République Tunisienne

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Carthage- Ecole Supérieure de la Statistique et de l'Analyse de l'Information

Rapport de Projet de Fin d'Études présenté pour l'obtention du

Diplôme National d'Ingénieur en Statistique et Analyse de l'Information

Houssem ARIDHI

Modélisation et Développement d'un Outil de Prédiction de la LGD sous IFRS9

Soutenu le 12/06/2024 devant le Jury composé de :

- *Mr Ghazi Belmufti (Président)*
- *Mme Hajer Sellami (Rapporteuse)*
- *Mme Tasnim Hamdeni(Encadrante universitaire)*
- *Mme Leila Zaghdoud (Encadrante entreprise)*
- *Mr Marwen Ben Nasr (Co-Encadrant entreprise)*

Stage de Fin d'Études effectué à

Banque Zitouna

Année universitaire 2023-2024

Une dédicace à mes précieux parents, Noureddine ARIDHI et Malika GHAZOUANI, dont l'amour, le soutien et les innombrables sacrifices ont tracé le chemin de ma réussite, je rends hommage à travers ce travail. Votre présence constante et vos conseils avisés ont été mes piliers, et je vous exprime ma gratitude éternelle pour les valeurs nobles, l'éducation et l'encouragement que vous avez insufflés en moi. Je souhaite également dédier ce travail à mon frère Jasser, à mes sœurs Tasnim et Nour, ainsi qu'à toute ma famille, source inépuisable d'inspiration et de soutien. Leur exemple m'a guidé à chaque étape de ma vie. Enfin, un grand merci à tous mes professeurs et instituteurs, dont le dévouement et l'excellence dans la transmission du savoir sont reflétés dans ce travail, symbole de fierté pour nous tous.

REMERCIEMENTS

Je souhaite commencer ce rapport en exprimant ma gratitude envers toutes les personnes qui m'ont énormément appris au cours de ce projet de fin d'études et envers celles qui ont eu la gentillesse de rendre ce stage extrêmement enrichissant.

Je tiens à remercier chaleureusement mes encadrants en entreprise, Madame Leila ZAGHDOUD et Monsieur Marwen BEN NASR, pour leur gentillesse, leur patience, le temps qu'ils m'ont consacré et le partage quotidien de leur expertise. Grâce à leurs conseils et à leur confiance, j'ai pu m'accomplir pleinement dans mes missions.

Je souhaite également exprimer ma reconnaissance à mon encadrante universitaire, Madame Tasnim Hamdeni, pour son implication dans cette étude, son suivi attentif, et la valeur ajoutée qu'elle a apportée à ce projet.

Des remerciements spéciaux à mon collègue Abdennacer BOUABID pour sa grande contribution pendant la période de saisie manuelle des données.

Enfin, j'apprécie la présence de Monsieur Ghazi Belmufti en tant que président du jury et de Madame Hajer Sellami en tant que rapporteuse de ce travail.

Résumé

Dans un contexte de transition vers l'IFRS 9, ce rapport se concentre sur la modélisation et le développement d'un outil de prédiction de la LGD (Loss Given Default) pour répondre aux exigences réglementaires et normatives en matière de provisionnement des risques. L'IFRS 9, adoptée à cause de la crise financière de 2008, requiert une estimation précise des pertes de crédit attendues (ECL), plaçant ainsi la LGD au cœur de ce processus.

En utilisant un portefeuille de crédits de la Banque Zitouna comme étude de cas, notre objectif principal est de concevoir un modèle robuste pour estimer la LGD conformément aux directives de l'IFRS 9. Après une évaluation approfondie de différentes techniques de modélisation, on a déterminé que le modèle de régression linéaire offrait la meilleure performance dans notre cas. On met en œuvre une méthodologie rigoureuse, incluant ce modèle de régression linéaire ainsi que des techniques statistiques avancées et une analyse approfondie des données, pour développer cet outil de prédiction.

Les résultats de notre étude démontrent l'importance critique de la LGD dans le cadre de l'IFRS 9, ainsi que l'efficacité de notre modèle de régression linéaire pour estimer cette mesure.

ABSTRACT

In a context of transitioning to IFRS 9, this report focuses on modeling and developing a Loss Given Default (LGD) prediction tool to meet regulatory and normative requirements in risk provisioning. IFRS 9, adopted in response to the 2008 financial crisis, necessitates precise estimation of Expected Credit Losses (ECL), thus placing LGD at the core of this process.

Using a credit portfolio from Zitouna Bank as a case study, our primary objective is to design a robust model to estimate LGD in accordance with IFRS 9 guidelines. After thorough evaluation of various modeling techniques, it was determined that the linear regression model provided the best performance in our case. We implement a rigorous methodology, incorporating this linear regression model along with advanced statistical techniques and comprehensive data analysis, to develop this prediction tool.

The results of our study demonstrate the critical importance of LGD within the framework of IFRS 9, as well as the effectiveness of our linear regression model in estimating this parameter.

Table des matières

Table des figures

Liste des tableaux

INTRODUCTION GÉNÉRALE

Le projet, intitulé "Modélisation et Développement d'un Outil de Prédiction de la LGD sous IFRS 9", adresse la nécessité de concevoir un outil capable de prédire avec précision la Loss Given Default (LGD) conformément aux exigences de la Banque Centrale de Tunisie pour la norme IFRS 9. Cette initiative vise à évaluer la perte financière en cas de défaut des emprunteurs, en utilisant des techniques avancées de modélisation et de prédiction, telles que les modèles d'apprentissage automatique. Ces techniques sont privilégiées pour leur fiabilité et leur aptitude à fournir des estimations précises et automatisées.

Le rapport est composé en quatre chapitres distincts. Le premier chapitre expose le contexte du projet, présente l'organisme d'accueil et énonce les principaux concepts fondamentaux sous-jacents. Le deuxième chapitre détaille le fonctionnement de chaque modèle d'apprentissage automatique employé dans le projet, les techniques de prétraitement et d'ingénierie des données utilisées pour garantir la robustesse des prédictions, ainsi que les outils utilisés pour créer l'outil de prédiction.

Le troisième chapitre se concentre sur l'analyse exploratoire des données, englobant les phases de prétraitement, de nettoyage et de description des données, ainsi que les méthodes de sélection des variables pertinentes. Enfin, le quatrième et dernier chapitre expose en détail les différents modèles d'apprentissage automatique déployés, les métriques d'évaluation de la performance des modèles et les étapes de création de l'outil de prédiction.

Chapitre 1

PRÉSENTATION GÉNÉRALE DE PROJET

1.1 Introduction

Dans un paysage financier en constante évolution, l'adoption de normes comptables internationales telles que l'IFRS 9 représente un défi crucial pour les institutions financières. La Banque Zitouna, en tant qu'acteur majeur du secteur bancaire en Tunisie, se trouve à un moment charnière de son parcours, cherchant à intégrer efficacement ces normes pour répondre aux exigences réglementaires tout en préservant ses valeurs islamiques fondamentales.[1]

1.2 Présentation de l'organisme d'accueil

La Banque Zitouna est une banque établie en Tunisie, qui se spécialise dans la finance islamique. Elle propose une variété de produits financiers respectant les principes de la charia., répondant ainsi aux attentes des clients désireux de bénéficier de services bancaires en accord avec leurs convictions religieuses.

Fondée en 2009, la Banque Zitouna a ouvert ses portes au public en 2010, s'engageant dès lors à fournir des solutions bancaires innovantes et éthiques sur le marché tunisien. Son approche axée sur les principes de la charia se reflète dans ses activités, de la conception de produits financiers à la gestion des transactions, garantissant ainsi une conformité totale aux exigences islamiques en matière de finances.

Cependant, l'histoire de la Banque Zitouna n'a pas été exempte de défis. En janvier

2011, la Banque Centrale de Tunisie a placé la banque sous administration provisoire, marquant une période de transition et de réorganisation. Malgré ces défis, la Banque Zitouna a maintenu son engagement envers ses clients et ses valeurs fondamentales, poursuivant ses efforts pour offrir des services bancaires conformes à la charia tout en assurant la stabilité et la fiabilité de ses opérations.

En 2018, un tournant majeur s'est produit lorsque le gouvernement tunisien a décidé de céder sa participation majoritaire de 69,15 % dans la Banque Zitouna au groupe qatari Majda. Cette acquisition a marqué une nouvelle ère pour la banque, renforçant sa position sur le marché et ouvrant de nouvelles opportunités de croissance et d'expansion.

Ainsi, la Banque Zitouna incarne l'engagement envers l'éthique et l'innovation dans le secteur financier tunisien, offrant des solutions bancaires qui allient performance économique et respect des valeurs religieuses. Son histoire témoigne de sa résilience et de sa capacité à s'adapter aux défis tout en restant fidèle à sa mission de servir ses clients et de contribuer au développement économique de la Tunisie.

1.3 **Présentation du projet**

1.3.1 **Contexte**

Dans ce contexte, la nécessité de développer des modèles précis pour estimer la Perte en cas de Défaut (LGD) devient impérative. Ces modèles constituent la base sur laquelle repose la gestion des risques de crédit de la Banque Zitouna, leur précision influençant directement la santé financière de l'institution et sa conformité aux exigences réglementaires.

1.3.2 **Objetifs**

Le présent projet se concentre sur le développement de modèles robustes pour estimer la LGD dans le cadre spécifique des produits financiers islamiques offerts par la Banque Zitouna. Les objectifs principaux sont les suivants :

Développement de Modèles LGD : Concevoir des modèles statistiques et analytiques pour estimer de manière précise la Perte en cas de Défaut (LGD) pour les différents types de produits financiers islamiques proposés par la Banque Zitouna.

Validation et Évaluation : Procéder à une validation rigoureuse des modèles développés en utilisant des données historiques pertinentes et des techniques appropriées de validation des modèles afin de garantir leur précision et leur fiabilité.

Création d'une Interface Graphique : Développer une interface graphique intuitive pour implémenter et utiliser le modèle de LGD, facilitant ainsi leur utilisation par les analystes et les décideurs de la Banque Zitouna.

1.3.3 **Problématique**

Sous l'IAS 39, antérieure à l'IFRS 9, les banques étaient contraintes de comptabiliser les pertes de crédit uniquement après avoir identifié une preuve manifeste de perte. Cela conduisait souvent à une sous-estimation des pertes de crédit. L'IFRS 9, quant à elle, privilégie une approche plus proactive en exigeant des banques qu'elles comptabilisent les pertes de crédit attendues (ECL) à tout moment, en tenant compte des événements passés, et aussi des conditions actuelles et des projections futures. Cette méthode anticipative facilite une reconnaissance plus précise des pertes de crédit, ce qui représente un changement significatif dans la gestion du risque de crédit.

Cependant, la détermination des pertes de crédit attendues, bien que conceptuellement simple, implique un processus complexe basé sur la modélisation statistique. Contrairement à l'IAS 39 qui ne définissait pas de méthodologie claire, l'IFRS 9 exige des institutions financières de mettre en place des modèles sophistiqués prenant en compte divers paramètres de risque tel que la perte en cas de défaut (LGD). Cette méthodologie nécessite une expertise approfondie en finance et une maîtrise des techniques de modélisation statistique.

1.4 **Définitions et notions de bases**

1.4.1 **Norme IFRS9**

L'IFRS 9, ou "International Financial Reporting Standards 9 - Instruments Financiers", est une norme comptable internationale qui est devenue effective pour les exercices débutant à partir du 1er janvier 2018. Elle régit la classification, l'évaluation et la dépréciation des actifs financiers, remplaçant et complétant la norme précédente, l'IAS 39 - Instruments Financiers : Comptabilisation et Évaluation. Cette évolution a pour objectif d'améliorer la qualité de l'information financière, en adoptant notamment une approche plus anticipative pour la comptabilisation des pertes attendues sur les actifs financiers, en réponse aux leçons tirées de la crise financière de 2008. L'International Accounting Standard Board (IASB) est à l'origine de cette initiative, visant à installer un système financier plus stable et à faciliter une comptabilisation plus rapide et efficace des pertes de crédit. En vertu de l'IFRS 9, les entités financières doivent prendre en compte la probabilité de défaut sur une période de crédit de 12 mois dans leurs évaluations, renforçant ainsi la transparence et la solidité du système financier international.

La norme IFRS 9 impose aux entités détenant des instruments financiers de comptabiliser les pertes de crédit attendues selon des règles très différentes de celles énoncées dans la norme IAS 39. En conséquence, l'IFRS 9 a entraîné une augmentation significative des provisions pour pertes de crédit. En effet, contrairement à l'IAS 39 qui se basait sur des pertes avérées, l'IFRS 9 impose une approche plus proactive et anticipative en évaluant les pertes attendues sur toute la durée de vie des actifs financiers. Cette méthode, appelée "modèle des pertes de crédit attendues", demande aux entreprises de prendre en compte non seulement les pertes de crédit déjà manifestes, mais aussi celles qui sont probables en fonction des conditions économiques actuelles et futures. Cela a conduit de nombreuses institutions financières à augmenter leurs provisions comme indique la figure 1.1, renforçant ainsi leur capacité à absorber les chocs financiers et à maintenir la

stabilité économique.[2][3]

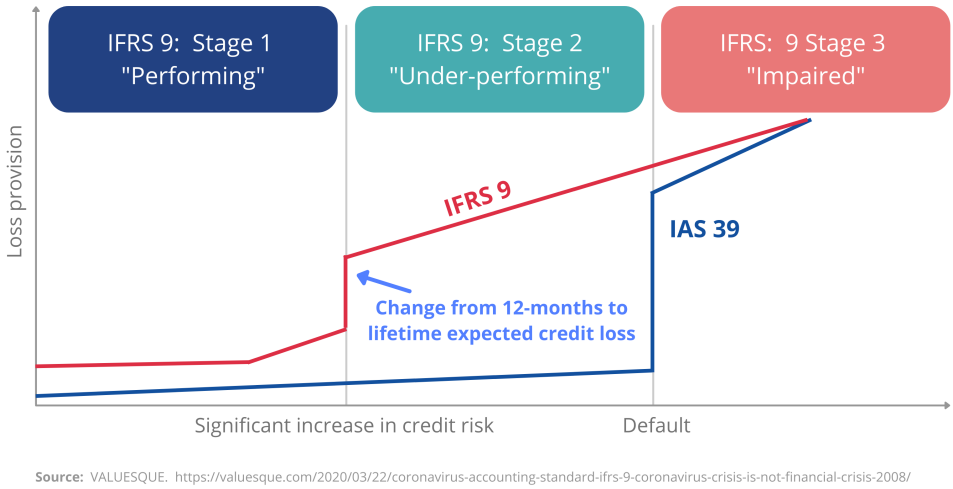

Figure 1.1: *IFRS9 vs IAS39*

1.4.2 IFRS9 dans le secteur bancaire tunisien

En 2018, le Conseil National de la Comptabilité a annoncé que le secteur financier tunisien adopterait les normes IFRS à partir de 2021. En avril 2019, lors d'une conférence sur la "Transition du secteur financier aux IFRS", le ministre des Finances a confirmé cette décision, soulignant que la Tunisie passerait progressivement aux normes IFRS à partir de 2021. Il a également annoncé l'intention de soumettre un projet de loi à l'Assemblée des Représentants du Peuple (ARP) pour réglementer cette transition. Cette adoption des normes IFRS représentait une opportunité pour accroître la transparence et l'efficacité des systèmes bancaires en Tunisie. [4]

1.4.3 Défaut

Dans le domaine des crédits bancaires, le terme "défaut" désigne une situation où un emprunteur est incapable de respecter ses engagements de remboursement envers la banque selon les conditions établies dans le contrat de prêt. Cette situation peut survenir

pour diverses raisons. Par exemple, l'emprunteur peut cesser de payer ses mensualités de prêt pendant une période prolongée en raison de difficultés financières ou d'une détérioration de sa situation économique. Dans d'autres cas, l'emprunteur peut ne pas être en mesure de rembourser le capital emprunté à la date d'échéance du prêt, ce qui constitue également un défaut.

Le défaut peut également survenir si l'emprunteur ne respecte pas d'autres conditions de paiement convenues, telles que le non-paiement des intérêts, des frais ou des charges associées au prêt. Quelle que soit la raison, le défaut représente un risque pour la banque, car cela peut entraîner une perte financière si l'emprunteur ne parvient pas à rembourser intégralement le prêt. En conséquence, les banques mettent en place des mesures pour évaluer et gérer ce risque, telles que l'utilisation de modèles de notation de crédit, l'évaluation de la solvabilité des emprunteurs et la mise en place de mécanismes de recouvrement en cas de défaut.

1.4.4 **Pertes de crédit attendues ECL**

L'Expected Credit Loss (ECL), ou Pertes de Crédit Attendues, est une mesure cruciale utilisée dans l'évaluation des actifs financiers, principalement dans le contexte de la norme comptable IFRS 9. L'ECL représente une estimation des pertes potentielles qui pourraient découler du défaut de paiement des débiteurs. Cette mesure est d'une importance primordiale pour les institutions financières, leur permettant d'évaluer de manière proactive les risques associés à leurs actifs financiers et de prendre des décisions éclairées en matière de gestion des risques.[5]

L'ECL est calculée en utilisant une équation qui prend en compte plusieurs composantes essentielles :

$$ECL = PD * LGD * EAD \tag{1.1}$$

Où :

ECL : Expected Credit Loss (Perte de crédit attendue)

PD : Probability of Default (Probabilité de défaut)

EAD : Exposure At Default (Exposition en cas de défaut)

LGD : Loss Given Default (Pertes en cas de défaut)

1.4.5 **Perte en cas de défaut(LGD)**

La "Loss Given Default" (LGD), ou perte en cas de défaut, est une mesure cruciale dans l'évaluation du risque de crédit. Elle représente la proportion de l'exposition totale d'un prêteur qui ne peut pas être récupérée en cas de défaut. La LGD inclut à la fois le montant principal du prêt ainsi que les intérêts ou frais impayés.[6]

Cette mesure prend en compte divers facteurs tels que la base d'actifs de l'emprunteur et les privilèges existants, comme les garanties fournies dans le contrat de prêt. La LGD est exprimée en pourcentage et indique la partie de l'exposition totale qui ne devrait pas être récupérée en cas de défaut. En d'autres termes, c'est la proportion des fonds prêtés qui est considérée comme irrécupérable.

Il est important de noter que chaque prêteur est exposé au risque que l'emprunteur fait défaut, en particulier dans des conditions économiques défavorables. Cependant, l'évaluation des pertes potentielles ne se résume pas uniquement à la valeur totale du prêt (l'exposition en cas de défaut, ou EAD), car cela dépend de facteurs variables tels que la valeur des garanties et les taux de recouvrement.[7]

En résumé, la LGD est un élément important dans l'évaluation du risque de crédit, car elle permet aux prêteurs et aux investisseurs de quantifier les pertes potentielles en cas de défaut des emprunteurs, et ainsi de prendre des décisions informées en matière de prêt et d'investissement.[8]

Processus de "Workout"

Le processus de "Workout" implique l'identification des prêts non performants (NPL) ou des actifs en difficulté, ainsi que l'évaluation de leur situation, y compris la valeur de toute garantie si le prêt est sécurisé. Des stratégies de restructuration sur mesure sont ensuite élaborées, souvent en négociant avec les emprunteurs pour parvenir à des modalités de remboursement mutuellement acceptables. Ces plans sont mis en œuvre, avec un suivi étroit de la progression des emprunteurs et des ajustements effectués si nécessaire. L'objectif ultime est de résoudre les NPL ou les actifs en difficulté de manière à maximiser la récupération pour les prêteurs tout en minimisant les pertes. Une communication efficace et un engagement proactif avec les emprunteurs sont essentiels tout au long du processus pour garantir des résultats réussis.[9]

"Workout" LGD

Le terme "Workout LGD" désigne une estimation de la Perte en cas de Défaut (LGD) spécifiquement pendant le processus de restructuration des prêts non performants ou d'autres instruments de crédit. En essence, il représente la portion du montant en défaut que les prêteurs ou créanciers sont susceptibles de perdre après avoir tenté de recouvrer la dette impayée par le biais de différentes stratégies de restructuration.

L'évaluation des prêts non performants implique souvent d'évaluer le Workout LGD dans le cadre de gestion des risques. Cette estimation aide les institutions financières à évaluer les pertes potentielles associées à leurs portefeuilles de crédit et à prendre des décisions informées concernant les provisions, l'allocation de capital et les stratégies de réduction des risques. En comprenant le Workout LGD, les prêteurs peuvent mieux gérer leurs expositions aux actifs non performants et optimiser leurs efforts de récupération pendant le processus de restructuration.

Calcul de "Workout" LGD pour un défaut

La "LGD Workout default" désigne l'estimation de la perte en cas de défaut spécifiquement pour un défaut pendant le processus de restructuration des prêts non performants ou d'autres instruments de crédit.

On a la formule de "Workout" LGD default est :

$$LGD_d = 1 - \frac{\sum_{i=1}^{n} VAN(recup) + VAN(fluxartificiel)}{EAD_d} \tag{1.2}$$

Dans cette formule, LGD_d représente la perte en cas de défaut, $VAN(Recup)$ est la valeur actuelle nette des recouvrements, $VAN(fluxartificiel)$ est la valeur actuelle nette de l'engagement au moment de sortie de défaut, et EAD_d est l'exposition en cas de défaut.

On a :

$$Recup = (Engagement_t - Engagement_{t+1}) + Max(Impayenprofit_{t+1} - Impayenprofit_t, 0) \tag{1.3}$$

Et

$$EAD_d = Engagement_d \tag{1.4}$$

Et

$$Engagement_d = Encours_d + Impaye_d \tag{1.5}$$

Remarque

Si on a plusieurs défauts pour un seul financement, la LGD de ce financements est la moyenne des LGDs pour ces défauts.

1.4.6 **Modélisation statistique**

La modélisation statistique est un processus analytique qui implique l'utilisation de modèles statistiques tels que la régression linéaire, la régression logistique, l'analyse des séries chronologiques, la modélisation par équations simultanées, etc., pour développer des modèles mathématiques ou probabilistes afin de représenter et d'expliquer les phénomènes observés dans le monde réel. Ces modèles sont utilisés pour prédire des résultats futurs, comprendre les relations entre les variables, estimer des paramètres inconnus et tester des hypothèses scientifiques.

1.5 **Environnement logiciel**

Dans l'étape de la modélisation, Google Colab a joué un rôle central et irremplaçable dans notre processus de travail. Cette plateforme interactive a été notre allié indispensable, offrant un environnement de développement de code flexible et puissant. On a utilisé Google Colab pour écrire et exécuter notre code Python, bénéficiant ainsi de ses fonctionnalités avancées de visualisation de données.

Grâce à son intégration native avec des bibliothèques telles que *Pandas, Scikit-learn* et *Statsmodels*, on a pu créer des graphiques et des visualisations de données riches et informatifs. Ces outils visuels ont été essentiels pour comprendre et interpréter nos résultats de manière approfondie.

Dans la partie de la création de l'interface graphique, on a opté pour l'utilisation de Visual Studio Code (VSCode). Cette plateforme de développement a été un choix judicieux pour concevoir une interface utilisateur conviviale et fonctionnelle. Avec ses nombreuses extensions et sa facilité d'utilisation, VSCode nous a permis de développer efficacement l'interface graphique de notre application. L'intégration fluide avec Flask a simplifié le processus de conception et de mise en œuvre d'une interface web interactive., nous permettant ainsi de fournir une expérience utilisateur optimale.

En résumé, tandis que Google Colab a été notre compagnon de choix pour la modé-

lisation et l'analyse des données, Visual Studio Code s'est révélé être l'outil idéal pour la création de l'interface graphique, offrant un environnement de développement intuitif et productif.

1.6 Méthodologie de travail

Afin d'établir une méthodologie de travail cohérente et mieux organisée, on a intégré la plateforme Mlflow. Cette dernière sert à gérer les différentes étapes de la modélisation. Mlflow nous a donné la capacité de suivre et d'enregistrer les résultats de nos expériences en conservant les métriques et les hyperparamètres utilisés dans chaque expérimentation menée. Cette solution nous a permis non seulement de suivre le progrès accompli lors du développement et de l'implémentation des modèles, mais aussi de stocker les modèles pour les réutiliser ultérieurement et les rendre reproductibles et simples à déployer.

1.7 Plan de travail

Pour illustrer les étapes d'avancement de ce projet de fin d'études, on a créé ce diagramme de Gantt (Figure 1.2), qui montre chaque tâche effectuée et sa durée.

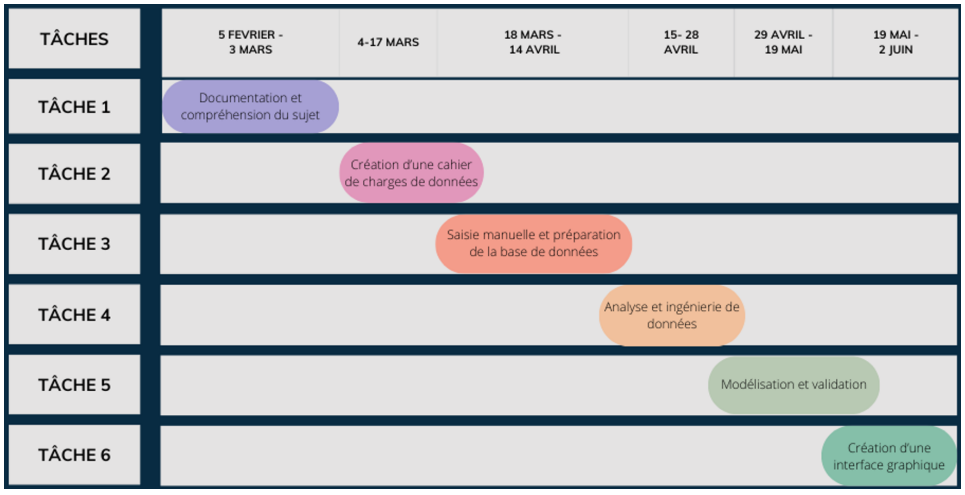

Figure 1.2: *Diagramme de Gantt*

1.8 Conclusion

Au cours de ce premier chapitre, on a présenté la banque Zitouna et son secteur d'activité, ainsi que les fondamentaux de la modélisation statistique et la norme IFRS 9 avec son application en Tunisie. On a annoncé la problématique et le cadre de ce projet ainsi que les objectifs visés à atteindre. Ce chapitre comprend également la méthodologie de travail ainsi que le planning du projet réalisé avec le diagramme de Gantt.

Chapitre 2

ÉTAT DE L'ART

2.1 Introduction

Dans ce chapitre, on expose les techniques de sélection des variables, les méthodes d'encodage, ainsi que le fonctionnement de chaque modèle de machine learning à utiliser lors de la phase de modélisation. En outre, on présente les métriques et les critères de performance choisis pour évaluer les modèles.

2.2 Sélection des variables

L'objectif est d'opter pour un nombre optimal de variables tout en améliorant les performances ou en maintenant des performances comparables au minimum. Concernant le choix des variables, l'objectif est de choisir celles qui ont une forte dépendance avec la variable cible. Cette réduction du nombre de variables rend les modèles plus interprétables et accélère leur processus d'entraînement.

2.2.1 Méthodes de filtrage

Les méthodes de filtrage font partie des techniques de sélection de variables utilisées dans le processus d'analyse de données et de modèlisation statistique. Elles consistent à évaluer chaque variable de manière indépendante, en se basant sur des mesures statistiques ou des tests spécifiques, pour déterminer leur pertinence par rapport à la variable cible ou pour identifier les variables qui présentent le moins de redondance entre elles.

Les étapes typiques d'une méthode de filtrage sont les suivantes :

Calcul des mesures de pertinence : Cette étape implique le calcul de mesures statistiques telles que la corrélation, l'information mutuelle ou le test du khi-deux entre chaque variable et la variable cible. Ces mesures permettent d'analyser la relation entre chaque variable et la variable cible.

Sélection des variables pertinentes : Une fois les mesures de pertinence calculées, les variables les plus pertinentes sont sélectionnées en fonction d'un seuil prédéfini. Les variables qui dépassent ce seuil sont conservées pour la modélisation, tandis que les autres sont éliminées.

Réduction de la redondance : En plus de sélectionner les variables les plus pertinentes, les méthodes de filtrage peuvent également être utilisées pour réduire la redondance entre les variables. Cela implique l'élimination de variables qui sont fortement corrélées entre elles, car elles fournissent des informations similaires au modèle.

Les avantages des méthodes de filtrage incluent leur simplicité et leur rapidité d'exécution, ce qui les rend adaptées aux ensembles de données de grande taille. Cependant, elles peuvent avoir tendance à sous-estimer la complexité des relations entre les variables et à ne pas prendre en compte les interactions entre celles-ci.

2.2.2 Méthodes d'encapsulation

Ces méthodes, familièrement appelées méthodes de sélection de variables de type Wrapper, parcourent l'ensemble des variables disponibles pour trouver le sous-ensemble qui offre le meilleur potentiel prédictif. Elles créent plusieurs modèles avec différents sous-ensembles de variables et sélectionnent celui qui obtient les meilleures performances selon une mesure de performance prédéfinie. Parmi les stratégies d'encapsulation les plus couramment utilisées, on trouve :

Forward Feature Selection (FFS) : Cette méthode commence avec un modèle contenant une constante, puis ajoute une à une les variables au modèle, en ne conservant que celles

qui améliorent la performance du modèle. Le processus se poursuit jusqu'à ce qu'aucune variable non choisie n'apporte une amélioration significative aux performances du modèle.

Backward Feature Elimination (BFE) : Contrairement à la FFS, cette méthode démarre avec toutes les variables et élimine progressivement celles qui ont le moins d'impact statistique sur le modèle, itération par itération.

Recursive Feature Selection (RFS) : Cette technique combine des aspects de la FFS et de la BFE. Elle explore toutes les combinaisons possibles de variables pour trouver celle qui offre la meilleure pertinence dans la construction du modèle.

2.2.3 Test F pour les modèles de régression linéaire

Le test F est couramment utilisé pour évaluer la signification globale d'un modèle de régression linéaire. Il permet de déterminer si les variables indépendantes, prises ensemble, ont une relation statistiquement significative avec la variable dépendante. Le test F compare un modèle de régression complet à un modèle nul (un modèle sans prédicteurs).

Hypothèses du test F

- **Hypothèse nulle (H0)** : Tous les coefficients de régression sont égaux à zéro (aucune relation entre les variables indépendantes et la variable dépendante).
- **Hypothèse alternative (H1)** : Au moins un des coefficients de régression est différent de zéro (au moins une variable indépendante a une relation avec la variable dépendante).

Formule du test F

Le test F se base sur la comparaison de la variance expliquée par le modèle (la somme des carrés expliquée, SCE) et la variance non expliquée par le modèle (la somme des carrés des résidus, SCR).

$$F = \frac{\frac{SCR}{p}}{\frac{SCE}{n-p-1}} \tag{2.1}$$

Où :

— SCR est la somme des carrés des résidus.

— $SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

— SCE est la somme des carrés expliquée.

— $SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

— p est le nombre de variables dans le modèle.

— n est le nombre total d'observations.

Interprétation

Une valeur de F élevée et une p-value faible (généralement $< 0{,}05$) indiquent que le modèle de régression est statistiquement significatif, c'est-à-dire qu'il existe une relation linéaire entre les variables indépendantes et la variable cible.

2.3 Manipulation des variables catégorielles

La présence de variables catégorielles dans les données peut souvent compliquer le processus d'apprentissage. En effet, la plupart des modèles d'apprentissage automatique nécessitent des données numériques en entrée. Par conséquent, il est nécessaire de trouver des méthodes pour convertir les différentes modalités de ces variables en valeurs numériques.

2.3.1 Encodage des variables catégorielles

Transformer les variables catégorielles en variables numériques, c'est ce qu'on appelle l'encodage. Cette étape est trés importante car elle permet de rendre les données manipulables par les modèles. Il est essentiel de choisir la méthode d'encodage la plus éfficace

pour chaque modèle. Ainsi, on accorde une attention particulière à cette phase de transformation. Nous explorerons deux méthodes d'encodage : le One Hot Encoder et le Label Encoder.

2.3.2 **One Hot Encoder**

Le One Hot Encoder est la méthode la plus utilisée pour convertir les variables catégorielles en variables numériques. Elle offre généralement des résultats satisfaisants et elle est simple à mettre en œuvre. Cette approche implique de remplacer chaque variable catégorielle par un ensemble de variables numériques correspondant au nombre de modalités de la variable initiale. Par exemple, si une variable catégorielle comporte k modalités, le One Hot Encoder génère k variables indicatrices. Chaque variable prend la valeur 1 pour la modalité correspondante et 0 pour les autres modalités.

Cependant, cette technique peut engendrer un nombre important de variables lorsque les variables catégorielles ont de nombreuses modalités. Cela conduit à un ensemble de données plus volumineux, nécessitant plus d'espace mémoire et pouvant rendre le traitement plus complexe pour les algorithmes d'apprentissage.

2.3.3 **Label Encoder**

Le Label Encoder est une autre méthode largement utilisée pour convertir les variables catégorielles en variables numériques. Contrairement au One Hot Encoder, il assigne à chaque modalité de la variable catégorielle une valeur numérique unique. Par exemple, si une variable catégorielle comporte k modalités, le Label Encoder assigne des entiers de 0 à k-1 à chaque modalité. Cette méthode est plus simple et crée un seul attribut numérique pour chaque variable catégorielle. Cependant, elle peut induire une relation d'ordre artificielle entre les catégories, ce qui peut être problématique pour certains algorithmes d'apprentissage. De plus, elle peut ne pas être adaptée aux variables catégorielles avec un grand nombre de modalités, car cela peut entraîner des valeurs numériques qui ne

reflètent pas de réelles différences entre les catégories.

2.3.4 Conclusion

Les variables catégorielles sont fréquentes dans les données et exigent une attention particulière. En effet, une manipulation adéquate de ces variables peut grandement améliorer les performances du modèle.

2.4 Régression linéaire

La régression linéaire est une méthode statistique fondamentale utilisée pour modéliser la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Elle est largement employée dans divers domaines, tels que l'économie, les sciences sociales, les sciences naturelles, l'ingénierie, et bien d'autres encore. Son objectif est de déterminer une relation linéaire optimale entre les variables en ajustant un modèle linéaire aux données observées.

Formellement, dans le cadre de la régression linéaire simple, l'équation du modèle est définie comme suit :

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.2}$$

Où y représente la variable dépendante, x la variable indépendante, β_0 l'ordonnée à l'origine, β_1 le coefficient de pente, et ε le terme d'erreur. Le but est de déterminer les coefficients β_0 et β_1 qui minimisent la somme des carrés des résidus, représentant la différence entre les valeurs observées et celles prédites par le modèle.

La régression linéaire multiple généralise ce concept en permettant l'inclusion de plusieurs variables indépendantes dans le modèle, ce qui permet de modéliser des relations plus complexes. La figure 2.1 montre un exemple simple de régession linéaire.

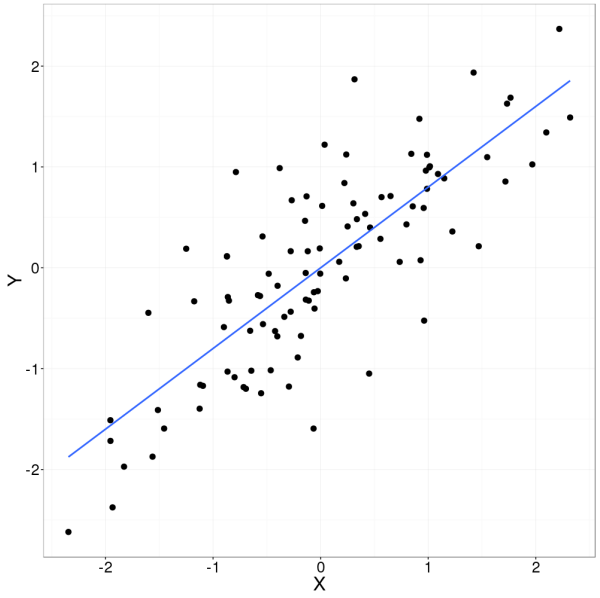

Figure 2.1: *Régression linéaire*

2.5 Stacking

Le stacking, ou empilement de modèles, est une technique d'apprentissage automatique qui consiste à combiner les prédictions de plusieurs modèles de base pour améliorer les performances de prédiction. Contrairement à d'autres techniques d'ensemble telles que le bagging et le boosting qui combinent les prédictions de plusieurs modèles de manière parallèle ou séquentielle, le stacking utilise un modèle de méta-apprentissage pour agréger les prédictions des modèles de base.

Le processus de stacking se déroule généralement en plusieurs étapes :

Division des données : Les données d'entraînement sont divisées en ensembles de validation et d'apprentissage.

Entraînement des modèles de base : Plusieurs modèles de base sont entraînés sur l'ensemble d'apprentissage. Ces modèles peuvent être de différents types ou utiliser des algorithmes d'apprentissage différents.

Génération des prédictions : Les modèles de base sont utilisés pour générer des

prédictions sur l'ensemble de validation.

Entraînement du modèle de méta-apprentissage : Les prédictions des modèles de base servent de caractéristiques d'entrée pour un modèle de méta-apprentissage (également appelé méta-modèle). Ce modèle de méta-apprentissage est ensuite entraîné sur les prédictions de l'ensemble de validation et les vraies valeurs cibles correspondantes.

Prédiction finale : Une fois que le modèle de méta-apprentissage est entraîné, il est utilisé pour faire des prédictions sur de nouvelles données.

Le stacking permet de tirer parti des points forts de différents modèles de base en les combinant de manière efficace. Il peut souvent produire des performances de prédiction plus élevées que celles des modèles individuels. Cependant, le stacking nécessite une gestion appropriée de la complexité du modèle, ainsi qu'une validation croisée rigoureuse pour éviter le surajustement. La figure 2.2 ci-dessous décrit le processus de stacking.

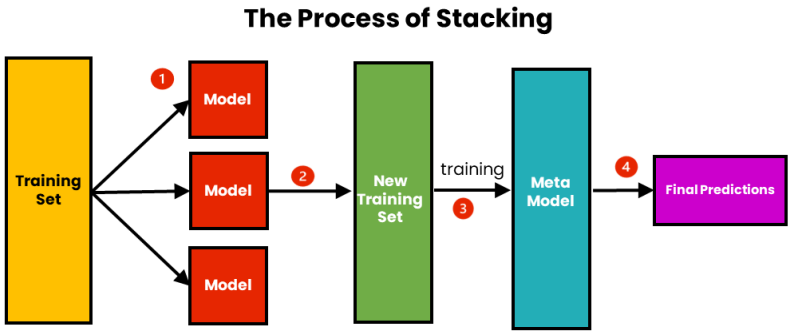

Figure 2.2: *Stacking*

2.6 Bagging

Le bagging, ou Bootstrap Aggregating, est une technique d'ensemble utilisée en apprentissage automatique pour améliorer la performance des modèles prédictifs. Il consiste à entraîner plusieurs modèles de base sur des sous-ensembles aléatoires des données d'entraînement et à combiner leurs prédictions pour obtenir une prédiction finale.

Le processus de bagging se déroule généralement en plusieurs étapes :

Bootstrap Sampling : Des échantillons de repliement sont générés à partir de l'ensemble de données d'entraînement. Ces échantillons sont des sous-ensembles aléatoires de la taille de l'ensemble de données d'entraînement, sélectionnés avec remplacement.

Entraînement des modèles de base : Sur chaque échantillon bootstrap, un modèle de base est entraîné indépendamment. Ces modèles de base peuvent être du même type ou de types différents, utilisant des algorithmes d'apprentissage différents.

Prédiction : Une fois les modèles de base entraînés, ils sont utilisés pour faire des prédictions sur l'ensemble de validation ou de test.

Agrégation des prédictions : Les prédictions des modèles de base sont agrégées pour obtenir une prédiction finale. Pour les problèmes de régression, la moyenne des prédictions est souvent utilisée, tandis que pour les problèmes de classification, un vote majoritaire est généralement appliqué.

Le bagging permet de réduire la variance des modèles individuels en moyennant leurs prédictions. Cela aide à réduire le surajustement et à améliorer la généralisation du modèle. De plus, le bagging peut être parallélisé efficacement, ce qui en fait une technique efficace pour les ensembles de données volumineux. La figure 2.3 ci-dessous explique le bagging.

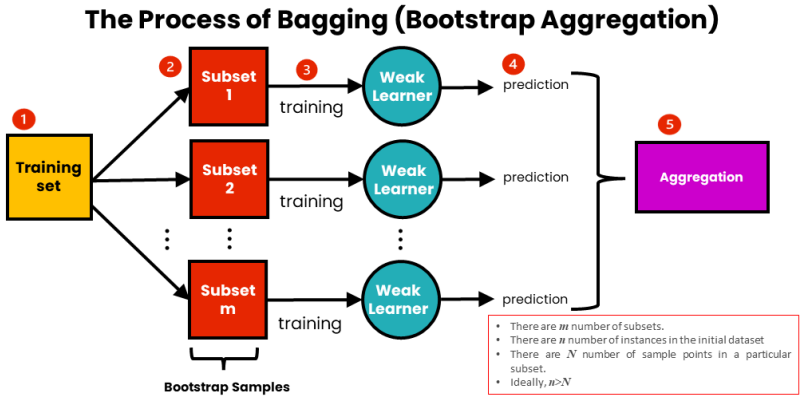

Figure 2.3*: Bagging*

2.7 Boosting

Le boosting est une technique d'ensemble en apprentissage automatique utilisée pour améliorer les performances des modèles prédictifs en combinant plusieurs modèles de base plus faibles pour former un modèle fort. Contrairement au bagging qui utilise des sous-ensembles aléatoires des données d'entraînement, le boosting utilise une approche itérative pour entraîner les modèles de base de manière séquentielle, en mettant l'accent sur les échantillons mal classés ou les résidus des modèles précédents.

Le processus de boosting se déroule généralement en plusieurs étapes :

Entraînement d'un modèle de base initial : Un premier modèle de base est entraîné sur l'ensemble des données d'entraînement.

Mise à jour des poids des échantillons : Les poids des échantillons d'entraînement sont mis à jour en fonction des erreurs de prédiction du premier modèle. Les échantillons mal classés reçoivent des poids plus importants, tandis que les échantillons correctement classés reçoivent des poids moins importants.

Entraînement de modèles supplémentaires : Des modèles de base supplémentaires sont entraînés séquentiellement sur les données d'entraînement pondérées. Chaque modèle se concentre sur les échantillons mal classés ou les résidus des modèles précédents.

Combinaison des modèles : Les prédictions de chaque modèle de base sont combinées pour former une prédiction finale. Dans les problèmes de classification, une méthode de vote pondéré est souvent utilisée, tandis que dans les problèmes de régression, les prédictions sont souvent pondérées en fonction de la performance des modèles individuels.

Le boosting vise à améliorer la performance du modèle global en se concentrant sur les erreurs des modèles précédents. En ajustant itérativement les modèles de base pour corriger les erreurs du modèle précédent, le boosting produit finalement un modèle fort qui peut mieux généraliser les données d'entraînement et fournir de meilleures performances de prédiction sur de nouvelles données. La figure 2.4 décrit le boosting.

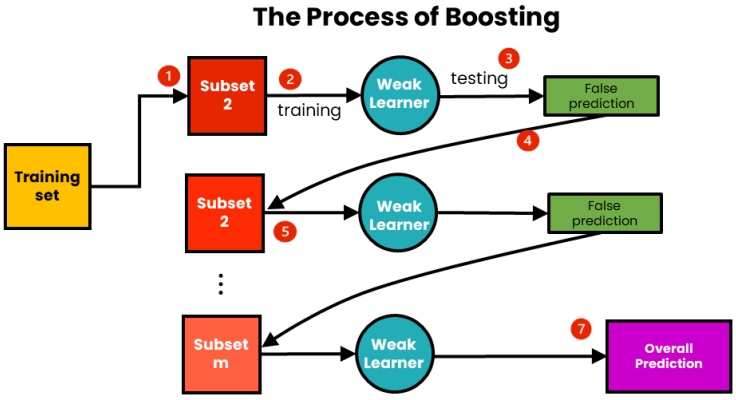

Figure 2.4: *Boosting*

2.8 Validation des modèles

Dans le contexte d'un problème de régression, notre but est également de concevoir un modèle capable de généraliser efficacement au-delà des données d'apprentissage disponibles. En d'autres termes, notre objectif est de créer un modèle qui puisse prédire avec précision les valeurs cibles pour de nouvelles observations, car une performance satisfaisante sur les données d'entraînement ne garantit pas nécessairement une bonne capacité de prédiction sur de nouvelles données.

Pour évaluer la capacité prédictive de notre modèle de régression, nous utilisons une variété de métriques adaptées à la nature continue des prédictions. Ces métriques nous permettent d'évaluer la précision de nos prédictions par rapport aux valeurs réelles et de déterminer dans quelle mesure notre modèle généralise efficacement.

2.8.1 Coefficient de détermination (R^2)

Le coefficient de détermination R^2 évalue la part de la variance de la variable dépendante que le modèle est capable d'expliquer. Il offre un aperçu de l'efficacité du modèle à représenter les fluctuations des données. Un R^2 proche de 1 suggère que le modèle

explique une grande partie de la variance, alors qu'un R^2 proche de 0 signale que le modèle ne fait pas mieux que la simple moyenne des valeurs observées pour expliquer la variance.

$$R^2 = 1 - \frac{SCR}{SCT} \tag{2.3}$$

Où :

SCR représente la somme des carrés résiduels, qui mesure la variance non expliquée par le modèle.

$$SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.4}$$

SCT représente la somme totale des carrés, qui mesure la variance totale de la variable cible.

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.5}$$

2.8.2 Erreur absolue moyenne (MAE)

En calculant la moyenne des écarts absolus entre les prédictions du modèle et les valeurs réelles, le MAE fournit une mesure directe de l'erreur moyenne de prédiction. Il est particulièrement utile pour évaluer la précision du modèle sans tenir compte de la direction des écarts.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2.6}$$

2.8.3 Erreur quadratique moyenne racine (RMSE)

Le RMSE est la racine carrée de la moyenne des carrés des écarts entre les prédictions et les valeurs réelles. Cette métrique exprime l'erreur dans les mêmes unités que la variable cible, ce qui la rend facilement interprétable. Comme il tient compte des erreurs de prédiction de manière proportionnelle à leur taille, le RMSE est une métrique populaire pour évaluer la précision globale du modèle.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.7}$$

2.9 Création de l'outil de prédiction : Framework Flask

Flask est une framework de développement web léger et flexible écrit en Python. Il est conçu pour être simple et facile à utiliser, permettant aux utilisateurs de créer des interfaces web rapidement et efficacement. Flask suit une architecture minimaliste et extensible, ce qui signifie qu'il ne force pas les développeurs à utiliser une structure spécifique et leur permet d'ajouter uniquement les composants dont ils ont besoin.

2.9.1 Caractéristiques principales de Flask

Léger et Minimaliste : Flask fournit les outils de base nécessaires pour construire une application web, sans ajout de fonctionnalités inutiles.

Extensible : Grâce à sa modularité, les développeurs peuvent intégrer des extensions pour ajouter des fonctionnalités.

Flexibilité : Flask ne dicte pas la manière dont les applications doivent être organisées, offrant ainsi une grande liberté aux développeurs pour structurer leur code comme ils le souhaitent.

Support pour les Templates : Flask utilise Jinja2, un moteur de templates puissant et flexible, pour permettre la création de pages web dynamiques.

Facilité d'utilisation : Avec une documentation complète et une communauté active, Flask est accessible aux développeurs de tous niveaux.

2.9.2 Backend de Flask(Python)

Flask est souvent utilisé comme un cadre backend pour le développement d'applications web en Python. En tant que tel, il offre un environnement robuste et flexible

pour la création de serveurs web et la gestion des requêtes HTTP. On utilise Flask pour créer des routes qui définissent les URL et les actions à effectuer lorsqu'un client fait une requête spécifique. Ces routes peuvent être associées à des fonctions Python qui manipulent les données et génèrent des réponses dynamiques pour l'utilisateur. Flask offre également un support intégré pour la gestion des cookies, les sessions utilisateur, l'authentification et l'accès aux bases de données, c'est pourquoi il est largement utilisé dans le développement. backend d'applications web de différentes tailles et complexités.

2.9.3 **Frontend de Flask(HTML,CSS et JavaScript)**

Bien que Flask soit principalement un cadre backend, il est couramment associé à des technologies frontend telles que HTML, CSS et JavaScript pour créer une expérience utilisateur complète. Dans le développement web moderne, Flask génère souvent des pages HTML dynamiques en utilisant des templates Jinja2 qui permettent d'insérer des données dynamiques dans des pages HTML. Ces pages peuvent être stylisées avec du CSS pour une présentation visuelle attrayante, tandis que JavaScript peut être utilisé pour ajouter des fonctionnalités interactives et des comportements dynamiques. Flask facilite l'intégration de ces technologies frontend en offrant des méthodes pour servir des fichiers statiques et en permettant une communication efficace entre le backend Flask et le frontend pour créer des applications web interactives et réactives.

2.10 **Conclusion**

Ce chapitre a été dédié à l'explication théorique des modèles adoptés dans ce projet, notamment la régression linéaire, le stacking, le bagging et le boosting. On a également abordé diverses techniques de sélection des variables et d'encodage des variables qualitatives, essentielles pour améliorer la qualité des modèles. En outre, on a discuté des critères de validation des modèles, indispensables pour évaluer leur performance et leur capacité de généralisation. On a également introduit le framework Flask, qui sera utilisé

pour le développement d'une interface web conviviale permettant d'interagir avec nos modèles prédictifs. Dans les prochains chapitres, on se penchera sur les résultats obtenus grâce à l'application de ces techniques, en analysant leur performance sur les jeux de données réels et en évaluant leur efficacité dans des scénarios pratiques.

Chapitre 3

TRAITEMENT ET ANALYSE EXPLORATOIRE DE DONNÉES

3.1 Introduction

Dans ce troisième chapitre, on va aborder les étapes de collecte, de prétraitement et d'exploration des données. On commencera par décrire la base de données mise en place, puis on examinera en détail les processus de nettoyage des données. Enfin, on présentera les résultats de la statistique descriptive ainsi que les méthodes de sélection des variables qu'on utilisera dans notre modélisation.

3.2 Collecte et saisie manuelle de données

La première étape du projet a été consacrée à l'exploration et à la collecte de données à partir de la base de données de la banque Zitouna, ainsi que des dossiers de financements en défaut de paiement de 2013 à 2017. On a élaboré des requêtes SQL pour extraire les variables pertinentes nécessaires à la construction de notre base de modélisation. Ces variables comprennent les engagements annuels des clients, classés en encours, impayés et contentieux, ainsi que des informations sur le type de financement et l'âge des clients.

En complément, on a procédé à la saisie manuelle d'autres variables cruciales pour la modélisation, telles que le revenu, l'autofinancement, le taux d'endettement, le montant des financements, et les données relatives aux garanties associées à chaque financement.

Cette approche a permis de rassembler une base de données exhaustive et diversifiée, nécessaire à une modélisation précise et efficace.

3.3 Calcul des LGDs et création de la base de données

On a constitué une nouvelle base de données à partir des données collectées, que l'on va employer pour appliquer les divers modèles. Cette étape se décompose en deux phases cruciales :

Sélection des variables pertinentes

Dans cette première phase, on doit choisir et identifier les variables pertinentes qui sont présents dans la base de données initiale fournie par la banque, ainsi que celles saisies manuellement.

Calcul des LGD (Loss Given Default) pour chaque financement

La deuxième phase consiste à calculer la LGD pour chaque financement en utilisant les formules préalablement mentionnées.

La base de données finale comprendra des variables explicatives comportementales, financières et des garanties, ainsi que la variable cible, la LGD.

3.4 Présentation de la base de données

La base de données finale comprend 16 variables et 238 entrées. Elle renferme des informations sur les financements ainsi que les situations financières et comportementales des clients. Ces clients ont connu des défauts entre les années 2013 et 2017.

3.4.1 Variables qualitatives

TRE (Tunisien résident à l'étranger) : Indique si le client réside à l'étranger ou non. Cette variable est importante pour comprendre les comportements de financement et les risques associés aux clients expatriés par rapport aux résidents locaux.

Caution solidaire et personnelle : Dénote la présence ou l'absence d'une caution solidaire et personnelle en tant que type de garantie. Une caution solidaire et personnelle signifie qu'une personne s'engage à rembourser la dette du client en cas de défaillance. Cette variable aide à évaluer le niveau de sécurité associé à chaque financement.

Hypothèque premier rang : Indique la présence ou l'absence d'une hypothèque de premier rang sur les biens immobiliers. Une hypothèque de premier rang est une garantie de premier niveau sur un bien immobilier, ce qui implique que le prêteur bénéficie d'une priorité sur les autres créanciers en cas de défaut de paiement. Cette variable est cruciale pour évaluer le risque et la valeur de la garantie.

Garantie financière : Indique la présence ou l'absence d'une garantie financière. Une garantie financière est une forme de sûreté où des actifs financiers (comme des dépôts ou des titres) sont mis en gage pour sécuriser le prêt. Cette variable permet d'estimer la liquidité et la sécurité des garanties offertes.

Secteur : Indique le secteur d'activité du client, soit public ou privé. Cette distinction est importante car elle peut influencer la stabilité de l'emploi et la capacité de remboursement du client.

Type de financement : Catégorise le type de financement accordé (par exemple, prêt hypothécaire, crédit à la consommation, prêt commercial, etc.). Cette variable aide à analyser les différents risques et comportements de remboursement associés à chaque type de financement.

CSP (Catégorie socioprofessionnelle) : Classe la catégorie socioprofessionnelle du client (par exemple, cadre, employé, artisan, etc.). Cette variable est utilisée pour comprendre le profil socio-économique des clients et son impact potentiel sur la LGD.

3.4.2 **Variables quantitatives**

Taux d'endettement : Indique le taux d'endettement du client au moment où le financement est accordé. Cette variable permet d'évaluer le rapport entre le niveau de dette du client et ses revenus, ainsi que d'estimer sa capacité de remboursement.

Nombre de financements : Représente le nombre total de financements obtenus par le client. Cette variable aide à évaluer l'expérience de crédit du client et sa capacité à gérer plusieurs prêts.

LtV Ratio (Loan to Value Ratio) : C'est le ratio entre le montant du financement et la valeur du bien financé. Il est égal à 1 moins le taux d'autofinancement. Ce ratio est crucial pour évaluer le risque de prêt par rapport à la valeur de l'actif sous-jacent.

Âge : L'âge du client au moment de la demande de financement. Cette variable peut influencer la capacité de remboursement et le profil de risque du client.

Année de financement : L'année où le financement est accordé. Cette variable permet d'examiner l'évolution des tendances de financement au fil du temps et les conditions économiques qui influent sur les prêts.

TIEG (Taux d'intérêt global) : C'est le taux d'intérêt global (TIEG) du financement. Cette variable inclut tous les coûts du prêt, offrant une vue complète du coût total du crédit pour le client.

Montant : Le montant du financement accordé. Cette variable est essentielle pour évaluer la taille des prêts et leur impact potentiel sur la solvabilité du client.

Revenu : Le revenu du client au moment de la demande de financement. Cette variable est utilisée pour évaluer la capacité de remboursement du client.

LGD (Loss Given Default) : La perte en cas de défaut de paiement du financement, et notre variable cible. Cette variable est utilisée pour quantifier les pertes potentielles pour la banque en cas de défaut du client.

3.5 Transformation et Analyse de données

3.5.1 Traitement des valeurs aberrantes

Les valeurs aberrantes sont des données extrêmes qui se distinguent nettement de la distribution des autres données de nos variables. Elles peuvent être détectées à l'aide de boxplots pour les différentes variables quantitatives. Ces boxplots sont des méthodes graphiques basées sur les quartiles et les intervalles interquartiles, permettant de définir les limites supérieure et inférieure au-delà desquelles les données sont considérées comme extrêmes.

3.5.2 Détection des valeurs aberrantes : Méthode de l'écart interquartile (IQR)

Cette méthode consiste à calculer :

- Le premier quartile (Q1) : la valeur en dessous de laquelle se trouvent 25 % des observations.

- Le troisième quartile (Q3) : la valeur en dessous de laquelle se trouvent 75 % des observations.

- L'écart interquartile (IQR) : la différence entre Q3 et Q1, représentant l'étendue de la moitié centrale des données.

- Écart interquartile (IQR) = Q3 - Q1.

- Limite inférieure = Q1 - 1,5 * IQR.

- Limite supérieure = Q3 + 1,5 * IQR.

Ensuite, on parcourt l'ensemble des valeurs de la variable. Les valeurs situées en deçà de la limite inférieure ou au-delà de la limite supérieure, comme indique la figure 3.1, sont considérés comme valeurs aberrantes.

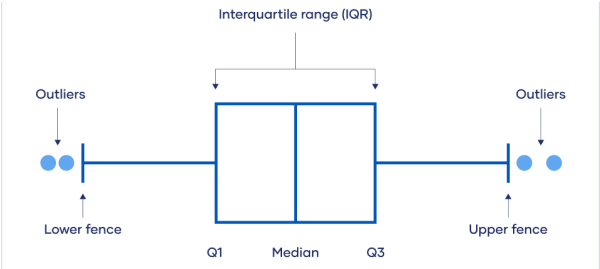

Figure 3.1*: Méthode de l'écart interquartile*

3.5.3 Regroupement et création de nouvelles variables

Dans le but de faciliter l'interprétation de certaines variables et d'améliorer la qualité de la modélisation, on a effectué plusieurs regroupements et créé de nouvelles variables en utilisant les données existantes.

Tout d'abord, on a regroupé les catégories socioprofessionnelles en une nouvelle classe appelée "Classe Activité", en utilisant la matrice de correspondance de la banque Zitouna. Cette démarche a permis de réduire le nombre de modalités de 45 à 12. Les nouvelles modalités comprenaient "Cadre Supérieur", "Cadre Moyen", "Commerçant", "Dirigeant d'entreprise", "Employé de bureau", "Enseignant", "Ingénieurs et cadres techniques", "Ouvrier", "Police et militaires", "Sans profession", "Techniciens" et "Autre".

Ensuite, on a regroupé les types de financements dans une nouvelle variable nommée "Catégorie", en se basant sur les documents de la banque Zitouna. Cela a permis de réduire le nombre de variables de 10 (telles que "Tamouil Manzel", "Tamouil Akarat Afrad", etc.) à 3 nouvelles catégories : "Immobilier", "Auto" ou "Consommation".

Par la suite, on a créé une nouvelle variable appelée "Note Garantie", en utilisant la base de garanties de la banque Zitouna. Cette variable comporte 6 modalités (A, B, C,

D, E, F) et dépend des garanties associées à chaque financement.

Enfin, on a introduit la variable "AtR" (Amount to Revenue), qui est simplement le montant du financement divisé par le revenu du client.

Ces ajustements ont été effectués dans le but d'optimiser la clarté et la compréhension des données, ainsi que la performance globale de notre modèle de modélisation.

3.5.4 **Analyse de données**

Variable cible LGD

La figure 3.2 montre la distribution de la variable cible LGD. Cette variable, que nous avons créée, représente la variable dépendante que nous cherchons à prédire dans les modèles de régression. La LGD est une variable numérique continue avec des valeurs comprises entre 0 et 1.

Observations de la Distribution :

Asymétrie :

La distribution de la LGD est fortement asymétrique, avec une concentration notable des valeurs près de zéro. Cela indique que la majorité des pertes en cas de défaut sont faibles.

Queue Longue :

On observe une queue longue vers la droite, signifiant qu'il y a des cas de défaut où les pertes sont significativement plus élevées, bien que moins fréquentes.

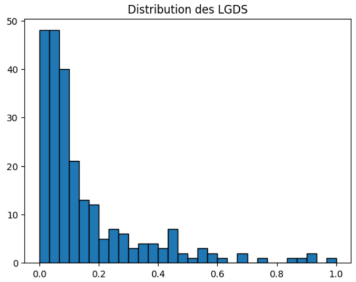

Figure 3.2*: Distribution de la variable cible LGD*

Categories de financement

La figure 3.3 ci-dessous montre la répartition des catégories de financement parmi les financements étudiés. Les catégories représentées sont ”Immobilier”, ”Consommation” et ”Auto”.

Observations :

Immobilier :

C’est la catégorie de financement la plus fréquente avec plus de 130 occurrences. Cela pourrait indiquer que la majorité des prêts dans le dataset sont destinés à l’immobilier.

Consommation :

La deuxième catégorie la plus fréquente, avec environ 75 occurrences. Les prêts à la consommation représentent une part significative du dataset.

Auto :

La catégorie la moins fréquente avec environ 25 occurrences, indiquant que les prêts auto sont moins courants dans ce dataset.

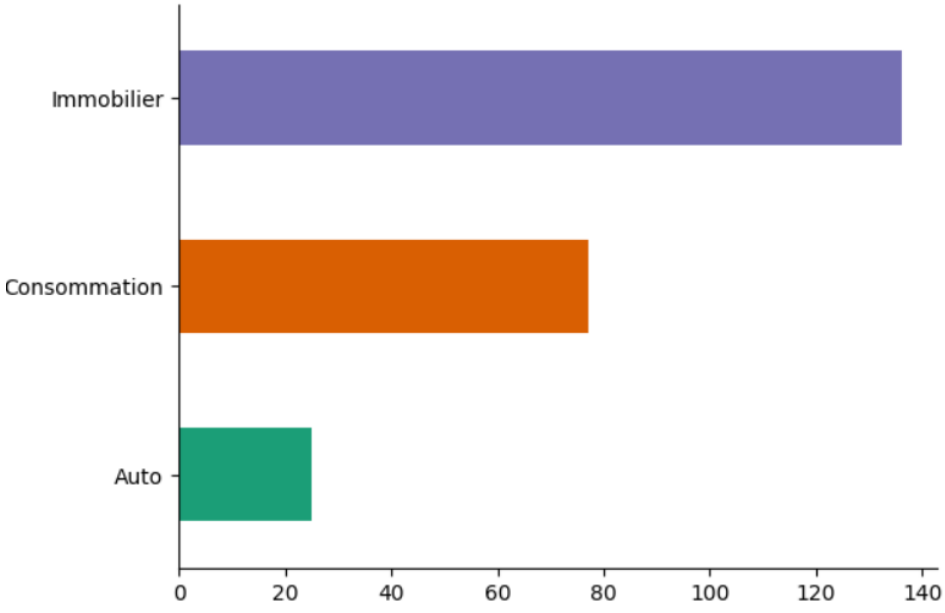

Figure 3.3: *Répartition des catégories*

Nombre de financements

La figure 3.4 montre la distribution du nombre de financements par emprunteur.

Observations :

Concentration des Financements :

La majorité des emprunteurs ont un seul financement (environ 130 occurrences), tandis que le nombre de financements décroît avec l'augmentation du nombre de financements.

Occurrences Multiples :

Il y a une diminution marquée après deux financements. Cependant, il existe des emprunteurs ayant jusqu'à six financements.

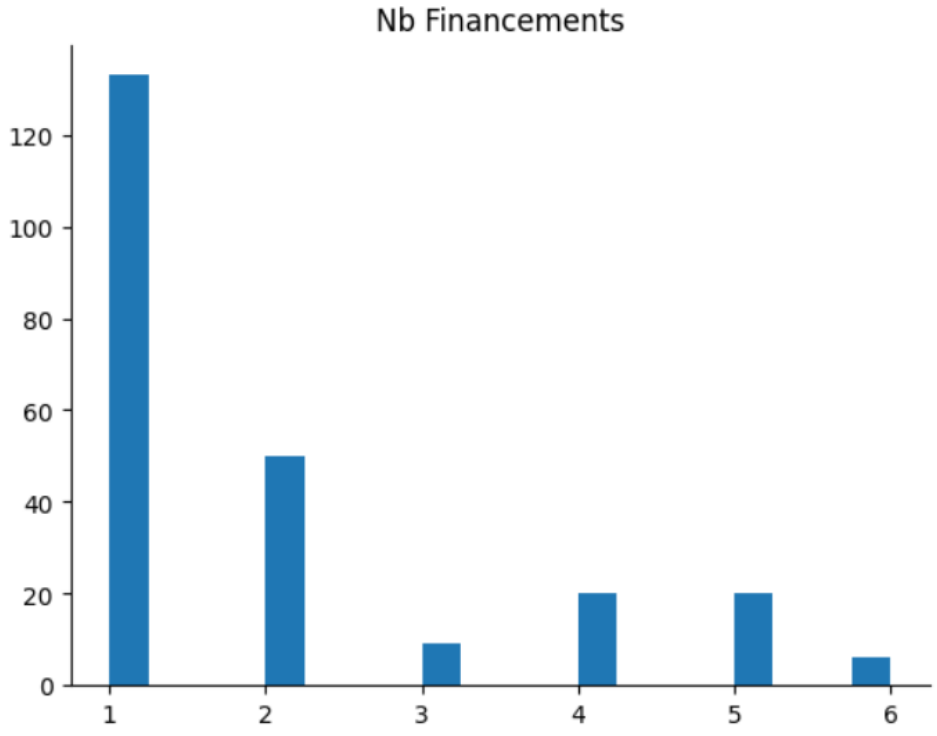

Figure 3.4: *Répartition de nombre de financements*

Note Garantie

La figure 3.5 montre la distribution de la "Note Garantie" des emprunteurs. Les notes sont classées de A à F et sont représentées par des barres horizontales de différentes couleurs. Chaque barre représente le nombre d'emprunteurs ayant reçu une certaine note.

Observations

Concentration des Notes :

La note B est la plus courante, avec environ 140 emprunteurs. Elle est attribuée aux garanties hypothécaires de premier rang. La note C suit avec environ 40 emprunteurs. Les notes F, E, A et D sont nettement moins fréquentes, avec la note A étant la moins courante.

Distribution :

La distribution montre une concentration élevée de notes B, ce qui suggère que la majorité des emprunteurs obtiennent cette note.

Les autres notes (A, C, E, F) sont beaucoup moins représentées, avec une forte diminution du nombre d'emprunteurs.

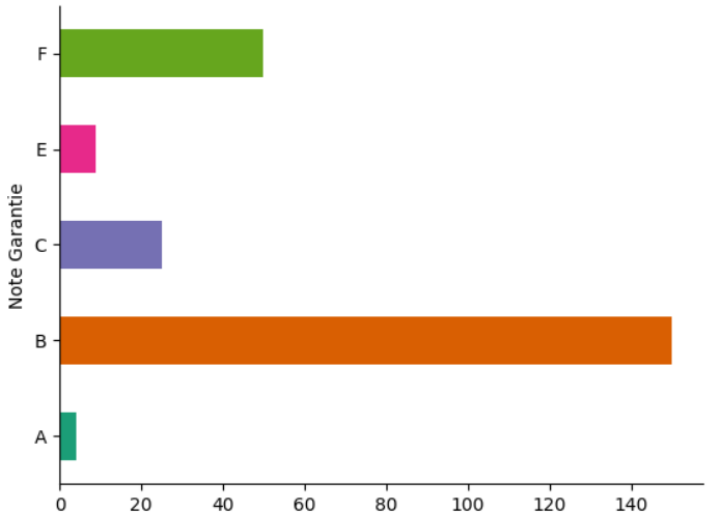

Figure 3.5: *Répartition des notes de garanties*

Matrice de corrélation

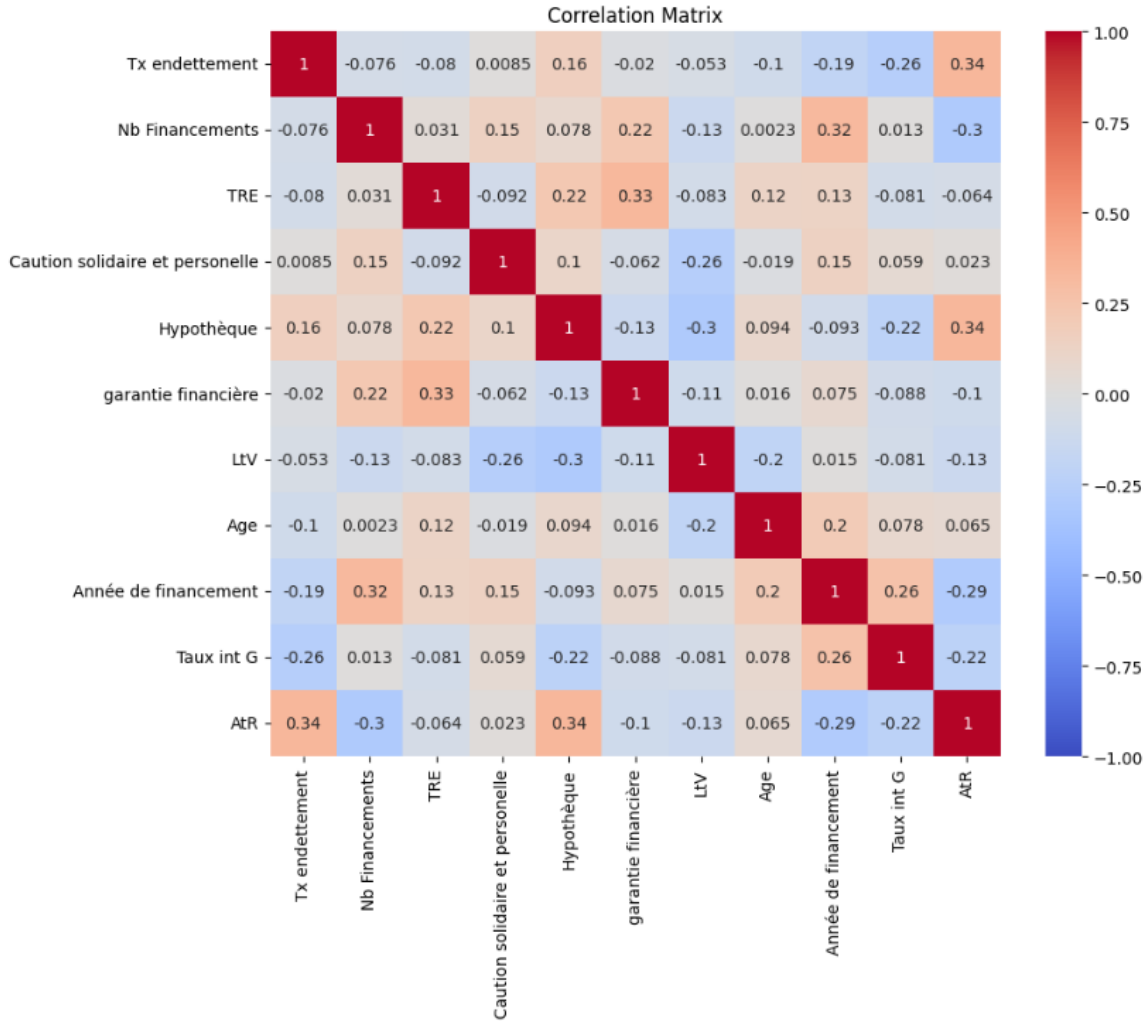

Figure 3.6*:* *Matrice de corrélation*

Cette matrice de corrélation (Figure 3.6) indique l'absence de corrélations fortes entre les variables. En effet, les coefficients de corrélation sont tous éloignés de 1 ou -1, valeurs qui signifieraient une corrélation positive ou négative parfaite. Les corrélations modérées les plus élevées, restent faibles. De plus, de nombreuses corrélations sont très faibles, indiquant peu ou pas de relation linéaire entre les variables. Ainsi, la matrice révèle que les variables étudiées n'entretiennent pas de relations linéaires significatives.

3.5.5 **Selection des variables**

Pour créer des modèles robustes, on va utiliser deux techniques pour la sélection des variables. La première méthode consiste à sélectionner les variables en utilisant la méthode d'élimination des caractéristiques en arrière (backward feature elimination), en se basant sur la valeur de p comme critère d'élimination. La deuxième méthode consiste à utiliser la RFS (Recursive Feature Selection) telle que décrite dans la Section 2.2.

La première méthode sera appliquée à la régression linéaire, tandis que la deuxième sera utilisée pour les méthodes d'ensemble.

En résumé, on va utiliser la régression linéaire avec la méthode d'élimination des caractéristiques en arrière basée sur la valeur de p pour sélectionner les variables. Pour les méthodes d'ensemble, on optera pour la Recursive Feature Selection (RFS) pour sélectionner les variables. Ces approches devraient nous permettre de construire des modèles robustes en identifiant les variables les plus importantes pour la prédiction.

3.5.6 **Sélection des variables par Backward Feature Elimination**

Nous avons utilisé la méthode de sélection des variables en utilisant la méthode d'élimination des caractéristiques en arrière (backward feature elimination). Nous avons choisi comme critère de division la p-value.

Les étapes à suivre sont :

- Commencer avec le modèle complet :

Inclure toutes les variables prédictives potentielles dans notre modèle de régression initial.

- Ajuster le modèle :

Ajuster le modèle à nos données et calculer les p-values pour chaque variable prédictive.

- Identifier la variable avec la p-value la plus élevée :

Trouver la variable prédictive avec la p-value la plus élevée, au-dessus de notre seuil

de signification.

- Retirer la variable :

Retirer cette variable du modèle.

- Réajuster le modèle :

Réajuster le modèle sans la variable retirée et recalculer les p-values pour les variables restantes.

- Répéter le processus :

Répéter les étapes jusqu'à ce que toutes les variables restantes aient des p-values en dessous du seuil de signification.

En suivant ces étapes, on peut progressivement éliminer les variables moins significatives pour améliorer la robustesse du modèle.

Les variables retenues pour notre modèle sont : TRE, Hypothèque, Secteur, Nombre de financements, Garantie financière, Classe Activité, Note Garantie, LtV, AtR, Âge et Catégorie.

3.5.7 **Selection des variables par Recursive Feature Selection**

On a également utilisé la méthode de sélection RFS décrite dans la Section 2.2.2. L'utilisation de cette technique nous a permis de générer des graphiques qui illustrent les variables retenues. Les variables ayant un rang égal à 1 seront sélectionnées par cette méthode comme indique les figures 3.7 et 3.8 ci-dessous.

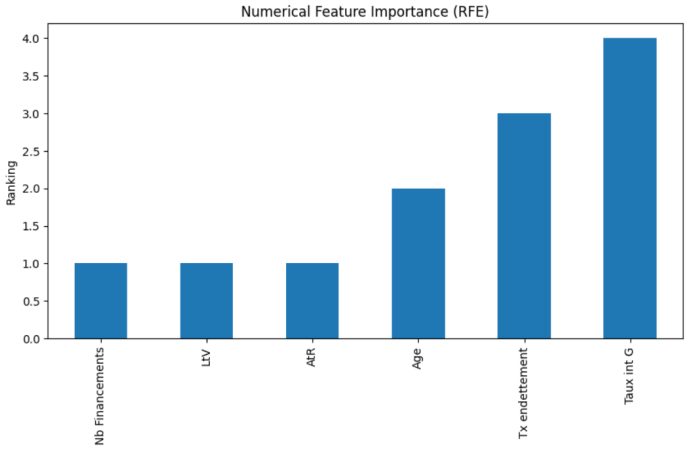

Figure 3.7: *Sélection des variables quantitatives avec RFS*

Les variables quantitatives retenues sont Nombre Financements, LtV et AtR.

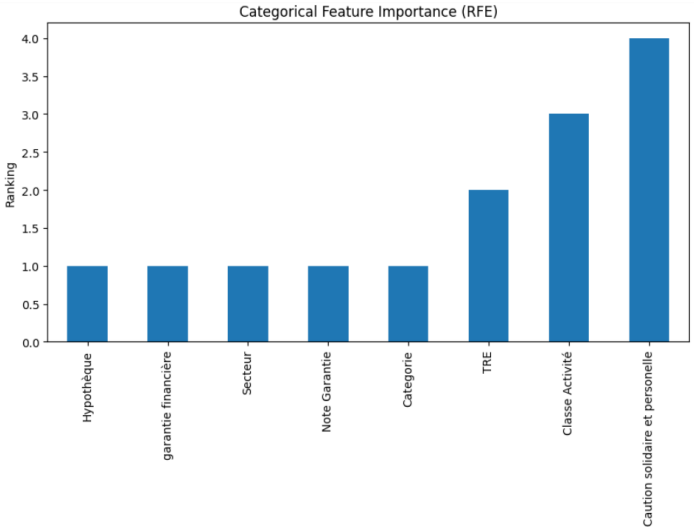

Figure 3.8: *Sélection des variables qualiitatives avec RFS*

Les variables qualiitatives retenues sont Hypothèque, garantie financière, Secteur, Note Garantie et Catégorie.

3.6 **Conclusion**

Dans ce chapitre, on a détaillé les étapes de la création et du prétraitement de notre base de données. Cela inclut notamment le calcul de la variable cible et le regroupement des modalités de certaines variables possédant un nombre très élevé de catégories. On a également présenté quelques graphiques décrivant notre jeu de données et exposé deux démarches de sélection de variables qu'on adoptera lors de la phase de modélisation.

Chapitre 4

MODÉLISATION ET CRÉATION DE L'INTERFACE GRAPHIQUE

4.1 Introduction

Ce chapitre présente deux sections principales : la première expose les résultats de la modélisation, tandis que la seconde décrit la création de l'interface graphique.

4.2 Modélisation de la LGD

4.2.1 Travail préliminaire

Avant d'entamer la phase de modélisation, on a commencé par le prétraitement des données en suivant les différentes étapes décrites dans le chapitre 3. Concernant le traitement des valeurs extrêmes et manquantes, on note que notre base de données ne contient pas de valeurs extrêmes ni de valeurs manquantes.

4.2.2 Résultats Empiriques

Dans cette partie, on va présenter les résultats obtenus par différents modèles de machine learning que l'on a implémentés. Les modèles utilisés incluent la régression linéaire, le stacking, le bagging et le boosting.

Pour chaque modèle, on a appliqué deux types d'encodage des variables qualitatives : le OneHotEncoder et le LabelEncoder, comme décrit dans la section 2.3.

— **Régression linéaire :** On a utilisé la méthode de Backward Feature Elimination pour sélectionner les caractéristiques pertinentes.

— **Méthodes d'ensemble (stacking, bagging et boosting) :** On a employé la sélection récursive de caractéristiques (Recursive Feature Selection) pour identifier les variables les plus significatives.

Ensuite, on a comparé les performances de ces modèles sélectionnés. La base de données a été séparée en deux parties : une partie pour l'entraînement (80 % des données) et une partie pour le test (20 % des données).

Enfin, une comparaison des résultats obtenus entre ces différents modèles sera réalisée afin de déterminer lequel offre la meilleure performance pour notre jeu de données.

4.2.3 Régression linéaire

Dans le cas de la régression linéaire, on a utilisé la méthode du Backward Feature Elimination (BFE) pour sélectionner les variables pertinentes, en fixant un seuil de significativité de 15 % basé sur les p-values des variables.

Encodage par Label Encoder

Le tableau 4.1 décrit la performance de model de régression linéaire en utilisant le Label Encoder comme méthode d'encodage de variables catégorielles.

Métrique	**Valeur**
R^2	0.552
MAE	0.1167
RMSE	0.1676

Table 4.1: *Tableau des métriques de Régression Linéaire (LE)*

Encodage par One Hot Encoder

Le tableau 4.2 décrit la performance de model de régression linéaire en utilisant le One Hot Encoder comme méthode d'encodage de variables catégorielles.

Métrique	**Valeur**
R^2	0.294
MAE	0.1132
RMSE	0.1663

Table 4.2: *Tableau des métriques de Régression Linéaire (1HE)*

4.2.4 Stacking

Dans le cadre des modèles d'ensemble, on a utilisé la méthode du Recursive Feature Selection (RFS) pour sélectionner les variables pertinentes.

Encodage par Label Encoder

Le tableau 4.3 décrit la performance de model de stacking en utilisant le Label Encoder comme méthode d'encodage de variables catégorielles.

Métrique	**Valeur**
R^2	0.3738
MAE	0.0967
RMSE	0.1468

Table 4.3: *Tableau des métriques de Stacking (LE)*

Encodage par One Hot Encoder

Le tableau 4.4 décrit la performance de model de régression linéaire en utilisant le One Hot Encoder comme méthode d'encodage de variables catégorielles.

Métrique	**Valeur**
R^2	0.4121
MAE	0.0924
RMSE	0.1422

Table 4.4: *Tableau des métriques de Stacking (1HE)*

4.2.5 **Bagging**

Encodage par Label Encoder

Le tableau 4.5 décrit la performance de model de bagging en utilisant le Label Encoder comme méthode d'encodage de variables catégorielles.

Métrique	**Valeur**
R^2	0.2443
MAE	0.1082
RMSE	0.1612

Table 4.5: *Tableau des métriques de Bagging (LE)*

Encodage par One Hot Encoder

Le tableau 4.6 décrit la performance de model de bagging en utilisant le One Hot Encoder comme méthode d'encodage de variables catégorielles.

Métrique	**Valeur**
R^2	0.2390
MAE	0.1097
RMSE	0.1618

Table 4.6: *Tableau des métriques de Bagging (1HE)*

4.2.6 **Boosting**

Encodage par Label Encoder

Le tableau 4.7 décrit la performance de model de boosting en utilisant le Label Encoder comme méthode d'encodage de variables catégorielles.

Métrique	**Valeur**
R^2	0.2411
MAE	0.1135
RMSE	0.1616

Table 4.7: *Tableau des métriques de Boosting (LE)*

Encodage par One Hot Encoder

Le tableau 4.7 décrit la performance de model de boosting en utilisant le One Hot Encoder comme méthode d'encodage de variables catégorielles.

Métrique	**Valeur**
R^2	0.2196
MAE	0.1177
RMSE	0.1639

Table 4.8: *Tableau des métriques de Boosting (1HE)*

4.2.7 **Comparaison des modèles**

On va maintenant examiner et comparer les diverses méthodes de modélisation, notamment la régression linéaire, le stacking, le boosting et le bagging, dans le but de déterminer le modèle le plus efficace. Voici un récapitulatif des métriques associées à chacune de ces méthodes.

Model	R^2	MAE	RMSE
Régression linéaire (LE)	**0.552**	0.1167	0.1676
Régression linéaire (1HE)	0.294	0.1132	0.1663
Stacking (LE)	0.3738	0.0967	0.1468
Stacking (1HE)	0.4121	**0.0924**	**0.1422**
Bagging (LE)	0.2443	0.1082	0.1612
Bagging (1HE)	0.2390	0.1097	0.1618
Boosting (LE)	0.2411	0.1135	0.1616
Boosting (1HE)	0.2196	0.1177	0.1639

Table 4.9*: Tableau de comparaison des modèles*

LE : Encodage par Label Encoder 1HE : Encodage par One Hot Encoder

Le tableau 4.9 présente une comparaison des modèles de régression en termes de coefficient de détermination (R^2), de l'erreur absolue moyenne (MAE) et de la racine carrée de l'erreur quadratique moyenne (RMSE). Parmi les modèles comparés, les deux modèles les plus performants sont la régression linéaire (LE) et le Stacking (1HE). La régression linéaire (LE) se distingue avec un R^2 de 0.552, le plus élevé parmi tous les modèles testés, ce qui indique qu'elle explique la plus grande proportion de la variance des données. Elle présente également un MAE de 0.1167 et un RMSE de 0.1676, ce qui en fait un modèle compétitif en termes de précision des prédictions.

Le modèle de Stacking (1HE), quant à lui, affiche des valeurs de MAE et de RMSE plus faibles, respectivement 0.0924 et 0.1422, indiquant une meilleure précision des prédictions. Cependant, son coefficient de détermination R^2 de 0.4121 est inférieur à celui de la régression linéaire (LE), signifiant qu'il explique moins de variance des données.

On a choisit le modèle de régression linéaire (LE), ce choix peut être justifié par son pouvoir explicatif supérieur, comme en témoigne son R^2 plus élevé. Bien que d'autres modèles présentent des erreurs de prédiction légèrement inférieures, le R^2 élevé de la

régression linéaire (LE) la rend particulièrement attrayante en raison de sa capacité à mieux expliquer les variations des données.

4.2.8 **Interprétation de model choisi**

Coefficients de model

Variable	**Coefficient**	**P-value**
Nb Financements	-0.0175	0.041
TRE	-0.0534	0.124
Hypothèque	-0.2240	0.005
Garantie financière	-0.2361	0.036
LTV (Loan to Value)	0.1385	0.065
Age	-0.0017	0.108
Secteur	-0.0572	0.153
Note Garantie	0.0783	0.000
Classe Activité	0.0108	0.002
Catégorie	-0.0911	0.017
AtR	0.0018	0.022

Table 4.10: *Coefficients et p-values des variables du modèle de régression linéaire*

Le tableau 4.10 montre le coefficient de régression ainsi que la p-value pour les variables de model de régression linéaire choisi. Voici une interprétation de ces coefficients.

Nb Financements :

Coefficient : -0.0175

Interprétation : Chaque augmentation du nombre de financements est associée à une diminution de 0.0175 de la LGD. Cela suggère que plus il y a de financements, plus la LGD diminue.

TRE :

Coefficient : -0.0534

Interprétation : Si un client est résident à l'étranger la LGD diminue de 0.0534 .

Hypothèque :

Coefficient : -0.2240

Interprétation : La présence d'une hypothèque est associée à une diminution de 0.2240 unités de la LGD. Cela montre que les hypothèques sont liées à une baisse significative de la LGD.

Garantie financière :

Coefficient : -0.2616

Interprétation : La présence d'une garantie financière est associée à une diminution de 0.2616 unités de la LGD. Les garanties financières réduisent significativement la LGD.

LTV (Loan to Value) :

Coefficient : 0.1385

Interprétation : Une augmentation de 10 % du ratio LTV est associée à une augmentation de 0.01385 unités de la LGD. Un LTV plus élevé augmente la LGD.

Age :

Coefficient : -0.0017

Interprétation : Chaque année supplémentaire d'âge est associée à une diminution de 0.0017 unités de la LGD. Les emprunteurs plus âgés tendent à avoir une LGD légèrement inférieure.

Secteur :

Coefficient : -0.0572

Interprétation : Le secteur est associé à une diminution de 0.0572 unités de la LGD. Cela suggère que les clients de secteur public ont un LGD plus bas que ceux de secteur privé.

Note Garantie :

Coefficient : 0.0783

Interprétation : Une augmentation d'une unité de la note de garantie est associée à une

augmentation de 0.0783 unités de la LGD. Plus la note est baisse plus la LGD augmente.

AtR :

Coefficient : 0.0018

Interprétation : Une augmentation d'une unité de l'AtR est associée à une augmentation de 0.0018 unités de la LGD. Un AtR plus élevé augmente la LGD.

F test

La valeur p du test F est de 7,67e-34. Cette valeur extrêmement faible indique que le modèle de régression est hautement significatif. Autrement dit, il y a une probabilité extrêmement faible que les coefficients des variables explicatives soient tous égaux à zéro simultanément. Cela signifie que les variables indépendantes dans le modèle apportent une contribution significative à la prédiction de la LGD.

Analyse des résidus

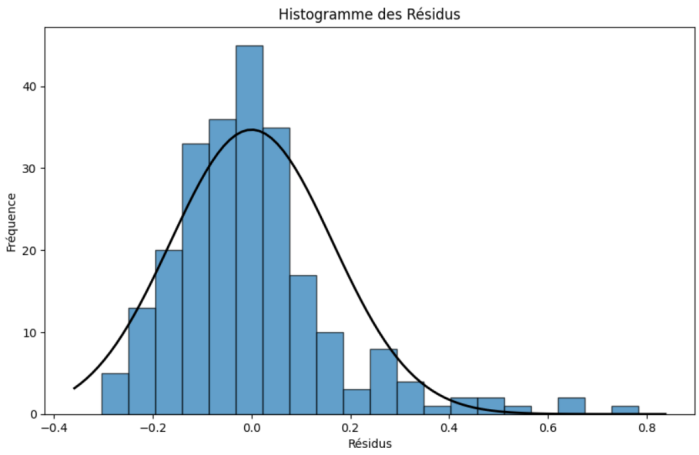

Figure 4.1*: Histogramme des résidus*

L'histogramme des résidus présenté ci dessus dans la figure 4.1 montre que la distribution des résidus suit approximativement une courbe normale, indiquée par la courbe de densité gaussienne superposée. Les résidus sont majoritairement centrés autour de zéro,

ce qui suggère que le modèle n'a pas de biais systématique et que les erreurs de prédiction sont aléatoires et relativement faibles. Cependant, une légère asymétrie à droite et quelques résidus plus importants (entre 0.4 et 0.8) indiquent la présence de points aberrants ou de facteurs non modélisés qui affectent les prédictions. En somme, l'histogramme révèle que le modèle est globalement bien ajusté.

4.3 Intérface graphique de prédiction de LGD

4.3.1 Interêt de l'interface graphique

Développer une interface graphique intuitive pour implémenter et utiliser le modèle de LGD (Loss Given Default) est essentiel pour plusieurs raisons :

Accessibilité et Utilisation Simplifiée

Facilitation de l'utilisation : Accessible même pour ceux sans expertise technique.

Réduction des erreurs : Moins de risques d'erreurs humaines.

Efficacité et Gain de Temps

Automatisation : Tâches répétitives automatisées, permettant aux analystes de se focaliser sur des tâches de plus grande importance.

Rapidité d'analyse : Décisions plus rapides grâce à des visualisations claires et des résultats instantanés.

Prise de Décision Informée

Visualisation : Données et résultats présentés de manière claire et compréhensible.

Accessibilité des résultats : Décideurs accédant facilement aux informations sans rapports complexes.

Collaboration et Communication

Partage d'informations : Résultats et analyses facilement partageables.

Uniformité des rapports : Présentation cohérente des informations.

Flexibilité et Adaptabilité

Personnalisation : Interface adaptable aux besoins spécifiques.

Mise à jour facile : Améliorations et mises à jour du modèle simplifiées.

4.3.2 Éléments de l'interface graphique

Page d'accueil

La page d'accueil de l'interface ci-dessous comme montre la figure 4.2 présente un design épuré et cohérent avec une palette de couleurs verte et blanche, renforçant l'identité visuelle de la banque. Le logo centralisé capte l'attention et favorise la reconnaissance de la marque. Deux boutons principaux, "Predict LGD" et "Dashboard", sont bien visibles et facilement accessibles. "Predict LGD" permet de prédire le Loss Given Default via le model choisi. "Dashboard" offre un accès à un tableau de bord analytique. L'interface, intuitive et accessible même pour les non-experts, est bien agencée avec des étiquettes claires, minimisant les erreurs et facilitant l'utilisation.

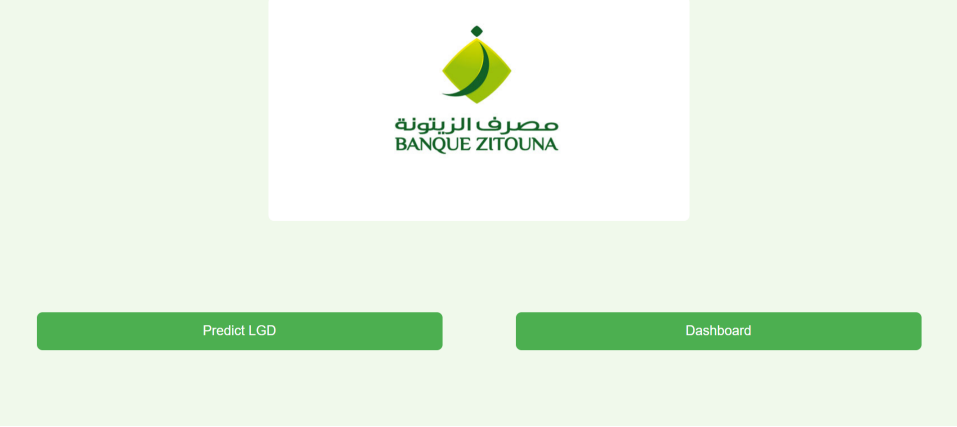

Figure 4.2*: Page d'accueil*

Page de prédiction LGD

Cette page (Figure 4.3) comprend des champs à remplir qui correspondent aux variables sélectionnées pour notre modèle. Ces champs serviront d'entrées pour le modèle.

Elle comporte également un bouton "Calculer LGD prédit", permettant de calculer la LGD (Loss Given Default) en utilisant des données fournies.

Figure 4.3*: Prédiction LGD*

Résultat de prédiction

Cette page (Figure 4.4) contient le résultat de la prédiction de la LGD selon notre model et nos variables d'entrée.

Figure 4.4*: Résultat de prédiction*

Tableau de bord

Cette page contient un tableau de bord interactif qui présente plusieurs visualisations comme indique la figure 4.5.

Distribution de LGD (Loss Given Default) :

Un histogramme montrant la distribution des pertes données en cas de défaut. La majorité des valeurs de LGD se situent entre 0 et 0,1, avec quelques valeurs plus élevées jusqu'à 1. Cela indique que les pertes sont généralement faibles, mais il y a des cas où les pertes peuvent être plus importantes.

Distribution de Secteur :

Un diagramme circulaire montrant la répartition entre les secteurs privé et public. Le secteur privé représente 90,8% tandis que le secteur public représente 9,24%.

Distribution de Nombre de financements :

Un autre diagramme circulaire montrant la distribution de nombre de financements pour chaque client.

Distribution de Categorie :

Ce diagramme circulaire montre la répartition des différentes catégories de financements. Immobilier (57,1%), Consommation (32,4%), et Auto (10,5%).

Distribution de Note Garantie :

Un histogramme montrant la distribution des notes de garantie. La majorité des garanties sont notées B, suivies par F, C, E, et A.

Moyenne LGD par Categorie :

Un graphique à barres montrant la moyenne de LGD pour chaque catégorie de financement. Les financements de consommation ont la moyenne LGD la plus élevée, suivis par Auto et Immobilier.

Ce tableau de bord permet de visualiser et d'analyser notre jeu de données de manière claire et concise, facilitant la prise de décisions informées.

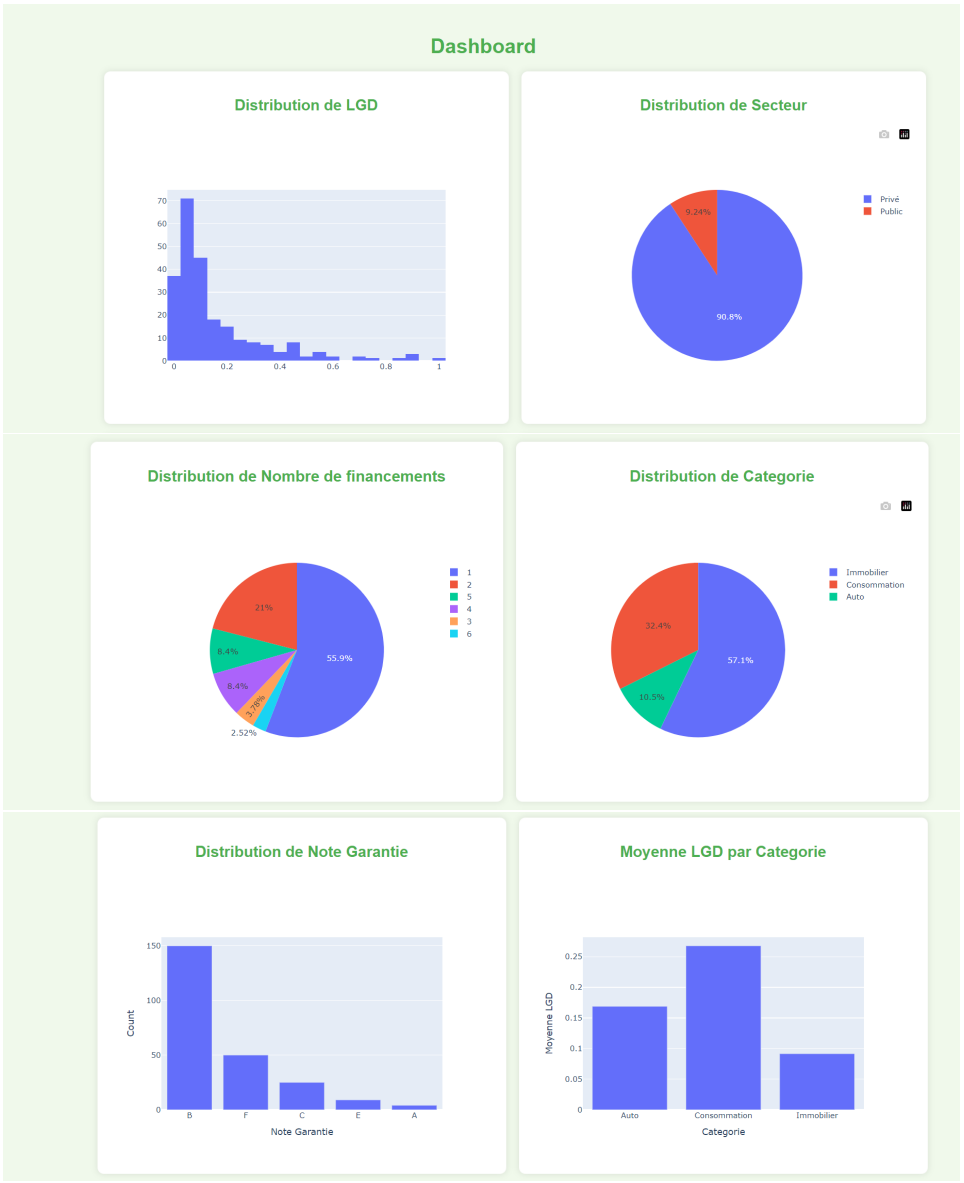

Figure 4.5: *Tableau de bord*

4.4 **Conclusion**

Dans ce chapitre, on a présenté les résultats de la modélisation en comparant différents modèles de régression et en sélectionnant le meilleur modèle lors de la première phase. Ensuite, on a analysé le modèle choisi en profondeur. Enfin, on a conclu en présentant l'interface graphique, qui comprend un outil de prédiction de la LGD et un tableau de bord.

Chapitre 5

CONCLUSION GÉNÉRALE

En conclusion, le projet "Modélisation et Développement d'un Outil de Prédiction de la LGD sous IFRS 9" a permis de concevoir un outil performant et précis pour prédire la Loss Given Default conformément aux normes IFRS 9. Ce travail a impliqué une compréhension approfondie des exigences de la Banque Centrale de Tunisie, ainsi qu'une application rigoureuse de techniques avancées de modélisation et d'apprentissage automatique.

Les quatre chapitres du rapport ont méthodiquement abordé chaque aspect essentiel du projet, depuis le contexte initial et les concepts fondamentaux, jusqu'à l'application pratique des modèles de prédiction. L'analyse exploratoire des données et le développement de modèles d'apprentissage automatique robustes ont été au cœur du processus, garantissant des prédictions fiables et automatisées. Pour ce projet, un modèle de régression linéaire a été choisi pour sa simplicité et son efficacité dans la prédiction de la LGD, et l'outil a été développé en utilisant Flask, un framework léger et flexible pour le développement web.

Cette initiative non seulement renforce la capacité de gestion des risques financiers des institutions, mais elle démontre également l'efficacité et la pertinence de l'intégration de technologies de pointe dans le domaine de la finance. Ainsi, l'outil de prédiction de la LGD développé dans ce projet constitue une avancée significative vers une meilleure gestion des risques et une conformité accrue aux réglementations financières internationales.

Chapitre 6

ANNEXES

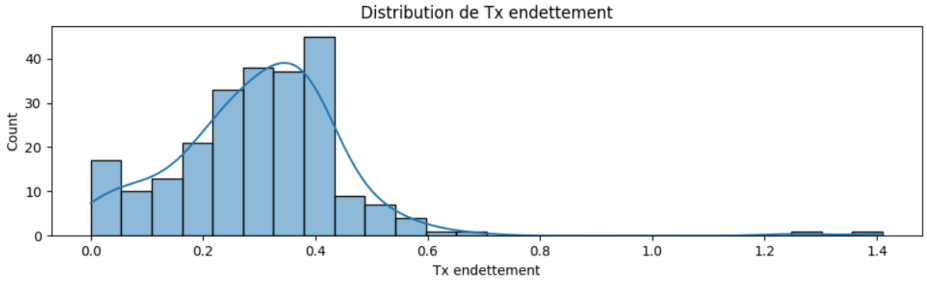

Figure 6.1: *Distribution de taux d'endettement*

La distribution du taux d'endettement (Figure 6.1) montre que la majorité des valeurs se situent entre 0.2 et 0.6. La courbe de densité indique une distribution légèrement asymétrique avec une longue queue à droite.

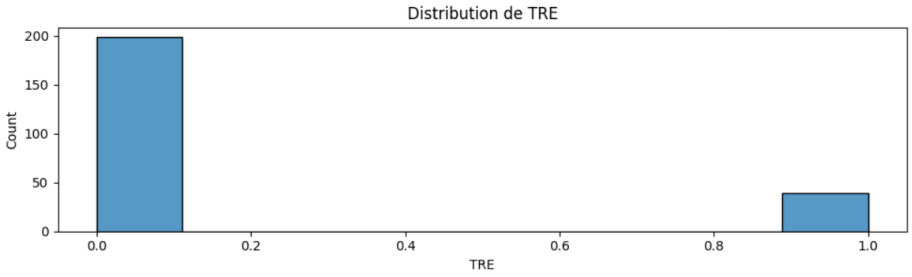

Figure 6.2: *Distribution de TRE*

La distribution de TRE (Figure 6.2) est fortement concentrée en 0, avec une petite proportion de valeurs de 1. Cela suggère que la plupart des observations sont résident à la Tunisie.

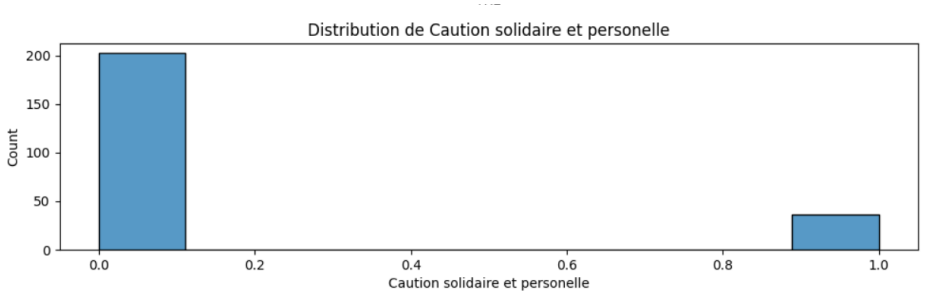

Figure 6.3: *Distribution de caution solidaire et personnel*

La figure 6.3 montre que la majorité des valeurs de caution solidaire et personnelle sont à 0, avec une petite fraction à 1. Cela montre que peu de cas incluent cette forme de caution.

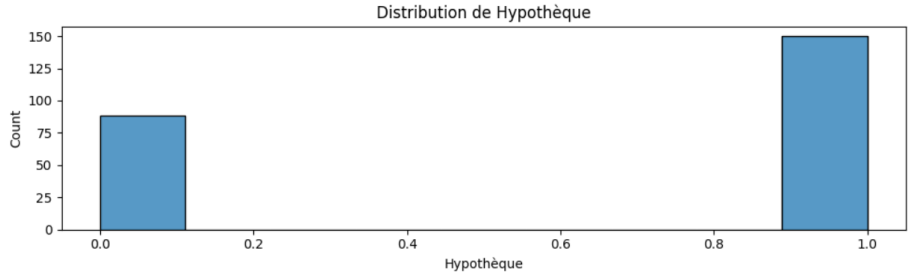

Figure 6.4: *Distribution d'hypothèque*

La distribution d'hypothèque (Figure 6.4) est binaire avec la plupart des valeurs à 1, indiquant que beaucoup des cas incluent une hypothèque.

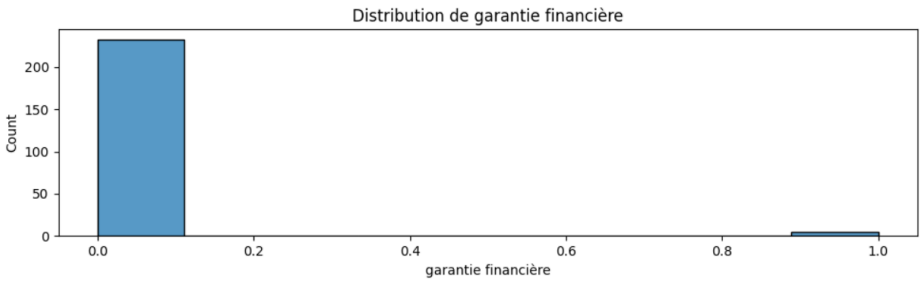

Figure 6.5*: Distribution de garantie financière*

Dans la figure 6.5, la distribution de garantie financière montre que la majorité des valeurs sont à 0, avec très peu à 1, signifiant que peu de cas utilisent une garantie financière.

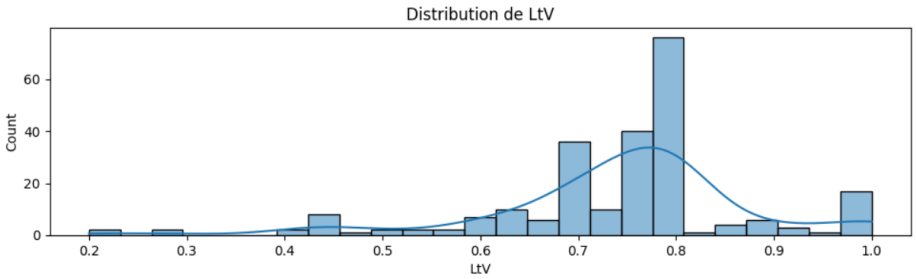

Figure 6.6*: Distribution de Loan to Value Ratio*

Dans la figure 6.6 la distribution du ratio Loan to Value (LTV) montre une concentration autour de 0.6 à 0.9, avec une distribution globalement symétrique. Quelques valeurs extrêmes existent en dessous de 0.3 et au-dessus de 1.

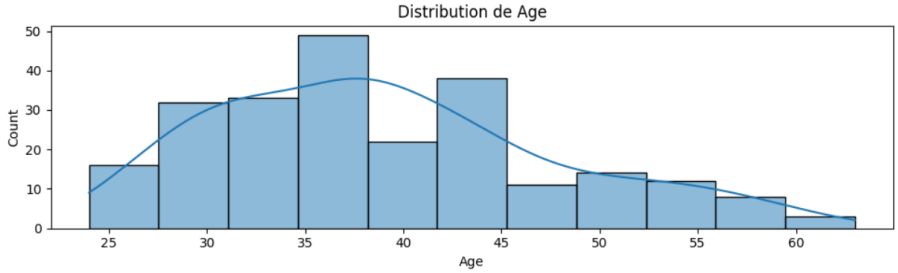

Figure 6.7*: Distribution d'âge*

La figure 6.7 montre la distribution des âges des individus dans l'ensemble de données. La majorité des individus ont entre 35 et 45 ans, avec une distribution légèrement asymétrique vers la droite, indiquant une plus grande concentration d'individus plus jeunes.

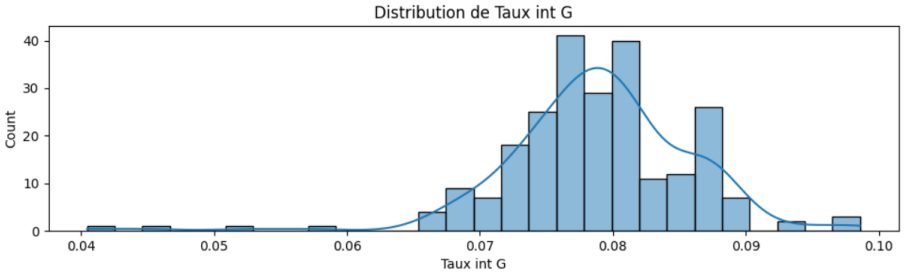

Figure 6.8*: Distribution de taux d'interet général*

La figure 6.8 illustre la distribution des taux d'intérêt généraux (Taux int G) dans l'ensemble de données. La distribution semble être normale, centrée autour de 0.07, avec une dispersion autour de cette valeur. Quelques valeurs extrêmes existent en dehors de cette plage principale.

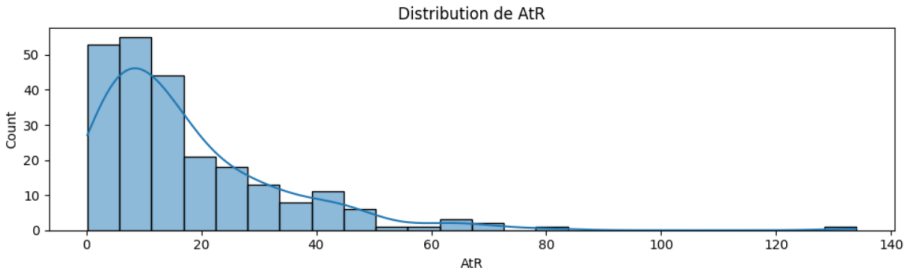

Figure 6.9*: Distribution de Amount to Revenue ratio*

La figure 6.9 représente la distribution du ratio montant/revenu (AtR). La majorité des valeurs se concentrent autour de 0 à 20, avec une queue longue à droite, indiquant des ratios élevés dans certains cas particuliers.

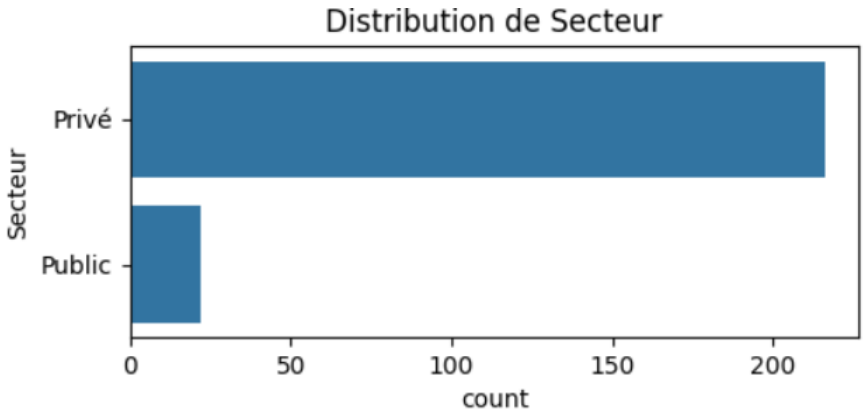

Figure 6.10*: Distribution de secteur d'activité*

La figure 6.10 montre la distribution des secteurs d'activité (public et privé). Il y a une prédominance des individus travaillant dans le secteur privé, beaucoup plus nombreux que ceux dans le secteur public.

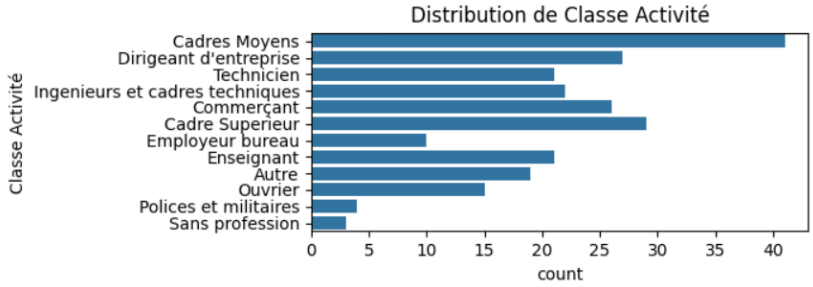

Figure 6.11*: Distribution de classe d'activité*

La figure 6.11 illustre la distribution des différentes classes d'activité professionnelle des individus. Les catégories "Cadres Moyens" et "Cadre Supérieur" semblent être les plus représentées, tandis que des catégories comme "Sans profession" ou "Policiers et militaires" sont les moins représentées.

Bibliographie

[1] Circulaire aux banques et aux etablissements financiers no. 2018-06. Banque centrale de Tunisie, 2018.

[2] Norme internationale d'information financiere 9 : Instruments financiers, 2014.

[3] AMEF CONSULTING. In point réglementation : Norme comptable IFRS 9. https ://www.amef-consulting.com/2020/01/06/point-reglementation-norme-comptable-ifrs-9/, 2020.

[4] ilboursa.com. L'enjeu et les prérequis de l'implémentation de la norme IFRS 9 pour le secteur financier tunisien. https ://www.ilboursa.com/marches/lenjeu-et-les-prerequis-de-limplementation-de-la-norme-ifrs-9-pour-le-secteur-financier-tunisien$_2$4237, *septembre*2020.

[5] Tiziano Bellini. *IFRS 9 and CECL Credit Risk Modelling and Validation*. Academic Press, 2018.

[6] Nadeem A. Siddiqi and Meiqing Zhang. A general methodology for modeling loss given default. *Journal of Banking & Finance*, 32(1) :74–82, 2008.

[7] Xiao Yao, Jonathan Crook, and Galina Andreeva. Modeling loss given default in sas/stat®. In *Proceedings of the SAS Global Forum 2014 Conference*, Washington, DC, 2014. SAS Institute Inc. Paper 1593-2014.

[8] Eugen Töws. *Advanced Methods for Loss Given Default Estimation*. PhD thesis, Universität zu Köln, Klimowka, 2015.

[9] Aida Salko and Rita D'Ecclesia. Decomposing loss given default : A closer look at recovery patterns. *Department of Economics and Social Sciences, Department of Statistics, Sapienza University of Rome*, 2018.