

REPUBLIQUE DE COTE D'IVOIRE



Union-Discipline-Travail

MINISTRE DE L'ECONOMIE, DU PLAN ET DU DEVELOPPEMENT



**ECOLE NATIONALE SUPERIEURE DE STATISTIQUE ET
D'ECONOMIE APPLIQUEE D'ABIDJAN**



BRIDGE BANK
GROUP CÔTE D'IVOIRE

BRIDGE BANK GROUP COTE D'IVOIRE

**APPLICATION DES METHODES DE MACHINE LEARNING
AU CREDIT SCORING BANCAIRE : CAS DES CLIENTS
PARTICULIERS DE LA BRIDGE BANK GROUP COTE
D'IVOIRE.**

Réalisé par :

ASSIMANY Kermah Moise
Elève Ingénieur Statisticien
Economiste

Sous la direction de :

M. KONE Brahima
Chef de service analyse
crédit de la Bridge Bank
Group CI

Année académique 2022-2023

DECHARGE

**L'ENSEA et la BRIDGE BANK GROUP COTE D'IVOIRE
n'entendent donner aucune approbation ni improbation aux
opinions émises dans ce mémoire. Ces opinions doivent être
considérées comme propres à l'auteur.**

AVANT-PROPOS

L'Ecole Nationale Supérieure de Statistique et d'Economie Appliquée d'Abidjan (ENSEA), centre d'excellence sous régional de l'UEMOA (2005), et centre d'excellence africaine de la Banque Mondiale (2015) a pour vocation de fournir une formation de haut niveau en statistique et économie appliquée aux cadres de plusieurs pays d'Afrique francophone, anglophone et lusophone. La formation à l'ENSEA est subdivisée en deux divisions dont celle des Ingénieurs Statisticiens Economistes. Ceux-ci reçoivent une formation théorique d'une part et une formation pratique d'autre part. Le passage à l'aspect pratique se fait à travers des travaux de recherche, une enquête académique sur le terrain, un stage d'immersion et un stage de fin de cycle. Le présent mémoire s'inscrit ainsi dans le cadre du stage de fin de cycle débuté depuis Mai 2023 au sein de la BRIDGE BANK GROUP CI.

Après un bon mois d'observation et d'acclimatation à l'univers de la banque, nous avons soumissionné l'idée de réaliser notre mémoire sur le thème : APPLICATION DES METHODES DE MACHINE LEARNING AU CREDIT SCORING BANCAIRE : CAS DES CLIENTS PARTICULIERS DE LA BRIDGE BANK GROUP COTE D'IVOIRE.

L'entreprise dans sa politique d'être à la pointe des techniques d'évaluation de risque de crédit, a toute suite acquiescé notre proposition et mis à notre disposition les conditions nécessaires à la réalisation de cette étude.

Par ailleurs, nous souhaitons souligner que la rédaction de ce document ne s'est pas faite sans difficulté, mais notre satisfaction a été d'être parvenue à surmonter ces difficultés et à arriver au sommet de notre espérance. Cependant, nous ne prétendons pas avoir bâti un travail parfait, et restons de ce fait ouvert aux remarques, aux critiques et aux suggestions.

Bonne lecture !

DEDICACES

A DIEU, maître suprême de l'univers ;

A mon Père pour ses sacrifices ;

A ma mère pour son amour et ses prières ;

*A mes petit-frères, pour la motivation qu'ils ne cessent de raviver
involontairement en moi ;*

A ma famille et mes amis proches ;

*A toutes les personnes qui de près ou de loin ont contribué au succès de mon
cursus académique.*

REMERCIEMENTS

Tout d'abord, nous souhaitons exprimer nos plus sincères remerciements et notre profonde gratitude à Dieu, source de toute connaissance et de toute inspiration, pour nous avoir accompagnés tout au long de cette entreprise intellectuelle.

Ensuite, nous tenons à exprimer notre reconnaissance à plusieurs personnes qui ont joué un rôle essentiel et direct dans la réalisation de ce mémoire. Il s'agit en premier lieu du Directeur de l'ENSEA, Dr KOUADIO Hugues à qui nous adressons nos remerciements les plus chaleureux. Son leadership éclairé a permis de créer un environnement propice à l'apprentissage et à la recherche. Nous sommes également reconnaissants envers le directeur des études, Dr COULIBALY Romaric pour ses conseils avisés et son suivi attentif tout au long de notre cursus. C'est grande partie grâce à lui que nous avons pu effectuer ce stage au sein de la BBG CI et que nous avons pu réaliser cette soutenance malgré toutes les contraintes qui se posaient.

Nos sincères remerciements vont bien évidemment à l'endroit de tout le personnel de la BBG CI, pour leur chaleureux accueil et leur confiance manifestée tout au long du stage. Nous tenons à remercier particulièrement M. Cédric BENDEY-DIBI, pour son leadership et sa confiance placée en ce projet ; M. Brahim KONE, pour son suivi rigoureux et son partage inconditionnel des connaissances du métier ; sans oublier bien sûr M. Siméon KOUADIO et M. Jacques Alexandre OUEGNIN, qui nous ont soutenu aussi bien professionnellement qu'humainement.

Nous ne saurions clore notre propos sans toutefois remercier famille, amis et surtout collègues Young Pro pour leur soutien indéfectible tout au long de la période de stage. Leurs conseils, leurs discussions stimulantes et leur soutien moral ont été d'une valeur inestimable.

SOMMAIRE

DECHARGE	ii
AVANT-PROPOS	iii
DEDICACES	iv
REMERCIEMENTS	v
SOMMAIRE	vi
SIGLES ET ABREVIATIONS.....	1
LISTE DES GRAPHIQUES	2
LISTE DES TABLEAUX	4
RESUME.....	5
ABSTARCT	5
INTRODUCTION GENERALE.....	8
CHAPITRE I : DÉFINITION DES CONCEPTS ET REVUE DE LA LITTÉRATURE	13
CHAPITRE II : PRESENTATION ET ANALYSE DE LA BASE DE DONNEES	38
CHAPITRE III : MÉTHODOLOGIE ET MODELISATION.....	54
CHAPITRE IV : PRÉSENTATION DES RÉSULTATS ET INTERPRÉTATIONS.....	63
CHAPITRE V : DISCUSSION DES RESULTATS, RECOMMANDATIONS ET PROPOSITION D'UNE APPLICATION WEB	70
CONCLUSION GENERALE	75
REFERENCES BIBLIOGRAPHIQUES	78
ANNEXES	82
TABLE DES MATIERES	95

SIGLES ET ABREVIATIONS

BBG CI: Bridge Bank Group Côte d'Ivoire;

BCEAO : Banque Centrale des Etats de l'Afrique de l'Ouest ;

BIC : Bureau d'Information sur le Crédit ;

CCT: Crédit Court Terme ;

CMT : Crédit Moyen Terme ;

DCE : Direction Crédit et Engagement ;

ENSEA : Ecole Nationale Supérieure de Statistique et d'Economie Appliquée ;

PME : Petites et Moyennes Entreprises ;

UEMOA : Union Economique et Monétaire Ouest-Africaine.

LISTE DES GRAPHIQUES

Figure 1 : Organigramme de la BBG CI (source : Direction des Ressources Humaines de la BBG CI)	6
Figure 2 : : Schéma théorique de l'apprentissage supervisé (source : Azencott, 2018).....	22
Figure 3 : Processus du crédit scoring (source : Yang, 2001)	33
Figure 4 : Tableau récapitulatif des variables de la base de données à l'étude	39
Figure 5 : Répartition de la population étudiée en fonction du défaut de paiement.....	44
<i>Figure 6 : Répartition de la population étudiée en fonction du sexe</i>	<i>45</i>
<i>Figure 7 : Répartition de la population étudiée en fonction du statut matrimonial</i>	<i>45</i>
Figure 8 : Répartition de la population étudiée en fonction du secteur d'activité.....	46
Figure 9 : Répartition de la population étudiée en fonction du top 10 des professions les mieux représentées	46
Figure 10 : Répartition de la population étudiée en fonction des types de prêt.....	47
Figure 11 : Nuage de mots des garanties des prêts en défaut de paiement	48
Figure 12 : Nuage de mots des garanties des prêts sains	48
Figure 13 : Courbe d'évolution annuelle des prêts octroyés aux particuliers de 2008 à 2021.	50
Figure 14 : Nombre de prêts octroyés aux particuliers repartis par mois de 2006 à 2022.	50
Figure 15 : Moyenne d'âge par rapport au défaut de paiement	
Figure 16 : Durée moyenne de remboursement par rapport au défaut de paiement.	51
Figure 17 : Moyenne des montant des prêts par rapport au défaut de paiement.	
Figure 18 : Moyenne des taux d'intérêt par rapport au défaut de paiement.....	51
Figure 19 : Répartition de la population étudiée en fonction du secteur d'activité et du statut de défaut de paiement.....	52

Figure 20 : Répartition de la population étudiée en fonction de la profession et du statut de défaut de paiement	52
--	----

Figure 21 : Guideline pour un projet de crédit scoring basé sur les techniques de machine learning (source : Dastile et al., 2020)	55
--	----

LISTE DES TABLEAUX

Tableau 1 : Actionnariat de la BBG CI.....	6
Tableau 2 : Statistiques descriptives des variables quantitatives de la base de données.....	47
Tableau 3 : Tableau récapitulatif des tests de kendall entre les variables quantitatives.....	53
Tableau 4 : Matrice de confusion appliquée à notre étude (source : nos travaux)	60
Tableau 5 : Scores des modèles de base	60
Tableau 6 : Résultat des prédictions (source : nos travaux, voir annexe 2)	64
Tableau 7 : tableau de recherche du seuil optimal (source : nos travaux).....	66

RESUME

L'objectif principal de ce mémoire est de développer un modèle de credit scoring pour les prêts accordés aux particuliers par la BBG CI en utilisant des techniques de machine learning. Cette étude repose sur une base de données portant sur 711 emprunteurs et les données financières afférents à leur prêt. En suivant la méthodologie proposée par Dastile (2020), diverses techniques de prétraitement des données ont été appliquées, y compris des approches originales telles que l'ingénierie des variables à partir de l'analyse textuelle sur une variable représentant les garanties. Les résultats de cette étude mettent en évidence la supériorité des modèles de type bagging et boosting par rapport aux autres modèles. De plus, en utilisant des outils d'interprétabilité des modèles comme le Shapley Summary Plot, cette recherche démontre que les caractéristiques démographiques des clients, le montant des prêts et leur durée ont un effet sur le risque de crédit. In fine, l'étude se conclut par la mise en place d'une application web de credit scoring déployée à l'aide de la librairie streamlit.

Mots clés : Credit scoring, Machine learning, Risque de crédit, Défaut de paiement.

ABSTARCT

The main objective of this thesis is to develop a credit scoring model for individual loans granted by BBG CI using machine learning techniques. This study is based on a database about 711 borrowers and the financial data related to their loans. Following the methodology proposed by Dastile (2020), various data preprocessing techniques were applied, including innovative approaches such as variable engineering through text analysis of a variable representing collateral. The results of this study highlight the superiority of bagging and boosting models over other models. Furthermore, using model interpretability tools such as the Shapley Summary Plot, this research demonstrates that customers' demographic characteristics, loan amounts, and durations have an impact on credit risk. Ultimately, the study concludes with the development of a web application for credit scoring, deployed using the Streamlit library.

Keywords: Credit scoring, Machine learning, Credit risk, Default.

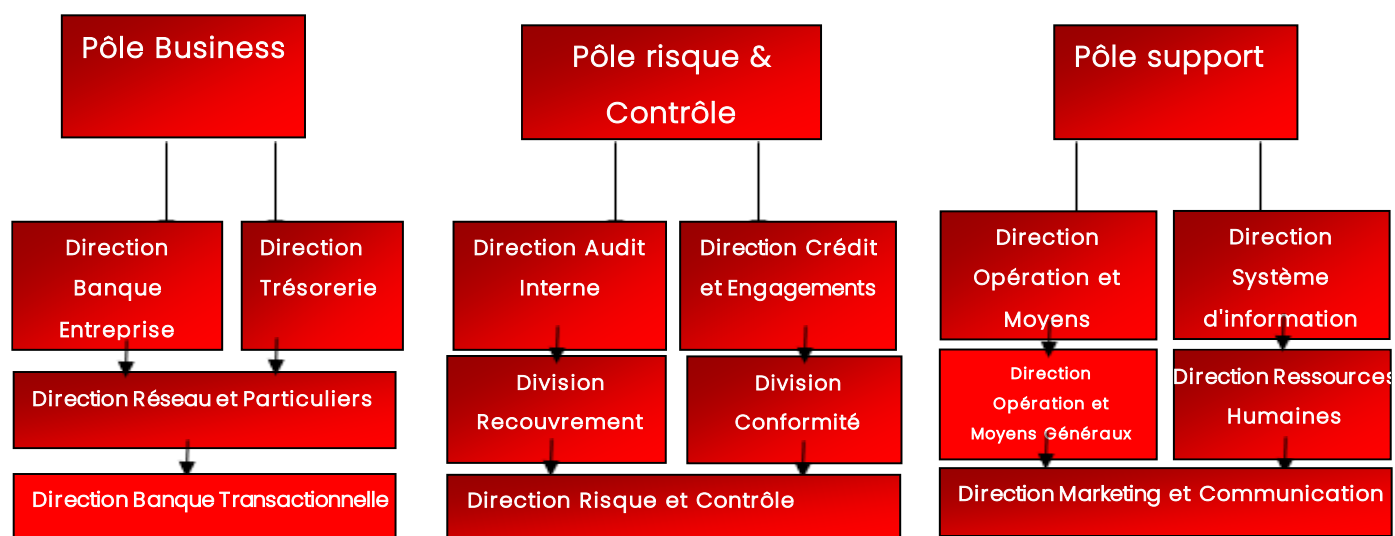
PRESENTATION DE LA STRUCTURE D'ACCUEIL

La Bridge Bank Group CI (BBG CI) est une banque commerciale ivoirienne ayant son siège à Abidjan Plateau. C'est une société anonyme avec un capital de 12,5 milliards de FCFA détenu majoritairement par le groupe Teyliom international à travers sa holding financière, Bridge Group West Africa Teyliom international. Au lancement de ses activités en juin 2006, la BBG CI s'est orientée vers le segment des PME fortement demandeur de solutions innovantes pour progressivement offrir ses services aux institutionnels et aux Particuliers moyen et haut de gamme. Après plus d'une décennie d'existence, la banque a progressivement pris sa place dans le paysage bancaire ivoirien au sein duquel elle représente désormais un acteur significatif. Dans un contexte d'amélioration du climat des affaires en Côte d'Ivoire, avec une plus grande attractivité du pays pour les partenaires étrangers, la banque s'appuie sur un réseau de correspondants extérieurs qui permet d'accompagner ses clients dans leurs opérations financières internationales. Toutes ces politiques mises en place proviennent d'un management de ladite Banque dont l'actionnariat se présente comme suit :

Tableau 1 : Actionnariat de la BBG CI

Actionnaires	Parts
Bridge Group West Africa	77,18%
CNPS	20%
Personnes Physiques	2,82%

Figure 1 : Organigramme de la BBG CI (source : Direction des Ressources Humaines de la BBG CI)



Au cours de ses 4 derniers mois, nous avons eu à effectuer notre stage au sein de la Direction Crédits et Engagements (DCE), plus précisément au sein du service analyse crédits. L'objectif était dans un premier temps de très vite nous initier au métier de la banque afin d'en comprendre clairement les rouages et dans un second temps d'apporter des solutions d'automatisation afin de faciliter et d'améliorer le traitement des dossiers de crédit.

Très vite, nous avons été placés sur l'analyse des dossiers de crédit des Très Petites Entreprises et des particuliers, tâche qui nous a permis de nous plonger dans les réalités de ce métier.

Pendant ces moments d'apprentissage, nous nous sommes également rendu compte de l'absence de processus de scoring interne à l'entreprise et applicable à tous les types de crédit. C'est donc en accord avec le chef de service, que nous nous sommes lancés comme objectif de proposer un modèle de scoring tenant compte de façon continue de l'historique et de la typologie des crédits accordés par la banque.

INTRODUCTION GENERALE

I. CONTEXTE ET JUSTIFICATION

Plus qu'un simple instrument financier, le crédit bancaire se positionne comme un levier fondamental, non seulement pour stimuler l'économie, mais surtout pour assurer la viabilité et la rentabilité des banques. Il constitue en effet une part prépondérante des activités bancaires, représentant environ 58,4 % du total bilan des banques de la zone UMOA.¹ Cette importance stratégique du crédit pour les banques se reflète dans l'évolution continue des volumes de crédits accordés dans la sous-région. Selon des données récentes de la BCEAO², le volume de crédit dans la sous-région a connu une croissance continue depuis plusieurs années et est estimé à 19 539,1 milliards de FCFA à fin 2022. Cette expansion s'est également accompagnée ces dernières années d'une amélioration notable dans la gestion des risques de crédit, avec une réduction du taux de dégradation des portefeuilles de 10,5% à 8,4% sur la même période. Toutefois, cette embellie cache des disparités et des défis persistants, en particulier en ce qui concerne l'aléa moral³ et le risque de crédit associé aux emprunteurs individuels.

En effet, la gestion de ces risques reste l'une des plus délicates, notamment parce que les informations financières relatives à ces types d'emprunteurs peuvent s'avérer insuffisantes voire inexistantes pour une évaluation précise du risque. En dépit des avancées en matière de systèmes d'information de crédit dans la sous-région, comme celles apportées par la base de données du Bureau d'Information sur le Crédit (BIC), une large frange de la clientèle bancaire demeure non couverte : seulement 13 053 020 individus sur un total de 18 221 904 clients dans le système bancaire de la sous-région sont fichés dans cette base.⁴ Cette situation pose ainsi un défi majeur aux institutions de crédit qui se voient contraintes d'adopter des systèmes internes d'évaluation du risque de crédit pour pallier ce manque d'information.

La Bridge Bank Group Côte d'Ivoire (BBG CI), banque commerciale ivoirienne créée en 2006, n'est pas en marge de ces systèmes internes d'évaluation du risque de crédit. En effet, depuis sa création la BBG CI s'est dotée de systèmes internes d'évaluation du risque de crédit basés à la fois sur l'expérience de ses analystes et sur des techniques statistiques telles que le scoring. S'agissant des crédits accordés aux particuliers, elle dispose d'un système de scoring

¹ Source : Rapport annuel de la Commission Bancaire de l'UMOA 2023

² Source : Rapport sur les conditions de banque dans l'UEMOA 2022

³ L'aléa moral décrit une situation d'asymétrie d'information où une partie n'observe pas parfaitement les actions entreprises par l'autre partie (source : https://fr.wikipedia.org/wiki/Al%C3%A9a_moral)

⁴ Source : Rapport annuel 2022 de la BCEAO

pour des montants assez bas. Toutefois, force est de constater que malgré ces systèmes d'évaluation de crédit, elle a enregistré en 2022, une hausse de 19% de la valeur de ces créances en souffrance avec pour talon d'Achille les créances contractées par des particuliers et des Très Petites Entreprises (TPE) représentant à eux deux 55% de la valeur totale des créances en souffrance.⁵ Au regard de cette situation, une étude visant à établir un système de credit scoring pour les TPE a déjà été réalisée mais aucune sur les particuliers n'a été entamée. Dès lors la nécessité d'en réaliser une qui soit propre à ce segment de la clientèle, c'est-à-dire les particuliers, prend tout son sens.

Traditionnellement basées des techniques statistiques classiques et beaucoup plus déterministes, les méthodes de scoring appliquées aux particuliers de ces dernières années tendent à évoluer vers l'intégration de processus de scoring plus élaborés et data-driven. Ainsi pour faire face à l'amas de données constitué de l'historique des prêts octroyés, de leurs caractéristiques financières et des données démographiques des contractants, les modèles de machine learning se révèlent être des véritables solutions. En effet, les modèles de machine learning (apprentissage automatique en français) sont basés sur des algorithmes permettant aux machines d'acquérir progressivement des connaissances à partir des données, aboutissant ainsi à de meilleures prédictions grâce à la détection de relations complexes et de tendances cachées au sein des données. Plusieurs études dans la littérature ont montré leur prédominance prédictive par rapport aux modèles traditionnels, avec une supériorité affirmée pour les modèles de boosting et de bagging (Abdou et al, 2011 ; Dastile et al., 2020, Wang et al. ,2015). Cependant, malgré le potentiel prometteur des méthodes de machine learning, aucune étude n'a encore été menée au sein de la BBG CI pour évaluer leur applicabilité pour l'amélioration du processus de scoring interne.

Ainsi ce mémoire se propose d'examiner comment les méthodes de machine learning pourraient améliorer le credit scoring au sein de la Bridge Bank Group Côte d'Ivoire, en se concentrant sur les clients particuliers.

II. PROBLEMATIQUE

La problématique de cette étude réside dans la nécessité de développer des modèles de machine learning pour prédire de manière précise le risque de défaut de paiement des

⁵ Source : Etudes menées en 2022 par la Direction Recouvrement de la BBGCI

particuliers qui sollicitent des prêts auprès de la BBG CI. Pour ce faire, il convient d'explorer les déterminants clés identifiés par ces modèles, qui permettront d'établir des prédictions fiables et de prendre des décisions de crédit éclairées.

Ainsi, les questions de recherche suivantes guideront cette étude :

- Comment prédire le risque de défaut de paiement d'un particulier qui sollicite un prêt auprès de la BBG CI à partir des méthodes de machine learning ?
- Quels sont les déterminants mis en avant par les modèles de machine learning pour prédire le risque de défaut de paiement des particuliers qui contractent un prêt bancaire auprès de la BBG CI ?

III. OBJECTIFS DE L'ETUDE

Cette problématique soulève les objectifs spécifiques suivant :

- Mettre en œuvre des modèles de classification basés sur les techniques de machine learning pour prédire la solvabilité des emprunteurs.
- Fournir à la BBG CI les déterminants du risque de défaut de paiement lié à sa clientèle.

IV. HYPOTHESES DE RECHERCHE

En se basant sur les résultats observés dans la littérature, nous formulons les hypothèses suivantes :

- Les modèles de machine learning de boosting et de bagging prédisent mieux le risque de défaut de paiement que les modèles classiques.
- Les caractéristiques démographiques des clients ont une influence sur leur risque de défaut de paiement.

V. PLAN DE L'ETUDE

La structure de ce mémoire est organisée comme suit : après cette introduction, la section suivante passera en revue les concepts clés liés au credit scoring et aux modèles de machine learning. Cette section se poursuivra par une revue théorique et empirique relative à la genèse du crédit scoring et aux différentes mutations qu'il a connues, tout en mentionnant les travaux antérieurs en lien avec le sujet. Par la suite, nous plongerons dans l'étude de cas de la BBG CI, en détaillant la méthodologie mise en place pour traiter, exploiter et valoriser les données mises à notre disposition. Les résultats obtenus et leur analyse seront présentés dans la section dédiée, suivis d'une discussion approfondie sur les implications et les recommandations pratiques. Enfin, l'étude se soldera par la mise en place d'une application web de crédit scoring basée sur le modèle retenu avant de terminer par une conclusion générale.

CHAPITRE I : DÉFINITION DES CONCEPTS ET REVUE DE LA LITTÉRATURE

I. DEFINITIONS DES NOTIONS ET CONCEPTS CLES

Dans cette section inaugurale, nous jetterons les bases conceptuelles de notre étude en définissant les termes fondamentaux qui serviront de socle tout au long de notre exploration. Cette démarche assurera une compréhension commune et solide des concepts clés au cœur de notre étude.

1. Les notions et concepts clés liés au secteur bancaire :

a. Les banques

Les banques sont l'un des types d'institutions financières les plus connus en raison de leur large champ d'action incluant une collaboration avec quasiment tous les types d'agents économiques. Selon L'article 511-1 du code monétaire et financier, la banque est un établissement de crédit dont l'activité consiste à recevoir des fonds remboursables du public et à octroyer des crédits. De façon plus large, elles sont responsables de la création monétaire, de la gestion des dépôts et des prêts, et de la fourniture de services financiers à la clientèle. Se faisant, on distingue principalement quatre types de banque en fonction de la nature de leurs activités. Ce sont entre autres :

✓ Les banques centrales :

Surnommée la banque des banques, une banque centrale est une institution publique indépendante chargée d'assurer la stabilité monétaire d'un pays. Chapeautant tous les autres types de banque, les banques centrales ont pour principales missions de garantir la stabilité des prix, d'émettre de la monnaie fiduciaire (c'est-à-dire mettre en circulation des pièces et billets), et de prêter de l'argent aux banques commerciales en vue de financer l'économie.

✓ Les banques commerciales :

Ce sont les plus courantes puisqu'étant des acteurs clés du financement de l'économie. Elles organisent le contact entre les agents à capacité de financement et les agents à déficit de financements. Des premiers, elles reçoivent des dépôts liquides à vue ou à terme contre une rémunération de ces dépôts ; aux seconds, elles octroient des crédits rémunérés à taux d'intérêt débiteurs supérieurs aux taux d'intérêt créditeurs dont bénéficient les déposants. Elles réalisent ces opérations par le biais de divers produits tels que les comptes courants, les

comptes d'épargne, les cartes de crédit, les prêts personnels, les prêts hypothécaires, etc. En outre, soulignons que ces types de banque sont généralement des sociétés constituées d'un capital détenu par des actionnaires dont leur principal but est de réaliser des profits commerciaux (PNB). Elles sont généralement subdivisées en deux départements à savoir les banques de détail destinées aux particuliers et aux petites entreprises et les banques d'affaire destinées aux moyennes et grandes entreprises.

✓ **Les banques d'investissement :**

Contrairement aux banques commerciales traditionnelles, les banques d'investissement se concentrent sur des activités telles que la gestion de fonds, les opérations de fusions et acquisitions, l'émission et la négociation de titres financiers (actions, obligations, dérivés), le conseil en stratégie financière pour les entreprises et les gouvernements, ainsi que le financement de projets complexes. Elles jouent un rôle essentiel dans les marchés financiers en facilitant la mobilisation de capitaux et en fournissant des solutions financières sophistiquées aux clients institutionnels, aux entreprises et aux particuliers fortunés.

✓ **Les banques multilatérales de développement (BMD) :**

En plus des types de banque énoncés plus haut, on dénombre aussi les banques multilatérales de développement (BMD) telles que la Banque Mondiale, la Banque Africaine de Développement, la Banque Asiatique de développement ou encore la Banque Européenne d'Investissement. Ces banques sont des institutions supranationales fondées par des États souverains qui en sont les actionnaires. Leurs mandats reflètent les politiques d'aide au développement et de coopération établie par ces États. Elles ont pour mission commune de favoriser le progrès économique et social dans les pays en développement, en finançant des projets, en soutenant des investissements et en générant des ressources financières au profit de l'ensemble de la population mondiale.

b. Les particuliers

Un particulier est défini comme un individu distinct des entreprises et des entités gouvernementales. Ce sont entre autres :

- Les travailleurs salariés (secteur public, para public et secteur privé) ;
- Les travailleurs indépendants (commerçant, exploitant agricole, artisan, etc.) ;

- Les professionnels libéraux (médecin, avocat, notaire, architecte, pharmacien, etc.) ;
- Les individus sans catégorisation professionnelle (retraité, étudiant, rentier, etc.)

Ainsi dans le contexte bancaire, un particulier est identifié comme étant une personne physique qui utilise les services financiers pour ses besoins personnels.

Les comptes bancaires, les cartes de crédit, les prêts personnels, les hypothèques et d'autres produits financiers sont conçus pour répondre aux besoins spécifiques de cette catégorie de clients. Les relations bancaires avec les particuliers sont axées sur la gestion de leurs finances personnelles, y compris la gestion de leurs dépôts, leurs emprunts, leurs investissements et leurs opérations courantes.

c. Les crédits bancaires aux particuliers

Le crédit bancaire représente un arrangement financier par lequel une institution bancaire accorde temporairement des fonds à un emprunteur, qu'il soit un particulier, une entreprise ou une entité gouvernementale, avec l'obligation de rembourser le montant emprunté assorti d'intérêts sur une période convenue. Cette pratique permet aux emprunteurs de disposer de liquidités immédiates pour financer des besoins variés, allant des achats courants aux projets d'investissement majeurs, tandis que les banques génèrent des revenus en facturant des intérêts sur les montants prêtés.

Le crédit bancaire joue un rôle essentiel dans l'économie en facilitant la circulation de la monnaie, en soutenant la croissance des entreprises et en permettant aux individus de gérer leurs finances de manière flexible, tout en étant soumis aux termes et conditions convenus entre les parties prenantes. Se faisant, on distingue plusieurs types de crédits bancaires allant des crédits spot aux crédits longs termes. Toutefois dans une dynamique de rester fidèle au thème de notre étude, nous nous contenterons de ne présenter que les crédits accordés aux particuliers.

Ipso facto, on distingue principalement deux types de crédit accordés aux particuliers : il s'agit des crédits à la consommation et des crédits immobiliers.⁶

⁶**Source** : Manuel de crédit de la Bridge Bank Group Côte d'Ivoire du 25/06/2023

❖ **Crédits à la consommation** : Le crédit à la consommation est un crédit court ou moyen terme que l'emprunteur peut utiliser librement sans qu'il soit obligé d'apporter une justification de son achat auprès de la banque. L'objectif de ce type de facilité est de répondre à la satisfaction d'un besoin, de résoudre un problème de trésorerie ou de pallier une dépense imprévue. Figurent notamment dans cette catégorie :

- ✓ Les prêts à l'équipement des ménages (moyens de transport, biens d'équipement ménagers) ;
- ✓ Les prêts personnels ;
- ✓ Les prêts afférents à l'usage de cartes de crédit ;
- ✓ Les comptes ordinaires débiteurs des particuliers autres que les entrepreneurs individuels.

En revanche, le crédit à la consommation s'adresse uniquement à la clientèle pouvant justifier d'un revenu jugé éligible par la banque.

❖ **Crédits immobiliers**

Le Crédit immobilier ou prêt immobilier est un emprunt à moyen ou long terme destiné à financer l'acquisition d'un bien immobilier, les travaux de construction ou les travaux de réaménagement sur un tel bien. Lorsque le crédit est adossé à la garantie sur la valeur du bien, on parle de crédit hypothécaire.

.

d. Les risques liés à l'activité bancaire

Les risques sont indissociables du domaine bancaire, englobant à la fois l'incertitude inhérente à l'évolution des gains et pertes, ainsi que les conséquences potentiellement négatives qui en découlent. Il est important de noter que le terme "risque" est rarement associé à la possibilité de gains, pourtant cette double facette de la notion de risque est encapsulée dans les mesures de risque couramment utilisées.

De manière générale, un risque dans le domaine bancaire correspond à un événement susceptible d'impacter les flux financiers d'une institution financière.⁷ Les risques peuvent être catégorisés en risques spécifiques liés à un débiteur particulier et en risques systémiques qui concernent l'ensemble du système bancaire. On distingue le risque spécifique du risque

⁷ **Source** : Vivien BRUNEL, « Gestion des risques et risques de crédit », 2009.

systémique, ce dernier pouvant perturber le fonctionnement global du système bancaire en affectant les banques, les institutions financières, les mécanismes de compensation et les cadres réglementaires.

Les risques relatifs aux banques se déclinent en deux catégories majeures : les risques non financiers et les risques financiers. Les risques non financiers englobent des éléments tels que les risques opérationnels (fraudes, pannes informatiques, sinistres majeurs), les risques réglementaires et légaux, ainsi que les risques liés aux marges ou au volume d'activité.

Parallèlement, on distingue six types de risques financiers (Brunel, 2009). Ce sont entre autres :

- **Le risque de Liquidité** : Représentant une menace majeure, il peut se matérialiser par une ruée des épargnants pour retirer leurs dépôts, en réaction à des rumeurs de non-solvabilité.
- **Le risque de Taux d'Intérêt** : Ce risque découle des fluctuations de la volatilité des taux d'intérêt. Les dépôts collectés à court terme sont souvent placés à moyen et long terme, exposant les banques à d'importants risques de taux d'intérêt.
- **Le risque de Change** : Pour les établissements internationaux, les variations des taux de change peuvent altérer leurs résultats, étant donné qu'ils publient leurs bilans dans une seule devise.
- **Le risque de Marché** : Résultant des opérations d'intermédiation sur les marchés financiers, il découle des fluctuations des marchés, y compris les taux d'intérêt, les taux de change, le crédit, les actions et les matières premières.
- **Le risque des options Cachées** : Notable en banque commerciale, ce risque est lié aux options de taux d'intérêt détenues par les particuliers, comme dans le cas des prêts immobiliers. La non-exercice optimal de ces options rend leur couverture complexe à calibrer.
- **Le risque de Crédit** : Le risque de crédit demeure intrinsèquement lié à l'essence même des activités bancaires, notamment les métiers de base tels que les prêts et le financement. Une banque engagée dans le prêt à un gouvernement ou une entreprise présentant des signes de risque court le péril de ne pas recouvrer l'intégralité du capital

prêté. Ce risque s'entremêle étroitement avec la qualité de la signature de l'emprunteur. En effet, ce risque se segmente en quatre catégories distinctes, à savoir :

- **Le risque de Défaut du Client** : Ce risque se matérialise lorsque l'emprunteur n'est pas en mesure de rembourser intégralement le montant dû, entraînant potentiellement des pertes financières pour la banque.
- **Le risque de Dégradation de la Qualité de Signature** : Il survient lorsque la qualité de la signature de l'emprunteur, généralement évaluée par des agences de notation, se détériore, ce qui peut affecter la probabilité de remboursement futur.
- **Le risque de Marché sur la Qualité de Signature** : Aussi appelé risque de spread, il se rapporte aux variations des écarts entre les taux d'intérêt sur les obligations d'État et les obligations d'entreprise. Ces fluctuations peuvent influencer la perception de la qualité de signature de l'emprunteur.
- **Le risque de Contrepartie sur les Contrats Dérivés** : Ce risque découle des contrats dérivés avec une contrepartie présentant un risque potentiel. En cas de défaut de la contrepartie, la banque pourrait subir des pertes sur ces contrats.

Dans cet éventail de risques susmentionnés, le risque de crédit joue un rôle de premier plan en ce sens où il incarne le risque principal lié à l'activité la plus lucrative des banques commerciales : **le crédit**. Les diverses nuances de ce risque illustrent son impact potentiel sur les opérations et la stabilité de la banque, ce qui souligne l'importance cruciale d'une évaluation minutieuse et d'une gestion proactive de ce risque. Dans le cadre de cette étude, notre évaluation de ce risque s'axera sur sa première dimension c'est-à-dire la dimension qui s'intéresse uniquement à la probabilité de défaut de paiement.

e. Le défaut de paiement

Selon le Comité de Bâle, on parle de défaut de paiement pour un emprunteur donné lorsqu'un ou les deux événements ci-dessous se produisent :

- La banque estime qu'il est peu probable que le débiteur rembourse intégralement son crédit au groupe bancaire sans nécessiter de mesures appropriées telles que la réalisation d'une garantie (si elle existe).

- Le retard de paiement d'un débiteur sur un crédit important dû au groupe bancaire dépasse 90 jours. Les découverts sont considérés comme des créances échues dès que le client a dépassé une limite autorisée ou qu'il a été informé qu'il disposait d'une limite inférieure à l'encours actuel. (Comité de Bâle, 2003, p. 73).

Selon les accords de Bâle III, les banques doivent adopter cette définition du défaut pour l'évaluation de leurs modèles internes de crédit.

Dans le cadre de notre étude et conformément aux pratiques en vigueur à la BBG⁸, la notion de défaillance retenue est pratiquement la même que celle proposée par le comité de Bâle. En effet, le processus formel permettant à la banque d'identifier les défauts de paiement avérés est le déclassement des comptes. **Ainsi, dans toute la suite de cette étude le concept de défaut de paiement sera strictement apparié au déclassement du compte du client.**

2. Les notions et concepts liés au machine learning

a. Le machine learning

Le machine learning ou « apprentissage automatique » en français, est une branche de l'intelligence artificielle qui s'est imposée comme l'un des piliers technologiques majeurs du 21e siècle. Il a façonné de nombreuses avancées technologiques, telles que la recommandation de produits en ligne, la reconnaissance vocale et faciale, la conduite autonome et très récemment les chats bots intelligents (Chat GPT, Gemini, Microsoft Copilot, etc.).

Avec des fondements basés sur la statistique et l'informatique, le machine learning permet aux machines d'acquérir des connaissances à partir des données, sans être explicitement programmées pour chaque tâche⁹. Cette méthode révolutionnaire repose sur la capacité des algorithmes à détecter des motifs et des relations dans les données, ce qui les permet de prendre des décisions et à générer des prédictions précises. En substance, cette méthode habilite les machines à apprendre de l'expérience passée et à généraliser ces enseignements

⁸ **Source** : Manuel de crédit de la Bridge Bank Group Côte d'Ivoire du 25/06/2023

⁹ **Source** : https://fr.wikipedia.org/wiki/Apprentissage_automatique

pour prendre des décisions ou effectuer des prédictions dans de nouvelles situations, sans qu'une programmation explicite soit nécessaire.

Le cœur du machine learning réside dans la capacité à automatiser l'apprentissage à partir des données. Lorsqu'on lui fournit un ensemble de données d'entraînement contenant des exemples et des réponses attendues, un modèle de machine learning explore et découvre progressivement les motifs sous-jacents. Ces modèles sont ensuite utilisés pour faire des prédictions, classer des éléments, effectuer des regroupements, ou même générer de nouvelles données à partir de celles existantes.

Cependant, il ne s'agit pas d'une solution miracle : il nécessite une sélection soignée des données, un choix judicieux d'algorithmes, et une validation rigoureuse pour garantir des résultats fiables.

Se faisant, les techniques de machine learning se subdivisent en trois grandes catégories, chacune adaptée à des types de problèmes spécifiques. Ce sont :

- **L'apprentissage supervisé :** Dans ce type d'apprentissage, les modèles sont formés à partir d'ensembles de données d'entraînement comprenant des exemples étiquetés, c'est-à-dire des données associées à des réponses attendues. Les modèles apprennent à partir de ces exemples pour prédire correctement les étiquettes ou les valeurs pour de nouvelles données. Les tâches courantes de l'apprentissage supervisé incluent la classification, où les modèles catégorisent les données en classes prédéfinies, et la régression, où ils prédisent des valeurs numériques continues en fonction des caractéristiques d'entrée.
- **L'apprentissage non supervisé :** Dans l'apprentissage non supervisé, les modèles sont exposés à des ensembles de données non étiquetées. Leur objectif est de découvrir des structures, des tendances ou des groupements intrinsèques dans les données. Les algorithmes d'apprentissage non supervisé sont souvent utilisés pour l'exploration de données, la détection d'anomalies, la réduction de dimension et le regroupement (clustering).
- **L'apprentissage par renforcement :** L'apprentissage par renforcement s'inspire du concept d'apprentissage humain basé sur des récompenses et des sanctions. Les modèles d'apprentissage par renforcement apprennent en interagissant avec un

environnement dynamique et en ajustant leurs actions pour maximiser une récompense à long terme. Cette approche est couramment utilisée dans les jeux, la robotique et la prise de décision.

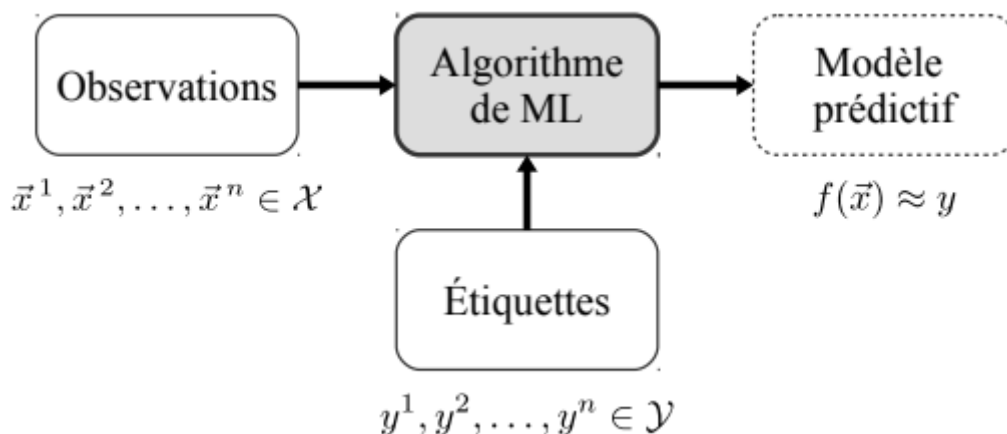
Ces types d'apprentissage ne sont pas mutuellement exclusifs, et certains problèmes peuvent nécessiter une combinaison de plusieurs approches. Par exemple, des approches semi-supervisées peuvent être utilisées lorsque l'on dispose à la fois de données étiquetées et non étiquetées.

Comme cela sera signifié plus loin dans notre travail, notre étude portera sur des données étiquetées et de facto, fera appel à l'apprentissage supervisé. Il convient de ce fait de présenter le formalisme mathématique qui sous-tend cette technique.

a) Le formalisme mathématique de l'apprentissage supervisé¹⁰

On appelle apprentissage supervisé la branche du machine learning qui s'intéresse aux problèmes pouvant être formalisés de la façon suivante : étant données n observations $\{\vec{x}^i\}_{i=1,\dots,n}$ décrites dans un espace X , et leurs étiquettes $\{\vec{y}^i\}_{i=1,\dots,n}$ décrites dans un espace Y , on suppose que les étiquettes peuvent être obtenues à partir des observations grâce à une fonction $\varphi : X \rightarrow Y$ fixe et inconnue : $y^i = \varphi(\vec{x}^i) + \epsilon_i$, où ϵ_i est un bruit aléatoire. Il s'agit alors d'utiliser les données pour déterminer une fonction $f : X \rightarrow Y$ telle que, pour tout couple $(\vec{x}, \varphi(\vec{x})) \in X \times Y$, $f(\vec{x}) \approx \varphi(\vec{x})$. L'espace sur lequel sont définies les données est le plus souvent $X = R_p$ où $p \in \mathbb{N}$, l'ensemble des entiers naturels.

Figure 2 : : Schéma théorique de l'apprentissage supervisé (source : Azencott, 2018)



¹⁰ Azencott, C.A (2018). Introduction au Machine Learning

- Dans le cas où les étiquettes sont binaires, elles indiquent l'appartenance à une classe. On parle alors de classification binaire soit $Y = \{0, 1\}$.
- Dans le cas où les étiquettes sont discrètes, et correspondent donc à plusieurs classes, on parle de classification multi-classes soit $Y = \{1, 2, \dots, C\}$, avec $C \in \mathbb{N}$.
- Dans le cas où les étiquettes sont à valeurs réelles, on parle de régression soit $Y = \mathbb{R}$

b. Quelques modèles classiques de machine learning¹¹

❖ La régression logistique :

La régression logistique est une technique couramment utilisée en machine learning pour résoudre des problèmes de classification binaire, où l'objectif est de prédire la probabilité d'appartenance à une classe donnée en fonction des caractéristiques des données. Lorsque les caractéristiques d'entrée sont représentées par un vecteur X et la variable d'intérêt binaire par Y , la régression logistique modélise la probabilité $P(Y=1|X)$ que l'événement appartienne à la classe 1 (ou la classe positive).

Mathématiquement, la régression logistique utilise la fonction logistique (ou sigmoïde) pour transformer la combinaison linéaire des caractéristiques d'entrée en une probabilité.

La fonction logistique est définie comme suit :

$$g(z) = \frac{1}{1 + e^{-z}}$$

où z représente la combinaison linéaire des caractéristiques d'entrée et des coefficients associés: $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$ avec n représentant le nombre de variables explicatives. Ici, $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ sont les coefficients à estimer à partir des données d'entraînement, et $X_1, X_2, \dots, X_n$ sont les valeurs des caractéristiques d'entrée.

La probabilité $P(Y=1|X)$ est ensuite obtenue en appliquant la fonction logistique à z :

$$P(Y = 1|X) = g(z) = \frac{1}{1 + e^{-z}}$$

Puisque la probabilité d'appartenir à la classe 0 (ou la classe négative) est $(P(Y = 0|X) = 1 - P(Y = 1|X))$, la régression logistique effectue la classification en attribuant les données à la classe ayant la probabilité la plus élevée.

¹¹ Azencott, C.A (2018). *Introduction au Machine Learning*

Pour estimer les coefficients β à partir des données d'entraînement, la méthode de maximum de vraisemblance est généralement utilisée. L'objectif est de trouver les valeurs de β qui maximisent la vraisemblance des données observées, c'est-à-dire les données pour lesquelles on connaît la véritable classe.

❖ Les arbres de décision

Les arbres de décision représentent l'une des méthodes les plus utilisées en machine learning pour la classification et la régression. Le principe de cette méthode consiste à créer des structures hiérarchiques d'ensembles de règles de décision basées sur les caractéristiques des données.

Mathématiquement, la construction d'un arbre de décision implique de diviser les données en fonction de questions posées sur les caractéristiques, tout en visant à maximiser l'homogénéité des sous-groupes résultants.

Le processus de construction d'un arbre de décision commence par le choix d'une caractéristique qui divise le mieux les données. Cela se fait en utilisant des critères de sélection tels que l'indice de Gini ou l'entropie. L'indice de Gini mesure l'impureté d'un ensemble de données et varie de 0 (ensemble pur) à 0,5 (ensemble équilibré). L'entropie, quant à elle, mesure la dispersion des classes dans un ensemble de données. L'objectif est de trouver la division qui minimise l'indice de Gini ou l'entropie, ce qui entraîne des sous-groupes plus homogènes.

Si nous prenons un exemple simple d'indice de Gini, la formule serait :

$$Gini(D) = 1 - \sum_{d=1}^C (p_d)^2$$

où D représente l'ensemble de données, C le nombre de classes, et p_d est la proportion d'éléments de classe k dans l'ensemble D .

Une fois que l'arbre est construit, il peut être utilisé pour classer de nouvelles données en suivant les chemins des questions jusqu'à atteindre une feuille, où une classe ou une valeur est attribuée.

❖ Le modèle de Naïve Bayes

Le modèle de classification Naïve Bayes est une méthode probabiliste largement utilisée pour la classification de texte, de documents et d'autres données catégorielles. Il repose sur le théorème de Bayes et suppose une indépendance naïve entre les caractéristiques, ce qui simplifie le calcul des probabilités conditionnelles. Mathématiquement, le Naïve Bayes utilise les probabilités a priori et conditionnelles pour estimer la probabilité d'appartenance à une classe donnée pour une nouvelle observation.

Supposons que nous ayons un ensemble de données d'entraînement $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, où $\forall i \in [1; n]$, x_i est un vecteur de caractéristiques et y_i est l'étiquette de classe associée. Le but est de prédire l'étiquette y pour une nouvelle observation x .

Le théorème de Bayes énonce que la probabilité conditionnelle $P(y|x)$ peut être calculée à partir des probabilités a priori $P(y)$ et des probabilités conditionnelles $P(x|y)$:

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)}$$

Dans le contexte de Naïve Bayes, l'hypothèse de base est que les caractéristiques x sont indépendantes conditionnellement à la classe y . Cela simplifie considérablement le calcul de $P(x|y)$, car il permet de décomposer la probabilité conditionnelle en produits de probabilités marginales :

$$P(x|y) = P(x_1|y) \cdot P(x_2|y) \cdot \dots \cdot P(x_n|y)$$

❖ Le modèle des SVM

Le modèle des Support Vectors Machines (SVM) ou Machines à Vecteurs de Support en français est une méthode de machine learning utilisée pour résoudre des problèmes de classification et de régression. Mathématiquement, les SVM cherchent à trouver l'hyperplan optimal qui sépare les données en maximisant la marge entre les classes. L'hyperplan est déterminé par les vecteurs de support, qui sont les points les plus proches de la frontière de décision.

Dans le cas de la classification binaire, où les classes sont étiquetées +1 et -1, l'hyperplan est représenté par l'équation :

$$w \cdot X + b = 0$$

où w est un vecteur de poids, b le terme de biais et $X = (x_1, x_2, \dots, x_p)$, un vecteur de variables d'entrée et p le nombre de variables d'entrée. Les données de chaque classe sont projetées sur cet hyperplan, et la distance entre les projections des vecteurs de support des deux classes, appelée marge, est maximisée. Ladite marge est donnée par la formule :

$$\text{Marge} = \frac{2}{\|w\|}$$

où $\|w\|$ représente la norme euclidienne du vecteur de poids.

Cependant, dans la réalité, il n'est pas toujours possible de trouver un hyperplan qui sépare parfaitement les classes. Les SVM introduisent donc des variables d'écart, appelées variables de slack, pour permettre des erreurs de classification. Cela conduit à un problème d'optimisation quadratique qui vise à minimiser la norme du vecteur de poids tout en satisfaisant les contraintes dues aux variables de slack. L'optimisation est formulée comme suit :

$$\text{Min} \begin{cases} \frac{1}{2} \|w\|^2 + C \sum_{k=1}^n \xi_k \\ S/C \\ y_k(w \cdot X_k + b) \geq 1 - \xi_k \text{ et } \xi_k \geq 0 \end{cases}$$

Où N est le nombre d'observations, C est le paramètre de régularisation et ξ_k sont les variables de slack associées à chaque échantillon (X_k, y_k) avec y_k l'étiquette associée au vecteur des variables d'entrée de la k -ème observation.

En somme, les SVM exploitent un formalisme mathématique sophistiqué pour trouver l'hyperplan optimal qui sépare les classes en maximisant la marge tout en autorisant des erreurs de classification contrôlées. Cette approche en fait une technique efficace pour résoudre des problèmes de classification et de régression, même lorsque les données sont non linéairement séparables en utilisant des astuces comme le noyau.

❖ Les forêts aléatoires

Les Forêts Aléatoires (Random Forest) sont une méthode de machine learning qui combine plusieurs arbres de décision pour améliorer la précision et la robustesse des prédictions. Le principe du Random Forest consiste à construire un ensemble d'arbres de décision en utilisant des échantillons de données et des variables d'entrée aléatoires. Chaque arbre vote pour une prédiction et la classe avec le plus de votes est choisie.

L'idée clé des Random Forests est le concept de bootstrap agrégé (bagging) et l'introduction de la variabilité en sélectionnant des sous-ensembles aléatoires de données et de caractéristiques pour former chaque arbre. Le processus de bagging implique de créer plusieurs ensembles de données d'entraînement à partir de l'ensemble de données d'origine en effectuant des échantillonnages aléatoires avec remplacement. Chaque ensemble est ensuite utilisé pour former un arbre de décision.

Mathématiquement, supposons que nous ayons un ensemble de données d'entraînement $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, où x_i représente un vecteur de caractéristiques et y_i est l'étiquette associée.

Le processus de construction d'un Random Forest commence par générer B ensembles de données d'entraînement D_b , chacun créé en effectuant un échantillonnage aléatoire avec remplacement à partir de D . Ensuite, pour chaque ensemble D_b , un arbre de décision est construit en utilisant ces données, mais en introduisant une variabilité supplémentaire : à chaque nœud, seulement un sous-ensemble aléatoire des caractéristiques est pris en compte pour la division.

Mathématiquement, supposons que F_b représente le b -ème arbre dans le Random Forest. Le processus de création d'un arbre F_b commence par le choix d'une caractéristique k à partir d'un sous-ensemble aléatoire de caractéristiques, puis le choix d'un seuil t pour diviser les données. Ces choix sont faits pour minimiser une mesure d'impureté, comme l'indice de Gini.

Ensuite, pour la phase de prédiction, chaque arbre F_b dans le Random Forest donne une prédiction $\hat{y}_b$ pour une nouvelle observation x . Dans le cas de la classification, le vote majoritaire des $\hat{y}_b$ détermine la classe prédite. Dans le cas de la régression, la moyenne des $\hat{y}_b$ est prise comme la valeur prédite finale.

L'ensemble B de prédictions $\hat{y}_b$ peut être noté comme :

$$\widehat{Y}_B = \{\widehat{y}_1, \widehat{y}_2, \dots, \widehat{y}_B\}$$

Enfin, les Forêts Aléatoires profitent de l'effet de la moyenne et de la diversité des prédictions d'arbres pour produire des modèles plus stables et performants.

❖ Les modèles de boosting

Les modèles de boosting sont des techniques avancées en machine learning visant à améliorer la précision des prédictions en combinant plusieurs modèles de base plus simples. Mathématiquement, le boosting ajuste séquentiellement des modèles de base, en attribuant davantage d'importance aux exemples mal classés à chaque itération. Cette approche itérative permet aux modèles de base ultérieurs de se concentrer sur les échantillons difficiles à prédire, améliorant ainsi la performance globale du modèle.

Supposons que nous ayons un ensemble de données d'entraînement $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, où x_i représente un vecteur de caractéristiques et y_i est l'étiquette associée. Le processus de boosting commence par attribuer des poids initiaux égaux à tous les exemples d'entraînement. À chaque itération, un modèle de base est ajusté en minimisant l'erreur pondérée, où les exemples mal classés reçoivent un poids plus élevé.

Mathématiquement, supposons que F_t représente le modèle de base à l'itération t . Le modèle F_t est ajusté pour minimiser l'erreur pondérée E_t :

$$E_t = \sum_{i=1}^n w_{i,t} \cdot I(y_i \neq F_t(x_i))$$

où $w_{i,t}$ est le poids de l'exemple (x_i, y_i) à l'itération t , et $I(y_i \neq F_t(x_i))$ est une fonction indicatrice qui vaut 1 si la prédiction du modèle F_t est incorrecte pour (x_i, y_i) , sinon elle vaut 0. Après avoir ajusté F_t , les poids $w_{i,t}$ sont mis à jour pour attribuer une importance accrue aux exemples mal classés :

$$w_{i,t+1} = w_{i,t} \cdot e^{-\alpha_t y_i F_t(x_i)}$$

où α_t est un facteur d'ajustement calculé en fonction de l'erreur pondérée E_t .

Le modèle global final $F(x)$ est obtenu en agrégeant les prédictions des modèles de base pondérés :

$$F(x) = \sum_{t=1}^T \alpha_t \cdot F_t(x)$$

où T est le nombre total d'itérations.

Différentes variantes de l'algorithme de boosting ont été développées au fil du temps, chacune avec ses spécificités mathématiques. Voici quelques-unes de ces variantes et leurs caractéristiques :

- **AdaBoost (Adaptive Boosting)** : AdaBoost est l'une des premières méthodes de boosting et reste l'une des plus connues. Mathématiquement, AdaBoost attribue initialement des poids égaux à tous les échantillons d'entraînement et ajuste séquentiellement des modèles de base en mettant davantage l'accent sur les échantillons mal classés. L'erreur pondérée de chaque modèle de base est utilisée pour calculer le poids α_t du modèle dans la prédiction globale. La formule d'ajustement des poids $w_{i,t+1}$ est caractéristique d'AdaBoost :

$$w_{i,t+1} = w_{i,t} \cdot e^{-\alpha_t \cdot y_i \cdot F_t(x_i)}$$

- **Gradient Boosting** : L'algorithme de gradient boosting ajuste des modèles de base en minimisant les résidus du modèle précédent. Mathématiquement, le modèle F_t est ajusté en minimisant la fonction de perte $L(y_i, F_{t-1}(x_i) + F_t(x_i))$, où L est une fonction de perte dépendante de la tâche (par exemple, la perte quadratique pour la régression). Les modèles de base ultérieurs se concentrent sur les résidus non encore capturés.
- **XGBoost (Extreme Gradient Boosting)** : XGBoost est une extension optimisée de l'algorithme de gradient boosting. Il introduit des régularisations L1 et L2 pour éviter le surajustement, utilise un algorithme de recherche d'arbre plus efficace et permet le traitement parallèle pour accélérer le calcul.
- **LightGBM (Light Gradient Boosting Machine)** : LightGBM est une autre variante du boosting qui se concentre sur l'efficacité en utilisant des techniques de partitionnement des données et de construction d'arbres. Il utilise une approche de construction d'arbre en profondeur, ce qui peut réduire le temps de construction et la consommation de mémoire.

En somme, les modèles de boosting partagent l'idée de base d'ajuster séquentiellement des modèles de base pour améliorer les prédictions. Cependant, chaque variante a ses spécificités mathématiques et techniques pour gérer les erreurs, les résidus et la régularisation, ce qui les rend utiles dans différentes situations.

❖ Les modèles de réseau de neurones

Les réseaux de neurones sont des modèles profonds en machine learning calqués sur le fonctionnement des neurones dans le cerveau humain. Conceptuellement, un réseau de neurones est constitué de plusieurs couches de neurones interconnectés, chacune exécutant des opérations linéaires et non linéaires sur les données d'entrée pour générer des prédictions ou des représentations transformées. Les réseaux de neurones peuvent être spécifiés par leur architecture, déterminée par le nombre de couches et le nombre de neurones dans chaque couche.

De façon plus mathématique, supposons que nous ayons un ensemble de données d'entraînement $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, où n est un entier naturel, x_i est un vecteur de caractéristiques en entrée et y_i est l'étiquette associée. Un réseau de neurones typique comporte trois types de couches : les couches d'entrée, les couches cachées et la couche de sortie.

La sortie z_q du q -ème neurone dans une couche cachée peut être calculée comme la composition d'une opération linéaire pondérée et d'une fonction d'activation non linéaire :

$$z_q = \sigma \left(\sum_{i=1}^n w_{iq} x_i + s_q \right)$$

où w_{iq} est le poids associé à la connexion entre l'entrée x_i et le neurone q , s_q est le biais du neurone q , et σ est la fonction d'activation, généralement non linéaire comme la fonction sigmoïde, la fonction ReLU (Rectified Linear Unit) ou d'autres fonctions.

La sortie des neurones dans une couche cachée devient ensuite l'entrée pour la couche suivante, et ainsi de suite jusqu'à la couche de sortie. Pour une tâche de classification binaire, la couche de sortie peut utiliser une fonction d'activation telle que la sigmoïde, qui renvoie une valeur entre 0 et 1, représentant la probabilité d'appartenance à la classe positive.

La formation des réseaux de neurones consiste à ajuster les poids w_{iq} et les biais s_q pour minimiser une fonction de perte, généralement définie en fonction de la différence entre les prédictions du réseau et les étiquettes réelles y_i . Cette optimisation est généralement réalisée à l'aide de techniques de descente de gradient, qui ajustent itérativement les paramètres du réseau pour atteindre le minimum global de la fonction de perte. La méthode de rétropropagation, basée sur la règle de la chaîne en calcul différentiel, joue un rôle crucial dans ce processus.

Le principe de cette méthode est le suivant : Supposons que $L(w, s)$ représente la fonction de perte associée aux poids w et aux biais s du réseau de neurones. L'objectif est de trouver les valeurs de w et s qui minimisent L . Il s'agira ainsi d'implémenter une méthode itérative qui ajuste les poids dans la direction opposée du gradient de la fonction de perte par rapport aux poids :

$$w_{ij} \rightarrow w_{ij} - \alpha \frac{\partial L}{\partial w_{ij}}$$
$$s_q \rightarrow s_q - \alpha \frac{\partial L}{\partial s_q}$$

où α est le taux d'apprentissage, contrôlant la taille des pas effectués lors de l'optimisation.

La rétropropagation est la clé pour calculer les dérivées partielles $\frac{\partial L}{\partial w_{ij}}$ et $\frac{\partial L}{\partial s_q}$. Elle repose sur la règle de la chaîne : pour calculer la dérivée d'une fonction composée, on multiplie les dérivées des fonctions individuelles le long du chemin. Pour chaque couche cachée h , les dérivées partielles sont calculées à partir de la couche de sortie.

Cette rétropropagation des dérivées est effectuée en remontant le réseau à partir de la couche de sortie jusqu'à la couche d'entrée. Une fois que les dérivées partielles sont calculées, les poids et les biais sont mis à jour selon la formule de descente de gradient.

In fine, retenons que la formation des réseaux de neurones est un processus mathématique complexe qui implique l'optimisation itérative des poids et des biais pour minimiser une fonction de perte. La descente de gradient et la rétropropagation sont des techniques fondamentales utilisées pour calculer les gradients et ajuster les paramètres du réseau.

A présent que nous avons une idée claire des notions et concepts qui seront au cœur de notre étude, intéressons-nous maintenant aux études et travaux antérieurs en rapport avec notre thème.

II. REVUE DE LA LITTÉRATURE

1. Revue théorique

Le phénomène de l'emprunt et du prêt a une longue histoire associée au comportement humain (Thomas et al., 2002). Malgré cette histoire très longue du crédit remontant à environ 2000 av. J.-C. ou même plus tôt, l'histoire du credit scoring se révèle être relativement plus courte, ne commençant que six décennies en arrière. En effet, les origines de cette méthode remontent au début du XXe siècle, avec la publication de la première grille de notation pour des obligations commerciales par John MOODY (1909) Par conséquent, la littérature bibliographique à ce sujet demeure récente mais assez fournie par la pléthore d'auteurs qui s'y sont intéressé.

Ainsi, avant d'explorer les travaux effectués dans cette littérature récente, nous allons d'abord accorder un point d'honneur aux différentes définitions et conceptions du credit scoring.

a. Définition du scoring par la littérature

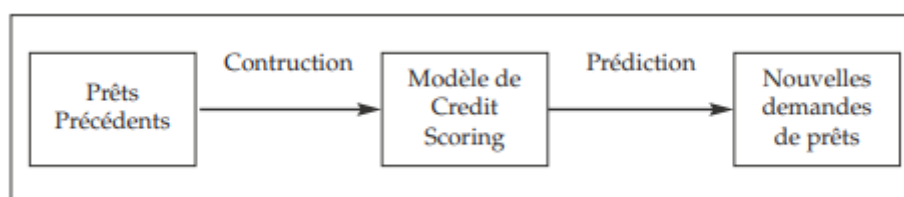
Bien que les définitions attribuées au crédit scoring diffèrent d'un auteur à un autre, il est généralement admis que le crédit scoring est un outil de gestion des risques visant à anticiper la probabilité de défaut d'un nouveau prêt en se basant sur les prêts antérieurs. Ainsi, conformément à la perspective de Feldman (1997), le crédit scoring est « le processus d'assignation d'une note ou d'un score à un emprunteur potentiel pour estimer la performance future de son prêt ». Cette démarche repose sur des mesures quantitatives de performance et les caractéristiques des prêts précédents, permettant de prédire la performance des prêts à venir.

Le crédit scoring se concentre principalement sur la prédiction de la probabilité de défaut des emprunteurs individuels, excluant cependant les prêts à responsabilité conjointe (Schreiner, 2002). Il ne prend pas de décisions directes sur l'approbation ou le rejet d'une demande de prêt, mais il est spécifiquement conçu pour anticiper la probabilité d'occurrence d'une

mauvaise performance, définie par le prêteur, telle que le défaut ou le retard de remboursement (Caire et al., 2003).

Thomas et al. (2002) considèrent le crédit scoring comme un ensemble de modèles de décision et de techniques sous-jacentes qui facilitent la prise de décision en matière d'octroi de crédits à la consommation. Dans le contexte de la microfinance, Viganò (1993) n'offre pas de définition précise de ces modèles, les considérant plutôt comme des instruments permettant aux analystes d'obtenir une vision approfondie des caractéristiques de l'emprunteur. Ainsi, le crédit scoring devient un processus complexe impliquant une analyse minutieuse des informations relatives aux emprunteurs afin d'estimer la probabilité que le prêt sollicité soit remboursé de manière régulière. Cependant, les systèmes de scoring ne cherchent pas forcément à déterminer le risque de non-remboursement d'un crédit, ils peuvent aussi chercher à déterminer la probabilité de retard de remboursement ou la probabilité de demande de renouvellement de crédit par un client ayant soldé un précédent crédit. Les différentes explications fournies pour la définition du crédit scoring peuvent être résumées par le graphique ci-après :

Figure 3 : Processus du crédit scoring (source : Yang, 2001)



b. Les méthodes d'évaluation du risque de crédit : Approche traditionnelle vs. Approche statistique

L'évaluation du risque de crédit d'un client repose principalement sur la comparaison de ses caractéristiques avec celles d'autres clients dans le passé, certains ayant remboursé leurs prêts et d'autres ayant fait défaut. Ainsi le principe de base est que si les caractéristiques du client sont similaires à celles des défaillants, la demande est généralement rejetée, tandis qu'elle est acceptée si elles se rapprochent de celles des clients qui se sont montrés solvables (Thomas, 2000). Deux méthodes sont généralement utilisées pour cette évaluation : l'approche subjective impliquant une analyse individuelle par un analyste de crédit et l'approche statistique via le credit scoring.

L'approche basée sur le jugement humain repose sur l'expertise et l'expérience des analystes de crédits dans l'évaluation du risque de crédit (Thomas, 2000). Elle implique une évaluation individuelle de chaque demande de prêt par un décideur agissant en tant que créancier (Sullivan, 1984 ; Bailey, 2004). Le succès de cette approche dépend dès lors de l'expertise et de l'intuition de ce dernier. Bien que sujette à la subjectivité, cette méthode présente des avantages tels que la prise en compte des aspects qualitatifs et l'utilisation de l'expérience passée de l'analyste de crédits (Al Amari, 1979). Cependant, cette approche peut être biaisée en raison des préjugés individuels et de l'incohérence dans les décisions prises par différents analystes de crédit.

L'approche statistique est quant à elle basée sur l'utilisation de modèles mathématiques et statistiques pour évaluer le risque de crédit. Cette approche objective permet de réduire la subjectivité et les erreurs liées au jugement humain (Altman, 1968 ; Beaver, 1966). Elle permet une évaluation plus objective en évitant les biais potentiels liés au jugement humain, et peut traiter de grandes quantités de données rapidement et efficacement (Abdou et al., 2011) .

Toutefois, dans la pratique, il est recommandé d'introduire le crédit scoring comme outil de raffinement dans le processus de prêt, afin de combiner à la fois techniques statistiques et jugement humains (Gool et al. 2011).

Bien que le credit scoring permette de distinguer les demandeurs de crédit solvables, il est critiqué pour ses problèmes statistiques et les hypothèses sous-jacentes à la technique utilisée pour réaliser la prédiction. Malgré cela, il reste l'un des outils les plus utilisés dans le domaine financier (Sullivan, 1981 ; Bailey, 2004).

Afin d'aiguiser notre opinion sur le sujet, nous consacrerons la partie suivante à analyser les avantages et inconvénients du credit scoring.

c. Les avantages et inconvénients du scoring

Le credit scoring suscite un débat riche et nuancé sur ses avantages et inconvénients par divers auteurs.

D'une part, plusieurs arguments favorables émergent de la littérature. Altman (1968) met en avant l'objectivité accrue qu'offre le credit scoring, en soulignant la nécessité d'une évaluation impartiale basée sur des données quantitatives et qualitatives pour améliorer la précision des

prévisions de risque de crédit. Cette objectivité permet non seulement aux prêteurs de prendre des décisions plus éclairées, mais également de réduire les biais subjectifs dans le processus d'octroi de crédit. Par ailleurs, Eisenbeis (1996) souligne l'efficacité du credit scoring en permettant un traitement rapide et efficient des demandes de crédit, ce qui réduit considérablement les coûts d'évaluation pour les prêteurs. Cette efficacité contribue de facto à améliorer l'expérience client par la réduction des délais de traitement des demandes de crédit, tout en diminuant dans le même temps les coûts associés.

En outre, Thomas et al. (2002) soulignent que le credit scoring peut élargir l'accès au crédit en évaluant le risque pour des emprunteurs sans historique de crédit. Se faisant, cette démarche favorise l'inclusion financière en permettant aux prêteurs d'évaluer le risque de crédit pour des emprunteurs qui n'ont pas d'historique de crédit, levant ainsi un obstacle majeur à leur accès au crédit.

Cependant, malgré ces avantages, des inconvénients non négligeables existent. A cet effet, Viganò (1993) met en garde contre la discrimination potentielle engendrée par le credit scoring, soulignant que des biais dans les données utilisées peuvent conduire à une discrimination injuste envers certains groupes d'emprunteurs. Cette discrimination, souvent involontaire, peut entraîner des refus de crédit injustifiés, affectant ainsi négativement l'accès au crédit et l'inclusion financière de certains groupes.

De plus, Viganò (1993) évoque également le risque de déshumanisation du processus d'octroi de prêt. En se basant principalement sur des statistiques, le credit scoring peut entraîner une perte de contact humain et une détérioration de la relation client. Cette déshumanisation peut, à son tour, compromettre la flexibilité et le jugement dans le processus d'octroi de prêt, ce qui peut altérer la satisfaction des emprunteurs.

Enfin, des auteurs comme Schreiner et Nagarajan (1998) ainsi qu'Eisenbeis (1996) soulignent les risques associés à une dépendance excessive aux modèles de credit scoring. Cette dépendance peut mener à des décisions de crédit inexactes et à des refus de prêts injustifiés, impactant directement l'accès au crédit et la solvabilité des emprunteurs.

2. Revue empirique

Au fil des années, diverses techniques de machine learning ont été appliquées pour améliorer la précision et l'efficacité des modèles credit scoring. L'un des travaux pionniers sur le sujet a été réalisé par Abdou et al. (2011). En effet, ces derniers ont réalisé une revue complète de

214 articles explorant les techniques statistiques et des critères d'évaluation utilisés dans le credit scoring. Ainsi, il en est ressorti l'utilisation fréquente des modèles de régression logistique, des arbres de décision, des réseaux neuronaux et des méthodes d'ensemble telles que les forêts aléatoires et les algorithmes de boosting pour la prédiction des probabilités de défaut. De plus, les auteurs ont fait ressorti que globalement, la performance des modèles de credit scoring était généralement évaluée en fonction de critères tels que la précision, la sensibilité, la spécificité et l'aire sous la courbe ROC.

Dans une autre étude arpentant la même problématique, réalisée par Thomas et al. (2017), les auteurs soulignaient l'importance de la sélection des variables et du prétraitement des données, qui judicieusement effectuées permettraient l'exploitation du plein potentiel des différents modèles aboutissant ainsi à une comparaison des performances plus objectives. En outre, ces derniers identifient l'utilisation de techniques d'apprentissage par ensemble telles que le boosting de gradient et le bagging comme des moyens d'améliorer la robustesse et la précision des modèles de credit scoring. Dans la même continuité, des recommandations sont formulées quant aux bonnes pratiques dans l'établissement d'un modèle de scoring via machine learning à savoir l'utilisation de la validation croisée pour éviter le surajustement (overfitting) et l'optimisation des hyperparamètres des modèles afin d'affiner leurs performances. Ils ont également mentionné l'utilisation de sources de données alternatives, telles que les médias sociaux et les données transactionnelles, pour améliorer la puissance prédictive des modèles de notation de crédit.

Dans cette même veine de bonnes pratiques, Johnson et al. (2019) ont également souligné la nécessité d'interprétabilité et de transparence dans les modèles de deep learning pour le credit scoring, ainsi que l'explication de la manière dont les modèles effectuent des prédictions. L'étude a également discuté de l'importance de la confidentialité des données et des considérations éthiques lors de l'utilisation de modèles de deep learning dans le credit scoring. De plus, la revue de la littérature a souligné l'importance de la surveillance continue de la performance du modèle et de sa mise à jour pour garantir que les modèles restent précis et pertinents dans le paysage dynamique du credit scoring.

Par ailleurs, Thomas et al. (2017) ont également comparé les performances de divers modèles d'apprentissage automatique dans la notation de crédit. Ils ont constaté que le gradient boosting, les machines à vecteurs de support (SVM) et les réseaux neuronaux surpassaient systématiquement les modèles traditionnels de régression logistique et d'arbres de décision en

termes d'exactitude prédictive et de robustesse. De même, Johnson et al. (2019) ont examiné l'utilisation de modèles de deep learning dans le credit scoring et ont constaté que ces modèles étaient capables de capturer des motifs complexes dans les données de crédit et obtenaient une précision prédictive plus élevée par rapport aux modèles traditionnels. De plus Aledanjro (2016) a constaté que la régression logistique traditionnelle ne performait pas aussi bien que les méthodes précédentes lorsque les différentes variables caractéristiques présentaient des relations non linéaires compliquées. Toutefois, l'auteur n'a pas manqué de mentionner que bien que le modèle de régression logistique ne soit aussi performant que le modèle d'apprentissage automatique en termes de précision de prédiction, il présente un avantage considérable en termes d'interprétabilité et de stabilité des variables. Par conséquent, certains chercheurs ont amélioré la régression logistique et l'ont appliquée à la prédiction du comportement de défaut des emprunteurs.

En outre, on note également des travaux réalisés pour le cas des banques africaines. C'est le cas de Kouassi (2021) qui a élaboré un modèle de scoring pour la Société Générale Côte d'Ivoire, basé sur les PPO (Prêts Personnels Ordinaires). Son étude a abouti à la prédominance du Random Forest sur les autres modèles, atteignant ainsi une F1-score de 0,90. Par ailleurs, Dastile et al. (2020), des chercheurs de l'université de Witwatersrand, à Johannesburg en Afrique du Sud, ont montré qu'un ensemble de classificateurs fonctionne mieux que des classificateurs individuels. Leur étude basée sur un récapitulatif de 74 études principales, telles que des articles de revues et des actes de conférences, publiées entre 2010 et 2018, a montré également que bien que les modèles de deep learning n'aient pas été largement utilisés dans la littérature sur le scoring, ceux-ci affichaient des résultats prometteurs pour le peu d'études dans lesquelles ils étaient utilisés. Dans le même article, ils ont également proposé une guideline au regard des résultats des 74 articles étudiés, afin de faciliter les projets de credit scoring. Se faisant, nous nous servons cette guideline pour mener à bien notre étude.

CHAPITRE II : PRESENTATION ET ANALYSE DE LA BASE DE DONNEES

L'objectif de ce chapitre est de présenter la base de données et de réaliser l'analyse exploratoire de ces données.

I. PRESENTATION DE LA BASE DONNEES

La base de données sur laquelle se fonde notre étude est extraite du core-banking¹² de l'entreprise. Elle retrace les informations liées aux crédits accordés par la BBG CI aux particuliers de 2008 à 2022 et dont la date d'échéance finale est inférieure au 31 Décembre 2022. Elle n'est pas exhaustive en raison de certains changements internes au sein du système informatique de la banque rendant impossible l'extraction de la totalité des informations enregistrées avant 2016. Toutefois l'échantillon reçu, composé de 828 lignes apparaît représentatif pour l'étude que l'on souhaite mener en ce sens où il est exhaustif pour les années récentes (2016 à 2021).

Se faisant, les informations transcrites dans la base de données sont retracées à partir de 15 variables qui sont énumérées dans le tableau ci-dessous.

Figure 4 : Tableau récapitulatif des variables de la base de données à l'étude

Libellé des Variables	Type de variable
INDICE	Variable quantitative discrète
SEXE	Variable qualitative nominale
AGE	Variable quantitative discrète
SITUATION_MAT	Variable qualitative nominale
ACTILIB	Variable qualitative nominale
PROFESSION	Variable qualitative nominale
AGENCELIB	Variable qualitative nominale
RM	Variable qualitative nominale
LIBELLE	Variable qualitative nominale
MNTprt	Variable quantitative continue
MARGE	Variable quantitative continue
TXT1	Variable textuelle
DATDEP	Variable temporelle
DERNDAT	Variable temporelle
STATUT	Variable binaire

- ✓ **INDICE** : L'indice du compte bancaire présent dans notre base est en réalité la racine du numéro du compte bancaire du client.
- ✓ **SEXE** : Cette variable indique si le client est un homme ou une femme. Elle prend la valeur H s'il s'agit d'un homme et F s'il s'agit d'une femme.

¹² Le **Core Banking** désigne le système informatique (logiciel) utilisé pour traiter et gérer les opérations d'une banque.

- ✓ **AGE** : Il s'agit de l'âge du client durant l'année où le prêt lui a été octroyé.
- ✓ **SITUATION_MAT** : Cette variable indique le statut matrimonial de l'emprunteur. Elle prend les valeurs suivantes :
 - M si l'emprunteur est légalement marié ;
 - C s'il est célibataire;
 - D s'il est divorcé;
 - V s'il est veuf ;
 - A s'il se trouve dans un autre cas que ceux énoncés plus haut.
- ✓ **ACTLIB** : Cette variable indique le secteur d'activité dans lequel le client exerce. On distingue au total 17 secteurs classifiés selon la politique de financement de la banque.
- ✓ **PROFESSION** : Il s'agit de la profession exercée par le client. Aucune classification spécifique n'a été effectuée par la banque pour cette variable. On dénombre ainsi plus de 140 modalités.
- ✓ **RM** : Cette variable indique le chargé d'affaire qui a accompagné le client pour sa demande de financement. A titre de rappel, notons qu'un chargé d'affaires est un employé de la banque qui s'assure de la gestion des relations avec les clients, en leur proposant des solutions bancaires adaptées à leurs besoins et en s'assurant du suivi et du bon déroulement du financement jusqu'à son remboursement total.
- ✓ **LIBELLE** : Cette variable indique le type de prêt contracté par le client. Les modalités présentées par cette variable sont essentiellement les produits proposés par la banque aux particuliers. Ce sont :
 - 'CMT CONSO BONNE GAMME'/ 'CCT CONSO BONNE GAMME'/'CMT CONSO PERSO'/'CCT CONSO PER AUTRE'/ 'CMT CONSO'/ 'CCT CONSO' : Il s'agit globalement des formules de financement consentis à une personne physique dans le but de satisfaire des besoins de consommation (sauf immobilier, aménagement et équipement). Le montant et le taux sont variables en fonction de la situation commerciale et du type de produit.

- **'CMT SCOLAIRE'/'CCT SCOLAIRE'** : Il s'agit d'un prêt permettant à un salarié (exclusion des salariés mis en disponibilité) de financer des études (y compris les frais annexes) soit pour lui ou soit pour un tiers à charge. Le financement est plafonné à 30 millions de FCFA au taux de 7,5 % et remboursable sur une durée maximale de 5 ans.
- **'CMT HAB PATRIMONIALE' /'CMT HAB BONNE GAMME'** : Il s'agit de formules de financement destinées à l'acquisition, l'aménagement, la construction ou le refinancement d'un bien immobilier. Sont exclus les employeurs Tier 3 (employeurs jugés risqués selon les critères d'évaluation de la banque).
- **'CCT RESTRUCTURES' /'CMT RESTRUCTURES'** : Il s'agit d'une option permettant à un titulaire de prêt de réduire ses mensualités à mi-parcours du remboursement de son prêt.
- **'BRIDGE PRET EXPRESS'** : Le prêt à la consommation d'urgence est une formule de financement consenti à une personne physique dans le but de satisfaire un certain nombre de besoin courant et personnel, sans justificatifs particuliers (factures...). Le financement est plafonné au montant d'un mois de salaire au taux de 12 % et remboursable sur une durée maximale de 6 mois.
- **'BRIDGE PRET RELAIS'** : Le prêt relais est une formule de financement qui est basée sur l'obtention d'une prime, gratification ou autres avantages devant servir à son remboursement. Le financement est plafonné à 5 millions de FCFA au taux de 12 % et remboursable sur une durée maximale de 5 ans.
- **'CCT HORS PP'/'CMT HORS PP'** : Il s'agit des prêts accordés aux particuliers mais ne respectant pas les garanties exigées. Ces prêts sont généralement accordés au regard d'autres facteurs que les garanties.
- **'CCT AUTRES CRD' ou 'CMT AUTRES'** : Il s'agit des autres types de crédits accordés aux particuliers, non mentionnés plus haut.

- ✓ **MNTPRT** : Cette variable indique le montant du crédit accordé.
- ✓ **TXT1** : Cette variable indique les différentes garanties associées au prêt.
- ✓ **DATDEP** : Il s'agit de la date de la mise à disposition du crédit au client.
- ✓ **DERNDAT** : Il s'agit de la date d'échéance finale du crédit.
- ✓ **STATUT** : Il s'agit de la variable permettant de savoir si le prêt a été correctement remboursé ou non. Elle prend la valeur 0 si le prêt a été correctement remboursé et 1 s'il ne l'a pas été.

II. APUREMENT DE LA BASE DE DONNÉES

Avant de plonger dans la méthodologie proposée par Dastile et al.(2019), nous nous devons d'avoir une base cohérente et exploitable. Ainsi, nous avons effectué un travail préliminaire d'apurement de la base de données dont le déroulé se décline selon les points subséquents.

1. Tests de cohérence

Dans l'optique de déceler d'éventuelles erreurs contenues dans les données, nous avons appliqué quelques tests de contrôle.

Le premier test de contrôle appliqué est celui entre le type de prêt et la durée du prêt. En effet, la typologie des prêts à notre disposition implique des durées de remboursement maximales spécifiques. Ainsi, il s'est agi de se rassurer que les crédits court terme avait effectivement une durée de remboursement inférieur à 24 mois et les crédits moyen terme, une durée de remboursement supérieur à 24 mois et inférieur à 84 mois. D'autres types de prêts tels que le bridge prêt express et le bridge prêt relais étaient également concernés par ces tests.

Par la suite, un second contrôle a été appliqué sur les dates de début et de fin d'échéance des prêts. Ce contrôle consistait d'une part à s'assurer que la date de début était effectivement moins récente que celle de fin et d'autre part vérifier que ces dernières étaient effectivement comprises entre 2006 et 2022

2. Suppression des doublons

Étant en présence de plusieurs doublons dans la base, essentiellement causé par des problèmes propres au core banking utilisé par la banque, nous avons procédé à la suppression des doublons. Ainsi après déduction des doublons, notre base de données comptabilisait 711 lignes contenant chacune les caractéristiques des différents prêts distincts.

3. Autres manipulations

La base brute extraite du core banking comprenait des variables et des métadonnées que nous avons jugé non pertinentes pour notre étude. Il s'agissait entre autres des variables telles que le nom du client (qui a été supprimé pour des raisons de confidentialité des données), la devise (qui était uniquement le XOF), le pays où s'est fait le crédit (qui était uniquement la Côte d'Ivoire) et bien d'autres variables qui présentaient ces mêmes caractéristiques.

Par la suite, nous avons déterminé de façon manuelle, la durée de remboursement des prêts, en effectuant la différence entre la date de décaissement et celle de l'échéance finale. Cette tâche relevant normalement de la partie features engineering a été effectuée en amont afin de faciliter l'analyse descriptive.

Un autre pré-traitement appliqué aux données était le format des variables dates qui ressortait sous le format MM/JJ/AAAA, en lieu et place du format JJ/MM/AAAA habituellement utilisé. Ainsi, ces variables dates ont été reparamétrées sous le format convenable sur Excel avant importation sur Python.

4. Création de la variable d'intérêt : STATUT

Les données provenant directement du core banking de la banque, aucune classification réelle n'a été faite entre clients sains et clients déclassés. Ainsi, pour mener à bien notre étude il a fallu créer une variable dichotomique qui nous permettrait de savoir si le client était sain ou déclassé. Ainsi pour reconnaître les clients déclassés des clients sains, il nous a fallu tout simplement identifier ceux qui avaient comme chargé d'affaire des responsables du service recouvrement. En effet, lors du déclassement du compte d'un client, le core banking remplace automatiquement le nom de son chargé d'affaire initial par celui du responsable du recouvrement désigné pour poursuivre l'affaire. Ainsi, les manipulations nécessaires ont été

menées afin de générer notre variable dichotomique STATUT, prenant comme valeur 1 si le client a été déclassée et 0 sinon.

III. ANALYSE EXPLORATOIRE DES DONNÉES

Dans toute cette partie, les graphiques présentés seront exclusivement extraits de nos travaux réalisés sur python à l'aide des librairies plotly, seaborn et matplotlib. Nous nous abstenons donc de préciser la source de chaque graphique, comme il est de coutume.

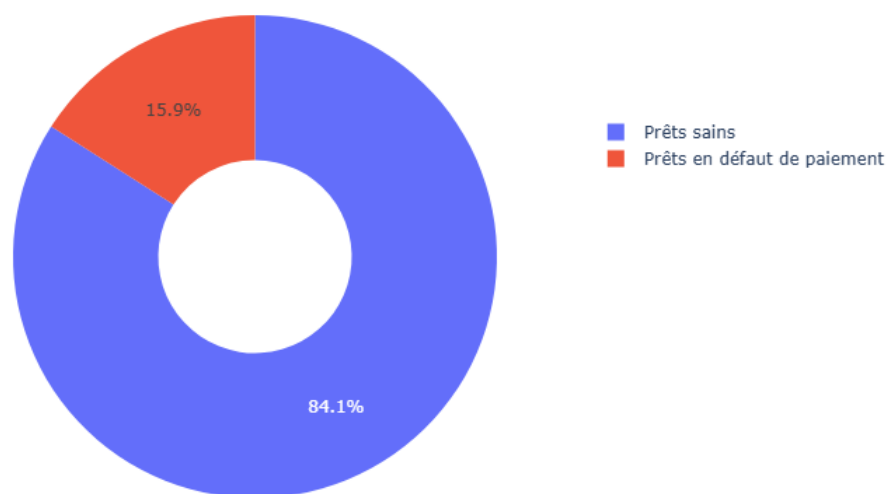
1. Analyse univariée

Nous débutons l'exploration de nos données par une visualisation unidimensionnelle de ses composantes.

➤ Analyse des variables qualitatives

Comme toute banque à vocation commerciale dans une économie plutôt stable, la Bridge Bank Group enregistre naturellement un pourcentage beaucoup plus élevé de prêts qui sont remboursés avec succès par les emprunteurs que de prêts qui se terminent par le déclassement du compte de l'emprunteur. En effet, dans notre population d'étude, cette réalité se reflète de manière significative, avec 84,1% des emprunts parvenant à un dénouement positif, tandis que seulement 15,9% connaissent des défauts de paiement, entraînant le déclassement du compte de l'emprunteur.

Figure 5 : Répartition de la population étudiée en fonction du défaut de paiement



Pour revenir aux spécificités démographiques de notre population, nous notons qu'elle est constituée en majorité d'hommes représentant environ 73,2% de la population totale (figure 6). Quant à la situation matrimoniale au sein de notre population, nous observons une distribution quasi-équilibrée entre les personnes mariées et celles qui sont célibataires, bien que légèrement en faveur de cette dernière catégorie (figure 7).

Figure 6 : Répartition de la population étudiée en fonction du sexe

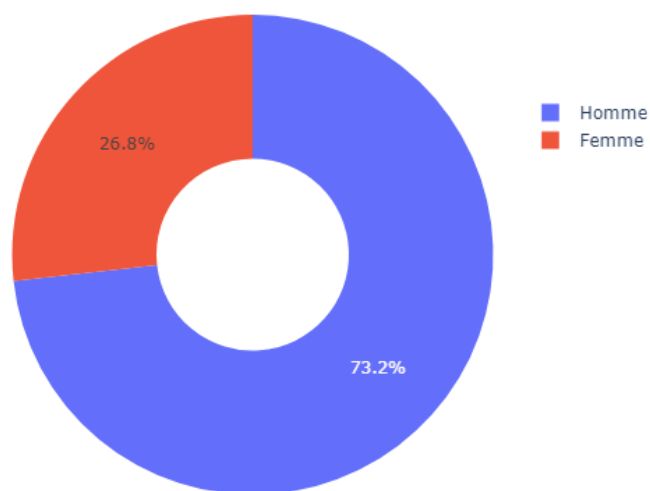
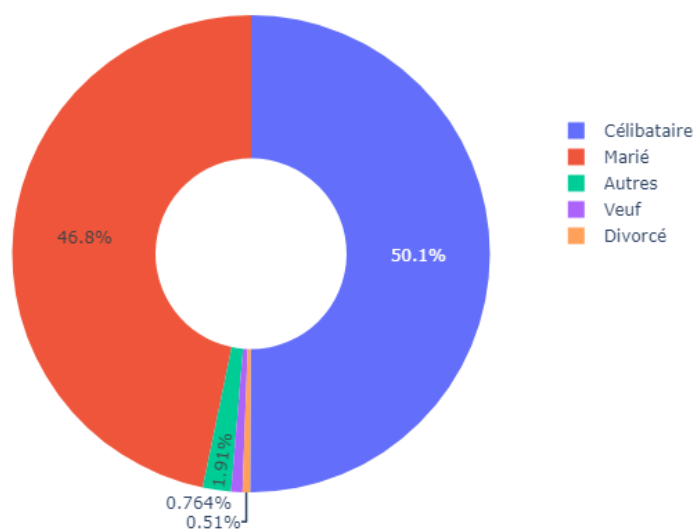


Figure 7 : Répartition de la population étudiée en fonction du statut matrimonial



Hormis la modalité « Divers » pour laquelle on a peu de précisions, on note que les clients sont à majorité dans des secteurs à fort niveau d'éducation telles que le domaine de la finance, de la banque et de l'informatique (figure 8 et 9).

Figure 8 : Répartition de la population étudiée en fonction du secteur d'activité

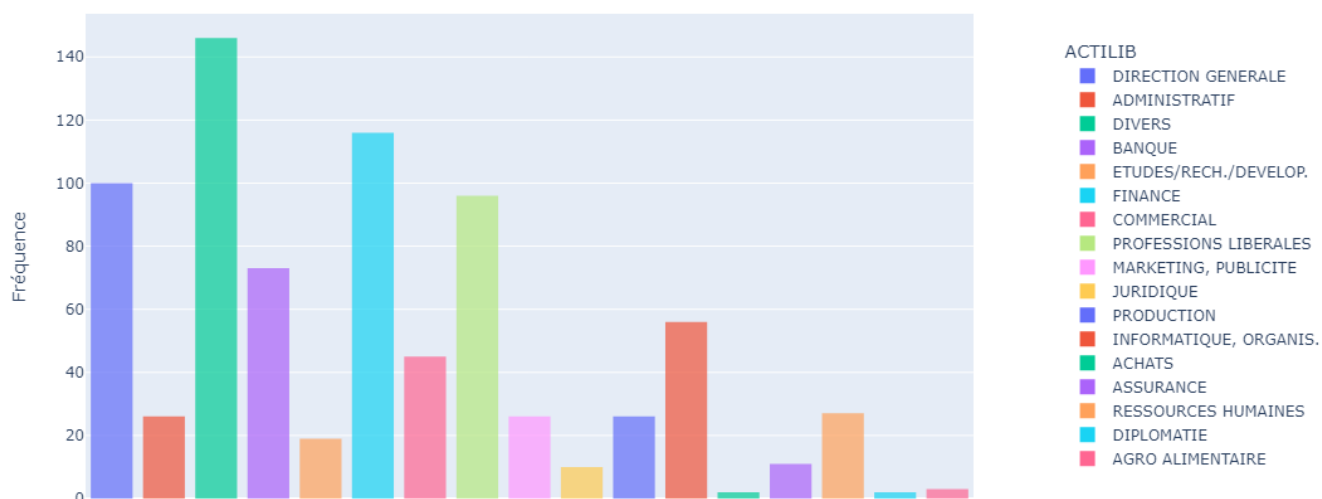
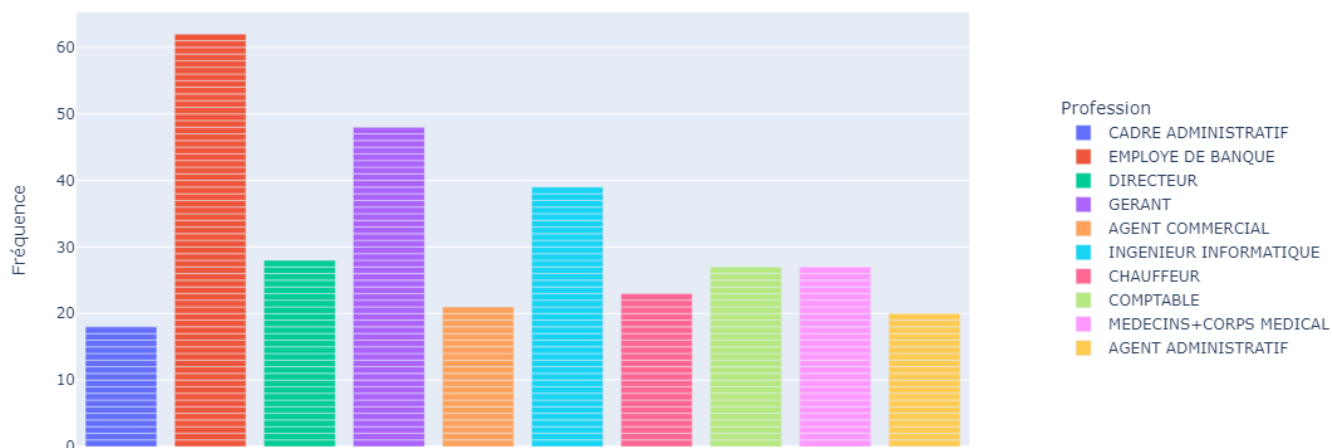
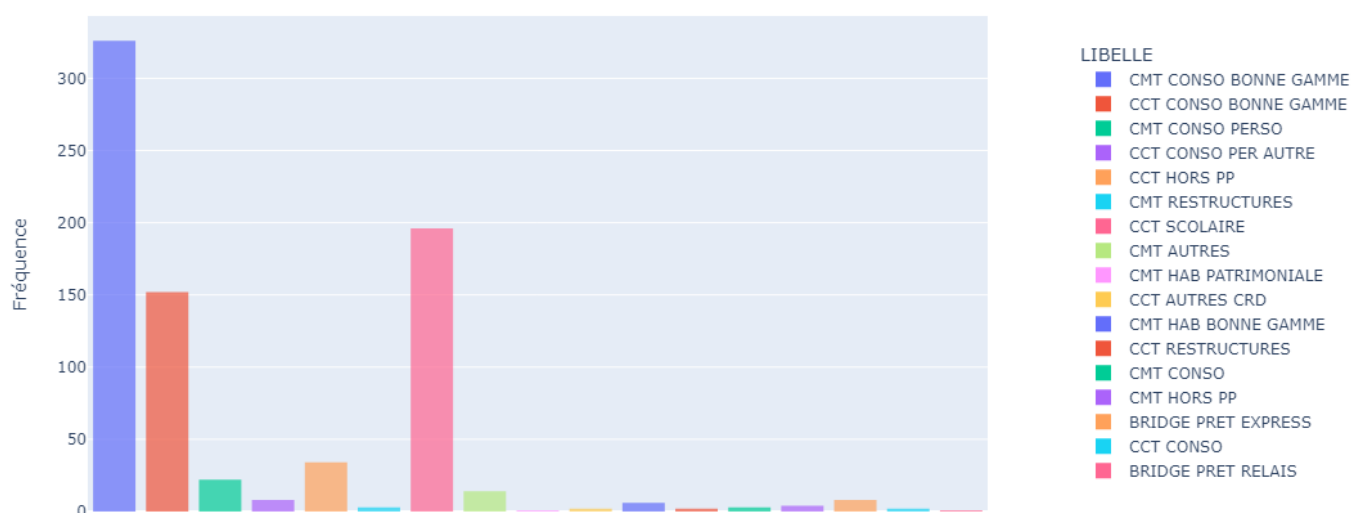


Figure 9 : Répartition de la population étudiée en fonction du top 10 des professions les mieux représentées



Par ailleurs, on note que les prêts à la consommation sont les prêts les plus octroyés suivi de très loin par les prêts scolaires (Figure 10).

Figure 10 : Répartition de la population étudiée en fonction des types de prêt.



➤ Analyse des variables quantitatives

Au niveau des variables quantitatives, on note que la moyenne d'âge des clients est de 48 ans et les montants des crédits octroyés présentent une fourchette très large, allant de 200 milles FCFA à 800 millions de FCFA. Le montant moyen octroyé s'élève à environ 14 millions de FCFA pour un taux d'intérêt moyen de 7,91% et une durée moyenne de remboursement d'environ 32 mois.

Tableau 2 : Statistiques descriptives des variables quantitatives de la base de données

Libellés	AGE (en années)	DUR_REMB (en années)	MNTPRT (en FCFA)	MARGE (en %)
Moyenne	48,14	31,54	13 913 784,70	7,91
Ecart-type	9,49	20,21	39 243 990,61	2,83
Min	26	1	200 000	0
1er quartile	42	12	1 500 000	7,5
Médiane	47	24	5 950 000	8,5
3ème quartile	53	59	15 000 000	9
Max	123	83	800 000 000	13

Par ailleurs, on observe des valeurs extrêmes au niveau des montants des prêts octroyés et de l'âge des clients. Ces variables seront de ce fait retraitées dans la suite de l'étude.

➤ Analyse de la variable textuelle

L'unique variable textuelle de notre base est celle des garanties. A l'aide des techniques de text mining, nous avons réalisé un nuage des mots permettant de se faire une idée des

Ainsi, il ressort de ces nuages de mots que les garanties sont d'apparence les mêmes quel que soit le statut du prêt. Par ailleurs, de ces nuages, il ressort de façon prépondérante les garanties suivantes :

- La domiciliation du salaire et des loyers (des revenus en général) ;
- Les DAT nantis (Dépôt À Terme) ;
- Les cash coll (progressif et up front) ;
- Les ordres de virement permanent ;
- Les assurances décès et invalidités ;
- Les hypothèques ;
- Les épargnes bloquées ;
- Les billets à ordre (BAO).

Ces garanties occurrence ainsi relevées nous permettront de créer de nouvelles variables dummies en vue d'enrichir notre base de données.

➤ **Analyse des variables temporelles**

La courbe d'évolution des prêts octroyés aux particulier affiche une croissance dans sa première phase jusqu'à l'atteinte d'un pic en 2018 avant de décroître jusqu'en 2020 et enfin connaître une dernière phase ascendante en 2021. Le faible nombre de prêts octroyés de 2008 à 2015 (moins de 50 prêts octroyés par an) s'explique par la non-exhaustivité des données pour ces années en raison d'un changement du core-banking de la BBGCI en 2015. De facto, les données obtenues entre 2016 et 2021 sont beaucoup plus réalistes car elles enregistrent la totalité des prêts contractés durant cette période et échus avant le 31 Décembre 2022. Par ailleurs la baisse du nombre de prêts en 2020 pourrait s'expliquer par le ralentissement des activités économiques causé par la COVID 19 (figure 12).

En outre, le mois le plus prolifique, toutes années confondues, est incontestablement celui de septembre avec 178 prêts octroyés (figure 13).

Figure 13 : Courbe d'évolution annuelle des prêts octroyés aux particuliers de 2008 à 2021.

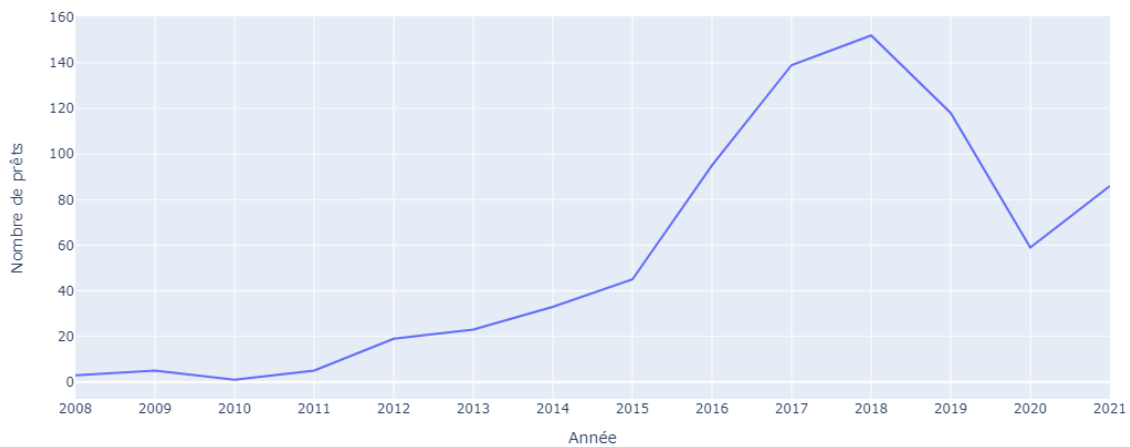
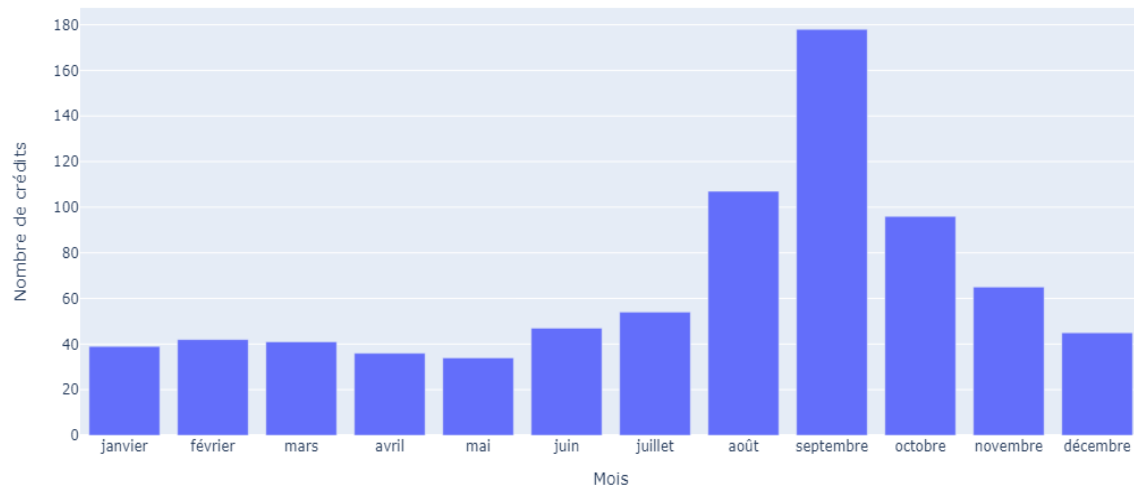


Figure 14 : Nombre de prêts octroyés aux particuliers repartis par mois de 2006 à 2022.



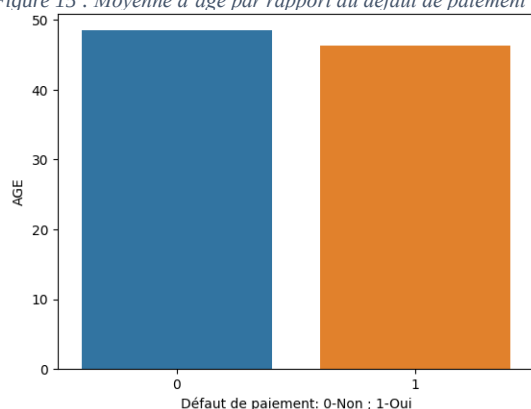
2. Analyse bivariée

a. Variables quantitatives Vs. Variable cible

Notre variable cible semble être quelque peu liée à certaines variables quantitatives présentes dans notre base de données. En effet, selon la figure 14, les personnes moins âgées semblent être plus susceptibles d'être en défaut de paiement d'autant plus que le test de student associé à cette analyse affiche une p-value significative. Il en est de même pour la durée moyenne de remboursement qui semble influencer le défaut de paiement en ce sens où les clients avec des comptes déclassés ont en moyenne une durée de remboursement plus longue que ceux dont les crédits se sont correctement dénoués (Figure 15). Même constat entre le montant des prêts et notre variable cible (Figure 16). Toutefois, le test de student associé conclu à une différence non significative entre la moyenne des montants des prêts sains et ceux présentant des défauts

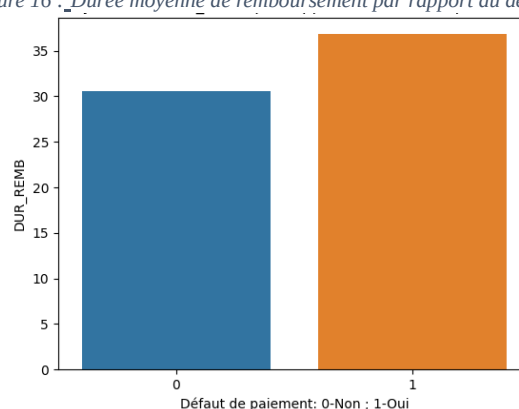
de paiement. De même, aucune différence significative n'est observée entre la moyenne des taux d'intérêt des deux groupes d'individus. (Figure 17).

Figure 15 : Moyenne d'âge par rapport au défaut de paiement



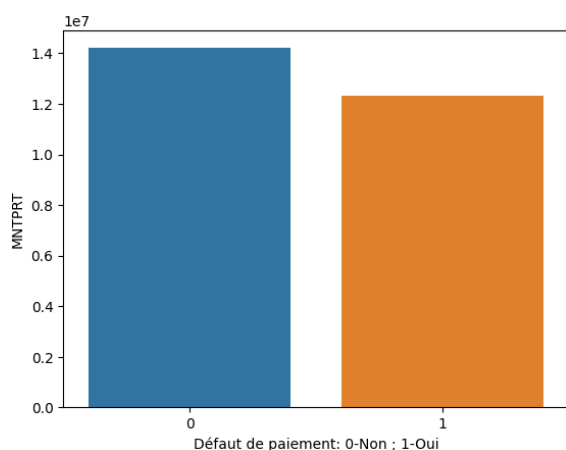
Statistique de test $t : 2,340$ $P\text{-value} : 0.020^{**}$

Figure 16 : Durée moyenne de remboursement par rapport au défaut de paiement.



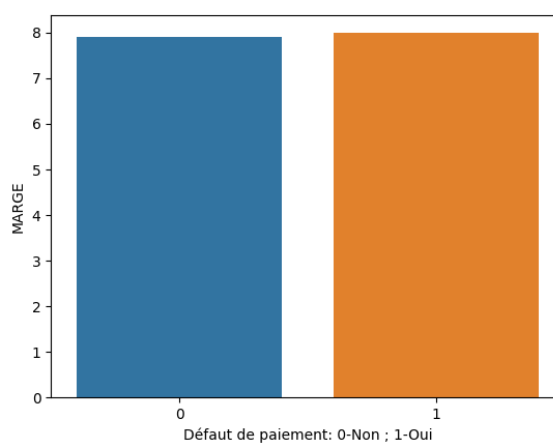
Statistique de test $t : -3.195$ Valeur de $p : 0.002^{***}$

Figure 17 : Moyenne des montant des prêts par rapport au défaut de paiement.



Statistique de test $t : 0,488$ $P\text{-value} : 0.625$

Figure 18 : Moyenne des taux d'intérêt par rapport au défaut de paiement.



Statistique de test $t : -0.321$ Valeur de $p : 0.748$

b. Variable qualitative Vs. Variable cible

A l'analyse des liaisons entre les variables qualitatives et notre variable cible, une liaison en particulier attire notre attention : Il s'agit de la liaison entre notre variable cible et la modalité BANQUE de la variable SECTEUR. En effet, la modalité en question enregistre à elle seule 26 déclassés sur seulement un total de 73 individus appartenant à cette modalité. Cette liaison se perçoit également à travers la variable PROFESSION, pour laquelle plus de 40% des individus ayant pour profession « Employé de banque » affiche un compte déclassé.

Figure 19 : Répartition de la population étudiée en fonction du secteur d'activité et du statut de défaut de paiement

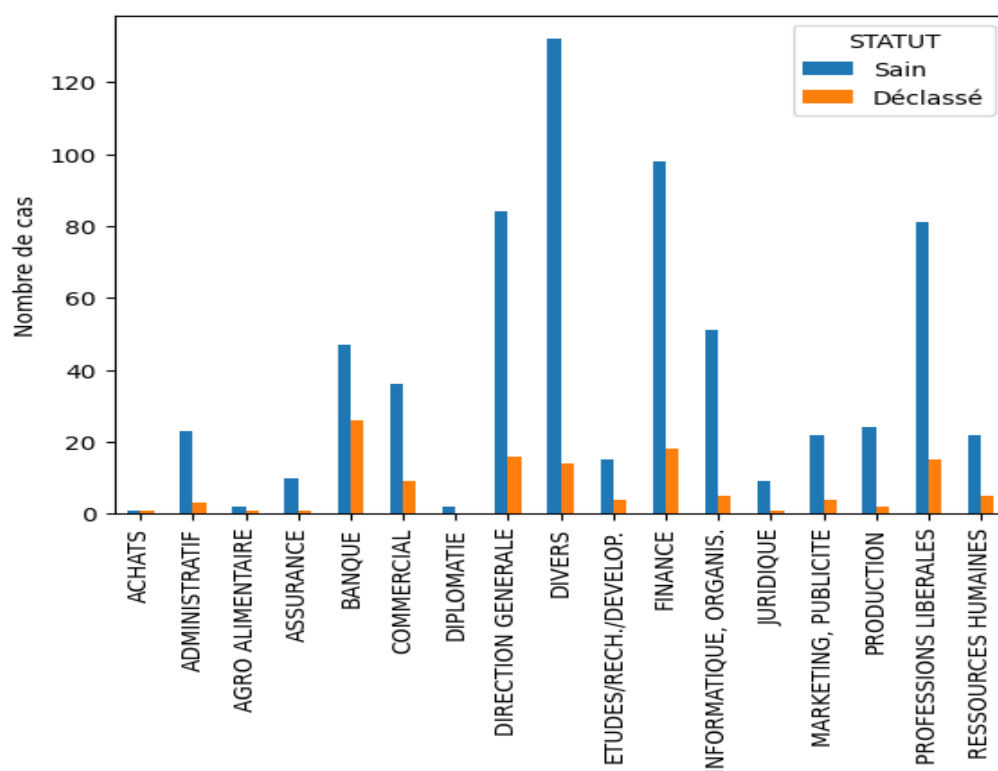
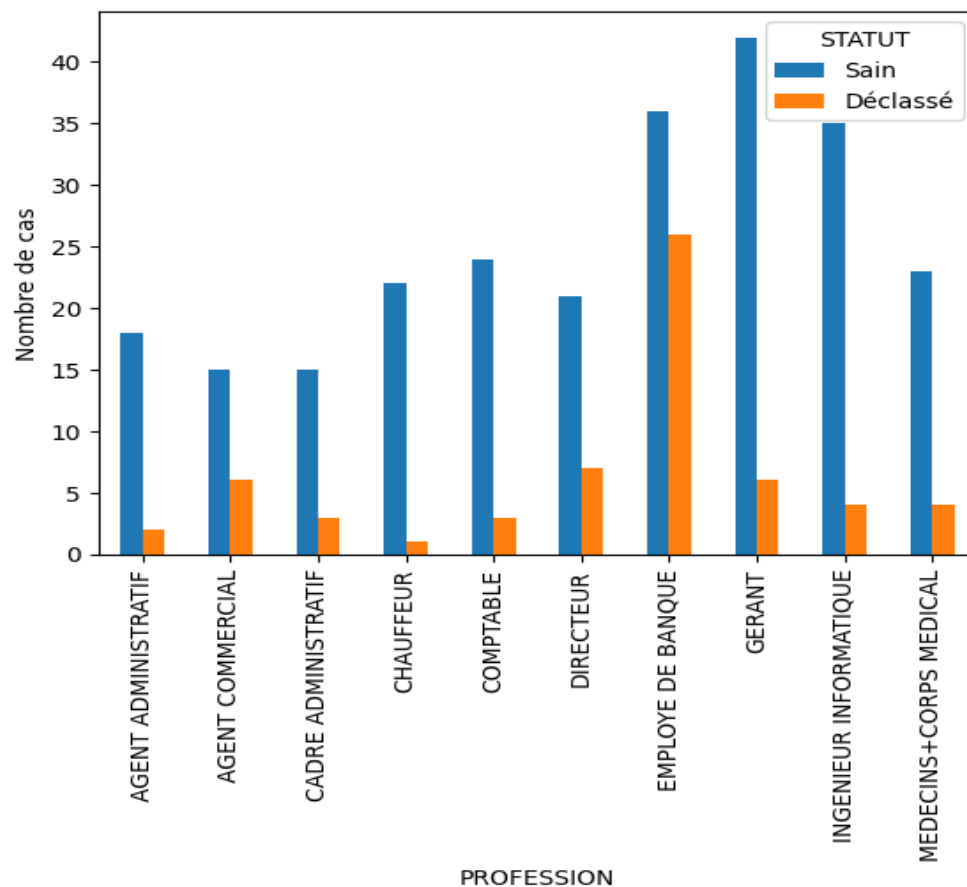


Figure 20 : Répartition de la population étudiée en fonction de la profession et du statut de défaut de paiement



c. Variables quantitatives Vs. Variables quantitatives

En dehors du couple de variables âge et durée de remboursement, toutes les autres paires de variables quantitatives présentent une corrélation statistiquement significative. Cependant, il est à noter que les coefficients de corrélation restent relativement faibles, tous inférieurs à 0,5. La corrélation la plus forte est observée entre la durée de remboursement du prêt et le montant du prêt, avec un coefficient de corrélation de 0,485.

Tableau 3 : Tableau récapitulatif des tests de kendall entre les variables quantitatives

Variable 1	Variable 2	Coefficient de Kendall	P-value
AGE	DUR_REMB	-0.0148	0.5580
AGE	MNTPRT	0.1259	0.0000***
AGE	MARGE	0.1266	0.0000***
DUR_REMB	MNTPRT	0.4851	0.0000***
DUR_REMB	MARGE	0.2773	0.0000***
MNTPRT	MARGE	0.3282	0.0000***

CHAPITRE III : MÉTHODOLOGIE ET MODELISATION.

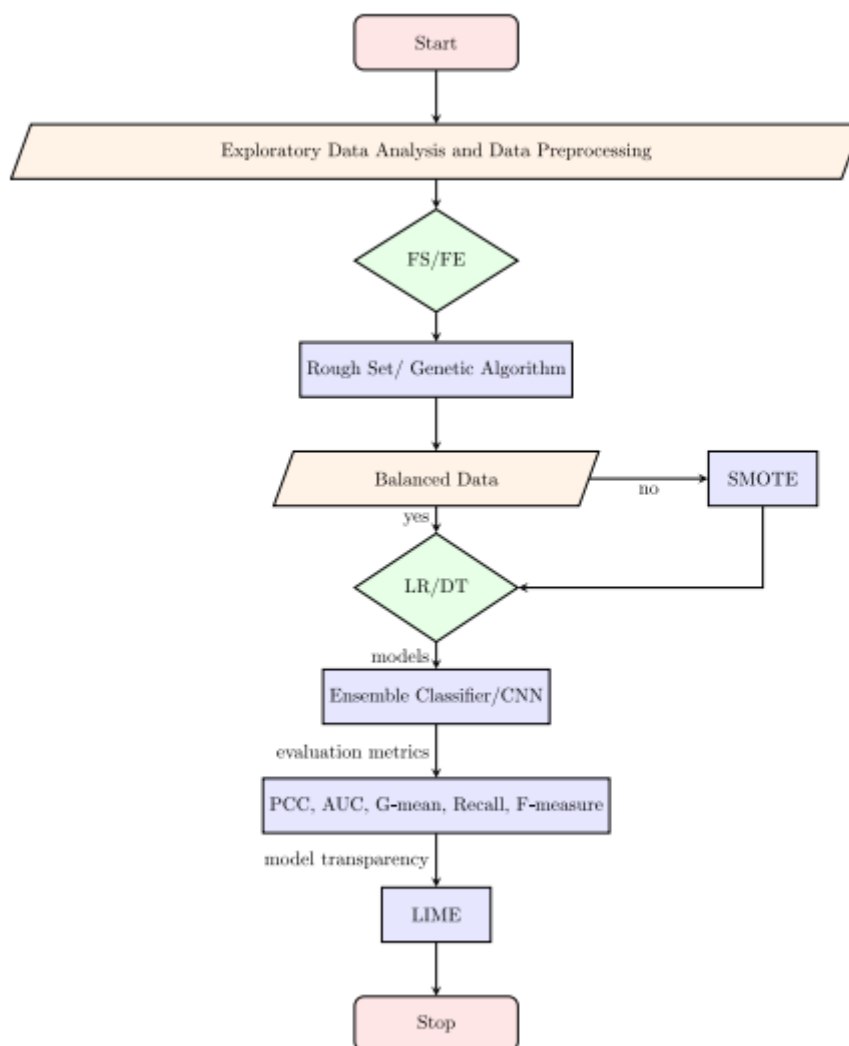
Cette section présente la méthodologie adoptée dans cette étude ainsi que la justification des modèles utilisés pour atteindre les résultats escomptés.

VI. MÉTHODOLOGIE DE L'ÉTUDE

La méthodologie adoptée dans cette présente étude sera celle de Dastile et al. (2020).

En effet, en s'appuyant sur les méthodologies et les résultats de 74 articles de recherche portant sur l'application des méthodes de machine learning au credit scoring et publiés entre 2010 et 2018, Dastile et al. (2020) ont proposé une guideline pour réaliser du credit scoring à partir des techniques de machine learning.

Figure 21 : Guideline pour un projet de crédit scoring basé sur les techniques de machine learning (source : Dastile et al., 2020)



Par ailleurs, des étapes supplémentaires n'entravant pas la logique de la guideline principale ont été ajoutées en vue d'affiner l'étude. L'analyse exploratoire étant déjà effectuée au chapitre précédent, nous passons directement au Data Preprocessing.

1. Data Preprocessing

Cette étape vise à préparer les données brutes en vue de les rendre aptes à l'entraînement de modèles.

a. Gestion des données manquantes :

Notre base ne contient aucune valeur manquante. Aucun traitement particulier de gestion des valeurs manquantes n'a été effectué à cet effet.

b. Encodage des variables catégorielles :

Il s'agit ici d'encoder les variables qualitatives numériquement. Pour ce faire, nous avons utilisé la méthode `get_dummies` de Pandas permettant de faire l'encodage one-hot (ou dummy) des variables catégorielles. Cette méthode prend en entrée un DataFrame et une liste de colonnes à encoder et crée par la suite pour chaque variable catégorielle de la liste de nouvelles variables binaires correspondantes aux différentes modalités de la variable. Elle prend le soin en outre de ne générer que n-1 nouvelles variables pour une variable à n modalités afin d'éviter la colinéarité entre les variables. La colonne d'origine est par la suite supprimée laissant place à plusieurs variables binaires correspondantes aux différentes modalités de la variable catégorielle. Pour chaque variable binaire, la valeur 1 est attribuée à l'individu lorsque celui-ci présente cette modalité et 0 sinon.

Par ailleurs, notons qu'en raison du nombre de modalités relativement élevé de la variable profession, nous décidons de ne pas lui appliquer cette transformation et de l'écarter de facto de notre analyse.

Ainsi, à l'issue de cette transformation, notre base de données affiche au total 54 variables.

c. Normalisation des valeurs numériques :

Afin de réduire l'impact des différences de grandeur entre les variables numériques dans l'ensemble de données, une mise à l'échelle robust scaler est appliquée pour remettre toutes les valeurs numériques à des échelles comparables. Le choix du robust scaler se justifie par la présence de données aberrantes dans la distribution de certaines variables numériques telles que le montant du prêt. Ainsi grâce à sa mise à l'échelle robuste qui utilise la médiane et l'écart interquartile (IQR) pour normaliser les données, cette méthode ressort résistante aux valeurs aberrantes.

d. Features Engineering

L'ingénierie des variables ou features engineering en anglais est le processus de création de nouvelles variables ou de la transformation de variables existantes dans un jeu de données afin d'améliorer les performances d'un modèle d'apprentissage automatique.

Dans le cadre de notre étude, quatre (04) techniques de features engineering ont été appliquée aux données.

❖ La technique basée sur la discrétisation des variables numériques en classe

Cette technique consistait à discrétiser la variable âge en classe d'âge afin de faciliter l'analyse. Ainsi elle a été divisée en 6 classes d'amplitude 10 partant de la classe [20 ans, 29 ans] à la classe [70 ans, et + [.

❖ La technique basée sur les opérations arithmétiques entre les différentes variables

Comme souligné dans la partie apurement des données, nous avons généré une nouvelle variable durée de remboursement en mois à partir des variables indiquant la date d'octroi du prêt et la date d'échéance finale du prêt. Nous avons jugé que cette variable serait mieux interprétable que les dates uniquement.

❖ La technique basée sur les données temporelles

Cette technique consistait à extraire le mois et l'année de la date d'octroi du prêt afin de permettre au modèle d'extraire des informations utiles pour sa performance. Cela se justifie par ailleurs par le fait que la majorité des modèles de machine learning n'ont pas la capacité de faire des prédictions avec des variables au format date.

❖ La technique basée sur l'analyse textuelle et l'expérience métier

En référence à l'analyse textuelle effectuée dans l'analyse exploratoire des données, des garanties assez courantes ont été repérées. En se basant sur l'expérience métier bancaire, certaines de ces garanties se révèlent plutôt pertinentes pour l'octroi d'un crédit. Se faisant, nous avons décidé de générer des variables binaires indiquant pour chaque crédit les garanties présentes dans la variable textuelle garantie, tout en limitant les variables générées qu'aux garanties pertinentes relevées. Ces garanties pertinentes sont :

- e. La domiciliation du salaire et des loyers (des revenus existants en général) ;
- f. Les DAT nanti (Dépôt À Terme) ;
- g. Les cash collatéraux progressifs ;
- h. Les cash collatéraux up front

- i. Les assurances multirisques, décès et invalidités ;
- j. Les hypothèques ;
- k. Les billets à ordre (BAO) ;

En outre on note aussi des prêts qui sont faits sans garanties. Ces prêts sont qualifiés de prêts à garantie clean. Ainsi, pour tenir compte de cette spécificité, nous rajoutons une autre variable binaire CLEAN pour identifier ces types de prêts.

In fine, notre base de données présente à cette étape 71 variables.

2. Découpage de la base de données

Après toutes ces étapes supra, nous scindons à présent notre base en données d'entraînement et données test : 70% de l'ensemble de données est utilisé pour créer la base d'entraînement et 30% de l'ensemble de données est utilisé pour créer la base de test des modèles, la répartition se faisant en tenant compte du déséquilibre des classes de notre variable cible.

3. Traitement des déséquilibres de classe :

Comme constaté dans l'analyse exploratoire, notre base présente un déséquilibre de classes pour notre variable cible. Pour traiter ce déséquilibre, des techniques de sous-échantillonnage et de sur-échantillonnage peuvent être appliquées pour surmonter ce problème.

Le sous-échantillonnage est la technique qui réduit la population majoritaire pour la ramener à l'effectif de la population minoritaire qui est celle des clients en défaut de paiement dans notre cas. A l'opposé le sur-échantillonnage est la technique qui augmente la proportion de la classe minoritaire pour la porter à égalité à celle de la classe majoritaire. Suivant la guideline méthodologique de base, nous opterons pour le sur-échantillonnage.

En effet, Chawla et al. (2002) présentent une méthode de sur-échantillonnage de la classe minoritaire qui consiste à créer des exemples synthétiques de classe minoritaire. La technique de suréchantillonnage des minorités (SMOTE) fonctionne par le biais de l'algorithme des plus proches voisins (k-nearest neighbour) pour améliorer la proportion de la classe minoritaire en créant des données synthétiques.

Cette méthode sera ainsi appliquée uniquement sur les données d'entraînement et non sur les données test afin d'éviter d'avoir des scores fallacieux sur des données test synthétiques et déjà drainées par une logique bien connue à l'avance (le KNN).

4. Métrique d'évaluation

Afin d'évaluer et de mesurer la performance de classification des modèles, nous avons utilisé les métriques suivantes.

a. L'exactitude (Accuracy) :

L'exactitude est la métrique la plus intuitive. Elle permet de déterminer la proportion de prédictions correctes parmi l'ensemble total des prédictions. Elle est calculée en divisant le nombre de prédictions correctes par le nombre total d'échantillons. L'exactitude est une mesure globale de la performance du modèle et convient aux ensembles de données équilibrés car elle peut être trompeuse avec des ensembles de données déséquilibrés, car un modèle prédisant toujours la classe majoritaire obtiendra une exactitude élevée. Ainsi à cette mesure, nous ajouterons d'autres métriques beaucoup plus précises.

b. La précision ou spécificité :

La précision évalue le nombre de vrais positifs (échantillons correctement prédits comme positifs) parmi tous les échantillons prédits comme positifs (vrais positifs + faux positifs). Elle mesure la capacité du modèle à éviter de classer à tort des échantillons négatifs comme positifs. Appliquée à notre étude, la précision représente la proportion des crédits jugés risqués alors qu'ils le sont vraiment. La précision se révèle importante lorsque les faux positifs c'est-à-dire les crédits jugés risqués à tort ont un coût élevé.

c. Le rappel ou sensibilité :

Le rappel, également appelé sensibilité, évalue le nombre de vrais positifs parmi tous les échantillons réellement positifs (vrais positifs + faux négatifs). Il mesure la capacité du modèle à identifier correctement les échantillons positifs. En d'autres termes, cette métrique permet d'évaluer la capacité du modèle à bien prédire les crédits effectivement risqués. Le rappel est essentiel lorsque les faux négatifs c'est-à-dire les crédits jugés sains à tort ont un coût élevé.

Toutefois l'activité à nature commerciale de la banque impose des coûts élevés à la fois pour les crédits sains rejetés à tort et pour les crédits risqués octroyés à tort. Ainsi, l'idéal serait d'avoir une métrique qui permettra de mesurer simultanément ces deux phénomènes. C'est le principe du F1 score.

d. Le F1-score :

Le F1-score est une mesure combinée de la précision et du rappel. Il est calculé en utilisant la

$$\text{formule } F1 = 2 \times \frac{(\text{précision} * \text{rappel})}{(\text{précision} + \text{rappel})}.$$

De plus le score F1 est particulièrement adapté pour évaluer la performance d'un modèle sur des données assez déséquilibrées.

e. La matrice de confusion :

Cette matrice est en quelque sorte le tableau sommaire des 4 métriques présentées plus haut. Elle présente une vue claire des prédictions correctes (vrais positifs et vrais négatifs) ainsi que des erreurs (faux positifs et faux négatifs). Elle est essentielle pour comprendre les points forts et points faibles du modèle.

Dans notre cas d'étude, elle pourrait se présenter comme suit :

Tableau 4 : Matrice de confusion appliquée à notre étude (source : nos travaux)

Libellés	Crédits prédits risqués	Crédits prédits sains
Crédits effectivement risqués	Nombre de crédits risqués bien prédits	Nombre de crédits risqués mal prédits
Crédits effectivement sains	Nombre de crédits sains mal prédits	Nombre de crédits sains bien prédits

f. Aire sous la courbe ROC (AUC-ROC) :

L'AUC-ROC est une mesure de la performance d'un modèle de classification binaire en analysant la courbe ROC (Receiver Operating Characteristic). Cette courbe trace la relation entre le taux de vrais positifs et le taux de faux positifs à différents seuils de classification. Plus l'AUC-ROC est proche de 1, meilleur est le modèle pour séparer les classes.

5. Modèles de base et sélection de variables

Les modèles de base retenus pour réaliser la première évaluation sont la régression logistique et les arbres de décision. Le choix de ces modèles se justifie par leur simplicité, leur interprétabilité et leurs performances avérées par la littérature

Tableau 5 : Scores des modèles de base

Modèle	Exactitude	Précision	sensibilité	F1-Score
Régression Logistique	0,86	0,59	0,40	0,48
Arbre de décisions	0,70	0,16	0,20	0,18

Comme on peut le constater les modèles de base présentent des scores pas très reluisants. Le F1-score qui est la métrique la plus utilisée dans la littérature n'atteint pas les 0,5 quel que soit le modèle. Dans l'optique d'améliorer ce score, nous allons dans un premier temps faire une sélection de variables, puis dans un second temps, utiliser et optimiser d'autres modèles plus complexes.

L'algorithme utilisé pour la sélection de variables sera le Select KBest. Ce choix se justifie par la facilité de compréhension de l'algorithme et de son temps de calcul raisonnable.

L'idée fondamentale derrière cette méthode est de calculer une statistique de test, généralement basée sur l'analyse de la variance ou des tests statistiques spécifiques à la régression, pour chaque variable par rapport à la variable cible. Ensuite, les variables sont classées en fonction de leurs scores de test, et les meilleures variables, celles qui ont les scores les plus élevés, sont sélectionnées pour être incluses dans le modèle final.

L'un des avantages de cette méthode est sa capacité à améliorer l'efficacité des modèles en éliminant les caractéristiques inutiles ou redondantes, ce qui peut réduire le surajustement (overfitting) et accélérer le processus d'apprentissage. Cependant, il est important de noter que le choix du paramètre K (le nombre de variables à sélectionner) peut avoir un impact significatif sur les performances du modèle, et il est souvent nécessaire de l'ajuster à l'aide d'une validation croisée.

Ainsi à l'aide de la méthode SelectKBest importée de la librairie Scikit Learn, le K optimal procurant le meilleur score était de 40 avec un f1 score de 0,55.

VII. MODÉLISATION

1. Modèles

En se basant sur la revue de littérature, notamment sur les travaux des auteurs qui ont effectué des analyses de synthèse des modèles les plus performants pour le scoring (Dastile et Al., 2020 ; Abdou et Al., 2018 ; Brown, 2012), nous postulons pour le choix des 8 modèles suivants :

- La Régression Logistique ;
- Les Forêts aléatoires (Random Forest Classifier) ;
- Les Machines à Vecteurs de Support (SVM) ;
- L'Extrême Gradient Boosting (XGBoostClassifier) ;
- Le Light Gradient Boosting (LGB) ;

- La méthode des réseaux de neurones ;
- La méthode de Naive Bayes ;
- L'Adaptative Boosting (AdaBoost).

2. Cross Validation

Dans l'optique d'estimer notre modèle de façon robuste et de lui permettre de se généraliser plus aisément, nous utiliserons la technique de la validation croisée (ou cross validation en anglais).

Le principe de base de la validation croisée est de diviser l'ensemble de données en plusieurs sous-ensembles, généralement appelés "plis" (ou "folds"), puis d'entraîner et de tester le modèle plusieurs fois. Plus précisément, le fonctionnement se déroule comme suit :

- **Division des données** : L'ensemble de données est divisé en k plis de taille égale (typiquement $k = 5$ ou $k = 10$, mais cela peut varier). Chaque pli sert successivement de jeu de test, tandis que les $k - 1$ autres plis sont combinés pour former le jeu d'entraînement.
- **Entraînement et test** : Le modèle est entraîné sur le jeu d'entraînement, puis testé sur le jeu de test. Cette étape est répétée k fois, chaque pli étant utilisé comme jeu de test exactement une fois.
- **Evaluation** : À chaque itération, les performances du modèle sont mesurées à l'aide d'une métrique spécifique, telle que la précision, le rappel, le F1-score, ou l'exactitude. Ces mesures sont ensuite moyennées pour obtenir une estimation globale de la performance du modèle.

Cette méthode permet d'obtenir une évaluation plus robuste de la capacité de généralisation du modèle, car elle utilise l'ensemble du jeu de données pour l'entraînement et la validation. Elle réduit également le risque de surajustement (overfitting) en exposant le modèle à différentes combinaisons de données d'entraînement et de test.

3. Optimisation des hyperparamètres

Afin d'obtenir le meilleur modèle possible, nous avons effectué une optimisation des hyperparamètres des différents modèles à tester. Cette optimisation consistait à trouver les combinaisons d'hyperparamètres qui maximisaient les performances du modèle sur des données test, tout en évitant le surajustement.

La méthode qui a été utilisée est celle de la GridSearchCV de la librairie Scikit Learn avec comme métrique d'évaluation le F1-score.

CHAPITRE IV : PRÉSENTATION DES RÉSULTATS ET INTERPRÉTATIONS

I. PRESENTATION DES RESULTATS

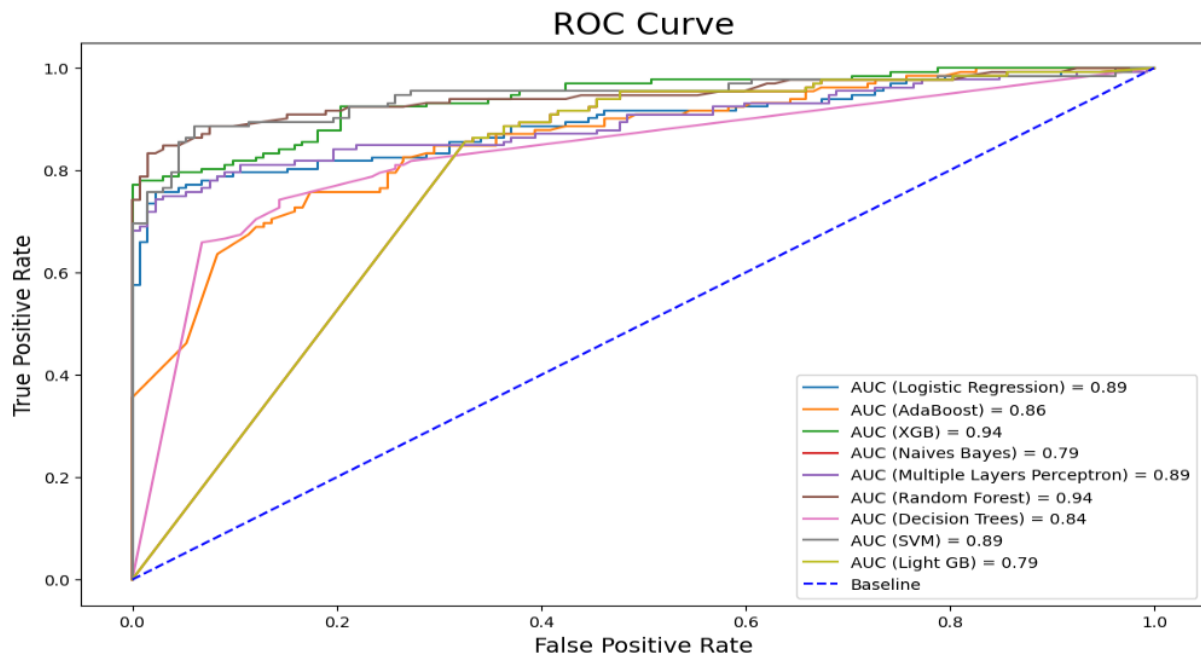
1. Détermination du modèle optimal

A l'issue de la validation croisée avec optimisation des hyperparamètres sur la base d'entraînement, nous avons obtenu les différentes combinaisons d'hyperparamètres qui optimisaient nos différents modèles. Par la suite, nous avons appliqué ces modèles optimisés sur la base de données test pour prédiction et avons obtenu les scores suivants :

Tableau 6 : Résultat des prédictions (source : nos travaux, voir annexe 2)

Model	Exactitude	Précision	Sensibilité	F1 Score
Régression Logistique	0,833	0,836	0,838	0,837
Random Forest	0,900	0,916	0,887	0,901
SVM	0,883	0,915	0,851	0,882
XGBoost	0,888	0,889	0,892	0,890
LightGBM	0,879	0,880	0,885	0,882
Naive Bayes	0,749	0,687	0,939	0,792
Neural Network	0,510	0,510	1,000	0,676
AdaBoost	0,847	0,829	0,885	0,855

Figure 21 : Courbe ROC des modèles (source : nos travaux)



Ainsi le modèle avec les meilleurs scores est incontestablement le Random Forest avec une f1-score de 0,901 et dont les hyperparamètres optimaux sont :

- **max_depth: None** signifie que les nœuds de l'arbre ne sont pas limités par une profondeur maximale, permettant ainsi aux arbres de croître jusqu'à ce que le nombre minimum d'échantillons dans une feuille soit atteint.
- **min_samples_leaf: 1** indique le nombre minimum d'échantillons requis pour être dans une feuille de l'arbre. Ici, il est défini à 1, ce qui signifie qu'une feuille peut contenir un seul échantillon.
- **min_samples_split: 2** spécifie le nombre minimum d'échantillons requis pour diviser un nœud interne. Un nœud ne sera divisé que s'il contient au moins 2 échantillons.
- **n_estimators: 50** représente le nombre d'arbres dans la forêt aléatoire.

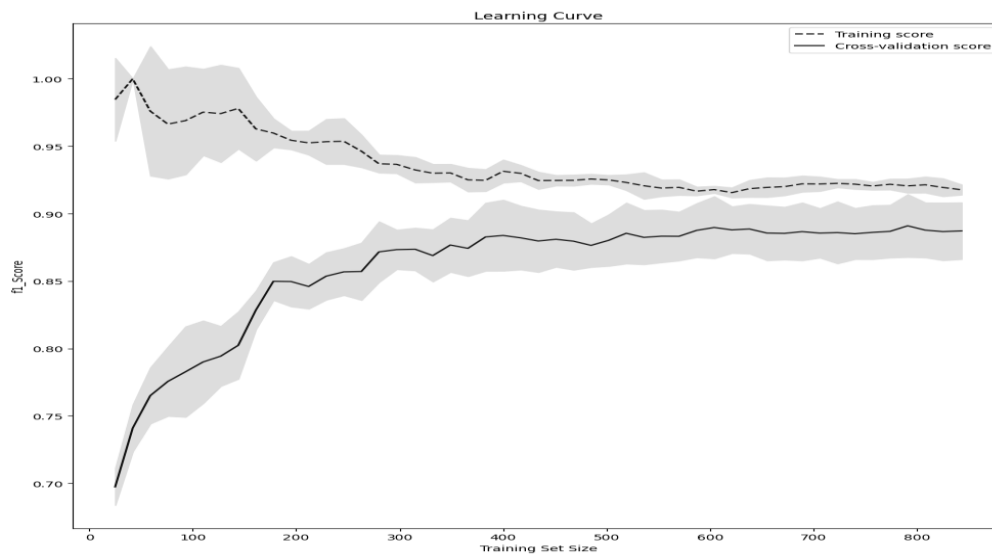
De plus, ce modèle témoigne d'une bonne capacité à pr au regard de la courbe ROC, le modèle Random Forest présente une fois de plus le meilleur score AUC. C'est donc ce modèle qui sera éligible pour la réalisation de notre scoring.

2. Capacité de généralisation du modèle

Il s'agit dans cette partie de vérifier la capacité de notre modèle à généraliser ses bonnes prédictions sur de nouvelles données. En d'autres termes, il s'agit de s'assurer que notre modèle n'est pas en situation de surapprentissage.

Se faisant, l'on a utilisé la courbe d'apprentissage (learning curve) qui nous a permis de comparer l'évolution de la performance du modèle en termes de score sur l'ensemble d'entraînement et l'ensemble de test à mesure que la taille de l'ensemble d'entraînement augmente. Ainsi il en est ressorti qu'à mesure que la taille de l'ensemble d'entraînement augmentait, la performance sur l'ensemble d'entraînement tendait à diminuer, car il devient plus difficile pour le modèle de s'ajuster parfaitement à un grand ensemble de données, tandis que la performance sur l'ensemble de test continuait à s'améliorer, indiquant ainsi une meilleure généralisation (figure 22).

Figure 22 : Courbe d'apprentissage du modèle (source : nos travaux)



II. MISE EN PLACE DU SCORE

En utilisant le modèle retenu c'est-à-dire le random forest aux paramètres optimisés, nous définissons un ensemble de seuils de décision représentant la valeur de la probabilité de défaut au-dessus duquel le crédit est jugé risqué. Par la suite, nous choisissons le seuil qui maximise la F1-score et qui minimise le taux d'erreur. Ainsi il ressort de cette optimisation que le seuil optimal est établi à 0,70, conduisant à un taux d'erreur de 0,03 et un F1-score de 0,88.

Tableau 7 : tableau de recherche du seuil optimal (source : nos travaux)

Seuil	F1 Score	Taux d'erreur
0,1	0,84	0,08
0,2	0,85	0,06
0,3	0,85	0,05
0,4	0,85	0,05
0,5	0,86	0,05
0,6	0,88	0,05
0,7	0,88	0,03
0,8	0,86	0,03
0,9	0,81	0,03

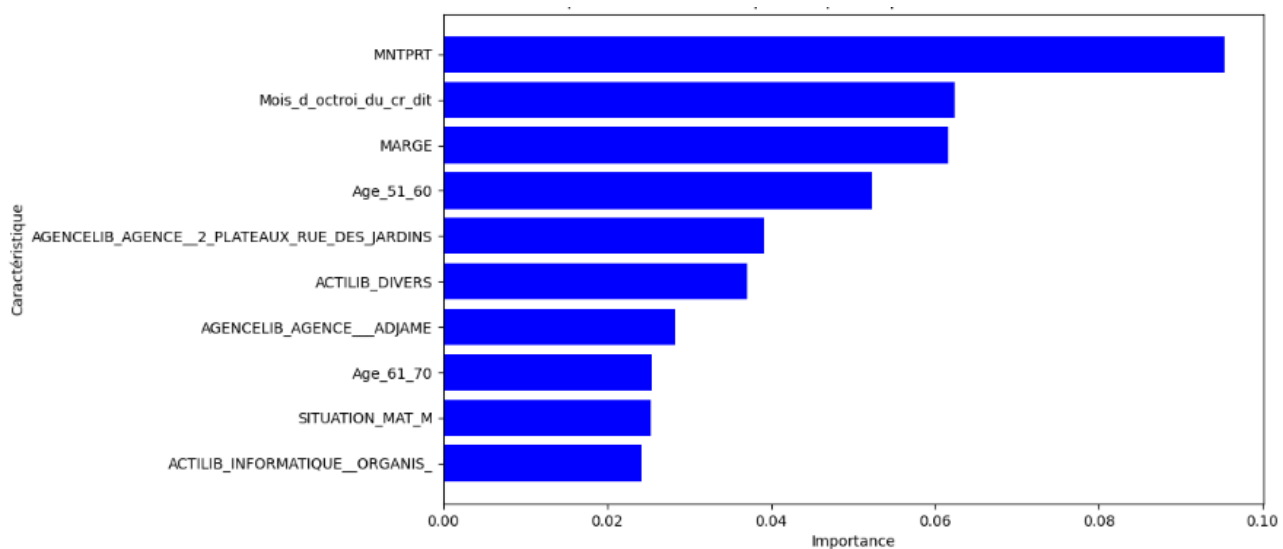
III. INTERPRÉTABILITÉ DU MODÈLE

1. Features Importance

La notion de « feature importance », introduite par Breiman (2001), est un concept fondamental dans la mesure de l'influence d'une variable au sein d'un modèle. Cette mesure

repose sur l'idée d'évaluer comment la perturbation ou la permutation des valeurs d'une variable affecte les performances du modèle, en rompant la relation initiale entre cette variable et la valeur cible (Nexialog, 2022)¹³. En d'autres termes, une variable est considérée comme importante dans un modèle si son altération entraîne une augmentation significative de l'erreur de prédiction. À l'inverse, une variable est jugée peu influente si sa perturbation n'a que peu ou pas d'impact sur la qualité des prédictions du modèle. Ainsi, appliquée à notre étude, le montant du prêt, le mois d'octroi du crédit et le taux d'intérêt du crédit ressortent comme étant les variables les plus importantes pour la stabilisation du taux de bonne prédiction.

Figure 23 : Top 10 des variables les plus importantes du modèle (source : nos travaux)



2. La Shapley Summary Plot

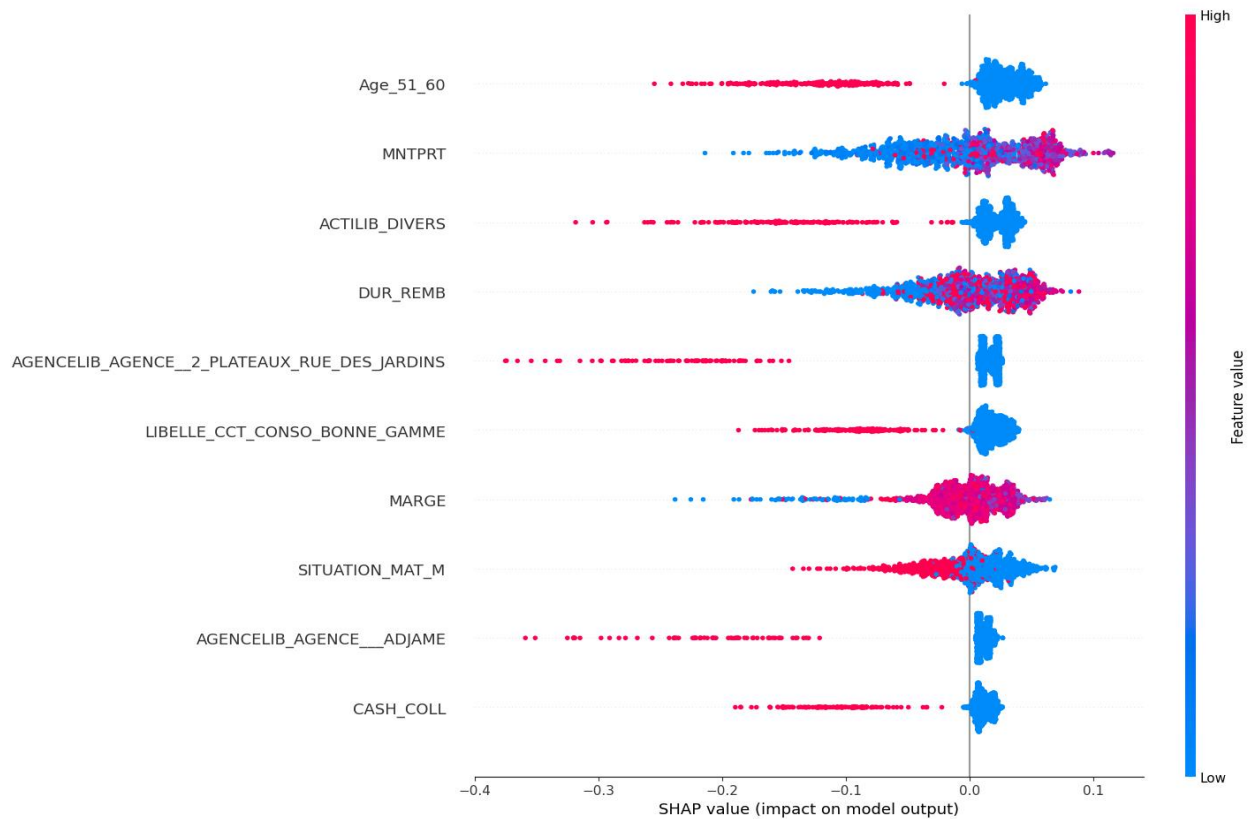
Les valeurs Shapley trouvent leur origine dans la théorie des jeux coopératifs, où elles ont été introduites par Lloyd Shapley en 1953. À l'origine, ces valeurs étaient utilisées pour attribuer une contribution équitable à chaque joueur d'un jeu coopératif, en quantifiant la valeur ajoutée que chaque joueur apportait à la coalition.

Dans le contexte de l'interprétation des modèles de machine learning, les valeurs Shapley ont été adaptées pour mesurer la contribution de chaque caractéristique à une prédiction spécifique. L'idée principale est de répondre à la question : "Comment chaque caractéristique a-t-elle contribué à la différence entre la prédiction finale et la valeur attendue ?"

¹³ <https://www.nexialog.com/wp-content/uploads/2022/03/Interpretabilite-du-ML-Nexialog-Consulting.pdf>

La Shapley Summary Plot est un outil visuel qui synthétise ces valeurs Shapley pour chaque caractéristique. Elle permet de classer les caractéristiques par ordre d'importance décroissante en fonction de leur impact global sur les prédictions du modèle.

Figure 24 : Shapley summary plot du modèle final (source : nos travaux)



Dans ce graphique, l'axe des abscisses représente la valeur SHAP et l'axe des ordonnées, toutes les variables étudiées. Chaque point du graphique correspond à une valeur SHAP pour une prédiction et une caractéristique. Dans notre cas, la caractéristique est le défaut de paiement. La couleur rouge signifie que le risque de défaut est plus élevé et la couleur bleue signifie qu'elle est plus faible. Ainsi la distribution des points rouges et bleus nous donne une idée générale de l'effet des variables sur le risque de défaut de paiement.

Se faisant, il ressort de ce graphique certains constats sur lesquels le modèle s'est basé pour réaliser les prédictions. Ce sont :

- ✓ **Appartenir à la classe d'âge moyenne augmente le risque de crédit.**
- ✓ **Le montant des prêts a un effet positif sur le risque de crédit :** Ce qui laisse comprendre à priori que plus le montant du prêt est élevé, plus le risque de défaut l'est également;

- ✓ **Une durée de remboursement plus longue accroît le risque de crédit.** En d'autres termes, plus la durée de remboursement du prêt augmente, plus celui-ci est risqué.
- ✓ **Les prêts de court terme ayant pour objet la consommation présente un faible risque de crédit.**
- ✓ **Le sexe masculin a un effet négatif sur le risque de crédit.** En d'autres termes le financement des hommes apparaît plus risqué que celui des femmes.

CHAPITRE V : DISCUSSION DES RESULTATS, RECOMMANDATIONS ET PROPOSITION D'UNE APPLICATION WEB

I. DISCUSSION DES RÉSULTATS

1. Age et risque de crédit

Selon les résultats affichés par le modèle, les adultes de la classe d'âge moyenne sont généralement considérées comme étant plus risquées que les personnes des autres classes d'âge. En effet, cela pourrait par le fait que cette tranche d'âge est généralement associée à des charges financières plus lourdes, telles que des prêts hypothécaires, des prêts étudiants, et des responsabilités familiales, y compris l'éducation des enfants et les dépenses liées au logement. Ces engagements financiers accrus peuvent augmenter la pression sur les ressources financières de ces emprunteurs et potentiellement accroître le risque de défaut de paiement. De plus, la stabilité de l'emploi peut varier davantage au sein de cette tranche d'âge, ce qui peut influencer la capacité à honorer les engagements financiers.

Cette conclusion s'aligne avec les résultats de Debbaut et al. (2013) qui se sont penchés sur la relation entre l'âge des emprunteurs et le défaut de paiement sur les cartes de crédit. Ils ont constaté que, de manière générale, le risque de défaut de paiement sur les cartes de crédit est plus élevé pour les emprunteurs d'âge moyen et moins élevé pour les jeunes et les personnes âgées.

2. Montant des prêts et risque de crédit

Les prêts dont les montants sont élevés sont naturellement considérés comme étant plus risqués que les petits prêts. Cela s'explique par le fait que des montants de crédit plus élevés impliquent généralement des engagements financiers plus importants pour les emprunteurs. Cela signifie que les emprunteurs doivent gérer des remboursements plus importants, ce qui peut être plus difficile s'ils rencontrent des problèmes financiers inattendus tels que la perte d'emploi, des dépenses médicales imprévues, ou d'autres crises financières. De plus, les montants de crédit élevés sont souvent associés à des prêts à plus long terme, ce qui augmente la période pendant laquelle l'emprunteur est exposé à des risques potentiels.

3. Durée de remboursement et risque de crédit

Les résultats du modèle ont montré qu'une durée de remboursement accroît le risque de crédit. Cela se justifie par le fait qu'une durée de remboursement plus longue signifie que l'emprunteur est engagé financièrement sur une période prolongée. Cela accroît la probabilité que des événements imprévus, tels que des perturbations économiques, des changements de revenus ou des dépenses inattendues, puissent survenir pendant la durée du prêt, augmentant ainsi le risque de défaut.

De plus, avec des prêts à long terme, les taux d'intérêt ont tendance à être plus élevés, ce qui signifie que l'emprunteur paiera davantage d'intérêts sur la durée de vie du prêt. Cela peut rendre le remboursement plus difficile, en particulier si l'emprunteur est confronté à des variations de taux d'intérêt ou à une évolution de sa situation financière.

En outre, la durée plus longue d'un prêt peut également augmenter le risque de changements dans la situation personnelle de l'emprunteur, tels que des changements d'emploi, des déménagements ou des circonstances familiales, qui pourraient affecter sa capacité à honorer ses engagements financiers.

4. Taux d'intérêt et risque de crédit

Le taux d'intérêt a un effet positif sur le risque de crédit. En effet, un taux d'intérêt plus élevé rend le remboursement du prêt plus coûteux, ce qui peut augmenter le risque de défaut de paiement. Cependant, l'effet du taux d'intérêt sur le risque de crédit est généralement limité. En effet, les emprunteurs sont généralement conscients du coût du crédit et ils l'intègrent dans leur décision de contracter un prêt. Dès lors, la justification de la relation positive entre taux d'intérêt et risque de défaut serait directement déduite de celle entre le montant de crédit et le risque de défaut.

5. Sexe et risque de crédit

Selon les résultats du modèle, les hommes sont considérés comme étant plus risqués que les femmes. Toutefois, ces résultats pourraient être biaisés par le fait que notre population soit constituée en grande majorité d'hommes et de très peu de femmes.

II. LIMITES DE L'ÉTUDE

Les principales limites de l'étude sont :

- **La non-exhaustivité des données** : N'ayant travaillé que sur un échantillon de 711 prêts, étant très peu représentés pour les années antérieures à 2016, notre étude n'est pas complètement généralisable. Toutefois, elle peut être améliorée avec un renforcement de la base de données.
- **La non prise en compte des variables sur l'historique financier des emprunteurs** : En effet, selon la littérature, les variables portant sur l'historique financier des emprunteurs telles que l'historique de crédit, le niveau d'endettement, les revenus de l'emprunteur ou même le paiement des factures d'électricité ressortaient comme étant assez pertinentes

pour une étude de ce genre. Toutefois, nous n'avons pas pu être en mesure de collecter ces informations dans notre étude.

III. RECOMMANDATIONS

Au regard de ces résultats, nous proposons les recommandations suivantes :

- ✓ **Optimisation de Modèles Prédicatifs** : La BBG CI pourrait intégrer les modèles prédictifs élaborés dans cette étude dans son processus de prise de décision en matière de prêts. Ces modèles peuvent aider à évaluer rapidement le risque de crédit des demandeurs et à prendre des décisions plus éclairées.
- ✓ **Amélioration de la collecte des données** : La BBG CI devrait mettre en place des systèmes de collecte de données accessibles et efficaces afin de pouvoir enrichir sa base de données sur ses clients pour faciliter la réalisation d'éventuelles études au sein de la banque.
- ✓ **Segmentation de la Clientèle** : La banque devrait envisager une segmentation plus fine de sa clientèle en fonction des facteurs identifiés comme influençant le risque de crédit. Ainsi, elle pourrait élaborer des offres de prêts spécifiquement adaptées aux besoins et au profil de différents groupes de clients.
- ✓ **Optimisation des Taux d'Intérêt** : Étant donné que les taux d'intérêt ont été identifiés comme un facteur contribuant au risque de défaut, la banque pourrait envisager une stratégie de tarification plus dynamique. Elle pourrait offrir des taux d'intérêt plus compétitifs aux emprunteurs présentant un profil de risque plus faible, tout en ajustant les taux pour ceux présentant un risque plus élevé. Cela permettrait de mieux aligner les revenus d'intérêts sur le risque réel.

IV. APPLICATION WEB DE CREDIT SCORING

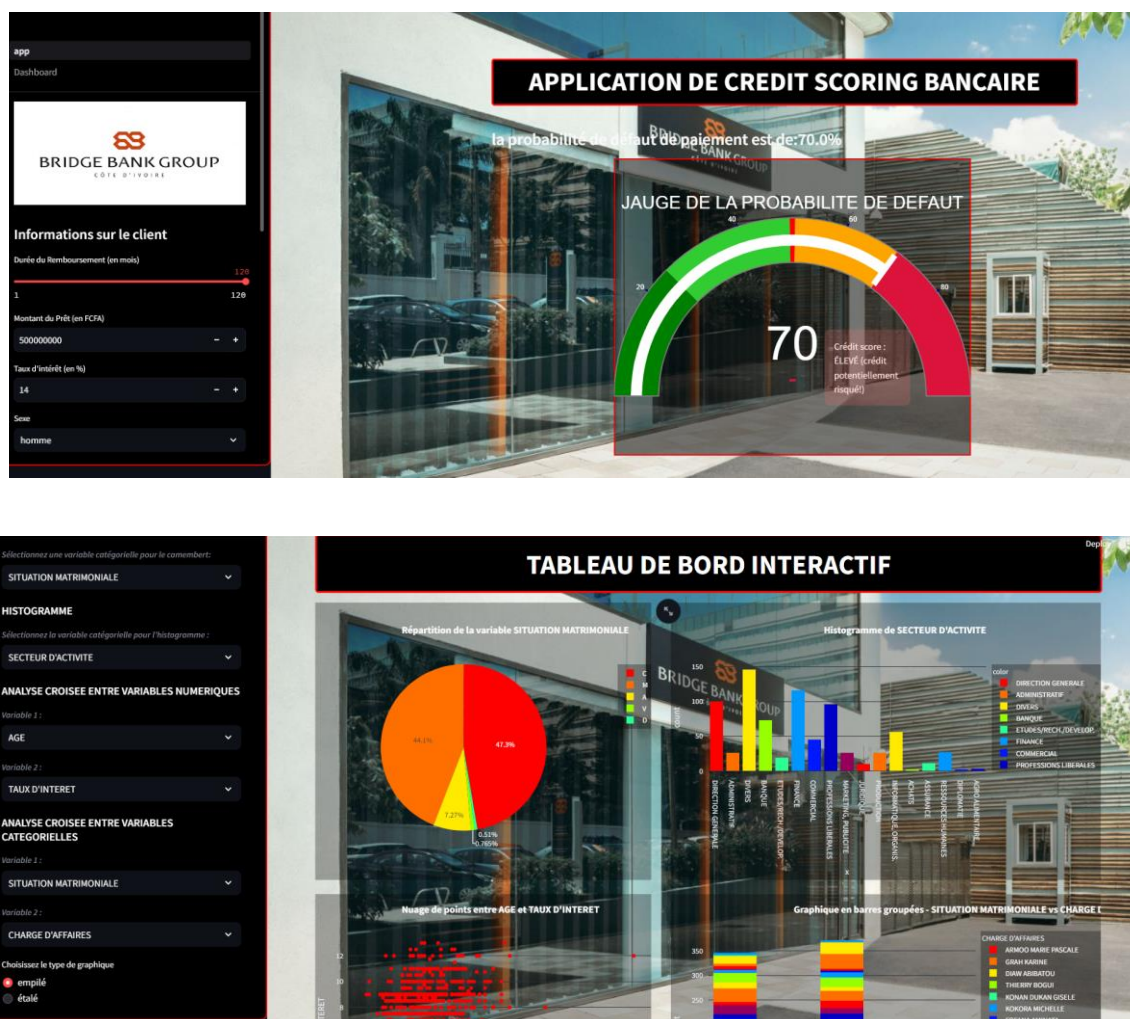
L'application a été construite à l'aide de la librairie Streamlit, qui permet de créer des applications web interactives avec du code Python. Cette librairie offre une expérience de développement simplifiée, permettant aux utilisateurs de se concentrer sur la logique métier plutôt que sur la gestion des détails d'implémentation d'une interface utilisateur. Ainsi l'application est constituée des fonctionnalités suivantes :

- ✚ **Entrée Utilisateur Intuitive** : L'application propose une interface utilisateur intuitive permettant aux utilisateurs de spécifier les détails du prêt, tels que la durée de

remboursement, le montant du prêt, la marge, le sexe, la situation matrimoniale, l'activité professionnelle, le libellé, l'agence, l'âge et les garanties.

- ✚ **Prédiction Instantanée** : Une fois que l'utilisateur a fourni les informations nécessaires, le modèle de Credit Scoring prédit instantanément la probabilité de défaut de remboursement. Cette prévision est ensuite présentée à l'utilisateur de manière claire et lisible à l'aide d'une jauge semi-circulaire.
- ✚ **Personnalisation des Garanties** : L'utilisateur peut sélectionner plusieurs garanties, offrant ainsi une personnalisation flexible et une prise en compte des scénarios divers.
- ✚ **Visualisation des Données d'Entrée** : L'application dispose également d'une fonctionnalité permettant de visualiser graphiquement la base de données de laquelle est issue le modèle de scoring.

Figure 25 : Interface graphique de l'application de crédit scoring (source : nos travaux)



CONCLUSION GENERALE

Pour sonner le glas de notre étude, retenons que le risque de crédit demeure une préoccupation majeure pour les banques et les institutions financières, c'est pourquoi beaucoup d'efforts sont fournis afin de trouver les moyens les plus efficaces pour maîtriser et atténuer le risque de crédit. Les accords de Bâle II ont donné la latitude aux banques de disposer de leurs propres systèmes d'estimation de la probabilité de défaut de leurs portefeuilles. Dans ce contexte, l'objectif principale de cette étude était d'établir un modèle de scoring basé sur des modèles de machine learning permettant de prédire la probabilité de défaut des particuliers qui contractent un prêt auprès de la Bridge Bank Group CI, tout en mettant en lumière les variables qui influencent le plus la probabilité de défaut d'un particulier.

Nous avons commencé notre étude par la clarification des notions inhérentes aux métiers de la banque, au risque de crédit, ainsi qu'au machine learning. Par la suite, nous avons procédé à la définition de la notion de défaut dans le cadre du crédit, ainsi que de celle qui a été considérée dans notre étude. Après quoi, nous nous sommes attelés à parcourir de façon globale la littérature sur les méthodes d'évaluation du risque de crédit avec un focus sur les travaux de credit scoring qui étaient basés sur les méthodes de machine learning afin de se forger une opinion solide sur le problème.

Ainsi, il est ressorti que les modèles de scoring crédit sont généralement fondés sur les caractéristiques démographiques et financières de l'emprunteur. Se faisant, nous nous sommes munis d'une base de données prenant en compte les caractéristiques démographiques de l'emprunteur et financières du prêt. Après un traitement rigoureux de la base et une exploration en profondeur des variables qui la compose suivant la méthodologie proposée par Dastile (2020), nous avons obtenu des scores assez satisfaisants avec une F-1 score de 89%. En outre, à l'aide des modèles d'interprétabilité, nous avons découvert que l'âge, le montant des prêts, la durée de remboursement, le taux d'intérêt et le sexe étaient des facteurs significatifs dans l'évaluation du risque de crédit dans notre population d'étude. Plus spécifiquement le modèle a montré que :

- ❖ Les adultes d'âge moyen présentaient plus de risques de défaut de paiement ;
- ❖ Les prêts élevés et les durées de remboursement plus longues augmentaient le risque de défaut de paiement.
- ❖ Les taux d'intérêt plus élevés pouvaient également accroître ce risque, bien que de manière limitée.

- ❖ Et enfin les hommes apparaissaient bien plus susceptibles d'être en défaut de paiement, bien que ce dernier résultat ait été fortement influencé par la composition de la population étudiée.

In fine, nous pouvons affirmer que nos objectifs pour cette étude ont été atteints. Toutefois cette étude pourrait subir des améliorations par exemple en incluant des variables sur les antécédents financiers des emprunteurs ou en débouchant sur la conception d'une application permettant à la fois de prédire le risque de défaut des demandes de crédit quotidienne et de renforcer la base de données d'entraînement avec l'issue de ces différentes demandes.

REFERENCES

BIBLIOGRAPHIQUES

- Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature: credit scoring, techniques & evaluation criteria: a literature review. *Intelligent Systems in Accounting, Finance and Management*, 18(2–3), 59–88. <https://doi.org/10.1002/isaf.325>
- Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications*, 35(3), 1275–1292. <https://doi.org/10.1016/j.eswa.2007.08.030>
- Al Amari, A. (2002). *The credit evaluation process and the role of credit scoring: A case study of Qatar*. Ph.D. Thesis, University College Dublin
- Azencott, C. (2018). *Introduction au Machine Learning*, ed. Dunod
- Bailey, M. (2001). *Credit scoring: The principles and practicalities*. Kingswood, Bristol: White Box Publishing.
- Bailey, M. (2004). *Consumer credit quality: Underwriting, scoring, fraud prevention and collections*. Kingswood, Bristol: White Box Publishing.
- Bardos, M. (2001), *Analyse discriminante, application au risque et scoring financier*, ed. Dunod.
- Bardos, M. et Zhu, W. (1997), *Comparaison de l'analyse discriminante linéaire et réseaux de neurones, applications à la détection de défaillance des entreprises*, *Revue de statistique appliquée*, XLV4, p. 65-92.
- Breiman, L. (2001). "Random Forests." *Machine Learning* 45 (1). Springer: 5–32
- Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446–3453. <https://doi.org/10.1016/j.eswa.2011.09.033>
- Brunel, V., (2019), *gestion des risques et risque de crédit*, <http://leguidedurentier.free.fr/documents/risquescredit.pdf>
- Caire et Kossmann, in Ntoutoume Obiang-Ndong F., (2006). *Scoring credit: une application comparative de la régression logistique et des réseaux de neurones*, Université Cheikh Anta Diop (UCAD) - Master Méthodes Statistiques et Économétries.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Oversampling TEchnique. *Journal of Artificial Intelligence Research*, 16:321-357.
- Crouhy, M. (2000). *La gestion du risque de crédit et la stabilité du système financier international*. École des hautes études commerciales.
- Chandler G., Coffman J., (1979). A comparative analysis of empirical vs judgmental credit evaluation. *Journal of Retail Banking*, pp. 12-26
- Crook, J. (1996). Credit Scoring: An Overview. *European Journal of Operational Research*, 95, 24-37.
- Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91, 106263. <https://doi.org/10.1016/j.asoc.2020.106263>
- Debbaut, P., Ghent, A. C., Kudlyak, M., & Romero, J., (2013), *How Risky Are Young Borrowers?*, "Federal Reserve Bank of Richmond", EB13-12.
- Desai, Vijay S. & Crook, Jonathan N. & Overstreet, George A., (1996). "[A comparison of neural networks and linear scoring models in the credit union environment](#)," *European Journal of Operational Research*, Elsevier, vol. 95(1), pages 24-37, November.
- Eisenbeis, Robert A., (1978). "[Problems in applying discriminant analysis in credit scoring models](#)," *Journal of Banking & Finance*, Elsevier, vol. 2(3), pages 205-219, October.
- Eisenbeis, R. A., (1996), *Recent developments in the application of credit scoring techniques to the evaluation of commercial loans*, *IMA J. Math. Appl. Business Indust.*, 7, 271–290.
- Feldman R., (1997), "Small Business Loans, Small Banks and Big Change in Technology Called Credit Scoring", *The Region*, September, pp. 19-25, Federal Reserve Bank of Minneapolis.

- Gool, J V., Verbeke, W., Sercu, P., & Baensens, B. (2011). Credit scoring for microfinance: is it worth it?. *International journal of finance & economics/International journal of finance and economics*, 17(2), 103-123. <https://doi.org/10.1002/ijfe.444>
- Gouiaa, R., & Ouedraogo, M. (2022). L'impact de la gestion du risque de crédit sur la performance des banques commerciales canadiennes. *Revue Organisations & territoires*, 31(1), 69-91. <https://doi.org/10.1522/revueot.v31n1.1449>
- Grablowsky, Bernie J. ; Talley, Wayne K., (1981) *Journal of economics & business*. - Amsterdam [u.a.] : Elsevier, ISSN 0148-6195, ZDB-ID 716757-X. - Vol. 33.1981, 3, p. 254-261
- Greene W., (1998), Sample selection in credit-scoring models1Working Paper No. 194 presented at The Japan–U.S. Technical Symposium in Japan, November 13, 1995.1, Japan and the World Economy, Volume 10, Issue 3, 1998, Pages 299-316, ISSN 0922-1425, [https://doi.org/10.1016/S0922-1425\(98\)00030-9](https://doi.org/10.1016/S0922-1425(98)00030-9).
- Heffernan, S. (2005),. *Modern Banking*. Chichester, West Sussex: John Wiley & Sons, Inc
- Hjelkrem, L. O., De Lange, P. E., & Nettet, E. (2022),. The Value of Open Banking Data for Application Credit Scoring: Case Study of a Norwegian Bank. *Journal of Risk and Financial Management*, 15(12), 597. <https://doi.org/10.3390/jrfm15120597>
- Hurlin, C., & Pérignon, C. (2020),. Machine learning et nouvelles sources de données pour le scoring de crédit : *Revue d'économie financière*, N° 135(3), 21-50. <https://doi.org/10.3917/ecofi.135.0021>
- Johnson, K. N. & Stafford, M. (2019), Examining the Use of Alternative Data in Underwriting and Credit Scoring to Expand Access to Credit. Tulane Public Law Research Paper No. 19-7. <http://dx.doi.org/10.2139/ssrn.3481102>
- Kouassi, G., S., (2021), Scoring du risque de crédit par la méthode de Machine Learning. [Mémoire non publié], Institut National Polytechnique Félix Houphouët Boigny de Yamoussoukro.
- Lotfi, S., & Mesk, H. (2020). Prédiction du risque de crédit : Étude comparative des techniques de Scoring Prediction of credit risk: Comparative study of scoring techniques. 1(2).
- Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2), 117-134. <https://doi.org/10.1016/j.sorms.2016.10.001>
- Łuczak, A., Ganzha, M., & Paprzycki, M. (2021). Probability of Loan Default—Applying Data Analytics to Financial Credit Risk Prediction. In M. Paprzycki, S. M. Thampi, S. Mitra, L. Trajkovic, & E.-S. M. El-Alfy (Eds.), *Intelligent Systems, Technologies and Applications* (Vol. 1353, pp. 1-16). Springer Singapore. https://doi.org/10.1007/978-981-16-0730-1_1
- Petrescu, A., M., Évaluation des risques des prêts bancaires, [Mémoire non publié], https://chairegestiondesrisques.hec.ca/wp-content/uploads/pdf/equipe/Memoire_Madalina_Petrescu.pdf
- Schreiner, M. et Nagarajan, G., (1998), « Predicting the Creditworthiness with publically observable characteristics: Evidence with ASCRAs and ROSCAs in the Gambia. Savings and Development »
- Schreiner M. (2004). Scoring at Prizma: How to Prepare, report to Prizma and CGAP, Mostar, Bosnia-Herzegovina.
- Schumpeter J.A (1954) : « Histoire de l'Analyse Economique », Paris, Gallimard, 3 vol., 1983.
- Soares, J O., Pina, J P., Ribeiro, M S., & Lopes, M C. (2011). Quantitative vs. Qualitative Criteria for Credit Risk Assessment. *Social Science Research Network*. <https://doi.org/10.2139/ssrn.2012443>
- Sullivan, A. C. (1981). Consumer finance. In E. I. Altman (Ed.), *Financial handbook* (pp. 9.3-9.27). New York : John Wiley & Sons
- Thomas, L., Crook, J., & Edelman, D. (2002), "Credit scoring and its applications", SIAM, 2002.

- Thomas, L., Crook, J., & Edelman, D. (2017). *Credit Scoring and its Applications*. (2nd ed.) SIAM, 2017.
- Viganò, L., (1993). *A credit scoring model for development banks: an African case study*, *Savings and Development*, N4-XVII.
- Wang, X., Xu, M., & Pusatli, Ö. T. (2015). *A Survey of Applying Machine Learning Techniques for Credit Rating: Existing Models and Open Issues*. In S. Arik, T. Huang, W. K. Lai, & Q. Liu (Eds.), *Neural Information Processing* (Vol. 9490, pp. 122–132). Springer International Publishing. https://doi.org/10.1007/978-3-319-26535-3_15
- Yang, L., (2001). *“New issues in credit scoring application, Working Paper No 16/2001, Institut Fur Wirtschaftsinformatik*

ANNEXES

Annexe I : Code source du traitement des données

```
import openpyxl
data=openpyxl.load_workbook(r"C:\Users\DELL I5\Documents\RECOVERED
FILES\TELECHARGEMENTS\BD finale.xlsx")
datas=data.active
donnees = []

for ligne in datas.iter_rows(values_only=True):
    donnees.append(list(ligne))
import pandas as pd
df = pd.DataFrame(donnees)

en_tetes = donnees[0]
donnees = donnees[1:]

# Créer un DataFrame avec les données, en spécifiant les en-têtes et l'index
df = pd.DataFrame(donnees, columns=en_tetes)
df.set_index(en_tetes[0], inplace=True) # En faire l'index
df['AGE'] = df['AGE'].astype(int)
df['MNTprt'] = df['MNTprt'].astype(int)
df["MARGE"].replace('null', 0, inplace=True)
df["MARGE"]=df["MARGE"].astype(float)
# Convertir la colonne "Date de déblocage" en format de date
df['DATDEP'] = pd.to_datetime(df['DATDEP'], errors='coerce' )
df["DERNDAT"] = pd.to_datetime(df["DERNDAT"], errors='coerce')

#Création des variables dummies
variables_catégorielles = ['SEXE','SITUATION_MAT',
'ACTILIB','LIBELLE','AGENCELIB']
df = pd.get_dummies(df, columns=variables_catégorielles, drop_first=True)

#Normalisation
from sklearn.preprocessing import RobustScaler
scaler = RobustScaler()
df[["DUR_REMB", "MNTprt", "MARGE"]] = scaler.fit_transform(df[["DUR_REMB",
"MNTprt", "MARGE"]])

#Discrétisation de la variable âge
age_bins = [19, 29, 39, 49, 59, 69, np.inf]
age_labels = ['20-29', '30-39', '40-49', '50-59', '60-69', '70+']

df['AgeGroup'] = pd.cut(df['AGE'], bins=age_bins, labels=age_labels,
right=False)

# Utilisez pd.get_dummies pour créer les variables dummies
df = pd.get_dummies(df, columns=['AgeGroup'], prefix='Age')

# Supprimez la colonne d'origine AGE si nécessaire
```

```

df.drop(columns=['AGE'], inplace=True)

#Garanties
df["TXT1"] = df["TXT1"].apply(lambda x: x.replace("\n", " "))
stopWords=['xof',"nothing", "de", "m","a","sur"]

# Création d'une fonction pour supprimer les stopWords
def no_stop_word(string, stopWords):A

    """
    Supprime les stop words d'un texte.

    Paramètres
    -----

    string : chaine de caractère.

    stopWords : liste de mots à exclure.
    """
    string_split=string.lower().split(" ")
    str1=[mot for mot in string_split if mot not in stopWords]
    string =" ".join(str1)
    return string

def stem_cleaner(pandasSeries, stopWords):

    print("#### Nettoyage en cours ####") # Mettre des print vous permet de
    comprendre où votre code rencontre des problèmes en cas de bug

    # confirmation que chaque article est bien de type str
    pandasSeries = pandasSeries.apply(lambda x : str(x))

    ### COMMENCEZ A CODER ICI! remplacer les 'None' par votre code ###

    # Passage en minuscule
    print("... Passage en minuscule")
    pandasSeries = pandasSeries.apply(lambda x : x.lower())

    # Suppression des accents
    print("... Suppression des accents")
    pandasSeries = pandasSeries.apply(lambda x : unicode(x))

    # Suppression des caractères spéciaux et numériques
    print("... Suppression des caractères spéciaux et numériques")
    pandasSeries = pandasSeries.apply(lambda x : re.sub(r"[^a-z]+", ' ', x))

```

```

# Suppression des stop words
print("... Suppression des stop words")
pandasSeries = pandasSeries.apply(Lambda x : no_stop_word(x, stopWords))
return pandasSeries

df['GARANTIES_c'] = stem_cleaner(df['TXT1'], stopWords)

df[['TXT1', 'GARANTIES_c']]

# Définissez une fonction pour créer les variables dichotomiques en fonction
des critères
def create_dichotomous_variables(text):
    text = text.lower() # Convertir le texte en minuscules pour une
correspondance insensible à la casse

    domiciliation_revenus = int("domiciliation" in text)
    cash_coll = int("cash" in text)
    dat = int("dat" in text)
    garantie_progressive = int("epargne" in text or "odvp" in text or
"progressif" in text or "virement" in text)
    cash_coll_multirisque= int("assurance" in text or "deces" in text or
"invalidite" in text or "assur" in text or "multirisque" in text)
    billet_a_ordre= int("bao" in text or "billet" in text)
    garantie_hypothèque = int("hypothèque" in text)
    clean_var = int("clean" in text)

    return pd.Series({
        "domiciliation des revenus": domiciliation_revenus,
        "CASH COLL": cash_coll,
        "DAT": dat,
        "garantie hypothécaire": garantie_hypothèque,
        "garantie progressive": garantie_progressive,
        "assurance multirisque": cash_coll_multirisque,
        "Billet à Ordre": billet_a_ordre,
        "clean": clean_var
    })

# Appliquez la fonction pour créer les variables dichotomiques
dichotomous_vars = df["GARANTIES_c"].apply(create_dichotomous_variables)

# Ajoutez les nouvelles variables au DataFrame original
df = pd.concat([df, dichotomous_vars], axis=1)

# Créez une nouvelle colonne pour le mois
df["Mois_d'octroi_du crédit"] = df["DATDEP"].dt.month

# Créez une nouvelle colonne pour l'année

```

```

df["Année d'octroi du crédit"] = df["DATDEP"].dt.year

# Sélectionner les colonnes booléennes
colonnes_bool = df.select_dtypes(include=bool).columns

# Convertir les colonnes booléennes en entiers (0 et 1)
df[colonnes_bool] = df[colonnes_bool].astype(int)

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score

# Supprimer les variables inutiles
variables_inutiles = ['NOM', 'PROFESSION', 'RM', 'TXT1', 'DERNDAT',
'DATDEP', 'GARANTIES_c']
df.drop(columns=variables_inutiles, inplace=True)

# Séparation des données en ensembles d'entraînement et de test
X = df.drop(columns=['STATUT'])
y = df["STATUT"]

from imblearn.over_sampling import SMOTE
# Créez une instance de SMOTE avec les paramètres souhaités
smote = SMOTE(sampling_strategy='auto', random_state=42)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42, stratify='STATUTS')

X_train, y_train = smote.fit_resample(X_train, y_train)

```

Annexe II : Code source de la modélisation

```

models = {
    'Logistic Regression': LogisticRegression(),
    'Random Forest': RandomForestClassifier(),
    'SVM': SVC(),
    'XGBoost': XGBClassifier(),
    'LightGBM': LGBMClassifier(),
    'Naive Bayes': GaussianNB(),
    'Neural Network': MLPClassifier(),
    'AdaBoost': AdaBoostClassifier()
}

```

```

# Définir les métriques
scoring = {
    'accuracy': make_scorer(accuracy_score),
    'precision': make_scorer(precision_score),
    'recall': make_scorer(recall_score),
    'f1': make_scorer(f1_score)
}

# Définir les grilles de paramètres pour chaque modèle
param_grids = {
    'Logistic Regression': {'C': [0.001, 0.01, 0.1, 1, 10, 100], 'penalty':
['l1', 'l2']},
    'Random Forest': {'n_estimators': [50, 100, 200], 'max_depth': [None, 10,
20, 30], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4]},
    'SVM': {'C': [0.1, 1, 10], 'kernel': ['linear', 'rbf'], 'gamma': ['scale',
'auto']},
    'XGBoost': {'learning_rate': [0.01, 0.1, 0.2], 'n_estimators': [50, 100,
200], 'max_depth': [3, 5, 7], 'subsample': [0.8, 1.0]},
    'LightGBM': {'learning_rate': [0.01, 0.1, 0.2], 'n_estimators': [50, 100,
200], 'max_depth': [3, 5, 7], 'subsample': [0.8, 1.0]},
    'Naive Bayes': {},
    'Neural Network': {'hidden_layer_sizes': [(50, 50), (100, 100), (50, 100,
50)], 'activation': ['relu', 'tanh'], 'alpha': [0.0001, 0.001, 0.01]},
    'AdaBoost': {'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1,
0.2]}
}

# Initialiser un DataFrame pour stocker les résultats
results = pd.DataFrame(columns=['Model', 'Optimal F1 Score', 'Optimal
Precision', 'Optimal Recall', 'Optimal Accuracy', 'Optimal Parameters'])

# Boucle sur chaque modèle et effectuer une recherche sur grille
for model_name, model in models.items():
    print(f"Training {model_name}...")

    # Définir la grille de paramètres
    param_grid = param_grids[model_name]

    # Initialiser GridSearchCV
    grid_search = GridSearchCV(model, param_grid, scoring=scoring, cv=5,
refit='f1', n_jobs=-1)

    # Effectuer la recherche sur grille
    grid_search.fit(X_train, y_train)

    # Obtenir les résultats optimaux
    optimal_f1 = grid_search.best_score_

```

```

    optimal_precision =
grid_search.cv_results_['mean_test_precision'][grid_search.best_index_]
    optimal_recall =
grid_search.cv_results_['mean_test_recall'][grid_search.best_index_]
    optimal_accuracy =
grid_search.cv_results_['mean_test_accuracy'][grid_search.best_index_]
    optimal_params = grid_search.best_params_

# Ajouter les résultats au DataFrame
results = pd.concat([results, pd.DataFrame({'Model': [model_name],
'Optimal F1 Score': [optimal_f1], 'Optimal Precision': [optimal_precision],
                                         'Optimal Recall':
[optimal_recall], 'Optimal Accuracy': [optimal_accuracy], 'Optimal
Parameters': [optimal_params]})])

# Afficher les résultats finaux
results

```

Annexe III : Code source de l'application

```

i
ii #Définition des fonctions
iii
iv # Appliquer le RobustScaler aux nouvelles données
v def apply_robust_scaler(value):
vi     params = {
vii         'median': df['MNTPRT'].median(),
viii        'q1': df['MNTPRT'].quantile(0.25),
ix         'q3': df['MNTPRT'].quantile(0.75)
x     }
xi     scaled_value = (value - params['median']) / (params['q3'] -
xii     params['q1'])
xiii     return scaled_value
xiv
xv def adjust_categorical_values1(new_values, saved_categories):
xvi     adjusted_values = [1 if value == new_values else 0 for value in
xvii     saved_categories]
xviii     return adjusted_values
xix
xx def reverse_onehot_encode1(new_values, categories_mapping):
xxi     reversed_values = []
xxii     for variable, categories in categories_mapping.items():
xxiii        adjusted_values =
adjust_categorical_values1(new_values[variable], categories)
reversed_values.extend(adjusted_values)
return reversed_values

```

```

def catégorielle(values_f):
    categories_mapping = {
        'SEXE': ['homme'],
        'SITUATION_MAT': ['célibataire', 'marié(e)',
        'divorcé(e)', 'veuf(ve)'],
        'ACTILIB': ['ADMINISTRATIF', 'AGRO
ALIMENTAIRE', 'ASSURANCE', 'BANQUE', 'COMMERCIAL', 'DIPLOMATIE', 'DIRECTION
GENERALE', 'DIVERS', 'ETUDES/RECH./DEVELOP.', 'FINANCE', 'INFORMATIQUE,
ORGANIS.', 'JURIDIQUE', 'MARKETING, PUBLICITE', 'PRODUCTION', 'PROFESSIONS
LIBERALES', 'RESSOURCES HUMAINES'],
        'LIBELLE': ['BRIDGE PRET RELAIS', 'CCT AUTRES CRD', 'CCT
CONSO', 'CCT CONSO BONNE GAMME', 'CCT CONSO PER AUTRE', 'CCT HORS PP', 'CCT
RESTRUCTURES', 'CCT SCOLAIRE', 'CMT AUTRES', 'CMT CONSO', 'CMT CONSO BONNE
GAMME', 'CMT CONSO PERSO', 'CMT HAB BONNE GAMME', 'CMT HAB
PATRIMONIALE', 'CMT HORS PP', 'CMT RESTRUCTURES'],
        'AGENCELIB': ['AGENCE - ADJAME', 'AGENCE - AG PRINCIPALE', 'AGENCE
- COCODY', 'AGENCE - II PLATEAUX 8IEME TRANCHE', 'AGENCE - MARCORY
RESIDENTIEL', 'AGENCE - PLATEAU SEEN HOTEL', 'AGENCE - RIVIERA 3', 'AGENCE
- RIVIERA GOLF', 'AGENCE - SAN PEDRO', 'AGENCE - TREICHVILLE ZONE
3', 'AGENCE - ZONE 4 DR BLANCHARD', 'AGENCE- 2 PLATEAUX RUE DES
JARDINS', 'AGENCE- DEUX PLATEAUX LATRILLE', 'AGENCE-TREICHVILLE NANAN
YAMOUSO']
    }
    return reverse_onehot_encode1(values_f, categories_mapping)

def adjust_categorical_values2(new_values, saved_categories):
    adjusted_values = [1 if value in new_values else 0 for value in
saved_categories]
    return adjusted_values

def reverse_onehot_encode2(new_values, categories_mapping):
    reversed_values = []
    for variable, categories in categories_mapping.items():
        adjusted_values =
adjust_categorical_values2(new_values[variable], categories)
        reversed_values.extend(adjusted_values)
    return reversed_values

categories_mapping_garanties = {
    'domiciliation des revenus': 0,
    'CASH COLL': 1,
    'DAT': 2,
    'garantie hypothécaire': 3,
    'garantie progressive': 4,
    'assurance multirisque': 5,
    'Billet à Ordre': 6,
    'clean': 7
}

```

```

lv  def modif_garanties(new_values_garanties):
lvi  # Exemple d'utilisation pour la variable catégorielle 'garanties'
lvii
lviii  # Ajuster les valeurs catégorielles pour 'garanties'
lix  adjusted_values_garanties = reverse_onehot_encode2({'garanties':
lxx  new_values_garanties}, {'garanties': categories_mapping_garanties})
lxxi
lxxii  return adjusted_values_garanties
lxxiii
lxxiv  def map_age_interval_vector(age_value):
lxxv  age_intervals = {
lxxvi  'Age_20-29': [20, 29],
lxxvii  'Age_30-39': [30, 39],
lxxviii  'Age_40-49': [40, 49],
lxxix  'Age_50-59': [50, 59],
lxxx  'Age_60-69': [60, 69],
lxxxi  'Age_70+': [70, float('inf')]}
lxxxii
lxxxiii  vector = np.zeros(len(age_intervals), dtype=int) # Initialisez un
lxxxiv  vecteur de zéros
lxxxv
lxxxvi  for i, (interval, bounds) in enumerate(age_intervals.items()):
lxxxvii  if bounds[0] <= age_value <= bounds[1]:
lxxxviii  vector[i] = 1 # Affectez 1 si la valeur appartient à
lxxxix  l'intervalle
lxxxx
lxxxxi  return vector.tolist()
lxxxxii
lxxxxiii  model = joblib.load("best_model1.pkl")
lxxxxiv
lxxxxv  def prédire(x):
lxxxxvi  pred = model.predict(x)
lxxxxvii  proba = model.predict_proba(x)
lxxxxviii  proba=np.round(proba*100, 2)
lxxxix  return pred,proba
lxxxxxi
lxxxxxii  # Créer une fonction pour l'application Streamlit
lxxxxxiii
lxxxxxiv  def main():
lxxxxxv
lxxxxxvi  @st.cache_data
lxxxxxvii  def get_img_as_base64(file):
lxxxxxviii  with open(file, "rb") as f:
lxxxxxix  data = f.read()
lxxxxc  return base64.b64encode(data).decode()
lxxxxci
lxxxxcii
lxxxxciii  page_bg_img = f"""
lxxxxciv  <style>
lxxxxcvi
lxxxxcvii
lxxxxcviii

```

```

[data-testid="stAppViewContainer"] > .main {{
  background-image: url(https://www.sygnum.com/wp-
content/uploads/2022/11/AdobeStock_453778834-scaled.jpeg);
  background-size: cover;
  background-position: center;
  background-repeat: no-repeat;
  background-attachment: no-fixed;
  height: 100vh;
  margin: 0;
  display: flex;

}}
.ribbon {{
  background-color: #000;
  color: #fff;
  text-align: center;
  padding: 10px;
  font-size: 24px;
  font-family: 'Arial', sans-serif; /* Choisir une belle police
*/
  font-weight: bold; /* Mettre le texte en gras */
  border: 2px solid #ff0000; /* Bordures rouges */
  border-radius: 10px; /* Coins arrondis */
  margin-top: -50px; /* Ajuster la position vers le haut */
  position: relative;
  z-index: 1; /* S'assurer que le ruban est au-dessus du contenu
*/
}}
[data-testid="stSidebar"] {{
  background-color: #000 !important; /* Fond noir */
  border: 2px solid #ff0000 !important; /* Bordure rouge */
  border-radius: 10px; /* Coins arrondis */
  margin-top: -30px; /* Ajuster la position vers le haut */
  position: relative;
  z-index: 1; /* S'assurer que la barre latérale est au-dessus du
contenu */
  padding: 10px;
}}
[data-testid="stHeader"] {{
  background: rgba(0, 0, 0, 0);
  color: white;
}}
[data-testid="stToolbar"] {{
  right: 2rem;
}}
</style>
"""

```

```

    st.markdown(page_bg_img, unsafe_allow_html=True)
    st.markdown('<div class="ribbon">APPLICATION DE CREDIT SCORING
BANCAIRE</div>', unsafe_allow_html=True)
    st.sidebar.image('logo.jpg', use_column_width='always')
    st.sidebar.title("Informations sur le client")
    duration = st.sidebar.slider("Durée du Remboursement (mois)",
min_value=1, max_value=120, value=60)
    amount = st.sidebar.number_input("Montant du Prêt", min_value=0)
    margin = st.sidebar.number_input("Taux d'intérêt", min_value=0)
    sex = st.sidebar.selectbox("Sexe", ['homme', 'femme'])
    marital_status = st.sidebar.selectbox("Situation Matrimoniale",
['célibataire', 'marié(e)', 'divorcé(e)', 'veuf(ve)'])
    job = st.sidebar.selectbox("Activité", ['ADMINISTRATIF', 'AGRO
ALIMENTAIRE', 'ASSURANCE', 'BANQUE', 'COMMERCIAL', 'DIPLOMATIE',
'DIRECTION GENERALE', 'DIVERS', 'ETUDES/RECH./DEVELOP.', 'FINANCE',
'INFORMATIQUE, ORGANIS.', 'JURIDIQUE', 'MARKETING, PUBLICITE',
'PRODUCTION', 'PROFESSIONS LIBERALES', 'RESSOURCES HUMAINES'])
    label = st.sidebar.selectbox("Type de prêt", ['BRIDGE PRET RELAIS',
'CCT AUTRES CRD', 'CCT CONSO', 'CCT CONSO BONNE GAMME', 'CCT CONSO PER
AUTRE', 'CCT HORS PP', 'CCT RESTRUCTURES', 'CCT SCOLAIRE', 'CMT AUTRES',
'CMT CONSO', 'CMT CONSO BONNE GAMME', 'CMT CONSO PERSO', 'CMT HAB BONNE
GAMME', 'CMT HAB PATRIMONIALE', 'CMT HORS PP', 'CMT RESTRUCTURES'])
    agency = st.sidebar.selectbox("Agence", ['AGENCE - ADJAME', 'AGENCE
- AG PRINCIPALE', 'AGENCE - COCODY', 'AGENCE - II PLATEAUX 8IEME
TRANCHE', 'AGENCE - MARCORY RESIDENTIEL', 'AGENCE - PLATEAU SEEN HOTEL',
'AGENCE - RIVIERA 3', 'AGENCE - RIVIERA GOLF', 'AGENCE - SAN PEDRO',
'AGENCE - TREICHVILLE ZONE 3', 'AGENCE - ZONE 4 DR BLANCHARD', 'AGENCE-
2 PLATEAUX RUE DES JARDINS', 'AGENCE- DEUX PLATEAUX LATRILLE', 'AGENCE-
TREICHVILLE NANAN YAMOUSSO'])
    age = st.sidebar.number_input("Age", min_value=18)
    garantie = st.sidebar.multiselect("Sélectionnez les garanties :",
list(categories_mapping_garanties.keys()))
    month = st.sidebar.number_input("Mois_d'octroi du crédit",
min_value=1, max_value=12)
    year = st.sidebar.number_input("Année d'octroi du crédit",
min_value=2006, max_value=2050)

    # Créer un dataframe temporaire pour stocker les valeurs entrées par
l'utilisateur
    l=[duration, apply_robust_scaler(amount), margin]
    l.extend(categorielle({'SEXE': sex, 'SITUATION_MAT': marital_status,
'ACTILIB': job, 'LIBELLE': label, 'AGENCELIB': agency}))
    l.extend(map_age_interval_vector(age))
    l.extend(modif_garanties(garantie))
    l.extend([month,year])
    user_input_df = pd.DataFrame(l).T

```

```

# Afficher le dataframe résultant
#st.subheader("Données d'entrée utilisateur :")
#st.write(user_input_df)
if st.sidebar.button("Prédire"):
    predict, probability=prédire(user_input_df)
    with st.spinner('Wait for it...'):
        time.sleep(3)
    st.success('Prédiction effectuée!')
    p=pd.DataFrame(probability)[1][0]
    st.subheader(f"la probabilité de défaut de paiement est
de:{p}%")

fig_jauge = go.Figure(go.Indicator(
    mode='gauge+number+delta',
    value=p,
    domain={'x': [0, 1], 'y': [0, 1]},
    title={'text': 'probabilité de défaut', 'font': {'size':
24}},
    delta={'reference': 70,
        'increasing': {'color': 'Crimson'},
        'decreasing': {'color': 'Green'}},
    gauge={'axis': {'range': [None, 100],
        'tickwidth': 3,
        'tickcolor': 'darkblue'},
        'bar': {'color': 'white', 'thickness': 0.25},
        'bgcolor': 'white',
        'borderwidth': 2,
        'bordercolor': 'gray',
        'steps': [{'range': [0, 25], 'color': 'Green'},
            {'range': [25, 49.49], 'color':
'LimeGreen'},
            {'range': [49.5, 50.5], 'color': 'red'},
            {'range': [50.51, 75], 'color': 'Orange'},
            {'range': [75, 100], 'color': 'Crimson'}],
        'threshold': {'line': {'color': 'white', 'width': 10},
            'thickness': 0.8,
            'value':p}}))

fig_jauge.update_layout(paper_bgcolor='black',
    plot_bgcolor='black',
    height=400, width=500,
    font={'color': 'white', 'family':
'Arial'},
    margin=dict(l=0, r=0, b=0, t=0, pad=0),
    showlegend=False,
    xaxis=dict(showgrid=False,
zeroline=False),

```

```

ccxii.         yaxis=dict(showgrid=False,
ccxiii.         zeroline=False),
ccxiv.         shapes=[
ccxv.             dict(
ccxvi.                 type='rect',
ccxvii.                 xref='paper',
ccxviii.                 yref='paper',
ccxix.                 x0=0,
ccxx.                 y0=0,
ccxxi.                 x1=1,
ccxxii.                 y1=1,
ccxxiii.                 fillColor='black',
ccxxiv.                 opacity=1,
ccxxv.                 layer='below',
ccxxvi.                 line=dict(width=2, color='red'),
ccxxvii.             )
ccxxviii.         ]
ccxxix.         )
ccxxx.
ccxxx1.
ccxxxii.         col1, col2 = st.columns([1.5, 1])
ccxxxiii.         with col1:
ccxxxiv.             st.plotly_chart(fig_jauge)
ccxxxv.         with col2:
ccxxxvi.             st.write("")
ccxxxvii.             st.write("")
ccxxxviii.             st.write("")
ccxxxix.             st.write("")
ccl.             st.write("")
ccli.             # Texte d'accompagnement de la jauge
cclii.             if 0 <= p < 25:
ccliii.                 score_text = 'Crédit score : EXCELLENT'
ccliv.                 st.success(score_text)
cclv.             elif 25 <= p < 50:
cclvi.                 score_text = 'Crédit score : BON'
cclvii.                 st.success(score_text)
cclviii.             elif 50 <= p < 75:
cclix.                 score_text = 'Crédit score : MOYEN'
ccli.                 st.warning(score_text)
ccli.             else:
cclii.                 score_text = 'Crédit score : BAS'
ccliii.                 st.error(score_text)
ccliv.
cclv.
cclvi.         # Exécuter l'application
cclvii.         if __name__ == '__main__':
cclviii.             main()

```

TABLE DES MATIERES

DECHARGE	ii
AVANT-PROPOS	iii
DEDICACES	iv
REMERCIEMENTS	v
SOMMAIRE	vi
SIGLES ET ABREVIATIONS.....	1
LISTE DES GRAPHIQUES	2
LISTE DES TABLEAUX	4
RESUME.....	5
ABSTARCT	5
INTRODUCTION GENERALE.....	8
I. CONTEXTE ET JUSTIFICATION	9
II. PROBLEMATIQUE.....	10
III. OBJECTIFS DE L'ETUDE	11
IV. HYPOTHESES DE RECHERCHE.....	11
V. PLAN DE L'ETUDE.....	12
CHAPITRE I : DÉFINITION DES CONCEPTS ET REVUE DE LA LITTÉRATURE	13
I. DEFINITIONS DES NOTIONS ET CONCEPTS CLES	14
1. Les notions et concepts clés liés au secteur bancaire :.....	14
a. Les banques	14
b. Les particuliers	15
c. Les crédits bancaires aux particuliers.....	16
d. Les risques liés à l'activité bancaire.....	17
e. Le défaut de paiement	19
2. Les notions et concepts liés au machine learning	20

a.	Le machine learning	20
b.	Quelques modèles classiques de machine learning.....	23
II.	REVUE DE LA LITTERATURE	32
1.	Revue théorique	32
a.	Définition du scoring par la littérature	32
b.	Les méthodes d'évaluation du risque de crédit : Approche traditionnelle vs. Approche statistique	33
c.	Les avantages et inconvénients du scoring.....	34
2.	Revue empirique	35
	CHAPITRE II : PRESENTATION ET ANALYSE DE LA BASE DE DONNEES	38
I.	PRESENTATION DE LA BASE DONNEES.....	39
II.	APUREMENT DE LA BASE DE DONNÉES	42
1.	Tests de cohérence	42
2.	Suppression des doublons	43
3.	Autres manipulations	43
4.	Création de la variable d'intérêt : STATUT	43
III.	ANALYSE EXPLORATOIRE DES DONNÉES	44
1.	Analyse univariée.....	44
➤	Analyse des variables qualitatives.....	44
➤	Analyse des variables quantitatives.....	47
➤	Analyse de la variable textuelle	47
➤	Analyse des variables temporelles	49

2.	Analyse bivariée.....	50
a.	Variables quantitatives Vs. Variable cible	50
b.	Variable qualitative Vs. Variable cible	51
c.	Variables quantitatives Vs. Variables quantitatives	53
CHAPITRE III : MÉTHODOLOGIE ET MODELISATION.....		54
VI.	MÉTHODOLOGIE DE L'ÉTUDE	55
1.	Data Preprocessing.....	56
a.	Gestion des données manquantes :.....	56
b.	Encodage des variables catégorielles :	56
c.	Normalisation des valeurs numériques :	56
d.	Features Engineering.....	57
2.	Découpage de la base de données	58
3.	Traitement des déséquilibres de classe :	58
4.	Métrique d'évaluation	59
a.	L'exactitude (Accuracy) :	59
b.	La précision ou spécificité :	59
c.	Le rappel ou sensibilité :	59
d.	Le F1-score :	60
e.	La matrice de confusion :	60
f.	Aire sous la courbe ROC (AUC-ROC) :	60
5.	Modèles de base et sélection de variables.....	60
VII.	MODÉLISATION	61

1. Modèles.....	61
2. Cross Validation	62
3. Optimisation des hyperparamètres.....	62
CHAPITRE IV : PRÉSENTATION DES RÉSULTATS ET INTERPRÉTATIONS.....	63
I. PRESENTATION DES RESULTATS	64
1. Détermination du modèle optimal	64
2. Capacité de généralisation du modèle.....	65
II. MISE EN PLACE DU SCORE.....	66
III. INTERPRÉTABILITÉ DU MODÈLE	66
1. Features Importance.....	66
2. La Shapley Summary Plot	67
CHAPITRE V : DISCUSSION DES RESULTATS, RECOMMANDATIONS ET PROPOSITION D'UNE APPLICATION WEB	70
I. DISCUSSION DES RÉSULTATS	71
1. Age et risque de crédit	71
2. Montant des prêts et risque de crédit	71
3. Durée de remboursement et risque de crédit.....	71
4. Taux d'intérêt et risque de crédit.....	72
5. Sexe et risque de crédit	72
II. LIMITES DE L'ÉTUDE	72
III. RECOMMANDATIONS.....	73
IV. APPLICATION WEB DE CREDIT SCORING	73
CONCLUSION GENERALE	75

REFERENCES BIBLIOGRAPHIQUES	78
ANNEXES	82
Annexe I : Code source du traitement des données.....	83
Annexe II : Code source de la modélisation	86
Annexe III : Code source de l'application	88
TABLE DES MATIERES	95