

RAPPORT DE PROJET

Analyse prédictive de l'emprunt à la
médiathèque en fonction des
caractéristiques des adhérents

Sommaire

Introduction.....	1
I. Choix du jeu de données.....	2
II. Description des données.....	3
1. Description des variables.....	3
2. Mise en qualité des données.....	4
3. Statistiques descriptives.....	6
4. Liaisons entre les variables.....	11
III. Détails du travail.....	14
IV. Modélisation des emprunts de la médiathèque.....	15
1. Etablissement du modèle.....	15
2. Qualité du modèle.....	15
3. Estimation du modèle.....	17
V. Evaluation de la performance du modèle.....	21
Conclusion.....	23

Introduction

Dans un monde de plus en plus digitalisé, où les bibliothèques et médiathèques jouent un rôle crucial dans la diffusion de la connaissance, il devient impératif d'optimiser l'expérience des adhérents en anticipant leurs besoins.

Notre projet novateur se concentre sur l'utilisation de l'apprentissage automatique et de l'analyse prédictive pour élaborer un système permettant de prédire avec précision si un adhérent est susceptible d'emprunter des ouvrages ou pas en fonction de ses caractéristiques personnelles.

En combinant les richesses de la technologie moderne avec la richesse du savoir traditionnel, notre initiative vise à transformer la médiathèque en un espace encore plus adapté et réactif aux attentes de ses usagers, tout en maximisant l'efficacité de son offre culturelle.

Pour mener à bien cette mission, elle s'articulera autour de trois axes majeurs : tout d'abord, une description approfondie des données recueillies sur les adhérents de la médiathèque; ensuite, une explication détaillée du processus de travail adopté pour élaborer le modèle prédictif; enfin, une évaluation critique de la performance du modèle, mettant en lumière ses réussites et ses limites.

I. Choix du jeu de données

Pour réaliser notre étude, nous avons sélectionné un jeu de données provenant du portail open data opendata.roubaix.fr. Ce jeu de données concerne les caractéristiques des adhérents à la médiathèque ***La Grand Plage*** au cours de l'année 2020. Localisée au cœur de Roubaix, dans la métropole lilloise, cette médiathèque est mise à la disposition du public par la Ville de Roubaix.

La médiathèque *La Grand-Plage* a pour missions principales de rendre accessibles à tous les moyens de formation, d'information, de culture et de divertissement, de promouvoir le vivre ensemble en encourageant les rencontres et la diversité, de jouer un rôle actif dans l'inclusion sociale, de faciliter l'accès, l'accompagnement et la formation aux outils et enjeux du numérique, ainsi que de collecter, préserver et faire connaître le patrimoine écrit et graphique de la ville de Roubaix.

L'ensemble de données que nous avons choisi englobe une variété d'informations détaillées sur les adhérents, offrant ainsi une base solide pour la création de modèles prédictifs. Notre objectif est d'exploiter ces données exhaustives afin de développer des modèles capables de déterminer avec précision si un adhérent de la médiathèque de Roubaix empruntera ou non des ouvrages.

Notre étude vise à apporter des insights significatifs pour optimiser l'expérience des adhérents tout en contribuant à une gestion plus efficiente des ressources culturelles de cet établissement.

II. Description des données

1. Description des variables

Au sein de notre base de données, nous avons recueilli un ensemble de 11 673 données relatives aux adhérents de 2020, détaillées autour de multiples variables, à la fois quantitatives et qualitatives. Dans le cadre de notre étude, notre attention se concentrera spécifiquement sur une variable cible *activite_emprunteur* qui distingue les emprunteurs des non emprunteurs, ainsi que sur plusieurs variables explicatives décrivant :

- l'activité de l'adhérent dans la médiathèque ;
- l'information relative à l'inscription ;
- la situation socio-démographique de l'adhérent ;
- la fréquence d'utilisation des services proposés par la médiathèque.

<code>activite</code>	
<code>activite_emprunteur_bus</code>	<code>inscription_attribut_action</code>
<code>activite_emprunteur_med</code>	<code>inscription_attribut_zèbre</code>
<code>activite_salle_etude</code>	<code>inscription_carte</code>
<code>activite_utilisateur_postes_informatiques</code>	<code>nombre d'années d'adhésion</code>
<code>activite_utilisateur_wifi</code>	<code>type_inscription</code>
<code>tranches d'âge (1)</code>	<code>nb_venues</code>
<code>tranches d'âge (2)</code>	<code>nb_venues_postes_informatiques</code>
<code>roubaisien ou non</code>	<code>nb_venues_prets</code>
<code>code IRIS de Roubaix</code>	<code>nb_venues_prets_bus</code>
<code>nom de l'IRIS à Roubaix</code>	<code>nb_venues_prets_mediatheque</code>
<code>commune de résidence</code>	<code>nb_venues_salle_etude</code>
<code>sexe</code>	<code>nb_venues_wifi</code>

Tableau 1 - Liste des variables

Il est important de souligner que l'absence d'un dictionnaire des variables a constitué un défi, certaines variables étant définies de manière vague. Afin de pallier cette lacune, nous avons minutieusement sélectionné les variables les plus pertinentes pour notre étude. Cette sélection a été réalisée après une consultation approfondie du site officiel de la médiathèque (<http://www.mediathequederoubaix.fr>), garantissant ainsi une meilleure compréhension contextuelle et la pertinence des variables retenues pour notre analyse.

2. Mise en qualité des données

Avant de débuter notre analyse, il est impératif d'effectuer une étape cruciale de nettoyage des données. Au cours de cette exploration, nous avons repéré des données manquantes, caractérisées par une information inconnue, ainsi que des données aberrantes, ou anormales par rapport au reste de l'ensemble de données. Ces éléments pourraient compromettre la validité de l'interprétation statistique.

Ce jeu de données présente des données manquantes pour plusieurs variables, et plus particulièrement nous avons deux variables (`inscription_attribut_action`, `inscription_attribut_zèbre`) pour lesquelles nous n'avons aucune information. Afin d'assurer l'intégrité de notre étude et d'éviter tout biais potentiel, nous avons décidé de les exclure de notre étude. Nous avons également identifié trois variables (`Nom de l'IRIS à Roubaix`, `Code IRIS de Roubaix'`, `geo_point_2d`) qui ne semblent pas pertinentes dans le contexte de notre prédiction. Nous avons décidé de ne pas les prendre en compte.

De plus, nous avons constaté la présence de deux variables (`tranche d'âge (1)`, `tranche d'âge (2)`) fournissant toutes deux des informations sur la tranche d'âge des adhérents. Bien que ces deux variables diffèrent par leurs intervalles, nous avons choisi celle offrant une précision plus adéquate. Malgré la présence de 128 données manquantes dans cette variable, nous avons décidé de la conserver en l'état, en précisant que l'âge demeure inconnu.

Simultanément, nous avons identifié la présence de données aberrantes liées à plusieurs variables, notamment le nombre d'années d'adhésion. En effet, la valeur 43 semble extrêmement atypique par rapport aux autres données de notre base. Afin de maintenir la cohérence de notre ensemble de données, nous avons procédé à la suppression de cette valeur

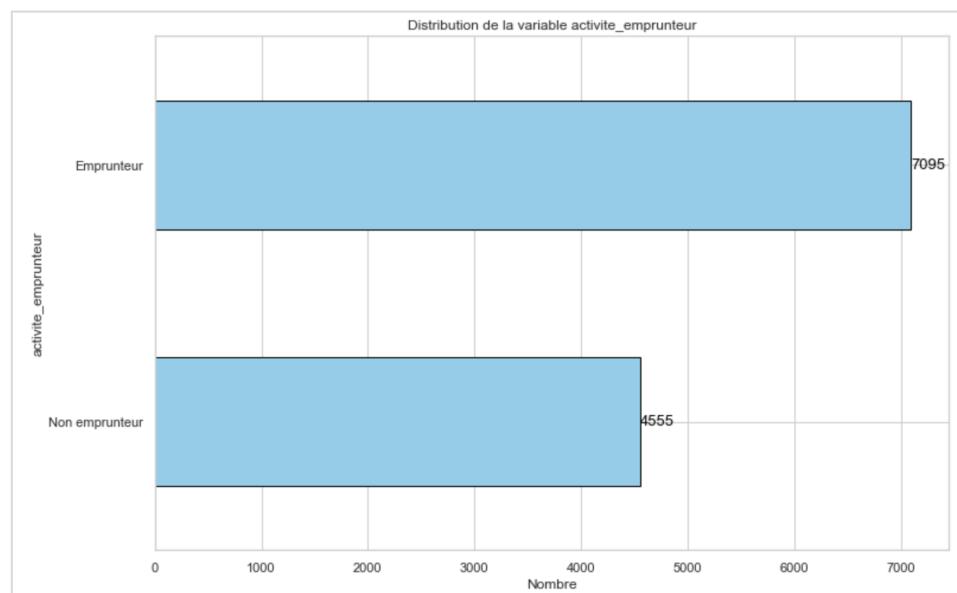
jugée anormale. Nous avons étendu cette démarche à l'ensemble des variables présentant des valeurs aberrantes (`nb_venues_postes_informatiques`, `nb_venues_prets_mediatheque`, `nb_venues_salle_etude`, `nb_venues_prets_bus`, `nb_venues_wifi`, `nb_venues`), en utilisant une méthode d'identification fondée sur l'analyse de la distribution des données à l'aide de boîtes à moustaches.

Finalement, suite à cette mise en qualité de nos données, nous avons une base contenant 11 650 observations décrites autour de 26 caractéristiques.

3. Statistiques descriptives

Nous allons à présent regarder les statistiques de nos données.

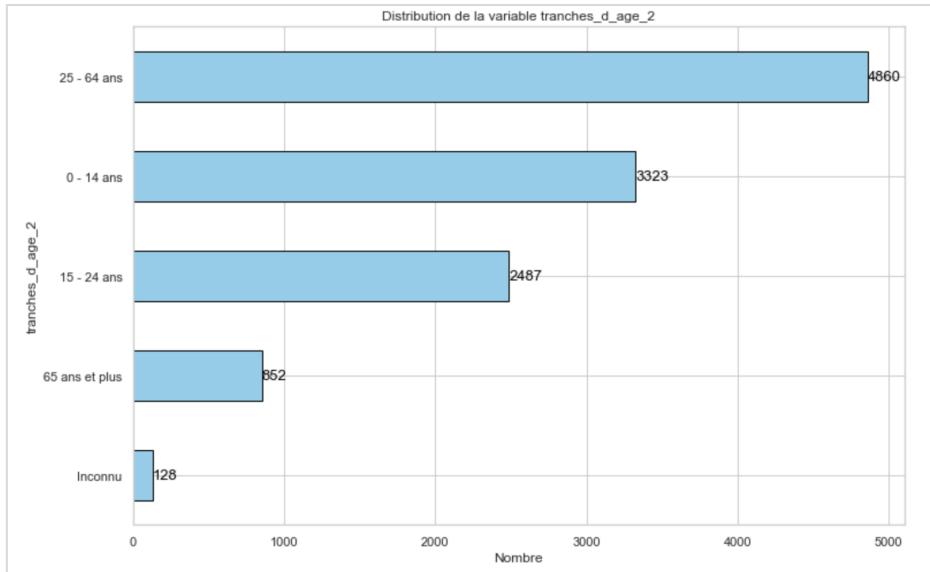
- Emprunt : en observant la variable cible, nous constatons que parmi l'ensemble des adhérents de la médiathèque, plus de la moitié sont emprunteurs, soit 60% et le reste n'ont pas emprunté.



Graphique 1 - Distribution des emprunts

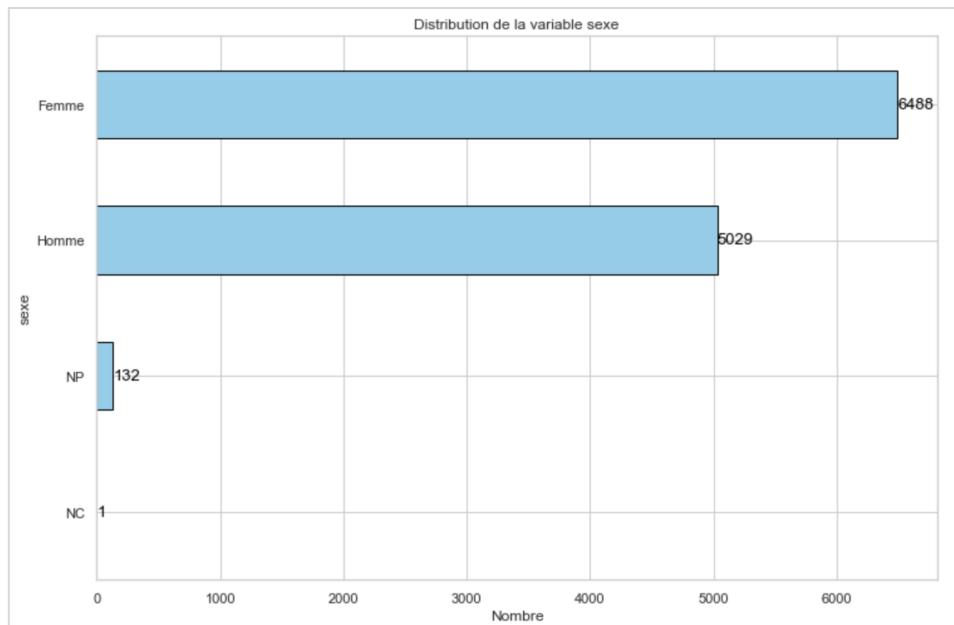
- Tranche d'âge : en analysant de plus près les caractéristiques démographiques de nos adhérents, une observation notable concerne la répartition par tranche d'âge. Environ 40% des adhérents se situent dans la fourchette d'âge de 25 à 64 ans, ce qui peut s'expliquer par la largeur de cet intervalle englobant une portion significative de la population adulte.

La deuxième catégorie la plus représentée est celle des enfants de moins de 14 ans, constituant 28% des adhérents, soulignant ainsi l'importance des jeunes lecteurs au sein de la communauté. Les adolescents, âgés de 15 à 24 ans, occupent la troisième position en termes de fréquentation. Par ailleurs, il est intéressant de noter la présence de membres plus âgés au sein de notre base d'adhérents, reflétant une diversité générationnelle au sein de la médiathèque.



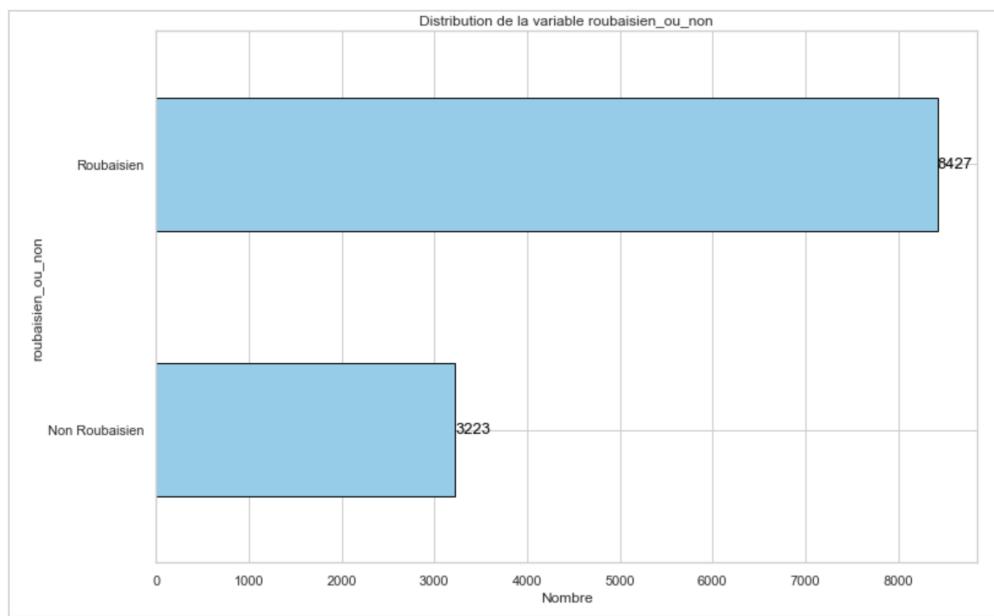
Graphique 2 - Répartition des adhérents selon leur âge

- Sexe : nous constatons une prédominance des adhérentes de sexe féminin par rapport aux adhérents masculins, avec une représentation féminine atteignant 55% de l'ensemble des adhérents.



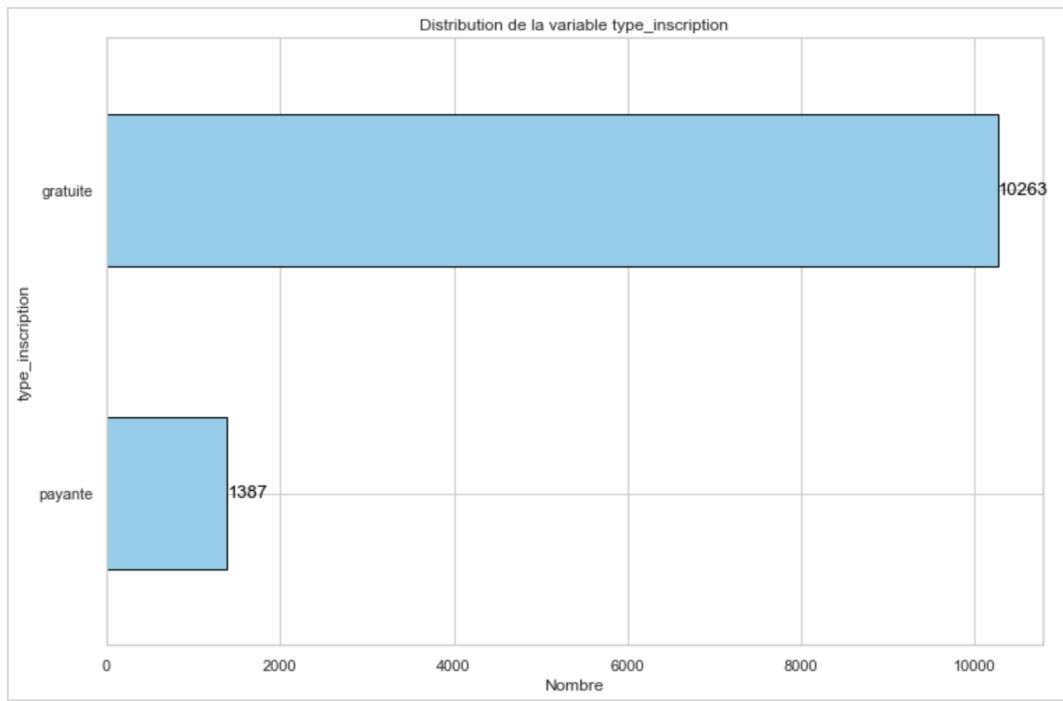
Graphique 3 - Répartition des adhérents selon leur sexe

- Appartenance à la communauté de Roubaix : il est intéressant de constater que la majorité des adhérents de la médiathèque, soit 72%, proviennent de Roubaix. Toutefois, la portée de la médiathèque s'étend au-delà de ses frontières municipales, attirant également des résidents des communes avoisinantes telles que Croix, Hem, Tourcoing, Wasquehal, et Wattrelos. Cette diversité géographique témoigne de l'attrait régional de la médiathèque, créant ainsi un espace culturel intercommunal au service d'une audience variée.



**Graphique 4 - Répartition des adhérents selon leur appartenance
à la communauté de Roubaix**

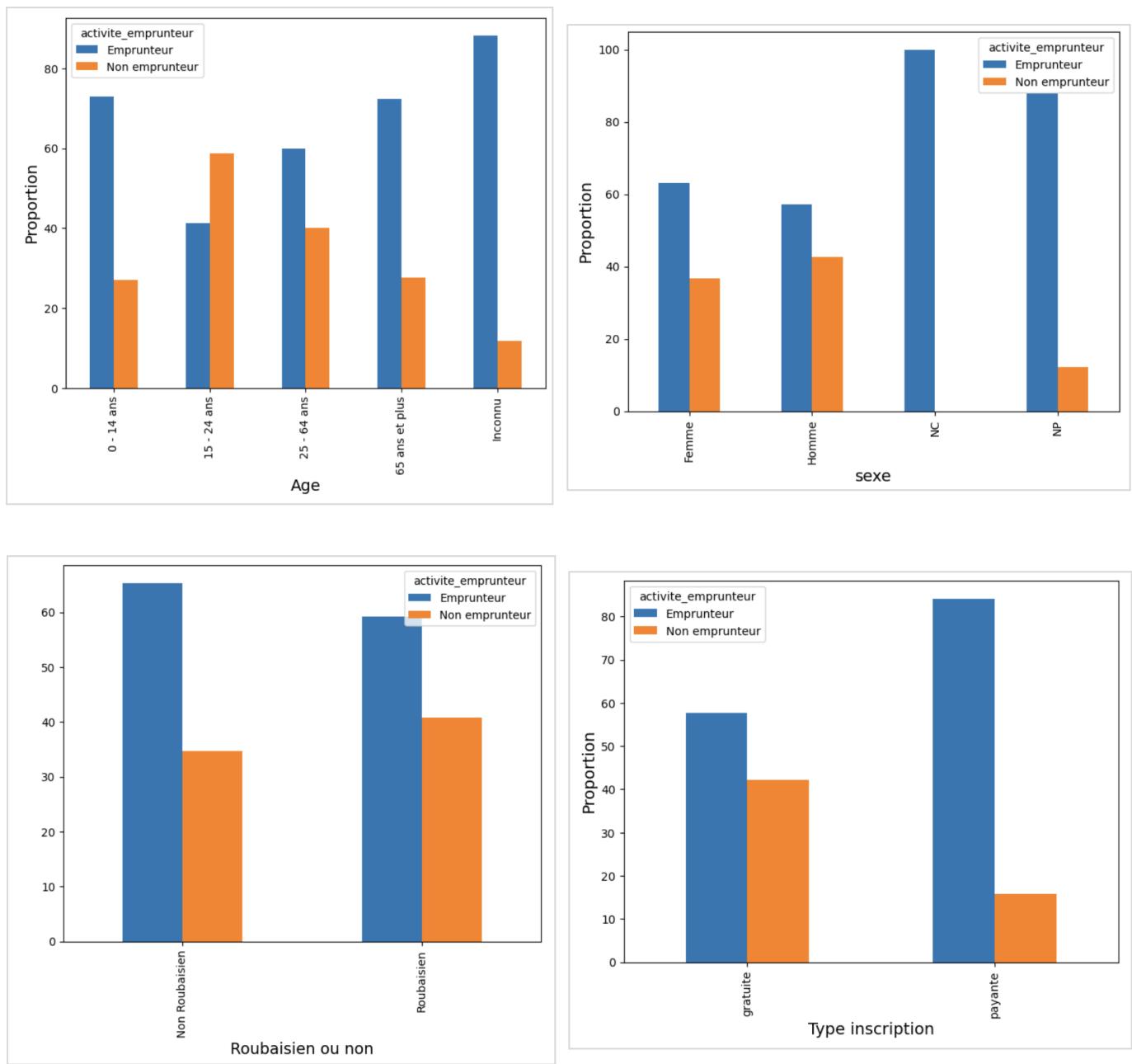
- Type d'inscription : si nous nous focalisons sur les variables relatives à l'inscription des adhérents. Nous pouvons voir que la grande majorité, 88% de la population étudiée, n'a pas eu à débourser de frais d'inscription pour obtenir leur carte d'adhérent. Parmi les différentes options de cartes payantes, une gamme tarifaire variée est offerte, offrant des tarifs différenciés en fonction des services proposés. Cette diversité de tarifs permet notamment aux utilisateurs de bénéficier de la possibilité d'emprunter un plus grand nombre de documents.



Graphique 5 - Répartition des adhérents selon leur type d'inscription

En effectuant une analyse croisée de nos différentes variables avec la variable cible, plusieurs observations majeures émergent :

- Au sein des différentes tranches d'âges, la majorité, soit plus de 60%, est constituée d'emprunteurs, à l'exception de la tranche d'âge entre 15 et 24 ans où cette proportion diminue ;
- Malgré une prédominance d'adhérentes féminines, la proportion d'emprunteurs est presque équivalente à celle des hommes. Ce constat souligne une participation active des deux sexes dans les emprunts, malgré les différences numériques ;
- Qu'ils soient résidents roubaisiens ou non, la majorité des adhérents, soit plus de 60%, sont des emprunteurs ;
- Lorsque les adhérents ont souscrit à un abonnement payant, nous observons une nette réduction du nombre de non-emprunteurs, la proportion de ceux-ci s'élevant à environ 15%.

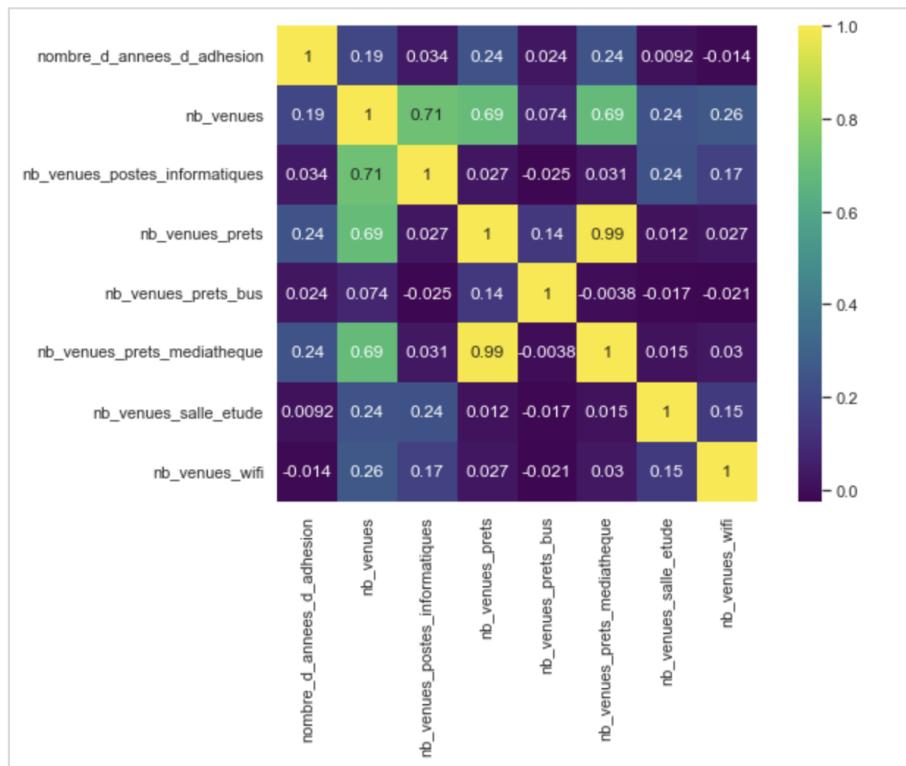


Graphique 6 - Analyse croisée avec la variable emprunt

4. Liaisons entre les variables

Nous entamons à présent l'analyse de la relation entre les différentes variables en vue de la préparation à la modélisation. Il est important que nos variables ne présentent pas une corrélation excessive entre elles, évitant ainsi la redondance d'informations. Simultanément, elles doivent influencer notre variable cible, à savoir le comportement d'emprunt ou non de l'adhérent.

Afin d'évaluer ces relations, nous avons réalisé une matrice de corrélation spécifiquement pour les variables numériques. Cette matrice est représentée graphiquement par une échelle de couleurs allant du violet au jaune, mettant en lumière l'intensité de la corrélation entre deux variables données. Une couleur plus proche du jaune indique une corrélation plus forte, signifiant une liaison significative entre les deux variables. Dans ces cas, il devient impératif de faire un choix judicieux en conservant uniquement l'une des deux variables afin d'éviter une redondance d'informations dans notre modèle.



Graphique 7 : Matrice de corrélation

Nous avons observé une corrélation significative entre la variable `nb_venues` et plusieurs autres variables (*cf graphique 7*), notamment le nombre de visites aux postes informatiques, le nombre d'emprunts et le nombre de visites à la médiathèque. Nous considérons que la variable `nb_venues` demeure la plus pertinente parmi celles mentionnées. Par conséquent, nous avons décidé de supprimer les autres variables mentionnées, estimant qu'elles n'apportent pas autant de valeur prédictive à notre modèle. Toutes les autres corrélations semblent appropriées et seront prises en compte dans le processus de modélisation.

En ce qui concerne les variables catégorielles, nous nous sommes appuyés sur le V de Cramer pour évaluer l'intensité de la dépendance entre ces variables. Une valeur élevée de ce coefficient indique une dépendance plus forte entre les variables.

	<code>activite</code>	<code>activite_emprunteur</code>	<code>activite_emprunteur_bus</code>	<code>activite_emprunteur_med</code>	<code>activite_salle_etude</code>	a
<code>activite</code>	1.0	0.999399	0.189174	0.95606	0.999399	
<code>activite_emprunteur</code>	0.999399	0.99982	0.176165	0.956148	0.065427	
<code>activite_emprunteur_bus</code>	0.189174	0.176165	0.999031	0.019787	0.040276	
<code>activite_emprunteur_med</code>	0.95606	0.956148	0.019787	0.999823	0.055213	
<code>activite_salle_etude</code>	0.999399	0.065427	0.040276	0.055213	0.999012	
<code>activite_utilisateur_postes_informatiques</code>	0.999399	0.197279	0.075318	0.172148	0.137355	
<code>activite_utilisateur_wifi</code>	0.999399	0.092486	0.040158	0.079236	0.210781	
<code>tranches_d_age_1</code>	0.165798	0.169925	0.138598	0.143468	0.113011	
<code>tranches_d_age_2</code>	0.177192	0.242826	0.151442	0.215494	0.147587	
<code>roubaisien_ou_non</code>	0.135872	0.055326	0.097286	0.076157	0.0	
<code>code_iris_de_roubaix</code>	0.063958	0.147987	0.248369	0.158082	0.037819	
<code>nom_de_l_iris_a_roubaix</code>	0.069153	0.148456	0.243247	0.157312	0.046326	
<code>commune_de_residence</code>	0.06132	0.150795	0.0	0.154203	0.069607	
<code>inscription_carte</code>	0.143143	0.39872	0.080323	0.389311	0.077378	
<code>nombre_d_annees_d_adhesion</code>	0.056616	0.161143	0.02593	0.163195	0.040285	

Graphique 8 : Matrice de V Cramer (partielle)

La commune de résidence et la variable indiquant si l'adhérent était originaire de Roubaix ou non présentent une corrélation significative. Suite à cette observation, nous avons pris la décision de conserver uniquement la variable décrivant si l'adhérent était roubaïen ou non, considérant qu'elle offre une représentation plus concise et pertinente de l'information recherchée.

Toutes comptes fait, les variables suivantes ont fait l'objet d'une suppression :

- `activite`, car elle apporte des informations similaires que la variable cible `activite_emprunteur`
- `nb_venues_postes_informatiques`, car elle apporte moins d'information que `nb_venues`
- `nb_venues_prets`, car elle apporte moins d'information que `nb_venues`
- `nb_venues_prets_mediatheque`, car même information que `nb_venues_prets`
- `activite_emprunteur_med`, car trop corrélée avec `activite_emprunteur`
- `nb_venues_prets_bus`, car même information que `activite_emprunteur_bus`
- `nb_venues_salle_etude`, car même information que `activite_salle_etude`
- `nb_venues_wifi`, car même information que `activite_utilisateur_wifi`
- `nb_venues`, car trop corrélée avec `activite_emprunteur`

Finalement, notre prédiction porte sur les 11 650 observation et prendra en compte l'ensemble des 7 variables suivantes :

<code>activité_emprunteur</code>
<code>type_inscription</code>
<code>inscription_carte</code>
<code>nombre d'années d'adhésion</code>
<code>tranches d'âge (2)</code>
<code>roubaisien ou non</code>
<code>sexe</code>

Tableau 2 - Liste des variables de notre prédiction

III. Détails du travail

Nous avons procédé à un recodage de nos variables en binaire à travers la création des variables dummy pour chacune des modalités des variables que nous avons retenues. La nouvelle variable correspondante prend la valeur 1 lorsque la modalité est présente pour l'observation en question et sinon 0.

Nous avons appliqué une méthode d'échantillonnage à notre base de données, la fractionnant en deux ensembles distincts : un échantillon d'apprentissage et un échantillon de test. L'échantillon d'apprentissage, constitué de 8 155 observations, sera utilisé pour élaborer le modèle. En parallèle, l'échantillon de test, comprenant 3 495 observations, sera déployé pour évaluer la performance du modèle. Cette démarche permet de valider la capacité de généralisation du modèle sur des données non incluses dans sa phase d'apprentissage, ainsi que de mesurer sa robustesse temporelle.

Nous avons poussé notre code sur la plateforme GitHub correspondante pour le partager avec l'équipe de manière collaborative, afin de faciliter la gestion des versions et permettre une meilleure traçabilité des modifications : https://github.com/bentouhamiloubna/projet_mlops

IV. Modélisation des emprunts de la médiathèque

La modélisation statistique est entre autres un moyen d'expliquer approximativement le lien qui existe entre une variable dite d'intérêt et une ou des variables dites explicatives. Le modèle statistique est l'équation mathématique utilisée pour décrire le mécanisme qui a généré les données étudiées.

La régression logistique permet d'étudier les relations entre un ensemble de variables qualitatives et une variable qualitative dichotomique. Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien. Ce modèle permet aussi de prédire la probabilité qu'un événement arrive ou non à partir de l'optimisation des coefficients de régression. Ce résultat varie toujours entre 0 et 1.

Nous allons de ce fait utiliser la régression logistique qui répond parfaitement à notre problématique d'étude pour prédire si un adhérent de la médiathèque est emprunteur ou non. Car le meilleur classifieur constant est celui qui a l'erreur de test minimale, nous allons prendre le classifieur qui prédit 0, avec un taux d'erreur de 38.92 %.

1. Etablissement du modèle

L'équation du modèle fondé sur la fonction logistique est la suivante :

$$P(Y_i = 1) = 1/(1 + e^{-X_i \alpha})$$

avec X_i : les variables explicatives de l'étude

2. Qualité du modèle

Qualité de la prédition

Nous allons utiliser le R^2 de Mcfadden qui mesure combien de déviances est expliqué par le modèle. La déviance a un rôle similaire à la variance résiduelle dans le cas de la variance dans un modèle linéaire. Si le R^2 obtenu est bas, nous pouvons dire que le modèle n'est pas très prédictif, nous n'avons donc pas identifié tous les facteurs qui prédisent le fait d'avoir un salaire élevé.

Dans notre analyse, le R² de McFadden atteint 0.18, démontrant un ajustement du modèle à un niveau significatif. Cette constatation est renforcée par des résultats satisfaisants observés dans d'autres mesures de performance telles que la courbe ROC par exemple.

Significativité globale

Tester la significativité globale d'un modèle revient à évaluer si l'ensemble des variables du modèle, collectivement, ont un impact statistiquement significatif sur la variable d'intérêt, le salaire.

Pour ce faire, nous allons utiliser le test du maximum de vraisemblance. Nous admettons les hypothèses suivantes:

H0 : Toutes les restrictions posées sont vraies (le modèle n'est pas globalement significatif)

H1 : Au moins une des restrictions posées est fausse (le modèle est globalement significatif)

L'hypothèse nulle est rejetée si la probabilité obtenue est inférieure à 0.05. Dans notre cas, nous obtenons une p-valeur égale à 0.69, ce qui suggère que le modèle effectué n'est globalement pas très significatif.

Significativité individuelle

La significativité individuelle d'une variable dans un modèle vise à évaluer si cette variable en particulier a un impact statistiquement significatif sur la variable dépendante (le salaire), indépendamment des autres variables présentes dans le modèle.

Le test du Khi2 de Wald permet de tester l'effet des variables explicatives sur le salaire. Il consiste à vérifier si la nullité de chacune des modalités de la variable enlève ou non une information. Si le test s'avère significatif (p-value<0.05) alors la variable ajoute une information supplémentaire.

H0 : Aucune des variables explicatives n'a d'effet sur le salaire

H1 : Au moins l'une des variables explicatives a un effet sur le salaire

Le résultat obtenu offre des perspectives importantes sur les variables non significatives. Ces variables, dont les p-values dépassent le seuil conventionnel de 0.05, ne montrent pas d'association significative avec la probabilité d'être emprunteur. Parmi ces variables non significatives, nous avons :

- inscription_carte_Collectivités (Classes maternelles et primaires)
- inscription_carte_Médiathèque
- inscription_carte_Médiathèque Plus (17 €)
- inscription_carte_Médiathèque Plus (Personnel médiathèque)
- type_inscription_payante
- sexe_NC
- sexe_NP
- inscription_carte_Médiathèque Plus (35 €)
- inscription_carte_Médiathèque Plus (5 €)
- inscription_carte_Médiathèque Plus (Personnel médiathèque)

Ces modalités n'ont donc pas augmenté ou diminué significativement la probabilité d'un individu par rapport aux autres modalités qui sont significatives.

3. Estimation du modèle

Passons à présent à l'estimation des coefficients par l'analyse du maximum de vraisemblance. En effet, pour estimer les paramètres du modèle, nous avons utilisé la méthode du maximum de vraisemblance qui consiste à choisir une valeur de probabilité qui maximise la probabilité d'avoir un emploi de qualité. Les estimateurs ainsi obtenus ont comme propriétés d'être asymptotiquement sans biais, efficaces et convergents.

L'interprétation des coefficients ne peut pas se faire directement dans ce type de modèle, elle doit être ramenée en termes de rapport de chance ou de probabilité, ce qui s'opère avec les formules suivantes :

$$\text{rapport de chance} = \exp(\text{coefficient estimé})$$

$$\text{probabilité} = \text{rapport de chance}/(\text{rapport de chance} + 1)$$

C'est ainsi que nous obtenons le tableau suivant. Pour faciliter la lecture, les modalités non significatives ont été ombrées.

	Coeff	Rapport de chances	Probabilité	P> z
const	-4.06	0.02	1.70%	
tranches_d_age_2_15 - 24 ans	-1.22	0.30	22.83%	0.00
tranches_d_age_2_25 - 64 ans	-0.61	0.54	35.19%	0.00
tranches_d_age_2_65 ans et plus	-0.36	0.70	41.18%	0.00
tranches_d_age_2_Inconnu	-1.09	0.34	25.11%	0.39
roubaisiens_ou_non_Roubaisiens	-0.42	0.66	39.61%	0.00
inscription_carte_Collectivités (Classes maternelles et primaires)	0.10	1.11	52.51%	0.91
inscription_carte_Collectivités (Structures non scolaires)	-0.05	0.96	48.87%	0.97
inscription_carte_Consultation sur place	-23.58	0.00	0.00%	1.00
inscription_carte_Médiathèque	5.74	310.72	99.68%	1.00
inscription_carte_Médiathèque Plus (17 €)	1.44	4.23	80.89%	1.00
inscription_carte_Médiathèque Plus (35 €)	1.84	6.31	86.32%	1.00
inscription_carte_Médiathèque Plus (5 €)	1.06	2.89	74.31%	1.00
inscription_carte_Médiathèque Plus (Conservatoire)	0.25	1.29	56.32%	1.00
inscription_carte_Médiathèque Plus (Personnel médiathèque)	1.13	3.09	75.57%	1.00
inscription_carte_Prêt en nombre	-2.01	0.13	11.86%	0.03
nombre_d_annees_d_adhesion_1.0	-0.40	0.67	40.22%	0.00
nombre_d_annees_d_adhesion_2.0	-0.25	0.78	43.88%	0.00
nombre_d_annees_d_adhesion_3.0	-0.31	0.73	42.36%	0.00
nombre_d_annees_d_adhesion_4.0	-0.38	0.68	40.51%	0.00
nombre_d_annees_d_adhesion_5.0	-0.40	0.67	40.23%	0.00
nombre_d_annees_d_adhesion_6.0	-0.15	0.86	46.16%	0.18
nombre_d_annees_d_adhesion_7.0	0.03	1.03	50.67%	0.84
nombre_d_annees_d_adhesion_8.0	-0.14	0.87	46.49%	0.28
nombre_d_annees_d_adhesion_9.0	0.10	1.11	52.58%	0.46
nombre_d_annees_d_adhesion_10.0	-0.18	0.83	45.45%	0.23
nombre_d_annees_d_adhesion_11.0	0.18	1.19	54.39%	0.33
nombre_d_annees_d_adhesion_12.0	0.39	1.48	59.71%	0.05
nombre_d_annees_d_adhesion_13.0	0.46	1.59	61.32%	0.03
nombre_d_annees_d_adhesion_14.0	0.20	1.23	55.07%	0.37
nombre_d_annees_d_adhesion_15.0	0.57	1.77	63.95%	0.00
type_inscription_payante	5.73	308.12	99.68%	1.00
sexe_Homme	-0.29	0.75	42.75%	0.00
sexe_NC	11.51	100197.65	100.00%	0.97
sexe_NP	8.05	3135.99	99.97%	1.00

Tableau 3 : Coefficients estimés

Notre référence est la suivante = une fille âgée de moins de 15 ans, non roubaisienne, possédant une carte de collectivité (BCD) non payante, et adhérente depuis moins de 12 mois.

Lecture par signe :

La lecture des coefficients par signe consiste à se référer au signe du coefficient obtenu afin de conclure à une action négative ou positive de la modalité correspondante sur la qualité de l'emploi.

- être âgé de plus de 14 ans exerce un effet négatif sur la possibilité d'emprunter par rapport à un individu âgé de moins de 15 ans ;
- posséder une carte médiathèque plus, n'a pas de significativité sur la possibilité d'emprunter par rapport à la référence ;
- avoir un nombre d'années d'adhésion inférieur à 7 ans exerce un effet négatif sur la possibilité d'emprunter par rapport à la référence ;
- être un homme exerce un effet négatif sur le fait d'emprunter par rapport à une femme, toutes choses égales par ailleurs.

Lecture par coefficient

La lecture par coefficient consiste à évaluer la probabilité ou les chances d'obtenir un emploi de qualité par rapport à la situation de référence correspondante. Les coefficients obtenus nous amènent à la conclusion suivante :

- être âgé de 25 à 64 ans divise les chances d'emprunter par 2 ($1/0.54$) par rapport à un individu âgé de maximum 14 ans toutes choses égales par ailleurs ;
- être inscrit à la médiathèque augmente les chances de 100% d'emprunter par rapport à la référence ;
- être roubaisien augmente les chances de 40% d'emprunter par rapport à un non roubaisien toutes choses égales par ailleurs ;
- être adhérent depuis 15 ans à la médiathèque augmente les chances de 64% d'emprunter par rapport à un adhérent de moins d'un an ;

- avoir une inscription payante augmente les chances de 100% d'emprunter par rapport à une inscription gratuite ;
- la probabilité qu'un homme emprunte est de 43% par rapport à une femme toutes choses égales par ailleurs.

Lecture de la référence

La situation de référence ne peut pas s'interpréter par rapport de chance (r), celui-ci est basculé en probabilité ($p = r / (r+1)$).

Ainsi, la probabilité que l'individu de référence emprunte à la médiathèque de Roubaix est de 1.7%. C'est-à-dire, une fille de moins de 15 ans, non roubaisienne et qui à une adhésion de moins de 1 an.

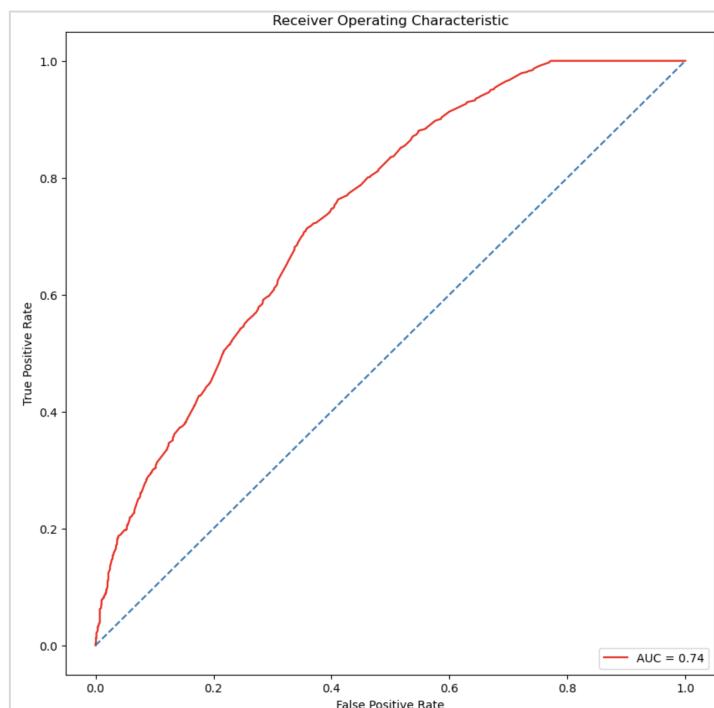
V. Evaluation de la performance du modèle

Pour évaluer la performance de notre modèle de prédiction, plusieurs possibilités s'offrent à nous, notamment à travers les différents indicateurs de mesure.

Nous avons appliqué notre modèle de prédiction sur notre échantillon de test et nous sommes venues évaluer le taux d'erreur entre celle prédictive et celle réelle concernant le comportement d'emprunt de l'adhérent. L'erreur sur la base de test représente la proportion d'observations mal classées par le modèle. Dans notre cas, environ 29.15% des prédictions du modèle sont incorrectes.

En parallèle de ce taux, nous avons le taux de bon classement qui représente la proportion d'observations qui ont été bien prédictes par le modèle. Ici, nous sommes environ à 70.84% de prédictions correctes.

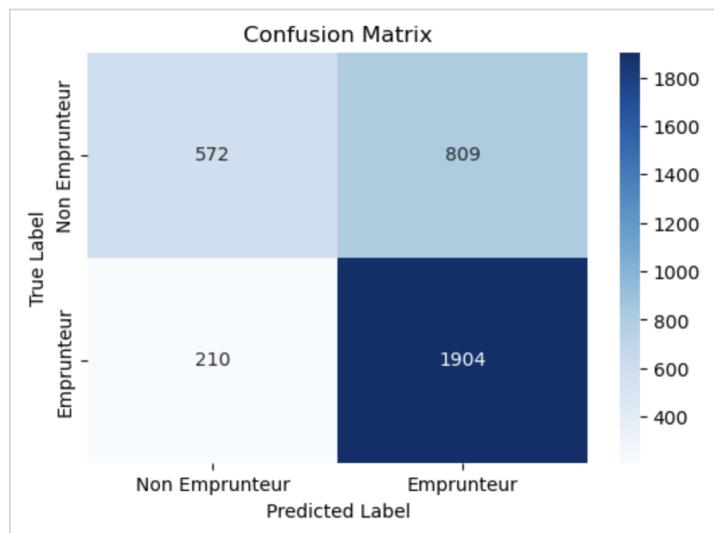
L'aire sous la courbe ROC mesure la capacité d'un modèle à distinguer entre les classes. Elle varie de 0 à 1, où 1 indique une performance parfaite. Une valeur de 0.7396 suggère que le modèle a une capacité modérée à discriminer entre les classes.



Graphique 9 - Courbe de ROC

La matrice de confusion résume les performances du modèle en termes de vrais positifs (TP), faux positifs (FP), vrais négatifs (TN) et faux négatifs (FN) :

- 1904 vrais positifs (TP) : observations correctement classées comme positives ;
- 809 faux positifs (FP) : observations incorrectement classées comme positives ;
- 210 faux négatifs (FN) : observations incorrectement classées comme négatives ;
- 572 vrais négatifs (TN) : observations correctement classées comme négatives.



Graphique 10 - Matrice de confusion

La sensibilité (Recall ou True Positive Rate) correspond à la proportion des vrais positifs parmi toutes les observations réellement positives. Elle mesure la capacité du modèle à détecter les cas positifs. Notre modèle prédit à hauteur de 90% les emprunteurs de la médiathèque : $1904 / (1904 + 210) \approx 0.90$

La précision quant à elle sort la proportion des vrais positifs parmi toutes les observations prédictes comme positives. Elle mesure la précision du modèle parmi ses prédictions positives. Nous avons ici un modèle qui prédit à 70% des emprunteurs : $1904 / (1904 + 809) \approx 0.70$

Par conséquent, notre modèle appliqué sur l'échantillon de test nous montre qu'il a un peu plus de mal à prédire les non emprunteurs que les emprunteurs et prédit très bien les adhérents qui empruntent. Nous aboutissons au fait que notre modèle est bon mais n'est pas parfait et comporte des erreurs.

Conclusion

Dans cette étude approfondie visant à appréhender les dynamiques d'emprunt à la médiathèque de Roubaix à travers l'analyse des caractéristiques des adhérents, nous avons entrepris une démarche rigoureuse basée sur la mise en œuvre d'une régression logistique.

L'objectif principal était de dévoiler les liens complexes entre diverses variables et la probabilité d'emprunter ou non, une quête qui nous a conduit à des résultats à la fois instructifs et nuancés. Notre source de données, issue du portail Open Data Roubaix s'est révélée être une mine d'informations riches sur les adhérents de la médiathèque *La Grand Plage* au cours de l'année 2020. Située au cœur de la ville de Roubaix, cette institution culturelle, mise à la disposition du public par la municipalité, a servi de toile de fond pour notre exploration des comportements d'emprunt.

Les coefficients estimés du modèle ont émergé comme des phares éclairant les facteurs prédictifs. Ainsi, être âgé de 25 à 64 ans, par exemple, divise les chances d'emprunter par 2 par rapport à un individu âgé de maximum 14 ans, toutes choses égales par ailleurs. L'inscription à la médiathèque augmente les chances de 100%, tandis qu'être roubaisien confère un avantage de 40% par rapport à un non-roubaisien. L'ancienneté d'adhésion, le type d'inscription et le genre ont également émergé comme des déterminants significatifs, avec des nuances intéressantes dans leurs implications.

Quant à l'évaluation de la qualité du modèle, le R² de McFadden a atteint 0.18, indiquant un ajustement significatif du modèle. Cependant, la significativité globale s'est maintenue à 0.69, signalant que les facteurs étudiés ne sont peut-être pas statistiquement significatifs. Les résultats de performance du modèle, avec 30% d'erreurs, dévoilent une réalité où les non-emprunteurs sont moins bien prédits que les emprunteurs.

En conclusion, bien que notre modèle de régression logistique ait démontré sa pertinence en éclairant les déterminants des emprunts à la médiathèque, il présente des limites. Ces dernières, sous forme de 30% d'erreurs, nous rappellent la complexité des comportements humains et la nécessité continue d'affiner nos modèles pour mieux capturer la diversité des motivations et des décisions des adhérents de la médiathèque *La Grand Plage* de Roubaix.

Il est recommandé de configurer le modèle de manière à optimiser l'identification des adhérents susceptibles de ne pas emprunter de documents, afin d'initier des actions ciblées visant à les encourager à modifier leurs habitudes d'utilisation de la médiathèque.