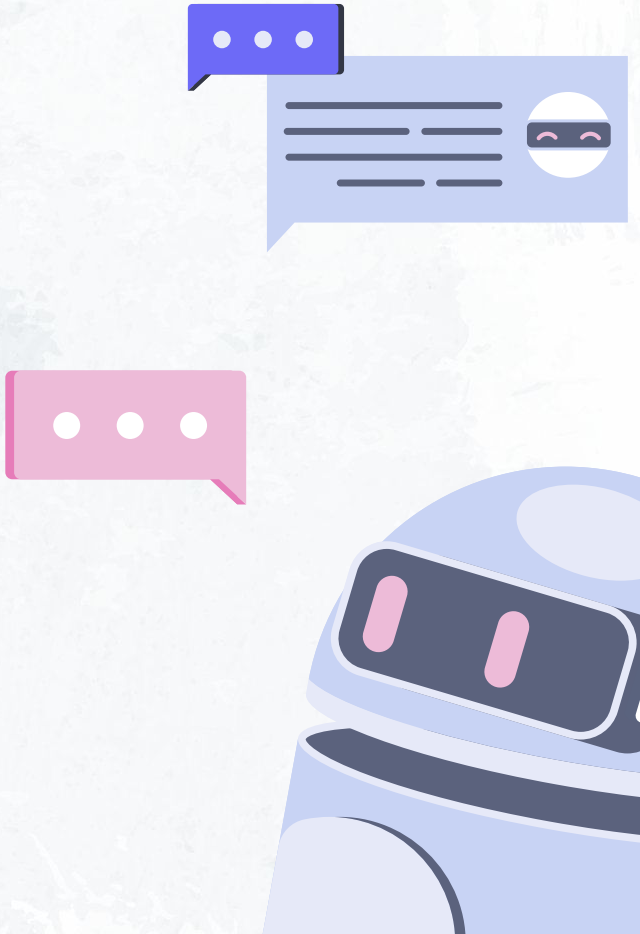
Construisez un modèle de scoring ⟶



Chapitres

01 →

Rappel du contexte et de l’objectifs

Une société financière, nommée **"Prêt à dépenser"**, propose des crédits à la consommation pour des personnes ayant peu ou pas d'historique de prêt.

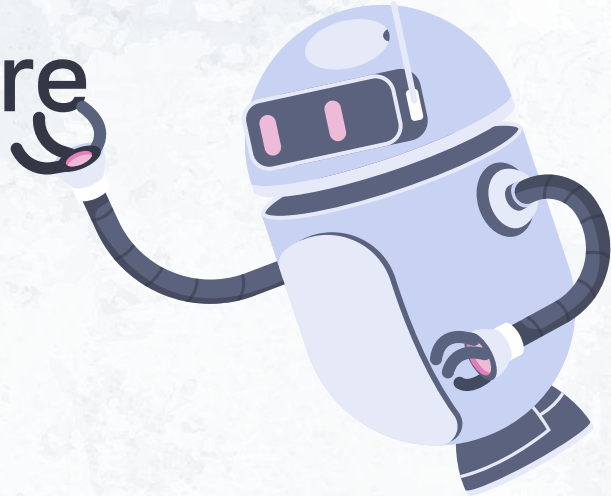
Pour accorder un crédit à la consommation, l'entreprise souhaite mettre en œuvre un outil de "scoring crédit" qui calcule la probabilité qu'un client le rembourse ou non, puis classifie la demande : crédit accordé ou refusé.

Elle souhaite donc développer un **algorithme de classification** pour aider à décider si un prêt peut être accordé à un client.

Les **chargés de relation client** seront les utilisateurs de l'outil de scoring. Puisqu'ils s'adressent aux clients, ils ont besoin que votre modèle soit **facilement interprétable**.

Les chargés de relation souhaitent, en plus, disposer d'**une mesure de l'importance des variables** qui ont poussé le modèle à donner cette probabilité à un client.

02 → Analyse Exploratoire

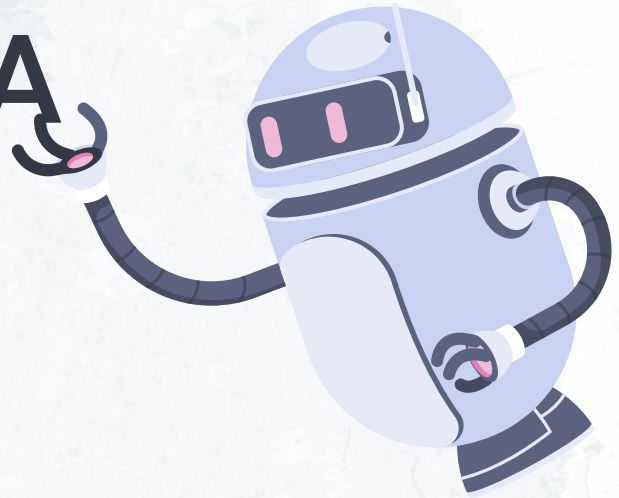


Au cours de l'analyse exploratoire, nous avons examiné les dimensions des données, les types de données, les valeurs manquantes, ainsi que les premières lignes de chaque jeu de données.

Cette étape initiale nous a permis de mieux comprendre la nature des données avec lesquelles nous travaillons et de préparer le terrain pour la création d'un modèle de scoring crédit efficace.

Nom du dataframe	Nombre d'entrées	Nombre de colonnes	Nombre de valeurs manquantes	Nombre de colonnes de type float64	Nombre de colonnes de type int64	Nombre de colonnes de type object
APPLICATION_TEST	48 744	121	140 441 900	65	40	16
APPLICATION_TRAIN	307 511	122	9 152 465	65	41	16
BUREAU	1 716 428	17	3 939 947	8	6	3
BUREAU_BALANCE	27 299 925	3	X	2	X	1
CREDIT_CARD_BALANCE	3 840 312	23	5 877 356	15	7	1
HOMECREDIT_COLUMNS_DESCRIPTION	219	5	133	1	X	4
INSTALLMENTS_PAYMENTS	13 605 401	8	5 810	5	3	X
POS_CASH_BALANCE	10 001 358	8	52 158	2	5	1
PREVIOUS_APPLICATION	1 670 214	37	11 109 336	15	6	16

03 → EDA



Sous - sections

A → Nettoyage des jeux de données

B → Analyse Univariée

C → Analyse Bivariée

D → Analyse Multivariée

E → Features Engineering

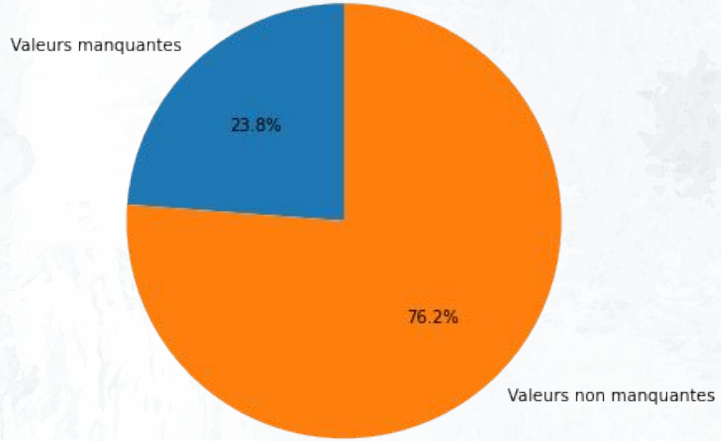
F → Features Selection

(A) Nettoyage des jeux de données ⟶

Dans cette section, nous avons examiné la présence de données manquantes au sein des ensembles de données 'application_test' et 'application_train'.

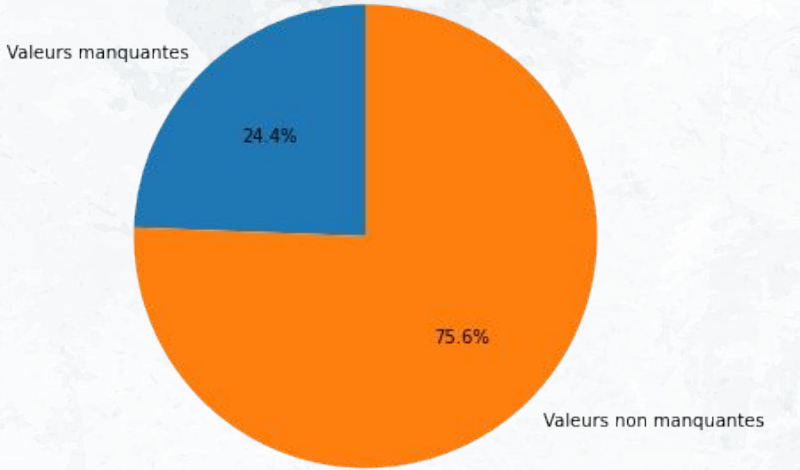
application_test

Comparaison du nombre total de valeurs avec le nombre de valeurs manquantes



application_train

Comparaison du nombre total de valeurs avec le nombre de valeurs manquantes



(A) Nettoyage des jeux de données ⟶

Notre étude a révélé que le taux de valeurs manquantes dans le jeu de données de test est de 23,8%, tandis que celui du jeu de données d'entraînement est de 24,4%.

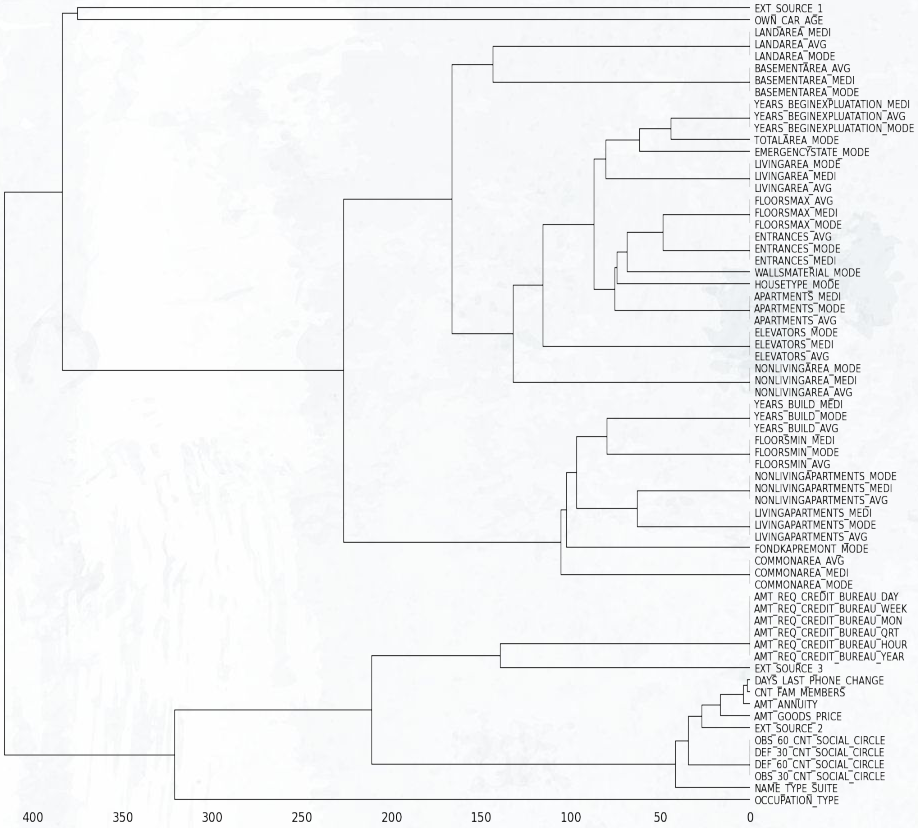
Conformément aux recherches menées par Dr. Graham et Dr. Schafer, il est généralement conseillé d'éviter l'imputation des données lorsque le taux de données manquantes est inférieur à 30%.

Cette ligne directrice est fondée sur les risques associés à l'imputation dans des situations où le pourcentage de données manquantes est relativement faible.

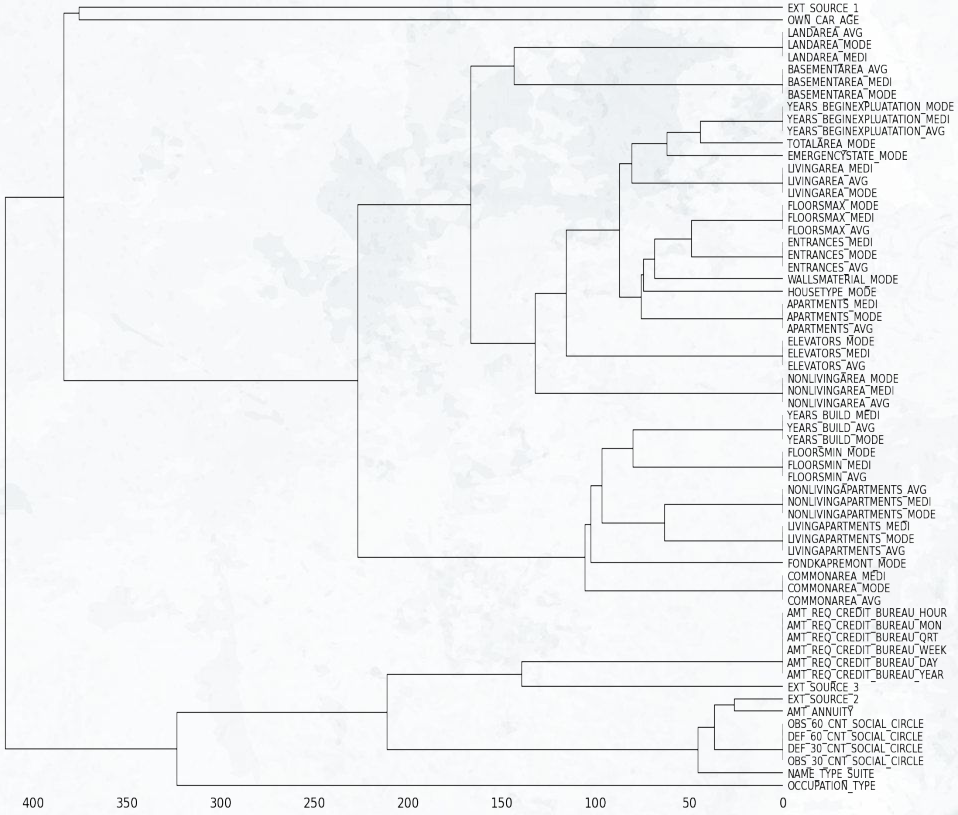
- The prevention and handling of the missing data
- When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts

(A) Nettoyage des jeux de données →

application_train



application_test



(A) Nettoyage des jeux de données ⟶

L'analyse détaillée du dendrogramme, créé via l'outil missingno, a mis en évidence plusieurs groupes de variables affichant des motifs de valeurs manquantes similaires.

Cette découverte laisse supposer une corrélation entre ces ensembles de variables, suggérant l'existence d'un mécanisme commun influençant l'apparition des données manquantes.

Pour chaque groupe identifié, nous avons déterminé le pourcentage de données manquantes, afin d'évaluer l'ampleur du problème.

Lorsque ce pourcentage était inférieur à 30%, nous avons privilégié une approche d'imputation, en sélectionnant la méthode la plus adaptée à la nature des variables concernées pour maintenir la fiabilité des données.

En revanche, pour les variables où le taux de valeurs manquantes dépassait 30%, nous avons choisi de les exclure de notre analyse.

Cette décision visait à réduire le risque d'introduire des biais dans nos analyses ultérieures

(B) Analyse Univariée ⟶

Dans notre étude des caractéristiques, nous avons adopté une démarche analytique rigoureuse, en distinguant clairement les variables qualitatives des quantitatives.

Cette distinction s'est avérée essentielle, guidant notre stratégie d'analyse et le choix des techniques statistiques adéquates.

Pour les variables qualitatives, telles que le genre, le niveau d'éducation, le statut familial et le type de logement, nous avons porté notre attention sur leur distribution empirique.

Cette méthode nous a permis d'examiner la fréquence et la répartition des différentes catégories au sein de notre échantillon.

À cet effet, nous avons utilisé des graphiques tels que des diagrammes en camembert et des histogrammes, fournissant une visualisation claire et intuitive des proportions de chaque catégorie.

Concernant les variables quantitatives, notre approche a été plus approfondie.

Nous avons employé des histogrammes et des graphiques de densité pour observer leur distribution, des boîtes à moustaches pour détecter les valeurs aberrantes, et des indicateurs statistiques tels que la moyenne, la médiane, la variance, l'écart-type, l'asymétrie et le kurtosis pour une compréhension exhaustive.

Cette analyse minutieuse nous a permis de comprendre non seulement la tendance centrale de la distribution, mais aussi sa dispersion et sa forme.

Dans les diapositives suivantes, nous présenterons de manière concise ces approches, en prenant comme exemple les indicateurs financiers.

Il est important de noter que cette analyse a été appliquée à l'ensemble des catégories de variables

(B) Analyse Univariée ⟶

Dans notre recherche visant à évaluer le risque de crédit, une importance cruciale a été accordée à l'analyse univariée des variables jugées essentielles.

Ces variables ont été soigneusement choisies pour leur pertinence potentielle dans l'appréciation de la solvabilité et des habitudes de remboursement des emprunteurs.

Notre étude univariée s'est focalisée sur les catégories suivantes :

Indicateurs Financiers :

AMT_INCOME_TOTAL : Revenu total de l'emprunteur.

AMT_CREDIT : Montant du crédit octroyé.

AMT_ANNUITY : Montant de l'annuité.

AMT_GOODS_PRICE : Valeur des biens pour lesquels le crédit est accordé.

Données Démographiques :

CODE_GENDER : Genre de l'emprunteur.

DAYS_BIRTH : Âge de l'emprunteur.

NAME_EDUCATION_TYPE : Niveau d'éducation de l'emprunteur.

NAME_FAMILY_STATUS : Statut familial de l'emprunteur.

NAME_HOUSING_TYPE : Type de logement de l'emprunteur.

Historique Professionnel :

DAYS_EMPLOYED : Durée de l'emploi actuel.

OCCUPATION_TYPE : Profession de l'emprunteur.

ORGANIZATION_TYPE : Type d'organisation employant l'emprunteur.

Indicateurs de Propriété :

FLAG_OWN_CAR : Possession d'une voiture.

FLAG_OWN_REALTY : Possession d'un bien immobilier.

OWN_CAR_AGE : Âge du véhicule possédé.

Évaluations Régionales :

REGION_RATING_CLIENT : Évaluation du client au niveau régional.

REGION_RATING_CLIENT_W_CITY : Évaluation du client au niveau de la ville.

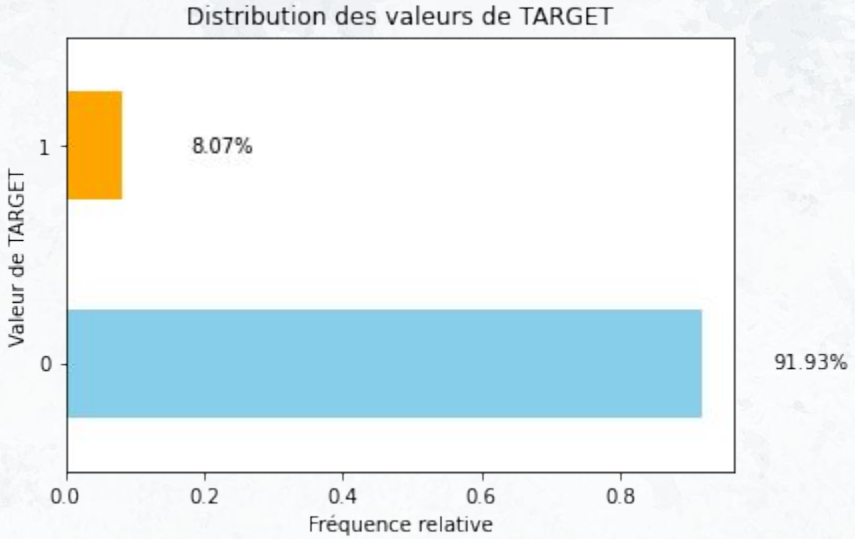
(B) Analyse Univariée ⟶

Distribution de la variable target :

Pour appréhender la distribution des cas dans notre jeu de données, une analyse approfondie de la variable cible, représentant l'issue du risque de crédit, a été menée.

Cette variable est d'une importance capitale pour notre modèle prédictif, car elle détermine l'objectif de prédiction.

Nous avons observé une distribution déséquilibrée de cette variable, ce qui a indique la nécessité de recourir à des techniques spécifiques dans la modélisation pour éviter un biais en faveur de la classe majoritaire.



(B) Analyse Univariée ⟶

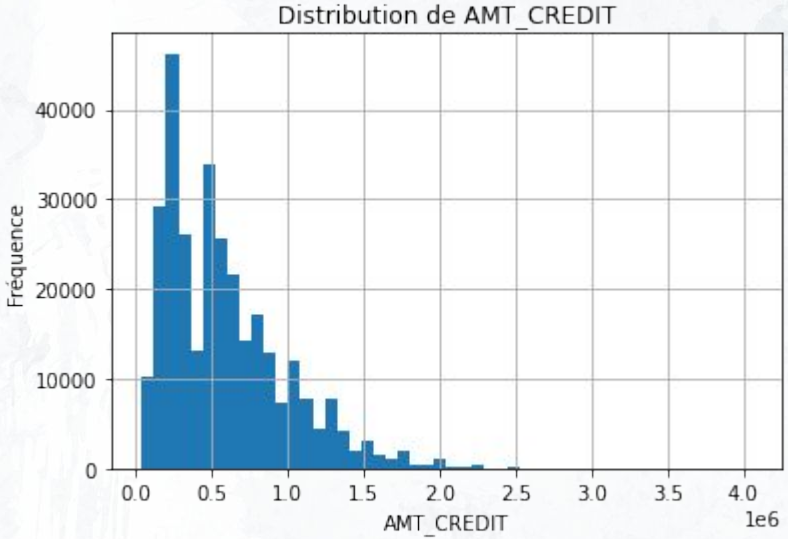
Exploration de la Distribution des Variables :

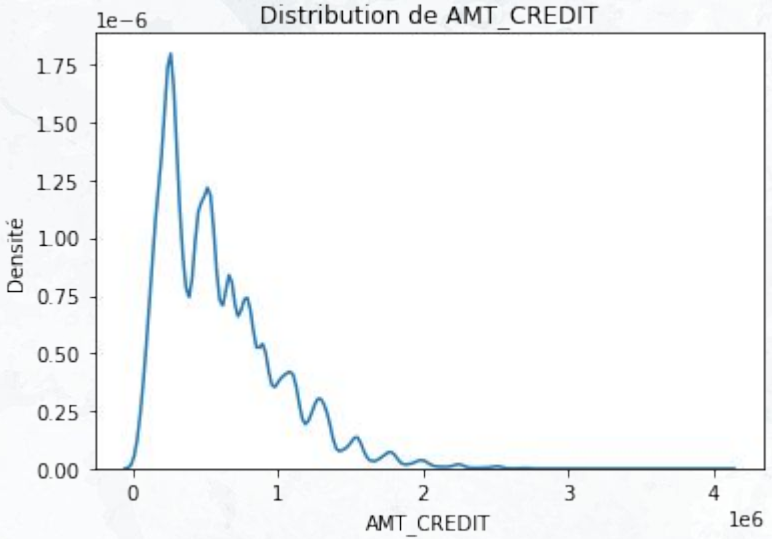
Nous avons entamé notre analyse en explorant la distribution des différentes variables.

Pour ce faire, nous avons utilisé des histogrammes et des diagrammes de densité, qui nous ont permis de visualiser la distribution de chaque variable.

Cette approche a facilité notre compréhension de la répartition des données et a aidé à identifier les tendances préliminaires.

Exemple d'histogrammes et diagrammes de densité pour AMT_CREDIT:





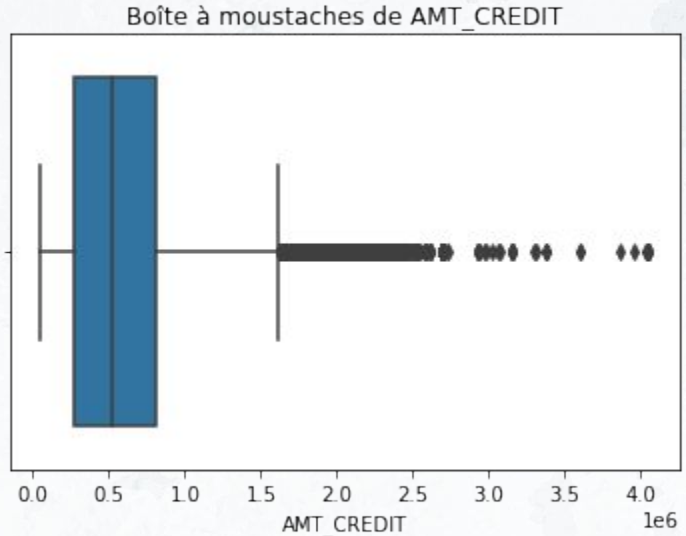
(B) Analyse Univariée ⟶

Analyse des Boîtes à Moustaches :

Pour une analyse plus approfondie de la dispersion des données et la détection des valeurs aberrantes, nous avons employé des boîtes à moustaches.

Cette technique nous a fourni une visualisation claire de l'étendue et de la concentration des données, tout en mettant en évidence les points extrêmes.

Exemple de boite a moustaches pour AMT_CREDIT:



(B) Analyse Univariée ⟶

Analyse des Tendances Centrales :

Pour chaque variable, nous avons déterminé les mesures de tendance centrale, telles que la moyenne et la médiane, ainsi que les mesures de dispersion, incluant la variance et l'écart-type.

Ces calculs offrent des insights cruciaux sur la localisation centrale et la variabilité au sein de nos données.

Exemple pour les Indicateurs Financiers :

	Variable	Moyenne	Médiane	Variance	Écart-Type	Plage	Asymétrie	Kurtosis
0	AMT_INCOME_TOTAL	168797.919297	147150.0	5.622739e+10	237123.146279	116974350.0	391.559654	191786.554381
1	AMT_CREDIT	599025.999706	513531.0	1.619988e+11	402490.776996	4005000.0	1.234778	1.934041
2	AMT_ANNUITY	27108.573909	24903.0	2.100684e+08	14493.737315	256410.0	1.579777	7.707320
3	AMT_GOODS_PRICE	538396.207429	450000.0	1.364907e+11	369446.460540	4009500.0	1.349000	2.431916

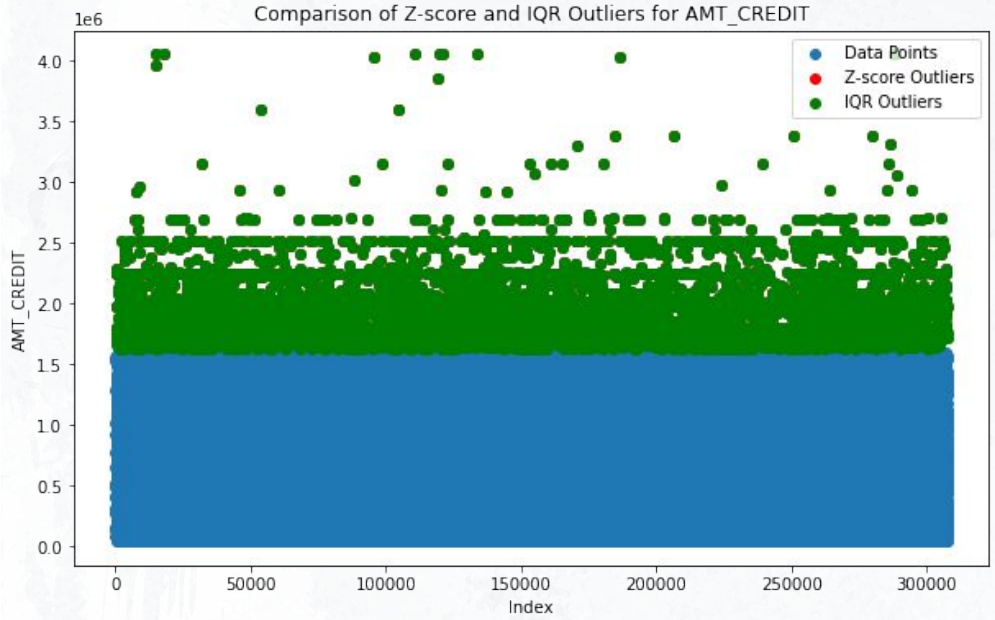
(B) Analyse Univariée ⟶

Détection et Analyse des Valeurs Aberrantes :

L'identification des valeurs aberrantes a été effectuée à travers les méthodes du score Z et de l'intervalle interquartile (IQR).

Cette démarche est essentielle pour garantir l'intégrité des données et prévenir les distorsions potentielles dans les analyses futures.

Exemple pour les Indicateurs Financiers :



	Variable	Nombre d'outliers détectés par Z-score	Pourcentage d'outliers détectés par Z-score	Nombre d'outliers détectés par IQR	Pourcentage d'outliers détectés par IQR
0	AMT_INCOME_TOTAL	454	0.147637	14035	4.564064
1	AMT_CREDIT	3255	1.058499	6562	2.133907
2	AMT_ANNUITY	0	0.000000	7504	2.440238
3	AMT_GOODS_PRICE	0	0.000000	14728	4.789422

(B) Analyse Univariée ⟶

Impact de la Suppression des Valeurs Aberrantes :

Suite à l'identification des valeurs aberrantes, nous avons procédé à leur élimination afin de purifier l'ensemble de données.

Par la suite, une comparaison des statistiques descriptives a été réalisée, avant et après cette suppression, afin d'illustrer l'effet significatif de cette opération sur les propriétés des données.

Exemple pour les Indicateurs Financiers :

	count_before	mean_before	std_before	min_before	25%_before	50%_before	75%_before	max_before	count_after	mean_after	std_after	mir
JRR	307511.0	278180.518577	102790.175348	100002.0	189145.5	278202.0	367142.5	456255.0	280858.0	278183.076736	102796.889497	10
GET	307511.0	0.080729	0.272419	0.0	0.0	0.0	0.0	1.0	280858.0	0.083248	0.276258	
REN	307511.0	0.417052	0.722121	0.0	0.0	0.0	1.0	19.0	280858.0	0.414163	0.720319	
TAL	307511.0	168797.919297	237123.146279	25650.0	112500.0	147150.0	202500.0	117000000.0	280858.0	151308.973070	62473.968682	2
EDIT	307511.0	599025.999706	402490.776996	45000.0	270000.0	513531.0	808650.0	4050000.0	280858.0	541660.818636	331592.420637	4
...	...	...	...	...	...	...	...	...	...	...	...	
DAY	265992.0	0.007000	0.110757	0.0	0.0	0.0	0.0	9.0	242313.0	0.006974	0.110229	
EEK	265992.0	0.034362	0.204685	0.0	0.0	0.0	0.0	8.0	242313.0	0.034026	0.203234	
MON	265992.0	0.267395	0.916002	0.0	0.0	0.0	0.0	27.0	242313.0	0.256412	0.887695	
QRT	265992.0	0.265474	0.794056	0.0	0.0	0.0	0.0	261.0	242313.0	0.263102	0.612131	
EAR	265992.0	1.899974	1.869295	0.0	0.0	1.0	3.0	25.0	242313.0	1.910950	1.872412	

(C) Analyse Bivariée

Toujours dans l'optique de développer un modèle de scoring crédit précis et efficace, la sélection des caractéristiques pertinentes est cruciale.

Notre analyse s'est concentrée sur les catégories suivantes :

Indicateurs Financiers :

AMT_INCOME_TOTAL : Les revenus totaux de l'emprunteur peuvent grandement influencer sa capacité à rembourser le prêt, faisant de cette variable un choix évident.

Données Démographiques :

CODE_GENDER : Bien que la sélection de cette variable soit délicate en raison de considérations éthiques, certaines études ont montré des différences dans les taux de défaut entre les genres. Il est crucial de traiter cette variable avec prudence et transparence.

DAYS_BIRTH : L'âge de l'emprunteur peut influencer la stabilité financière, et donc la probabilité de remboursement, surtout en tenant compte des risques associés à l'âge avancé

NAME_EDUCATION_TYPE : Le niveau d'éducation peut être corrélé à la stabilité de l'emploi et au revenu, impactant ainsi la capacité de remboursement.

NAME_FAMILY_STATUS : La situation familiale et le type de logement peuvent également influencer la stabilité financière de l'emprunteur.

Historique Professionnel :

DAYS_EMPLOYED : La durée de l'emploi est souvent un indicateur de stabilité financière.

OCCUPATION_TYPE et ORGANIZATION_TYPE :
Ces variables peuvent refléter le type et la stabilité de l'emploi, ainsi que le niveau de revenu.

Indicateurs de Propriété :

FLAG_OWN_CAR,
FLAG_OWN_REALTY,OWN_CAR_AGE

La possession de biens peut indiquer une certaine stabilité financière, ce qui peut influencer la capacité de l'emprunteur à rembourser le prêt.

(C) Analyse bivariée ⟶

Les recherches et publications suivantes éclairent les disparités de genre en matière de taux de défaut de prêt et de comportements financiers associés :

- Performance des Prêts et Disparités de Genre en Littératie Financière : L'étude de Chen et al. (2018) révèle que les prêts souscrits par des femmes présentent des taux de défaut plus élevés et des rendements inférieurs comparés à ceux des investisseurs masculins. Cette constatation souligne l'importance de considérer le genre dans l'analyse des performances de prêt[1].

- Disparités de Genre dans le Remboursement des Microcrédits : Les études menées dans plusieurs pays par D'Espallier et al. (2011) et Gul et al. (2017) indiquent qu'une proportion plus élevée de clientes dans les institutions de microfinance est corrélée à un risque de portefeuille réduit, moins de radiations de prêts, et une diminution des provisions pour pertes. Néanmoins, les mécanismes sous-jacents à ces différences de genre demeurent peu élucidés[2].

Ces études fournissent un fondement pour comprendre et justifier l'inclusion de la variable 'CODE_GENDER' dans notre analyse. Il est essentiel d'aborder l'utilisation de cette variable avec prudence et transparence, en prenant en compte les implications éthiques.

Il est important de reconnaître que les différences de genre en matière de taux de défaut de prêt peuvent être influencées par une variété de facteurs, y compris la littératie financière et les biais potentiels dans les pratiques de prêt. De plus, il est crucial d'interpréter ces différences avec discernement pour éviter toute forme de discrimination ou de stigmatisation.

[1]: Financial literacy and gender difference in loan performance

[2]: Gender differences in the repayment of microcredit: The mediating role ...

(C) Analyse bivariée →

Classification des Variables

Nous avons classé nos variables en deux catégories : quantitatives et qualitatives.

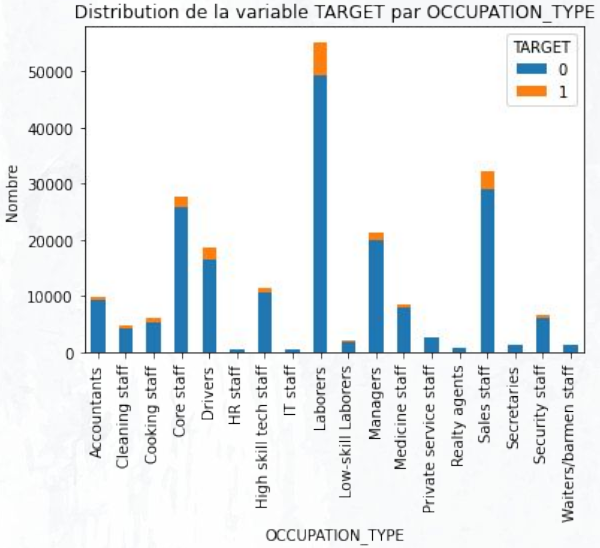
Variables Qualitatives	Variables Quantitatives
CODE_GENDER	AMT_INCOME_TOTAL
NAME_EDUCATION_TYPE	DAYS_BIRTH
NAME_FAMILY_STATUS	DAYS_EMPLOYED
NAME_HOUSING_TYPE	OWN_CAR_AGE
OCCUPATION_TYPE	AMT_CREDIT
ORGANIZATION_TYPE	AMT_ANNUITY
FLAG_OWN_CAR	
FLAG_OWN_REALTY	

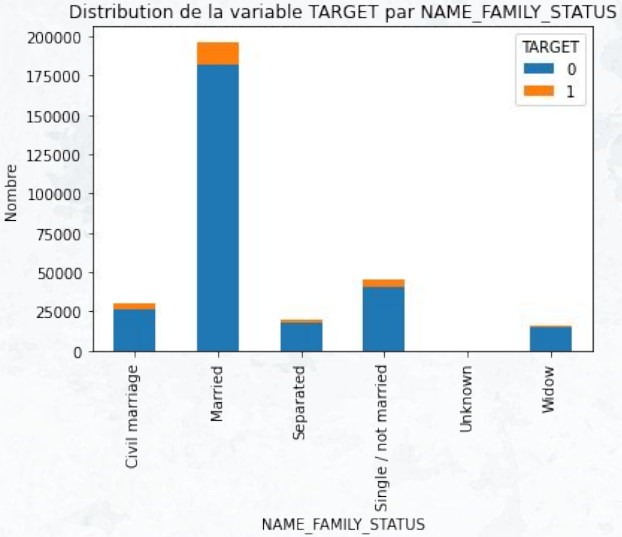
(C) Analyse bivariée

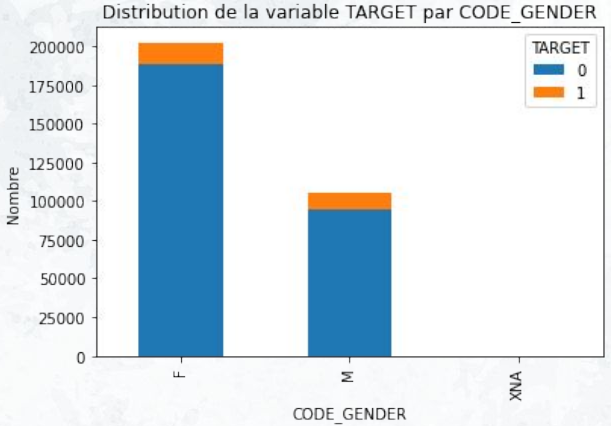
Analyse de la Variable Target

Nous avons utiliser des diagrammes pour visualiser la distribution de la variable target par rapport a chaque variable qualitative.

Exemple de visualisation







(C) Analyse bivariée → Conclusions analyse de la Variable Target

CODE_GENDER	Les hommes et les femmes ont des distributions similaires de TARGET, sans différence significative en fonction du sexe.
NAME_EDUCATION_TYPE	Les individus ayant une éducation supérieure semblent avoir une légèrement meilleure performance (moins de la catégorie orange de TARGET).
NAME_FAMILY_STATUS	Les célibataires ou non mariés ont une proportion légèrement plus élevée de défauts (catégorie orange) par rapport à ceux qui sont mariés.
NAME_HOUSING_TYPE	Les individus vivant dans un appartement loué sont légèrement plus susceptibles d'avoir des défauts
OCCUPATION_TYPE	Les travailleurs non qualifiés ont une proportion plus élevée de défauts par rapport à d'autres types d'occupations.
ORGANIZATION_TYPE	Certains types d'organisations, tels que "Transport: type 3", montrent une proportion plus élevée de défauts.
FLAG_OWN_CAR	Posséder une voiture n'a pas d'impact majeur sur la distribution de TARGET.
FLAG_OWN_REALTY	Posséder un bien immobilier n'a pas d'impact majeur sur la distribution de TARGET.
NAME_TYPE_SUITE	La majorité des demandes sont faites sans accompagnement. Il n'y a pas de tendance clairement identifiable entre les différentes catégories.
WEEKDAY_APPR_PROCESS_START	Le jour de début du processus de demande n'a pas d'impact significatif sur la distribution de TARGET.
HOUR_APPR_PROCESS_START	L'heure de début du processus de demande n'a pas d'impact majeur sur la distribution de TARGET.
REGION_RATING_CLIENT et REGION_RATING_CLIENT_W_CITY	Les clients ayant une note de région de 3 (avec ou sans prise en compte de la ville) sont plus susceptibles d'avoir des défauts par rapport à ceux avec des notes de 1 ou 2.
REG_CITY_NOT_LIVE_CITY	Les clients dont la ville de résidence est différente de leur ville d'enregistrement sont légèrement plus susceptibles d'avoir des défauts.

(C) Analyse bivariée

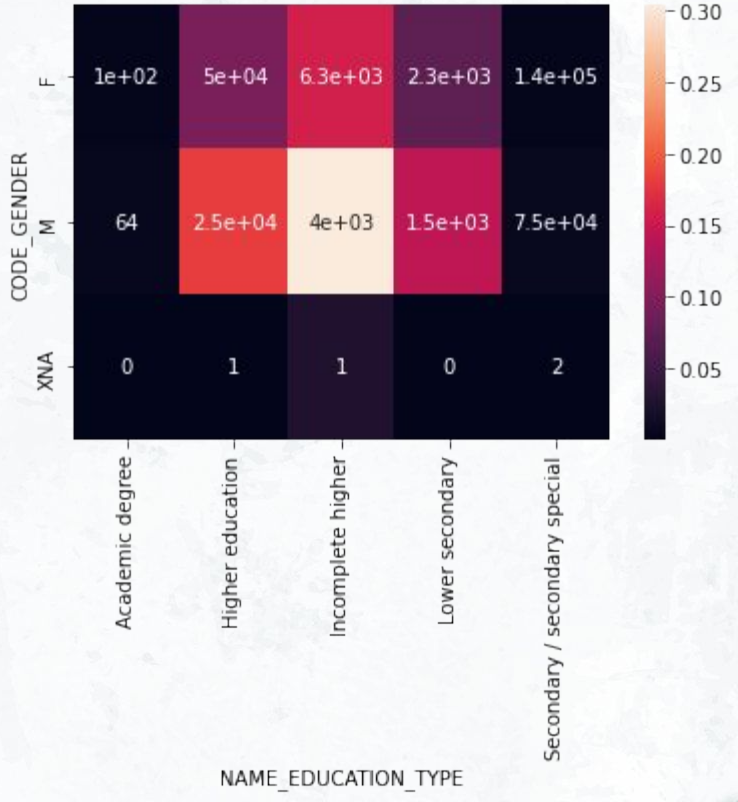
Analyse des Variables Qualitatives

Pour les variables qualitatives, nous avons utilisé des tableaux de contingences, puis nous avons analysé les corrélations avec des carte de chaleurs

Exemple de visualisation

'Tableau de contingence pour CODE_GENDER et NAME_EDUCATION_TYPE :'

	Secondary / secondary special	Higher education	Incomplete higher	Lower secondary	Academic degree
M	74924.0	24573.0	3993.0	1505.0	64.0
F	143465.0	50289.0	6283.0	2311.0	100.0
XNA	2.0	1.0	1.0	0.0	0.0



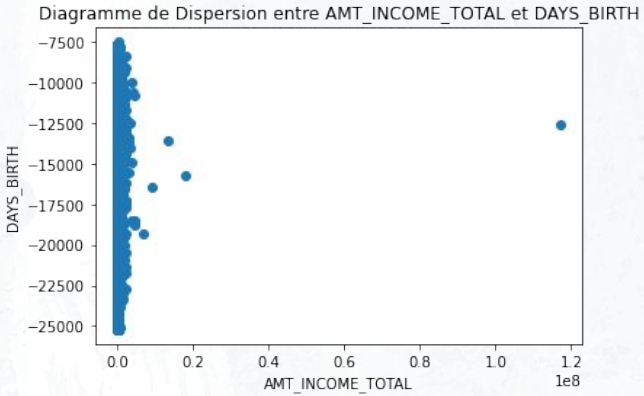
(C) Analyse bivariée

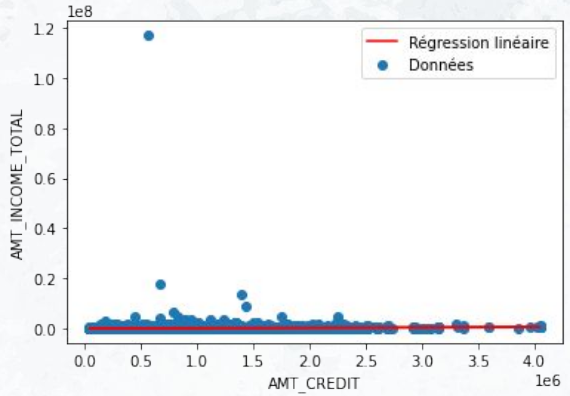
Analyse des Variables Quantitatives

Pour les variables quantitative, nous avons utilisé des diagrammes de dispersion, puis nous avons calculer le coefficient de Pearson et la covariance pour chaque paire de variables.

Enfin nous avons analysé les corrélation en utilisant la régression linéaire

Exemple de visualisation





	Pearson_with_DAYS_BIRTH	Covariance_with_DAYS_BIRTH	Pearson_with_DAY
AMT_INCOME_TOTAL	0.009099	5.069871e+06	
DAYS_BIRTH	0.000000	0.000000e+00	
DAYS_EMPLOYED	0.000000	0.000000e+00	
OWN_CAR_AGE	0.000000	0.000000e+00	
AMT_CREDIT	0.000000	0.000000e+00	

(D) Analyse multivariée ⟶

ANOVA - Analyse de la Variance

Nous avons examiné les interactions entre plusieurs variables pour mieux comprendre leur impact conjoint.

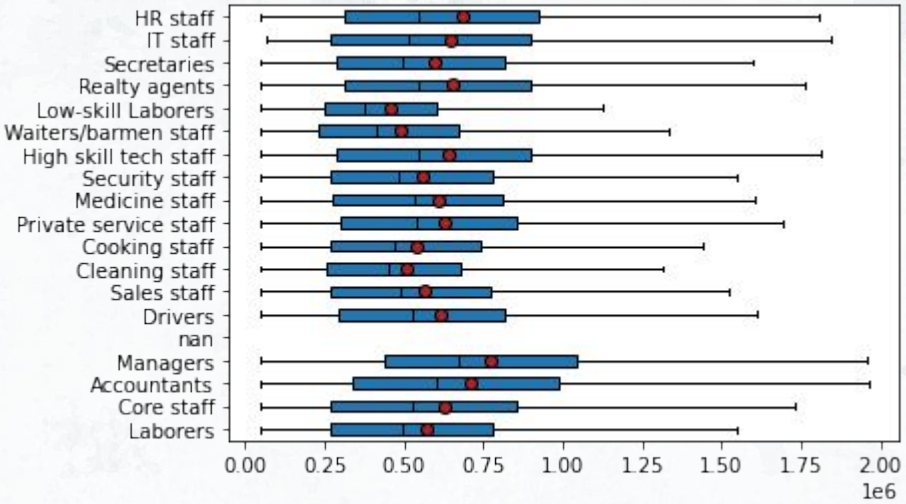
Nous avons commencé par une ANOVA pour évaluer l'impact de la variable qualitative 'OCCUPATION_TYPE' sur la variable quantitative 'AMT_CREDIT'. Seulement 2.94% de la variance du montant du crédit est expliquée par le type d'occupation, suggérant une association modeste entre ces variables.

La boîte à moustaches fournie montre la distribution du montant du crédit pour chaque modalité de OCCUPATION_TYPE.

Les médianes et moyennes de certains groupes diffèrent des autres, indiquant des variations dans les montants de crédit en fonction du type d'occupation.

Cependant, ces différences ne sont pas extrêmement marquées.

En somme, bien qu'il y ait des différences observables dans les montants de crédit selon le type d'occupation, cette variable n'explique qu'une petite partie de la variabilité totale du montant du crédit



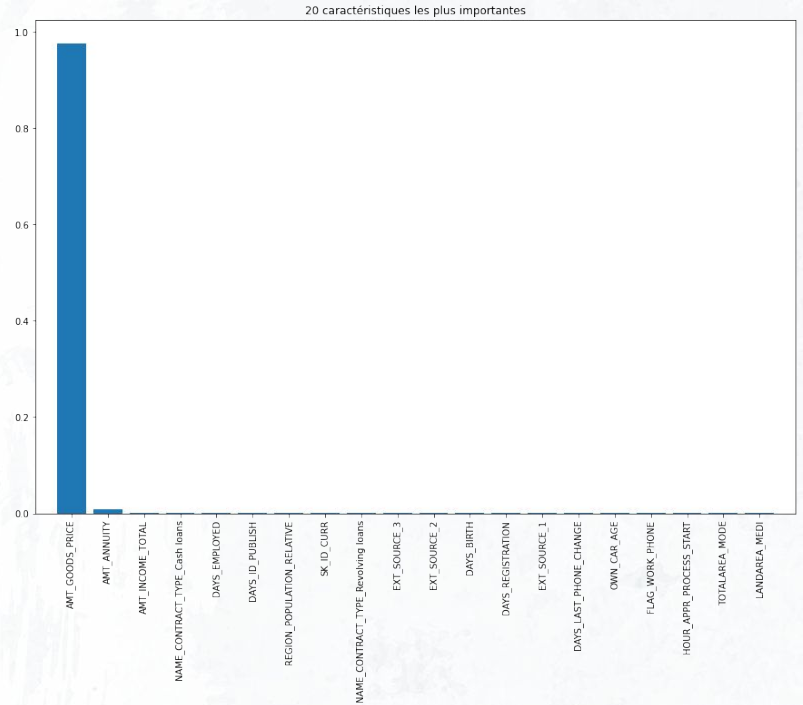
(D) Analyse multivariée ⟶

Forêts Aléatoires - Préparation des Données

Pour évaluer l'influence des variables sur 'AMT_CREDIT', nous avons utilisé un modèle de forêts aléatoires.

Les résultats montrent que 'AMT_GOODS_PRICE' est de loin la caractéristique la plus influente, suivie par 'AMT_ANNUITY' et 'AMT_INCOME_TOTAL'.

Cela indique que ces variables sont des prédicteurs clés pour le montant du crédit



	Importance
AMT_GOODS_PRICE	0.976351
AMT_ANNUITY	0.008996
AMT_INCOME_TOTAL	0.000762
NAME_CONTRACT_TYPE_Cash loans	0.000539
DAYS_EMPLOYED	0.000518
...	...
ORGANIZATION_TYPE_Industry: type 8	0.000000
FLAG_DOCUMENT_4	0.000000
ORGANIZATION_TYPE_Trade: type 4	0.000000
FLAG_DOCUMENT_21	0.000000
FLAG_DOCUMENT_7	0.000000

(E) Feature Engineering ⟶

Dans cette section, nous avons exploré le 'Feature Engineering', une étape cruciale pour améliorer la précision de notre modèle de prédiction du risque de crédit.

Nous avons introduit trois nouvelles caractéristiques basées sur nos données : CREDIT_TO_INCOME_RATIO, OUTSTANDING_DEBT, et RATE_OF_PREVIOUS_TIMELY_PAYMENTS.

Feature	Description	Source
CREDIT_TO_INCOME_RATIO	Ce ratio donne une indication de la taille du prêt par rapport au revenu de l'emprunteur. Un ratio élevé pourrait indiquer que l'emprunteur pourrait avoir des difficultés à gérer un prêt significatif par rapport à son revenu.	application_train.csv
OUTSTANDING_DEBT	Cette caractéristique représente la somme totale des montants de crédit actifs pour chaque client. Un montant élevé pourrait indiquer un risque accru si le client a déjà une dette significative.	bureau.csv, application_train.csv
RATE_OF_PREVIOUS_TIMELY_PAYMENTS	Il s'agit du ratio du nombre total de paiements effectués à temps au nombre total de paiements précédents. Un taux élevé pourrait indiquer un historique de remboursement solide, tandis qu'un taux faible pourrait suggérer un risque accru.	installments_payments.csv, previous_application.csv

(F) Features Selection ⟶

Pour optimiser la sélection des variables de notre modèle, nous avons adopté une approche hybride, combinant des critères métier spécifiques et des techniques de sélection automatique.

Cette stratégie visait à capitaliser sur les forces respectives de chaque méthode, garantissant ainsi une sélection de caractéristiques à la fois robuste et pertinente.

Dans un premier temps, nous avons intégré des connaissances métier pour identifier les variables potentiellement influentes.

Parallèlement, des méthodes automatiques, telles que l'analyse de la variance ou des algorithmes de sélection de caractéristiques, ont été employées pour évaluer l'importance des variables de manière objective.

Après avoir combiné ces deux approches, nous avons procédé à un examen minutieux des résultats.

Cette étape critique nous a permis d'affiner notre sélection, en éliminant les variables redondantes ou peu informatives, et en conservant celles qui apportaient une valeur ajoutée significative au modèle.

Suite à cette optimisation, le DataFrame application_train_cleaned a été mis à jour pour refléter notre sélection finale de variables.

Cette version épurée et améliorée du DataFrame a ensuite été exportée depuis ce notebook, la rendant prête pour une utilisation ultérieure dans le processus de modélisation, spécifiquement dans le notebook 01_Modelisation.

(F) Features Selection →

Sélection métier

Dans notre approche métier, nous avons sélectionné des caractéristiques clés pour évaluer la solvabilité des emprunteurs.

Ces caractéristiques sont choisies pour leur pertinence dans la prédiction du risque de crédit et la compréhension globale de la situation financière et personnelle de l'emprunteur.

CREDIT_TO_INCOME_RATIO	Ce ratio est crucial pour évaluer la capacité de l'emprunteur à gérer un prêt par rapport à son revenu. Un ratio élevé peut signaler un risque potentiel de surendettement.
AMT_CREDIT_SUM_DEBT (Dette en cours)	La dette totale actuelle de l'emprunteur nous donne une idée de ses engagements financiers existants, essentielle pour évaluer sa capacité à rembourser de nouveaux crédits.
PAID_ON_TIME (Taux de paiements ponctuels précédents)	Cette donnée reflète la fiabilité de l'emprunteur dans le remboursement de ses dettes passées, un indicateur clé de son comportement de paiement futur.
Fonctionnalités Financières	Des éléments tels que le revenu total, le montant du crédit, et le prix des biens financés sont fondamentaux pour comprendre la situation financière de l'emprunteur.
Fonctionnalités Démographiques	L'âge, le sexe, le niveau d'éducation, le statut familial, et le type de logement fournissent un contexte important sur la stabilité et le profil de vie de l'emprunteur.
Historique de l'Emploi	La durée de l'emploi, le type d'occupation, et le type d'organisation où l'emprunteur est employé nous aident à évaluer sa stabilité professionnelle.
Fonctionnalités de Propriété	La possession de biens tels qu'une voiture ou un bien immobilier peut influencer la capacité de l'emprunteur à gérer ses finances.
Évaluations Régionales	Les évaluations du client dans sa région et sa ville offrent une perspective sur le contexte économique et social de l'emprunteur.

(F) Features Selection →

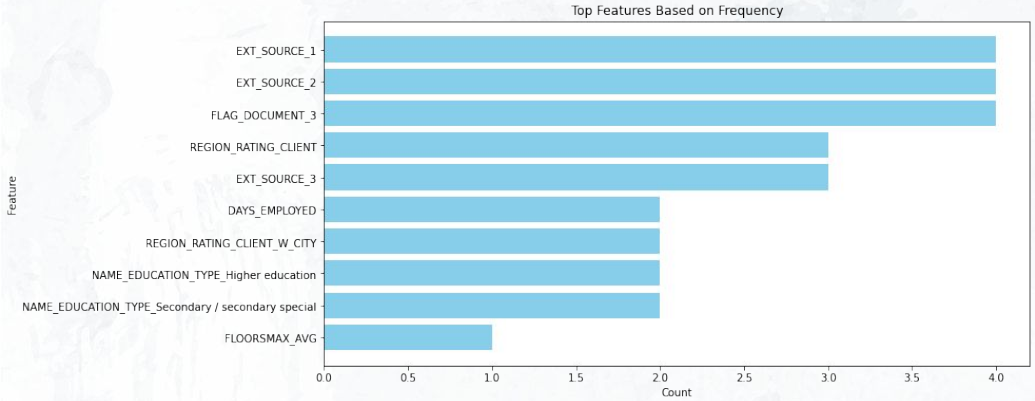
Approche par filtrage

Notre processus de sélection des caractéristiques a été rigoureux et diversifié, visant à identifier les variables les plus influentes pour notre modèle prédictif.

Nous avons d'abord écarté les variables présentant une faible variance, puis avons employé des méthodes statistiques avancées pour évaluer l'impact des caractéristiques restantes sur notre variable cible.

L'analyse de variance (ANOVA) et le test de Kruskal-Wallis ont été utilisés pour les variables continues, tandis que les tests du Chi-Deux, V de Cramér et T de Tschuprow ont examiné les associations entre les variables catégorielles et la variable cible.

Enfin, l'Information Value a été appliquée pour évaluer le pouvoir prédictif de chaque caractéristique, garantissant ainsi une sélection de variables à la fois pertinente et puissante pour notre modèle.



Notre sélection de caractéristiques a combiné pertinence statistique et considérations métier, aboutissant à un ensemble robuste et fiable pour notre modèle.

Nous avons d'abord identifié les variables les plus influentes par leur fréquence d'apparition, puis consolidé ces données dans un DataFrame dédié.

Cette approche a assuré un équilibre optimal entre les exigences analytiques et les connaissances spécifiques du domaine, garantissant ainsi la précision et la pertinence de notre modèle.

(F) Features Selection →

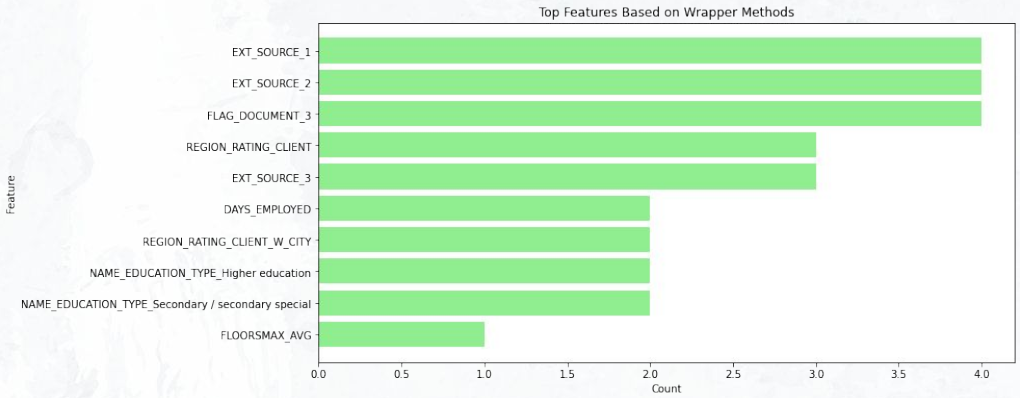
Approche Wrapper

Dans notre stratégie de sélection de caractéristiques de type wrapper, nous avons mis en œuvre des algorithmes spécifiques pour déterminer les ensembles de données les plus influents pour notre modèle prédictif.

Cette approche s'est appuyée directement sur la performance du modèle pour guider la sélection.

Nous avons adopté trois méthodes distinctes : la sélection progressive (Forward Selection), la sélection régressive (Backward Selection) et l'élimination récursive des caractéristiques (Recursive Feature Elimination).

Chacune de ces méthodes a été appliquée pour affiner notre ensemble de données, en se concentrant sur l'optimisation de la performance du modèle prédictif.



Chaque méthode employée dans notre stratégie wrapper a identifié un ensemble distinct de caractéristiques, mettant en lumière le rôle crucial de différentes variables dans la prédiction de notre variable cible.

Pour synthétiser ces résultats, nous avons procédé au décompte de la fréquence d'apparition de chaque caractéristique à travers les différentes méthodes.

Sur cette base, nous avons créé un dataframe regroupant les caractéristiques les plus fréquemment sélectionnées, fournissant ainsi un aperçu consolidé des variables les plus influentes pour notre modèle prédictif.

(F) Features Selection

Approche embedded

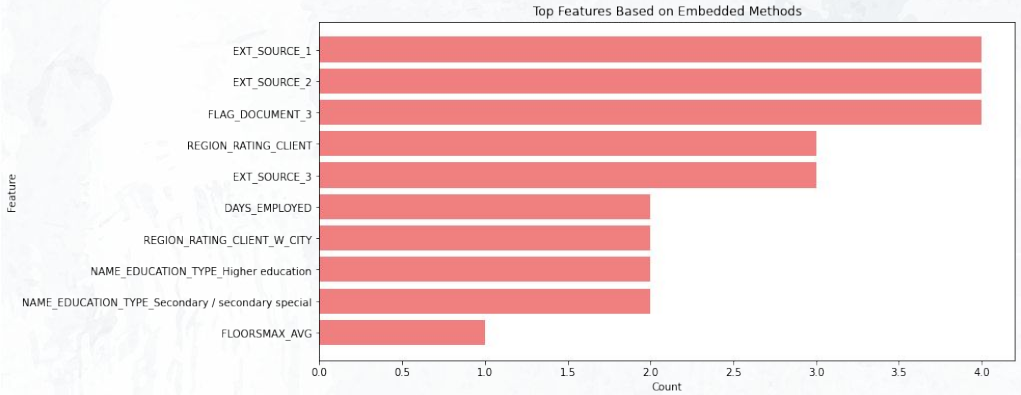
Dans le cadre de notre stratégie embedded, nous avons intégré la sélection des caractéristiques directement dans le processus de développement du modèle.

Cette approche s'appuie sur des algorithmes dotés de mécanismes internes de sélection de caractéristiques.

Nous avons d'abord mis en œuvre la régression LASSO, qui applique une pénalité aux coefficients de régression.

Cette pénalité conduit à réduire certains coefficients à zéro, éliminant ainsi ces caractéristiques du modèle. Ensuite, nous avons employé un modèle de forêt aléatoire (Random Forest) pour évaluer l'importance des caractéristiques.

Dans ce contexte, les caractéristiques jugées les plus importantes sont celles qui contribuent le plus significativement à la prédiction de notre variable cible.



Chaque méthode de sélection a identifié un ensemble distinct de caractéristiques, mettant en lumière l'importance de différentes variables dans la prédiction de notre variable cible.

Nous avons ensuite procédé au décompte de la fréquence d'apparition de chaque caractéristique.

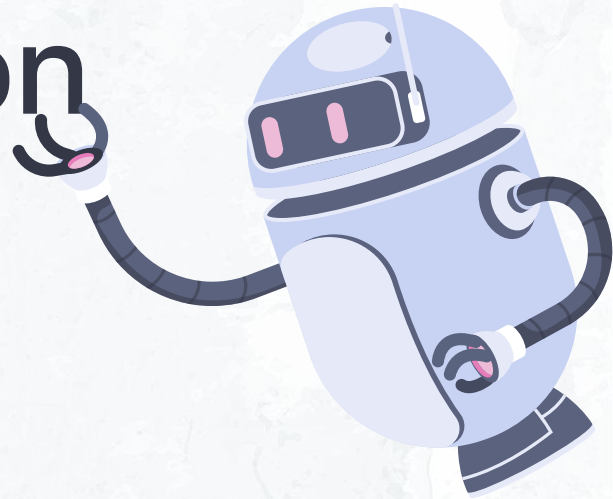
Sur cette base, nous avons créé un dataframe regroupant les caractéristiques les plus fréquemment sélectionnées, offrant ainsi une vue consolidée des variables les plus pertinentes pour notre modèle.

Pour finaliser notre analyse exploratoire des données, nous avons examiné les corrélations entre les variables et la variable cible.

Les variables présentant une corrélation excessive ont été exclues pour éviter les problèmes de multicollinéarité.

Ensuite, nous avons créé des dataframes distincts pour chaque approche adoptée, assurant ainsi une organisation claire et une analyse ciblée des variables retenues.

04 → Modélisation



Sous - sections

A → Paramétrage

B → Préparation des données

C → Approche naïve

D → Évaluation des Performances des Modèles sur Différents Ensembles de Données

E → Évaluation et optimisation du Modèle de régression logistique sur le jeu de données 'filtered'

F → Feature importance

(A) Paramétrage →

Vérification de data-leakage

Nous avons développé une fonction spécifique pour identifier toute fuite potentielle de données dans nos différents ensembles de données sélectionnés.

La fonction parcourt chaque ensemble de données et examine les corrélations entre les caractéristiques numériques et la variable cible.

Cette étape de vérification est cruciale pour maintenir la qualité et la fiabilité de notre modèle de prédiction. Elle nous assure que les caractéristiques sélectionnées sont appropriées pour une modélisation sans biais

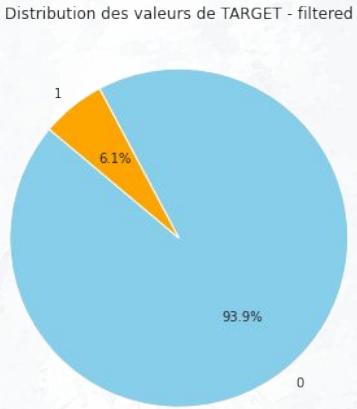
	index	Data Leakage
0	filtered	False
1	wrapper	False
2	embedded	False
3	business	False

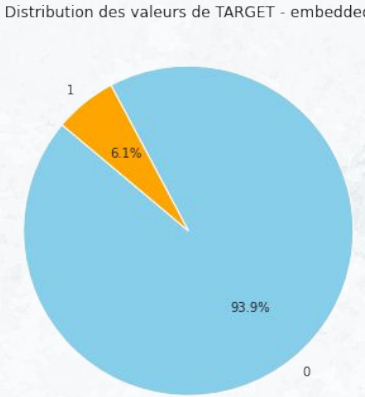
(A) Paramétrage →

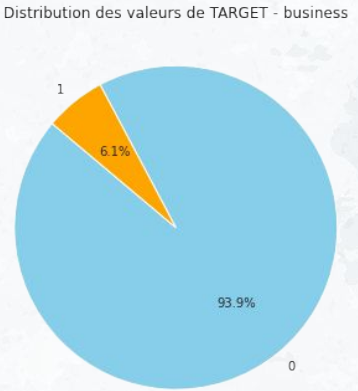
Visualisation de la Distribution de la Variable Cible

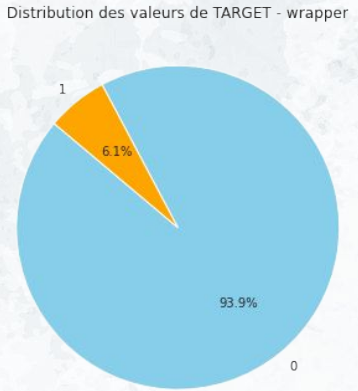
Nous avons mis en place une série de visualisations pour examiner la distribution de notre variable cible, 'TARGET', dans différents ensembles de données.

Ces ensembles incluent ceux obtenus par les méthodes de sélection de caractéristiques filtrage, wrapper, embedded, et basées sur des critères métier









(B) Préparation des données ⟶

Étapes Clés de la Préparation des Données

Dans notre projet, nous avons suivi une série d'étapes méthodiques pour préparer nos ensembles de données.

Nous avons commencé par appliquer l'encodage One-Hot aux variables catégorielles de nos quatre ensembles de données (filtered, wrapper, embedded, business)

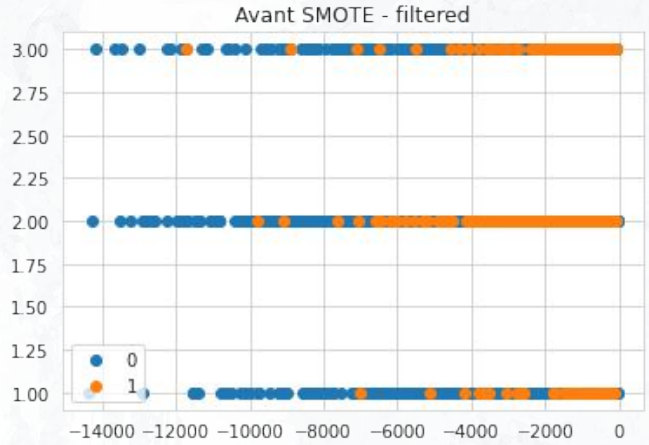
Cependant, il s'est avéré que nos ensembles de données ne contenaient pas de variables catégorielles nécessitant cet encodage.

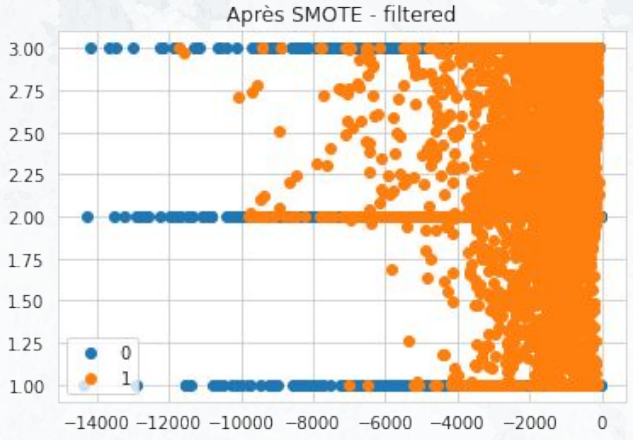
Chaque ensemble de données a été divisé en sous-ensembles d'entraînement et de test, en utilisant une répartition de 80% pour l'entraînement et 20% pour le test.

Nous avons identifié un déséquilibre dans la distribution des classes de notre variable cible.

Pour y remédier, nous avons utilisé la technique SMOTE (Synthetic Minority Over-sampling Technique), qui a permis de rééquilibrer les classes dans chaque ensemble de données.

Enfin, nous avons normalisé les caractéristiques de chaque ensemble de données en utilisant le StandardScaler.





	dataset	mean_first_feature	std_first_feature	mean_second_feature	std_second_feature
0	filtered	1.575073e-15	1.0	9.644023e-15	1.0
1	wrapper	1.575073e-15	1.0	9.644023e-15	1.0
2	embedded	1.575073e-15	1.0	9.644023e-15	1.0
3	business	-3.811193e-15	1.0	2.463801e-16	1.0

(C) Approche naïve ⟶

Nous avons débuté notre analyse de modélisation en évaluant une approche naïve avec le DummyClassifier.

Cette étape sert de point de référence pour mesurer l'efficacité des modèles plus sophistiqués.

Le DummyClassifier a été appliqué sur nos quatre ensembles de données (filtered, wrapper, embedded, business)

Les résultats mettent en lumière l'inaptitude du DummyClassifier à fournir une classification utile.

Il prédit systématiquement la classe majoritaire, échouant ainsi à capturer la complexité et les nuances des données.

Ces métriques servent de base pour évaluer les performances des modèles plus avancés.

Tout modèle significatif doit surpasser ces résultats pour être considéré comme efficace.

	dataset	model	accuracy	precision	recall	f1_score	roc_auc	business_cost	roc_curve	ela
0	filtered	Dummy	0.938989	0.0	0.0	0.0	0.5	1050	([0.0, 1.0], [0.0, 1.0])	
1	wrapper	Dummy	0.938989	0.0	0.0	0.0	0.5	1050	([0.0, 1.0], [0.0, 1.0])	
2	embedded	Dummy	0.938989	0.0	0.0	0.0	0.5	1050	([0.0, 1.0], [0.0, 1.0])	
3	business	Dummy	0.938989	0.0	0.0	0.0	0.5	1050	([0.0, 1.0], [0.0, 1.0])	

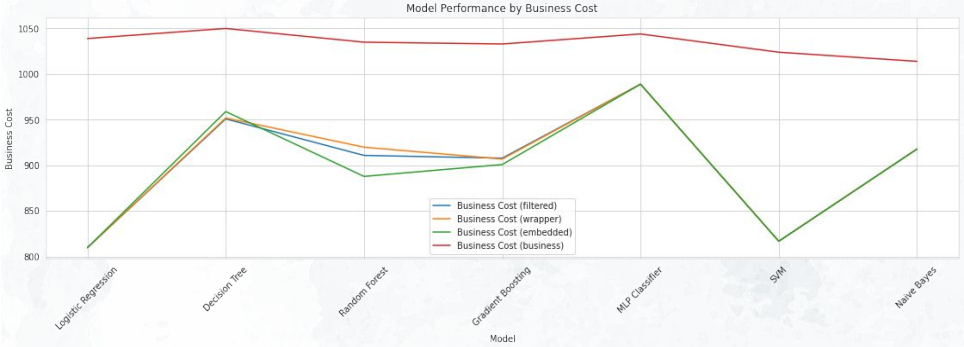
(D) Évaluation des Performances des Modèles sur Différents Ensembles de Données

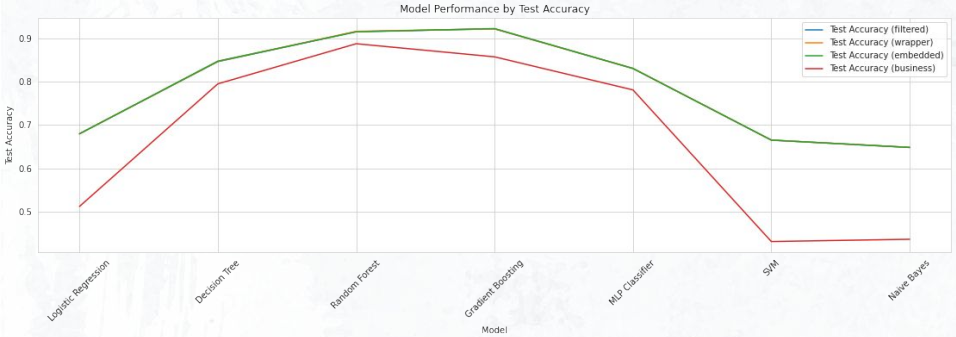
Nous avons évalué plusieurs modèles de machine learning sur quatre ensembles de données distincts - "filtered", "wrapper", "embedded" et "business" - pour identifier le modèle le plus performant pour notre cas d'usage spécifique.

Les modèles ont montré des performances variables sur les différents ensembles de données.

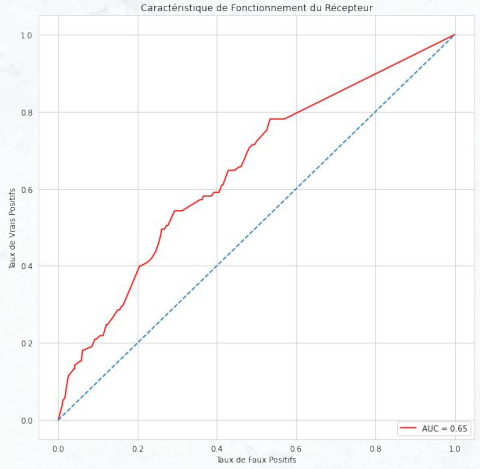
La Régression Logistique sur l'ensemble de données "filtered" s'est avérée être le choix optimal, offrant un équilibre entre précision, coût métier, et les autres métriques.

Cette combinaison offre une performance robuste tout en maintenant la simplicité et la clarté, facilitant l'interprétation et la gestion du modèle dans des scénarios pratiques.





	model	dataset	train_accuracy	test_accuracy	precision	recall	f1_score	roc_auc	t
0	Logistic Regression	filtered	0.706269	0.679837	0.112847	0.619048	0.190896	0.736315	
1	Decision Tree	filtered	0.912616	0.846020	0.111650	0.219048	0.147910	0.647186	
2	Random Forest	filtered	1.000000	0.915166	0.098039	0.047619	0.064103	0.678188	
3	Gradient Boosting	filtered	0.971749	0.921557	0.142857	0.057143	0.081633	0.705169	



(E) Évaluation et optimisation du Modèle de régression logistique sur le jeu de données 'filtered

Le modèle final avec validation croisée, a été ajusté avec des paramètres spécifiquement choisis pour optimiser la performance.

Ces paramètres incluaient une gamme de valeurs pour la force de régularisation (Cs), le nombre de plis pour la validation croisée (cv), l'algorithme d'optimisation (solver), et le nombre maximal d'itérations pour la convergence (max_iter)

Les résultats obtenus sont prometteurs.

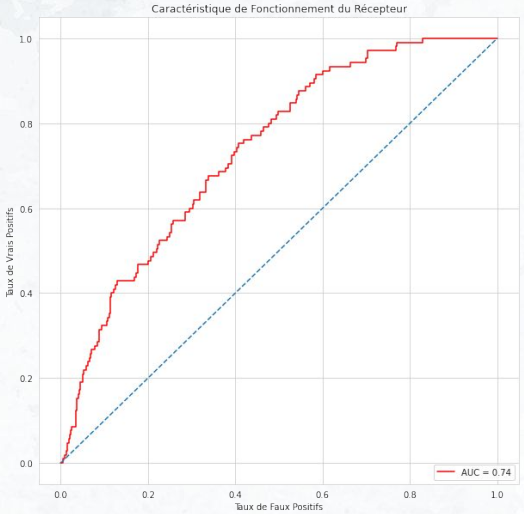
Le modèle a démontré une précision de 70.63% sur l'ensemble d'entraînement et de 67.98% sur l'ensemble de test, indiquant une bonne généralisation.

De plus, des scores de précision, de rappel, de F1 et de ROC-AUC respectivement de 11.28%, 61.90%, 19.09% et 73.63% soulignent la capacité du modèle à distinguer efficacement entre les classes.

Le score métier calculé, s'est élevé à 810, fournissant une mesure supplémentaire de l'efficacité du modèle.

Le modèle de régression logistique optimisé représente une avancée significative dans notre objectif de prédiction précise du risque de crédit. Il allie efficacité et robustesse, tout en offrant une base solide pour des analyses plus poussées ou des améliorations futures.

	model	dataset	train_accuracy	test_accuracy	precision	recall	f1_score	roc_auc
0	LogisticRegression	filtered	0.706269	0.679837	0.112847	0.619048	0.190896	0.736315

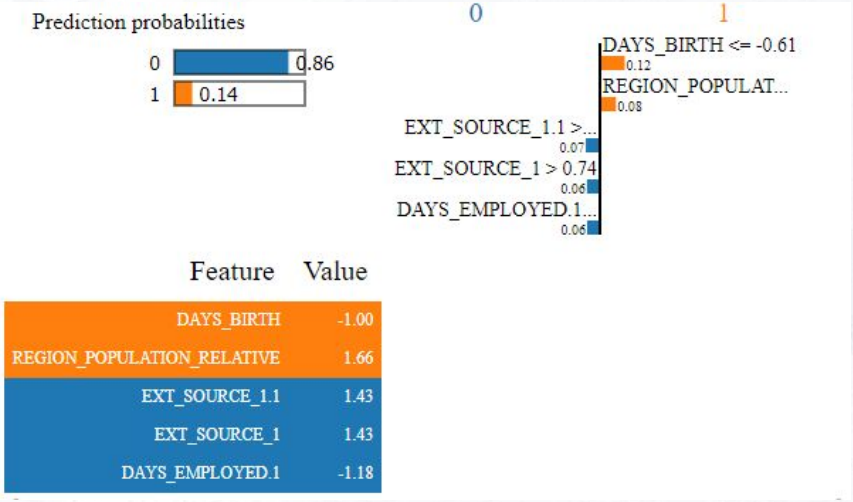


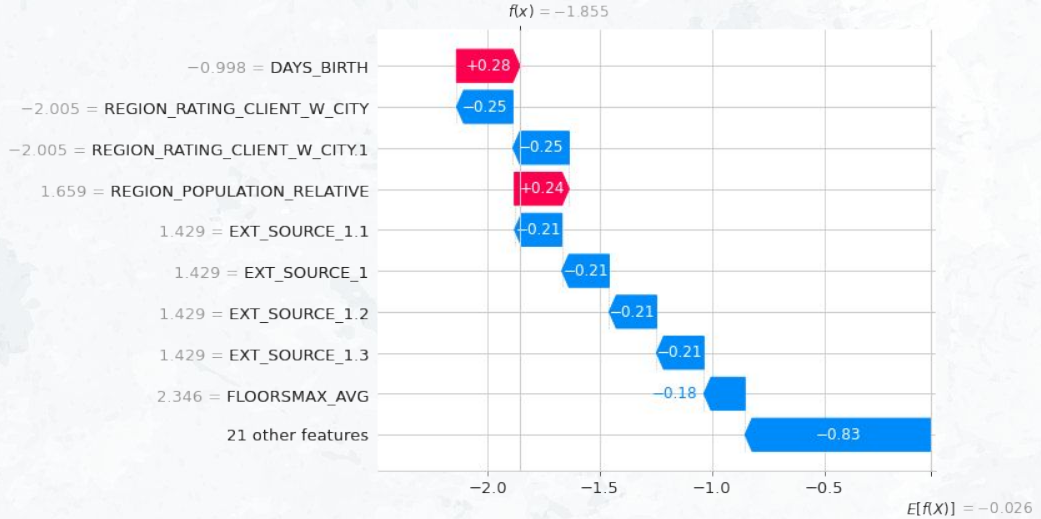
(F) Feature importance →

Interprétabilité Locale

Une attention particulière a été portée à l'interprétabilité locale des prédictions, un aspect fondamental pour élucider les mécanismes de décision spécifiques de notre modèle.

À cet effet, nous avons intégré deux méthodologies principales : LIME et SHAP





Les analyses via LIME et SHAP révèlent l'importance cruciale de l'âge du client, des facteurs démographiques régionaux et des scores de sources externes dans les prédictions du modèle.

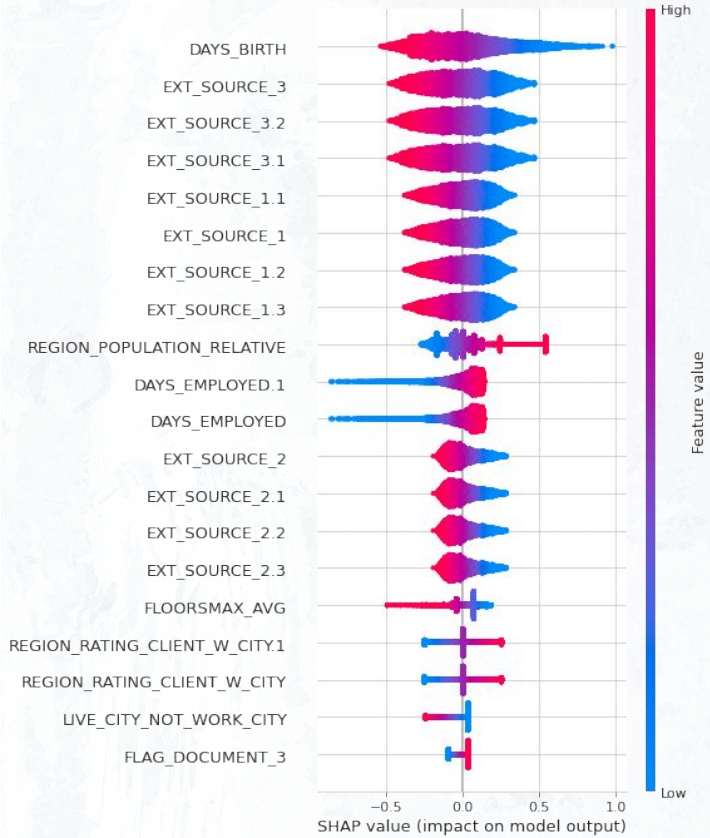
Malgré de légères différences, ces méthodes s'accordent sur le rôle majeur de ces caractéristiques.

LIME génère un modèle local interprétable autour d'une prédiction spécifique pour évaluer l'importance des caractéristiques.

En contraste, SHAP utilise la théorie des jeux pour attribuer des valeurs de Shapley à chaque caractéristique, reflétant leur contribution moyenne à la prédiction sur toutes les combinaisons possibles de caractéristiques.

(F) Feature importance →

Interprétabilité Global



Cette visualisation révèle que l'âge du sujet, représenté par la variable DAYS_BIRTH, est le prédicteur le plus significatif, exerçant une influence prédominante sur l'issue prédite par le modèle.

Les sources externes, notamment EXT_SOURCE_3 et ses dérivées, montrent une forte corrélation négative, indiquant que des valeurs élevées de ces caractéristiques tendent à diminuer la probabilité de l'événement étudié.

La variabilité des impacts, indiquée par la dispersion des points, suggère une hétérogénéité des effets des variables sur les prédictions.

L'alignement des résultats entre la feature importance local et global attestent de la robustesse de ces variables comme déterminants prédictifs majeurs dans notre modèle de régression logistique.

Merci →

Avez-vous des questions ?

Site Web