Université de Montpellier - Faculté D'Économie
Mention Monnaie Banque Finance Assurance
Parcours Actuariat
2023-2024

Application du Machine Learning au Provisionnement en Assurance : Modèle de détection automatique des dossiers à clôturer

---

Réalisé par : AZOULAY Nathan
Tuteur académique : M. SADEFO KAMDEM Jules
Tuteur d'entreprise : M. BOKRETA Rachid

Remerciements

Je tiens à exprimer ma sincère gratitude au Professeur Jules SADADEFO KAMDEM, responsable de notre Master 2, ainsi qu'à mon professeur et tuteur de stage, M. Rachid BOKRETA. Leur dévouement, leurs conseils judicieux et leur bienveillance ont été d'une aide et d'un soutien précieux tout au long de ce projet.

Je souhaite également remercier mes collègues pour leurs remarques et directives pertinentes qui ont grandement contribué à l'avancement de ce projet. Enfin, un merci tout particulier à ma famille pour leur soutien et leurs encouragements, sans lesquels cette réalisation n'aurait pas été possible.

Résumé

Ce mémoire plonge dans l'application des techniques avancées de machine learning pour transformer la gestion des dossiers sinistres au sein des compagnies d'assurance. L'observation principale révèle que certains dossiers sinistres restent ouverts malgré leur éligibilité à la clôture, engendrant ainsi des inefficacités opérationnelles et des charges administratives superflues. Ce projet vise à développer un modèle de détection automatique capable d'identifier les dossiers nécessitant une clôture, en exploitant les outils sophistiqués du machine learning.

Le travail est articulé autour des axes suivants :

1. **Analyse des Données Historiques :** Une étude exhaustive des données historiques des dossiers sinistres a été menée pour déceler les motifs récurrents et les caractéristiques des dossiers injustement non clôturés. Cette analyse englobe l'examen des types de sinistres, des montants engagés, des délais de traitement, ainsi que d'autres variables pertinentes. L'objectif est de discerner les patterns sous-jacents et les facteurs communs liés aux dossiers nécessitant une clôture.
2. **Développement de Modèles Prédictifs :** Grâce aux techniques avancées de machine learning telles que la régression logistique Elastic Net, les forêts aléatoires et les arbres de décision, un modèle prédictif a été conçu pour classer automatiquement les dossiers sinistres selon leur besoin de clôture. Ce modèle est conçu pour être à la fois robuste et précis, avec une capacité à généraliser sur des données nouvelles tout en minimisant les erreurs de classification, c'est-à-dire les faux positifs et les faux négatifs.
3. **Implémentation et Évaluation :** Le modèle développé a été intégré au système d'information de l'entreprise et soumis à une évaluation rigoureuse sur des jeux de données réels. Cette évaluation a inclus des métriques de performance telles que la précision, le rappel, la spécificité, ainsi qu'une analyse détaillée des faux positifs et des faux négatifs. Des études de cas ont également été réalisées pour illustrer l'efficacité du modèle dans un contexte opérationnel réel.
4. **Impact et Implications :** En cas de succès, ce modèle de détection automatique pourrait transformer les processus internes de gestion des sinistres, en réduisant les délais de clôture et en assurant une provision plus fiable. Il permettrait aussi de rationaliser les coûts administratifs associés à la gestion des sinistres, tout en améliorant l'efficacité opérationnelle globale de l'entreprise.

Les modèles de machine learning développés ont démontré leur pertinence et leur efficacité, comme le confirme la validation croisée. Les économies potentielles en provisions pourraient atteindre environ 80 000 euros pour les dossiers identifiés comme nécessitant une clôture, ce qui représente environ 5,8 % des dossiers actuellement ouverts.

Mots clés : arbre de décision, machine learning, régression logistique, Elastic Net, gestion des sinistres, prévision, optimisation, analyse de données, provision, efficience opérationnelle, réduction des coûts

Abstract

This thesis delves into the application of advanced machine learning techniques to revolutionize the management of insurance claims within insurance companies. It has been observed that certain claims remain open despite being eligible for closure, leading to operational inefficiencies and unnecessary administrative burdens. The primary objective of this project is to develop an automatic detection model to identify claims that require closure, utilizing sophisticated machine learning tools.

The thesis is organized around the following key areas :

1. **Historical Data Analysis :** An in-depth analysis of historical claims data was conducted to uncover recurring patterns and characteristics of unjustifiably open claims. This analysis includes studying the types of claims, amounts involved, processing times, and other relevant variables. The goal is to identify the underlying patterns and common factors associated with claims that need closure.
2. **Development of Predictive Models :** Using advanced machine learning techniques such as Elastic Net logistic regression, random forests, and decision trees, a predictive model was developed to automatically classify claims based on their need for closure. The model is designed to be both robust and accurate, with the capability to generalize to new data while minimizing classification errors, i.e., false positives and false negatives.
3. **Implementation and Evaluation :** The developed model was integrated into the company's information system and rigorously evaluated on real datasets. This evaluation included performance metrics such as precision, recall, specificity, as well as a detailed analysis of false positives and false negatives. Case studies were also conducted to illustrate the model's effectiveness in a real operational setting.
4. **Impact and Implications :** If successful, this automatic detection model could transform internal claims management processes, reducing closure times and ensuring more reliable provisioning. It would also streamline administrative costs associated with claims management and improve overall operational efficiency.

The developed machine learning models proved to be relevant and effective, as demonstrated by cross-validation. Potential savings in provisions could amount to approximately 80,000 euros for claims identified as needing

closure, representing about 5.8% of currently open claims.

Keywords : decision tree, machine learning, logistic regression, Elastic Net, claims management, forecasting, optimization, data analysis, provision, operational efficiency, cost reduction

Table des matières

Chapitre 1

Introduction

Le provisionnement en assurance constitue un pilier fondamental pour l'évaluation et la gestion des réserves nécessaires à la couverture des indemnités futures pour sinistres en cours. Selon l'Autorité de Contrôle Prudentiel et de Résolution (ACPR), le provisionnement est défini comme suit : *"Ce sont les provisions techniques correspondant à l'évaluation des engagements des entreprises d'assurance et de réassurance vis-à-vis des assurés et des bénéficiaires de contrats."*

Il apparaît donc évident que ce processus est crucial non seulement pour garantir la solvabilité de l'assureur, mais aussi pour assurer une gestion efficace des fonds et la conformité aux régulations en vigueur.

Avec l'avènement rapide des technologies et l'explosion des volumes de données disponibles, les méthodes traditionnelles de provisionnement montrent leurs limites. C'est dans ce contexte que la Data Science, et plus précisément le Machine Learning, offrent des perspectives innovantes pour affiner la précision des estimations de provisions et optimiser leur gestion.

1.1 L'Importance du Provisionnement

Stabilité Financière

Le provisionnement est un élément clé pour la stabilité financière des compagnies d'assurance. Il assure la solvabilité, permet de se conformer aux réglementations et optimise les fonds disponibles. Une estimation précise des réserves garantit que les assureurs peuvent couvrir les sinistres futurs, respecter les exigences légales et gérer efficacement leurs investissements. En

somme, le provisionnement est un levier crucial pour la robustesse financière, la fiabilité opérationnelle et la rentabilité des compagnies d'assurance.

Avant tout, le provisionnement joue un rôle vital dans la stabilité financière des compagnies d'assurance. Une évaluation précise des réserves est essentielle pour plusieurs raisons majeures :

— **Maintien de la Solvabilité :** La solvabilité d'une compagnie d'assurance se mesure à sa capacité à honorer ses engagements financiers envers les assurés. Une évaluation des provisions trop basse peut entraîner un manque de fonds, menaçant ainsi la capacité de l'assureur à payer les indemnités nécessaires.

— **Gestion des Investissements :** À l'inverse, une évaluation trop élevée des provisions immobilise des fonds qui pourraient être investis dans des actifs générateurs de rendement. Les assureurs doivent donc équilibrer cette évaluation pour ne pas compromettre leur rentabilité en conservant des réserves excessives.

— **Préparation aux Incertitudes :** Les sinistres peuvent apparaître longtemps après la souscription d'une police et peuvent varier considérablement. Une estimation précise des provisions est donc cruciale, bien que soumise à des facteurs multiples. Elle permet aux assureurs de mieux se préparer à ces incertitudes et de maintenir une marge de sécurité adéquate.

Conformité Réglementaire

Les exigences réglementaires en matière de provisionnement sont strictes et varient selon les juridictions. Ces régulations ont pour but de garantir que les assureurs disposent des fonds nécessaires pour satisfaire leurs obligations. Deux cadres réglementaires majeurs doivent être pris en compte :

— **Normes Comptables Internationales :** Les International Financial Reporting Standards (IFRS) imposent des normes rigoureuses pour le provisionnement. Ces normes exigent que les assureurs évaluent leurs provisions sur une base actuarielle solide.

— **Régulations Locales :** En France, le Code des Assurances et les régulations de l'ACPR imposent des exigences spécifiques concernant le provisionnement. Le non-respect de ces exigences peut entraîner des

sanctions ou des restrictions d'activités.

Optimisation des Fonds

La gestion efficace des fonds provisionnés est cruciale pour la santé financière d'une compagnie d'assurance. Une gestion optimisée des réserves permet non seulement de répondre aux obligations futures mais aussi de maximiser les rendements et d'assurer une allocation efficace des ressources. Trois aspects clés doivent être pris en compte :

— **Gestion des Actifs et Passifs :** Les assureurs doivent gérer leurs actifs et passifs de manière à optimiser le rendement tout en garantissant la liquidité nécessaire pour couvrir les indemnités futures.

— **Analyse et Ajustement :** Les assureurs doivent régulièrement analyser leurs réserves et ajuster leurs stratégies de provisionnement en fonction des évolutions des sinistres et des conditions économiques.

— **Impact sur les Coûts et la Rentabilité :** Une gestion efficace des provisions influence directement les coûts opérationnels et la rentabilité d'une compagnie d'assurance.

En conclusion, le provisionnement est une composante essentielle de la gestion financière des compagnies d'assurance, jouant un rôle crucial dans la stabilité financière, la conformité réglementaire et l'optimisation des fonds.

Une estimation précise des provisions permet aux assureurs de maintenir leur solvabilité, de répondre aux exigences réglementaires et de maximiser leur rentabilité, tout en se préparant aux incertitudes inhérentes à leur activité. Les défis liés au provisionnement nécessitent une approche rigoureuse et une gestion proactive pour assurer le succès et la pérennité des compagnies d'assurance.

1.2 Contexte et Importance de la Gestion des Dossiers Sinistres

Introduction Générale à l'Assurance et à la Gestion des Sinistres

L'assurance non-vie est un vaste domaine couvrant une multitude de risques, allant des accidents aux catastrophes naturelles. Son objectif premier est de protéger les assurés contre des événements imprévus et potentiellement dévastateurs. La gestion des sinistres, au cœur de cette industrie, comprend plusieurs étapes cruciales, nécessitant à la fois rigueur et efficacité pour assurer la satisfaction des assurés et la pérennité financière des compagnies d'assurance.

— **Réception des Déclarations de Sinistre :** Lorsqu'un sinistre survient, l'assuré soumet une déclaration qui marque le début du processus de réclamation. Cette étape est cruciale pour recueillir toutes les informations nécessaires à la gestion du dossier.

— **Évaluation des Dommages :** Des experts procèdent à l'évaluation des pertes, en s'appuyant sur des inspections physiques et l'analyse des documents soumis. Cette étape permet de déterminer la nature et l'étendue des dommages.

— **Détermination des Indemnités :** Suite à l'évaluation, le montant des indemnités est calculé selon les termes du contrat d'assurance. Cette phase peut parfois nécessiter des négociations avec l'assuré pour parvenir à un accord.

— **Suivi du Dossier jusqu'à sa Clôture :** Le dossier est suivi jusqu'à sa clôture, garantissant que toutes les obligations contractuelles sont remplies et que l'assuré est pleinement satisfait.

Bien que ces étapes puissent sembler linéaires, elles sont en réalité complexes et nécessitent une gestion précise et rapide pour garantir l'efficacité opérationnelle des compagnies d'assurance.

Importance de la Clôture des Dossiers Sinistres pour l'Efficacité Opérationnelle et Financière

La clôture efficace des dossiers sinistres est cruciale pour plusieurs raisons, influençant directement l'efficacité opérationnelle et la stabilité financière des compagnies d'assurance.

Premièrement, la gestion des sinistres est particulièrement gourmande en ressources, tant humaines que financières. Une clôture rapide des dossiers permet de réallouer ces ressources vers d'autres tâches, augmentant ainsi l'efficacité globale de l'entreprise.

Ensuite, les provisions pour sinistres représentent une part substantielle des réserves d'une compagnie d'assurance. Une clôture efficace des dossiers permet de libérer ces fonds, améliorant ainsi la liquidité et facilitant une gestion plus agile des flux de trésorerie.

De plus, une gestion rigoureuse des sinistres assure que les données restent à jour et fiables. Cette précision est essentielle pour la qualité des prévisions actuarielles, permettant d'éviter des erreurs de provisionnement et contribuant à préserver la rentabilité et la solvabilité de l'assureur.

Enfin, la rapidité et la transparence dans la clôture des dossiers sinistres jouent un rôle déterminant dans la satisfaction des assurés. Un traitement efficace renforce la confiance des clients et améliore la réputation de l'assureur.

En résumé, la gestion efficace des sinistres et la clôture rapide des dossiers sont essentielles non seulement pour optimiser les ressources et améliorer la liquidité, mais aussi pour garantir la précision des prévisions actuarielles et accroître la satisfaction des clients. Dans ce contexte, l'application du machine learning offre des opportunités prometteuses. Les algorithmes de machine learning peuvent, par exemple, identifier automatiquement les dossiers mûrs pour la clôture, réduisant ainsi les délais et les coûts associés à la gestion des sinistres, tout en améliorant l'efficacité opérationnelle des compagnies d'assurance.

1.3 Rôle du Machine Learning dans l'Innovation

Rôle du Machine Learning dans l'Innovation en Assurance

Historiquement, les compagnies d'assurance se sont reposées sur des méthodes statistiques classiques pour le provisionnement, telles que le **Chain-Ladder**, introduit dans les années 1960. Cette méthode repose sur l'hypothèse que les ratios de développement des sinistres, c'est-à-dire l'évolution des sinistres au fil du temps, restent constants à travers les périodes d'évaluation (Mack, 1993). Bien que le Chain-Ladder ait montré son efficacité dans un contexte de données stables, il présente des limites, notamment en ne tenant pas compte des changements économiques, des évolutions comportementales des assurés, ou des nouvelles régulations. De plus, cette méthode n'est pas conçue pour gérer des données non linéaires ou des interactions complexes entre variables.

Avec l'avènement du big data, les méthodes traditionnelles comme le Chain-Ladder révèlent leurs limites face à des ensembles de données volumineux, hétérogènes et dynamiques. Ces limitations peuvent mener à des estimations imprécises, compromettant la solvabilité de l'entreprise ou mal allouant les ressources. C'est dans ce contexte que le machine learning se distingue comme une avancée majeure. Contrairement aux méthodes classiques, le machine learning apprend directement des données, affinant ainsi la capacité de prévision.

Les algorithmes de machine learning tels que les forêts aléatoires, les régressions logistiques et les arbres de classification offrent des capacités prédictives supérieures en traitant efficacement de grandes quantités de données tout en capturant des interactions complexes. Par exemple, les forêts aléatoires, qui combinent les résultats de plusieurs arbres de décision, réduisent le risque de sur-apprentissage et améliorent la robustesse des prévisions. Les arbres de classification, en segmentant les données selon divers critères, et la régression logistique, en modélisant les probabilités d'événements binaires, renforcent également la précision des estimations.

En somme, le machine learning représente une avancée essentielle pour les compagnies d'assurance, permettant non seulement d'améliorer la précision des prévisions mais aussi de détecter des comportements anormaux, tout en

gérant efficacement des volumes de données toujours croissants.

Transformation de la Gestion des Dossiers Sinistres par le Machine Learning

Le machine learning transforme de manière significative la gestion des dossiers sinistres dans le secteur de l'assurance, apportant des innovations qui améliorent tant l'efficacité que la précision des processus.

Tout d'abord, le machine learning facilite l'automatisation de la détection des dossiers prêts à être clôturés. Les modèles de machine learning, comme ceux basés sur les forêts aléatoires, peuvent analyser des données historiques pour prédire avec une grande précision quand un dossier doit être fermé. Ces modèles identifient des schémas dans les données des sinistres, tels que le temps moyen de traitement ou la complexité des cas, pour déterminer les dossiers qui ont atteint leur stade de résolution. Cela permet de libérer des ressources humaines pour se concentrer sur des tâches plus critiques tout en réduisant les erreurs humaines et en accélérant le processus de clôture.

Ensuite, le machine learning améliore la précision de la gestion des dossiers sinistres grâce à des modèles prédictifs avancés. Les algorithmes, comme les forêts aléatoires et les réseaux de neurones, capturent des relations non linéaires et des interactions complexes entre variables non visibles avec les méthodes traditionnelles. Par exemple, une forêt aléatoire peut combiner les résultats de plusieurs arbres de décision pour affiner l'évaluation de la probabilité qu'un sinistre nécessite une intervention prolongée ou complexe. Cette capacité à modéliser des interactions complexes permet une meilleure gestion des ressources et une réduction des délais de traitement.

En matière de détection de la fraude, le machine learning joue également un rôle crucial. Les techniques de détection d'anomalies, telles que celles utilisées dans les forêts aléatoires ou les algorithmes de clustering, analysent les données des sinistres pour identifier des comportements atypiques ou des réclamations suspectes. Par exemple, les modèles peuvent signaler des dossiers présentant des caractéristiques inhabituelles, comme des réclamations répétées ou des incohérences dans les documents soumis. Cette identification précoce des anomalies permet une investigation plus approfondie et contribue à la réduction des pertes financières liées aux fraudes.

Enfin, l'intégration du machine learning optimise l'évaluation des dom-

mages et le traitement des sinistres. Par exemple, des systèmes basés sur le machine learning utilisent des techniques de traitement d'image pour analyser les photos des sinistres, telles que celles de véhicules endommagés, et estimer les coûts de réparation de manière automatique et précise. De même, les algorithmes de traitement du langage naturel peuvent extraire et analyser les informations pertinentes des rapports de sinistre, facilitant ainsi une évaluation plus rapide et précise. Cette automatisation réduit les coûts opérationnels en limitant les tâches manuelles répétitives et améliore la satisfaction des clients en accélérant le processus de règlement des sinistres.

En résumé, le machine learning révolutionne la gestion des dossiers sinistres en assurant une clôture plus rapide et précise, en détectant efficacement les fraudes, et en optimisant l'évaluation des dommages. Ces innovations contribuent à une meilleure efficacité opérationnelle et à une réduction des coûts, tout en améliorant l'expérience des clients.

1.4 L'Assurance Garantie Loyers Impayés

Les Risques Associés aux Loyers Impayés

L'assurance Garantie Loyers Impayés (GLI) représente une solution indispensable pour les propriétaires bailleurs, leur permettant de se prémunir contre les risques financiers associés au non-paiement des loyers par leurs locataires. Ce produit d'assurance, conçu pour sécuriser les revenus locatifs, prend une importance particulière dans un contexte où les incertitudes économiques et les tensions sur le marché immobilier augmentent.

Les loyers impayés constituent une menace significative pour les propriétaires. En cas de défaillance du locataire, le propriétaire se trouve dans une situation délicate où il doit non seulement faire face à une perte de revenus, mais aussi couvrir les frais supplémentaires liés à la gestion du sinistre, tels que les coûts juridiques pour engager une procédure d'expulsion. Dans les situations où le loyer constitue une part importante des revenus du propriétaire, par exemple pour ceux ayant contracté des prêts immobiliers, les impayés peuvent entraîner des difficultés financières graves, voire conduire à des situations de surendettement.

Ce phénomène des loyers impayés ne se limite pas à un problème individuel pour les propriétaires ; il a des répercussions plus larges sur le marché locatif et sur l'économie en général. Les assureurs doivent gérer un volume

croissant de sinistres, surtout en période de crise économique où les taux d'impayés tendent à augmenter. Cette situation exige une gestion rigoureuse des risques et un provisionnement adéquat pour faire face aux sinistres futurs. Les fluctuations économiques, telles que la montée du chômage ou une récession, peuvent entraîner une hausse soudaine des loyers impayés, mettant à l'épreuve la capacité des assureurs à maintenir leur viabilité financière.

Gestion et Prévention des Risques pour les Assureurs

La gestion des sinistres liés aux loyers impayés est donc un enjeu central pour les assureurs. Un provisionnement insuffisant peut conduire à des pertes financières considérables, tandis qu'un provisionnement excessif pourrait immobiliser des capitaux de manière inefficace, réduisant ainsi la rentabilité de l'assureur. Pour cette raison, il est crucial de développer des modèles prédictifs robustes qui prennent en compte une multitude de facteurs, tels que les caractéristiques socio-économiques des locataires, les dynamiques du marché locatif, et les tendances macroéconomiques. L'utilisation de techniques avancées de machine learning peut améliorer la précision des prévisions de sinistres et permettre aux assureurs d'ajuster leurs provisions de manière dynamique.

En parallèle, l'assurance GLI pose des défis en termes d'accessibilité pour les locataires. Les assureurs imposent souvent des critères d'éligibilité stricts, tels qu'un niveau minimum de revenus ou une situation professionnelle stable, afin de minimiser les risques. Si ces critères permettent effectivement de limiter les sinistres pour les assureurs, ils peuvent également restreindre l'accès au logement pour certaines catégories de locataires, notamment ceux en situation précaire ou en début de carrière. Cette sélection rigoureuse des locataires peut contribuer à accentuer les tensions sur le marché locatif, en excluant les plus vulnérables.

Les répercussions économiques et sociales des loyers impayés sont également notables. Lorsque les locataires rencontrent des difficultés à honorer leurs paiements, cela peut créer une spirale négative où les propriétaires deviennent de plus en plus sélectifs, exacerbant ainsi la pénurie de logements disponibles pour les ménages à revenus modestes. Cette situation peut renforcer les inégalités sociales et accroître la pression sur les dispositifs publics de soutien au logement, tels que les aides sociales et les subventions au logement.

Les assureurs, de leur côté, doivent continuellement adapter leurs stratégies de gestion des risques pour faire face à ces défis. L'une des approches

consiste à développer des méthodes de provisionnement plus sophistiquées, capables de refléter les risques réels de manière précise. Cela implique non seulement une meilleure compréhension des facteurs qui influencent les loyers impayés, mais aussi une capacité à réagir rapidement aux changements du marché. Par exemple, en utilisant des données en temps réel et des analyses prédictives, les assureurs peuvent ajuster leurs provisions de manière proactive, réduisant ainsi leur exposition aux risques.

Dans cette optique, les innovations technologiques jouent un rôle crucial. L'intelligence artificielle et les systèmes d'analyse de données avancés permettent aux assureurs de détecter des schémas complexes dans les données des locataires et d'anticiper les comportements à risque. Par exemple, l'analyse prédictive peut identifier les locataires susceptibles de rencontrer des difficultés de paiement bien avant que les impayés ne surviennent, permettant ainsi aux assureurs de prendre des mesures préventives, telles que la proposition de plans de paiement adaptés ou l'ajustement des garanties.

Enfin, la gestion des sinistres liés aux loyers impayés nécessite des processus rigoureux pour assurer une clôture efficace des dossiers. La rapidité et l'efficacité dans la gestion des réclamations sont essentielles pour minimiser les coûts associés et améliorer la satisfaction des clients. Dans ce cadre, l'intégration de solutions numériques, comme les plateformes de gestion des sinistres automatisées, peut considérablement réduire les délais de traitement et les erreurs humaines, tout en offrant une meilleure transparence aux parties prenantes.

En somme, l'assurance Garantie Loyers Impayés occupe une position stratégique dans le secteur de l'assurance et du marché locatif. Elle répond à une demande croissante de sécurité financière de la part des propriétaires, tout en posant des défis complexes en termes de gestion des risques, de provisionnement, et d'accessibilité au logement. Les évolutions économiques, technologiques et sociales continueront de façonner ce secteur, rendant nécessaire une adaptation continue des pratiques actuarielles et des politiques d'assurance pour répondre aux besoins du marché tout en garantissant une protection efficace contre les risques de loyers impayés.

1.5 Énoncé de la Problématique Actuarielle

Pour résumer, dans le domaine de l'assurance, la prédiction de la clôture des dossiers de sinistres est un processus complexe et crucial. Ce proces-

sus joue un rôle déterminant dans l'estimation des provisions nécessaires, influençant directement la solvabilité et la stabilité financière des compagnies d'assurance. Toutefois, les méthodes traditionnelles utilisées pour cette prédiction, bien qu'elles aient servi le secteur pendant des décennies, commencent à montrer leurs limites face à l'évolution rapide des données et des exigences actuelles.

Un des principaux défis est l'hétérogénéité des données. Chaque dossier de sinistre est unique et peut varier considérablement en fonction de multiples facteurs tels que le type de risque, les caractéristiques des assurés, les conditions économiques, et les régulations locales. Cette diversité complique la création de modèles prédictifs fiables. Les méthodes traditionnelles, qui ont tendance à agréger les données pour simplifier l'analyse, risquent de perdre des nuances importantes. Par exemple, un sinistre automobile survenu dans une zone urbaine densément peuplée peut avoir une dynamique très différente d'un sinistre similaire en zone rurale, avec des impacts distincts sur le temps de clôture et les coûts associés. Cette hétérogénéité des sinistres rend les modèles classiques moins efficaces pour capturer la complexité nécessaire à une prévision précise et fiable.

De plus, les modèles de sinistres ne sont pas statiques ; ils évoluent continuellement sous l'influence de divers facteurs. Parmi ceux-ci, l'évolution technologique, les changements dans les comportements des assurés, et les modifications réglementaires jouent un rôle central. Par exemple, l'émergence des véhicules autonomes a le potentiel de transformer radicalement le paysage de l'assurance automobile, modifiant les types de sinistres, leur fréquence et leur gravité. Les méthodes traditionnelles, souvent fondées sur l'analyse de données historiques, peuvent être trop rigides pour s'adapter efficacement à ces évolutions dynamiques. Par conséquent, elles risquent de produire des prévisions obsolètes ou inexactes, compromettant ainsi la précision du provisionnement et, par extension, la stabilité financière des compagnies d'assurance.

Un autre défi réside dans le rythme accéléré des affaires modernes, qui exige des décisions rapides, souvent en temps réel. Les méthodes traditionnelles de provisionnement reposent sur des cycles de reporting périodiques et des processus manuels, rendant difficile une réactivité adéquate aux événements imprévus. Par exemple, lors d'une catastrophe naturelle, les assureurs doivent être capables d'évaluer rapidement les pertes potentielles et de provisionner les ressources nécessaires. L'incapacité à effectuer des analyses en temps réel peut non seulement ralentir les processus de règlement des si-

nistres, mais aussi nuire à la satisfaction des clients et augmenter les risques financiers pour l'assureur.

Face à ces défis, la question centrale de cette recherche se pose ainsi : **Comment le machine learning peut-il être appliqué efficacement pour prédire la clôture des dossiers en assurance afin d'optimiser le provisionnement ?**

Cette question vise à explorer l'application des techniques avancées de machine learning pour améliorer la précision et l'efficacité des prévisions de clôture des dossiers, un aspect critique pour le provisionnement en assurance. Pour répondre à cette question, plusieurs objectifs spécifiques doivent être atteints, chacun contribuant à une meilleure compréhension et une meilleure application du machine learning dans ce contexte.

Chapitre 2

Présentation des Modèles de Machine Learning

L'objectif central de cette recherche est de développer un modèle prédictif capable de prévoir avec précision la clôture des dossiers de sinistres. Ce modèle devra être suffisamment robuste pour s'adapter aux variations des données et aux évolutions des modèles de sinistres. Le développement de ce modèle nécessitera une approche itérative, combinant la sélection de variables pertinentes, l'entraînement d'algorithmes de machine learning, ainsi que l'optimisation des hyperparamètres pour maximiser la performance.

2.1 Méthodologie

Identification et Intégration de Données Variées

L'un des premiers objectifs de cette recherche est de collecter et d'intégrer des sources de données variées. Les données traditionnelles, telles que les historiques de sinistres et les caractéristiques des contrats, seront complétées par des indicateurs économiques, des variables liées à l'évaluation des provisions, et d'autres données contextuelles. L'intégration de ces différentes sources de données permettra de capturer la complexité et la diversité des sinistres de manière plus précise, offrant ainsi une base solide pour l'entraînement des modèles de machine learning.

Enrichir les modèles prédictifs avec une telle variété de données vise à surmonter les limitations des méthodes traditionnelles, qui tendent à simplifier excessivement la réalité des sinistres. Cette approche permettra de mieux comprendre les facteurs influençant la durée de vie d'un dossier de sinistre,

d'identifier les motifs sous-jacents qui pourraient passer inaperçus dans des modèles plus simples, et ainsi d'améliorer la précision des prévisions de clôture.

Développement de Modèles Adaptatifs

La prochaine étape de notre travail de recherche consiste à développer des modèles de machine learning capables de s'adapter aux évolutions des modèles de sinistres et aux nouvelles tendances du marché. Les algorithmes de machine learning offrent une flexibilité qui leur permet de s'ajuster en continu aux nouvelles données et aux changements de comportement. Par exemple, un modèle de machine learning pourrait être conçu pour détecter et s'ajuster automatiquement aux variations dans la fréquence et la gravité des sinistres, ou aux changements dans les profils des assurés.

Le développement de ces modèles adaptatifs nécessite également la capacité de ces algorithmes à reconnaître et intégrer rapidement de nouvelles variables qui peuvent émerger à mesure que le contexte des sinistres évolue. Cette capacité d'adaptation repose sur une architecture de modélisation non seulement robuste, mais aussi apte à gérer des relations complexes et hétérogènes entre les différentes variables. les modèles concernés sont :la régression logistique Elastic Net, les arbres de décision, et les forêts aléatoires

Évaluation de la Performance

Enfin, un objectif essentiel de cette recherche est d'évaluer la performance des modèles développés. Cette évaluation se fera en comparant la précision des prévisions obtenues grâce aux algorithmes de machine learning avec celles des méthodes traditionnelles. Des techniques rigoureuses de validation croisée seront employées, accompagnées de l'utilisation d'ensembles de test indépendants, pour garantir que les modèles sont fiables et capables de généraliser correctement à de nouvelles données.

Des métriques spécifiques, telles que la précision, le rappel, et l'AUC-ROC (Area Under the Receiver Operating Characteristic Curve), seront utilisées pour quantifier la performance des modèles. Ces métriques permettront d'évaluer la capacité des modèles à prédire efficacement la clôture des sinistres et de démontrer les avantages du machine learning en termes de précision, de robustesse, et d'efficacité dans le cadre du provisionnement en assurance.

En résumé, la méthodologie adoptée dans cette recherche met l'accent sur une approche intégrée et adaptative pour le développement de modèles prédictifs, en tirant parti des avancées du machine learning pour améliorer la gestion des sinistres dans le secteur de l'assurance.

2.2 Biais, Variance, et Complexité du Modèle

Dans l'évaluation de la performance d'un modèle de machine learning, les concepts de biais et de variance jouent un rôle crucial. Le biais fait référence à l'erreur systématique d'un modèle, soit l'écart moyen entre les prédictions du modèle et la valeur réelle cible. Lorsqu'un modèle présente un biais élevé, cela indique qu'il est sous-ajusté (*underfitting*), c'est-à-dire qu'il est trop simple pour capturer la complexité inhérente des données d'entraînement. Par exemple, un modèle linéaire appliqué à des données non linéaires est typiquement sujet à un biais élevé. Ce manque de flexibilité limite la capacité du modèle à saisir les nuances des données, entraînant des prédictions erronées et une mauvaise performance globale.

En parallèle, la variance mesure la sensibilité du modèle aux variations des données d'entraînement. Un modèle avec une variance élevée est très flexible, ce qui lui permet de s'ajuster de manière détaillée aux données d'entraînement, y compris aux variations dues au bruit aléatoire. Ce phénomène est connu sous le nom de sur-ajustement (*overfitting*). Un modèle sur-ajusté capte non seulement les tendances générales des données, mais aussi les fluctuations spécifiques à l'échantillon d'entraînement, ce qui nuit à sa capacité de généraliser à de nouvelles données. L'existence de ce compromis entre biais et variance, souvent appelé le dilemme biais-variance, est au cœur de la construction de modèles robustes. À mesure que la complexité du modèle augmente, le biais tend à diminuer, mais la variance augmente. L'objectif est donc de trouver un équilibre optimal qui minimise l'erreur de généralisation.

2.3 Régression Logistique Pénalisée Elastic Net

La régression logistique est utilisée pour les problèmes de classification binaire, où l'objectif est de prédire la probabilité d'un événement. Dans notre contexte, elle permet de prédire si un dossier sinistre doit être clôturé. La probabilité $P(y = 1 \mid \mathbf{x})$ est calculée à partir des variables explicatives, telles que le type de sinistre, le montant en jeu, et le délai de traitement, via la

fonction logistique suivante :

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp\left(-\left(\mathbf{w}^\top \mathbf{x} + b\right)\right)}$$

où $\mathbf{x}$ est le vecteur des caractéristiques, $\mathbf{w}$ est le vecteur des coefficients à estimer, et b est le biais.

Régularisation Elastic Net

Pour éviter le sur-apprentissage (*overfitting*) dans des contextes avec de nombreuses variables, la régression logistique Elastic Net intègre une régularisation mixte des pénalisations L_1 (Lasso) et L_2 (Ridge). La fonction de coût régularisée pour la régression logistique Elastic Net est :

$$\begin{aligned} J(\mathbf{w}) = -\frac{1}{m} \sum_{i=1}^{m} \big[& y^{(i)} \log\left(P(y^{(i)} = 1 \mid \mathbf{x}^{(i)})\right) \\ & + (1 - y^{(i)}) \log\left(1 - P(y^{(i)} = 1 \mid \mathbf{x}^{(i)})\right) \big] \\ & + \lambda_1 \|\mathbf{w}\|_1 + \frac{\lambda_2}{2} \|\mathbf{w}\|_2^2 \end{aligned} \tag{2.1}$$

où :
— m est le nombre d'observations,
— λ_1 est le paramètre de régularisation pour la pénalisation L_1,
— λ_2 est le paramètre de régularisation pour la pénalisation L_2,
— $\|\mathbf{w}\|_1$ est la norme L_1 des coefficients,
— $\|\mathbf{w}\|_2^2$ est la norme L_2 au carré des coefficients.

L'Elastic Net permet à la fois la sélection des variables et la gestion de la multicolinéarité en combinant les avantages du Lasso et du Ridge. Il possède deux hyperparamètres principaux à ajuster : λ_1 et λ_2, contrôlant respectivement les contributions des régularisations L_1 et L_2. Un autre paramètre, α, ajuste la combinaison entre λ_1 et λ_2 :

$$\alpha \in [0, 1] \quad \text{avec} \quad \lambda_1 = \alpha\lambda, \quad \lambda_2 = (1 - \alpha)\lambda$$

où λ est un paramètre global de régularisation. Lorsque $\alpha = 1$, il correspond à une régularisation Lasso pure, et lorsque $\alpha = 0$, il correspond à une régularisation Ridge pure.

Avantages de l'Elastic Net

L'Elastic Net est particulièrement utile pour :

— **Sélection de Variables** : Élimine complètement certaines variables inutiles grâce à la composante Lasso.
— **Gestion de la Multicolinéarité** : La composante Ridge aide à gérer la multicolinéarité entre les variables explicatives.
— **Flexibilité** : Permet de choisir entre une pénalisation Lasso pure, une pénalisation Ridge pure, ou une combinaison des deux en ajustant α.

Applications Pratiques

Dans le contexte de la détection automatique des dossiers sinistres nécessitant une clôture, la régression logistique Elastic Net est particulièrement adaptée. Les packages `glmnet`, `penalized`, et `glmpath` sont recommandés pour implémenter ces modèles, offrant flexibilité et performance. La régression logistique Elastic Net se révèle ainsi être un modèle robuste, alliant complexité et interprétabilité tout en limitant le sur-apprentissage et en gérant la multicolinéarité. Sa transparence dans l'identification des facteurs déterminants en fait un choix pertinent pour des applications opérationnelles où la compréhension des décisions est essentielle.

2.4 Arbres de Décision

Les arbres de décision sont des modèles non linéaires puissants qui divisent l'espace des données en régions homogènes selon les caractéristiques explicatives. Ils sont capables de modéliser des interactions complexes entre les variables, ce qui en fait un outil flexible et puissant. Cependant, ils sont sensibles aux variations des données d'entraînement et peuvent facilement s'ajuster au bruit aléatoire, ce qui peut entraîner une variance élevée et un risque de sur-ajustement (*overfitting*).

Techniques pour Limiter le Sur-Ajustement

Pour limiter le sur-ajustement dans les arbres de décision, plusieurs techniques peuvent être appliquées :
— **Élagage (Pruning)** : Suppression de certaines branches après la construction de l'arbre pour réduire la complexité du modèle. Cela peut être réalisé soit par *pré-pruning* (arrêt de la croissance de l'arbre avant qu'il ne se développe complètement), soit par *post-pruning* (réduction d'un arbre complet).
— **Contrôle de la Profondeur** : Limitation de la profondeur maximale de l'arbre pour réduire le nombre de divisions et la variance. Une

profondeur maximale est fixée pour éviter une complexité excessive.

Arbre de Classification

L'arbre de classification, ou méthode CART (Classification and Regression Tree), est largement utilisé pour les problèmes de classification. Il se distingue par sa simplicité et son interprétabilité, offrant des règles décisionnelles explicites et gérant efficacement des données hétérogènes, manquantes, ainsi que des effets non linéaires entre les variables.

L'arbre de classification repose sur une série de règles décisionnelles simples organisées en une structure hiérarchique. À chaque nœud, une condition est évaluée sur une variable explicative pour diviser les données en sous-ensembles de plus en plus homogènes. Chaque feuille de l'arbre correspond à une classe prédite.

L'arbre de classification offre plusieurs avantages :

- **Interprétabilité** : Offre une visualisation claire des règles décisionnelles, facilitant l'explication et la justification des décisions.
- **Adaptabilité** : Gère bien les données hétérogènes sans nécessiter de transformations complexes.
- **Flexibilité** : Adapté à la gestion des données déséquilibrées et à la détection des cas minoritaires critiques.

La construction d'un arbre de classification commence par la segmentation de la population en sous-ensembles homogènes. Cette segmentation se fait en cherchant la variable et le seuil qui optimisent la séparation des données. L'arbre est découpé jusqu'à ce qu'une condition d'arrêt soit rencontrée, comme un nombre minimal d'observations par feuille ou un seuil de pureté des nœuds.

Critères de Division

Le choix du critère de division est crucial pour la performance du modèle. Les critères principaux sont l'indice de Gini, l'entropie, et l'impureté. Voici leur définition et utilisation :

Indice de Gini L'indice de Gini mesure la probabilité qu'un élément choisi au hasard soit mal classé s'il était étiqueté selon la distribution des classes dans un ensemble de données. Il est défini par :

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2 \tag{2.2}$$

où p_i est la proportion d'éléments appartenant à la classe i dans l'ensemble D, et C est le nombre total de classes. Un faible indice de Gini indique une forte homogénéité des classes dans le nœud.

Entropie (Gain d'Information) L'entropie mesure l'incertitude ou l'imprévisibilité d'une distribution. Elle est définie par :

$$Entropy(D) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{2.3}$$

Le gain d'information, qui mesure la réduction d'entropie après une division, est défini par :

$$Gain(D, A) = Entropy(D) - \sum_{v \in V(A)} \frac{|D_v|}{|D|} Entropy(D_v) \tag{2.4}$$

où A est l'attribut utilisé pour diviser l'ensemble de données D, $V(A)$ est l'ensemble des valeurs possibles de A, et D_v est le sous-ensemble des données où A prend la valeur v. Un gain d'information élevé indique une réduction significative de l'incertitude.

Impureté (Misclassification Error) L'impureté mesure le taux d'erreur de classification dans un nœud. Elle est définie par :

$$Impurity(D) = 1 - \max(p_1, p_2, \ldots, p_C) \tag{2.5}$$

Dans cette étude, nous privilégierons l'indice de Gini pour sa simplicité et sa rapidité de calcul.

2.5 Forêt Aléatoire

La forêt aléatoire est une méthode d'ensemble puissante qui combine plusieurs arbres de décision pour améliorer la performance du modèle tout en réduisant la variance. Chaque arbre de la forêt est construit à partir d'un sous-échantillon aléatoire des données d'entraînement, avec remplacement (*bootstrap*), et les prédictions sont agrégées par vote majoritaire pour les tâches de classification ou par moyenne pour les problèmes de régression.

Cette approche permet de surmonter les principales faiblesses des arbres de décision individuels, telles que le surapprentissage (*overfitting*) et la variance élevée.

Fonctionnement Mathématique

L'erreur quadratique moyenne (MSE) d'une forêt aléatoire peut être exprimée comme suit :

$$\text{MSE} = (\text{Biais})^2 + \frac{\sigma^2}{B} + \frac{(1-\rho)\sigma^2}{B}$$

où σ^2 est la variance des prédictions individuelles des arbres, ρ est la corrélation moyenne entre les prédictions des arbres, et B est le nombre d'arbres dans la forêt. L'augmentation du nombre d'arbres B tend à diminuer la variance, car les erreurs non corrélées entre les arbres s'annulent. Cependant, la profondeur des arbres individuels joue également un rôle crucial ; des arbres trop profonds peuvent augmenter la corrélation ρ entre eux, limitant ainsi l'effet de réduction de la variance.

Construction et Aléatorisation des Arbres

La forêt aléatoire se distingue par sa double randomisation :

— **Échantillonnage Bootstrap** : Chaque arbre est construit à partir d'un sous-échantillon aléatoire des données d'entraînement avec remplacement.

— **Sélection Aléatoire des Caractéristiques** : À chaque nœud de l'arbre, un sous-ensemble aléatoire des caractéristiques est sélectionné pour déterminer la meilleure division.

Ces étapes réduisent la corrélation entre les arbres, augmentant ainsi la diversité et améliorant la capacité de généralisation du modèle global. La prédiction finale de la forêt aléatoire est obtenue par :

$$\hat{y} = \text{majority\_vote}\left\{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_B(\mathbf{x})\right\}$$

où $h_i(\mathbf{x})$ représente la prédiction de l'arbre i pour l'instance $\mathbf{x}$. Cette agrégation par vote majoritaire réduit l'impact des erreurs des arbres individuels.

Importance des Variables

La forêt aléatoire peut estimer l'importance des variables en utilisant la réduction moyenne de l'impureté (*Mean Decrease in Impurity, MDI*) obtenue

lors des divisions impliquant chaque caractéristique à travers tous les arbres. Cette fonctionnalité est particulièrement utile pour identifier les variables les plus influentes dans la prise de décision, notamment dans le contexte des dossiers sinistres nécessitant clôture.

Optimisation du Modèle

La forêt aléatoire permet de gérer les données déséquilibrées et de maintenir un bon équilibre entre les classes, ce qui est essentiel dans les cas où certaines classes sont sous-représentées. Pour optimiser le compromis entre biais et variance, il est important d'ajuster les hyperparamètres, notamment le nombre de prédicteurs sélectionnés à chaque scission q. En général, pour les problèmes de régression, q est souvent fixé à la partie entière de $\frac{p}{3}$, tandis que pour les problèmes de classification, il est recommandé de prendre q égal à la racine carrée de p, où p est le nombre total de prédicteurs.

Pour implémenter la forêt aléatoire, nous utiliserons le package `randomForest` dans notre analyse.

La forêt aléatoire surmonte les limitations des arbres de décision individuels en agrégeant un grand nombre d'arbres, ce qui améliore la robustesse du modèle tout en conservant ses avantages. Elle fournit un pouvoir discriminant élevé et une bonne généralisation aux nouvelles données, tout en restant relativement simple à mettre en œuvre. La capacité de la forêt aléatoire à réduire la variance et à estimer l'importance des variables en fait un outil précieux pour la détection et l'analyse des dossiers sinistres nécessitant clôture.

Comparatif des Performances

Le tableau ci-dessous présente un résumé des caractéristiques clés des modèles : régression logistique Elastic Net, arbre de décision, et forêt aléatoire

Critères	Régression Logistique Elastic Net	Arbre de Décision	Forêt Aléatoire
Performance Générale	Bonne sur données avec multicolinéarité	Performances variables, selon la profondeur de l'arbre	Excellente, robuste aux différentes structures de données
Biais et Variance	Biais modéré, variance contrôlée via α et λ	Biais faible, variance élevée	Biais modéré, variance réduite par agrégation
Interprétabilité	Haute, coefficients faciles à interpréter, mais avec une sélection variable	Moyenne, intuitive mais complexe si profond	Faible, difficile d'interpréter l'ensemble d'arbres
Complexité Computationnelle	Faible, rapide même sur grands jeux de données	Faible à modérée, dépend de la profondeur	Élevée, plus de temps et de ressources nécessaires
Gestion des Données Manquantes	Nécessite une imputation préalable	Peut gérer certaines données manquantes intrinsèquement	Gère efficacement les données manquantes en moyenneant sur plusieurs arbres
Sensibilité aux Outliers	Moins sensible que la régression Lasso, mais sensible comparé aux arbres	Sensible, affecte les divisions de l'arbre	Moins sensible, les effets sont dilués sur plusieurs arbres
Capacité à Capturer les Relations Non Linéaires	Limitée, nécessite la transformation des variables	Bonne, capture naturellement les relations complexes	Excellente, combine plusieurs arbres pour modéliser des relations complexes
Prévention de l'Overfitting	Contrôlée via la régularisation Elastic Net (α et λ)	Risque élevé d'overfitting sur données d'entraînement	Faible risque grâce à l'agrégation et la randomisation
Traitement des Variables Catégoriques	Nécessite un encodage préalable	Gère naturellement les variables catégoriques	Gère efficacement les variables catégoriques
Utilisation dans les Problèmes Multi-classes	Bien adaptée avec extensions appropriées	Supporte naturellement les classifications multi-classes	Très performante dans contextes multi-classes

TABLE 2.1 – Comparaison des Modèles de Classification

2.6 Métriques d'Évaluation des Modèles

Introduction aux Métriques d'Évaluation

Pour évaluer la performance des modèles dans la détection des dossiers sinistres nécessitant clôture, plusieurs métriques sont essentielles. Elles offrent un aperçu de la qualité des prédictions des modèles. Les principales métriques utilisées sont la spécificité, le rappel, la précision, la balanced accuracy, et l'AUC. Nous allons définir chacune de ces métriques, en considérant que les "positifs" sont des dossiers ouverts et les "négatifs" sont des dossiers clôturés.

Spécificité (True Negative Rate)

La spécificité, ou True Negative Rate (TNR), mesure la proportion de dossiers clôturés correctement identifiés comme ne nécessitant pas de clôture parmi tous les dossiers effectivement clôturés. Elle est définie par la formule :

$$\text{Spécificité} = \frac{\text{Vrais Négatifs}}{\text{Vrais Négatifs} + \text{Faux Positifs}} \tag{2.6}$$

Un vrai négatif est un dossier clôturé correctement qui ne nécessitait pas de clôture, tandis qu'un faux positif est un dossier ouvert identifié à tort comme nécessitant une clôture.

Rappel (Sensibilité ou True Positive Rate)

Le rappel, ou True Positive Rate (TPR), mesure la proportion de dossiers ouverts nécessitant une clôture qui ont été correctement identifiés parmi tous les dossiers nécessitant effectivement une clôture. Il est défini par :

$$\text{Rappel} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Négatifs}} \tag{2.7}$$

Un vrai positif est un dossier ouvert nécessitant une clôture qui a été correctement identifié, tandis qu'un faux négatif est un dossier ouvert nécessitant une clôture qui n'a pas été détecté.

Balanced Accuracy

La balanced accuracy combine la spécificité et le rappel pour fournir une vue équilibrée de la performance du modèle, surtout en présence de classes déséquilibrées. Elle est définie par :

$$\text{Balanced Accuracy} = \frac{\text{Spécificité} + \text{Rappel}}{2} \tag{2.8}$$

Cette métrique permet d'évaluer la capacité du modèle à détecter à la fois les dossiers nécessitant une clôture et ceux ne nécessitant pas de clôture.

AUC (Area Under the Curve)

L'AUC (Area Under the ROC Curve) mesure la capacité du modèle à discriminer entre les dossiers nécessitant une clôture et ceux ne nécessitant pas de clôture. Elle est définie par l'intégrale de la courbe ROC :

$$\text{AUC} = \int_0^1 \text{ROC}(t)\, dt \tag{2.9}$$

où ROC(t) représente la courbe des caractéristiques de fonctionnement du récepteur. Une AUC proche de 1 indique une excellente capacité du modèle à séparer correctement les dossiers ouverts nécessitant une clôture des dossiers clôturés.

Précision

La précision mesure la proportion de dossiers identifiés comme nécessitant une clôture qui nécessitent effectivement une clôture. Elle est définie par :

$$\text{Précision} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Positifs}} \tag{2.10}$$

Un vrai positif est un dossier ouvert nécessitant une clôture correctement identifié, tandis qu'un faux positif est un dossier ouvert incorrectement classé comme nécessitant une clôture. La précision est importante pour évaluer la pertinence des dossiers marqués comme nécessitant clôture et éviter les erreurs coûteuses de faux positifs.

Chapitre 3

Analyse des Données

3.1 Périmètre d'Étude

Présentation des Types de Données

Pour cette étude, il est fondamental de bien comprendre la nature et l'importance des données disponibles afin de construire un modèle efficace pour la clôture automatique des dossiers. Les données ont été extraites via des requêtes SQL depuis la base de données interne de l'entreprise, englobant un total de 1 989 190 enregistrements relatifs à des sinistres. Ces enregistrements comprennent diverses variables telles que les dates de survenance et de clôture, les montants des indemnités, le type de garantie couverte, ainsi que les modalités de gestion des sinistres, que nous détaillerons plus tard.

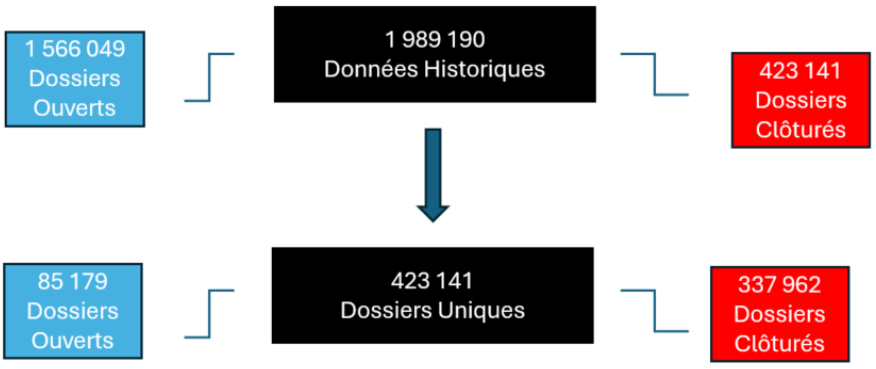

FIGURE 3.1 – Diagramme de positions des dossiers

Les sinistres sont classifiés en deux catégories : **dossiers clôturés** et **dossiers ouverts**. Nous désignerons cette variable de "situation du dossier" comme *position*, qui constituera également notre variable cible. Cette distinction est cruciale pour notre étude, dont l'objectif est de prédire quels sinistres ouverts devraient être clôturés.

Pour limiter le volume de données à traiter et garantir la pertinence des analyses, l'étude se concentre uniquement sur les dossiers dont l'année de survenance est postérieure ou égale à 2015. Les données couvrent donc la période de 2015 à 2024, avec une mise à jour annuelle des dossiers. Par exemple, un dossier ouvert en 2018 apparaîtra chaque année dans le dataset jusqu'à sa clôture en 2024, ce qui explique les multiples occurrences des mêmes sinistres. Les dossiers sont identifiés par la variable *NUMERO_SINISTRE*.

NUMERO_SINISTRE	**Année**	**Position**
1758000	2015	Ouvert
1758000	2016	Ouvert
1758000	2017	Ouvert
1758000	2018	Ouvert
1758000	2019	Ouvert
1758000	2020	Ouvert
1758000	2021	Ouvert
1758000	2022	Ouvert
1758000	2023	Clôt
1758000	2024	Clôt

L'exemple ci-dessus illustre un dossier ouvert en 2015 et finalement clôturé en 2023. Il est pertinent de s'intéresser aux caractéristiques des variables lorsqu'un dossier est en statut Clôt ou Ouvert.

Affichons un aperçu de la répartition de nos données au fil des années.

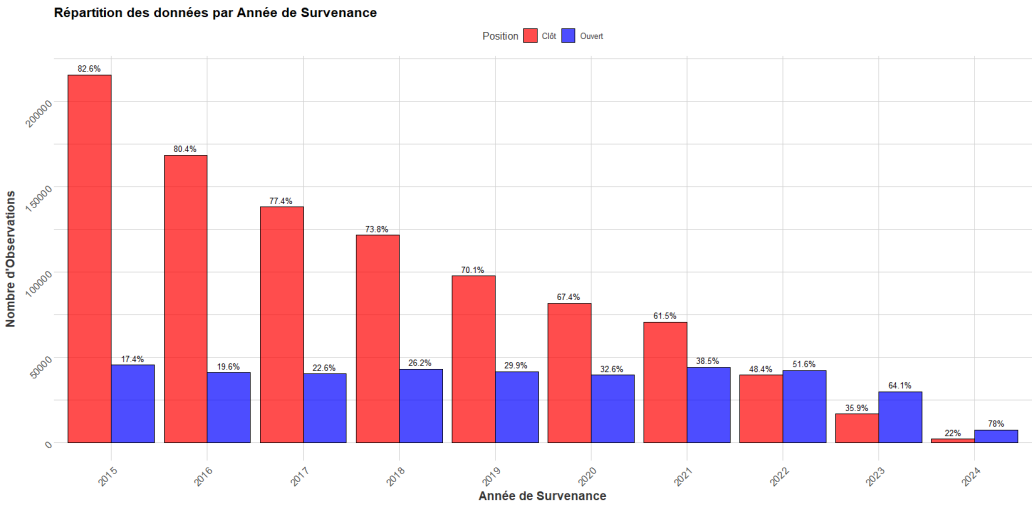

FIGURE 3.2 – Répartition des dossiers selon l'année de survenance du sinistre

Comme le montre l'histogramme,'une proportion significative des sinistres survenus antérieurement sont actuellement clôturés. Par exemple, 82,6% des dossiers ouverts en 2015 sont clôturés, contre 80,4% de ceux survenus en 2016. Cependant, cette tendance s'inverse pour les dossiers plus récents : 64,1% des dossiers ouverts en 2023 sont encore en cours, et 78% des sinistres survenus cette même année sont toujours ouverts. Ces données historiques enrichiront l'apprentissage de nos modèles de prédiction.

En effet, le nombre de dossiers dont l'année de survenance est récente est plus élevé pour les dossiers ouverts.

Les données utilisées dans cette étude proviennent directement des bases de données internes de l'entreprise, garantissant leur fiabilité. Les modèles de machine learning doivent être capables de gérer ces données en termes de scalabilité et d'interprétabilité pour fournir des résultats compréhensibles.

3.2 Description des Données Disponibles

Intéressons-nous maintenant aux variables disponibles pour chaque dossier, résumées dans le tableau suivant :

Variable	Définition	Type	Nombre de Modalités
NUMERO_SINISTRE	Identifiant unique du sinistre	character	282094
CD_Operateur_Modif	Code de l'opérateur ayant modifié les données	character	134
Date_Debut	Date de dernière modification du dossier	Date	NA
Date_Fin	Date de fin du sinistre	Date	NA
position	Statut du dossier : Ouvert ou Clôt	character	2
Date_Ouverture	Date d'ouverture du dossier	POSIXct	NA
Date_Survenance	Date de survenance du sinistre	POSIXct	NA
Date_Cloture	Date de clôture du dossier	Date	NA
Date_Reouverture	Date de réouverture du dossier	Date	NA
CD_Motif_Modif	Code du motif de modification	numeric	18
LB_Motif_Modif	Libellé du motif de modification	character	33
annee_survenance	Année de survenance du sinistre	numeric	NA
Solde_Reg	Solde des règlements associés au sinistre	character	NA
Eval_Reg	Évaluation des règlements à venir pour le sinistre	character	174536
Solde_Rec	Solde des recouvrements associés au sinistre	character	NA
Eval_Rec	Évaluation des recouvrements à venir pour le sinistre	character	37138
Solde_Eval_nette	Solde net de l'évaluation après déduction des règlements et recouvrements	character	180280
provision	Montant de la provision associée au sinistre	character	NA
TYPEMC_Sinistre	Type de sinistre : Matériel ou Corporel	character	2
LB_Produit	Libellé du produit d'assurance	character	7
garantie	Garantie associée au sinistre	character	26
Courtier_Delegataire	Présence du courtier ou délégataire	character	2
Suivi_Judiciaire	Date de suivi judiciaire associé au sinistre, le cas échéant	POSIXct	NA
ID_Mois	Identifiant numérique du mois	numeric	NA
DT_Premier_Jour_Mois	Date du premier jour du mois	Date	NA
DT_Dernier_Jour_Mois	Date du dernier jour du mois	Date	NA
ID_Jour_Debut	Identifiant numérique du jour de début du sinistre	numeric	NA
ID_Jour_Fin	Identifiant numérique du jour de fin du sinistre	numeric	NA
ID_Annee	Identifiant numérique de l'année	numeric	NA
ID_Mois_Annee	Identifiant numérique du mois et de l'année	numeric	NA

TABLE 3.2 – Définitions des variables du dataset, leurs classes et leurs nombres de modalités

Sélection et Transformation des Données

Dans un premier temps, certaines variables ont été jugées inutiles ou redondantes et seront supprimées. Par exemple, les variables temporelles détaillées telles que *ID_Jour_Debut*, *ID_Jour_Fin* sont redondantes avec les dates correspondantes, ce qui les rend superflues. De même, *NUMERO_SINISTRE* ne sera pas utilisé dans le modèle de prédiction mais restera conservé pour permettre l'identification des dossiers.

Il est aussi pertinent de traiter les variables catégorielles présentant un grand nombre de modalités. À titre d'exemple, *CD_Operateur_Modif* sera regroupé ou binarisé afin de faciliter l'exploitation par les modèles. Certaines variables, comme les différentes évaluations des règlements et recouvrements, ainsi que l'année de survenance des sinistres, sont maintenues pour leur potentielle pertinence. Cependant, des analyses plus approfondies seront nécessaires pour en déterminer l'impact exact sur le modèle. D'autres variables, telles que *ID_Mois*, *DT_Premier_Jour_Mois*, et *DT_Dernier_Jour_Mois*, seront supprimées en raison de leur redondance ou de leur manque de valeur informative.

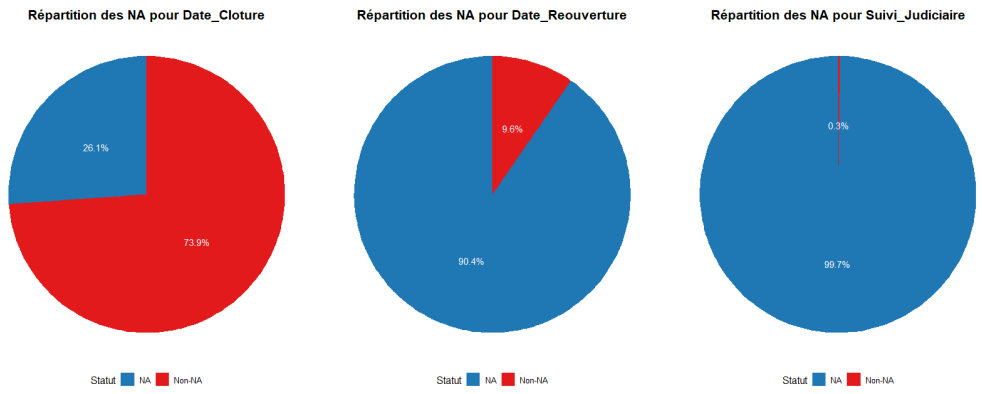

FIGURE 3.3 – Répartition des valeurs manquantes pour les variables : *Date_Cloture*, *Date_Reouverture*, et *Suivi_Judiciaire*

L'analyse des valeurs manquantes révèle que trois variables présentent des absences explicables. Les valeurs manquantes pour *Date_Cloture* sont associées aux dossiers encore ouverts. Les 90,4% de valeurs manquantes pour *Date_Reouverture* correspondent aux dossiers qui n'ont jamais été réouverts,

tandis que 99,7% des absences pour *Suivi_ Judiciaire* concernent des dossiers sans suivi judiciaire. Ainsi, il n'y a pas de valeurs manquantes inexpliquées dans nos données.
Les variables retenues sont listées dans le tableau 3.4, avec leur état de transformation prévu. Ces variables sont classées en quatre catégories de traitement :

Variable	**État**
NUMERO_SINISTRE	À conserver pour l'identification des dossiers
CD_Operateur_Modif	Réduire le nombre de modalités
Date_Debut	À transformer
Date_Fin	À supprimer
position	À conserver
Date_Ouverture	À transformer
Date_Survenance	À transformer
Date_Cloture	À supprimer
Date_Reouverture	À transformer
CD_Motif_Modif	À transformer
LB_Motif_Modif	À transformer
annee_survenance	À conserver
Solde_Reg	À conserver
Eval_Reg	À conserver
Solde_Rec	À conserver
Eval_Rec	À conserver
Solde_Eval_nette	À conserver
provision	À conserver
TYPEMC_Sinistre	À conserver
LB_Produit	À conserver
garantie	Réduire le nombre de modalités
Courtier_Delegataire	À conserver
Suivi_Judiciaire	À transformer
ID_Mois	À supprimer
DT_Premier_Jour_Mois	À supprimer
DT_Dernier_Jour_Mois	À supprimer
ID_Jour_Debut	À supprimer
ID_Jour_Fin	À supprimer
ID_Annee	À supprimer
ID_Mois_Annee	À supprimer

TABLE 3.3 – État des variables du dataset

Pertinence des Données

La sélection des variables retenues repose sur leur capacité à contribuer efficacement à la construction d'un modèle prédictif robuste. Les variables clés, telles que *position*, *annee_ survenance*, *Solde_ Reg*, et *provision*, sont directement liées à la gestion des sinistres, ce qui les rend essentielles pour l'objectif du modèle. Les variables jugées moins pertinentes seront modifiées ou exclues afin d'optimiser les performances du modèle.

Pour les variables à transformer, les étapes suivantes seront appliquées :
- *Date_ Reouverture* sera transformée en une variable booléenne *Reouverture*, prenant la valeur TRUE si le dossier a été réouvert, FALSE sinon.
- *Suivi_ Judiciaire* sera également transformée en booléen : TRUE si le dossier est en suivi judiciaire, FALSE sinon.

Ces transformations nous permettront de tirer parti des valeurs manquantes. Par exemple, *Date_ Reouverture* sera renommée *Reouverture*.

L'objectif est de focaliser l'analyse sur les aspects évolutifs des dossiers plutôt que sur les dates précises. Par exemple, nous envisagerons de comparer l'évolution des dossiers d'une année à l'autre, en sélectionnant entre *Date_ Survenance* et *annee_ survenance* afin d'éviter les redondances. La variable *Date_ Debut* sera transformée en *delai_ Modif_ ouverture*, qui représentera le délai entre la dernière modification et la date actuelle.

De nouvelles variables seront également créées pour affiner l'analyse :

Ces transformations et sélections de données contribueront à développer un modèle prédictif efficace pour la clôture automatique des dossiers de sinistres. Par ailleurs, une attention particulière sera portée aux variables avec un grand nombre de modalités, comme *LB_ Motif_ Modif* et *CD_ Motif_ Modif* . Après analyse, il apparaît que ces variables sont parfaitement corrélées, le code étant une abréviation du motif de modification. De plus après analyse approfondie nous constatons que les variables *CD_ Operateur_ Modif* et *CD_ Motif_ Modif* ne sont pas entièrement fiables en raison de la codification arbitraire. Nous nous passerons donc de ces trois variables.

Nous allons également créer les variables suivantes :
- `delai_SituationDerniereModif` : le nombre de jours écoulés depuis la dernière modification du dossier jusqu'à la date du jour.

— `Annee_Developement_survenance` : le nombre d'années depuis l'enregistrement du dossier dans le système.
— *Hors_Norme* : variable booléenne valant TRUE si l'évaluation du règlement est supérieure à 100.000 euros, FALSE sinon.
— `Solde_reg_Flux` : l'évolution du règlement d'une année à l'autre pour un même dossier.
— `Solde_Rec_Flux` : l'évolution du recouvrement d'une année à l'autre pour un même dossier.
— `Solde_reg_Flux_2` et `Solde_Rec_Flux_2` : l'évolution du règlement et du recouvrement sur un intervalle de deux années.
— `Solde_reg_Flux_3` et `Solde_Rec_Flux_3` : l'évolution du règlement et du recouvrement sur un intervalle de trois années.
— `Diff_Reg` : la différence entre l'évaluation des règlements (`Eval_Reg`) et le solde des règlements (`Solde_Reg`). Calculée comme `Diff_Reg` = `Eval_Reg` - `Solde_Reg`.
— `Diff_Rec` : la différence entre l'évaluation des recours (`Eval_Rec`) et le solde des recours (`Solde_Rec`). Calculée comme `Diff_Rec` = `Eval_Rec` - `Solde_Rec`.

3.3 Analyse et Stratification des Garanties

L'analyse des données en assurance demande une maîtrise fine des garanties proposées ainsi qu'une analyse rigoureuse pour développer des modèles de machine learning performants.

Cette section se penche sur les garanties présentes dans le dataset, détaille la stratégie de stratification adoptée et souligne l'importance de cette démarche dans le cadre du projet. Une attention particulière est accordée à la garantie "Loyers Impayés", qui constitue un cas d'étude pertinent pour la prédiction de la clôture automatique des sinistres.

Présentation des Garanties

Dans les contrats d'assurance, les garanties spécifient les protections offertes face à différents risques. Le dataset étudié comporte 26 garanties distinctes, chacune correspondant à un type particulier de sinistre.

La diversité et la complexité des garanties imposent une analyse méthodique pour en tirer des conclusions précises. Pour simplifier cette analyse et faciliter la modélisation des données, les garanties peuvent être regroupées en plusieurs catégories principales.

Les garanties sont organisées selon la nature des risques couverts, en plusieurs grandes catégories :

- **Dégâts des Eaux (DDEA, DDE)** : Couvre les dommages matériels causés par des fuites d'eau ou des ruptures de canalisations. Il s'agit d'une garantie courante dans les contrats d'assurance habitation.
- **Loyers Impayés (LOYS, LGRL)** : Protège les propriétaires contre la perte de loyers en cas d'impayés, un aspect crucial pour la rentabilité des biens immobiliers.
- **Incendie (INCD, IBAT, INCC, ICON)** : Couvre les dommages causés par le feu, avec des sous-catégories spécifiques pour les bâtiments, les contenus, etc.
- **Vol (VOLD)** : Offre une protection contre le vol de biens personnels ou professionnels.
- **Électrique (ELCT)** : Couvre les dommages dus à des incidents électriques.
- **Responsabilité Civile (RCIM)** : Protège contre les réclamations de tiers pour des dommages causés, essentiel pour les entreprises comme pour les particuliers.
- **Bris de Glace (BDGL, BDGP)** : Couvre les dommages aux vitrages, pertinent pour les commerces et les habitations.
- **Catastrophes Naturelles (CNAT, CNAP)** : Protège contre les dommages dus à des événements naturels majeurs, tels que les inondations ou les séismes.
- **Protection Juridique (PJFP, PRJU, PJBQ)** : Offre une couverture des frais juridiques et une assistance en cas de litige.
- **Autres Garanties** : Inclut des protections moins fréquentes comme les Risques Technologiques.

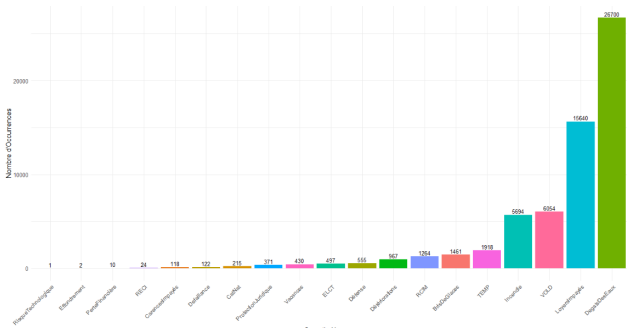

FIGURE 3.4 – Répartition du nombre de dossiers ouverts en 2024 par garantie

Stratégie de Stratification des Garanties

Pour optimiser l'analyse et la modélisation des sinistres, une approche de stratification des données par type de garantie a été retenue. Cette méthode consiste à segmenter les données en sous-groupes homogènes, permettant ainsi de traiter efficacement des volumes importants tout en préservant une pertinence analytique accrue.

En regroupant les sinistres selon leur garantie, la stratification permet de réduire le volume global des données et de concentrer l'analyse sur des sous-ensembles plus restreints mais représentatifs. Cette approche facilite la création de modèles de machine learning plus précis et économes en ressources, adaptés aux spécificités de chaque type de garantie.

Pour des garanties comme "Loyers Impayés", cette méthode permet de développer des modèles prédictifs spécifiques, en tenant compte des caractéristiques distinctes des sinistres associés à cette garantie par rapport à d'autres, telles que les "Dégâts des Eaux". Ainsi, la stratification améliore la précision des prédictions et réduit les erreurs de classification.

Elle est également essentielle pour éviter les biais potentiels dans les modèles, causés par une sous-représentation de certaines garanties. La stratification assure une représentation équilibrée, évitant que les modèles ne soient dominés par les garanties plus fréquentes mais potentiellement moins pertinentes pour l'analyse ciblée.

Focus sur la Garantie "Loyers Impayés"

Parmi les diverses garanties présentes dans le dataset, la garantie "Loyers Impayés" se distingue par son impact économique significatif et sa complexité. Les sinistres liés aux loyers impayés présentent des caractéristiques spécifiques : une fréquence variable, des montants de provision potentiellement élevés, et des durées de règlement souvent longues. Ces éléments rendent la prédiction de la clôture automatique particulièrement pertinente pour optimiser la gestion des sinistres.

La stratification permet d'analyser les sinistres liés aux loyers impayés de manière isolée, même s'ils sont moins fréquents comparés à d'autres types de sinistres dans le dataset global. En se concentrant sur cette garantie, il devient possible de développer des modèles prédictifs plus précis pour estimer la probabilité de clôture automatique des sinistres, améliorant ainsi l'efficacité des processus de gestion et la satisfaction des assurés.

Dans le cadre d'une analyse spécifique à la garantie "Loyers Impayés", les

variables `garantie` et `LB_Produit` seront exclues des variables explicatives. En effet, `LB_Produit` est directement lié à la garantie "Loyers Impayés" sous le produit IMMOCLA.

De même, la variable `TYPEMC_Sinistre`, qui distingue les sinistres matériels des sinistres corporels, est ici redondante. Pour les dossiers relatifs aux Loyers Impayés, seule la modalité "matériel" est concernée, rendant cette variable binaire sans intérêt pour la modélisation des sinistres.

L'absence de diversité dans les modalités de `TYPEMC_Sinistre` signifie qu'elle n'apporte pas de valeur ajoutée à la différenciation des dossiers ou à la modélisation des sinistres. Par conséquent, cette variable sera également écartée. Cette décision contribue à clarifier les données et à améliorer l'efficacité de la modélisation en se concentrant sur les variables réellement pertinentes pour atteindre le résultat souhaité.

En conclusion, cette étude se concentrera exclusivement sur les sinistres liés à la garantie "Loyers Impayés", offrant ainsi une analyse détaillée et ciblée pour optimiser les processus de gestion et de provisionnement associés.

3.4 Statistiques Descriptives et Préparation des Données

Après harmonisation des types de variables choisies, nous nous retrouvons avec les variables restantes suivantes :

Variable	**Type**	**Nombre de Modalités**
Position	Factor	2
Annee_Survenance	Factor	NA
Solde_Reg	Numeric	NA
Eval_Reg	Numeric	NA
Solde_Rec	Numeric	NA
Eval_Rec	Numeric	NA
Solde_Eval_Nette	Numeric	NA
Provision	Numeric	NA
Courtier_Delegataire	Factor	2
Suivi_Judiciaire	Logical	NA
Delai_ModifOuverture	Numeric	NA
Delai_SituationDerniereModif	Numeric	NA
Reouverture	Logical	NA
Annee_Developpement_Survenance	Factor	NA
Solde_Rec_Flux	Numeric	NA
Solde_Reg_Flux_2	Numeric	NA
Solde_Rec_Flux_2	Numeric	NA
Solde_Reg_Flux_3	Numeric	NA
Solde_Rec_Flux_3	Numeric	NA
Solde_Reg_Flux	Numeric	NA
Annee_Ouverture	Factor	10
Diff_Reg	Numeric	NA
Diff_Rec	Numeric	NA
Hors_Norme	Factor	2

TABLE 3.4 – Description des Variables du Dataset

Rappelons que nos données sont divisées en deux ensembles : les données historiques et les données de l'année en cours. Les données historiques regroupe les itérations des dossiers dans notre dataset, et donc leurs évolutions au fil des années, une itération par année de durée de vie. et Nous nous intéresserons ici aux données historiques, car ce sont elles qui constitueront notre dataset d'apprentissage. les données de l'année actuelle représentent létat de

chaque dossier aujourd'hui et constitueront notre dataset de test.

Le but est d'entraîner nos modèles en utilisant les données datant de 2015 à 2023, qui sont les données historiques. En effet, ces données historiques permettent de capturer les tendances et les comportements passés nécessaires pour prédire le comportement des dossiers de sinistres. Une fois les modèles entraînés sur ces données, nous les appliquerons aux dossiers ouverts actuels (2024) pour déterminer lesquels devraient être clôturés.

Ainsi, le processus d'apprentissage consistera à :

— Utiliser les données historiques (de 2015 à 2023) pour entraîner les modèles. Ces données serviront de base pour apprendre à prédire quels dossiers de sinistres nécessitent une clôture.
— Appliquer les modèles ainsi entraînés sur les dossiers ouverts actuels afin de déterminer lesquels doivent être clôturés.

3.5 Analyse des Variables Continues

Commençons notre analyse exploratoire avec l'affichage de statistiques sommaires afin de mieux comprendre nos variables.

Eval_Reg et Eval_Rec

Variable	Minimum	Maximum	Moyenne	Écart-Type
Eval_Reg	0	1,601,163	7,527.39	16,825.20
Eval_Rec	0	89,223.01	1,507.51	2,568.24

TABLE 3.5 – Statistiques descriptives des évaluations des règlements et recouvrements des sinistres

(Boxplot visible en Annexe 7.1)

La variable **Eval_Reg** correspond à l'évaluation des règlements à venir pour un sinistre. Les valeurs observées varient considérablement, avec un minimum de 0 et un maximum dépassant 1,6 million, reflétant la diversité des montants provisionnés selon la gravité ou la complexité des sinistres. La moyenne de 7,527.39, associée à un écart-type élevé de 16,825.20, souligne

une grande dispersion, suggérant la présence de sinistres à très haute valeur, qui peuvent fortement influencer les résultats des modèles prédictifs..

La variable **Eval_Rec**, représentant l'évaluation des recouvrements à venir pour le sinistre, affiche des montants généralement plus modestes, avec un maximum de 89,223.01 et une moyenne de 1,507.51. L'écart-type de 2,568.24, bien que significatif, reste proportionnellement plus faible comparé à celui de **Eval_Reg**. Cela indique que les prévisions de recouvrement sont moins dispersées et potentiellement plus prévisibles. Cependant, l'écart entre les valeurs minimales et maximales suggère qu'il existe tout de même des variations notables qui pourraient affecter l'efficacité du recouvrement dans les processus de gestion des sinistres.

Solde_Reg et Solde_Rec

Variable	Minimum	Maximum	Moyenne	Écart-Type
Solde_Reg	0	1,601,163	6,429.53	17,038.58
Solde_Rec	0	89,223.01	898.32	2,680.55

TABLE 3.6 – Statistiques descriptives des soldes des règlements et recouvrements des sinistres

(Boxplot visible en Annexe 7.2)

La variable **Solde_Reg** représente le solde des règlements associés au sinistre. Les valeurs varient d'un minimum de 0 à un maximum de 1,601,163, ce qui indique une grande variabilité dans les montants déjà réglés. La moyenne de 6,429.53, légèrement inférieure à celle de **Eval_Reg**, suggère que les montants restant à régler peuvent souvent être supérieurs aux montants déjà payés, surtout dans les cas de sinistres complexes ou coûteux. L'écart-type de 17,038.58, proche de celui de **Eval_Reg**, confirme la dispersion importante des montants de règlement, ce qui pourrait indiquer une distribution hétérogène des sinistres en termes de coûts.

La variable **Solde_Rec** correspond au solde des recouvrements associés au sinistre. Cette variable présente une amplitude de valeurs plus restreinte, avec un maximum de 89,223.01 et une moyenne de 898.32, ce qui est inférieur à la moyenne de **Eval_Rec**. Cela peut indiquer que les recouvrements effectivement réalisés sont souvent inférieurs aux évaluations initiales, ou que les

recouvrements sont souvent partiels. L'écart-type de 2,680.55, bien que relativement faible, montre qu'il existe toujours une certaine variabilité dans les recouvrements effectués. Cette variabilité pourrait être liée à des facteurs tels que la solvabilité des débiteurs ou l'efficacité des procédures de recouvrement.

Solde_Eval_Nette et Provision

Variable	Minimum	Maximum	Moyenne	Écart-Type
Solde_Eval_Nette	-42,390.00	1,599,558	6,019.88	16,527.11
Provision	-42,697.40	53,009.67	488.67	1,547.75

TABLE 3.7 – Statistiques descriptives du solde d'évaluation nette et des provisions des sinistres

(Boxplot visible en Annexe 7.3)

La variable **Solde_Eval_Nette** représente la différence nette entre les évaluations des règlements et des recouvrements d'un sinistre. Cette variable peut prendre des valeurs négatives, comme le montre le minimum de -42,390.00, ce qui indique que, pour certains sinistres, les recouvrements prévus pourraient dépasser les règlements restants, un scénario potentiellement favorable pour l'assureur. Le maximum de 1,599,558 et une moyenne de 6,019.88 indiquent toutefois que, dans de nombreux cas, les règlements surpassent les recouvrements, soulignant l'importance de bien gérer les prévisions pour éviter des pertes imprévues. L'écart-type de 16,527.11 reflète une variabilité importante, ce qui nécessite une attention particulière dans les modèles de prédiction pour capter ces écarts significatifs.

La variable **Provision** correspond aux provisions constituées pour couvrir les pertes potentielles liées à un sinistre. Les valeurs négatives, avec un minimum de -42,697.40, suggèrent des ajustements ou des réévaluations à la baisse de certaines provisions précédemment établies. Le maximum de 53,009.67, bien que beaucoup plus modéré que les autres montants évalués, montre que les provisions peuvent être significatives dans certains cas. La moyenne de 488.67, associée à un écart-type de 1,547.75, indique une distribution majoritairement concentrée autour de petites valeurs, mais avec quelques provisions plus élevées qui influencent la variabilité. Ces provisions doivent être précisément calibrées dans les modèles prédictifs pour s'assurer qu'elles reflètent correctement le risque associé à chaque sinistre, en particulier dans les cas complexes tels que les loyers impayés.

Solde_Rec_Flux et Solde_Reg_Flux

Variable	Minimum	Maximum	Moyenne	Écart-Type
Solde_Rec_Flux	-7,708.74	665,000.46	507.14	3,142.29
Solde_Reg_Flux	-7,708.74	665,000.46	518.04	3,141.11

TABLE 3.8 – Statistiques descriptives des flux de règlements et recouvrements des sinistres

(Boxplot visible en Annexe 7.4)

La variable **Solde_Rec_Flux** représente l'évolution des recouvrements d'une année à l'autre pour un même dossier. Les valeurs varient d'un minimum de -7,708.74 à un maximum de 665,000.46, montrant une large amplitude dans les ajustements annuels des recouvrements. Une moyenne de 507.14, associée à un écart-type de 3,142.29, suggère que bien que les variations moyennes soient relativement modestes, il existe des fluctuations importantes qui peuvent influencer le résultat final du recouvrement pour certains sinistres.

La variable **Solde_Reg_Flux** quant à elle, mesure l'évolution des règlements d'une année à l'autre pour un même dossier. Les caractéristiques statistiques de cette variable sont très similaires à celles du **Solde_Rec_Flux**, avec un minimum et un maximum identiques, ainsi qu'une moyenne légèrement plus élevée à 518.04 et un écart-type de 3,141.11. Ces similitudes indiquent que les variations annuelles dans les règlements et les recouvrements suivent des tendances similaires, ce qui peut être dû à une réévaluation conjointe des deux éléments au cours du processus de gestion des sinistres.

Solde_Rec_Flux_2 et Solde_Reg_Flux_2

Variable	Minimum	Maximum	Moyenne	Écart-Type
Solde_Reg_Flux_2	-7,708.74	63,729.52	522.16	2,002.89
Solde_Rec_Flux_2	-7,708.74	63,729.52	465.65	2,010.39

TABLE 3.9 – Statistiques descriptives des flux de règlements et recouvrements sur deux ans

(Boxplot visible en Annexe 7.5)

Les variables **Solde_Reg_Flux_2** et **Solde_Rec_Flux_2** mesurent l'évolution des règlements et des recouvrements respectivement, sur une période de deux ans pour un même dossier.

La **Solde_Reg_Flux_2** montre une variation des règlements allant de -7,708.74 à 63,729.52, avec une moyenne de 522.16 et un écart-type de 2,002.89. Ces valeurs indiquent que les ajustements sur deux ans tendent à être moins extrêmes que ceux observés sur une base annuelle (comme noté précédemment), mais ils demeurent significatifs, suggérant que les sinistres peuvent connaître des révisions financières importantes sur cette période.

De manière similaire, la **Solde_Rec_Flux_2** présente une variation de recouvrements comprise entre -7,708.74 et 63,729.52, avec une moyenne légèrement inférieure à celle des règlements, à 465.65, et un écart-type de 2,010.39. Cela montre que les recouvrements, bien que souvent alignés sur les règlements, peuvent évoluer différemment au fil des ans, avec un certain degré d'incertitude ou de réévaluation dans les processus de recouvrement.

Solde_Rec_Flux_3 et Solde_Reg_Flux_3

Variable	Minimum	Maximum	Moyenne	Écart-Type
Solde_Reg_Flux_3	-5,633.51	63,729.52	338.31	1,219.97
Solde_Rec_Flux_3	-5,633.51	63,729.52	265.50	1,223.79

TABLE 3.10 – Statistiques descriptives des flux de règlements et recouvrements sur trois ans

(Boxplot visible en Annexe 7.6) Les variables **Solde_Reg_Flux_3** et **Solde_Rec_Flux_3** reflètent l'évolution des règlements et des recouvrements sur une période de trois ans pour un même dossier.

Pour la **Solde_Reg_Flux_3**, les variations des règlements s'étendent de -5,633.51 à 63,729.52, avec une moyenne de 338.31 et un écart-type de 1,219.97. Ces chiffres montrent une tendance à la baisse en termes de moyenne et d'écart-type par rapport aux périodes plus courtes, ce qui peut indiquer une stabilisation des ajustements financiers après plusieurs années, ou une diminution des variations extrêmes.

De même, la **Solde_Rec_Flux_3** présente une variation de recouvrements allant de -5,633.51 à 63,729.52, avec une moyenne de 265.50 et un écart-type de 1,223.79. La moyenne plus faible par rapport aux flux sur deux ans indique que les recouvrements deviennent de moins en moins volatils au fil du temps, ce qui pourrait signifier que les montants recouvrés sont plus prévisibles sur des périodes plus longues.

delai_ModifOuverture et delai_SituationDerniereModif

Variable	Minimum	Maximum	Moyenne	Écart-Type
delai_ModifOuverture	0.000	3,048.00	502.75	520.81
delai_SituationDerniereModif	0.000	3,037.00	554.03	581.98

TABLE 3.11 – Statistiques descriptives des délais de modification des dossiers de sinistres

(Boxplot visible en Annexe 7.7 & 7.8)

La variable **delai_ModifOuverture** représente le nombre de jours écoulés entre l'ouverture du dossier et sa dernière modification. Les délais observés varient de 0 à 3,048 jours, avec une moyenne de 502.75 jours et un écart-type de 520.81 jours. Cette dispersion importante suggère que certains dossiers sont modifiés très rapidement après leur ouverture, tandis que d'autres restent longtemps sans modification.

Quant à la variable **delai_SituationDerniereModif**, qui mesure le nombre de jours écoulés depuis la dernière modification du dossier jusqu'à la date d'analyse, les valeurs vont de 0 à 3,037 jours, avec une moyenne de 554.03 jours et un écart-type de 581.98 jours. Cela indique une variation significative dans la fréquence des mises à jour des dossiers, certaines étant très récentes, tandis que d'autres n'ont pas été modifiées depuis longtemps.

Diff_Rec et Diff_Reg

Variable	Minimum	Maximum	Moyenne	Écart-Type
Diff_Rec	0.000	44,970.00	325.58	526.54
Diff_Reg	0.000	54,059.67	586.75	1,238.75

TABLE 3.12 – Statistiques descriptives des différences entre évaluations et soldes

La variable **Diff_Rec** mesure la différence entre l'évaluation des recours (`Eval_Reg`) et le solde des recours (`Solde_Reg`). Les valeurs vont de 0 à 44,970.00, avec une moyenne de 325.58 et un écart-type de 526.54. Cette grande dispersion indique que les différences entre l'évaluation et le solde des recours peuvent être significatives, ce qui peut refléter une variabilité importante dans les estimations des recours par rapport aux soldes réellement enregistrés. Une évaluation plus précise pourrait améliorer la gestion et la prévision des recours.

La variable **Diff_Reg** quantifie la différence entre l'évaluation des règlements (`Eval_Rec`) et le solde des règlements (`Solde_Rec`). Les valeurs varient de 0 à 54,059.67, avec une moyenne de 586.75 et un écart-type de 1,238.75. La dispersion élevée suggère des variations importantes dans les différences entre les évaluations et les soldes des règlements. Les grandes différences pourraient indiquer des ajustements nécessaires dans les évaluations ou des écarts importants entre les prévisions et les règlements réels, nécessitant une analyse plus approfondie pour optimiser les prévisions et le provisionnement.

3.6 Test de Kruskal-Wallis

Le test de Kruskal-Wallis est une méthode non paramétrique utilisée pour comparer la distribution d'une variable quantitative entre plusieurs groupes indépendants. Ce test est particulièrement adapté lorsque les hypothèses de normalité ne sont pas respectées, ce qui est souvent le cas dans les données réelles utilisées en machine learning.

Dans notre étude, nous appliquons le test de Kruskal-Wallis pour évaluer comment la variable cible `position` influence les variables numériques du dataset. Un aperçu visuel de la répartition de ces variables par rapport à la variable cible est présenté pour obtenir une intuition préliminaire sur l'impact de cette variable sur les variables quantitatives.

Fonctionnement Mathématique

Le test de Kruskal-Wallis est une extension du test de Wilcoxon-Mann-Whitney pour comparer plus de deux groupes. Il se base sur les rangs des données plutôt que sur leurs valeurs brutes, ce qui le rend robuste aux violations de l'hypothèse de normalité. Voici les principales étapes du test :

1. **Classement :** Les observations de toutes les conditions sont combinées et classées en ordre croissant.
2. **Calcul des rangs moyens :** Les rangs moyens sont calculés pour chaque groupe.
3. **Calcul de la statistique de test :** La statistique de test H est calculée selon la formule :

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \bar{R}\right)^2$$

   où n_i est la taille du groupe i, $\bar{R}_i$ est le rang moyen du groupe i, $\bar{R}$ est le rang moyen global, et N est le nombre total d'observations.
4. **Distribution de H :** Sous l'hypothèse nulle (pas de différence entre les groupes), H suit une distribution du chi-carré avec $k-1$ degrés de liberté.

Résultats du Test

Le tableau 3.13 présente les résultats du test de Kruskal-Wallis appliqué aux variables numériques de notre dataset en fonction de la variable cible `position`.

Sachhant que les p-value à 0 sont des résultatsa arrondis, les résultats montrent que toutes les variables présentent une p-value très inférieure au seuil de significativité généralement utilisé (0.05), indiquant que les distributions des variables numériques diffèrent de manière significative entre les groupes définis par la variable cible `position`.

— `Solde_Reg` : La différence marquée entre les groupes suggère que le montant des règlements varie de manière significative selon la situation du dossier (`position`). Cela pourrait indiquer que les dossiers ouverts ont des montants de règlement plus élevés ou plus variables par rapport aux dossiers clôturés.
— `Eval_Reg` : La forte statistique de test et la p-value proche de zéro indiquent une variabilité significative dans les évaluations des règlements entre les groupes de `position`. Les différences peuvent refléter

Variable	Statistique de test	P-Value
Solde_Reg	21,950.07	0.0000
Eval_Reg	66,342.69	0.0000
Solde_Rec	329.20	1.43e-73
Eval_Rec	48,711.94	0.0000
Solde_Eval_Nette	68,681.71	0.0000
Provision	3,306.12	0.0000
delai_ModifOuverture	10,369.59	0.0000
delai_SituationDerniereModif	48,778.22	0.0000
Solde_Rec_Flux	12,207.94	0.0000
Solde_Reg_Flux_2	33,509.37	0.0000
Solde_Rec_Flux_2	33,498.86	0.0000
Solde_Reg_Flux_3	39,500.28	0.0000
Solde_Rec_Flux_3	39,646.50	0.0000
Solde_Reg_Flux	12,233.26	0.0000
Diff_Reg	57,902.36	0.0000
Diff_Rec	81,341.06	0.0000

TABLE 3.13 – Résultats du test de Kruskal-Wallis pour les variables numériques par rapport à la variable cible `position`.

des pratiques d'évaluation distinctes ou des seuils de provision différents pour les dossiers ouverts et clôturés.

— `Solde_Rec` : Cette variable, bien que présentant une statistique de test plus basse, montre une p-value extrêmement significative, suggérant que les montants des recouvrements varient également de manière notable entre les groupes. Les dossiers ouverts pourraient avoir des soldes de recouvrement plus élevés en raison de recouvrements en cours.

— `Eval_Rec` : Comme avec `Eval_Reg`, les différences significatives entre les groupes peuvent être dues à des évaluations de recouvrement très variables. Cela pourrait refléter des approches différentes dans l'évaluation des potentiels recouvrements futurs.

— `Solde_Eval_Nette` : La statistique élevée et la p-value très basse indiquent que la différence de solde net est significative entre les groupes. Cela pourrait indiquer une gestion des provisions nettement différente entre les dossiers ouverts et clôturés.

— `Provision` : Les différences significatives suggèrent que les montants provisionnés varient considérablement entre les groupes, ce qui pourrait influencer la gestion des sinistres.

— `delai_ModifOuverture` et `delai_SituationDerniereModif` : Les délais de modification et de mise à jour du dossier montrent des diffé-

rences significatives, ce qui peut indiquer que les dossiers ouverts sont modifiés plus fréquemment ou plus récemment que les dossiers clôturés.

— `Solde_Rec_Flux`, `Solde_Reg_Flux_2`, `Solde_Rec_Flux_2`, `Solde_Reg_Flux_3`, `Solde_Rec_Flux_3` : Ces variables montrent des différences significatives dans les évolutions annuelles des soldes de règlement et de recouvrement, ce qui peut refléter des tendances de modification dans les dossiers ouverts comparés aux dossiers clôturés.

— `Diff_Reg` et `Diff_Rec` : Les différences entre les évaluations et les soldes des règlements et des recouvrements sont significatives, ce qui suggère des variations importantes dans la manière dont les montants sont ajustés ou recouverts selon la situation du dossier.

Le test de Kruskal-Wallis révèle des différences significatives entre les groupes de la variable cible `position` pour toutes les variables numériques examinées. Ces résultats soulignent l'importance de la variable `position` dans la compréhension des caractéristiques des dossiers de sinistre et de la gestion des provisions. En effet, la variabilité significative observée dans les variables telles que les soldes, les évaluations et les différences entre les prévisions et les soldes suggère que la variable `position` est un déterminant clé des montants et des évolutions des sinistres.

3.7 Discrétisation des Variables Continues

La discrétisation des variables continues en déciles est une technique utile pour convertir des données numériques en catégories discrètes. Cette méthode est particulièrement bénéfique pour les modèles de machine learning tels que la régression logistique, les arbres de décision et les forêts aléatoires. Nous analysons l'efficacité de cette approche pour chaque variable explicative et présentons les observations basées sur la répartition des déciles.

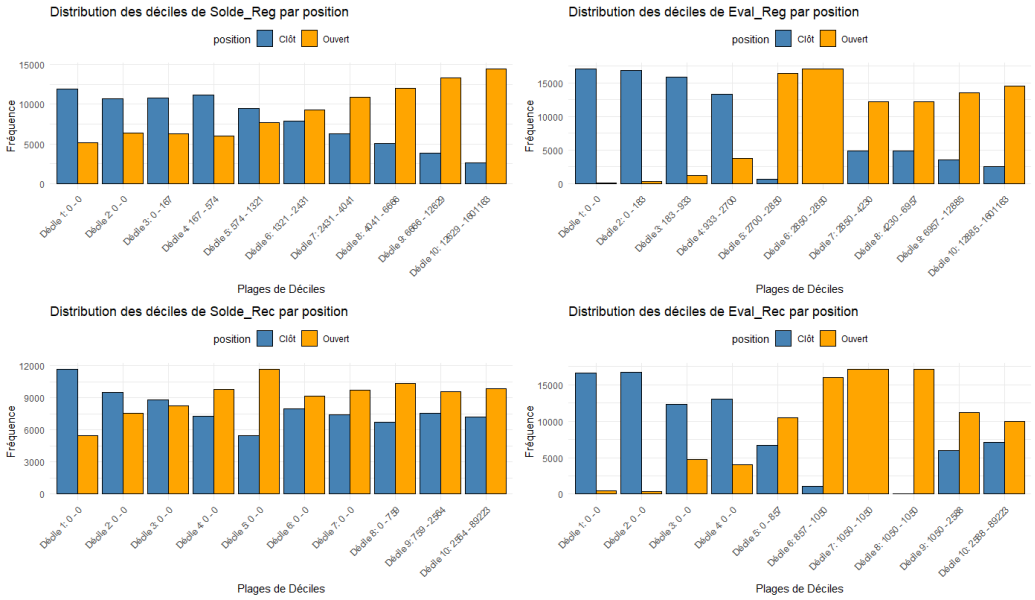

FIGURE 3.5 – Répartition par déciles des variables Solde_Reg, Solde_Rec, Eval_Reg, Eval_Rec.

Solde_Reg

La répartition par déciles montre que les plages supérieures de `Solde_Reg` sont associées à des valeurs plus élevées pour `Ouvert`, tandis que les plages inférieures sont associées à `Clôt`.

Eval_Reg

Les plages supérieures des déciles présentent des valeurs extrêmes, avec des différences marquées entre les fréquences de `Clôt` et `Ouvert`.

Solde_Rec

Les fréquences varient de manière significative entre les plages de déciles, montrant une bonne séparation entre `Clôt` et `Ouvert`.

Eval_Rec

Les plages de déciles montrent une concentration élevée dans les déciles inférieurs et supérieurs, avec des valeurs distinctes pour `Clôt` et `Ouvert`.

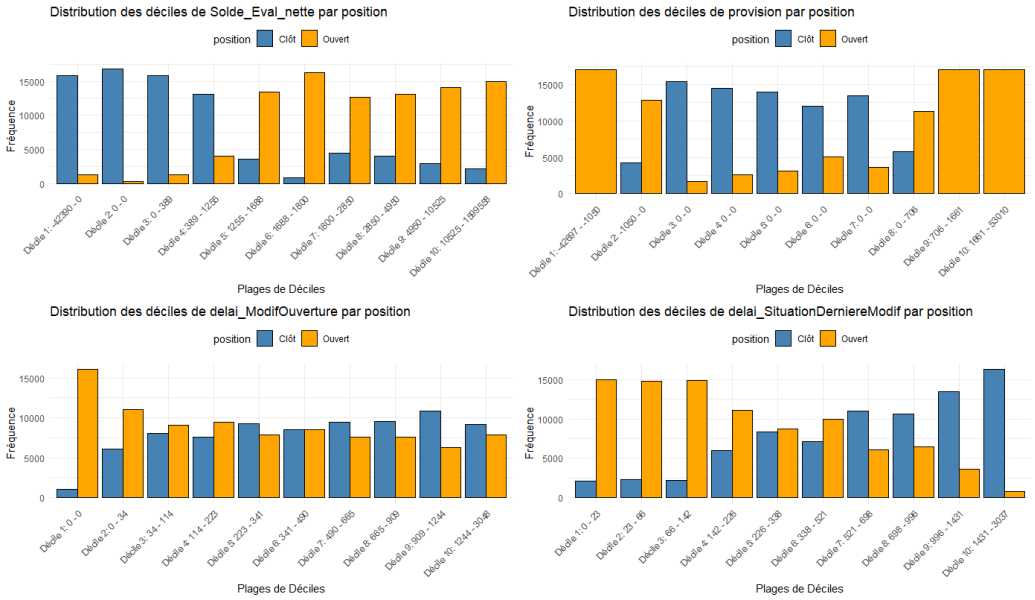

FIGURE 3.6 – Répartition par déciles des variables Provision, Solde_Eval_Nette, Delai_Modif_Ouverture et Delai_SituationDerniereModif.

Solde_Eval_Nette

La discrétisation en déciles montre une bonne séparation entre `Clôt` et `Ouvert`.

Provision

La variable présente une concentration significative dans les déciles extrêmes avec une variation importante dans les plages centrales.

Delai_Modif_Ouverture

Bonne séparation des fréquences entre les plages de déciles, avec des différences notables entre les catégories.

Delai_SituationDerniereModif

Les plages de déciles montrent des variations substantielles entre `Clôt` et `Ouvert`.

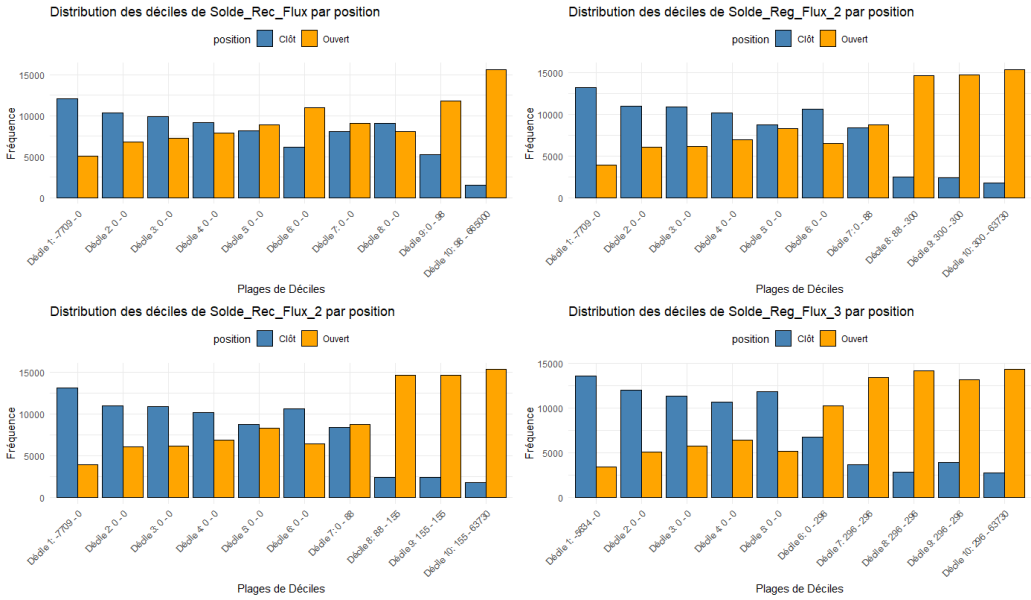

FIGURE 3.7 – Répartition par déciles des variables Solde_Rec_Flux, Solde_Rec_Flux_2, Solde_Reg_Flux_2 et Solde_Reg_Flux_3.

Solde_Rec_Flux

Les valeurs élevées dans les plages supérieures pour `Ouvert` et dans les plages inférieures pour `Clôt`.

Solde_Reg_Flux_2

On observe des Différences notables entre `Clôt` et `Ouvert`, avec des variations dans les plages supérieures et inférieures.

Solde_Rec_Flux_2

Bonne séparation entre les plages de déciles, avec des valeurs distinctes dans les plages supérieures.

Solde_Reg_Flux_3

éparation claire entre `Clôt` et `Ouvert`, avec des valeurs extrêmes dans les plages supérieures.

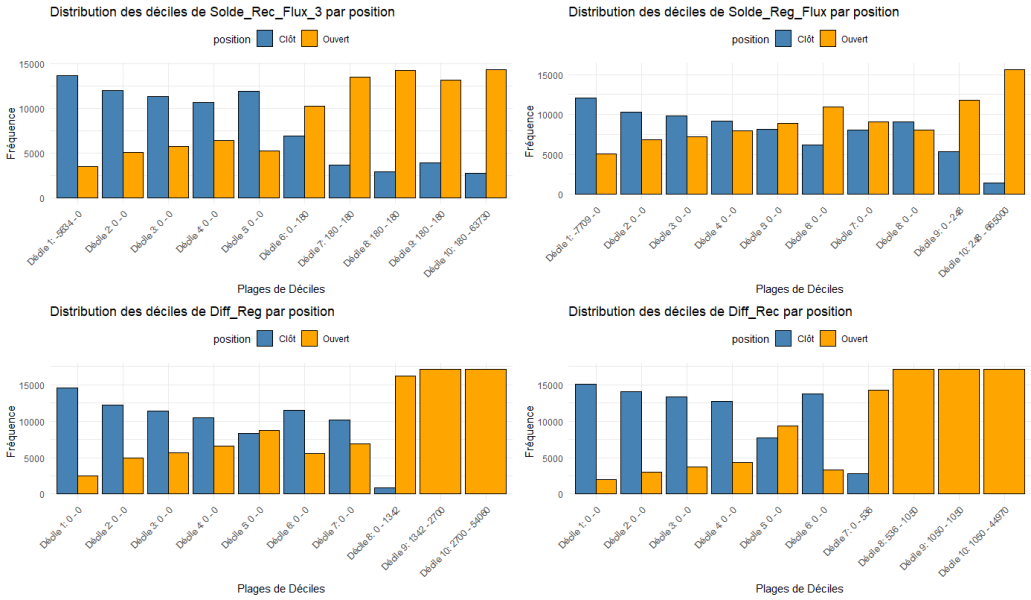

FIGURE 3.8 – Répartition par déciles des variables Diff_Reg, Diff_Rec, Solde_Reg_Flux et Solde_Rec_Flux_3.

Solde_Rec_Flux_3

Différences significatives entre les plages de déciles, avec des valeurs distinctes pour `Clôt` et `Ouvert`.

Solde_Reg_Flux

Variations marquées dans les fréquences entre `Clôt` et `Ouvert`, avec des valeurs extrêmes dans les plages supérieures.

Diff_Reg

Les déciles supérieurs montrent des valeurs extrêmes, avec une concentration élevée de `Clôt` dans les plages inférieures.

Diff_Rec

Les plages de déciles montrent des valeurs élevées pour `Clôt` dans les plages inférieures et des valeurs élevées pour `Ouvert` dans les plages supérieures.

Tableau récapitulatif des Variables et Possibilités de Discrétisation

Nom Variable	À Discrétiser	Motif
Solde_Reg	Non	Conservation des détails
Eval_Reg	Non	Grande variabilité
Solde_Rec	Non	Distributions variées
Eval_Rec	Non	Distributions variées
Solde_Eval_Nette	Oui	Large plage de valeurs
Provision	Oui	Réduire les valeurs extrêmes
Delai_Modif_Ouverture	Oui	Large plage de valeurs
Delai_SituationDerniereModif	Oui	Large plage de valeurs
Solde_Rec_Flux	Oui	Réduire les valeurs extrêmes
Solde_Reg_Flux_2	Oui	Réduire les valeurs extrêmes
Solde_Rec_Flux_2	Oui	Réduire les valeurs extrêmes
Solde_Reg_Flux_3	Oui	Réduire les valeurs extrêmes
Solde_Rec_Flux_3	Oui	Réduire les valeurs extrêmes
Solde_Reg_Flux	Oui	Réduire les valeurs extrêmes
Diff_Reg	Non	Conservation des détails
Diff_Rec	Non	Conservation des détails

TABLE 3.14 – Discrétisation des variables et justification

3.8 Méthodologie pour l'Analyse des Variables Continues

Dans cette section, nous détaillons la méthodologie employée pour discrétiser les variables continues sélectionnées. Pour ce faire, nous avons opté pour l'utilisation du clustering k-means. Cependant, il est primordial de déterminer le nombre de clusters optimal avant d'appliquer cette méthode. Chaque cluster représentera une plage de données dans laquelle les variables seront discrétisées. Nous avons utilisé plusieurs techniques pour déterminer ce nombre optimal de clusters : la méthode du coude, le Gap Statistic et le BIC (Bayesian Information Criterion).

Les variables continues analysées sont les suivantes :
— **Solde_Eval_nette**
— **Provision**
— **Delai_ModifOuverture**
— **Delai_SituationDerniereModif**
— **Solde_Rec_Flux**

— **Solde_Reg_Flux_2**
— **Solde_Rec_Flux_2**
— **Solde_Reg_Flux_3**
— **Solde_Rec_Flux_3**
— **Solde_Reg_Flux**

Pour déterminer le nombre optimal de clusters, nous avons employé trois méthodes distinctes, chacune ayant ses propres caractéristiques et avantages. Nous illustrons ces méthodes avec les résultats obtenus pour la variable "Provision".

Méthode du Coude

La méthode du coude est une technique visuelle qui consiste à tracer la somme des carrés intra-cluster (WSS) en fonction du nombre de clusters. La WSS est définie par :

$$WSS = \sum_{i=1}^{n} \sum_{j=1}^{k} \|x_{ij} - \mu_j\|^2$$

où x_{ij} est l'observation i-ème dans le j-ème cluster, et μ_j est la moyenne des points dans le j-ème cluster. En observant le graphique de WSS en fonction du nombre de clusters, nous identifions le point où la diminution de WSS devient moins significative, connu sous le nom de "coude". Ce point indique le nombre optimal de clusters, équilibrant complexité et précision.

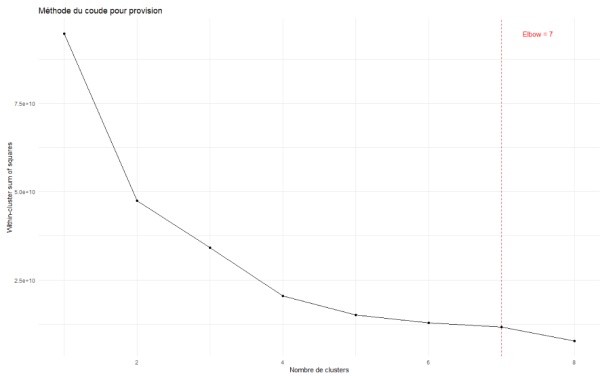

FIGURE 3.9 – Méthode du coude pour la variable Provision

Le résultat technique de la méthode du coude pour la variable "Provision" suggère 7 clusters.

Gap Statistic

Le Gap Statistic évalue la performance du clustering en comparant le clustering réel à un clustering aléatoire. Il est défini par :

$$\text{Gap}(k) = \frac{1}{B} \sum_{b=1}^{B} \log(WSS_b) - \log(WSS_k)$$

où WSS_k est la WSS pour k clusters, et WSS_b est la WSS pour un échantillon aléatoire avec B répétitions. Le nombre optimal de clusters est celui qui maximise le Gap Statistic, indiquant un ajustement significativement meilleur par rapport à un clustering aléatoire.

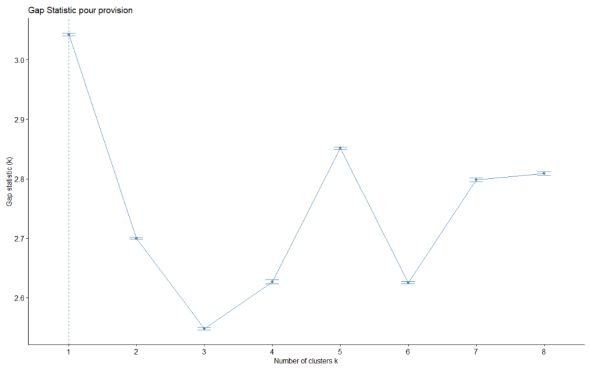

FIGURE 3.10 – Gap Statistic pour la variable Provision

Le résultat du Gap Statistic pour la variable "Provision" indique 5 clusters. Cependant, il convient de noter que le Gap Statistic montre également une valeur très faible pour 1 cluster, suggérant que cette méthode pourrait ne pas être entièrement adaptée à la variable "Provision".

Bayesian Information Criterion (BIC)

Le BIC est un critère de sélection de modèle qui pénalise les modèles plus complexes pour éviter le sur-ajustement. Il est calculé par :

$$\text{BIC} = -2\log(L) + k\log(n)$$

où L est la vraisemblance du modèle, k est le nombre de paramètres estimés, et n est le nombre d'observations. Le nombre optimal de clusters est celui qui minimise le BIC, représentant le modèle le plus parcimonieux qui explique les données de manière efficace.

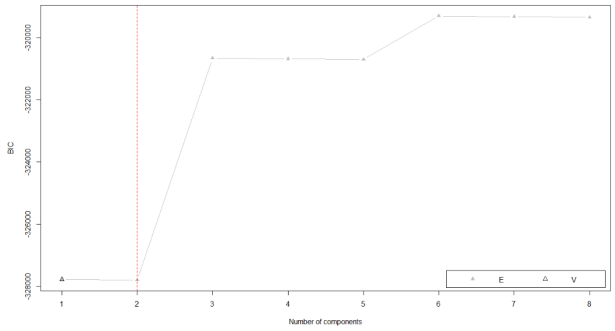

FIGURE 3.11 – BIC pour la variable Provision

Le résultat avec la méthode BIC pour la variable "Provision" indique 2 clusters.

Analyse des Résultats

Les résultats obtenus par les différentes méthodes montrent des divergences notables :

— **Méthode du Coude** : Suggère 7 clusters, ce qui peut indiquer une complexité plus élevée pour capturer la variabilité des données.
— **Gap Statistic** : Recommande 5 clusters mais aussi une valeur très basse pour 1 cluster, ce qui pourrait indiquer une limitation de la méthode pour cette variable spécifique.
— **BIC** : Indique 2 clusters, favorisant un modèle plus simple et potentiellement plus généralisable.

Ces divergences suggèrent que le choix du nombre de clusters peut être influencé par la méthode utilisée et la nature des données. On se basera sur de multiples essais et notre intuition pour poursuivre.

3.9 Clustering avec la Méthode des K-means

Principe de la Méthode K-means

La méthode des K-means est une technique de clustering largement utilisée pour partitionner un ensemble de données en k clusters distincts, de manière à minimiser la variance intra-cluster. Chaque cluster est défini par son centre, ou centroïde, qui est la moyenne des points de données apparte-

nant à ce cluster.

Mathématiquement, l'objectif de K-means est de minimiser la somme des distances au carré entre chaque point de données et le centre du cluster auquel il appartient. Cette fonction objective est exprimée par :

$$J(C) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

où C_i représente le i-ème cluster, μ_i est le centroïde du i-ème cluster, et x est un point de données. La minimisation de cette fonction permet de trouver une partition qui regroupe les points de données les plus similaires.

Intérêt et Faisabilité

La méthode K-means présente plusieurs avantages, notamment :

— **Simplicité et Efficacité** : K-means est facile à comprendre et à implémenter. Il est également rapide, même pour des ensembles de données volumineux, ce qui le rend adapté à des applications à grande échelle.
— **Clarté** : Les résultats de K-means sont facilement interprétables, offrant une partition claire en groupes distincts.
— **Utilisation Répandue** : Cette méthode est couramment utilisée dans divers domaines, tels que le marketing, l'analyse des données, et le traitement d'image, notamment pour la segmentation et la réduction de dimensionnalité.

Cependant, K-means présente également certaines limitations :

— **Scalabilité** : Bien que K-means puisse gérer efficacement des ensembles de données volumineux, il nécessite que le nombre de clusters k soit spécifié à l'avance, ce qui peut nécessiter une évaluation préliminaire.
— **Sensibilité aux Initialisations** : Les résultats de K-means peuvent varier en fonction des points initiaux choisis pour les centres des clusters. Des méthodes d'initialisation comme k-means++ peuvent améliorer cette étape en réduisant le risque d'obtenir une solution locale optimale.

3.10 Analyse des Variables Discrétisées : Application Pratique

Discrétisation de la Variable *délai_modifOuverture*

La variable *délai_modifOuverture* a été discrétisée en utilisant la méthode de clustering K-means pour générer trois plages distinctes. Cette discrétisation nous permet de transformer une variable continue en catégories qualitatives, facilitant ainsi une analyse plus approfondie et significative des données.

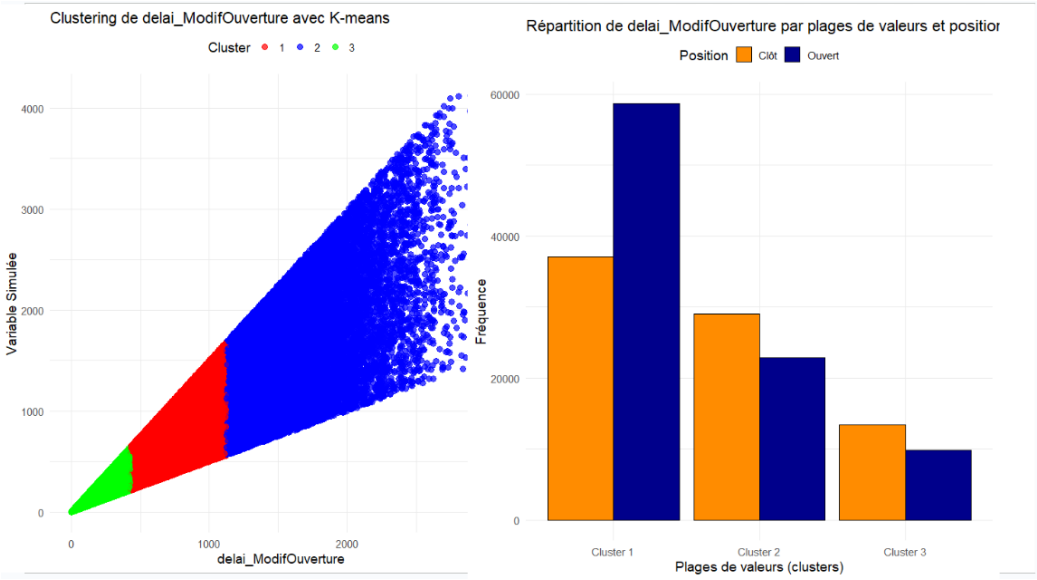

FIGURE 3.12 – Discrétisation par clustering de la variable *délai_modifOuverture* avec histogramme des plages

Le clustering révèle une segmentation efficace des observations :

Cluster	Valeur Min	Valeur Max	Nombre d'observations
1	429	1123	51892
2	1124	3048	23220
3	0	428	95759

TABLE 3.15 – Résultats pour le clustering de la variable *délai_modifOuverture*

— **Cluster 1** : Regroupe les valeurs de 429 à 1123, englobant 51,892 observations. Ce cluster représente une plage intermédiaire qui pourrait correspondre à des délais modérés dans le processus d'ouverture.
— **Cluster 2** : Capture les valeurs de 1124 à 3048, avec 23,220 observations. Ce groupe représente des délais plus longs, reflétant potentiellement des cas où le processus d'ouverture prend plus de temps.
— **Cluster 3** : Englobe les valeurs de 0 à 428, regroupant 95,759 observations. Ce cluster correspond à des délais plus courts et fréquents.

Cette discrétisation permet de segmenter la variable de manière à mieux comprendre les tendances et anomalies dans les délais d'ouverture, offrant ainsi une vue plus claire sur les processus étudiés.

Discrétisation de la Variable *délai_SituationDernièreModif*

La variable *délai_SituationDernièreModif* a également été discrétisée en quatre clusters distincts, fournissant une vue détaillée des plages de valeurs pour cette variable continue :

Cluster	Valeur Min	Valeur Max	Nombre d'observations
1	0	374	91397
2	1606	3037	12344
3	375	900	39448
4	901	1605	27682

TABLE 3.16 – Résultats pour le clustering de la variable *délai_SituationDernièreModif*

Les résultats pour cette variable montrent une discrétisation utile :

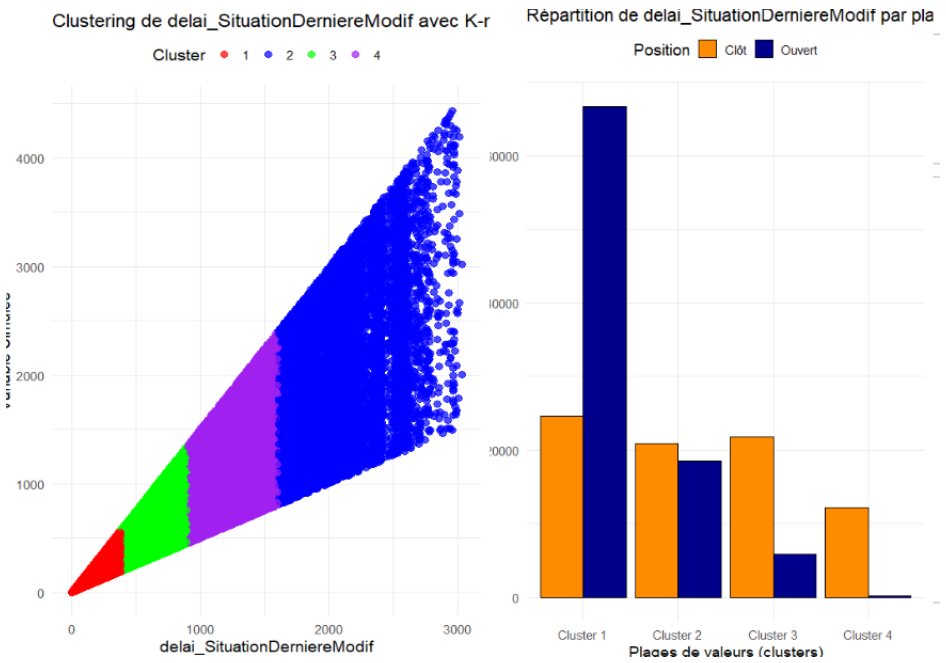

FIGURE 3.13 – Discrétisation par clustering de la variable *délai_SituationDernièreModif* avec histogramme des plages

— **Cluster 1** : Comprend les valeurs de 0 à 374, avec 91,397 observations. Ce groupe représente des délais relativement courts, suggérant une fréquence élevée des modifications dans cette plage.
— **Cluster 2** : Couvre les valeurs de 1606 à 3037, avec 12,344 observations. Ce cluster montre des délais plus longs entre les situations, probablement moins fréquents.
— **Cluster 3** : Englobe les valeurs de 375 à 900, regroupant 39,448 observations. Ce cluster se situe entre les délais courts et longs, représentant des délais modérés.
— **Cluster 4** : Inclut les valeurs de 901 à 1605, avec 27,682 observations. Ce groupe montre des délais plus longs, indiquant moins de fréquence dans les modifications.

Ces clusters offrent une vision plus détaillée des délais de modification, facilitant une compréhension plus fine des processus liés à ces variables.

Conclusion de la discrétisation

La discrétisation par clustering s'est révélée particulièrement bénéfique pour certaines variables, telles que *délai_modifOuverture* et *délai_SituationDernièreModif*. Les résultats ont permis de transformer des variables continues en catégories

significatives, offrant des perspectives plus claires et pertinentes pour les analyses actuarielles.

En revanche, pour la variable *Provision*, la discrétisation par clustering n'a pas montré de gains significatifs en termes d'analyse ou de performance des modèles. Cette observation souligne l'importance d'évaluer les méthodes de discrétisation en fonction des spécificités des données et des objectifs d'analyse.

Ainsi, bien que la méthode K-means soit un outil précieux pour la discrétisation, il est essentiel de choisir judicieusement le nombre de clusters et d'évaluer les résultats en fonction du contexte spécifique des données et des analyses prévues. Les méthodes de discrétisation appropriées peuvent grandement influencer la qualité et la pertinence des analyses dans le domaine actuarielles et au-delà.

3.11 Méthode de l'ACP

Analyse Préliminaire des Corrélations

Avant d'entamer l'Analyse en Composantes Principales (ACP), il est crucial de saisir les relations sous-jacentes entre les variables quantitatives de notre dataset. Pour ce faire, nous avons généré une heatmap des corrélations en utilisant le coefficient de Spearman, un outil particulièrement adapté pour appréhender des relations monotones, qu'elles soient linéaires ou non.

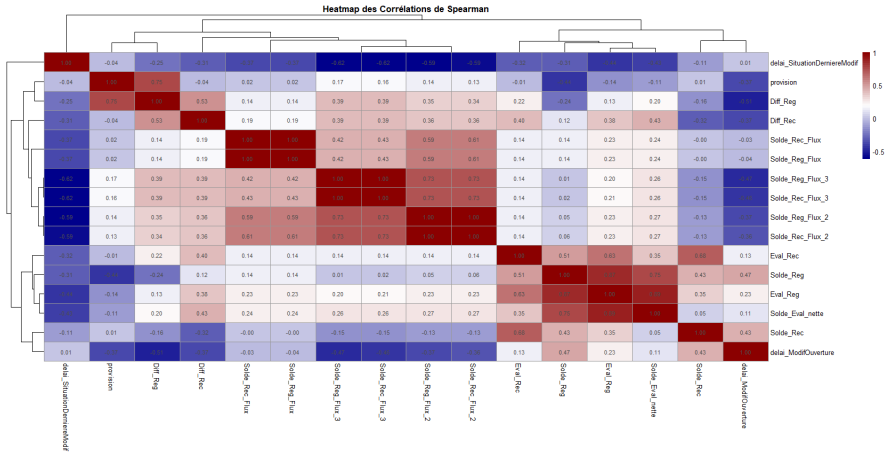

FIGURE 3.14 – Heatmap des corrélations de Spearman entre les variables quantitatives

Comme l'illustre la figure 3.14, les variables liées aux flux financiers révèlent des corrélations positives très marquées entre elles. De plus, ces variables semblent présenter une corrélation négative notable avec la variable *délai_ ModifOuverture*. Les variables *Eval Reg* et *Solde Reg* sont également fortement corrélées entre elles. Ces observations soulignent la nécessité d'appliquer l'ACP pour réduire la dimensionnalité tout en préservant l'essentiel de l'information contenue dans ces corrélations.

Introduction à l'Analyse en Composantes Principales

L'Analyse en Composantes Principales (ACP) est une méthode statistique puissante, conçue pour réduire la dimensionnalité d'un ensemble de données tout en conservant le plus possible de variance. Cette technique transforme un ensemble de variables corrélées en un nouveau jeu de variables, appelées composantes principales, qui sont non corrélées entre elles. Les composantes principales sont des combinaisons linéaires des variables originales, ordonnées de manière à capturer la majorité de la variance avec un minimum de dimensions.

Dans les domaines du machine learning et des statistiques, l'ACP s'avère particulièrement utile pour simplifier les modèles, réduire le bruit et éviter les problèmes de multicolinéarité. Nous allons maintenant décrire les étapes de l'ACP appliquées à notre dataset, analyser les résultats obtenus, et discuter de leur pertinence pour l'intégration dans des modèles prédictifs.

L'ACP suit un processus méthodique en plusieurs étapes, permettant de transformer les données d'origine en un ensemble de variables simplifiées.

1. Calcul de la Matrice de Covariance

La première étape de l'ACP est le calcul de la matrice de covariance des variables quantitatives. Cette matrice résume la manière dont les variables varient ensemble. Pour des données centrées, la matrice de covariance $\mathbf{\Sigma}$ est exprimée par :

$$\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

où $\mathbf{X}$ représente la matrice des données centrées, après soustraction de la moyenne de chaque variable.

2. Décomposition en Valeurs Propres

Une fois la matrice de covariance obtenue, nous procédons à sa décomposition en valeurs propres et vecteurs propres. Les valeurs propres (ou *eigenvalues*) quantifient l'importance de chaque direction principale dans les données, tandis que les vecteurs propres (ou *eigenvectors*) définissent ces directions dans l'espace des variables. Cette décomposition est exprimée comme suit :

$$\mathbf{\Sigma} = \mathbf{W}\Lambda\mathbf{W}^T$$

où $\mathbf{W}$ est la matrice des vecteurs propres et Λ est la matrice diagonale des valeurs propres.

3. Sélection des Composantes Principales

Les composantes principales sont sélectionnées en fonction de leur contribution à l'inertie totale du dataset, c'est-à-dire la variance totale expliquée par les composantes. Une méthode courante consiste à retenir les composantes dont les valeurs propres surpassent la moyenne des valeurs propres. Cela garantit que chaque composante retenue contribue de manière significative à l'explication de la variance des données :

$$\text{Moyenne de l'inertie} = \frac{\sum_{i=1}^{p} \lambda_i}{p}$$

Formulation Mathématique de l'ACP

L'ACP repose sur des principes mathématiques fondamentaux qui assurent une transformation optimale des données corrélées en un ensemble de variables non corrélées.

- **Matrice des Données :** Nous commençons avec une matrice de données $\mathbf{X}$ de dimension $n \times p$, où n est le nombre d'observations et p le nombre de variables.
- **Centrage des Données :** Les données sont centrées en soustrayant la moyenne de chaque variable, donnant ainsi la matrice $\tilde{\mathbf{X}}$.
- **Calcul de la Matrice de Covariance :** La matrice de covariance $\mathbf{C}$ est ensuite calculée pour $\tilde{\mathbf{X}}$:

$$\mathbf{C} = \frac{1}{n-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$$

— **Valeurs Propres et Vecteurs Propres :** Les valeurs propres et vecteurs propres de $\mathbf{C}$ sont déterminés. Les vecteurs propres définissent les directions des nouvelles composantes principales, tandis que les valeurs propres indiquent l'importance relative de ces directions.
— **Projection des Données :** Enfin, les données originales sont projetées dans l'espace des composantes principales, produisant le jeu de données réduit :

$$\mathbf{Z} = \mathbf{XW}$$

où $\mathbf{W}$ est la matrice des vecteurs propres associés aux plus grandes valeurs propres.

Ainsi, les nouvelles variables $\mathbf{Z}$ obtenues sont les composantes principales, représentant des combinaisons linéaires des variables originales, ordonnées selon la quantité de variance expliquée. Ces composantes principales permettent de simplifier les modèles tout en conservant l'essentiel de l'information des données d'origine.

De plus, les résultats de l'ACP offrent plusieurs perspectives importantes :

— **Interprétation des Groupes :** Les groupes de variables identifiés peuvent être interprétés comme des ensembles de variables présentant des caractéristiques similaires. Les variables fortement corrélées au sein d'un groupe peuvent être redondantes, simplifiant ainsi l'analyse en combinant ces variables ou en sélectionnant celles ayant la contribution la plus significative.
— **Réduction Dimensionnelle :** Utiliser les résultats de l'ACP pour réduire la dimension des données en sélectionnant les composantes principales expliquant la majorité de la variance. Cela aide à simplifier les modèles ultérieurs et à améliorer l'efficacité des analyses.
— **Analyse de la Corrélation :** Examiner plus en détail les corrélations entre les variables au sein des groupes. Cela peut inclure des analyses de régression ou d'autres techniques statistiques pour mieux comprendre les relations et les impacts des variables sur les résultats.

En conclusion, l'ACP constitue une base robuste pour la réduction de dimension et l'exploration des relations entre variables.

Résultats de l'Analyse en Composantes Principales (ACP)

Les résultats de l'Analyse en Composantes Principales (ACP) offrent une vue détaillée sur la structure des données ainsi que sur la projection des variables et des individus dans les dimensions principales. Voici un résumé des principaux éléments fournis par l'ACP :

- **Valeurs propres** (`eig`) :
  - Les valeurs propres représentent la variance expliquée par chaque composante principale. Elles sont essentielles pour évaluer l'importance des dimensions principales.
- **Coordonnées des variables** (`var$coord`) :
  - Indiquent la position des variables dans le plan des composantes principales. Elles aident à visualiser les relations entre les variables.
- **Corrélations variables - dimensions** (`var$cor`) :
  - Montrent comment chaque variable est liée aux dimensions principales. Une forte corrélation indique une variable bien représentée par une composante principale.
- **Cos2 pour les variables** (`var$cos2`) :
  - Mesure la qualité de la représentation des variables sur le plan factoriel. Un cos2 élevé signifie que la variable est bien représentée dans le plan principal.
- **Coordonnées des individus** (`ind$coord`) :
  - Montrent la position des individus dans le plan des composantes principales. Utilisées pour visualiser les groupes d'individus et leurs relations.
- **Cos2 pour les individus** (`ind$cos2`) :
  - Mesure la qualité de la représentation des individus dans le plan principal. Un cos2 élevé indique que les individus sont bien représentés dans le plan principal.
- **Contributions des variables** (`var$contrib`) :
  - Mesurent l'importance relative de chaque variable pour chaque composante principale. Aident à identifier les variables les plus influentes.
- **Contributions des individus** (`ind$contrib`) :
  - Montrent l'importance relative de chaque individu pour chaque composante principale. Permettent de détecter les individus les plus influents.

L'analyse des valeurs propres (`eig`), des coordonnées (`var$coord` et `ind$coord`), des corrélations (`var$cor`), des cos2 (`var$cos2` et `ind$cos2`), et des contri-

butions (`var$contrib` et `ind$contrib`) est essentielle pour comprendre la structure des données. Ces éléments permettent de visualiser et d'interpréter les relations entre variables et individus, facilitant ainsi une analyse approfondie et une prise de décision éclairée basée sur les dimensions principales identifiées.

3.12 Présentation des Résultats de l'ACP

Variance Expliquée

L'Analyse en Composantes Principales (ACP) permet de réduire la dimensionnalité des données tout en conservant autant d'information que possible. Les valeurs propres (λ_i) quantifient la variance expliquée par chaque composante principale. La proportion de variance expliquée par la composante i est calculée comme suit :

$$\text{Proportion de variance expliquée par la composante } i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

où λ_i est la valeur propre associée à la composante i et $\sum_{j=1}^{p} \lambda_j$ représente la variance totale des données.

Le graphique ci-dessous montre que les 5 premières dimensions capturent plus de 80% de la variance totale :

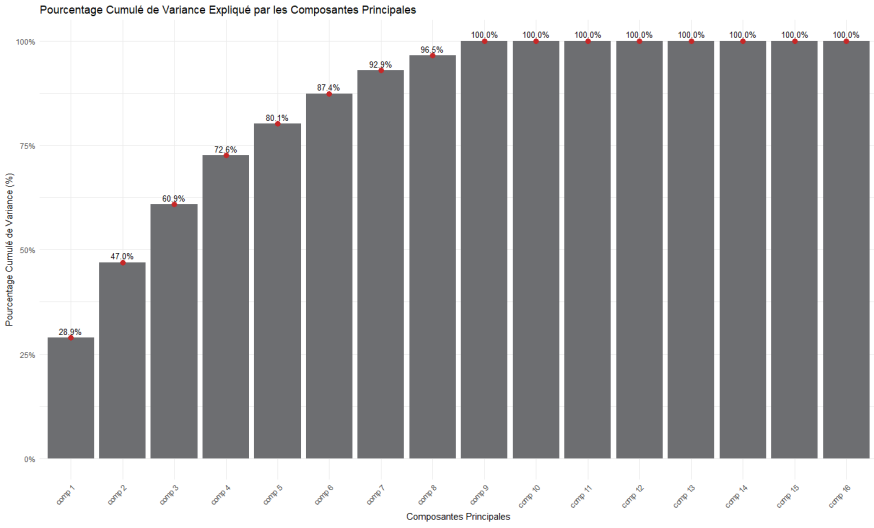

FIGURE 3.15 – Pourcentage cumulé de la variance expliquée par les composantes principales.

Ainsi, nous concentrerons notre analyse sur ces 5 premières dimensions, qui capturent l'essentiel de l'information contenue dans les données. Cette approche permet de simplifier le modèle tout en conservant la majorité de la variance, facilitant ainsi l'interprétation et la visualisation des résultats.

Interprétation des Coefficients de Chargement

Les coefficients de chargement (*loadings*) révèlent la contribution de chaque variable d'origine à chaque composante principale. Ces coefficients, extraits de la matrice **W**, permettent d'identifier quelles composantes principales capturent le mieux la variance associée à chaque variable.

Le tableau ci-dessous illustre les coefficients de chargement pour les cinq premières composantes principales dans notre analyse :

Variable	**Dim.1**	**Dim.2**	**Dim.3**	**Dim.4**	**Dim.5**
Solde_Reg	33.96	4.58	0.10	14.08	0.05
Eval_Reg	33.86	2.86	0.07	6.77	0.03
Solde_Eval_nette	32.05	18.52	0.01	1.34	0.00
Solde_Rec_Flux	0.03	0.12	37.43	0.07	12.05
Solde_Reg_Flux	0.03	0.12	37.44	0.05	12.06
Eval_Rec	0.02	35.93	0.13	2.09	0.01
Solde_Rec	0.02	37.14	0.21	0.66	0.09
Solde_Rec_Flux_2	0.01	0.09	10.60	0.04	27.04
Solde_Reg_Flux_2	0.01	0.07	10.61	0.10	26.99
Solde_Rec_Flux_3	0.00	0.04	1.66	0.12	10.49
Solde_Reg_Flux_3	0.00	0.03	1.70	0.20	10.58
delai_ModifOuverture	0.00	0.16	0.00	0.53	0.00
delai_SituationDerniereModif	0.00	0.02	0.03	0.42	0.12
provision	0.00	0.12	0.00	32.75	0.32
Diff_Rec	0.00	0.01	0.01	0.40	0.03
Diff_Reg	0.00	0.20	0.00	40.37	0.14

TABLE 3.17 – Résultats de l'ACP pour les variables

Les résultats dévoilent la contribution des variables aux dimensions principales. Les variables avec des valeurs élevées sur une composante principale signalent que cette composante est essentielle pour expliquer la variation de ces variables.

— **Dim.1 :** Les variables telles que *Solde_Reg* et *Eval_Reg* présentent des contributions marquées à la première composante, indiquant que cette composante explique une part substantielle de la variance dans les données.
— **Dim.2 :** Les variables *Solde_Rec_Flux* et *Eval_Rec* montrent des contributions significatives, suggérant que la deuxième composante est fortement influencée par ces variables.
— **Dim.3 à Dim.5 :** Les contributions varient davantage, mettant en évidence des influences spécifiques de certaines variables, telles que *Solde_Eval_nette* dans les dimensions ultérieures.

Les coefficients de chargement indiquent que les variables *Solde_Reg* et *Eval_Reg* ont des valeurs élevées pour la première composante principale (*Dim.1*), suggérant que cette composante capte une grande partie de la variance associée à ces deux variables. De même, *Provision* contribue de manière significative à la troisième composante principale (*Dim.3*), avec un coefficient de chargement de 0.660.

Ces observations suggèrent que les composantes principales peuvent remplacer certaines variables d'origine dans un modèle de machine learning, en capturant l'essentiel de l'information de manière plus concise. Par exemple, plutôt que d'utiliser *Solde_Reg* et *Eval_Reg* séparément, il serait plus judicieux d'utiliser *Dim.1* pour représenter ces deux variables de manière combinée.

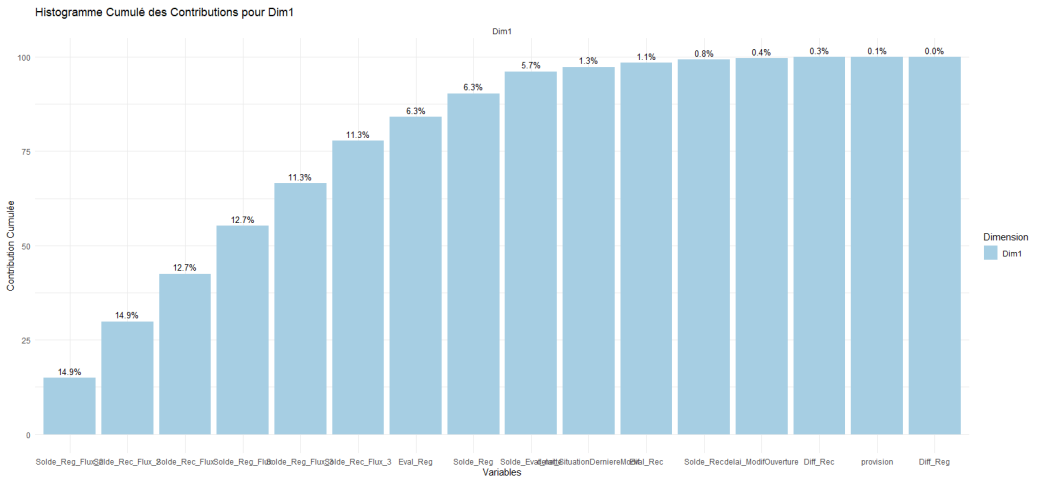

FIGURE 3.16 – Contribution des variables sur la Dimension 1

Coordonnées des Variables

Les coordonnées des variables dans les dimensions principales jouent un rôle crucial dans l'interprétation des résultats de l'Analyse en Composantes Principales (ACP). Elles montrent comment chaque variable se projette dans l'espace des composantes principales, fournissant ainsi des indications sur leur contribution relative à chaque dimension. Ces coordonnées permettent de comprendre quelles variables sont les plus influentes dans chaque dimension principale et comment elles sont corrélées entre elles.

Pour visualiser ces coordonnées, nous avons généré un graphique montrant la projection des variables sur les deux dimensions principales :

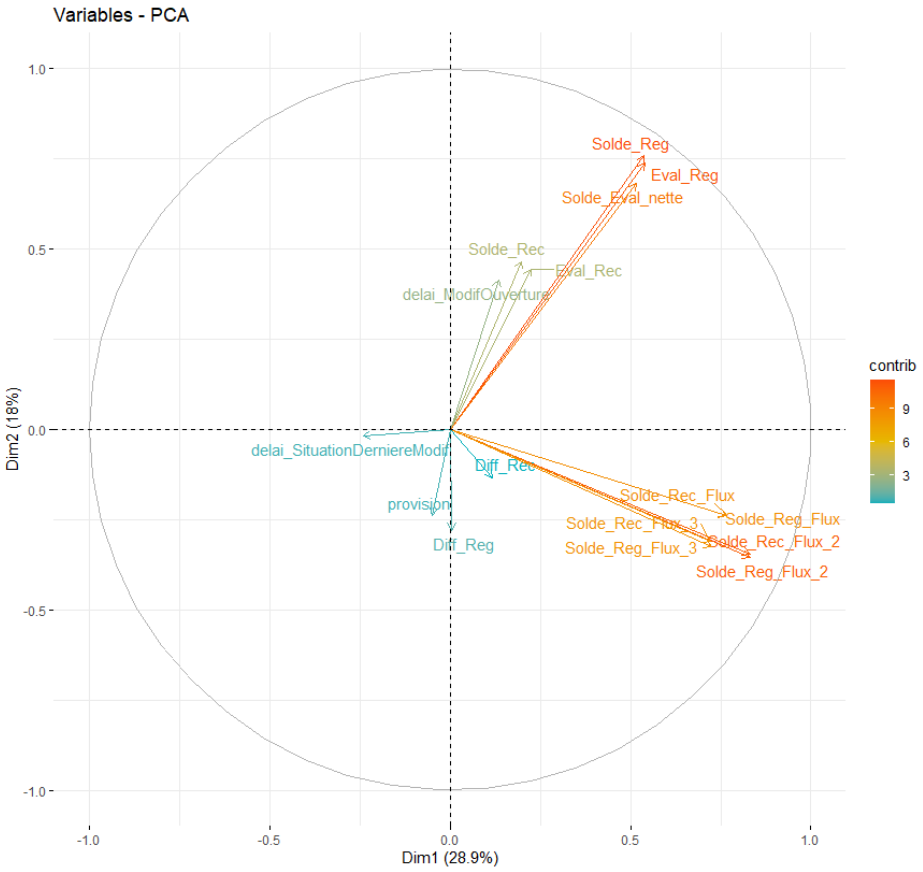

FIGURE 3.17 – Projection des variables sur les deux dimensions principales

Le tableau ci-dessous présente les coefficients de chargement pour les cinq premières composantes principales dans notre analyse :

Variable	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Solde_Reg	0.538	0.69	0.302	-0.190	-0.024
Eval_Reg	0.540	0.739	0.365	-0.151	-0.014
Solde_Rec	0.196	0.466	-0.417	0.719	0.107
Eval_Rec	0.223	0.445	-0.338	0.748	0.020
Solde_Eval_nette	0.515	0.682	0.429	-0.277	-0.018
Provision	-0.049	-0.238	0.660	0.494	0.328
Delai_ModifOuverture	0.135	0.415	-0.458	0.093	0.028
Delai_SituationDerniereModif	-0.241	-0.017	-0.211	-0.323	0.302
Solde_Rec_Flux	0.766	-0.237	-0.071	-0.118	0.462
Solde_Reg_Flux_2	0.831	-0.353	-0.110	-0.012	0.031
Solde_Rec_Flux_2	0.831	-0.345	-0.127	-0.019	0.038
Solde_Reg_Flux_3	0.724	-0.326	-0.075	0.105	-0.436
Solde_Rec_Flux_3	0.722	-0.311	-0.102	0.096	-0.428
Solde_Reg_Flux	0.765	-0.238	-0.068	-0.116	0.461
Diff_Reg	0.004	-0.279	0.788	0.500	0.126
Diff_Rec	0.117	-0.133	0.404	0.089	-0.424

TABLE 3.18 – Coefficients de chargement des variables sur les cinq premières composantes principales.

Les coordonnées des variables dans les dimensions principales révèlent deux aspects essentiels :

— **Magnitude des Coordonnées :** Les valeurs élevées dans une dimension principale indiquent qu'une variable est fortement alignée avec cette dimension. Par exemple, la variable `Solde_Rec_Flux` présente une coordonnée élevée dans Dim.1 (0.766), suggérant une forte représentation dans cette dimension.
— **Direction :** La direction (positive ou négative) de la coordonnée montre l'orientation de la variable par rapport à la dimension. Par exemple, `Diff_Reg` a une coordonnée élevée et positive dans Dim.3 (0.788), indiquant que Dim.3 capture bien l'information relative à `Diff_Reg`.

Les coordonnées des variables dans les dimensions principales correspondent aux valeurs des coefficients de chargements. L'analyse des coordonnées révèle plusieurs observations intéressantes :

— **Dimension 1 (Dim.1) :** Les variables `Solde_Reg` et `Eval_Reg` affichent des coordonnées élevées dans Dim.1 (0.538 et 0.540, respecti-

vement), suggérant une bonne représentation dans cette dimension. Cela indique que Dim.1 capte principalement l'information liée aux soldes et évaluations financières.

- **Dimension 2 (Dim.2) :** Les variables `Solde_Reg` et `Eval_Reg` ont également des valeurs élevées dans Dim.2 (0.69 et 0.739, respectivement), montrant qu'elles contribuent de manière significative à cette dimension, bien que moins fortement que Dim.1.
- **Dimension 3 (Dim.3) :** La variable `provision` a une coordonnée élevée dans Dim.3 (0.660) mais une coordonnée négative dans Dim.1 (-0.049). Cela indique que `provision` est une contribution significative à Dim.3, ce qui suggère que cette dimension est liée aux aspects spécifiques des provisions, tandis que Dim.1 reste plus général.
- **Dimension 4 (Dim.4) :** Les variables `Solde_Rec` et `Eval_Rec` présentent des coordonnées positives dans Dim.4 (0.719 et 0.748, respectivement), suggérant que Dim.4 capture des aspects des soldes et évaluations reçus.
- **Variables Spécifiques :** Les variables liées aux flux, telles que `Solde_Rec_Flux` et `Solde_Reg_Flux_2`, montrent des coordonnées élevées dans Dim.1, soulignant leur forte contribution à cette dimension. À l'inverse, des variables comme `delai_SituationDerniereModif` présentent des coordonnées faibles dans toutes les dimensions, signalant une contribution plus modeste.

Ces résultats permettent de discerner quelles variables sont les plus influentes dans chaque dimension principale et comment elles interagissent entre elles. Cette compréhension est essentielle pour interpréter les composantes principales et guider les décisions concernant la simplification et l'analyse ultérieure des données.

L'Analyse des Cosinus Carrés

L'Analyse des Cosinus Carrés ($\cos^2$) constitue un outil précieux pour évaluer la qualité de la représentation des variables dans les dimensions principales d'une Analyse en Composantes Principales (ACP). Les valeurs de $\cos^2$ mesurent combien chaque variable est capturée par chaque dimension principale, offrant ainsi une vue d'ensemble de la pertinence des variables par rapport aux dimensions extraites.

Le tableau suivant présente les valeurs de $\cos^2$ pour chaque variable à travers les différentes dimensions principales :

Variable	Dim.1	Dim.2	Dim.3	Dim.4
Solde_Reg	0.2896	0.5760	0.0911	0.0362
Eval_Reg	0.2914	0.5457	0.1335	0.0228
Solde_Rec	0.0385	0.2176	0.1740	0.5170
Eval_Rec	0.0500	0.1979	0.1142	0.5588
Solde_Eval_nette	0.2654	0.4651	0.1841	0.0769
Provision	0.0024	0.0568	0.4353	0.2445
Delai_ModifOuverture	0.0181	0.1726	0.2097	0.0087
Delai_SituationDerniereModif	0.0580	0.0003	0.0446	0.1043
Solde_Rec_Flux	0.5861	0.0561	0.0050	0.0139
Solde_Reg_Flux_2	0.6911	0.1249	0.0120	0.0001
Solde_Rec_Flux_2	0.6908	0.1187	0.0160	0.0004
Solde_Reg_Flux_3	0.5248	0.1062	0.0056	0.0111
Solde_Rec_Flux_3	0.5216	0.0966	0.0104	0.0092
Diff_Reg	0.0000	0.0780	0.6214	0.2498
Diff_Rec	0.0136	0.0178	0.1635	0.0079

TABLE 3.19 – Tableau des $\cos^2$ des variables dans les dimensions de l'ACP.

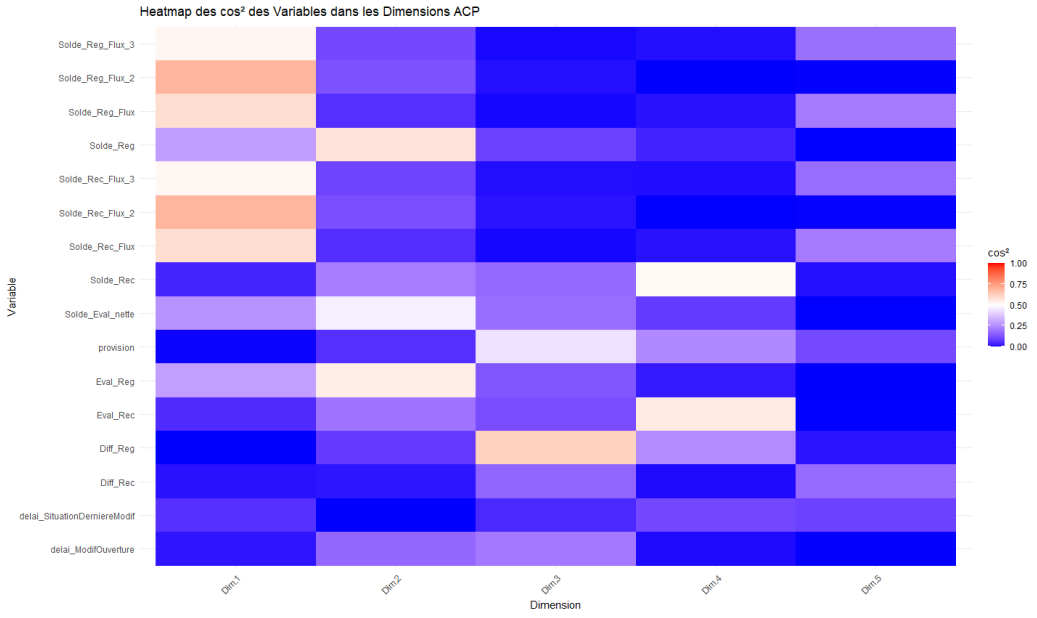

FIGURE 3.18 – Heatmap des valeurs de $\cos^2$ des variables pour les cinq premières dimensions

L'analyse des valeurs de $\cos^2$ fournit des informations clés sur la repré-

sentation des variables dans les dimensions principales de l'ACP.

Les variables avec des valeurs élevées de `cos²` dans la Dimension 1 sont :

— **Solde_Reg** : Avec un `cos²` de 0.2896, cette variable est fortement représentée dans Dim.1, soulignant sa contribution significative à cette dimension.
— **Eval_Reg** : Son `cos²` de 0.2914 indique également une forte présence dans Dim.1, confirmant son importance pour cette dimension.
— **Solde_Rec_Flux_2** : Avec un `cos²` de 0.6908, cette variable est particulièrement bien représentée dans Dim.1, marquant une contribution très significative.

Ces variables sont essentielles pour une interprétation approfondie de la Dimension 1 et doivent être considérées comme des facteurs majeurs expliquant cette dimension.

Les variables suivantes montrent une forte représentation dans la Dimension 3 :

— **Provision** : Avec un `cos²` de 0.4353, cette variable est fortement associée à Dim.3, indiquant une bonne capture des aspects relatifs à `provision`.
— **Diff_Reg** : Affichant un `cos²` de 0.6214 dans Dim.3, cette variable montre une très bonne représentation, soulignant son rôle significatif dans cette dimension.

Ces variables sont cruciales pour comprendre les caractéristiques de la Dimension 3 et méritent une attention particulière dans l'analyse de cette dimension.

Les variables qui se distinguent dans la Dimension 4 sont :

— **Solde_Rec** : Avec un `cos²` de 0.5170, `Solde_Rec` est bien représentée dans Dim.4, indiquant une contribution notable à cette dimension.
— **Eval_Rec** : Possédant un `cos²` de 0.5588 dans Dim.4, cette variable montre une bonne représentation, confirmant son importance pour cette dimension.

Ces variables jouent un rôle important dans l'explication de la Dimension 4 et doivent être prises en compte pour une analyse approfondie de cette

dimension.

Certaines variables montrent une répartition plus équilibrée de leurs valeurs de $\cos^2$ à travers plusieurs dimensions :

— **Solde_Eval_nette** : Présente des $\cos^2$ relativement équilibrés dans Dim.1 (0.2654), Dim.2 (0.4651), et Dim.3 (0.1841), suggérant une contribution polyvalente à plusieurs dimensions.
— **Delai_ModifOuverture** : Affiche des valeurs de $\cos^2$ réparties entre Dim.2 (0.1726) et Dim.3 (0.2097), indiquant une présence modérée dans plusieurs dimensions.

Ces variables méritent une attention particulière pour comprendre leur contribution à plusieurs dimensions principales.

Ainsi, nous obtenons cinq nouvelles variables quantitatives — Dim.1, Dim.2, Dim.3, Dim.4, et Dim.5 — qui s'avéreront certainement pertinentes pour les analyses futures.

3.13 Analyse des Variables Qualitatives avec Histogrammes

L'analyse des variables qualitatives est cruciale pour déchiffrer l'influence de ces variables sur la clôture des dossiers sinistres. Nous combinons ici les résultats des tests statistiques — notamment le V de Cramer et le test du chi^2 — avec une analyse visuelle à travers des histogrammes pour évaluer la relation entre chaque variable qualitative et la variable cible, « position » des dossiers (clôturé ou non).

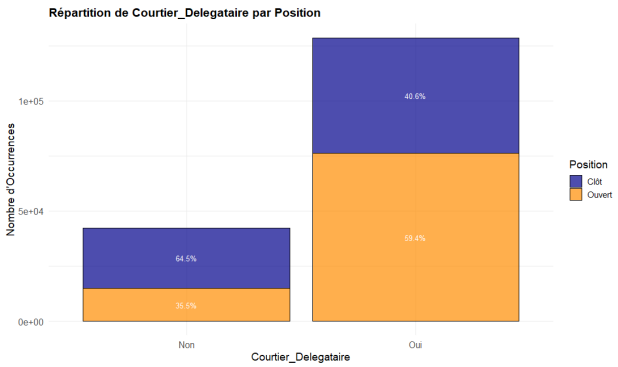

FIGURE 3.19 – Répartition de la variable *Courtier Délégataire* en fonction de la position des dossiers

Courtier Délégataire La variable « Courtier Délégataire » révèle une distinction significative : 65% des dossiers sans délégation courtier sont clôturés, contre 40% pour ceux avec délégation. Cette observation suggère que les dossiers traités en interne pourraient bénéficier d'une gestion plus efficace ou concerner des cas moins complexes.

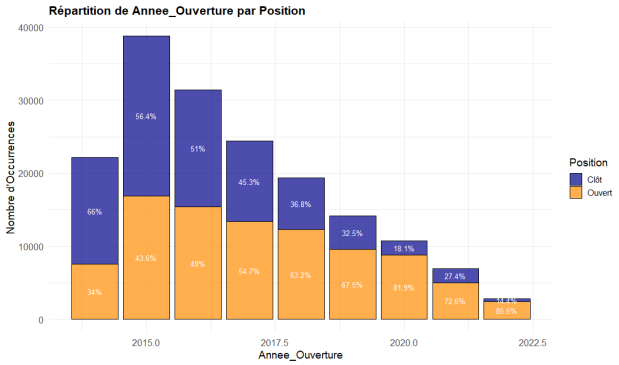

FIGURE 3.20 – Répartition de la variable *Année d'Ouverture* en fonction de la position des dossiers

Année d'Ouverture L'analyse de l'« Année d'Ouverture » montre une tendance intéressante : les dossiers ouverts en 2014 ont un taux de clôture de 66%, tandis que ce taux diminue pour atteindre 45,3% en 2017 et 18,1% en 2020, avec un léger pic à 27,4% en 2021. Les dossiers plus anciens ont eu davantage de temps pour être clôturés, alors que les dossiers plus récents nécessitent encore un suivi.

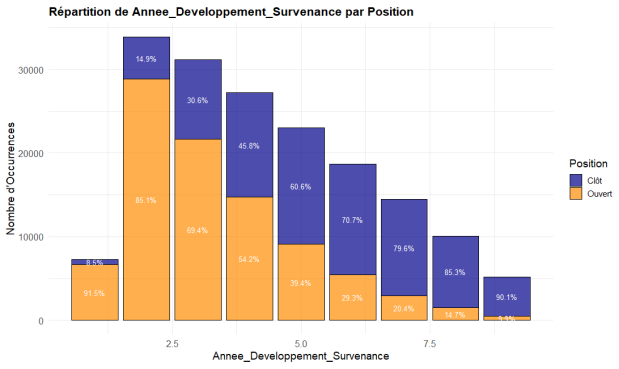

FIGURE 3.21 – Répartition de la variable *Année de Développement de Survenance* en fonction de la position des dossiers

Année de Développement de Survenance L'« Année de Développement de Survenance » suit une tendance similaire à l'« Année d'Ouverture ». Les dossiers plus anciens ont un taux de clôture plus élevé, ce qui reflète la dynamique où les dossiers anciens ont eu plus d'opportunités pour être traités et clôturés.

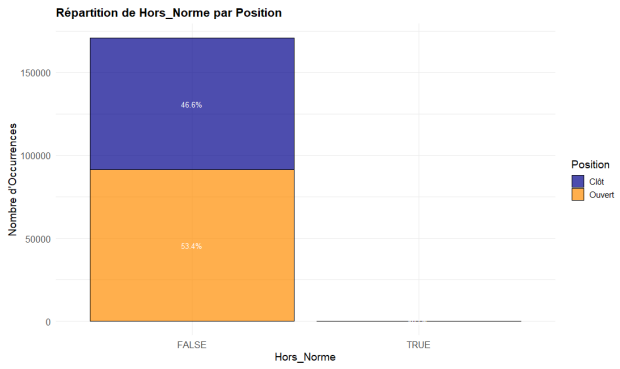

FIGURE 3.22 – Répartition de la variable *Dossier Hors Norme* en fonction de la position des dossiers

Dossier Hors Norme La variable « Dossier Hors Norme » montre une faible occurrence de valeurs « TRUE », rendant l'analyse statistique moins significative. Ces dossiers ne montrent pas de tendance claire à être clôturés plus ou moins souvent que les autres. La faible occurrence suggère que cette variable pourrait être moins pertinente pour les modèles de machine learning.

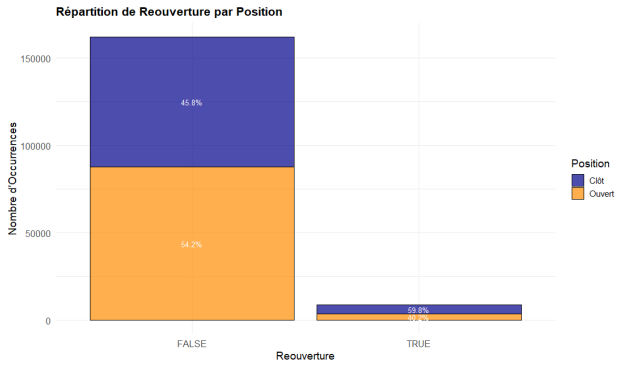

FIGURE 3.23 – Répartition de la variable *Réouverture du Dossier* en fonction de la position des dossiers

Réouverture du Dossier La variable « Réouverture du Dossier » indique que 60% des dossiers réouverts sont clôturés, contre 45,8% pour ceux qui n'ont jamais été réouverts. Cela suggère que la réouverture des dossiers favorise légèrement leur clôture, possiblement en raison d'une attention accrue après réouverture.

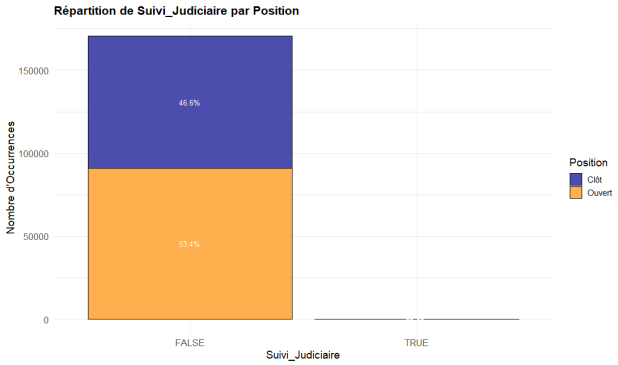

FIGURE 3.24 – Répartition de la variable *Suivi Judiciaire* en fonction de la position des dossiers

Suivi Judiciaire Le « Suivi Judiciaire » montre une très faible occurrence de dossiers clôturés. Cela pourrait refléter la complexité prolongée des procédures judiciaires qui retarde la clôture des dossiers concernés. Cette variable pourrait être indicative de la complexité des cas.

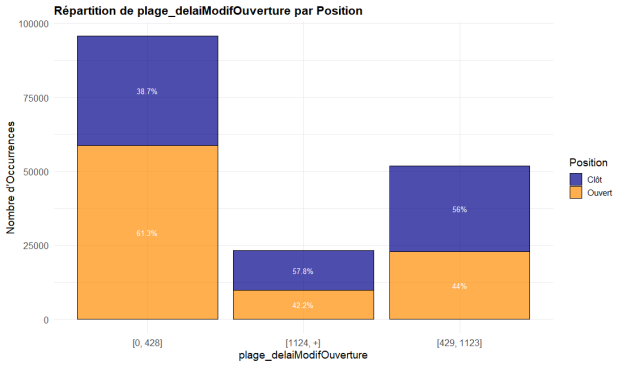

FIGURE 3.25 – Répartition de la variable *Plage Délais Modifications-Ouverture* en fonction de la position des dossiers

Plage Délais Modifications-Ouverture La variable « Plage Délais Modifications-Ouverture » indique une tendance vers la clôture lorsque les délais entre l'ouverture et la dernière modification sont plus longs. Cela suggère un lien potentiel entre le temps passé et la finalisation des dossiers.

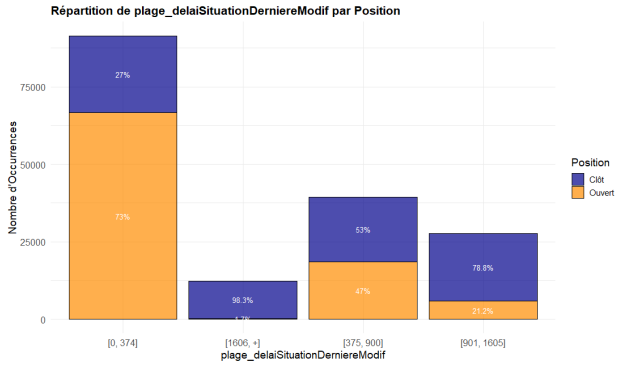

FIGURE 3.26 – Répartition de la variable *Plage Délais Survenance-Dernière Modification* en fonction de la position des dossiers

Plage Délais Survenance-Dernière Modification La « Plage Délais Survenance-Dernière Modification » montre un taux de clôture plus élevé pour les dossiers avec une durée entre 375 et 900 jours (53%) et encore plus élevé pour ceux entre 901 et 1400 jours (78,8%). Cela suggère qu'un suivi prolongé peut favoriser la clôture des dossiers.

Analyse Globale des Résultats Les analyses des variables qualitatives offrent des insights précieux pour la gestion des dossiers sinistres :

- **Temps comme Facteur de Clôture :** Les variables temporelles — année d'ouverture, année de développement de survenance, et plages de délais — montrent que les dossiers avec une durée de traitement plus longue ont une probabilité accrue d'être clôturés. Cela pourrait refléter le caractère plus complexe et prolongé de ces dossiers.
- **Complexité des Dossiers :** Les variables telles que « Courtier Délégataire » et « Suivi Judiciaire » révèlent que la complexité, qu'elle soit administrative ou légale, tend à diminuer le taux de clôture, soulignant l'importance d'une gestion efficace.
- **Réouverture des Dossiers :** La réouverture des dossiers semble offrir une opportunité supplémentaire pour leur clôture, même si le nombre de dossiers réouverts est limité.

3.14 V de Cramer et Test du Chi²

Le V de Cramer (V) et le test du chi² (χ^2) sont des outils statistiques essentiels pour évaluer l'association entre deux variables qualitatives. Tandis que le V de Cramer mesure l'intensité de cette association, le test du chi² permet de déterminer si cette association est statistiquement significative.

Définition et Calcul du V de Cramer Le V de Cramer est dérivé de la statistique du chi² et est calculé selon la formule suivante :

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

où :
— χ^2 représente la statistique du chi² obtenue,
— n est le nombre total d'observations,
— k est le nombre de colonnes dans le tableau de contingence,
— r est le nombre de lignes dans le tableau de contingence.

Le V de Cramer varie entre 0 et 1, où 0 indique une absence d'association et 1 une association parfaite. Une valeur élevée de V suggère une relation forte entre les variables analysées.

Interprétation du Test du Chi² Le test du chi² évalue l'hypothèse nulle selon laquelle les variables sont indépendantes. Une p-value inférieure à un seuil conventionnel (souvent 0,05) indique que l'association observée entre les variables n'est probablement pas due au hasard, suggérant une relation statistiquement significative.

Résultats des Tests Statistiques

Les résultats obtenus pour chaque variable qualitative par rapport à la variable cible sont résumés dans la figure ci-dessous, ainsi que dans le tableau 3.14, qui présente les valeurs du V de Cramer et les p-values associées.

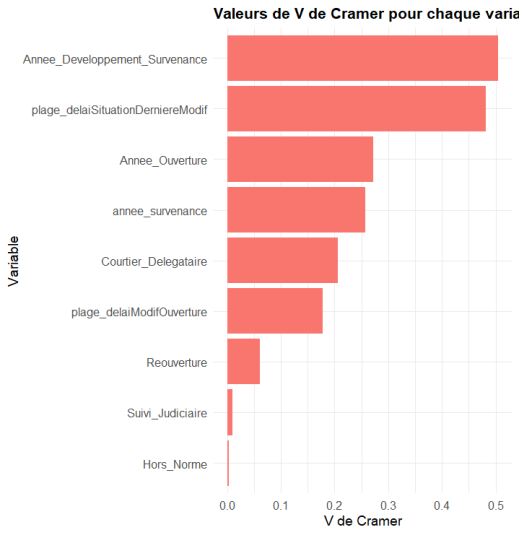

FIGURE 3.27 – V de Cramer entre la variable cible et les variables qualitatives

Variable	**V de Cramer**	**P-value**
annee_survenance	0.2567	0.0000
Courtier_Delegataire	0.2068	0.0000
Suivi_Judiciaire	0.0101	0.0000
Reouverture	0.0614	0.0000
Annee_Developpement_Survenance	0.5052	0.0000
Hors_Norme	0.0026	0.3344
Annee_Ouverture	0.2726	0.0000
plage_delaiSituationDerniereModif	0.4813	0.0000
plage_delaiModifOuverture	0.1779	0.0000

TABLE 3.20 – Tableau des V de Cramer et p-values (test du chi²) entre chaque variable qualitative et la variable cible

Les résultats des tests statistiques mettent en lumière l'importance de certaines variables qualitatives dans la clôture des dossiers sinistres. Parmi celles-ci, les variables temporelles se distinguent particulièrement. L'**année de développement depuis la survenance** et la **plage du délai entre la situation et la dernière modification** présentent des V de Cramer élevés (respectivement 0,5052 et 0,4813), témoignant d'une forte association avec la variable cible. Ces résultats corroborent les tendances observées dans les histogrammes précédents, où les dossiers plus anciens ou ayant un suivi prolongé montrent des taux de clôture plus élevés.

Année de Survenance et Année d'Ouverture Les V de Cramer pour les variables **année de survenance** et **année d'ouverture** sont modérés (0,2567 et 0,2726), mais indiquent une association significative avec la clôture des dossiers. Cette observation est en phase avec l'analyse des histogrammes, où les dossiers ouverts plus tôt (comme en 2014) avaient un taux de clôture plus élevé, probablement en raison du temps disponible pour leur traitement. Ainsi, ces variables restent pertinentes pour la modélisation prédictive, car elles capturent des éléments temporels cruciaux.

Courtier Délégataire et Suivi Judiciaire La variable **Courtier_Delegataire** présente un V de Cramer de 0,2068, signifiant une association modérée mais significative avec la variable cible. L'histogramme montre que les dossiers non délégués au courtier sont plus souvent clôturés, suggérant une efficacité potentiellement réduite pour les dossiers délégués. La variable **Suivi_Judiciaire**, bien que statistiquement significative, affiche un V de Cramer très faible (0,0101), ce qui, associé aux observations précédentes, indique que l'influence directe du suivi judiciaire sur la clôture est minime et peut nécessiter d'autres variables explicatives pour une compréhension complète.

Réouverture et Hors Norme Les variables **Réouverture** et **Hors_Norme** affichent des V de Cramer faibles, 0,0614 et 0,0026 respectivement. Malgré une tendance à la clôture des dossiers réouverts, cette variable présente un V de Cramer faible, suggérant un rôle secondaire dans la prédiction. La variable **Hors_Norme**, avec un V de Cramer très faible et une p-value non significative, semble avoir une influence négligeable sur la clôture des dossiers. Elle pourrait donc être exclue des modèles prédictifs pour éviter l'introduction de bruit inutile.

Implications pour la Modélisation L'analyse combinée des histogrammes et des résultats du V de Cramer et du chi² permet de mettre en évidence les variables qualitatives les plus pertinentes pour la modélisation prédictive. Les variables avec des V de Cramer élevés et des p-values significatives, telles que l'**année de développement depuis la survenance** et la **plage du délai entre la situation et la dernière modification**, doivent être priorisées dans la construction des modèles de machine learning. En capturant des dynamiques temporelles cruciales, ces variables sont susceptibles d'améliorer la précision et la spécificité des prévisions de clôture des dossiers sinistres.

En revanche, les variables comme **Hors_Norme**, qui ne montrent pas d'association significative, peuvent être omises sans compromettre la qualité

des prédictions. Cette approche rationalise le processus de sélection des variables et contribue à la création d'un modèle plus performant et robuste.

En conclusion, les analyses statistiques soulignent l'importance de se concentrer sur les variables qui capturent efficacement les aspects temporels et structurels des dossiers sinistres. Ces insights, enrichis par les résultats des tests statistiques, guideront la prochaine phase de modélisation.

3.15 Analyse de la Dépendance entre Variables Qualitatives

Pour explorer la force des associations entre les variables qualitatives, nous avons calculé les valeurs du V de Cramer pour chaque paire de variables et les avons visualisées sous forme de heatmap, comme le montre la figure **??**. Cette approche nous permet de saisir les nuances des relations entre les variables de manière intuitive et détaillée.

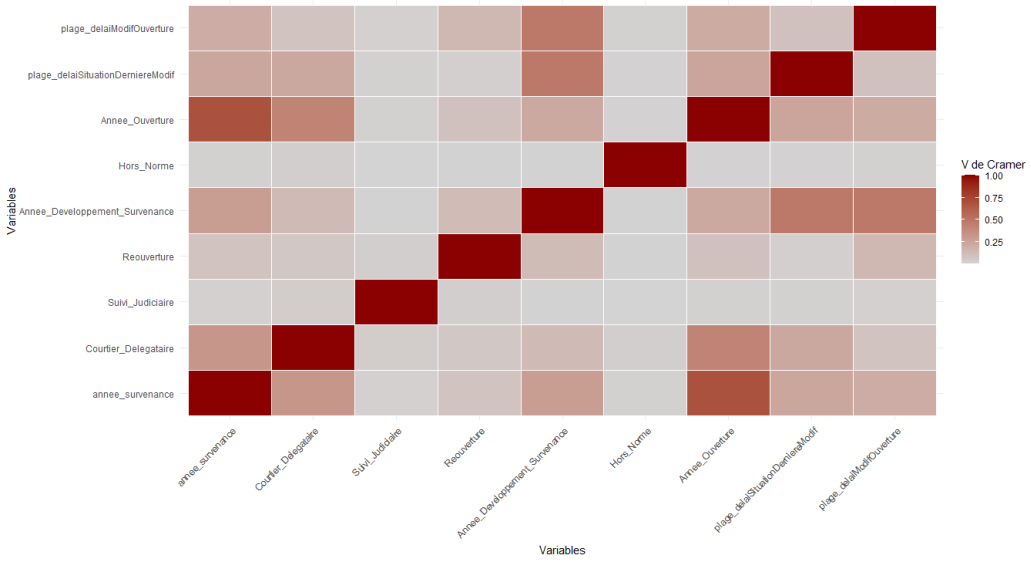

FIGURE 3.28 – Heatmap des valeurs du V de Cramer entre les variables qualitatives

La heatmap révèle une diversité de relations entre les variables qualitatives, chacune portant un poids différent dans l'analyse globale. Les valeurs du V de Cramer, qui varient entre 0 et 1, offrent un aperçu précieux de ces

relations :

- **Faibles Associations :** La majorité des paires de variables présentent des valeurs de V de Cramer relativement modestes, indiquant une association faible. Par exemple, la relation entre *Suivi_ Judiciaire* et *annee_ survenance* affiche une valeur de 0,017, suggérant une absence quasi totale de corrélation. Cette faible association confirme l'observation selon laquelle le suivi judiciaire n'influence pas de manière significative l'année de survenance des sinistres.
- **Associations Modérées à Fortes :** Certaines associations se révèlent plus prononcées. Notamment, la relation entre *Annee_ Ouverture* et *annee_ survenance* se distingue par une valeur de V de Cramer de 0,687, indiquant une forte association. Cette forte corrélation est intuitive, car l'année d'ouverture est directement influencée par l'année de survenance des sinistres, révélant ainsi une dépendance temporelle évidente entre ces deux variables.
- **Associations Maximales :** Comme attendu, les paires de variables identiques, telles que *annee_ survenance* comparée à elle-même, affichent une valeur maximale de V de Cramer (1,000). Ces valeurs maximales confirment l'absence de variabilité entre des variables identiques et soulignent l'importance d'éviter les redondances dans notre analyse.

L'analyse de la heatmap met en lumière des aspects essentiels de la dépendance entre variables qualitatives. Bien que de nombreuses associations apparaissent faibles, certaines paires révèlent des relations significatives, suggérant des interactions importantes entre les variables :

- **Associations Temporelles Significatives :** Les associations fortes observées, telles que celle entre *Annee_ Ouverture* et *annee_ survenance*, indiquent que ces variables capturent des dimensions temporelles importantes. Leur influence directe sur les résultats suggère qu'elles doivent être intégrées avec soin dans les modèles de prédiction pour améliorer leur précision.
- **Évaluation des Variables Faiblement Associées :** Les variables présentant des valeurs faibles de V de Cramer, telles que *Suivi_ Judiciaire*, pourraient nécessiter une réévaluation. Leur faible association avec d'autres variables pourrait indiquer un impact limité ou une interaction moins significative. Cette observation devrait guider la sélection des variables pour les étapes suivantes de modélisation.

Les insights tirés de cette analyse permettent de focaliser notre attention sur les variables les plus influentes pour la modélisation prédictive. En concentrant les efforts sur les variables avec des associations significatives, telles que celles révélées par les valeurs élevées du V de Cramer, nous pouvons renforcer la robustesse et la précision des modèles de machine learning. Les relations moins marquées, quant à elles, peuvent être examinées de manière plus approfondie pour déterminer leur contribution réelle et potentiellement exclues si leur apport est jugé négligeable.

En conclusion, cette analyse détaillée de la dépendance entre variables qualitatives, enrichie par la visualisation élégante de la heatmap, offre des orientations claires pour l'optimisation des modèles prédictifs. En intégrant les variables pertinentes et en excluant les moins influentes, nous visons à créer des modèles plus performants, capables de capturer avec finesse les dynamiques sous-jacentes aux dossiers sinistres.

Chapitre 4

Modélisation Automatique de Clôture des Dossiers

Dans cette partie, nous explorons l'application de méthodes de machine learning pour prédire la clôture des dossiers sinistres à partir d'un jeu de données d'apprentissage. Nous implémentons trois modèles : la régression logistique avec régularisation Elastic Net, l'arbre de décision et la forêt aléatoire. L'objectif est d'identifier le modèle qui fournit les meilleures performances pour cette tâche de classification.

Voici le code LaTeX avec les variables affichées deux à deux et les corrections de fautes de frappe :

"'latex

4.1 Présentation des Données

Présentation des Variables

Les variables explicatives sont réparties en deux catégories principales : quantitatives et qualitatives.

Les variables quantitatives représentent des mesures financières et temporelles essentielles pour prédire la clôture des dossiers sinistres. Elles incluent notamment :

- `Eval_Rec`, `Eval_Reg` : Ces variables financières fournissent des informations sur les évaluations des dossiers.
- `Solde_Eval_nette`, `Solde_Reg` : Informations sur les soldes des dossiers.

— `Solde_Rec`, `Solde_Reg_Flux_2` : Soldes financiers et leurs déclinaisons sur plusieurs périodes.
— `Solde_Rec_Flux_2`, `Solde_Reg_Flux_3` : Soldes financiers supplémentaires.
— `Solde_Rec_Flux_3`, `provision` : Indicateur des provisions financières des dossiers.
— `delai_ModifOuverture`, `delai_SituationDerniereModif` : Mesures temporelles représentant les délais entre différentes étapes du traitement des dossiers.
— `Dim_1`, `Dim_2` : Dimensions additionnelles.
— `Dim_3`, `Dim_4` : Dimensions additionnelles continues.
— `Dim_5`, `Diff_Reg` : Dimensions additionnelles et différences de soldes.
— `Diff_Rec` : Autre différence de soldes.

Ces variables jouent un rôle crucial pour capturer la dynamique des flux financiers et temporels dans le processus de clôture des dossiers.

Les variables qualitatives capturent des informations catégorielles, telles que :

— `Courtier_Delegataire`, `Suivi_Judiciaire` : Indiquent si le dossier est géré par un courtier ou un délégataire, et si un suivi judiciaire est en cours.
— `Reouverture`, `Annee_Developpement_Survenance` : Indiquent si le dossier a été réouvert et les années associées aux événements de développement.
— `Annee_Ouverture`, `plage_delaiSituationDerniereModif` : Indicateurs temporels associés aux dates d'ouverture et à la plage des délais.
— `plage_delaiModifOuverture` : Indicateur de la plage de délais pour la modification d'ouverture.

Ces variables nécessitent une transformation en variables factices (*dummies*) pour être utilisées efficacement dans les modèles prédictifs.

Transformation et Combinaison des Variables

Après avoir sélectionné les variables quantitatives et qualitatives, nous transformons ces dernières en variables factices pour leur traitement dans les modèles. Ensuite, nous combinons les deux types de variables afin de former un jeu de données complet.

```
% Préparation des variables quantitatives
```

```
X_quant_tree <- dataset[, c("Eval_Rec", "Eval_Reg",
                            "Solde_Eval_nette", "Solde_Reg",
                            "Solde_Rec", "Solde_Reg_Flux_2",
                            "Solde_Rec_Flux_2", "Solde_Reg_Flux_3",
                            "Solde_Rec_Flux_3", "provision",
                            "delai_ModifOuverture", "delai_SituationDerniereModi
                            "Dim_1", "Dim_2", "Dim_3", "Dim_4",
                            "Dim_5", "Diff_Reg", "Diff_Rec")]

% Transformation des variables qualitatives en dummies
X_qual_dummy_tree <- model.matrix(~ Courtier_Delegataire + Suivi_Judiciaire +
                                    Reouverture + Hors_Norme +
                                    Annee_Developpement_Survenance +
                                    Annee_Ouverture + annee_survenance +
                                    plage_delaiSituationDerniereModif +
                                    plage_delaiModifOuverture - 1,
                                    data = dataset_app_LI)

% Combinaison des variables quantitatives et qualitatives factices
X_tree <- cbind(X_quant_tree, X_qual_dummy_tree)
```

Validation Croisée

Une validation rigoureuse des modèles est nécessaire pour éviter le surapprentissage (*overfitting*) et obtenir des prédictions généralisables. Pour cela, nous avons utilisé une validation croisée à 5 sous-ensembles (*5-fold cross-validation*), une technique éprouvée qui consiste à diviser le jeu de données en cinq parties égales.

À chaque itération, le modèle est entraîné sur quatre sous-ensembles et testé sur le sous-ensemble restant. Ce processus est répété cinq fois, garantissant une évaluation robuste de la performance du modèle. La validation croisée permet d'obtenir une estimation fiable de l'erreur de généralisation et évite les biais liés à une simple division en échantillons d'entraînement et de test.

Métriques d'Évaluation

Les performances des modèles sont évaluées à l'aide de plusieurs métriques importantes pour cette problématique de classification binaire (clôturé/non

clôturé). Voici les principales métriques utilisées :

— **Sensibilité (Recall)** : Mesure la proportion de dossiers clôturés correctement identifiés par le modèle (vrais positifs).
— **Spécificité (Specificity)** : Mesure la proportion de dossiers ouverts correctement identifiés (vrais négatifs).
— **AUC (Area Under the ROC Curve)** : Évalue la capacité globale du modèle à discriminer entre les deux classes. Un score AUC élevé indique une meilleure séparation entre les dossiers clôturés et non clôturés.
— **Erreur OOB (Out-Of-Bag Error)** : Mesure l'erreur sur les échantillons non utilisés dans l'entraînement, utile pour estimer l'erreur de généralisation dans les forêts aléatoires.

Les métriques additionnelles seront :

— **Précision (Precision)** : Mesure la proportion de prédictions positives correctes parmi l'ensemble des prédictions positives effectuées.
— **F1-Score** : Combinaison de la précision et du rappel, utile dans les cas où les classes sont déséquilibrées.
— **Balanced Accuracy** : Moyenne de la sensibilité et de la spécificité, offrant une évaluation équilibrée des performances.

Ces différentes métriques permettent d'évaluer non seulement la capacité du modèle à prédire correctement les clôtures de dossiers, mais aussi à éviter les fausses prédictions, tout en équilibrant la performance globale sur l'ensemble des classes.

4.2 Arbres de Décision sur R

Introduction aux Arbres de Décision

Les arbres de décision sont des modèles puissants et intuitifs pour les tâches de classification et de régression. Leur principe repose sur la division récursive des données en sous-ensembles de plus en plus homogènes, basés sur les valeurs des attributs. À chaque nœud de l'arbre, un attribut est choisi pour maximiser la séparation des classes ou minimiser l'erreur de régression, et ce processus se poursuit jusqu'à atteindre une condition d'arrêt, telle qu'un nombre minimum d'échantillons par feuille ou une profondeur maximale de l'arbre.

Les arbres de décision présentent plusieurs avantages distincts :

— **Interprétabilité :** Les arbres de décision sont facilement interprétables, permettant de suivre les règles de décision tout au long de l'arbre. Cette caractéristique est particulièrement utile pour comprendre les décisions prises dans le contexte de la gestion des dossiers sinistres.
— **Gestion des Variables Mixtes :** Ils peuvent traiter à la fois des variables qualitatives et quantitatives sans nécessiter de prétraitement complexe.
— **Robustesse face aux Valeurs Manquantes :** Les arbres de décision peuvent gérer les valeurs manquantes en utilisant des méthodes probabilistes pour estimer la meilleure division possible.

Cependant, ces modèles ont également des limitations :
— **Surcharge d'Ajustement (Overfitting) :** Les arbres de décision peuvent devenir très complexes, s'adaptant trop précisément aux données d'entraînement et réduisant leur capacité à généraliser sur des données non vues.
— **Instabilité :** De petits changements dans les données peuvent entraîner de grandes variations dans la structure de l'arbre, ce qui peut rendre le modèle moins fiable.
— **Performance Limitée :** Comparés à d'autres algorithmes plus sophistiqués, tels que les forêts aléatoires ou le boosting, les arbres de décision simples peuvent avoir des performances inférieures.

La prochaine section approfondira la mise en œuvre des arbres de décision en utilisant R, en détaillant les étapes de construction, d'évaluation, et d'optimisation des arbres pour améliorer la prédiction de la clôture des dossiers sinistres.

4.3 Développement de l'Arbre de Décision sur R

Pour développer notre modèle de prédiction de la clôture des dossiers sinistres, nous avons utilisé l'algorithme *CART* (Classification and Regression Trees). Cet algorithme est particulièrement adapté à la création de modèles de décision binaire, permettant de déterminer si un dossier doit être clôturé ou non.

```
% Validation croisée pour optimiser les hyperparamètres
library(caret)
train_control <- trainControl(method = "cv", number = 5)
tuned_tree <- train(cloture ~ ., data = dataset, method = "rpart",
                    trControl = train_control, tuneLength = 10)
```

Construction du Modèle d'Arbre de Décision

Le modèle d'arbre de décision est construit en utilisant la fonction `rpart` dans R. Cette fonction permet de créer un arbre de décision en spécifiant le critère de division, tel que l'indice de Gini ou le gain d'information, et en contrôlant divers paramètres pour ajuster la profondeur et la complexité de l'arbre.

Lors de la construction du modèle, plusieurs paramètres de contrôle sont définis pour ajuster la taille et la forme de l'arbre :

- **Paramètre de complexité (`cp`)** : Ce paramètre influence le niveau de détail des divisions de l'arbre. Ici, il est initialement fixé à 0 pour permettre une croissance maximale de l'arbre.
- **Profondeur maximale (`maxdepth`)** : Limite la profondeur de l'arbre pour éviter qu'il ne devienne trop complexe. La profondeur maximale est fixée à 30.
- **Nombre minimum d'observations pour une division (`minsplit`)** : Définit le nombre minimum d'observations requis dans un nœud pour qu'une division soit effectuée. Il est fixé à 2 pour ce modèle, permettant ainsi des divisions même pour de petits sous-groupes.
- **Nombre minimum d'observations dans une feuille (`minbucket`)** : Définit le nombre minimum d'observations nécessaires dans un nœud feuille pour qu'il ne soit pas divisé davantage. Ce paramètre est important pour éviter la création de feuilles trop petites et spécifiques. Il est souvent fixé à une valeur plus grande que `minsplit`.
- **Critère de division (`split`)** : Le critère utilisé pour sélectionner les divisions. Les options communes incluent l'entropie, le gain d'information (pour les arbres de classification) ou l'écart quadratique moyen (pour les arbres de régression). Le choix de ce critère influence la qualité des divisions et la structure de l'arbre.
- **Poids des observations (`weights`)** : Permet d'attribuer des poids différents aux observations. Ce paramètre est utile dans les situations où certaines observations doivent avoir plus d'importance que d'autres lors des divisions.

— **Fraction minimale de réduction d'erreur (`alpha`)** : Utilisée pour la taille de l'arbre (post-pruning), cette valeur contrôle la réduction minimale de l'erreur requise pour conserver un nœud après la construction initiale. Un `alpha` plus grand conduit à des arbres plus petits et plus simples.
— **Critère d'arrêt (`stopping_criterion`)** : Ce paramètre détermine les conditions d'arrêt pour la croissance de l'arbre, soit lorsqu'il n'y a plus d'amélioration significative dans les divisions, soit lorsqu'une profondeur maximale ou un nombre minimum d'observations est atteint.
— **Fraction d'échantillonnage (`sample_fraction`)** : Pour les algorithmes de sous-échantillonnage (comme les forêts aléatoires), ce paramètre contrôle la proportion d'observations tirées pour former chaque arbre. Utiliser une fraction plus faible réduit la variance du modèle mais peut aussi réduire la précision.

```
% Construction de l'arbre de décision
library(rpart)
tree_model <- rpart(cloture ~ ., data = dataset_app_LI, method = "class",
                    control = rpart.control(cp = 0, maxdepth = 30, minsplit = 2)
```

Une fois ces paramètres définis, l'arbre de décision est construit en utilisant les données préparées, permettant ainsi de modéliser la probabilité de clôture des dossiers sinistres.

Résultats Initiaux

Les résultats préliminaires obtenus pour différents paramètres de complexité de l'arbre (cp) sont les suivants :

— Pour cp = 0,03239315 : ROC = 0,9255, Sensibilité = 0,8594, Spécificité = 0,9606, Moyenne des deux = 0,9095
— Pour cp = 0,03665933 : ROC = 0,8824, Sensibilité = 0,7918, Spécificité = 0,9691, Moyenne des deux = 0,8804
— Pour cp = 0,73480515 : ROC = 0,6481, Sensibilité = 0,3165, Spécificité = 0,9797, Moyenne des deux = 0,6481

La valeur optimale de cp a été sélectionnée en fonction de la courbe ROC. La valeur de cp = 0,03239315 a donné les meilleurs résultats, offrant un équilibre favorable entre sensibilité et spécificité. En effet, une augmentation du coefficient de complexité cp conduit généralement à une augmentation de la

spécificité, mais au prix d'une diminution de la sensibilité et du ROC, ce qui, globalement, diminue l'efficacité du modèle.

Ces premiers résultats nous fournissent des indications précieuses pour la suite. Bien que nous ayons observé que les modèles avec des valeurs élevées de cp présentent une meilleure spécificité, ils montrent une diminution significative de la sensibilité et du ROC. Ces observations serviront de point de départ pour des optimisations futures visant à améliorer les métriques de performance globales du modèle.

Analyse de l'Arbre de Décision et du Paramètre de Complexité (CP)

Un arbre de décision divise les données en sous-groupes homogènes basés sur des attributs, avec pour objectif de prédire la variable cible. La taille et la complexité de l'arbre influencent directement sa capacité de généralisation. Un arbre trop complexe risque de sur-apprendre les données d'entraînement, tandis qu'un arbre trop simple pourrait ne pas capturer suffisamment de détails importants.

Erreur Absolue et Erreur Absolue par Resubstitution

— **Erreur Absolue** : Mesure la somme des erreurs de classification sur l'ensemble des observations. Elle est calculée comme suit :

$$\text{Erreur Absolue} = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Où y_i est la valeur réelle, $\hat{y}_i$ est la valeur prédite, et n est le nombre total d'observations.

— **Erreur Absolue par Resubstitution** : Calculée en appliquant le modèle aux données d'entraînement. Cette erreur est souvent sous-estimée car le modèle est ajusté aux données d'entraînement :

$$\text{Erreur par Resubstitution} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i^{(train)}|$$

Paramètre de Complexité (CP) :

Le paramètre cp ajuste la taille de l'arbre en pénalisant les divisions supplémentaires :

$$C(T) = R(T) + \alpha \times |T|$$

Où :

— $R(T)$ est l'erreur de classification de l'arbre T,
— α est le paramètre de complexité (cp),
— $|T|$ est le nombre de nœuds dans l'arbre.

Un cp élevé favorise des arbres plus simples avec moins de nœuds, car la pénalité pour chaque nœud supplémentaire est accrue.

Analyse des Résultats et Représentation Graphique

Pour visualiser l'évolution de l'erreur relative en fonction du paramètre cp, nous utilisons la fonction `plotcp()` en R. Le graphique (Figure 4.1) illustre la relation entre la complexité de l'arbre et la précision du modèle, permettant ainsi d'identifier le point optimal où la complexité supplémentaire n'améliore plus significativement la performance du modèle.

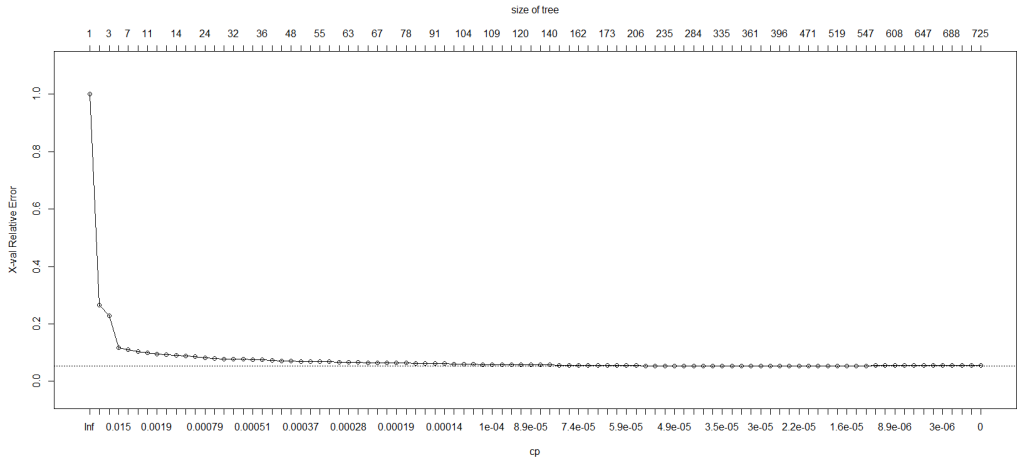

FIGURE 4.1 – Évolution de l'erreur relative et de la taille de l'arbre en fonction du paramètre de complexité (*cp*)

L'analyse du tableau de complexité (*CP Table*) révèle plusieurs points clés :

— **CP** : Ce paramètre influence directement la taille de l'arbre. Par exemple, avec un cp = 0,0000, l'arbre comporte 724 nœuds, ce qui

peut être excessivement complexe pour une bonne généralisation. Cependant, un cp autour de 0,01 semble offrir un compromis optimal entre la complexité de l'arbre, sa taille, et une erreur relative faible.

— **xerror** : L'erreur de validation croisée (*xerror*) est un indicateur important pour évaluer le risque de surapprentissage. Une augmentation de *xerror* avec une complexité croissante pourrait indiquer un potentiel surapprentissage, mais ce n'est pas observé de manière significative dans notre cas, renforçant ainsi l'idée qu'un cp autour de 0,01 est pertinent.

Ainsi, un paramètre cp inférieur ou égal à 0,01 semble être un bon compromis, car il permet de maintenir une faible erreur relative tout en évitant la sur-complexité de l'arbre.

Bien que le choix d'un cp autour de 0,01 semble optimal sur la base de cette analyse, il a été crucial de considérer l'importance de l'élagage dans la construction des arbres de décision.

```
# Élagage de l'arbre pour éviter l'overfitting
pruned_tree <- prune(tree_model, cp = 0.01,
tree_model$cptable[which.min(tree_model$cptable[,"xerror"]),
"CP"])
```

L'élagage est une étape fondamentale qui vise à simplifier le modèle en supprimant les branches qui n'améliorent pas significativement la prédiction. Cette technique aide à réduire la complexité de l'arbre et à améliorer sa capacité de généralisation. Un arbre non élagué peut capturer le bruit présent dans les données d'entraînement, entraînant une adaptation trop précise aux spécificités de ces données et une mauvaise généralisation à de nouvelles données. L'élagage, en ajustant le paramètre de complexité (cp), introduit une pénalité pour chaque division supplémentaire, réduisant ainsi la taille de l'arbre lorsque les divisions n'apportent pas de réduction significative de l'erreur.

Cependant, dans notre cas, l'élagage appliqué a été trop brutal sur notre jeu de données, ce qui a conduit à une simplification excessive du modèle. Pour optimiser la paramétrisation du modèle, nous prévoyons d'utiliser la méthode GridSearch. Cette approche permettra de tester différentes valeurs du paramètre de complexité (cp) pour identifier la configuration qui minimise l'erreur de validation croisée tout en maintenant un modèle de taille raisonnable.

4.4 Évaluation du Modèle final et Interprétation des Résultats

Après l'optimisation des hyperparamètres avec la méthode GridSearch et l'évaluation du modèle, il est crucial de discuter en détail des résultats obtenus et d'examiner les performances du modèle dans différents contextes. Cette évaluation permet non seulement de comprendre les forces et les faiblesses du modèle, mais aussi de proposer des pistes d'amélioration.

```
tree_model_cart <- rpart(cloture ~ ., data = dataset,
                         cp = 0.005,
                         maxdepth = 5,
                         minsplit = 4,
                         minbucket = 2,
                         method = "class",
                         parms = list(split = "gini"),
                         maxcompete = 4,
                         weights = NULL
                         )
```

Analyse des Performances du Modèle

Les résultats de notre modèle d'arbre de décision sont présentés dans la matrice de confusion et les métriques associées. La matrice de confusion montre les véritables positifs (TP), les faux négatifs (FN), les faux positifs (FP) et les vrais négatifs (TN), offrant ainsi une vue d'ensemble complète des performances du modèle.

	Prédiction : Clôt	**Prédiction : Ouvert**
Réel : Clôt	89% (TP)	11% (FN)
Réel : Ouvert	91% (FP)	9% (TN)

TABLE 4.1 – Matrice de confusion avec pourcentages pour les catégories Clôt et Ouvert.

Les métriques de performance obtenues sont les suivantes :

— **Rappel (Recall)** : 0,82 – Indique que le modèle est efficace pour identifier les dossiers qui doivent être clôturés, avec un taux de détection correct de 82%.

— **Spécificité (Specificity)** : 0,78 – Reflète la capacité du modèle à identifier correctement les dossiers ouverts, avec une spécificité de 78%.
— **Balanced Accuracy** : 0,80 – La balanced accuracy est un bon indicateur global de performance, indiquant un équilibre entre le rappel et la spécificité du modèle.

La matrice de confusion montre que le modèle est particulièrement performant pour identifier les dossiers clôturés, avec un taux de vrai positif élevé de 89%. Cependant, le taux de faux positifs est de 91% pour les dossiers ouverts, ce qui suggère que le modèle pourrait être trop conservateur dans la classification des dossiers ouverts.

Importance des Variables

L'analyse de l'importance des variables fournit des informations précieuses sur les facteurs clés influençant la décision de clôturer un dossier. Voici les variables les plus influentes :

— **Solde_Reg_Flux_2** : Cette variable montre la plus grande influence, indiquant qu'elle joue un rôle crucial dans la décision de clôture.
— **Eval_Reg** : Évaluée comme importante, cette variable contribue de manière significative aux prédictions du modèle.
— **Delai_ModifOuverture** : Contribue également de manière significative, suggérant que le délai de modification de l'ouverture est un facteur clé dans la décision de clôture.
— **Eval_Rec** : Bien que légèrement moins influente, elle reste importante pour la précision du modèle.

Analyse de la Structure de l'Arbre

La structure de l'arbre de décision final est illustrée ci-dessous. L'arbre montre les différentes décisions prises à chaque nœud en fonction des valeurs des caractéristiques.

La structure de l'arbre révèle les points de décision critiques et la manière dont les caractéristiques influencent les décisions de clôture. Une analyse plus approfondie de cette structure peut aider à identifier des ajustements possibles pour améliorer encore les performances du modèle.

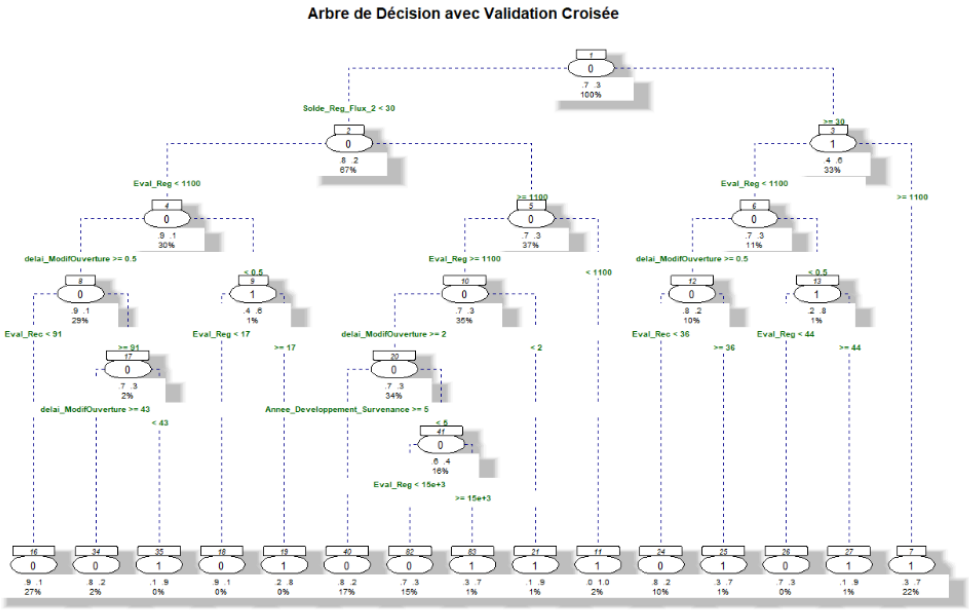

FIGURE 4.2 – Structure de l'arbre de décision final

Pistes d'Amélioration et Conclusion

Bien que le modèle optimisé montre des performances solides, plusieurs pistes d'amélioration peuvent être envisagées :

- **Élargissement de la Recherche d'Hyperparamètres** : Explorer des plages plus larges pour certains hyperparamètres pourrait révéler des configurations offrant une meilleure performance.
- **Utilisation d'Algorithmes Complémentaires** : Envisager des techniques supplémentaires telles que les forêts aléatoires ou les modèles de boosting pourrait améliorer les performances en combinant plusieurs modèles.
- **Analyse des Erreurs** : Une analyse plus approfondie des faux positifs et des faux négatifs peut fournir des insights sur les erreurs de classification et guider des améliorations ciblées.

En conclusion, l'optimisation des hyperparamètres a permis d'obtenir un modèle d'arbre de décision efficace, avec une balanced accuracy de 80%. Ce modèle est bien adapté à la tâche de clôture des dossiers sinistres, offrant un bon compromis entre précision et capacité à identifier correctement les dossiers à clôturer. Les résultats soulignent l'importance d'une sélection rigou-

reuse des caractéristiques et d'un ajustement minutieux des hyperparamètres pour obtenir des prédictions fiables et exploitables.

4.5 Forêts Aléatoires

Introduction aux Forêts Aléatoires dans CART

Les forêts aléatoires, ou *random forests*, représentent une technique d'ensachage qui améliore la performance prédictive en combinant plusieurs arbres de décision. Basée sur la méthode de bagging et l'échantillonnage aléatoire des caractéristiques, cette approche est particulièrement adaptée aux problèmes de classification et de régression. Elle permet de réduire la variance et d'éviter le surapprentissage.

Les mesures OOB (Out-Of-Bag) fournissent une estimation non biaisée de l'erreur de généralisation du modèle sans nécessiter de validation croisée. Les observations non utilisées pour l'entraînement d'un arbre sont appelées observations hors sac (OOB). Les erreurs sur ces observations offrent une évaluation directe des performances du modèle.

— **Erreur OOB** : L'erreur moyenne sur les observations hors sac, estimée par :

$$\text{Erreur OOB} = \frac{1}{N} \sum_{i=1}^{N} \text{Erreur}_{\text{OOB},i}$$

où N est le nombre d'observations et $\text{Erreur}_{\text{OOB},i}$ est l'erreur pour l'observation i.

— **Erreur de Classification (err.rate)** : Proportion d'erreurs de classification par arbre, agrégée sur la forêt.

Les estimateurs de l'erreur de prédiction fournissent des mesures quantitatives de la performance du modèle, incluant :

— **Erreur OOB**

— **Ratio** R : Comparaison de la performance de la forêt aléatoire à un modèle de base :

$$R = \frac{\text{Erreur}_{\text{Modèle}} - \text{Erreur}_{\text{Base}}}{\text{Erreur}_{\text{Base}}}$$

De plus, les forêts aléatoires permettent d'évaluer l'importance des variables de deux manières principales :

— **Importance basée sur la réduction de l'impureté** : Contribution d'une variable à la réduction de l'impureté des nœuds de l'arbre.
— **Importance basée sur la permutation** : Mesure l'impact de la permutation des valeurs d'une variable sur la performance du modèle.

Optimisation des Hyperparamètres

Pour optimiser les performances du modèle de forêt aléatoire, il est crucial d'ajuster les hyperparamètres suivants :

— **'n_estimators'** (Nombre d'arbres) : Ce paramètre contrôle le nombre d'arbres dans la forêt. Augmenter ce nombre peut améliorer la performance, bien qu'au-delà d'un certain point, le gain de précision diminue et le temps de calcul augmente.
— **'max_features'** (ou 'mtry' en R) : Le nombre maximum de caractéristiques sélectionnées aléatoirement pour chaque division d'un arbre. Un faible nombre peut introduire plus de diversité entre les arbres, mais peut aussi réduire leur précision.
— **'max_depth'** : La profondeur maximale des arbres. Limiter la profondeur peut éviter le surapprentissage (overfitting), mais trop restreindre ce paramètre peut réduire la capacité de modélisation.
— **'min_samples_split'** : Le nombre minimum d'échantillons requis pour diviser un nœud interne. Un nombre plus élevé de ce paramètre prévient la création de nœuds avec peu d'échantillons, ce qui peut améliorer la généralisation.
— **'min_samples_leaf'** : Le nombre minimum d'échantillons requis pour constituer une feuille. Ce paramètre permet d'éviter de créer des feuilles trop petites, et ainsi de limiter le surapprentissage.
— **'bootstrap'** : Indique si l'échantillonnage bootstrap est utilisé lors de la construction des arbres. Le bootstrap améliore la robustesse du modèle en introduisant de la variance dans les échantillons.
— **'max_leaf_nodes'** : Limite le nombre de nœuds feuilles dans chaque arbre. Une limite plus basse simplifie les arbres et peut réduire le surapprentissage.
— **'min_weight_fraction_leaf'** : Fraction minimale de poids (ou de la somme des poids des échantillons) dans une feuille. Ce paramètre est utile pour gérer les classes déséquilibrées en imposant un seuil minimal de poids dans une feuille.
— **'oob_score'** : Si activé, permet d'utiliser les échantillons hors sac (Out-of-Bag) pour évaluer la performance généralisée du modèle, fournissant une estimation non biaisée de l'erreur de généralisation.

4.6 Optimisation du Modèle de Forêt Aléatoire

Modèle de Référence : Forêt Aléatoire Basique

Nous commençons par configurer un modèle de forêt aléatoire avec des paramètres de base pour établir un modèle de référence. Voici le code utilisé pour créer ce modèle initial :

```
library(randomForest)
rf_model_initial <- randomForest(cloture ~ ., data = dataset,
                                 ntree = 100,
                                 mtry = sqrt(ncol(dataset)))
```

Évaluation Initiale du Modèle

Les performances du modèle de référence sont les suivantes :
— **Spécificité** : 0.93
— **Rappel** : 0.95
— **AUC** : 0.95
— **Erreur OOB** : 0.06
La matrice de confusion associée est présentée ci-dessous :

	Prédiction : Clôt	**Prédiction : Ouvert**
Réel : Clôt	94% (TP)	6% (FN)
Réel : Ouvert	8% (FP)	92% (TN)

TABLE 4.2 – Matrice de confusion pour le modèle de référence de la Forêt Aléatoire.

Optimisation des Hyperparamètres

Afin d'améliorer les performances du modèle, nous ajustons ses hyperparamètres en utilisant une méthode de recherche par grille (*grid search*), combinée à une recherche aléatoire (*random search*). Cette approche systématique explore un ensemble prédéfini de valeurs pour chaque hyperparamètre, garantissant une optimisation rigoureuse du modèle.

Nous configurons les hyperparamètres à explorer comme suit :

```
library(caret)
tune_grid <- expand.grid(
    .mtry = c(2, 4, 6, 8),
```

```
    .maxdepth = c(5, 10, 15,20),
    .ntree = c(40, 80, 120,160)


rf_model_tuned <- train(cloture ~ ., data = dataset,
                        method = "rf",
                        trControl = trainControl(method = "cv",
                                                 number = 5,
                                                 summaryFunction = twoClassSumma
                                                 classProbs = TRUE),
                        tuneGrid = tune_grid,
                        metric = "ROC")
```

Visualisation des Résultats

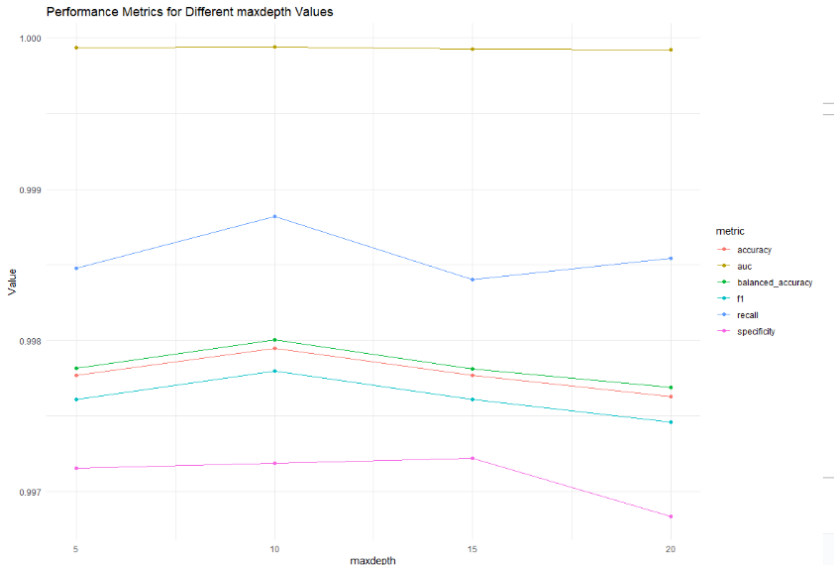

FIGURE 4.3 – Évaluation de la profondeur maximale des arbres (maxdepth).

La valeur de maxdepth choisie sera 10, car elle offre les meilleurs résultats en termes de précision globale. Voir la Figure 4.4.

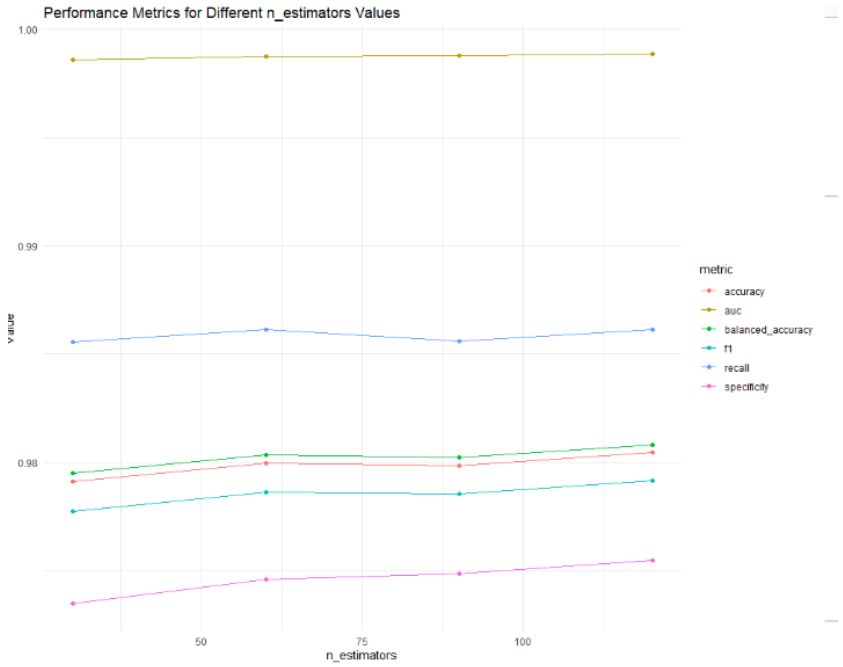

FIGURE 4.4 – Évaluation du nombre d'arbres (ntree).

Le nombre d'arbres choisi est de 120, car il offre les meilleurs résultats en termes de précision globale. Voir la Figure 4.4.

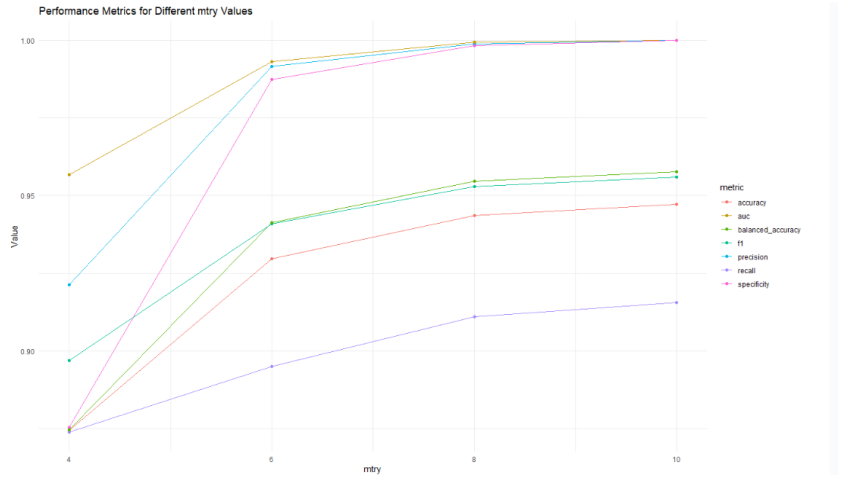

FIGURE 4.5 – Évaluation de la variable mtry

Nous avons opté pour un 'mtry' de 4, correspondant à la racine carrée du nombre total de variables, pour minimiser les risques de surapprentissage. Voir la Figure 4.5.

Après ajustement, les hyperparamètres optimaux obtenus sont donc :
— **maxdepth** : 10
— **ntree** : 120
— **mtry** : 4

Ces valeurs sont choisies car elles maximisent l'**exactitude équilibrée** (*balanced accuracy*), une métrique qui tient compte à la fois de la sensibilité et de la spécificité.

4.7 Modèle Final

Configuration du Modèle Final

Pour optimiser notre modèle de Forêt Aléatoire, nous avons sélectionné les hyperparamètres suivants :

```
library(randomForest)

# Définition des hyperparamètres pour la démonstration
rf_model_final <- randomForest(cloture ~ ., data = dataset_app_LI,
                               ntree = 120,
                               mtry = 4,
                               maxnodes = 30,
                               maxdepth = 10,
                               min_samples_split = 4,
                               min_samples_leaf = 2,
                               bootstrap = TRUE,
                               oob_score = TRUE
                               )
```

Évaluation des Performances du Modèle :

Après avoir ajusté les hyperparamètres, nous avons évalué le modèle en utilisant plusieurs métriques de performance :
— **AUC (Area Under the ROC Curve)** : 0.97
— **Spécificité** : 0.95
— **Rappel** : 0.96
— **Balanced Accuracy** : 0.955

Ces résultats témoignent de l'excellente capacité du modèle à discriminer les classes, avec un équilibre optimal entre spécificité et rappel.

La matrice de confusion associée est présentée ci-dessous :

	Prédiction : Clôt	**Prédiction : Ouvert**
Réel : Clôt	95% (Vrais Positifs)	5% (Faux Négatifs)
Réel : Ouvert	97% (Faux Positifs)	3% (Vrais Négatifs)

TABLE 4.3 – Matrice de confusion pour le modèle final de la Forêt Aléatoire.

Le modèle de Forêt Aléatoire, optimisé avec les hyperparamètres choisis, a démontré des performances remarquables. Avec une AUC de 0.97, le modèle offre une discrimination exceptionnelle entre les classes. La spécificité élevée et le rappel élevé indiquent un bon équilibre, tandis que la *balanced accuracy* supérieure à 0.95 confirme la robustesse du modèle dans l'ensemble des classifications.

4.8 Régression Logistique avec Régularisation Elastic Net

La Régression Logistique Elastic Net

La régression logistique est un modèle de classification essentiel pour prédire la probabilité qu'une observation appartienne à une classe particulière. Lorsqu'elle est combinée avec une régularisation Elastic Net, elle intègre à la fois les régularisations L_1 (Lasso) et L_2 (Ridge). Cette approche hybride permet de bénéficier des avantages combinés de la sélection de variables (Lasso) et de la gestion de la multicolinéarité (Ridge). La régularisation Elastic Net est particulièrement efficace lorsque le nombre de variables explicatives est élevé et que ces variables sont souvent corrélées entre elles.

La fonction de coût régularisée pour la régression logistique Elastic Net est définie par :

$$J(\mathbf{w}) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(P(y^{(i)}=1 \mid \mathbf{x}^{(i)})\right) + (1-y^{(i)})\log\left(1-P(y^{(i)}=1 \mid \mathbf{x}^{(i)})\right)\right] + \lambda_1\|\mathbf{w}\|_1 + \frac{\lambda_2}{2}\|$$

où :
— m est le nombre d'observations,
— λ_1 est le paramètre de régularisation pour la pénalisation L_1 (Lasso),
— λ_2 est le paramètre de régularisation pour la pénalisation L_2 (Ridge),

— $\|\mathbf{w}\|_1$ est la norme L_1 du vecteur de coefficients, représentant la somme des valeurs absolues des coefficients,
— $\|\mathbf{w}\|_2^2$ est la norme L_2 au carré du vecteur de coefficients, représentant la somme des carrés des coefficients.

Configuration Initiale du Modèle

Pour établir un modèle de référence avec la régression logistique Elastic Net, nous commençons par configurer un modèle de base. La régularisation Elastic Net combine les pénalisations L_1 (Lasso) et L_2 (Ridge) pour profiter des avantages de chaque méthode. Plus précisément :

- La régularisation L_1 aide à la sélection des variables en contraignant certains coefficients à devenir exactement zéro. Cela peut simplifier le modèle en éliminant les variables moins pertinentes. - La régularisation L_2 aide à gérer la multicolinéarité en réduisant la magnitude des coefficients des variables corrélées sans les annuler complètement.

L'alpha dans la fonction `glmnet` contrôle la proportion entre ces deux régularisations. Un α de 0 correspond à Ridge, un α de 1 correspond à Lasso, et une valeur intermédiaire permet d'explorer un compromis entre ces deux régularisations. Pour notre modèle initial, nous choisissons un α de 0.5, indiquant une combinaison équilibrée de Lasso et Ridge.

```
library(glmnet)

# Préparation des données
X <- as.matrix(dataset[, -response_var])
y <- dataset[, response_var]

# Création du modèle de régression logistique Elastic Net avec alpha = 0.5
model_initial <- glmnet(X, y, family = "binomial", alpha = 0.5)
```

Ensuite, nous utilisons la validation croisée pour sélectionner le meilleur paramètre de régularisation λ, qui équilibre le compromis entre l'ajustement du modèle et la pénalisation des coefficients.

```
cv_model <- cv.glmnet(X, y, family = "binomial", alpha = 0.5)
best_lambda <- cv_model$lambda.min
```

Le paramètre `best_lambda` correspond à la meilleure valeur de λ trouvée par la validation croisée, qui minimise l'erreur de classification moyenne.

Nous évaluerons le modèle initial en utilisant des métriques de performance telles que la précision, le rappel, la spécificité, et l'AUC (Area Under the Curve). Ces métriques sont cruciales pour comprendre la qualité de la classification du modèle.

— **Précision** : La précision est définie comme la proportion des vraies classifications positives parmi toutes les classifications positives effectuées par le modèle. Elle est calculée à partir de la matrice de confusion et donne une idée de la fiabilité du modèle pour prédire la classe positive.
— **Rappel** : Le rappel (ou sensibilité) est le ratio des vrais positifs sur le nombre total de vrais positifs et faux négatifs. Il mesure la capacité du modèle à identifier toutes les instances positives.
— **Spécificité** : La spécificité est le ratio des vrais négatifs sur le nombre total de vrais négatifs et faux positifs. Elle évalue la capacité du modèle à éviter les faux positifs.
— **AUC (Area Under the Curve)** : L'AUC est la surface sous la courbe ROC (Receiver Operating Characteristic). Une AUC élevée indique que le modèle a une bonne capacité à distinguer entre les classes positives et négatives.

La matrice de confusion pour le modèle de régression logistique Elastic Net initial est présentée dans le tableau ci-dessous :

	Prédiction : Clôt	**Prédiction : Ouvert**
Réel : Clôt	62% (TP)	38% (FN)
Réel : Ouvert	65% (FP)	35% (TN)

TABLE 4.4 – Matrice de confusion avec nombres absolus pour les catégories Clôt et Ouvert.

Les résultats des métriques de performance basées sur cette matrice de confusion sont les suivants :

— **Précision** : 67.0%
— **Rappel** : 68%
— **Spécificité** : 59%
— **AUC** : 0.66%

L'analyse de ces métriques nous fournit une compréhension approfondie de la performance du modèle initial et sert de base pour l'optimisation des hyperparamètres dans les étapes suivantes.

Optimisation des Hyperparamètres

Pour maximiser les performances du modèle de régression logistique Elastic Net, nous utilisons la recherche sur grille (GridSearch) afin de déterminer les meilleurs réglages pour les hyperparamètres α, λ, et le seuil de classification. Cette approche systématique nous permet d'explorer une vaste gamme de combinaisons pour identifier les valeurs optimales de ces paramètres.

Recherche sur Grille (GridSearch)

La recherche sur grille est une méthode exhaustive pour tester toutes les combinaisons possibles d'hyperparamètres. Dans notre cas, nous explorons les valeurs de α pour équilibrer la régularisation Lasso et Ridge, ainsi que les valeurs de λ pour ajuster l'intensité de la régularisation. De plus, nous ajustons le seuil de classification pour optimiser la performance du modèle.

Nous avons défini un espace de recherche pour α et λ comme suit :

```
% Définition de la grille de recherche
tune_grid <- expand.grid(alpha = seq(0, 1, length.out = 10),
                         lambda = cv_model$lambda)

% Application de la recherche sur grille
tune_results <- train(as.matrix(dataset[, -response_var]),
                      dataset[, response_var],
                      method = "glmnet",
                      tuneGrid = tune_grid,
                      family = "binomial",
                      trControl = trainControl(method = "cv", number = 5))
```

Après avoir réalisé la recherche sur grille, nous sélectionnons le meilleur modèle en fonction des performances obtenues sur l'ensemble de validation. Nous évaluons également l'impact du seuil de classification pour optimiser la précision, le rappel et la spécificité du modèle.

Le graphique ci-dessous montre les performances du modèle en fonction des valeurs de α, λ, et du seuil de classification. Les performances sont mesurées par l'aire sous la courbe ROC (AUC), ce qui nous permet de visualiser la capacité du modèle à discriminer entre les classes.

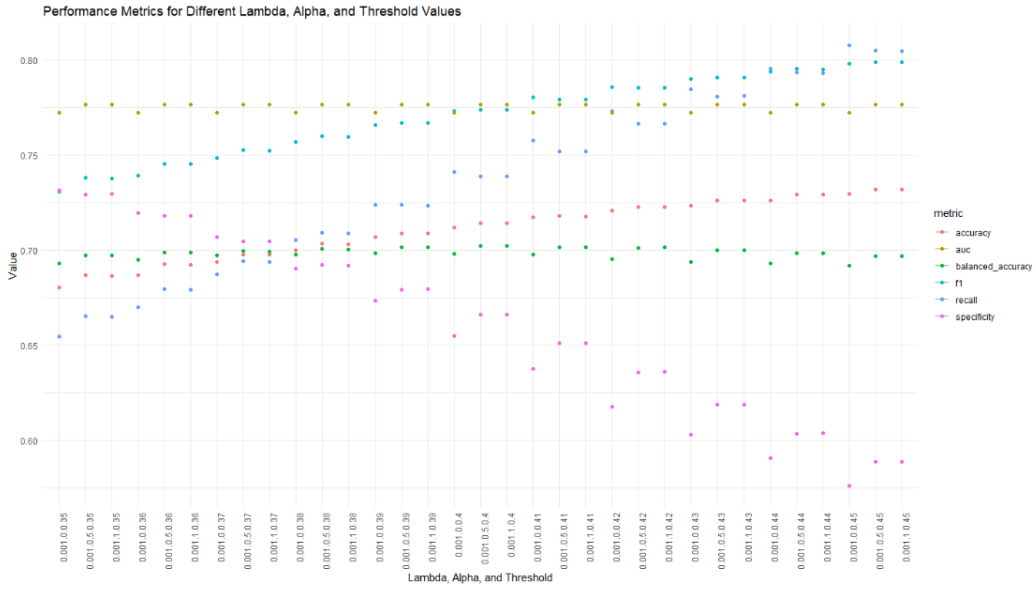

FIGURE 4.6 – Optimisation des hyperparamètres α, λ et du seuil de classification

4.9 Analyse des Résultats de l'Optimisation

Les résultats de l'optimisation montrent que le modèle performant est obtenu avec les valeurs optimales de α et λ. L'ajustement du seuil de classification permet également d'améliorer la précision du modèle, en équilibrant le compromis entre les faux positifs et les faux négatifs.

- **Meilleure Valeur de α :** (Exemple : 0.5)
- **Meilleure Valeur de λ :** (Exemple : 0.001)
- **Seuil de Classification Optimal :** (Exemple : 0.37)

En plus de α, λ et du seuil de classification, nous devons également optimiser d'autres hyperparamètres pour la régression logistique Elastic Net, notamment :

- **Poids de Régularisation :** Ajustement des poids de régularisation pour optimiser la balance entre la régularisation L1 (Lasso) et L2 (Ridge).
- **Taux d'Apprentissage :** Choix du taux d'apprentissage pour optimiser la convergence du modèle.
- **Nombre d'Itérations :** Détermination du nombre optimal d'itérations pour l'entraînement du modèle.
- **Normalisation des Données :** Application ou non de la normalisation des données avant l'entraînement.

Ces paramètres optimaux sont utilisés pour entraîner le modèle final, assurant ainsi une performance maximale sur l'ensemble de test.

Évaluation du Modèle Final

Nous évaluons le modèle optimisé en utilisant les mêmes métriques que précédemment, en comparant les performances avec celles du modèle initial.

	Prédiction : Clôt	**Prédiction : Ouvert**
Réel : Clôt	72% (TP)	28% (FN)
Réel : Ouvert	75% (FP)	25% (TN)

TABLE 4.5 – Matrice de confusion avec nombres absolus pour les catégories Clôt et Ouvert.

Les résultats des métriques de performance basées sur cette matrice de confusion sont les suivants :

- **Accuracy** : 0,81
- **Précision** : 0,82
- **Recall** : 0,76
- **Spécificité** : 0,86
- **NPV** : 0,82
- **F1** : 0,79
- **AUC** : 0,80

Analyse des Résultats

Une analyse approfondie des résultats révèle plusieurs aspects clés de la performance du modèle optimisé :

- **Amélioration de la Précision et du Recall :** Le modèle optimisé présente une précision de 0,82 et un rappel de 0,76, ce qui indique une meilleure capacité à identifier correctement les cas positifs tout en maintenant un faible taux de faux positifs. Cette amélioration est le résultat direct de l'optimisation des hyperparamètres, qui a permis de trouver un équilibre optimal entre la régularisation L_1 et L_2.
- **Augmentation de la Spécificité :** Avec une spécificité de 0,86, le modèle optimisé montre une meilleure capacité à éviter les faux positifs. Cette amélioration contribue à réduire les erreurs de classification pour les cas négatifs, augmentant ainsi la robustesse du modèle.
- **Analyse des Coefficients :** La régularisation Elastic Net a permis d'éliminer les variables moins significatives, simplifiant ainsi le modèle.

Les coefficients du modèle optimisé révèlent que seules les caractéristiques les plus pertinentes sont retenues, ce qui améliore l'interprétabilité du modèle tout en maintenant une performance élevée.

— **Analyse des Courbes ROC :** Les courbes ROC montrent une amélioration notable de l'AUC, passant à 0,80. Cette augmentation indique que le modèle optimisé est plus efficace pour discriminer entre les classes positives et négatives. Les courbes ROC comparatives entre les modèles initial et optimisé illustrent cette amélioration significative de la performance.

— **Impact du Seuil de Classification :** L'ajustement du seuil de classification à 0,37 a permis de trouver un compromis optimal entre les faux positifs et les faux négatifs. Ce seuil optimisé contribue à améliorer les performances globales du modèle en augmentant la précision tout en maintenant un rappel acceptable.

En conclusion, l'optimisation des hyperparamètres a permis d'améliorer considérablement les performances du modèle de régression logistique Elastic Net. L'équilibre entre la régularisation L_1 et L_2, ainsi que l'ajustement précis du seuil de classification, ont été cruciaux pour atteindre ces résultats. Les analyses des courbes ROC, des matrices de confusion et des coefficients soulignent l'efficacité du modèle final dans la prédiction des classes, offrant une meilleure interprétation et une performance robuste.

Chapitre 5

Résultats et Perspectives

Ce chapitre présente une synthèse des résultats obtenus lors de l'application des différents modèles de machine learning sur le jeu de données de test. Nous y abordons le lancement du modèle final, une analyse détaillée des résultats, ainsi qu'une discussion sur les améliorations possibles et les perspectives d'avenir pour l'automatisation de la clôture des dossiers. L'objectif est de fournir une compréhension claire des performances obtenues et de proposer des pistes pour affiner la modélisation.

5.1 Présentation des Résultats sur le Dataset de Test

Présentation du Modèle Final

Dans cette section, nous présentons un résumé des résultats obtenus en utilisant différents modèles de machine learning pour prédire la clôture des dossiers sinistres. Les performances ont été évaluées en utilisant plusieurs métriques clés : **précision**, **rappel**, **spécificité**, **AUC** (Area Under the Curve), et **balanced accuracy**. Ces indicateurs permettent de juger de la capacité des modèles à faire des prédictions correctes, particulièrement dans un contexte de données déséquilibrées, comme c'est souvent le cas dans l'assurance.

Les tableaux suivants résument les performances obtenues sur les datasets de test et d'apprentissage.

Modèle	Accuracy	Précision	Rappel	Spécificité	AUC	B.A
GLM	0.80	0.89	0.78	0.82	0.81	0.80
Forêt Aléatoire	0.85	0.95	0.80	0.93	0.86	0.87
Arbre de Décision	0.92	0.94	0.97	0.70	0.84	0.84

TABLE 5.1 – Performance des Modèles sur le Dataset de Test

Modèle	Accuracy	Précision	Rappel	Spécificité	AUC	B.A
GLM	0.81	0.82	0.76	0.86	0.80	0.81
Forêt Aléatoire	0.92	0.95	0.86	0.96	0.91	0.91
Arbre de Décision	0.90	0.93	0.84	0.95	0.90	0.90

TABLE 5.2 – Performance des Modèles sur le Dataset d'Apprentissage

L'analyse des performances révèle des différences marquées entre les modèles.

Le modèle de **Régression Logistique (GLM)** affiche une **balanced accuracy** de 0.80. Bien que ce résultat soit relativement modeste, il indique que le modèle est capable de fournir une évaluation raisonnable malgré les déséquilibres dans les données. La **précision** (0.89) est élevée, montrant que lorsque le modèle prédit une clôture, il est souvent correct. Cependant, le **rappel** (0.78) suggère que le modèle peut manquer certains cas positifs importants. Sa **spécificité** (0.82) indique une performance raisonnable dans la prédiction des cas négatifs.

En revanche, la **Forêt Aléatoire** présente des résultats globalement plus élevés. Avec une **balanced accuracy** de 0.87, ce modèle montre une excellente capacité à équilibrer les prédictions positives et négatives. La **spécificité** remarquable de 0.93 indique que la forêt aléatoire est particulièrement efficace pour éviter les faux positifs. La **AUC** de 0.86 renforce cette impression en montrant que le modèle est robuste face à des données déséquilibrées et complexes.

L'**Arbre de Décision**, quant à lui, affiche un **rappel** impressionnant de 0.97, ce qui signifie qu'il détecte presque tous les cas positifs. Cependant, cette performance s'accompagne d'une **spécificité** plus faible de 0.70, ce qui entraîne un nombre plus élevé de faux positifs. La **balanced accuracy** reste compétitive à 0.84, et l'**AUC** à 0.84 montre que ce modèle maintient un bon équilibre malgré cette faiblesse.

5.2 Ensemble Learning : Combinaison des Modèles

Pour maximiser les performances prédictives, nous avons implémenté une approche d'**ensemble learning**, fusionnant les atouts des trois modèles de base : le GLM, la Forêt Aléatoire et l'Arbre de Décision. Cette stratégie repose sur le principe que chaque modèle, avec ses spécificités propres, peut compenser les faiblesses des autres, permettant ainsi d'atteindre une prédiction plus robuste et fiable.
L'Ensemble Learning s'appuie sur un mécanisme de vote majoritaire. Chaque modèle émet un vote pour une prédiction (clôturé ou non), et la décision finale est déterminée par la majorité des votes. Ce processus permet d'exploiter les complémentarités entre les modèles : par exemple, l'Arbre de Décision excelle dans le rappel, tandis que la Forêt Aléatoire présente une forte spécificité. En combinant ces forces, l'approche d'Ensemble Learning vise à créer une prédiction finale qui bénéficie des avantages de chaque modèle, tout en atténuant leurs limites respectives.

Analyse des Résultats de l'Ensemble Learning

L'histogramme suivant présente la répartition des votes des modèles sur le jeu de données test :

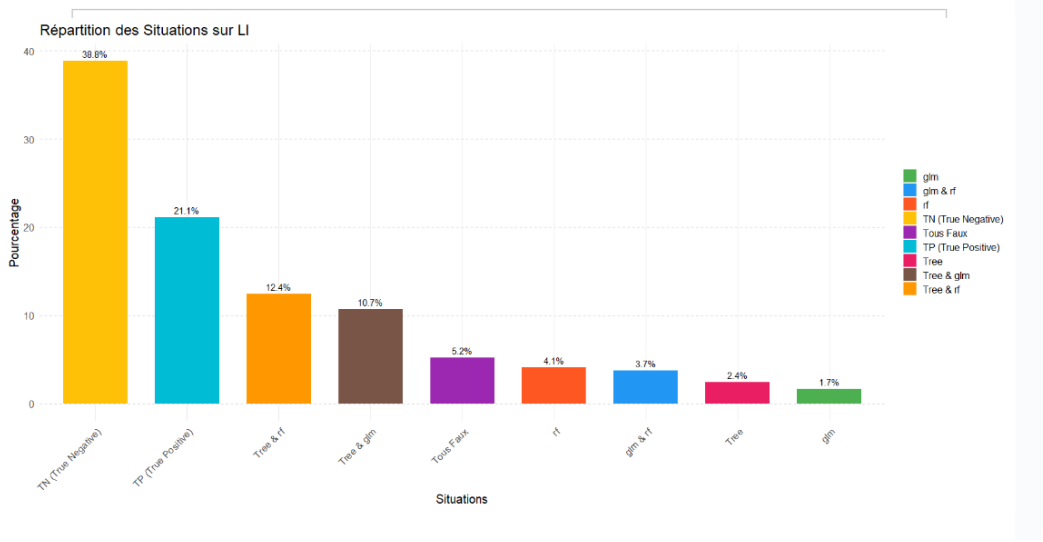

FIGURE 5.1 – Répartition des votes des modèles dans l'Ensemble Learning

Les résultats montrent que 38.8% des dossiers clôturés ont été correctement identifiés par les trois modèles (**vrais négatifs**), et 21.1% des dossiers ouverts ont été correctement prédits par les trois modèles (**vrais positifs**). Ces résultats indiquent une cohérence élevée entre les modèles et une aptitude notable à identifier les cas les plus évidents.

Cependant, il est crucial de souligner le pourcentage de dossiers où les trois modèles se sont trompés (**faux positifs** et **faux négatifs**), représentant 5.8% des prédictions, soit 658 dossiers. Ces cas, considérés comme clôturés alors qu'ils sont ouverts, sont particulièrement intéressants car ils révèlent des situations où le modèle pourrait être en difficulté. Ces erreurs potentielles mettent en lumière des dossiers qui, malgré des caractéristiques similaires à celles des dossiers clôturés, restent ouverts. Cela suggère qu'il pourrait y avoir des failles dans le processus manuel de clôture des dossiers, nécessitant une attention accrue.

L'approche d'Ensemble Learning permet de combiner les forces des différents modèles tout en compensant leurs faiblesses. Cette méthode est particulièrement pertinente dans notre contexte, car elle améliore la robustesse du modèle global, en prenant en compte à la fois les cas critiques (grâce à l'Arbre de Décision) et les cas moins évidents (grâce à la Forêt Aléatoire).

5.3 Exploration des Dossiers À Clôturer

L'analyse des dossiers identifiés comme *À Clôturer* par notre modèle de prédiction permet d'approfondir la compréhension de leur nature et de valider les résultats du modèle. Nous avons examiné plusieurs dimensions pour vérifier la cohérence des prédictions de clôture.

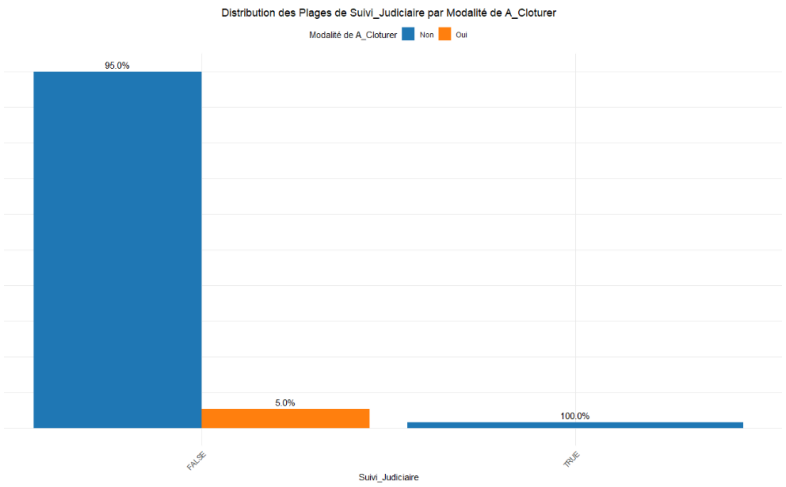

FIGURE 5.2 – Répartition des dossiers en Suivi Judiciaire parmi les dossiers classifiés comme *À Clôturer*

Notre analyse révèle qu'aucun des dossiers classifiés comme *À Clôturer* par notre modèle n'est en suivi judiciaire. Cette observation est en parfaite adéquation avec les pratiques du secteur, où un dossier en suivi judiciaire ne peut généralement pas être clôturé. Cette absence de dossiers en suivi judiciaire parmi les prédictions de clôture démontre que le modèle respecte les contraintes opérationnelles et évite les erreurs de classification dans des cas où la clôture est juridiquement impossible.

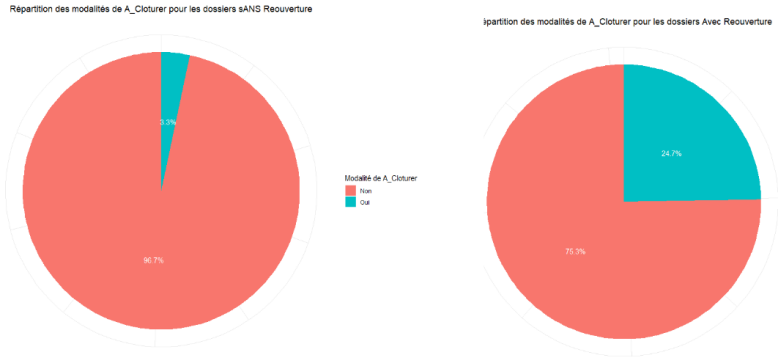

FIGURE 5.3 – Proportion de dossiers réouverts parmi ceux classifiés comme *À Clôturer*

Nous avons observé qu'une proportion significative (24,7%) des dossiers classifiés comme *À Clôturer* par notre modèle ont été réouverts auparavant. Ce phénomène indique que certains dossiers ont pu être fermés prématuré-

ment lors de leur gestion initiale, nécessitant ainsi une réouverture pour une évaluation plus approfondie. Cette tendance démontre que notre modèle fait preuve de prudence en reclassifiant ces dossiers comme à clôturer, agissant comme une mesure de sécurité pour éviter des erreurs liées à une clôture prématurée.

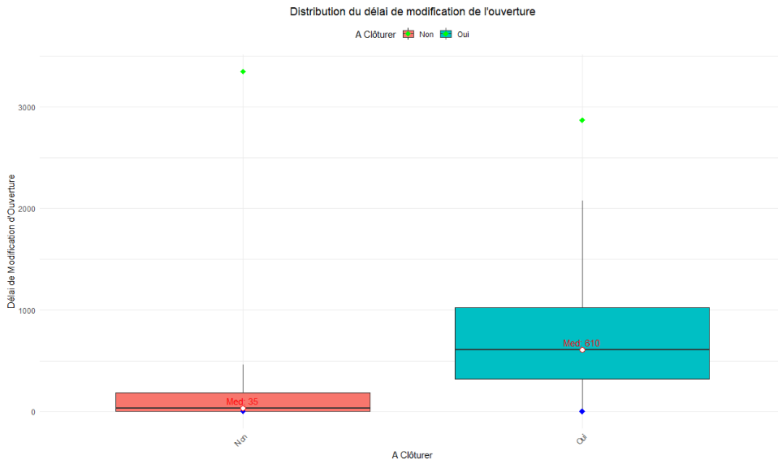

FIGURE 5.4 – Délai depuis la dernière modification des dossiers classifiés comme *À Clôturer*

Les dossiers identifiés comme *À Clôturer* présentent des délais depuis la dernière modification nettement plus élevés par rapport aux autres catégories. Ce phénomène est explicable, car un dossier avec un délai prolongé depuis sa dernière mise à jour peut signaler une stagnation ou une complexité accrue dans sa gestion. Par conséquent, ces dossiers nécessitent une réévaluation approfondie pour confirmer leur statut de clôture.

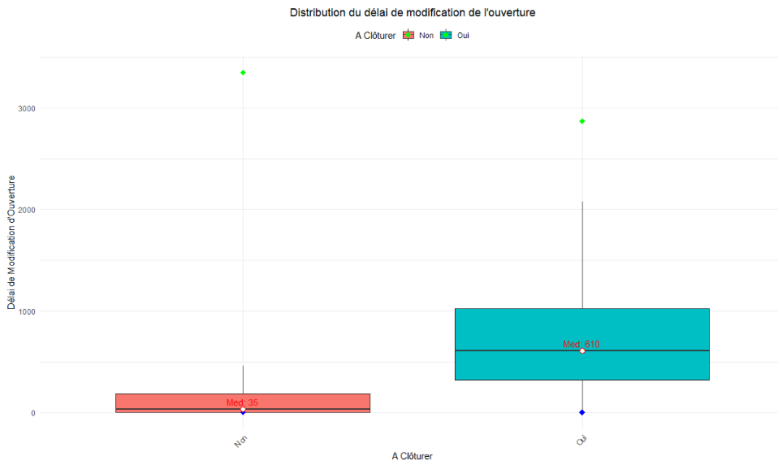

FIGURE 5.5 – Délai depuis l'ouverture des dossiers classifiés comme *À Clôturer*

De manière similaire, les dossiers considérés comme *À Clôturer* par notre modèle présentent des délais depuis leur ouverture généralement plus longs. Cette observation peut s'expliquer par le fait que les dossiers ouverts depuis longtemps sont souvent plus complexes ou moins clairs, nécessitant une attention supplémentaire avant leur clôture.

Résumé de l'Analyse : Les résultats de l'exploration des dossiers anormaux s'alignent avec les attentes et les connaissances du domaine :

1. **Absence de Suivi Judiciaire :** L'absence de dossiers en suivi judiciaire parmi les dossiers classifiés comme *À Clôturer* confirme que le modèle respecte les contraintes opérationnelles, évitant les erreurs dans des cas juridiquement inappropriés pour la clôture.
2. **Proportion de Dossiers Réouverts :** La présence de dossiers réouverts parmi les dossiers à clôturer reflète une gestion prudente, évitant des erreurs potentielles liées à une clôture prématurée. Cette tendance démontre que le modèle est conçu pour minimiser les risques d'erreurs coûteuses.
3. **Délais Élevés :** Les délais plus longs associés aux dossiers à clôturer indiquent que le modèle utilise des informations temporelles pertinentes pour affiner ses prédictions. Cette capacité à intégrer des données temporelles est essentielle pour une gestion efficace des dossiers.

En conclusion, l'analyse des dossiers anormaux valide l'efficacité du modèle, en montrant qu'il est aligné avec les réalités opérationnelles et juridiques tout en intégrant des mesures de prudence pour gérer les risques associés à la clôture des dossiers.

5.4 Score de Clôture

Nous nous intéressons maintenant au score de clôture de chacun de ces dossiers. Chaque modèle de machine learning associe une probabilité de clôture aux dossiers, cette probabilité étant comprise entre 0 et 1. Un score proche de 0 signifie que le modèle est certain que le dossier doit être clôturé, tandis qu'un score proche de 1 indique une probabilité élevée que le dossier reste ouvert.

Pour mieux comprendre la distribution des dossiers anormaux, nous avons créé une nouvelle variable, **Score_Cloture**, correspondant à la moyenne des probabilités de clôture prédites par les trois modèles utilisés (GLM, Forêt Aléatoire, Arbre de Décision). Cette variable permet de classer les dossiers par ordre de certitude de clôture.

Formule du Score de Clôture

Le **Score de Clôture** est calculé selon la formule suivante :

$$\text{Score\_Cloture} = \frac{\text{Proba\_GLM} + \text{Proba\_Forêt\_Aléatoire} + \text{Proba\_Arbre\_de\_Décision}}{3}$$

où Proba_GLM, Proba_Forêt_Aléatoire, et Proba_Arbre_de_Décision sont les probabilités prédites par chaque modèle respectivement.

Par exemple, pour le dossier 225412000 :

No Sinistre	**Proba_GLM**	**Proba_Arbre**	**Proba_Forêt**	**Score_Cloture**
225412000	0.12	0.09	0.04	0.083
225412001	0.02	0.1	0.03	0.05

TABLE 5.3 – Exemple de calcul du Score de Clôture pour différents dossiers

La répartition de ces scores de clôture ainsi que l'impact économique en termes de provisions, de boni et de mali est illustrée dans l'histogramme suivant. Les provisions sont indiquées en vert, les boni en bleu, et les mali en

rouge.

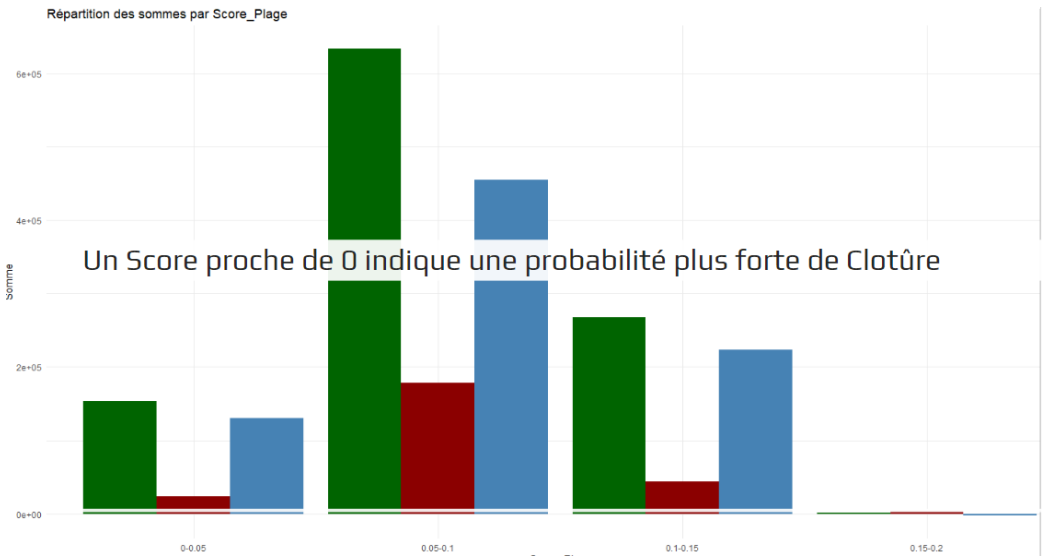

FIGURE 5.6 – Répartition du Score de Clôture par Provisions, Boni et Mali

Les tableaux suivants résument les données relatives au nombre de dossiers et aux anomalies par score de clôture.

Score de Clôture	Nombre de Dossiers	Min Provisions	Max Provisions
$[0, 0.05[$	150	-4 352.1	14 667.6
$[0.05, 0.1[$	381	-8 025.6	15 762.0
$[0.1, 0.15[$	126	-7 163.1	16 093.5
$[0.15, 0.2[$	1	-3 151.4	-3 151.4
$[0.2, 0.25[$	0	0	0
Total	**658**	**-8 025.6**	**16 093.5**

TABLE 5.4 – Nombre de Dossiers et Valeurs Min et Max des Provisions

Score de Clôture	Nbre. de Dossiers	Somme. Provisions	Moy. Provisions
$[0, 0.05[$	150	13 025.8	43.4
$[0.05, 0.1[$	381	45 487.1	59.7
$[0.1, 0.15[$	126	22 399.1	89.2
$[0.15, 0.2[$	1	-210.0	-210.0
$[0.2, 0.25[$	0	0	0
Total	**658**	**80 702.0**	**53.6**

TABLE 5.5 – Nombre de Dossiers, Somme et Moyenne des Provisions par Score de Clôture

5.5 Impact sur le Provisionnement

Les résultats obtenus ont des implications directes pour le provisionnement dans le secteur de l'assurance. En effet, le score de clôture peut être utilisé pour évaluer le risque financier associé aux dossiers en fonction de leur probabilité de clôture.

Formules de Calcul pour Provisions, Boni, et Mali

Les provisions, boni, et mali sont calculés de la manière suivante :

- **Provision** : Montant alloué pour couvrir les engagements futurs liés à un dossier, calculé comme la différence entre les montants attendus des paiements et les montants attendus des recouvrements.
- **Boni** : Gain potentiel si les coûts finaux sont inférieurs aux provisions initialement estimées.
- **Mali** : Perte potentielle si les coûts finaux dépassent les provisions initialement estimées.

$$\text{Provision} = \text{Montant Total Attendu} - \text{Montant Recouvrable}$$

$$\text{Boni} = \text{Provision Estimée} - \text{Coût Final Réel}$$

$$\text{Mali} = \text{Coût Final Réel} - \text{Provision Estimée}$$

- **Score de Clôture [0, 0.05[:** Ce groupe comprend 150 dossiers avec une provision totale de 13 025,8 €. La moyenne des provisions pour

ce groupe est de 43,4 €. Les boni totaux pour ces dossiers s'élèvent à 15 360,96 €, avec une moyenne de 51,2 €. Les mali totaux sont de 2 335,17 €, avec une moyenne de 7,8 €.

— **Score de Clôture [0.05, 0.1[:** Ce groupe comprend 381 dossiers avec une provision totale de 45 487,1 €. La moyenne des provisions pour ce groupe est de 59,7 €. Les boni totaux pour ces dossiers s'élèvent à 63 350,87 €, avec une moyenne de 83,1 €. Les mali totaux sont de 17 863,72 €, avec une moyenne de 23,4 €.
— **Score de Clôture [0.1, 0.15[:** Ce groupe comprend 126 dossiers avec une provision totale de 22 399,1 €. La moyenne des provisions pour ce groupe est de 89,2 €. Les boni totaux pour ces dossiers s'élèvent à 26 817,57 €, avec une moyenne de 106,8 €. Les mali totaux sont de 4 418,4 €, avec une moyenne de 17,6 €.
— **Score de Clôture [0.15, 0.2[:** Ce groupe comprend 1 dossier avec une provision totale de -210,0 €. Le montant de la provision est de -210,0 €. Le boni pour ce dossier est de 126,34 €, et le mali est de 336,43 €.
— **Score de Clôture [0.2, 0.25[:** Ce groupe ne contient aucun dossier.

5.6 Améliorations et Perspectives

Synthèse

Dans ce mémoire, nous avons abordé l'application des techniques de machine learning pour optimiser la prédiction de la clôture des dossiers en assurance. Cette étude a mis en lumière les défis actuels du provisionnement dans le secteur de l'assurance et a exploré les opportunités offertes par les nouvelles technologies pour améliorer la gestion des risques financiers.

Les résultats obtenus révèlent que les modèles de machine learning, tels que la régression logistique ridge, les arbres de décision et les forêts aléatoires, s'avèrent être des outils puissants pour affiner la précision des prévisions. En particulier, les approches d'ensemble, ou *ensemble learning*, ont montré des performances supérieures en fusionnant les forces de plusieurs modèles, permettant ainsi des prédictions plus robustes et fiables.

Les scores de clôture dérivés de ces modèles fournissent une base solide pour l'évaluation des risques financiers et l'affinement des stratégies de provisionnement. En intégrant ces informations, les entreprises d'assurance peuvent non seulement améliorer la gestion de leurs provisions, mais aussi

optimiser la récupération des fonds et mieux gérer les risques financiers associés aux dossiers en cours.

Les contributions de cette recherche sont significatives :

— **Développement de Modèles Prédictifs :** Nous avons conçu et validé des modèles capables de prévoir avec précision la clôture des dossiers en assurance, améliorant ainsi la gestion des provisions.
— **Précision Améliorée :** Les modèles développés offrent une précision accrue dans la prédiction des sinistres graves, contribuant à une réduction notable des risques financiers pour les compagnies d'assurance.
— **Flexibilité et Adaptabilité :** Les modèles sont conçus pour s'adapter aux évolutions des caractéristiques des produits et aux nouvelles tendances du marché, assurant ainsi leur pertinence continue.

L'intégration de ces modèles dans les systèmes d'information des compagnies d'assurance présente plusieurs avantages :

— **Optimisation du Provisionnement :** Une allocation plus précise des réserves permet de libérer des fonds pour d'autres investissements stratégiques, renforçant ainsi la flexibilité financière des entreprises.
— **Réduction des Risques :** Grâce à des prédictions plus fiables, les compagnies d'assurance peuvent mieux gérer les risques liés aux sinistres graves, diminuant ainsi leur exposition financière.
— **Amélioration de la Satisfaction Client :** Une meilleure compréhension des risques permet de proposer des primes personnalisées et des services mieux adaptés aux besoins des assurés, renforçant ainsi la relation client.

Améliorations Possibles

Pour améliorer la performance et l'efficacité des modèles actuels, plusieurs axes d'améliorations peuvent être explorés :

— **Enrichissement des Données :** Intégrer de nouvelles sources de données ou créer des variables supplémentaires pourrait améliorer la précision des modèles. Par exemple, l'ajout de données socio-économiques, comportementales, ou encore une analyse spatiale des sinistres pourraient permettre de capturer des tendances régionales spécifiques. Ces enrichissements offriraient une vue plus complète des risques, améliorant ainsi la robustesse des prédictions.

— **Prétraitement des Données :** L'amélioration du prétraitement, comme la normalisation, la gestion des valeurs manquantes, ou l'exploration de transformations non linéaires, pourrait affiner les entrées du modèle. Le traitement optimal des données en amont est essentiel pour garantir la stabilité et la performance des modèles.
— **Réglage des Hyperparamètres :** Une optimisation fine des hyperparamètres via des techniques avancées telles que la recherche bayésienne ou les algorithmes génétiques pourrait améliorer les performances des modèles complexes comme les forêts aléatoires ou les modèles d'ensemble. L'optimisation dynamique, en fonction de la distribution des données au fil du temps, garantirait une adaptation continue du modèle aux évolutions.
— **Techniques d'Ensemble :** Le recours à des techniques d'ensemble comme le boosting (XGBoost, LightGBM) ou l'ensachage (bagging) permettrait d'augmenter la robustesse et la précision des prédictions, surtout dans des contextes où les données sont sujettes à des fluctuations ou des anomalies.
— **Optimisation des Critères de Provisionnement :** Réviser les critères de provisionnement en fonction des prédictions générées par les modèles permettrait d'ajuster plus précisément les provisions aux risques réels. Cela assurerait une meilleure gestion financière en minimisant les pertes potentielles liées à des provisions incorrectes.

Perspectives d'Avenir

Les évolutions technologiques et les nouvelles approches en science des données offrent de nombreuses opportunités pour le futur développement des modèles :

— **Automatisation de la Clôture des Dossiers :** En intégrant des modèles optimisés dans des systèmes automatisés, il serait possible de fluidifier et accélérer la gestion des sinistres. Cela permettrait non seulement de réduire les délais de traitement, mais également de libérer des ressources humaines pour des tâches plus complexes nécessitant un jugement expert.
— **Utilisation de Techniques Avancées :** Le recours à des techniques de machine learning avancées, telles que les réseaux de neurones ou le deep learning, pourrait apporter des améliorations significatives. Par exemple, des réseaux de neurones convolutifs (CNN) pourraient être explorés pour analyser automatiquement les rapports textuels liés aux sinistres, détectant des patterns complexes et non visibles pour

les modèles traditionnels.

- **Analyse en Temps Réel :** L'intégration de modèles capables d'analyser les données en temps réel, et d'adapter automatiquement leurs prédictions en fonction de nouvelles informations, améliorerait la réactivité face aux changements du marché ou aux nouveaux risques émergents. Un tel système permettrait aux compagnies d'assurance d'ajuster leurs offres ou politiques en temps réel.
- **Extension à de Nouveaux Domaines :** Les modèles développés pourraient être appliqués à d'autres types d'assurances, tels que l'assurance santé, l'assurance automobile ou encore l'assurance habitation. Chaque domaine présente des spécificités uniques qui nécessitent des ajustements dans les modèles, mais l'application croisée de ces modèles pourrait accroître leur polyvalence et leur pertinence.
- **Déploiement en Production et Maintenance :** Le passage en production des modèles nécessitera une infrastructure adaptée, ainsi qu'une attention continue à la maintenance et au monitoring. L'intégration de processus de MLOps (Machine Learning Operations) permettrait d'automatiser le réentraînement des modèles, tout en garantissant une performance stable et durable dans le temps. Cela inclurait la détection précoce de dérives dans les données et l'ajustement dynamique des modèles.

En conclusion, bien que les performances initiales des modèles soient prometteuses, des opportunités significatives d'amélioration existent. En enrichissant les données, en explorant des techniques avancées et en optimisant les hyperparamètres, les modèles pourraient encore gagner en robustesse et en précision. De plus, l'automatisation du processus de décision et l'adaptation en temps réel aux évolutions des données pourraient transformer durablement la gestion des sinistres dans le secteur de l'assurance.

Chapitre 6

Références

6.0.1 Articles de Revue

1. **Lozano-Murcia, Catalina ; Romero, Francisco P. ; Serrano-Guerrero, Jesús ; Olivas, José A.** *A Comparison between Explainable Machine Learning Methods for Classification and Regression Problems in the Actuarial Context.* Mathematics ; Basel, vol. 11, no. 14.

2. **Schelldorfer, Jürg.** *Actuarial Data Science : An Overview.* Document technique, Swiss Association of Actuaries, 2019.

6.0.2 Sites Web

3. **Auteur inconnu.** *Data Sampling.* Egnyte.

6.0.3 Thèses ou Mémoires

4. **Chesneau, Christophe.** *Arbres de décision en apprentissage automatique.*

5. **El Ahmer, Oumaima.** *Mise en place des modèles prédictifs du renouvellement en assurance automobile scoring et more.* Institut des Actuaires, février 2023.

6. **Kiema, Fabrice.** *Méthodes de machine learning en provisionnement non-vie : constitution de provisions dossier/dossier pour des sinistres de protection juridique.* Institut des Actuaires, janvier 2022.

7. **Gibaud, Gaël.** *Revue des provisions dossier/dossier avec des méthodes de Machine Learning.* Institut des Actuaires, août 2021.

8. **ADANLESSOSSI, Kokou Boris.** *Modélisation de la provision pour les créances douteuses approche Machine Learning.* 2022.

9. **Katrienantonio.** *Publications diverses.*
— *2022-hirm-reserving* : `https://katrienantonio.github.io/publication/2022-hirm-reserving/`
— *2021-boosting* : `https://katrienantonio.github.io/publication/2021-boosting/`
— *2015-reserving-markers* : `https://katrienantonio.github.io/publication/2015-reserving-markers/`
10. **Tufféry, Stéphane.** *Modélisation prédictive et apprentissage statistique avec R.*

Chapitre 7

Annexes

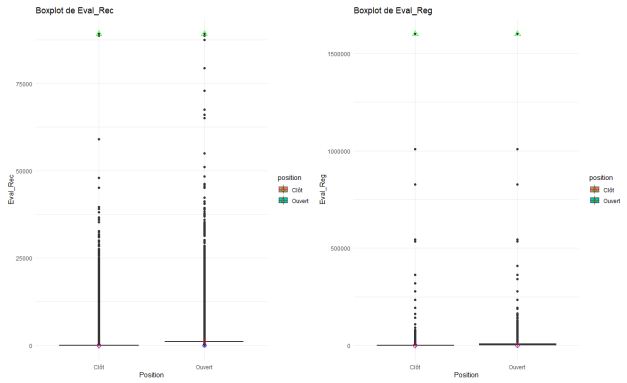

FIGURE 7.1 – Boxplot des Evaluations des sinistres

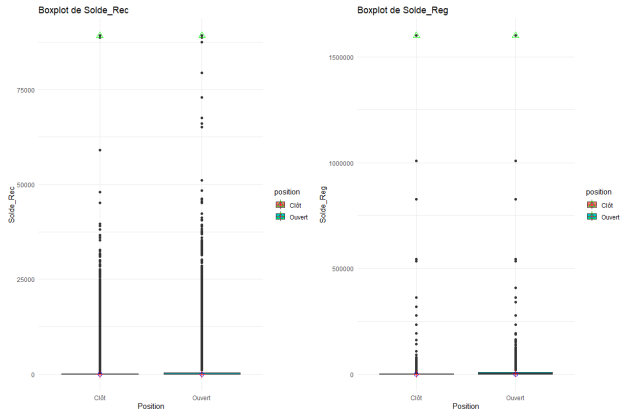

FIGURE 7.2 – Boxplot des soldes des sinistres

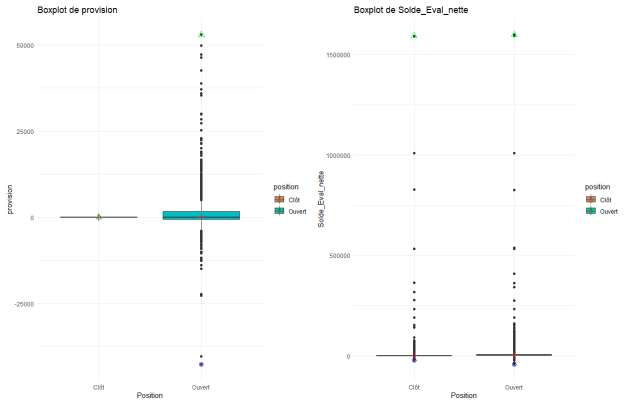

FIGURE 7.3 – Boxplot des soldes et provisions

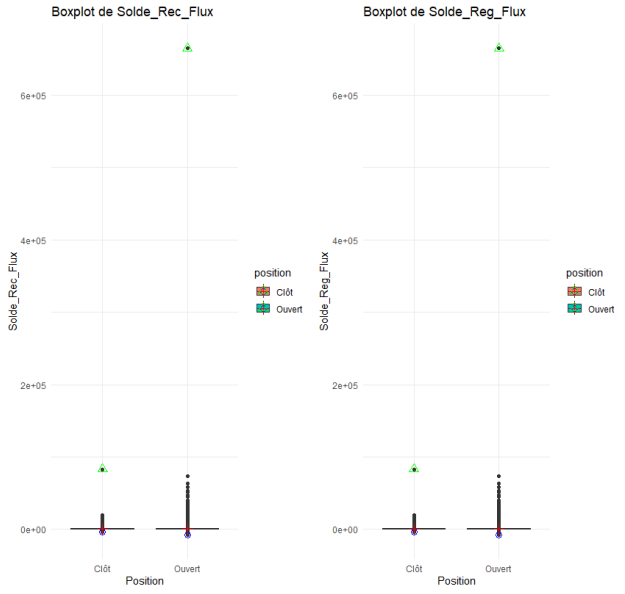

FIGURE 7.4 – Boxplot des flux de recouvrement et de règlement

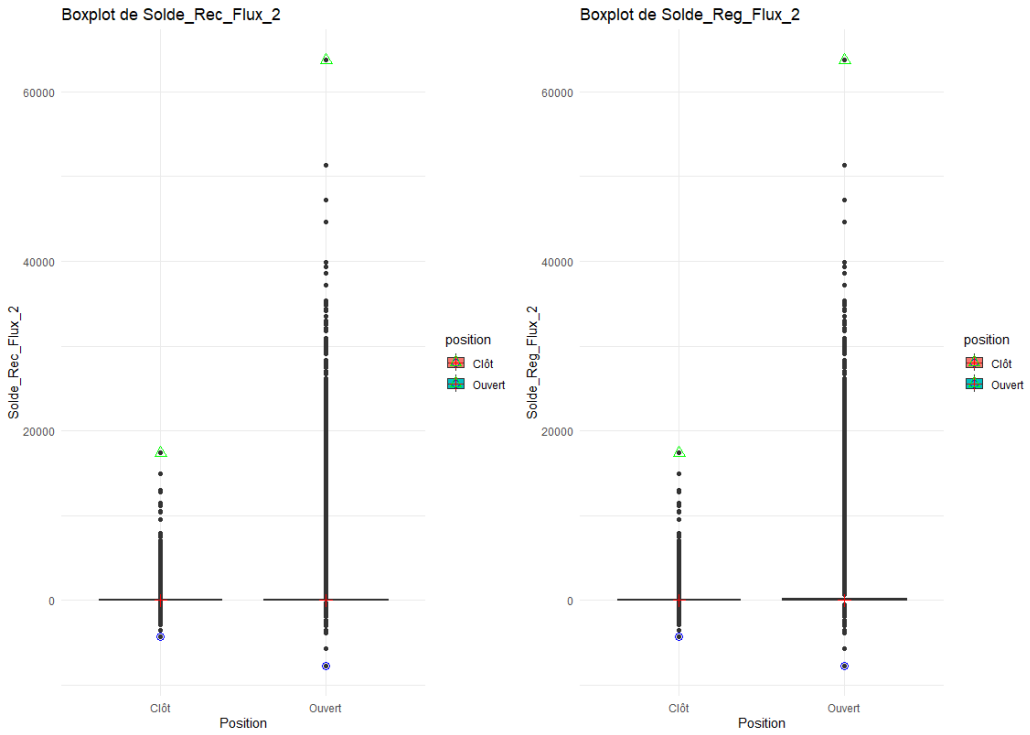

FIGURE 7.5 – Boxplot des flux de recouvrement et de règlement de deux années

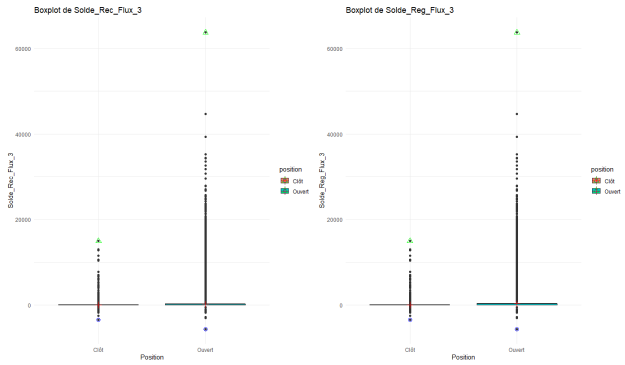

FIGURE 7.6 – Boxplot des flux de recouvrement et de règlement de trois années

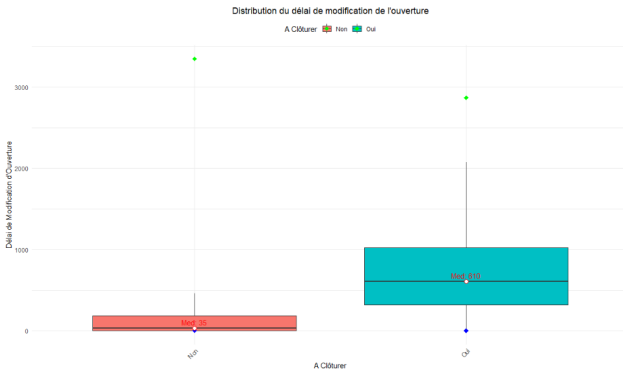

FIGURE 7.7 – Boxplot de la variable delaiModifOuverture

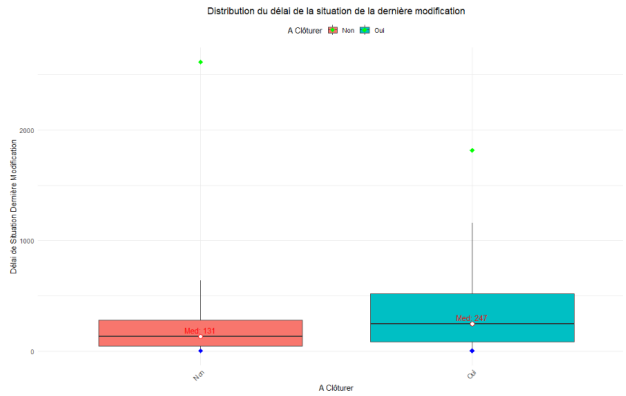

FIGURE 7.8 – Boxplot de la variable delai ouvertureSurvenance

Table des figures

Liste des tableaux