-2.2175
Achats_Catalogue_0	-2.1665
Plainte	-1.3319
Statut_Marital_Marié/Ensemble	-1.0039
VisitesWeb_Mois_0	-1.2012
Achats_Magasin_0	-1.1988
Anciennete_5	-1.6155
Categorie_Recence_Achat_Moyennement Ancien	-1.4010
Achats_Promos_0	-1.2998

TABLE 6.1 – Top 10 des coefficients positifs et négatifs - Régression Logistique

Les coefficients obtenus fournissent des indications sur les éléments à prendre en compte pour améliorer l'efficacité des campagnes marketing.

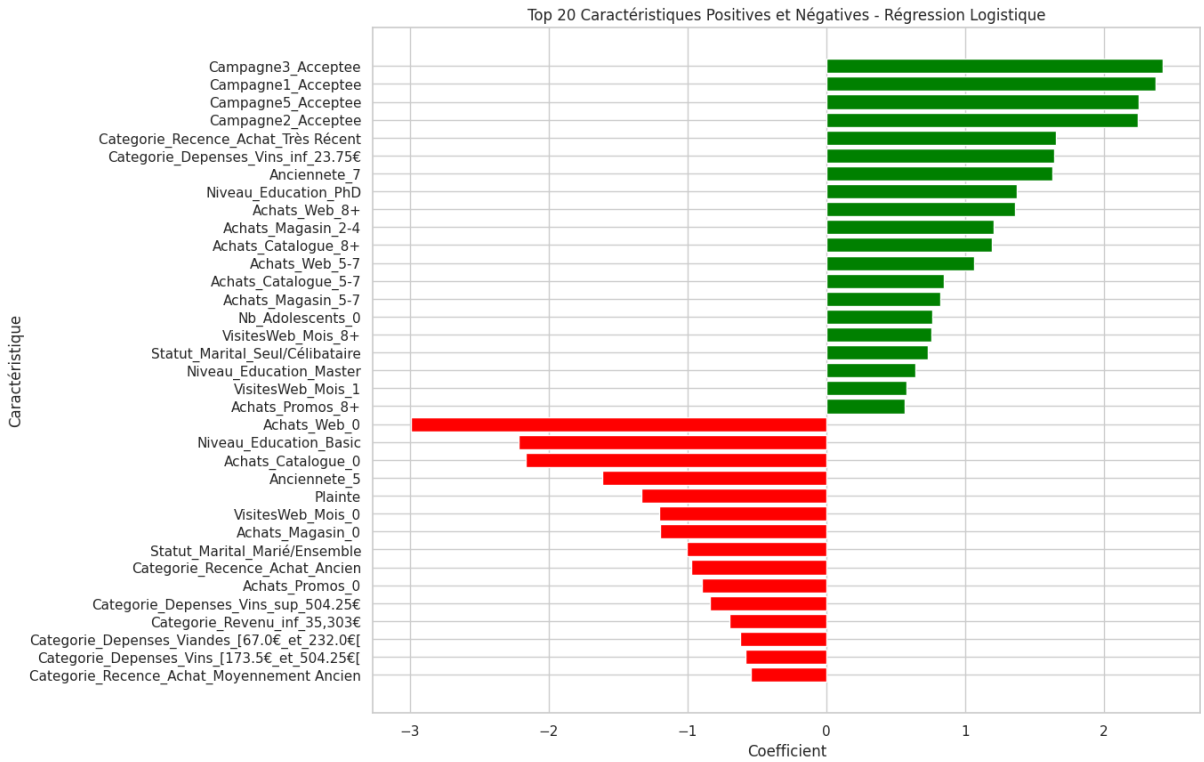

FIGURE 6.5 – Top 10 des caractéristiques positives et négatives - Régression Logistique

6.7 Interprétation des Coefficients de la Régression Logistique

L'examen des coefficients issus de notre modèle de régression logistique offre des éclairages sur les éléments qui ont un impact sur les réponses aux campagnes marketing. Des coefficients positifs élevés signalent une forte probabilité de réponse positive, tandis que des coefficients négatifs élevés indiquent le contraire.

6.7.1 Facteurs Positifs

Les variables comme l'acceptation des campagnes antérieures (Campagne3_Acceptee, Campagne1_Acceptee, etc.) affichent de forts coefficients positifs, suggérant une plus grande propension des clients à répondre de nouveau positivement. D'autres variables comme la récence d'achat et les dépenses élevées en vins sont également indicatives d'une réponse favorable, mettant en lumière l'importance de l'engagement récent et des préfé-

rences de consommation.

6.7.2 Facteurs Négatifs

En revanche, des caractéristiques comme l'absence d'achats sur le Web (Achats_Web_0) ou un niveau d'éducation de base (Niveau_Education_Basic) se révèlent négatives, ce qui pourrait refléter une moindre propension à interagir avec les campagnes en ligne ou une sensibilité différente aux offres marketing.

6.7.3 Implications pour les Stratégies Marketing

Ces données suggèrent que les campagnes devraient être personnalisées en fonction des antécédents d'interaction et des comportements d'achat des clients. L'accentuation des efforts sur les clients avec des interactions précédentes positives et une préférence pour des produits de haute valeur peut améliorer les résultats des campagnes.

En conclusion, notre modèle offre une vision approfondie des tendances de réponse des clients, permettant d'affiner les stratégies marketing pour une efficacité accrue et une meilleure fidélisation de la clientèle.

6.8 Modèles Challengers

Bien que le modèle de régression logistique optimisé (Modèle 2) ait été choisi comme le meilleur pour notre analyse des données marketing, d'autres modèles ont été évalués pour leur performance. Les modèles challengers, bien que non présentés en détail dans ce rapport, sont disponibles pour consultation dans le notebook de projet. Ces modèles incluent :

6.8.1 Arbre de Décision

Ce modèle simple effectue des décisions séquentielles et était testé pour évaluer sa capacité à gérer notre ensemble de données sans les complexités des méthodes ensemblistes.

6.8.2 Random Forest

En tant qu'ensemble d'arbres de décision, le Random Forest a été testé pour sa robustesse et sa précision. C'est une méthode qui peut souvent surpasser les modèles individuels grâce à sa capacité à réduire le surajustement.

6.8.3 Gradient Boosting

Le Gradient Boosting, qui affine les arbres de décision itérativement, a été considéré pour sa performance dans la correction des erreurs des arbres antérieurs et son efficacité potentielle dans notre contexte.

6.8.4 Évaluation Comparative

Chaque modèle a été évalué sur des métriques telles que la précision et le rappel pour déterminer son efficacité par rapport à notre modèle de régression logistique. Ces analyses sont essentielles pour assurer que nous utilisons le meilleur outil disponible pour nos prédictions de campagne marketing.

En conclusion, bien que ces modèles challengers aient fourni des perspectives intéressantes, le modèle de régression logistique reste notre choix préféré pour la prédiction des réponses aux campagnes marketing, en raison de sa performance supérieure et de sa capacité à équilibrer la détection des réponses positives.

Chapitre 7

Conclusion

Le projet d'analyse des campagnes marketing entrepris a porté ses fruits, permettant d'identifier les facteurs déterminants les réponses favorables des clients. Grâce au modèle de régression logistique optimisé pour tenir compte du déséquilibre des classes, nous avons pu extraire des informations pertinentes sur les éléments influençant positivement l'engagement des clients.

Les résultats obtenus soulignent l'importance de l'historique de réponses aux campagnes, des dépenses récentes, des investissements dans certaines catégories de produits comme les vins, et du niveau d'éducation. Ces facteurs semblent jouer un rôle prépondérant dans la prévision de l'engagement des clients dans les activités marketing.

Notre étude s'est démarquée par l'absence de sélection a priori des variables, laissant le modèle évaluer leur importance respective. Cette stratégie pourrait être affinée lors d'analyses futures pour améliorer la précision de nos prédictions.

La corrélation notable entre les réponses aux campagnes précédentes et les réponses actuelles mérite une attention particulière pour comprendre la dynamique de fidélisation et l'influence des pratiques marketing sur la rétention de la clientèle.

En définitive, bien que le modèle actuel constitue une solide fondation pour comprendre l'efficacité des campagnes marketing, il reste un potentiel d'amélioration. Des investigations plus poussées sur la sélection des variables et une analyse plus fine des effets des

campagnes antérieures pourraient contribuer à affiner nos stratégies marketing et à les rendre plus performantes et plus ciblées.