Moulin Vincent / Mamouni Ayoub

Mise en Place d’un processus de datamining

BUT 3A VCOD

Groupe 5 / Cas 2
06/12/2024

Table des matières

1. Nettoyage

1) Vérification des données

Variables
Id
Cus_subTyp
Avg_siz_Hld
Avg_age
Cust_Typ
Avg_Incom
Car_policie
Third_party_insur
Moto_policie
Life_insur
Accident_insur
Mobil_home_policie

Le jeu de données utilisé est un tableau Excel contenant des informations et diverses caractéristiques des familles aux États-Unis, basé sur les différents types d'assurances détenus par ces familles, jeu de données possédant des variables qui vont du N° de clients à la possession d'assurance mobil-Home.

Dans un premier temps nous devons vérifier si le jeu de données ne comporte pas des erreurs ou des valeurs manquantes qui pourrait biaiser les analyses plus tard.

2) Problème de syntaxe

Parmi les problèmes de syntaxe la variable Avg_Incom possède des valeurs qui contiennent des espaces au début, cela n'aura pas d'impact avec l'analyse sur R mais Excel est sensible à ce genre de problème lorsqu'on manipule ces données (la suppression des espaces permette aussi d'avoir une visibilité).

3) Valeur manquante

Après vérification des données, une grande partie du jeu de données ne comporte pas de valeur manquante à l'exception de la variable Avg_Incom ou Revenu moyen qui possède plusieurs valeurs manquantes, le problème étant que ces valeurs manquantes sont dispersées sur le jeu de données, c'est-à-dire qu'il n'est pas possible de les croisés simplement avec une autre variable et de données une seule valeur qui ne correspond pas forcément à la valeur réelle.

La variable Avg_age sera la variable va nous permettre de remplir les valeurs manquantes mais dans ce contexte on ne peut pas remplir le revenu moyen qui est le plus représenté.

Avg_age	Cust_Typ	Avg_Incom
3 40-50 years	2 Driven Growers	#N/A
3 40-50 years	1 Successful hedonists	#N/A
3 40-50 years	3 Average Family	#N/A
3 40-50 years	1 Successful hedonists	#N/A
3 40-50 years	2 Driven Growers	#N/A
3 40-50 years	6 Cruising Seniors	#N/A
3 40-50 years	2 Driven Growers	#N/A
3 40-50 years	1 Successful hedonists	#N/A

Pour compléter les valeurs il nous a fallu faire un pourcentage du Revenu moyen en fonction de la Moyenne d'âge afin de remplir les valeurs manquantes en fonction de ce pourcentage.

Avg_age	Cust_Typ	Avg_Incom
3 40-50 years	2 Driven Growers	45-75.000
3 40-50 years	1 Successful hedonists	45-75.000
3 40-50 years	3 Average Family	45-75.000
3 40-50 years	1 Successful hedonists	45-75.000
3 40-50 years	2 Driven Growers	75-122.000
3 40-50 years	6 Cruising Seniors	30-45.000
3 40-50 years	2 Driven Growers	30-45.000
3 40-50 years	1 Successful hedonists	30-45.000

Dans ce contexte, la moyenne d'âge de 40 à 50 ans possède 49% de personne possédant un revenu moyen de 45 à 75 000 dollars, en fonction de tous les pourcentages affichés on remplit les valeurs manquantes.

	40-50 years
< 30.000	0,75757576
45-75.000	49,7245179
75-122.000	6,4738292
>123.000	1,72176309
30-45.000	41,322314

Ce système ci-dessus est utilisé pour les moyennes d'âge 30-40 ans et 50-60 ans où certaines valeurs sont manquantes, les revenus moyens sont appliqués en fonction du pourcentage indiqués par les moyennes d'âge.

4) Modalités rares

Suites à la correction des différentes données qui peuvent potentiellement erroné la partie statistique, il a fallu dans notre analyse trouver les modalités rares qui sont les catégories qui apparaissent très rarement (généralement moins de 5%) dans les données et qui peuvent être exclues de certaines analyses pour éviter de biaiser les résultats, dans notre cas nous allons devoir les fusionner si possible à une autre modalité et que la modalité soit proche de celle qu'on veut fusionner.

La recherche des modalités rare se fera exclusivement sur les variables quantitatives partant de Avg_siz_Hld à Avg_Incom, la variable Cus_subTyp représente les détails du regroupement des clients en classe qui possèdent trop de modalité et sera retiré de l'analyse.

Avg_siz_Hld
Avg_age
Cust_Typ
Avg_Incom

a. Avg_siz_Hld

Cette variable correspond à la taille moyenne d'une famille au États-Unis, les modalités rares que nous avons décidé de regrouper les familles de 4 et 5 individus car généralement les familles de 5 individus possèdent 3 enfants ce qui est le même cas pour les familles de 4 individus avec 1 enfants en moins, afin de les regrouper nous avons décidé d'utiliser la modalité 4&5.

Taille famille		
1	284	4,87804878
2	2131	36,60254208
3	2646	45,44829955
4	693	11,90312607
5	68	1,167983511
4&5	761	13,07110958

b. Avg_age

Cette variable correspond à l'âge moyen au États-Unis, les modalités rares que nous avons décidé de regrouper sont les moyennes d'âge de 60-70 ans et 70-80 ans individus car le graphique AFCM nous indique une proximité et une progression équivalente, afin de les regrouper nous avons décidé d'utiliser la tranche d'âge de 60 à 80 ans.

1 20-30 years	74	1,271040879
2 30-40 years	1452	24,9398832
3 40-50 years	3000	51,5286843
4 50-60 years	1073	18,43009275
5 60-70 years	193	3,315012023
6 70-80 years	30	0,515286843
5 60-80 years	223	3,830298866

c. Cust_Typ

La variable Cust_Typ correspond à une description synthétique du ménage, les modalités rares que nous avons décidé de regrouper pour cette variable sont les Career Loners et Cruising Seniors car elles ont généralement la même taille moyenne de famille et le graphique AFCM nous indique une proximité, afin de les regrouper nous avons décidé de renommer les modalités à 4 Career Loners and Cruising Seniors.

1 Successful hedonists	552	9,481277911
10 Farmers	276	4,740638956
2 Driven Growers	502	8,622466506
3 Average Family	886	15,2181381
4 Career Loners	52	0,893163861
5 Living well	569	9,773273789
6 Cruising Seniors	205	3,521126761
7 Retired and Religeous	550	9,446925455
8 Family with grown ups	1563	26,84644452
9 Conservative families	667	11,45654414
4 Career Loners and Cruising Seniors	257	4,414290622

d. Avg_Incom

La variable qui représente le revenu moyen, ils nous à pas été possible de regrouper les deux modalités ensemble car elles sont beaucoup à l'opposer l'une de

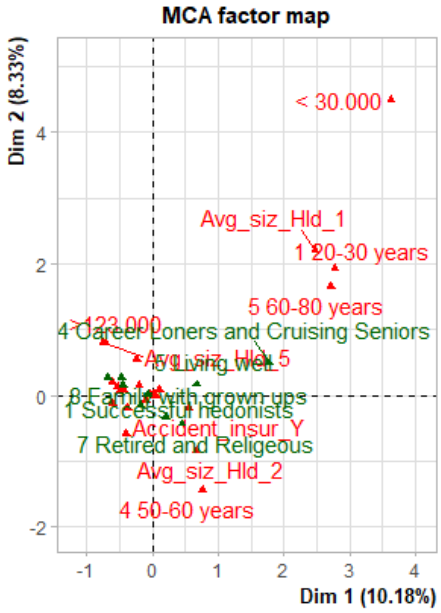
l'autre et qui ne peuvent pas être supprimer du fait que cela paraît logique dans un cas réel où une personne reçoit des revenus moyens très bas ou très élevé.

< 30.000	49	0,84163518
30-45.000	2594	44,5551357
45-75.000	2598	44,6238406
75-122.000	489	8,39917554
>123.000	92	1,58021299

2. Analyse des correspondance multiples

1) Graphique 1

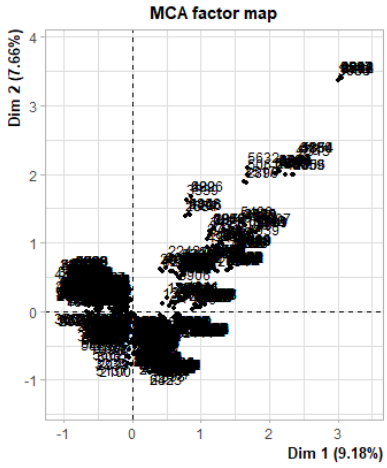
L'AFCM réalisée sur les données d'assurance nous as permis d'identifier des relations entre les modalités du jeu de données. Le premier graphique, qui représente les modalités, met en évidence des groupes distincts. Par exemple, les modalités associées aux seniors de 60 à 80 ans et au 20 à 30, avec de faibles revenus (« <30,000 ») et seul (« Avg_siz_Hld_1 »), se positionnent loin des autres modalités, suggérant un profil spécifique. Ce groupe peut correspondre à des retraités vivant seuls et à des jeunes salarié seuls. Ces groupes peuvent avoir des besoins précis en termes de produits d'assurance. D'un autre côté, les individus avec des revenus très élevés (">123,000") se distinguent également fortement et s'oppose à ceux à revenue faible. Ils sont situés loin du centre de l'espace factoriel, témoignant de leur singularité. Ce segment pourrait représenter une cible majeure pour des offres d'assurance spécifique comme les premiers groupes. Un autre groupe qui se



démarque est celui des individus âgés de 50 à 60 ans en couple (« Avg_siz_Hld_2 »), représentant une clientèle avec un potentiel de consommation important et des besoins spécifiques en matière d'assurance.

Le graphique met également en évidence des groupes divisés par type de client, comme les retraités actifs et les salariés célibataires. Toutefois, des groupes comme les « Successful Hédoniste » se positionnent différemment, indiquant des besoins distincts par rapport à ces autres segments. Cela suggère qu'ils partagent des caractéristiques communes, mais aussi des attentes spécifiques en matière d'assurance. D'un point de vue marketing, ces observations ouvrent des pistes stratégiques intéressantes. Par exemple, les seniors à faibles revenus pourraient bénéficier d'offres d'assurance simplifiées et abordables, tandis que les ménages à hauts revenus pourraient être approchés avec des campagnes sur mesure, mettant en avant l'exclusivité et la valeur ajoutée de l'offre. Enfin, les individus de 50 à 60 ans en couple pourraient être ciblés par des offres de produits plus complets, offrant une sécurité à long terme et répondant à des besoins de prévoyance.

2) Graphique 2



Le second graphique, représentant les individus. Ces individus nous montrent une forte concentration autour du centre des axes factoriels. Cela suggère une majorité d'individus partagent des caractéristiques en commun. Cependant, on peut observer certains points éloignés reflètent des profils spécifiques (en lien avec les modalités identifiées sur le premier graphique). Cette différence d'individus est essentielle dans une optique marketing, car elle permet de se concentrer sur des niches de client précises tout en couvrant les besoins de tous.

Cette AFCM nous a permis de dévoiler des parties de population définis, chacun présentant des besoins spécifiques, qui peuvent être exploité à des fin de créer des stratégies ciblées et adaptées.

3. Cluster et K-means

K-means est une méthode qui sert à regrouper des individus en groupes qui se ressemblent le plus possible, en se basant sur leurs caractéristiques. Donc après l'AFCM on a pu appliquer K-means sur les données résultant de celle-ci afin de scinder les clients en plusieurs groupe. Afin de déterminer le nombre de groupe optimal nous avons optimiser plusieurs méthodes.

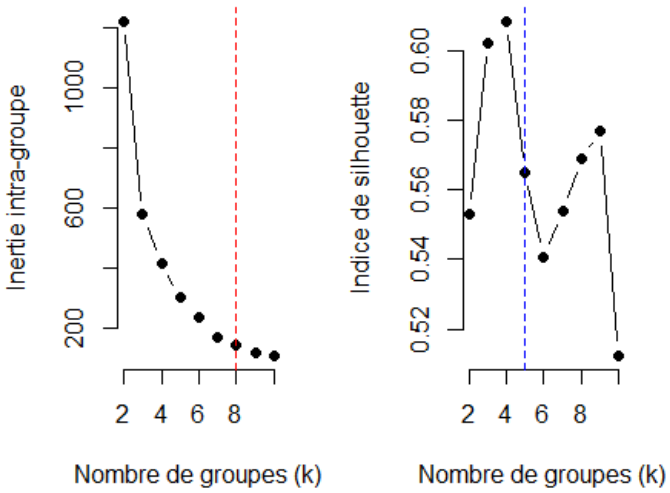
1) R^2

Afin de déterminer le nombre idéal de clusters, nous avons utilisé R^2, qui permet de calculer le coefficient de détermination. R^2 mesure la proportion de la variance totale expliquée par les clusters. Il varie entre 0 et 1, où une valeur est plus proche de 1 indique que les clusters représentent bien les données. Nous avons ainsi testé K-means avec plusieurs valeurs de K jusqu'à identifier celle offrant le meilleur coefficient. Cependant, pour valider nos résultats, nous avons également adopté une autre approche afin de nous assurer de ceci.

2) Méthode du coude et silhouette

Donc pour s'assurer du nombre de cluster à utiliser nous avons pu utiliser la méthode du coude et l'indice de silhouette, qui sont des techniques couramment employées en datamining afin d'évaluer la qualité des clusters.

La méthode du coude consiste à mesurer l'inertie intra-groupe, c'est-à-dire la somme des distances entre les points de chaque cluster et leur centre. Plus le nombre de clusters augmente, plus l'inertie diminue, car les groupes deviennent plus petits et plus compacts. Cependant, au-delà d'un certain point, la diminution devient marginale et forme un "coude" sur le graphique. Ce point indique le nombre optimal de clusters. Dans notre cas, nous avons identifié ce coude pour 5 clusters à partir du graphique généré par notre code R.



L'indice de silhouette, quant à lui, mesure la cohérence interne des clusters en évaluant à quel point les points sont proches des autres membres de leur propre cluster par rapport à ceux des clusters voisins. Un indice de silhouette élevé signifie que les clusters sont bien définis et bien séparés. En appliquant cette méthode avec R, nous avons calculé les indices pour différents nombres de clusters et constaté que la meilleure valeur était également obtenue pour 5 clusters.

Grâce à notre analyse R, combinant ces deux approches, nous avons validé que 5 clusters offraient la meilleure segmentation des individus.

3) Cluster

Après avoir déterminé que cinq clusters représentaient le choix optimal grâce aux différentes méthodes, nous avons approfondi notre analyse pour interpréter la

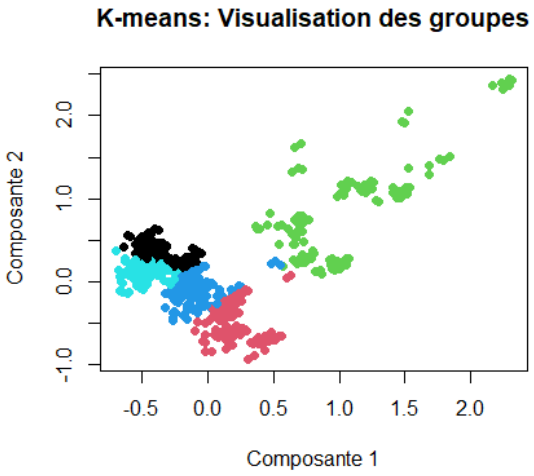
composition et la signification de ces groupes. Grâce à notre code R, nous avons pu observer la répartition des individus dans chaque cluster :

- Cluster 1 : 875 individus
- Cluster 2 : 1584 individus
- Cluster 3 : 476 individus
- Cluster 4 : 1355 individus
- Cluster 5 : 1532 individus

La distribution des individus révèle que certains clusters regroupent une grande proportion des individus, notamment les clusters 2, 4 et 5, tandis que d'autres, comme le cluster 3, correspondent à un segment plus restreint. Cette répartition met en évidence des différences marquées entre les groupes, avec certains segments plus généraux et d'autres plus spécifiques. Ces résultats reflètent bien les caractéristiques distinctes, telles qu'identifiées lors de l'AFCM.

4) Interprétation

Avec K-means nous avons pu dresser un graphique des points ainsi que le cluster d'appartenance afin de les analyser, nous avons pu définir ces différentes familles de clients.



a) Budget limité

Ce cluster est principalement composé de jeunes adultes et d'individus ayant un revenu relativement faible comme les 60 à 80ans. Ce groupe est constitué de

personnes de moins de 30 ans et plus de 60ans, avec des revenus inférieurs à 30 000, à la recherche de solutions abordables. Leur faible pouvoir d'achat les pousse à privilégier des offres à bas coût, telles que des assurances d'entrée de gamme

b) Famille Professionnel

Ce cluster représente des familles avec des enfants à charge et des moyens aisée. Ce cluster est très proche des autres montrent un même besoin pour les différentes familles comparé au retraité et au faible revenue.

c) Famille aisée

Ce cluster représente des personnes de 30 à 40 avec des moyens aisés avec un enfant à charge, avec des besoins d'offre plus adapté à leurs moyen et leur budget tel que des assurances haut de gamme.

d) Familles stables

Ce cluster regroupe des familles avec enfants à charges et des moyens suffisant. Les foyers de ces groupes sont souvent bien établis, avec des besoins plus spécifiques.

e) Retraité

Le cluster rouge regroupe des individus âgés de 50 à 60 ans sans enfants à charge. Ce groupe est constitué de retraités ou de personnes proches de la retraite. Ils sont probablement à la recherche de produits moins coûteux que des jeunes familles.

5) Conclusion

En conclusion, nous pouvons observer que quatre de ces clusters présentent des besoins similaires en matière d'offres d'assurance, tandis que certains se distinguent par la demande d'offres plus spécifiques. L'un des clusters se distingue particulièrement, puisqu'il regroupe des individus de tous âges avec un revenu faible, ce qui révèle un marché à fort potentiel pour des offres d'entrée de gamme adaptées à des clients à budget limité.

4. Classification ascendante hiérarchique

Nous avons décidé de segmenter nos données en 5 clusters, en utilisant d'abord l'algorithme de K-Means. Ce choix a été guidé par l'analyse des données effectué au part avant. Visuellement, ce graphique montre que 5 clusters représentent un bon compromis entre une segmentation précise et une cohérence des groupes formés.

Les Résultats obtenus avec K-Means montre que les 5 clusters ont des tailles relativement équilibrées. Ces tailles équilibrées facilitent l'exploitation de ces données, car chaque groupe est assez homogène et suffisamment grand pour justifier des actions ciblées.

1) Analyse avec la CAH

Nous avons complété cette segmentation avec une Classification Ascendante Hiérarchique. Comme pour K-Means, nous avons retenu 5 clusters, mais cette méthode a donné des tailles très différentes :

- Cluster 1 : 1 297 individus,

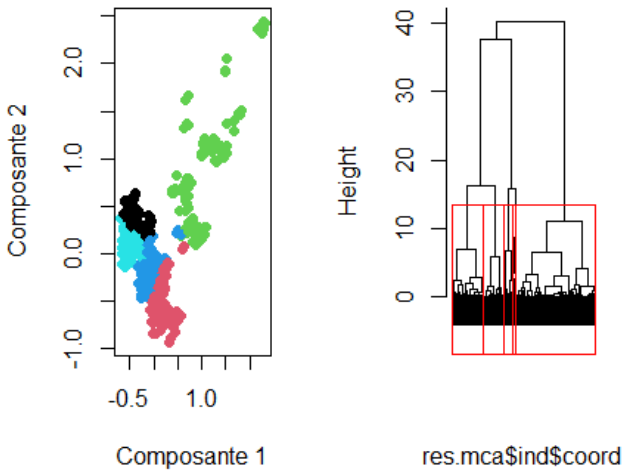
- Cluster 2 : 3 212 individus,
- Cluster 3 : 832 individus,
- Cluster 4 : 332 individus,
- Cluster 5 : 149 individus.

Ces résultats montrent que la CAH met en évidence des groupes aux caractéristiques plus variées. Par exemple, le cluster 5 (149 individus) pourrait représenter une niche très spécifique. Cette disparité dans les tailles est typique de la CAH, qui privilégie les regroupements naturels, même si cela mène à des clusters déséquilibrés.

2) Interprétation

Nous constatons que les 5 clusters obtenus semblent pertinents dans les deux approches, mais pour des raisons différentes.

Le graphique de gauche montre la répartition des individus sur les deux premières dimensions factorielles issues de l'analyse. Les clusters obtenus avec la CAH sont représentés par des couleurs distinctes. Nous pouvons observer que certains clusters, comme le cluster vert et le cluster rouge, occupent des zones bien définies et sont nettement séparés des autres groupes, ce qui indique une homogénéité forte au sein de ces clusters. Cependant, d'autres clusters, comme le cluster noir, semblent se chevaucher partiellement avec d'autres, suggérant que ces groupes partagent certaines caractéristiques similaires avec d'autres segments.



Le dendrogramme montre le processus de regroupement hiérarchique des individus en fonction de leur similarité. Ce graphique illustre que les groupes se forment progressivement, des sous-groupes très homogènes en bas de l'arbre vers

des regroupements plus larges et plus diversifiés en haut. Les cinq découpes effectuées (encadrées en rouge) révèlent des tailles de clusters très variées

3) Conclusion

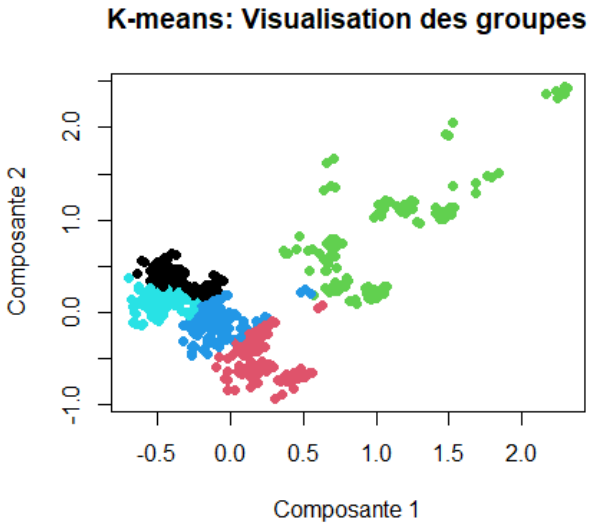
Le fait que la projection factorielle des clusters obtenus avec la CAH donne un résultat similaire à K-Means renforce notre confiance dans la qualité de la segmentation. Cela montre que, même si les tailles des clusters diffèrent, leur structure et leur différenciation sont cohérentes.

5. Analyse AFCM

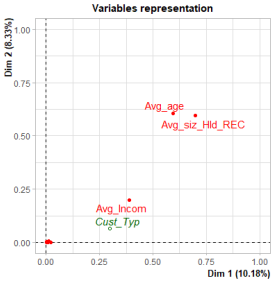
L'analyse de l'AFCM en relation avec les clusters permet de mieux comprendre les caractéristiques des groupes et de visualiser leurs relations sur le plan factoriel. Cette analyse nous apporte des clés pour interpréter le comportement et les préférences des segments identifiés.

1) Visualisation des groupes

Le graphique des clusters obtenus par K-means superposé à l'espace factoriel montre que les groupes se positionnent de manière distincte sur les axes principaux. Cela indique que les clusters sont bien différenciés en termes de caractéristiques clés. Par exemple, les clusters comme le groupe rouge ou vert se distinguent nettement sur les axes principaux, reflétant des différences marquées dans les attributs tels que les tranches d'âge, les revenus, ou encore la composition du foyer.



Les variables actives représentées dans le deuxième graphique renforcent cette interprétation. On observe une corrélation forte entre certaines variables et les dimensions principales.



2) Correspondance AFCM et les Clusters

Les clusters identifiés par K-means se répartissent bien dans cet espace factoriel, confirmant la robustesse de la segmentation. Bien que les clusters varient en termes de taille, leurs positions sur le graphique factoriel sont cohérentes avec les analyses descriptives des segments.

3) Conclusion

En conclusion, l'AFCM et K-means se complètent efficacement pour fournir une segmentation pertinente et exploitable. Malgré les différences dans la taille des clusters, l'interprétation reste cohérente, avec des segments bien différenciés dans l'espace factoriel.

6. Régression Logistique

Nous avons réalisé une régression logistique pour prédire si un individu souscrirait à une police d'assurance automobile (« Car_policie »). L'objectif est de développer une segmentation prédictive qui nous permette de mieux cibler nos campagnes marketing et d'optimiser nos ressources en identifiant les profils les plus susceptibles de souscrire. Voici une analyse détaillée de notre démarche et des résultats obtenus.

1) Choix des Variable Pivots

Dans ce cas nous allons devoir choisir une modalité pivot, pour se faire nous devons vérifier si une modalité en particulier qui possède le même pourcentage c'est-à-dire qu'il n'y a autant de personne ayant une assurance de voiture que de personne n'ayant pas d'assurance de voiture dans ce contexte.

Nombre de Id	Car_policie		
Avg_age	0	1	Total général
1 20-30 years	1,41%	1,14%	1,27%
2 30-40 years	24,29%	25,56%	24,94%
3 40-50 years	51,53%	51,53%	51,53%
4 50-60 years	18,52%	18,34%	18,43%
5 60-70 years	3,51%	3,12%	3,32%
6 70-80 years	0,74%	0,30%	0,52%
Total général	100,00%	100,00%	100,00%

Dans ce cas ci-dessus on peut voir que toute la modalité possède le même nombre de pourcentage, on pourrait dire que toutes ces modalités sont des modalités pivot mais dans notre cas nous avons choisies la tranche d'âge entre numéro 3 entre 40 et 50 ans, pour les autres variables que nous avons dû analyser qui sont Avg_siz_Hld qui est le numéro 3 et Avg_Incom dont le Revenu moyen choisie est de 30-45 000.

Dans le cas où ils nous ont fallu choisir une modalité pivot qui prend en compte les variables de Car_policie à Accident_insur ces variables étant binaire nous avons croisés les variables avec les différents variables la variable qui sort le plus du lot est la Third_party_insur qui nous indique que si une famille possède un assurance Voiture elle aura plus de chance de souscrire à une assurance au tiers.

Nombre de Id	Car_policie		
Third_party_insur	0	1	Total général
0	66,71%	53,21%	59,81%
1	33,29%	46,79%	40,19%
Total général	100,00%	100,00%	100,00%

2) Echantillonnage

Pour garantir que notre modèle de régression logistique est robuste et évite les biais, nous avons suivi une démarche structurée

a) Équilibrage des données

Les données initiales montraient un déséquilibre entre les classes (par exemple, plus de personnes n'ayant pas souscrit à une police que de souscripteurs). Pour résoudre ce problème, nous avons équilibré les classes en échantillonnant de manière aléatoire le même nombre d'individus dans chaque classe. Cela permet au modèle de

mieux apprendre les caractéristiques des deux groupes, sans favoriser une classe dominante.

b) Division en ensembles d'apprentissage et de test

Nous avons divisé les données pour les équilibrées en deux parties

- Ensemble d'entraînement (80 % des données) : Utilisé pour ajuster le modèle.
- Ensemble de test (20 % des données) : Utilisé pour évaluer la performance du modèle sur des données jamais vues, simulant des scénarios réels.

Cette division garantit que nous pouvons mesurer la capacité du modèle à généraliser, un critère essentiel pour son application.

3) Matrice de confusion

	Classe réelle	Classe réelle
Classe prédite	360	374
Classe prédite	190	214

a) Métriques calculées

Les résultats obtenus de notre modèle de régression logistique montrent des performances modestes, mais cela peut être expliqué par la nature universelle de l'assurance automobile, qui est un produit largement adopté. Avec une accuracy de 51 %, le modèle a correctement prédit un peu plus de la moitié des cas. Cela peut sembler

proche d'une prédiction aléatoire, mais reflète également la complexité du comportement d'achat, où de nombreux facteurs individuels influencent la décision de souscrire.

En termes de précision, le modèle indique que, parmi les individus identifiés comme souscripteurs potentiels, 53 % le sont réellement. Bien que ce chiffre puisse limiter la fiabilité des campagnes marketing si l'on se basait uniquement sur ces prédictions, il reste cohérent avec un produit d'assurance.

Le rappel, à 36 %, montre que le modèle a des difficultés à identifier tous les souscripteurs réels. Cela est en partie lié à l'échantillonnage équilibré que nous avons utilisé pour entraîner le modèle, où nous avons réparti équitablement les classes de souscripteurs et de non-souscripteurs afin de garantir une analyse représentative. Cependant, ce rappel plus faible reflète que le modèle ne capture pas complètement les subtilités propres aux souscripteurs.

Enfin, le F1-Score, à 43 %, montre un équilibre modéré entre précision et rappel. Ce score, bien qu'améliorable.

b) Utilisabilité

En l'état, le modèle n'est pas suffisamment performant pour une application marketing directe. Bien qu'il parvienne à identifier certains souscripteurs, il passe à côté d'un grand nombre de client potentiels, ce qui limiterait l'efficacité des campagnes. De plus, il présente un risque de sur-ciblage, en identifiant à tort de nombreux individus comme intéressés, ce qui pourrait entraîner des coûts inutiles pour des actions marketing.

c) Optimisation

Pour améliorer l'efficacité de ce modèle, nous pourrions enrichir les données en ajoutant des variables qui reflètent mieux les comportements et les historiques d'achat

des clients potentiels. Cela permettrait de mieux comprendre leurs motivations. Tester d'autres modèle de prédictions pourrait aider à mieux cerner les clients.

4) Conclusion

En conclusion, en analysant cette régression, nous constatons qu'il n'existe pas de variable déterminante permettant de prédire avec certitude la souscription à une assurance automobile. Parmi les facteurs étudiés, seule l'assurance responsabilité civile semble jouer un rôle notable, probablement parce qu'elle est souvent associée à une souscription d'assurance auto. Cela s'explique par le fait que l'assurance auto est un produit de masse, accessible à une grande variété de profils. Par conséquent, aucun facteur ne semble influencer significativement la décision, rendant difficile une segmentation prédictive précise pour ce produit.